Coláiste na hOllscoile, Corcaigh

THE NATIONAL UNIVERSITY OF IRELAND, CORK

School of Microbiology

Bacterial inhabitants of tumours: Methods for exploration and exploitation

Thesis presented by

Sidney Walker

Under the supervision of

Dr. Mark Tangney and Dr. Marcus Claesson

For the degree of

Doctor of Philosophy

2016-2020

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.

_____

Sidney Walker

Abstract

The presence of bacteria in patient tumours of various types has been reported by numerous groups since 2014, but the findings of these and many similar studies remain contentious. Tumour samples provide many obstacles to carrying out robust and reliable microbial surveys, primarily the anticipated low biomass of these samples, which leaves them vulnerable to environmental contamination. While the debate over the presence or absence of bacterial communities in these tumours continues, it impedes any research into how such bacteria might be utilised in medicine. Larger sample numbers are required, from diverse tumour tissues within the human body, and these must be analysed in a reproducible and accurate manner to allow for the drawing of definitive conclusions in this debate. To accommodate this requirement, the primary methodological aspects of this thesis were: i) The assembly and validation of a contamination control pipeline using recent advances is bioinformatic contamination control detection. ii) The development and validation of a bacterial DNA extraction protocol for formalin fixed, paraffin embedded (FFPE) samples, with accompanying FFPE biological standards for use as controls. A key aim of this thesis was to increase the accuracy and reproducibility of research into clinical tissue biopsies by eliminating the role of contamination, and to expand the applicability of FFPE tissues which represent an invaluable resource of samples for analysis.

In this thesis, ecological surveys of a variety of related environments were conducted with the common goal of characterising a detectable bacterial community and identify potential bacterial biomarkers unique to these host environments. Regardless of whether or not a consistently present and detectable tumour microbiome exists, tumours possess several phenotypes making them hospitable environments for bacteria to colonise. Where the unique physiology of tumours is seen as an obstacle for traditional cancer treatments, they represent an opportunity for bacterial-mediated solutions. Therefore, findings from sequencing-based research of host environments have potential to be translated into the use of administered bacteria as delivery vehicles to locally produce biomolecules.

There are two considerations in this context, requiring two very different applications of bioinformatics. i) The first is to identify which bacteria colonize the desired niche in body; this can be a 'foreign' body such as a tumour (Chapter 3 and 4), or parasite (Chapter 5), or a distal niche such as the gut. ii) The second, often under-considered parameter, relates to what these bacteria produce. Synthetic biology presents enormous scope for sophisticated medical therapy mediated by novel synthetic proteins. However, the task of getting a bacterial cell to successfully express and secrete a stable protein that it does not produce naturally is far from trivial, and is becoming a key aspect of the synthetic biology field. To facilitate this synthetic protein aspect, a novel strategy for the performance prediction of designed protein constructs was developed. This tool was able to predict the overall performance of a protein construct *in vitro* using only *in silico* derived data.

Thesis aim:

This thesis aimed to develop novel strategies for the analysis of bacterial communities within tumours by i) increasing the sample sizes available to future projects by enabling the use of FFPE samples and ii) improving the accuracy of analysis by designing bioinformatics analysis pipelines appropriate for these samples. This enabled further research concerned with finding differentially present taxa between the tumour and surrounding environment, which have the potential for use as therapeutic vectors. As the key aim is to establish the presence of potential bacterial therapeutic vectors rather than to establish the role these bacteria play in tumorigenesis, this approach is easily translatable to other foreign bodies, and could therefore be validated in a parasitic nematode model.

GLOSSARY:

16S ribosomal RNA gene: Transcribed form of the small subunit gene located in the 30S subunit of a prokaryotic ribosome. Contains 9 variable regions that can be targeted for amplification and used for profiling of microbial communities.

ASV: Amplicon Sequence Variant. Refers to individual DNA sequences recovered from high throughput marker gene sequencing following the removal of spurious sequences.

BER: Base Excision Repair

DADA2: Divisive amplicon denoising algorithm

FFPE: Formalin-Fixed, Paraffin-Embedded

LASSO: Least absolute shrinkage and selection operator

MICROBIOME: The genetic material of all microorganisms that live on or in an ecological niche, such as the human body.

MICROBIOTA: The ecological community of commensal, symbiotic and pathogenic microorganisms found in and on all multicellular organisms.

NGS: Next Generation Sequencing

NMR: Nuclear magnetic resonance

OTU: Operational Toxonomic Unit. Cluster of sequences with similarity above a specified threshold, eg. 97%.

PCR: Polymerase Chain Reaction

qPCR: Quantitative Polymerase Chain Reaction

QIIME: Quantitative insights into microbial ecology

READS: DNA segments obtained from a sequencing experiment.

SNP: Single Nucleotide Polymorphism.

VARIANT CALLING: Process of identifying variants between closely related sets of sequence data, typically taking the form of SNPs.

WGS: Whole Genome Sequencing

Chapter I

Literature Review

SECTION 1: SEQUENCING AND THE MICROBIOME

A portion of this section has been submitted as ***"Bioinformatics Platforms for Metagenomics",*** currently under review with the Elsevier editorial team as part of the book ***"Comprehensive Foodomics"***

Julia Eckenberger* 1,2 , Sidney Walker* 1,2,3 , Marcus J Claesson 1,2

*Authors contributed equally

The Microbiome

Introduction to the Microbiome

The term Microbiome refers to the cumulative genetic material found within a microbiota. The Microbiota is a term used to define a community of micro-organisms living within an ecological niche. These niches range from well researched environments such as the human gastro-intestinal tract (1) or vaginal tract(2) to some of the most extreme locations where life has been found. These include the Door to Hell gas crater in the Karakum Desert of Turkmenistan, Deep-sea brine lakes in the Gulf of Mexico, and the Permafrost of Siberia(3).

The Human Microbiome

Current estimates place the size of the human microbiota at between 10 and 100 trillion microbial cells, spanning the kingdoms Bacteria, Archaea, Fungi and Viruses(4). The study of these microbial communities within the human host owes its origins to Antonie van Leewenhoek. As early as the 1680's this Dutch scientist was comparing his faecal bacteria with his oral bacteria, although calling them "animalcules" at the time(5).

Research into the human microbiome has focused predominantly on the niches outlined in below;

- Oral Microbiome(6)
- Skin Microbiome(7)
- Gastro-Intestinal tract microbiome(1)
- Urogenital tract microbiome(8)
- Nasopharyngeal tract microbiome(9)

The number of human body sites found, or suspected to harbour endogenous microbial communities is constantly increasing and now potentially includes sites such as the brain, breast tissue(10), the lungs(10) and a variety of tumour sites(11). Sites such as these typically harbour considerably lower levels of micro-organisms than the more thoroughly researched tract-based microbiomes, as such there is a risk of environmental contamination mistakenly being recognised as biological signal(12).

Human Microbiome in Cancer

The human microbiome is known or suspected to play a role in a diverse spectrum of host indications. Due to the incredible potential for therapeutic use or intervention, a plethora of studies can be found investigating possible links between specific micro-organisms, or fluctuations in the overall microbial community at a given niche and cancer. This relates in particular to the gut microbiome (13). To date, the only definitively proven causative link between a bacteria and cancers is that of *H. pylori* and gastric adenocarcinoma as proven by Barry Marshall in 1983 (14) and mucosa associated lymphoid tissue lymphoma. This has led to *H. pylori* being the only bacteria identified as a class 1 carcinogen by the World Health Organisation (13).

Research is ongoing into the role played by other bacteria in cancers, and there are several interesting prospects that encourage further research. *Fusobacterium* has been consistently found enriched in patients with colorectal cancer(15) suggesting the strong possibility of a causative link. This is supported by a potential mechanism as *Fusobacterium nucleatum* in particular has been shown to recruit tumour-infiltrating immune cells, contributing to the generation of a pro-inflammatory environment conducive to the progression of colorectal neoplasia (16). Recently published work highlights how infection with *Salmonella enterica* serovar Typhi through a cascade of events can leave individuals with a considerably elevated risk of developing gallbladder cancer. This involves both the secretion of a typhoid toxin with carcinogenic potential, in conjunction with biofilm production promoting a persistent infection(17).

In addition to the numerous studies hypothesising causal relationships between bacteria known to commonly colonise human hosts and cancer, certain protective interactions may also exist. The most interesting of these is that despite being classified as a class 1 carcinogen, *H. pylori* infection is associated with a reduced risk of Barrett's Aesophagus, but considerable follow up work is required before any medically significant conclusions can be drawn from this inverse association(18).

Intratumoural bacteria?

There have been several conflicting studies in relation to the presence of a consistently detectable tumour microbiome since the concept was first postulated (19). Since then, the number of studies claiming to have identified bacterial

communities in new tumour sites (11,20,21) has been matched evenly number of studies urging caution when seeking to characterise novel environments of low biomass, due to their susceptibility to contamination (22,23). More high-quality research, accounting for the numerous sources of error when analysing samples of this type is required before this question can be definitively answered.

Sequencing

Three generations of sequencing strategies

Moore's Law states that the number of transistors on a microchip doubles every two years, along with a halving in the overall cost (24). As sequencing technology improves in accuracy and price in accordance with this law, it is an extremely dynamic field that is difficult to precisely define. To assist in this, different sequencing strategies are grouped together in different generations of the technology, a current snapshot of the state of the art is as follows.

Sanger sequencing, developed in 1977, is referred to as the first generation of sequencing. While still widely used for some projects due to its relatively long read length, on average 650 base pairs, and high accuracy, it is not an appropriate tool for metagenomics studies due to its low throughput and relatively high cost (25).

Next generation sequencing methods are massively parallel and can often produce millions of reads during a typical run, where genomes present are sequenced repeatedly in small random fragments. The two predominant NGS methods are Ion Torrent and Illumina but differences between them in terms of cost, underlying chemistry, output and accuracy mean they are not always suited to the same tasks. The Illumina sequencing platforms provide the most popular sequencing solutions owing to their low cost, high accuracy, and high output (26). They function by synthesising the complementary strand of DNA present in a sample followed by fluorescence based detection of DNA bases. The Illumina MiSeq platform offers low sample output of up to 15Gb but relatively long reads at an affordable price. The paired end functionality offers overlapping reads of up to 300bp each, making this the technology of choice for amplification based sequencing experiments such as 16S rRNA gene sequencing. Other Illumina platforms such as the HiSeq, NextSeq and NovaSeq offer much higher output, with the NovaSeq generating up to 6000Gb,

but with shorter read lengths of up to 150bp (27). Ion Torrent sequencing platforms also sequence by synthesis of the complementary strand, but detection of the base composition relies on pH meters measuring the release of hydrogen ions when the DNA is polymerised. The sequencing run is shorter, taking only hours compared to days for Illumina technology, and the read length yielded is up to 400bp, however only single end reads are available and the low total output of up to 15Gb makes this technology impractical for anything other than amplicon sequencing, this method also has a higher error rate when long repeat regions are present in the sample (28).

In the context of WGS, the short reads generated by second generation technologies often yield incomplete genome assemblies. In more complex cases such as genomes with long repeat regions, paralogs or bacteriophages, longer reads are often required to close the genomes (29). Third generation sequencing platforms targeting individual DNA molecules, have been developed to meet this demand. The two platforms currently available are the Oxford Nanopore sequencing methods, and the Pacific Biosciences PacBio platform. Nanopore sequencers identify DNA bases based on the changes in electrical conductivity due to a DNA strand passing through a biological pore. There are a variety of Nanopore solutions available based on the number of flow cells they contain, ranging from the portable minION, designed for use in the field to the scaled up promethION with on board data processing. These Nanopore methods produce reads of up to 100kb in length (30). The PacBio platform uses single molecule real-time sequencing technology (SMRT). Similarly to the Illumina short read technologies the sequencing is synthesis based and uses fluorescent dyes, but in this instance single stranded DNA molecules are sequenced individually by being deposited in wells with immobilised DNA polymerase (31). The PacBio Sequel2 is the current state of the art technology in this respect and can sequence reads up to 60kb in length. As these third generation strategies begin to reliably upscale their output, their use in metagenomics studies will become more widespread (28). A summary of the sequencing options available is found in Table 1 below.

Table 1.1: ***Summary of major sequencing solutions and their associated performances***

Generation	Method	Output Type	Max Read Length	Throughput	Error Rate	Runtime
1st	Sanger(25,27)	SE	650-900bp	62.4kb	0.0001%	1-3 hours
2nd	Illumina(25,27)	SE and PE	75-300bp	13.2-6000	0.26-0.8	Up to 6 days
2nd	Ion Torrent(25,27)	SE	150-400b	10Gb	1.785	7.3 hours
3rd	Nanopore(25,27)	SE	Up to 100kb	0.1-1Gb	12-38%	Real time data analysis
3rd	PacBio(25,27)	SE	Up to 60kb	1Gb	11-15%	2 hours

Sequencing an ecological niche

Strategies for characterising a microbial environment are split into two distinct categories, amplification based sequencing and whole genome shotgun sequencing. These two approaches differ considerably in terms of cost, information content and sample requirements. Unfortunately there is no one size fits all solution, and the decision on which strategy to pursue should be performed on a study by study basis.

Amplification based methods

Amplicon sequencing refers to the sequencing of PCR products, obtained by a targeted amplification of a variable region of interest. In human genomics this is often carried out to test for somatic mutations in specific exons, in metagenomics marker genes are targeted to characterise complex environments that may not be fully described with a whole genome sequencing approach. Common marker genes include the 16S rRNA gene sequence, hypervariable regions of which can be used to

discriminate bacteria and some archaea(32), and the internal transcribed spacer (ITS) region for fungi(33).

The 16S rRNA gene sequence is sequence within the 30S small subunit of the ribosome, a small subunit has an integral function in mRNA translation(34). The 16S rRNA gene sequence is almost ubiquitous in bacteria, and present in many archaea, it consists of highly conserved regions that can be targeted by primers, interspersed with hypervariable regions making it an ideal genetic marker(35) for bacterial characterisation. The sequence is ~1500bp in length, and contains nine hypervariable regions varying in length and conservation(36). As most microbial surveys are carried out using Illumina technology, which have a maximum read length of 2x300bp, a subset of these hypervariable regions are usually selected for amplification and eventual sequencing. No one hypervariable region reliably outperforms all others although considerable research has gone into comparing and contrasting the effectiveness with which different regions can resolve complex bacterial communities (36,37). The level of variability in each of these 9 regions is shown in the figure below, adapted from research by Bodilis, J et al(38). The higher this variability, the better the discriminatory power between bacterial taxa.

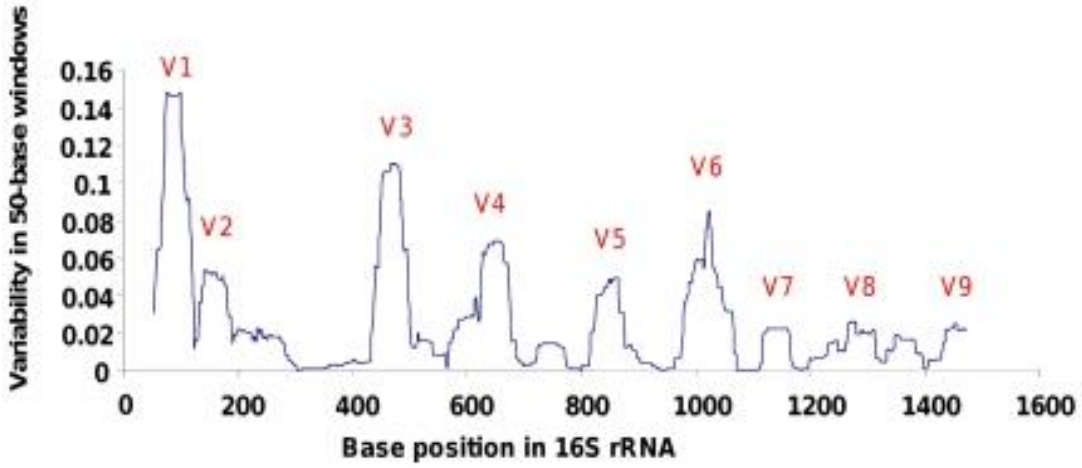


Figure 1.1: ***Variability over the length of the 16s rRNA gene sequence. Window size = 50 bases. (Adapted from (38).)***

The two most commonly used regions at present are V1-V2 and V3-V4 regions(37), although an eventual shift to sequencing the entire gene sequence with 3rd generation technology seems inevitable as studies have shown a 100-fold increase in resolution

when combining primers to cover a region spanning 80% of the 16S rRNA gene sequence(39).

WGS Sequencing

Whole genome shotgun metagenomics entails non targeted sequencing of all genetic material in a microbiome, resulting in the key difference that while marker gene studies can tell us what organisms are present, the presence of entire genomes yielded by WGS gives detailed information about metabolic potential, evolutionary relationships, and the structure and organisation of microbial genomes. Since the seminal work by Craig Venter in 2004(40) sequencing microbial populations present in the Sargasso Sea, this sequencing strategy has contributed to breakthroughs in a range of research areas. These range from bacterial associations in Inflammatory Bowel Disease(41) to tracking the outbreak of human and foodborne pathogens(42).

WGS sequencing vs 16S sequencing for characterising a bacterial community

Before deciding on which approach to take, there are advantages and disadvantages to both approaches that must be considered, which are outlined in the table below.

Table 1.2*: Comparison of 16S rRNA gene sequencing and Whole Genome Shotgun sequencing, for exploration of bacterial communities.*

	16S rRNA	WGS
Cost	Minimum 10x cheaper per sample than WGS metagenomics sequencing Analysis more accessible as less computational power needed	Prohibitively expensive for smaller labs Requires considerable processing power for most analysis platforms
Information Content	Provides data specific to amplicon used, ie. Bacteria/archaea with 16S rRNA Genus level resolution with some species level resolution possible Taxonomy only, with some	Provides data on all micro-organisms present in a sample Species and strain level resolution possible Detailed taxonomic and functional information available as full genomes

	inferred metagenomics information using tools such as PICRUST2	generated with sufficient sequencing coverage
Suitability	Amplification step can introduce bias in samples of low biomass due to presence of contaminant DNA or large quantities of host DNA in biopsies. Databases more comprehensive when characterising novel environments, as they are easier to populate	Less susceptible to bias as no amplification is required. No amplification means that in biopsies where host DNA can make up ~99% of all DNA, WGS is unsuitable without microbial enrichment.

These methods are not mutually exclusive, and often an effective trade-off is to carry out a broad analysis with an amplicon sequencing method, before proceeding to WGS metagenomics with a subset of interest. Figure 1.2 below outlines the two potential pathways for characterising a bacterial community using next generation sequencing.

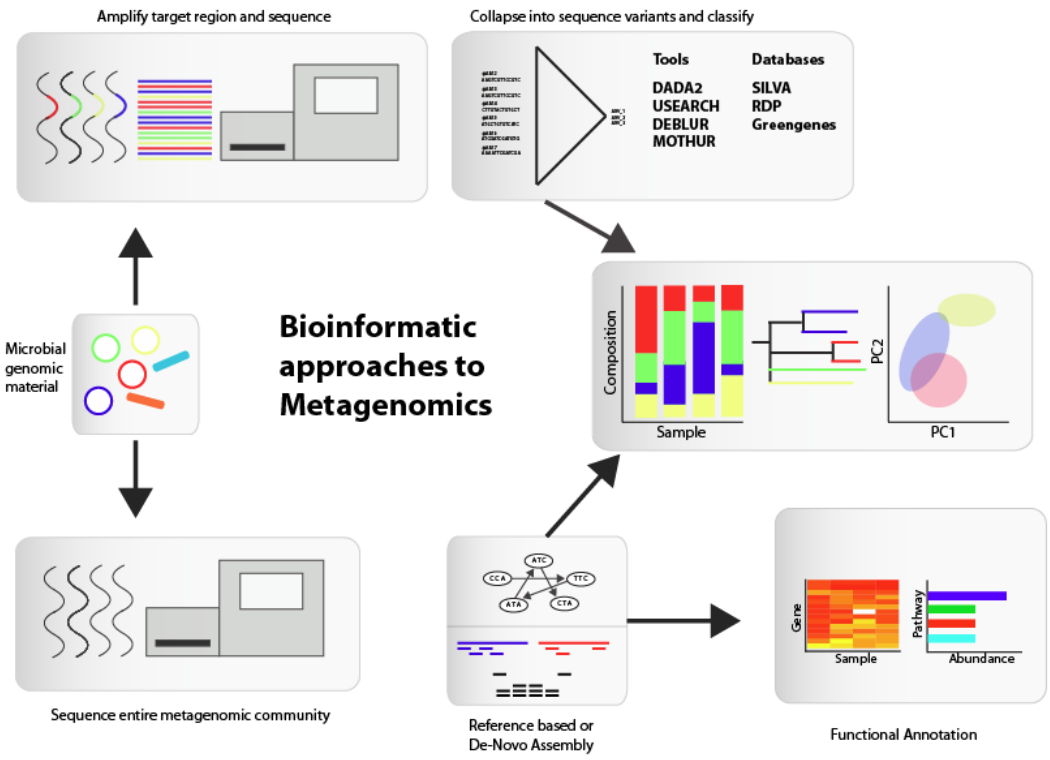


Figure 1.2: ***The divergence and convergence of WGS vs Amplicon sequencing strategies***

Analysis of Bacterial Sequence Data

Quality Filtering of sequence Data

As mentioned, the quality of sequencing data returned can be variable, so quality filtering of this data is always a crucial first step in any analysis. This is particularly important when working with 2nd generation sequence data. Although the error rate in 3rd generation technologies is higher, it remains stable across the length of the sequence, whereas 2nd generation sequencing technologies, Illumina in particular, have a quality profile that depreciates with sequence length, particularly in the reverse read of a pair(43). This makes quality filtering prior to downstream analysis a critical step in any analysis pipeline.

The typical output of a sequencing experiment is paired or unpaired reads, in the fastq format (Illumina and Ion Torrent). This format contains sequence information and a corresponding per base quality score called a Phred score. This ranges from 1-40 and is translated as -10 x log10(P), with P being the probability of a base being called erroneously. For example a Phred score of 10 means there is a 1:10 chance of

the base being called incorrectly, whereas the maximum score of 40 means a 1:10,000 probability of error.

The fastq file format, contains sequence data with machine readable error annotations for each sequenced base and serves as input for most programmes which are utilized to trim away lower quality reads, either based on manual visual analysis and trimming to a certain length, setting a minimum quality cut off, or both. Freely accessible technologies for quality filtering sequence data include Trimmomatic(44) and Fastx toolkit(45).

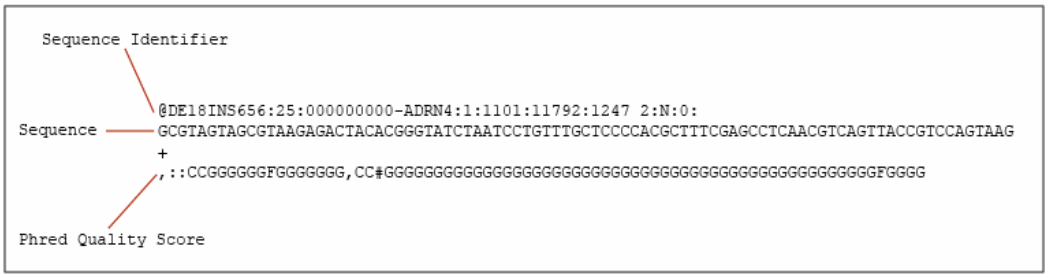


Figure 1.3: ***The FastQ format***

16S Sequence Analysis

The operational taxonomic unit, as coined in 1965(46) is generally accepted to refer to closely related groups of organisms, that are clustered into one "OTU" based on a threshold of identity, typically 97%. 16S sequence data can be collapsed into these units either by clustering sequences against a 16S reference database, in what is called "closed reference" OTU picking, or clustered based on pairwise similarity, which is called *de novo* OTU picking. While opinion is divided on which method returns the more biologically relevant taxonomic unit, the increased computational demands of "closed reference" picking means that "open reference" is the only option available to those without access to a computing cluster(47). A shortcoming that applies to OTUs regardless of which picking method is used is that a certain amount of diversity is obscured during the clustering process. The original purpose of clustering was to shield against spurious false speciation events that are a function of sequencing errors, but as this accuracy continues to increase there is an argument

that OTUs can be considered obsolete and a new method able to explore the full diversity of bacteria in an environment is more suitable.

New methods have been developed resolving sequence data into Amplicon Sequence Variants (ASVs). Each ASV directly relates to a bacterial or archaeal sequence found in the original sample, prior to amplification. These methods function on the assumption that biological sequences are more likely to be repeated than sequences containing erroneous bases. These require the constructon of an error model, built on the reads present in the sequencing run that correct errors by a process termed “denoising” (48).

The two premier facilities for generation of Amplicon Sequence Variants are The Divisive Amplicon Denoising Algorithm (DADA) 2 and DeBLUR. DADA2DADA2 first constructs an error model trained on a user defined subset of the dataset, before collapsing the reads found into ASVs. The key advantage of training the error model on the entire dataset is that it allows for the merging of different sequencing runs prior to analysis by accounting for any run bias(49). Deblur also employs a “denoising” approach facilitated by an error model. In this instance the model is constructed on a per sample basis, which significantly reduces computational memory requirements, but means the algorithm is unable to compensate for batch effects when merging samples(50).

Regardless of which unit is used to represent the genetic material present in a sample, the process of taxonomic classification remains constant. Several databases of 16S rRNA gene sequences exist, the most popular of which are SILVA(51), the Ribosomal Database Project (RDP) (52), and Greengenes(53). Of these, SILVA is the most regularly updated. Aligning sequence data to these databases would cause considerable computational bottlenecks, and to alleviate this a variety of tools exist to merge sequence data with the information provided by these databases as efficiently as possible.

Table 1.3: ***Databases available for analysis of meta-barcoding sequencing data.***

Database Name	Size	Latest revision	Target Organism
SILVA(54)	695,171 Non-Redundant, 2,090,668 Total	Dec 2017	85% Bacteria, some Archaea and Eukaryota
RDP	3,356,809 Total	Sep 2016	Comprehensive Bacteria and Fungi, with some Archaea
Greengenes(53)	92,684 Non-Redundant, 1,012,863 Total	May 2013	Bacteria and some Archaea
UNITE(55)	98,183 Non-Redundant, 1,750,261 Total	Sep 2019	Eukaryota

The *classify.seqs* tool within Mothur is a versatile taxonomic classifier. It allows the user to dictate which database is used, and also whether to employ a k-mer or k-nearest neighbour approach. The k-mer approach examines each query sequence as a collection of 8 base k-mers and assigns taxonomy based on their cumulative probabilistic classification. The k-nearest neighbour approach first finds the 10 most similar sequences to the query sequence in the selected database, and then uses thee to generate a consensus classification(56).

As mentioned, classification of 16S rRNA gene sequence reads beyond genus level is unreliable and some cases impossible. For example, *E.coli* have seven different copies of this sequence(57). SPINGO is a stand-alone tool dedicated to this challenge. It uses a customised modification of the RDP database containing 95210 sequences which represent 12394 species. This is queried using k-mer fragments of the ASVs/OTUs generated prior to this(58).

More detailed analysis of different steps and considerations involved in 16S rRNA gene sequence analysis more pertinent to the scope of this thesis is provided in section 4 of this Literature Review, entitled "Characterising intratumoural bacteria."

WGS Sequence Analysis

The expansion in applicability of whole genome shotgun sequencing has been mirrored only by the rapid advances in the number of bioinformatics platforms available for analysis of the ensuing data. Be it for aligning reads to reference genomes, assembling reads into contiguous segments of overlapping DNA (contigs), or functionally annotating a metagenome, there is a myriad of potential tools, none of which outperform all others in every circumstance. A comprehensive review of the relevant literature is recommended before undertaking any analysis of shotgun sequence data.

Whole genome sequencing (WGS) strategies fall under two umbrella terms, based on whether the target DNA is well characterised, or if novel DNA is expected. Simply put, if the sample is expected to contain previously characterised bacteria, taxonomic and functional annotation can be performed by referring to available databases. If these are not available, sequencing reads must be manually assembled into contigs using a variety of methods. There are advantages and disadvantages to both approaches that must be considered before proceeding beyond this point.

The depth of read coverage of genomes present can dictate which method to use, as 20x coverage is generally recommended as a minimum for genome assembly(28). This limits the effectiveness of assembly-based methods in analysing complex microbial communities given the current sequencing technologies available.

Genome assembly, and the associated downstream tasks are computationally intensive which can often limit the size of the study, whereas most reference based methods emphasise efficiency allowing for large scale metagenomics analyses(59).

As is indicated by the name, reference-based methods require a database which contains at a minimum closely related microbes to those found in the samples, whereas assembly-based strategies are able to resolve genomes of novel organisms(60). While some degree of manual supervision is required for reference-based assembly, the tools involved require minimal intervention when compared to the degree of curation required for accurate contig assembly and taxonomic

binning(28). The theory of metagenomics assembly is the same in principle as genomic assembly, and the same underlying principles are observed.

Assembly of metagenomics samples

Most metagenomics assemblers use a modification of the *De Bruijn Graph* approach, however as sequencing errors dramatically increase the size of the graph, and therefore the processing power required, *overlap-layout-consensus* methods have returned to the fore when assembling data from single molecule technologies such as PACBIO and Nanopore (61).

Metagenome assembly is still undoubtedly an imperfect science, and no one assembler can be relied upon to outperform all others in every situation. As such, a variety of assemblers have been designed to supplement the existing established assemblers which have added metagenome specific extensions. These include the Iterative De Bruijn Graph De Novo Assembler (IDBA) and the Saint Petersburg genome assembler, (SPAdes). IDBA-UD(62) is built as an extension on to IDBA**,** and was specifically designed to take into account the Uneven read Depth of typical metagenomics sequence data. A drawback in terms of implementing this tool is that it was originally designed for read lengths up to 100bp, and a k-mer size of 120bp. Modification of this to suit a longer fragment length requires the modification of some accessory scripts when compiling the tool, which may be beyond the casual user. MetaSPAdes(28,63) first constructs a *De Bruijn* graph of all reads available, using SPAdes. This is followed by a variety of graph simplification strategies to transform this into an assembly graph. This step constructs the paths corresponding with fragments within the genomes sequenced. Variations between highly similar contigs, potentially due to strain level variation are not considered by MetaSPAdes. The tool instead aims to assemble reliable consensus sequences giving the most accurate representation of species present, without accounting for strain level variation. Unlike the two previously mentioned tools, BBAP(64) is an overlap-consensus based genome assembler, making it more suited to sequence data with high error rates, such as those from SMRT technologies or highly polymorphic DNA.

Considering no tool consistently outperforms all others, multiple assembly methods should be attempted and compared before proceeding in the analysis.

Assembly free taxonomic profiling

Many tools combine both assembly and classification/function analysis, particularly in the case of reference-based methods. Profiling of microbial species present, and their abundance, in an environmental sample can also be carried out without prior genome assembly. Assembly-free profiling is similar in principle to the tools used for single genome alignments to a reference, but adapted for metagenomics use. This approach has a number of advantages, it is less computationally demanding and can provide information on low abundance organisms that would not be sequenced to sufficient depths for assembly(28). Also, reference based approaches generally require less manual intervention than assembly-based methods. The principal limitation is that even with improvements to the algorithms used, if the biological material present in the sample has not at the minimum had a close ancestor sequenced, they are impossible to identify (28). Despite this, the complex community structure of common sample sites such as the human gut, make assembly free profiling the more suitable method.

At a fundamental level, assembly free profiling means comparing sequence data yielded from a sequencing experiments to existing databases of micro-organism genomes in a way that is accurate and computationally efficient. In practise, this is carried out in four different ways. Sequences can be classified by sequence similarity to a reference genome, similarities in composition such as codon usage, hybrid methods that combine elements of the first two approaches, or finally, marker-based methods. These classify sequences based on presence of specific marker sequences such as the 16S rRNA gene fragment in bacteria, or the internal transcribed spacer (ITS) region in fungi (65). These tools all rely on models derived from reference sequences of existing sequenced genomes.

Simple “brute force” mapping of reads to sequenced genomes in a database leads to spurious false positives in terms of taxonomic classification. A selection of more reliable approaches are outlined below; Kraken (66) which uses as default the REFseq database hosted by NCBI extracts k-mers of default length 31 from sequence data and finds the lowest common ancestor in which this is present in the

database. It is therefore still a similarity-based method, but considerably faster than BLAST based methods. Relative synonymous codon usage (RCSU), based on the established fact that codons are differentially favoured in different organisms, is a computationally efficient way to discriminate between microbes, particularly at higher levels of taxonomic resolution (67).

MetaPhilan (68), and the recently released extension MetaPhilan2 (69) use clade specific marker genes to characterise microbial communities. As of the most recent release, this classification programme contains over oneone million clade specific markers, which equates to approximately 145 markers per bacterial species for over 7,000 commonly identified species. Additional functionality for classification of viral and eukaryotic components of metagenomics samples has also been added.

When classifying microbial reads a balance must be found between accuracy and speed. Similarity based classification methods based on BLAST are often the slowest methods, but modifications of similarity-based classification using short regions, considerably improves on this speed without sacrificing accuracy.

Functional Profiling

As with taxonomic classification, functional annotation is at its most fundamental, the process of identifying coding regions within sequenced genomes, and aligning these to a translated protein database. There are several databases containing functional information relating to genes and genomes. The Kyoto Encyclopaedia of Genes and Genomes (KEGG) (70) is an online database of genomes and genes with the primary aim of assigning functional meaning to both. The information is stored in a hierarchy of different levels as Kegg Orthology (KO) containing molecular level functional annotations, with each annotated KO being homologous to a gene or protein. Higher level functional information for a gene or protein is kept in BRITE hierarchies and KEGG pathway maps (70). The Clusters of Orthologous Groups (COGs) of proteins database is curated by clustering together orthologues from different genomes, with the hypothesis that orthologous genes can be expected to have a conserved function. Functional prediction of proteins is performed by querying which cluster the protein falls into, using the COGNITOR program (71). UniProt(72) is a vast database containing both manually annotated, curated database

of protein sequences and functional information in UniProtKB/Swiss-Prot, and a much larger database of automatically annotated records in UniProtKB/TrEMBL. This database is provided by the UniProt Consortium, which is comprised of the subsidiaries Swiss Institute of Bioinformatics, the Protein Information Resource, and the European Bioinformatics Institute (73).

Without refinement, analysis of this kind leads to considerable computational bottlenecks, particularly in a large scale metagenomics dataset. It is impossible to manually cross-reference all available data with databases of this type, intermediary programmes such as HuManN2 (74) and MEGAN (75) are therefore required to bridge the gap. MEGAN, currently in its 6th iteration has an advantage over many other metagenomics platforms in that it is compatible with all Windows, Mac and Linux operating systems. MEGAN first uses the DIAMOND alignment tool to align all reads to a database, typically the NCBI-nr database. MEGAN then takes this alignment file as a reference for binning the reads, functionally and taxonomically. The lowest common ancestor (LCA) algorithm assigns each sequenced read to the lowest taxonomic rank of common ancestor of all organisms the read in question aligns to. This is repeated for all reads in the dataset. Functional annotation is carried out by searching for the best alignment between a sequenced read, and a functionally annotated DNA sequence from one of the following databases;; SEED (76), KEGG (77), InterPro2Go (78) or eggnog (75).

HuManN2 (74) performs species level functional annotation of metagenomes and metatranscriptomes. Unlike taxonomic profiling, functional profiling quantifies the metabolic potential of a microbial community. HuManN2 uses a “tiered search” strategy to rapidly profile the functional composition of a metagenome. Initially, MetaPhIAn2 is used to identify previously characterised microbes in the sample, and constructs a database per sample, merging existing data with pan genomes of identified species. Following this, reads in a given sample are mapped against the samples pan genome database at the nucleotide level. The reads that do not align are then translated and used to query a protein database which as by default either UniRef90 or UniRef50. The alignments created by this tiered search strategy, once weighted by sequence length and quality of alignment, are used to generate per-organism and total community, gene family abundance (74).

As with any other aspect of metagenomics research, a thorough understanding of the tools available and their strengths and weaknesses is recommended.

Variant Calling

Reference based sequence assembly also allows for variant calling analysis. This is the process of identifying variants between closely related sets of sequence data, which typically take the form of single-nucleotide polymorphisms (SNPs). Variant calling analysis based on WGS data has superseded more traditional methods such as PFGE or MLST, as the level of sensitivity to small, localised changes is much higher (79).

A typical workflow for variant calling involves aligning WGS sequence data to a reference genome, creating BAM files. This is done using a genome aligner such as Bowtie(80), or Burrows-Wheeler aligner (81). Following this, differences between the aligned reads and the reference genomes are identified and written to a variant call file (VCF),,using tools within the SAMtools package (82). Lastly, this VCF file must be filtered to ensure results are significant, and not resulting from artefacts of the sequencing process also performedperformed within the SAMtools package. Lastly, this VCF file must be filtered to ensure results are significant, and not resulting from artefacts of the sequencing process. This can also be done within the SAMtools package.

Variant calling allows for the differentiation of micro-organisms at the strain level (83), which can be of crucial significance in metagenomics. For example, the fact that some members of the E.coli genus are harmless commensals and others are major pathogens such as the Shiga toxin-producing E.coli O157:H7(84) makes analysis of this nature to differentiate between them invaluable. Variant calling can be scaled upwards to process entire metagenomic datasets, and tools such as StrainPhilan(85) work off the same principle, but are tuned for the complexities of large mixed samples.

This facility of variant calling to detect minor differences or mutations between closely related sequences is also used to detect DNA damage. This is a challenging task as although most organisms expose their DNA to potential sources of DNA damage regularly, only a small proportion of the sites in a given genome are

damaged. As such it can be difficult to differentiate these from the general noise of sequencing miscalls(86). When reliably performed, variant calling can be a valuable tool for assessing DNA damage due to formalin fixation in WGS sequencing samples, and thus can be used to compare different strategies of DNA repair.

Statistical analysis of microbiome data

Metagenomics data is usually summarized as a table of read counts per OTU /ASV or gene/genome per sample. Those tables tend to be very sparse, where counts of zeroes may mean the true absence of an OTU/ASV or gene/genome or that its presence is below the detection limit. This detection limit can vary between samples due to differences between sequencing runs and an unequal representation of samples in pooled sequencing libraries. One way of bioinformatically dealing with this problem is to discard instances which are observed in less than a certain percent of samples, proportion of reads or given number of independent samples (87).

Once a metagenomics dataset has been characterised taxonomically and functionally, statistical comparisons between groups, experimental conditions or time series experiments among others can be carried out using regular parametric and non-parametric methods within traditional multivariate statistical approaches. Beyond this, several traditional ecological methods can also be applied to microbial ecology, and packages such as vegan(88) and phyloseq (89) exist within the R environment to facilitate their use. These can be complemented by multidimensional scaling tools which are extremely important for the visual representation of high dimensional data and are facilitated by the ape (90) package within the R environment.

These measurements are typically broken down into alpha and beta diversity. Alpha diversity describes the diversity within a sample of environment. At its most simple, this means the number unique species observed at a given site and is therefore scaled from 0 to infinity. Beta diversity allows for the comparison of diversity between samples, again at its most simple this strategy counts the number of species unique to one environment being compared, and adds this to the number of unique species in a second environment, giving an eventual score of beta diversity, or dissimilarity between the two samples or environments (91).

In practice, there are a variety of strategies for measuring both alpha and beta diversity, each of which lend or subtract weight from certain aspects of the

comparison, such as phylogeny. Some of the more common measures for alpha diversity used in microbial ecology are as follows. The Shannon diversity index takes into account both the evenness and the abundance of species in an environment (92). If a sample is dominated by a small number of species, it is not considered diverse, and both the number of species obvserved and an evenness in their abundance is required for an increase in diversity using this metric. Simpson's diversity index uses a similar principle. Chao1 species richness belongs to a class of methods called nonparametric estimators, which are adapted from mark-release-recapture ratio approaches in macro-ecology(93). This means that the number of observed species is added to the ratio of species only seen once versus species seen twice. This index is noted for its accuracy with sparse datasets, such as microbiome data.

Despite the concept of beta or between sample diversity being quite simple, there is no gold standard methodolgy for its measurement. The most common metric used in microbial ecology is the Bray-Curtis dissimilarity(94). Always a number between zero and one, this measure of dissimilarity between two samples is measured by subtracting two times the sum of lesser counts of species shared between both sites, divided by the total number of counts of species in both sites, from one. Therefore, a score of one indicates that the samples are identical and zero that they have no species in common. The Jaccard index is similar to Bray-Curtis, with the exception that it does not account for the quantities of species observed, instead working off a binary view of presence or absence(95). The number of shared species between two samples, are divided by the cumulative number of species found in both samples, this number is then subtracted from one to give a measure of dissimilarity.

While the previous examples are common ecological techniques that have stood the test of time and have now been adapted for use in microbiology, Unifrac is a more modern method, designed specifically with microbial communities in mind. A common complaint of metrics such as the previously described Bray-Curtis dissimilarity when analysing microbial data is that they treat sequences with 99% and 20% sequence similarity as equally different, resulting in a loss of information potentially useful for descrimination. UniFrac or the unique fraction(96), measures the phylogenetic distance between species in two different samples. There are two

different implementations of this measure based on whether (weighted) or not (unweighted) the abundances of these species are taken into account.

Specific tools tailored to metagenomics data also exist, which have been modified to be more sensitive to the specific nature of the data. For instance, the problem of calculating differential abundance across experimental conditions, be it of taxa or gene expression, is one that has been significantly improved on with the development of bespoke bioinformatics tools.The DeSeq algorithm was developed for RNAseq data, with the aim of finding genes that are differentially expressed across treatment groups, samples or time points based on the negative binomial distribution. In the second iteration (DeSeq2) this has been extended to other types of HTS data(97). MetagenomeSeq was specifically designed for marker gene surveys such as 16S rRNA gene sequencing but can equally be used for count tables generated by whole genome shotgun sequencing experiments. It addresses the effects of both normalization and under-sampling of microbial communities and also incorporates the testing of feature correlations (98). The Anova-Like differential expression tool for high through put sequencing data (ALDEX2) uses underlying assumptions of compositionality (99).

Reproducibility and Benchmarking

There are a multitude of bioinformatics platforms available for metagenomics analysis, and not only the tools of choice but also how a specific pipeline is used can have an effect on the conclusions drawn by the resulting data. As every step of a metagenomics study can bias the result and change our perception of the underlying microbial community, it is vital to keep all variables consistent throughout a study and include them in the method section. Apart from the DNA extraction method and choice of sequencing technologies, and in case of amplicon sequencing studies which 16S rRNA region was targeted with which exact primers, this also must include a description of how DNA contamination was controlled for, how the sequencing error rate was assessed and importantly an in depth description of the *in silico* analysis. It is not sufficient to only report the used tools but also the versions and specific parameters (if divergent from the defaults) have to be indicated. To allow reproducibility and comparison of studies it is recommend to not only make

the generated data of study available but also the implemented code used to analyse the data (100,101)

When developing a suite of tools or pipeline for analysis, the benchmarking of different potential platforms is an effective way to ensure the pipeline developed suits the experimental needs. Tools such as MetaSim (102) provide the raw materials for such a benchmarking project. This allows the user to define and simulate a sequencing dataset of known microbial composition, and consequently to assess the accuracy and speed of potential metagenomics platforms(102). In terms of reproducibility, an important initial step in promoting this is to deposit the results of any sequencing experiment into one of the online sequence data repositories available. Two of these are the Sequence Read Archive (SRA) (103) and Metagenomic Rapid Annotations using Subsystems Technology (MG-RAST) (104).

The sequence read archive, or SRA, is one of the most important platforms involved in metagenomics research. It is the primary repository of high throughput sequencing data hosted by the National Institute of Health in the United States, and part of the International Nucleotide Sequence Database Collaboration. A wide range of sequencing data is accepted, such as Roche454, Illumina and Pacific Biosystems data. All data submitted to this portal is publically available, and serves the purpose of aiding new discoveries by increasing access to data, and promoting the reproducibility of the field of metagenomics, which is at present one of the key weaknesses of the field(103). MG-RAST is another repository for sequence data. It currently stores over 150,000 datasets with over 23,000 of them in the public domain. As the name suggests, it also provides some limited functionality in metagenomics analysis. Raw reads can be uploaded in fastq format, which are then taxonomically and functionally annotated with minimal user input. Further analysis and visualisation is then possible through an interactive web server (105).

SECTION 2: CHARACTERISING INTRATUMOURAL BACTERIA

This chapter has been published as:

*"**Sequence-based Characterisation of Intratumoural Bacteria – A Guide to Best Practice"***

Sidney P Walker [a,b,c,d], Mark Tangney* [a,b,c], Marcus J Claesson* [c,d]

Frontiers in Oncology **10,** 179 (2020)

Abstract

Tumours environments are amenable to bacterial growth and several recent studies on cancer patient samples have introduced the concept of an endogenous tumour microbiome. For a variety of reasons, this putative tumour microbiome is particularly challenging to investigate, and a failure to account for the various potential pitfalls will result in erroneous results and thus false claims. Before this potentially significant habitat can be accurately characterised, a clear understanding of all potential confounding factors is required, and a best-practice approach should be developed and adopted.

This review summarises all of the potential issues confounding accurate bacterial DNA sequence analysis of the putative tumour microbiome, and offers solutions based on related research with the hope of assisting in the progression of research in this field.

The tumour microbiome: Current status and future challenges

The existence of a tumour bacterial microbiome is still a contentious concept, but an increasing number of articles are being published exploring this novel habitat, and simultaneously exploring the possible effects these bacteria could have. To inform the direction such research will take in the future, it is important to take stock of the research carried out to this point to learn from past mistakes, and similar analyses in relevant fields. Research to date has focused on two key questions; what is there, and what does it do? This has involved comparing the microbiota of malignant and non-malignant breast tissue (including non-cancer patient) in the original studies (19,106,107). Subsequent studies examined potential causative links between bacteria and their host tumours, or assessing their metabolic activity, for example their effect on chemotherapeutics (108,109). These concepts have important potential in cancer care, in terms of treatment regime, diagnosis or prevention, but rely on the field developing a thorough understanding of the microbial-related tumour microenvironment.

The key hurdles in accurately characterising these environments are outlined as follows.

- Tumour samples are regions of known low microbial biomass, a feature which complicates any metagenomic analysis. This review will include suggested methodologies for bioinformatic analysis of tumours, and also of low biomass samples in general. Linked to the issue of low biomass, tumour samples present an extremely high ratio of host to bacterial DNA, which can lead to bias in amplicon based sequencing strategies such as 16S rRNA sequencing, and can make whole genome sequencing impossible without a microbial enrichment strategy (110).

- A further problem relates to the quality and quantity of patient tumour-related samples. Sourcing high numbers of aseptically-collected samples to enable statistical power is challenging, due to potential impact on standard of care, the workload of healthcare professionals, and competing requirements of the hospital diagnostic and other research teams for a limited amount of sample. A resource with potential for higher sample throughput for tumour metagenomics analysis is formalin-fixed paraffin-embedded (FFPE) tissues, the international gold standard for tissue sample storage. A proof of concept study recently showed that FFPE tissues provided a reliable source of germline and malignant human DNA (111). It is hoped that FFPE

tissues can provide reliable bacterial DNA also, once the proper precautions are taken, not least distinguishing contamination inherent to this biobanking process. As with the low biomass characteristic, FFPE tissues would also present challenges to any bioinformatics analysis.

When performing library preparation and bacterial DNA sequence analysis to investigate the tumour microenvironment, the issues raised in (i) and (ii) manifest in a number of ways. Introduced environmental contamination is likely to be inherent given the sampling process, which, given the low biomass nature of this tissue, has the potential to obscure tumour-originating bacteria. Similarly, there are other issues associated with low biomass such as PCR bias caused by the high ratio of host to bacterial DNA. If FFPE samples are used, errors in the sequence data will occur due to DNA damage during the formalin fixation process (112,113).

In summary, as more research is carried out into the tumour microbiota, it is important to address the many potential pitfalls involved to ensure that these environments are reliably characterised, the scale of the problem is shown in Figure 1. The credibility of this field and other low biomass fields has been affected by recent publications highlighting methodological mistakes in previous research characterising the microbiome of tumours and other low biomass environments (23). Therefore, a robust strategy needs to be established to ensure that future results are as reliable as possible.

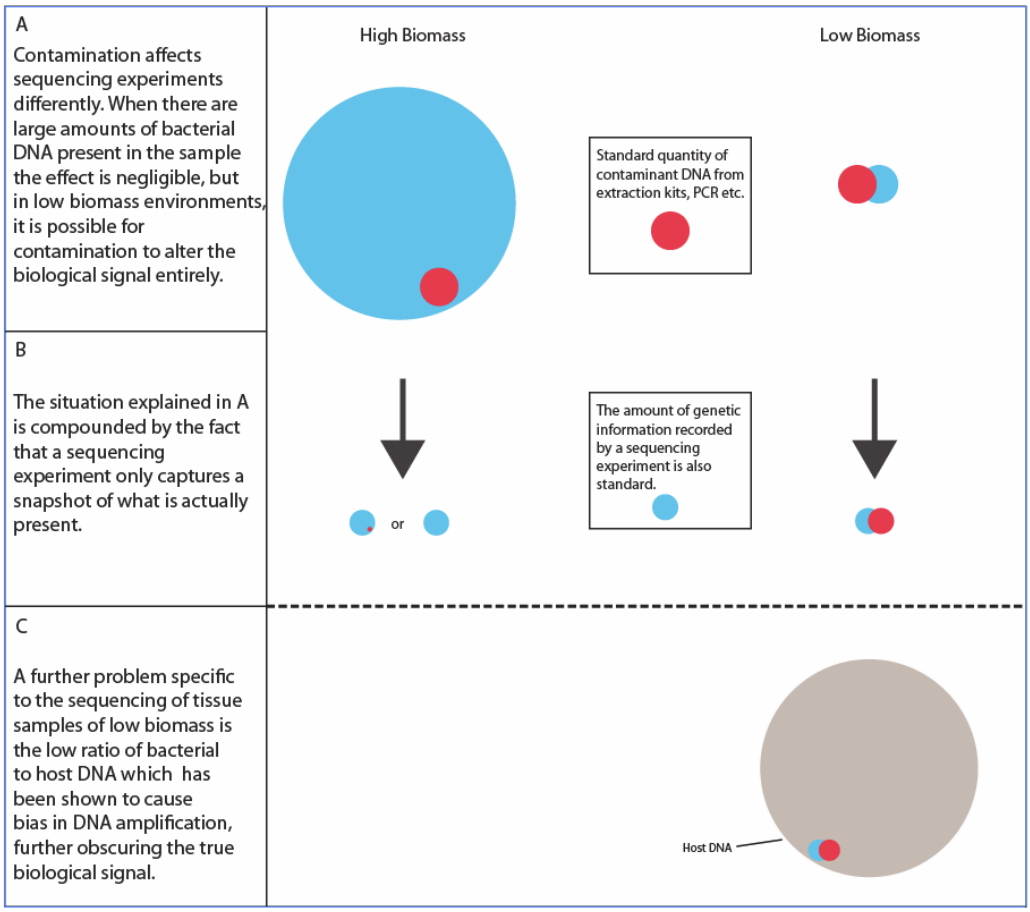


Figure 2.1: The scale of the problem. *Low biomass environments are considerably more susceptible to biological signal alteration arising from contaminant DNA than high biomass samples, along with the increased likelihood of PCR bias.*

Research on the Tumour Microbiome to date

The recent work characterising the microbiomes of solid tumours is outlined in Table 2.1 below. Due to the challenges posed in characterising the tumour microbiome, it is likely that some or all of the studies referenced have been negatively impacted in some way, reducing their accuracy. This caveat must be kept in mind when assessing the results, and reinforces the need for the introduction of a best practise methodology to make future research more reliable.

Table 2.1: *Tumour sites with suspected bacterial communities*

Tumour site	Description	Bacterial Community	Reference
Breast	Tumour tissue has microbial signature similar to surrounding tissue Tumour adjacent tissue significantly different to non-cancer patient breast tissue.	*Enterobacteriacae* spp. (Proteobacteria), *Gammaproteobacteria* spp. (Proteobacteria), *Acinetobacter* spp. (Proteobacteria), *Bacillus* spp. (Firmicutes), *Staphylococcus* spp. (Firmicutes), and *Lactococcus* spp. (Firmicutes). Bacteria found in healthy breast tissue: *Micrococcus* spp. (Actinobacteria) and *Prevotella* spp. (Bacteroidetes), and to lesser extend *Lactococcus spp.* and *Gammaproteobacteria* also found in cancer-related tissue.	[8-10]
Pancreatic Ductal Adenocarcinomas (PDAC)	*The* cancerous pancreas has a more abundant microbiota than healthy control.	Enterobacteriaceae, Pseudomonadaceae and to a lesser extent *Streptococcaceae*, *Staphylococcaceae* and *Micrococcaceae*	[5,11]
Prostatic Cancer	Microbiome analysis carried out by	Actinobacteria, Firmicutes and Proteobacteria,	[12]

	Pyrosequencing.	Lactobacillales and *Streptococcaceae* significantly elevated in healthy samples, *Staphylococcaceae* in tumour and peritumour.	
Others	Ovarian and lung cancer tumour microenvironments have also been characterised.		[13-15]

Is there an aetiological relationship between tumours and bacteria?

Considerable research has been conducted to demonstrate links between microbiota and a variety of proximal and distal cancers. Some associations were found to be directly causative, such as *H. pylori* and Gastric Adenocarcinoma (114) In other circumstances, reports suggesting certain bacteria being elevated in specific instances of cancer along with a variety of potential mechanisms for causing/progressing the cancer make a strong case, even if the final confirmation has yet to be found. An example of this is the constantly developing picture of the role *Fusobacterium* plays in colorectal cancer (115). Mycoplasma infection has also been shown to transform normal lung cells, affecting cell proliferation and differentiation (116). In many tumours, it may be that bacteria are simply opportunistic inhabitants (21,117). Tumours are uniquely amenable to bacterial colonisation, and unlike healthy tissues, conceivably provide a refuge for circulating bacteria, including non-invasive species (Figure 2.2). A collection of phenotypes unique to tumours which have been proposed to explain the phenomenon of selective tumour colonisation by bacteria are as follows: i) Angiogenesis associated with tumour growth is an imperfect process, resulting in disorganised or "leaky" vasculature. This could allow circulating bacteria to embed themselves in the tissue. ii) Tumours are immune privileged regions of the body. This characteristic means that bacteria which may be cleared by the host immune system at other body sites are able to proliferate within tumours. iii) Many solid tumour regions are hypoxic, this lower level of oxygen compared with healthy surrounding tissue provides an environment that suits the proliferation of facultative and anaerobic bacteria. iv) Necrotic regions within the tumour are nutrient rich, promoting bacterial proliferation.

What is the significance of endogenous bacteria residing within tumours? Beyond ongoing research into any causal relationships between bacteria and tumours, there are several other benefits to fully understanding these habitats. Understanding what bacteria colonise tumours could help with the development of more personalised or targeted treatment regimens for many tumours, maximising effect on the tumour and minimising the impact on the patient. A number of potential influences (both positive and negative) of resident intratumoural bacteria on tumour growth and responses to treatments have already been proposed by us and others, and include

effects on therapeutics, potential cross-talk between cancer cells and bacteria, and the potential for intratumoural bacteria to mediate therapy. *For example,* we were the first to report that a variety of unmodified bacteria found in tumours, with natural levels of endogenous enzymes can either positively or negatively affect the efficacy of various chemotherapeutics, such as gemcitabine, as evidenced by *in vitro* and *in vivo* cancer models (108). In parallel, given their unique capacity for selective growth in tumour tissue, therapeutics may be locally produced within the tumour by administered engineered bacteria (118,119). However, considerable challenges stand in the way of an approach such as this becoming a reality.

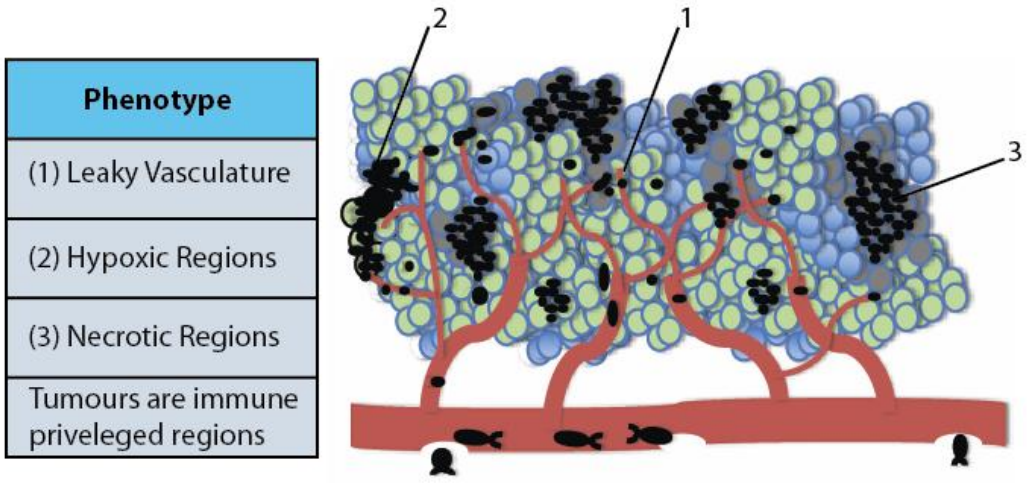


Figure 2.2: Tumours are uniquely hospitable environments for bacteria. *i) Leaky vasculature allows circulating bacteria to embed in tumour tissue; ii) Tumours are immune privileged regions; iii) Solid tumours possess low oxygen regions suitable for the proliferation of facultative and anaerobic bacteria; iv) High-turnover regions of tumours can be nutrient rich, promoting bacterial growth.*

Biological Considerations

Formalin-fixed paraffin-embedded tissue (FFPE) represents a resource that, if correctly harnessed, could exponentially increase the sample sizes and sites available for tumour microbiota studies. Crucially, these do not have to be obtained at the time of surgery, like fresh frozen tissue, although both fresh frozen and FFPE tissues involve difficulties.

Patient sample logistics

The realities of patient sample acquisition must be taken into account by researchers in this field.

- Sampling-related contamination e.g. from the patient, the operating theatre, or the pathology lab (tissue handling and processing) must be considered in the design of research workflows (see later).
- Broad-spectrum antibiotic administration can be routine in many hospitals immediately prior to tumour resection operations. While interfering with the clinical standard of care is difficult, antibiotic administration should be considered and reported in such studies.
- An under-considered parameter is that tissue is heterogenous within a tumour, and bacterial profiles are likely to differ (quantity and quality) intratumourally, with some tumour tissue providing different growth conditions to other regions. Hence, typical pathologist-preferred tumour regions required for diagnosis (e.g. ‘margins’) may not be representative of the holistic tumour microbiome.

Low biomass

Tumours represent low bacterial biomass samples. This poses a variety of challenges to the data generation process. This is a situation where the bacterial DNA that is the target of the study, is outnumbered by orders of magnitude by host DNA. Due to the targeted PCR amplification of bacterial DNA in 16S rRNA gene sequence analysis, this heavy ratio of host to bacterial DNA is commonly considered unimportant. This is not the case, with many studies demonstrating a reduction in PCR amplification efficiency in circumstances of high human nucleic acid and low bacterial 16S rRNA gene fragment copies, ultimately leading to sampling bias (120). Therefore, an

effective host DNA depletion strategy is an important component of a 16s rRNA gene sequencing library preparation.

Commercial kits for microbial enrichment by host DNA depletion were recently compared by Marotz *et al* (112). These included MolYsis, QIAamp and lyPMA kits. All were found to significantly improve the microbial yield. lyPMA was the most effective, having a mean of 8-10% of reads aligning to the human genome, and MolYsis the least, with an average of 60% of reads aligning to the human genome. It is inevitable that the microbial DNA would also be affected. For example, the MolYsis approach is suspected to degrade bacteria with weak cell walls, or cell walls that have been previously weakened by exposure to certain antibiotics, so a balance between host depletion, and bacterial degradation must be found.

Contamination

A recurring issue with low biomass samples is contamination, which poses a significant challenge in sequence analysis and interpretation. Often, the true microbiota can be masked by confounding bacterial DNA found in library preparation and DNA extraction kits. This feature is then often exacerbated by subsequent intensive amplification via PCR. Typical sources of contamination include environmental (surgery- and pathology-related), contaminants during the library preparation, and, as has been recently described, contamination from within the extraction kit itself (23). Since Salter *et al* published on this, there has been a general increase in awareness that reagent, laboratory and human contamination can have a serious impact on microbiome analysis (12). As water and soil associated bacteria are well documented contaminants associated with DNA extraction kits and PCR reagents, some contaminants are easily identified if they make it through the sample preparation, sequencing and bioinformatics contamination removal process. Genera such as *Bradyrhizobium*, which function in nitrogen fixation, are unlikely to be legitimate constituents of any human microbiome. The problem becomes more complex when sequences from *Escherichia* spp. and *Bacillus* spp. are found. Both have been shown to be artefacts of the library preparation process, but both are also common human pathogens (12). In 16S rRNA gene sequence analysis, taxonomic resolution to the species level is not always available, and never available in the

instance of *Escherichia* spp., which compounds the problem.

Summary of contaminants affecting 16s rRNA gene sequence analysis

The table below is a summary of recent articles addressing and discussing the problem of contamination in sequence analysis. It contains genera mentioned across all recent studies which include analysis of extraction and PCR kits, and also the ultra-pure water that is used in many kits and as a negative control.

Table 2.2: ***Previously identified bacterial contaminants as per publications: (121),(12), (23), (122) .***

Phylum	**Genus**
Actinobacteria	*Actinomyces, Aeromicrobium, Agrococcus, Arthrobacter, Atopobium, Beutenbergia, Bifidobacterium, Blastococcus, Brevibacterium, Candidatus, Planktoluna, Cellulosimicrobium, Clavibacter, Collinsella, Corynebacterium, Curtobacterium, Dietzia, Eggerthella, Geodermatophilus, Gordonia, Janibacter, Kocuria, Microbacterium, Micrococcus, Microlunatus, Patulibacter, Pilimelia, Propionibacterium, Pseudoclavibacter, Rhodococcus, Rothia, Slackia, Tsukamurella*
Bacteroidetes	*Alistipes, Bacteroides, Bergeyella, Capnocytophaga, Chryseobacterium, Cloacibacterium, Cytophaga, Dyadobacter, Flavisolibacter, Flavobacterium, Gelidibacter, Hydrotalea, Niastella, Olivibacter, Parabacteroides, Pedobacter, Porphyromonas, Prevotella, Wautersiella, Xylanibacter*
Deinococcus-Thermus	*Deinococcus, Meiothermus*

Firmicutes	*Abiotrophia, Anaerococcus, Anaerotruncus, Bacillus, Blautia, Brevibacillus, Brochothrix, Catenibacterium, Christensenella, Clostridium, Dialister, Dorea, Enterococcus, Erysipelatoclostridium, Eubacterium, Facklamia, Faecalibacterium, Fastidiosipila, Flavonifractor, Gemella, Geobacillus, Granulicatella, Halocella, Intestinibacter, Johnsonella, Lachnoanaerobaculum, Lachnoclostridium, Lachnospira, Lactobacillus, Listeria, Megasphaera, Moryella, Oscillospira, Paenibacillus, Papillibacter, Parvimonas, Peptococcus, Peptoniphilus, Pseudobutyvibrio, Pseudoflavonifractor, Quinella, Roseburia, Ruminococcus, Ruminosclostridium, Selenomonas, Solobacterium, Staphylococcus, Streptococcus, Trichococcus, Tumebacillus, Turicibacter, Tyzzerella,Veillonella*
Fusobacteria	*Fusobacterium, Leptotrichiaceae*
Proteobacteria	*Achromobacter, Acidovorax, Acinetobacter, Afipia, Alcanivorax, Alicycliphilus, Aquabacterium, Aquabacterium, Asticcacaulis, Aurantimonas, Azoarcus, Azospira, Beijernickia, Bosea, Bradyrhizobium, Brevundimonas, Burkholderia, Cardiobacterium, Caulobacter, Comamonas, Coprococcus, Craurococcus, Cupriavidus, Curvibacter, Delftia, Devosia, Diaphorobacter, Duganella, Enhydrobacter, Enterobacter, Eschericia, Geodermatophilus, Haemophilus, Herbaspirillum, Hoeflea, Janthinobacterium, Kingella, Klebsiella, Leptothrix, Limnobacter, Massilia, Matsuebacter, Mesorhizobium, Methylobacterium, Methylophilus, Methyloversatilis, Neisseria, Nevskia, Novosphingobium, Ochrobactrum, Oxalobacter, Paracoccus, Parasutterella, Pelomonas, Phyllobacterium, Polaromonas, Pseudomonas, Pseudorhodoferax, Pseudoxanthomonas, Psychrobacter, Ralstonia, Rhizobium, Rhodanobacter,*

	Roseateles, Roseomonas, Rubellimicrobium, Ruegeria, Schlegelella, Serratia, Sphingobacterium, Sphingobium, Sphingomonas, Sphingopyxis, Stenotrophomonas, Sulfuritalea, Terrimonas, Thiohalocapsa, Undibacterium, Variovorax, Xanthomonas
Tenericutes	*Mycoplasma*

FFPE tissue as a source of sample tissue

With more developed screening methods and constantly improving medical care, particularly in the developed world, the size of tumours at the time of excision is rapidly reducing. The average size of a breast tumour has shrunk to less than 1 cm in diameter in the United States. As mentioned previously, this means fewer fresh 'surplus to diagnostic' samples are available to research (123). Formalin fixation followed by paraffin embedding is the gold standard for preserving tissue samples after histological examination. FFPE blocks are stable at room temperature, and preserve the morphology and cellular details of tissue samples, along with the DNA. A unique problem when handling FFPE tissues is the degradation and mutation to which the DNA is subjected during the fixing and embedding process. FFPE blocks are undoubtedly a valuable resource due to the sheer quantity of samples available. However, there are several challenges involved in their effective use. Formalin fixation has been shown to cause cross-linking of histone-like proteins to DNA, DNA to formaldehyde adducts, and inter-strand DNA crosslinks (124). Generally, sequencing errors are caused by PCR mistakes, or miscalls during sequencing, but in a small set of circumstances, sequencing errors are caused predominantly by mutagenic DNA damage. These include ancient DNA from archaeological sites, circulating tumour DNA, and FFPE samples (125). The value of FFPE tissues as a sample type has begun to supersede the difficulty in their processing and analysis from a bacterial sequencing perspective. A recent study by Stewart *et al* successfully used formalin fixed, paraffin embedded tissue to characterise the intestinal microbiota of pre-term infants with necrotising enterocolitis, despite some of their samples being almost 10 years old (126).

Although the strategies for minimising and/or retroactively repairing this DNA damage mainly falls under the remit of the laboratory personnel carrying out the extraction and subsequent library preparation, there are some bioinformatics strategies that can be applied to lessen the impact of damaged DNA on the sequence data. Chen *et al* proposed a method of scoring the extent of the errors in sequencing caused by DNA damage, called the Global Imbalance Value (GIV) [41]. This method is based on the directional adapters used in Illumina sequencing. The principle behind this is that because the majority of DNA damage only affects one base in a pair, DNA damage caused by oxidation, for example, could cause G-T transversion errors when the forward read of sequence data is mapped to a reference genome, but the reverse read would show the reverse complement of this, so C-A errors. This causes a "global imbalance" (125). A slight modification of this method would allow for the user to screen the reads generated by 16S sequencing of bacterial DNA within the tumour and in a process similar to the quality filtering already employed, only retain reads that had a GIV score below a certain threshold.

Bacterial DNA extraction from FFPE samples

Despite these problems with using FFPE tissues for metagenomic analysis, there is a considerable history of bacterial identification in FFPE tissue in clinical settings, if not research settings.(127). The QIAamp DNA FFPE Tissue kit is a purpose-built kit for the extraction of total genomic DNA from FFPE blocks produced by Qiagen. This kit compensates somewhat for damage caused by formalin fixation by including an incubation at elevated temperature following a proteinase K digestion. However, the kit does not take into account the oxidative damage that can be caused, or the extreme ratio of host to bacterial DNA, both of which can affect marker gene sequence analysis such as 16S rRNA gene sequencing (128).

If reliable characterisation of the bacterial communities within tumours is to extend to FFPE samples, then a protocol for bacterial DNA extraction, repair and purification from these tissues is required to improve downstream analysis. A workflow of biological considerations for a sequencing experiment is shown here in Figure 2.3.

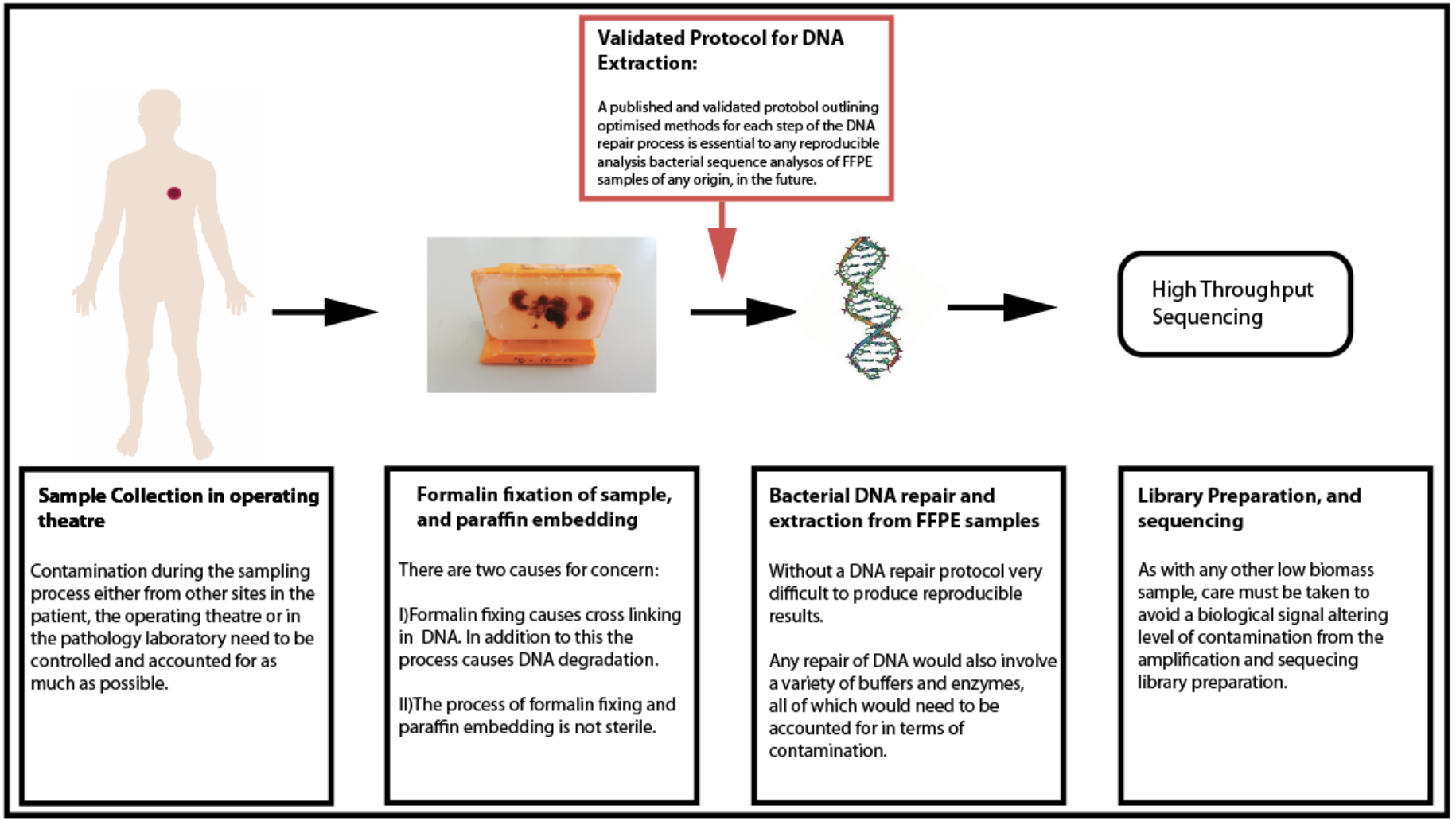


Figure 2.3: *Workflow of biological considerations prior to bioinformatic sequence analysis*

Bioinformatic Aspects

Detection of microbial communities

The two sequencing strategies employed in metagenomics analysis are WGS and amplicon sequencing. WGSWGS provides a high-resolution overview of all (or the most abundant, dependent on sequencing depth) DNA present in a sample. Bacterial genomes present will be characterised base by base, providing insights into bacterial taxonomy, function and rates of mutation, among other aspects. Host DNA present in the sample is also sequenced. Amplicon sequencing is a targeted approach allowing the targeting of specific regions within genomes, generally amplified by PCR. It is a two-stage process where primers are used to capture the target region, which is followed by high-throughput sequencing. Amplicon sequencing in bacterial microbiota studies typically targets the 16S rRNA gene subunit. This is the component of the 30S small subunit adjacent to the Shine-Delgarno sequence, a region noted for its slow rate of evolution, containing nine "hypervariable regions" which can be used to differentiate between bacteria with varying degrees of effectiveness (129).

Whole genome sequencing has several advantages over 16S rRNA gene sequencing, such as increased species and strain level resolution, enhanced ability to detect rare species, and the ability to detect organisms in other kingdoms of life, such as viruses and fungi (130). At present 16S rRNA gene sequencing may be more technically suitable to metagenomics analysis of low biomass environments, in addition to being significantly more cost-effective. In a typical sequencing run of a non-tract biopsy in humans, 97% of the reads generated can be expected to align to a human reference genome (131). This makes it extremely expensive to get sufficient sequencing depth of the bacterial DNA present in a sample (131). As mentioned earlier, 16S rRNA gene sequencing is still affected by the low ratio of bacterial DNA, but to a lesser extent than whole genome sequencing methods. This can be improved upon by incorporating the previously mentioned host depletion strategies.

Removal of Chimeric reads

Chimeras arise as aborted extension products from earlier PCR cycles and can end up being taken up as a primer in a subsequent cycle. Undetected chimeric DNA sequences can be misinterpreted as novel species, particularly in 16S rRNA gene sequence analysis. Therefore, the number of PCR cycles can influence chimera

formation (132). Given the low bacterial concentrations expected in tumour samples, the generation of chimeric reads is logically a significant cause for concern, and a robust protocol should be employed for their removal. Chimeras can be computationally identified and removed using one of a variety of programmes that fall into two groups. *De novo* methods usually work by identifying sequences which contain half of one abundant read and half of another, as evidenced by a difference in abundance between the start and the end of a sequence. Alternatively, reference-based methods compare reads identified to a curated database known to be chimera free, and attempts to find sequences that may have arisen from multiple samples (133). In this situation, where there is an elevated proportion of chimeras present, combining both methods would give the best chance of effective clearance of chimeras. Some of the most cited examples of chimera removal programmes across both categories include *Chimera Slayer* which is a referenced based method, *Is.Bimera.Denovo* which is the *de novo* chimera removal programme within the DADA2 pipeline, and UCHIME within the QIIME environment which has both reference based and *de novo* capabilities (133).

Removal of contamination

Two bioinformatics utilities have been developed recently, to retroactively solve this problem. SourceTracker, and Decontam (134,135). These methods have different functionality but can be used in conjunction to remove contaminant taxa. The SourceTracker algorithm utilises a Bayesian approach to provide an estimate of the proportion of contaminants that arise from possible source environments. Decontam looks for unusual relationships between DNA concentration in the original sample, and proportional abundance of sequence variants, and can add another layer by comparing samples with negative controls.

Analysing the outputs The traditional method of analysing 16S rRNA gene sequencing data by clustering reads together based on a pre-defined threshold of similarity is no longer necessary due to recent advances. New methods of error modelling allow for sequence variants to be distinguished by a single base, generating amplicon sequence variants (ASV) which are comparable to OTU's, but where OTUs are clustered by percentage sequence identity, ASV's correspond to an exact amplicon sequence variant in the sample (48). A major consideration when choosing a 16S rRNA gene sequence analysis pipeline is the degree of damage to the

DNA. As mentioned earlier, it is possible to measure this based on global imbalance value (125). DNA damage could cause ASV generating methods may be unsuitable as mutations, caused by formalin fixing for example could be incorrectly classified as different strains of bacteria. In these circumstances, the clustering-based OTUs may prove the more reliable method. Several of these are contained within the QIIME environment, such as Usearch (136). Samples can be analysed with both clustering and ASV methods, and a comparison of the number of observed species identified could inform the user on the level of damage. When combined with experimental knowledge for example laboratory based culturing from tumours, a large amount of closely related species reported by ASV generating methods but not clustering methods could indicate unrepaired DNA damage.

Best Practice

As there is currently no established best practice for sequence analysis of bacteria residing in tumour tissue, fresh or formalin fixed, the primary objective of this article is the proposal of such. The section below, along with Figure 2.3, summarises a methodology that falls in line with what is currently accepted for 16S rRNA gene sequence analysis, incorporating sample-specific modifications as outlined earlier.

Pre Analysis

During the extraction process, microbial enrichment and DNA repair, if the sample originates from FFPE tissue, should be carried out if possible. Since, in low biomass samples, the biological signal can be significantly altered by the presence of contaminants, extreme ‘aseptic’ care must be taken when preparing the samples for sequencing. A variety of controls to account for introduction of contamination should be used. Given the documented effects that a lack of controlling for contamination has had on previous tumour microbiota studies, this is of paramount importance. Eisenhofer *et al* recently published a comprehensive description of a robust strategy to control for contamination in low biomass studies (22). This suggests using a variety of negative controls to assess the degree of contamination introduced during the processes of sampling, DNA extraction and amplification. Positive controls are also recommended, such as mock communities of known microbial composition and amplification controls. This should be adhered to when

sequencing from FFPE tissues, with some additional steps as outlined below in Figure 2.4.

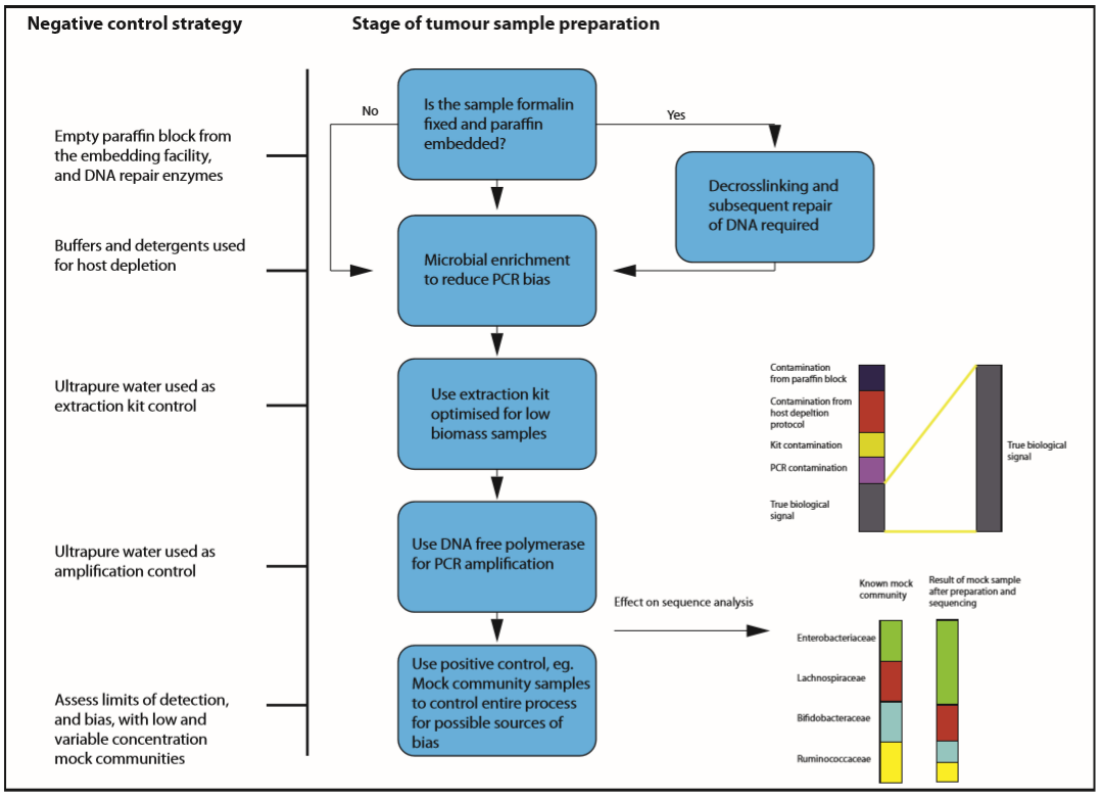


Figure 2.4: ***Overview of suggested sample preparation with appropriate control for contamination and bias.***

Bioinformatic analysis Figure 2.5 summarises the key points outlined previously in this article in relation to the required modifications to a bioinformatic pipeline required to ensure high quality reproducible analysis.

Figure 2.5: *Suggested bioinformatics workflow for bacterial sequence analysis from tumour tissue.*

Possible Improvements

Typically, hypervariable regions within the 16S rRNA gene fragment are targeted by primers, the most commonly targeted is the V3-V4 region as it is thought to provide the best resolution. While this method is effective to genus level in most cases, species level classification is often unsuccessful. An obvious solution would be to simply increase the length of the reads, as sequencing technologies such as Oxford Nanopore sequencing are capable of producing reads that are hundreds of kb in length, it should be straightforward to simply sequence the entire 16S rRNA gene fragment (137). Specifically in the case of sequencing samples from formalin fixed samples however, this is currently not possible, as the DNA will often be fragmented, preventing long read sequencing. A potential solution to this is to combine multiple, independently sequenced short regions within the 16S rRNA gene fragment. One way this has been implemented is in the Short Multiple Regions Framework (SMURF) method, by Fuks *et al* (39). This entails independent amplification and sequencing of multiple regions along the gene fragment, these are then computationally combined to provide a significantly more accurate assessment of the microbial community. When tested on a Human Microbiome Project "Mock" community, it was found that the increase in resolution was a function of the number of regions analysed. Using two different regions resulted in a two-fold increase in resolution, while using 6 resulted in a ~100 fold increase in resolution (39).

A further improvement was not directly related to the bioinformatic analysis but to sample preparation. As was mentioned earlier, while there are extraction kits for DNA in FFPE tissues, these do not take into account damage that may have occurred to the DNA during the fixation process, or the high ratio of host to bacterial DNA. To make metagenomic analysis of tumour samples from FFPE tissues a reliable and crucially reproducible option, there is a genuine need for the establishment of a validated protocol to extract bacterial DNA from FFPE tissues, repair the damage, and deplete the host DNA.

Concluding remarks

In conclusion, taking advantage of the presence of bacteria in tumours has the potential to contribute to cancer treatments in the future. As the field is still in its infancy, it is

important for data to be as truly representative as possible. It is the objective of this article to provide a guideline for more effective bioinformatic analysis of the tumour microbiota in future.

SECTION 3: UTILISING BACTERIA FOR THERAPEUTIC INTERVENTION

Regardless of whether or not a consistently present and detectable tumour microbiome exists, tumours are undeniably hospitable environments for bacteria to colonise. Where the unique physiology of tumours is seen as an obstacle for traditional cancer treatments, they represent an opportunity for bacterial-mediated solutions.

The use of bacterial cellular machinery to secrete proteins is far from novel. In the biotechnology industry, bacteria have been exploited for their ability to produce recombinant proteins such as human growth hormone and insulin (118). Industry and academia have also been exploring the potential for *in vivo* production of therapeutic proteins from bacteria. In this scenario, bacteria would either naturally, or through inducement, locate to the body site where they are required, for example a tumour, and once there would colonise the niche and produce therapeutic agents (118).

This are two considerations in this context, requiring two very different applications of bioinformatics. i) The first is to identify <u>which</u> bacteria colonize the desired niche in body; this can be a 'foreign' body such as a tumour, or parasite, or a distal niche such as the gut. ii) The second, often under-considered parameter, relates to <u>what</u> these bacteria produce. Synthetic biology presents enormous scope for sophisticated medical therapy mediated by novel synthetic proteins. However, the task of getting a bacterial cell to successfully express and secrete a stable protein that it does not produce naturally is far from trivial, and is becoming a key aspect of the synthetic biology field.

Microbiome research as an R&D tool

An appropriate workflow to develop such targeted therapeutic strategies involves combining the knowledge gained from microbiome analysis with the machinery available within bacteria. From a viewpoint of bacterial-mediated therapy, this is achieved by:

- Using microbiome research as an R&D tool to conduct an ecological survey of the target niche, the aim being to find candidate taxa which selectively colonise the niche in question (138).
- Modulation of an existing microbiome to create a niche for the bacterial vehicle to colonise.
- Artificially inoculate the same niche(139).

The desired end-result of an ecological survey of this kind is to be able to state with a degree of confidence that if a given bacterium is introduced into the host, it has a high probability of locating to the target niche. Following this, the niche-targeted bacterium, can be engineered to produce a therapeutic agent directly within this niche. Bacteria can produce toxins to directly kill tumour cells, release cytokines to attract immune cells to the niche, or produce synthetic proteins to interact with receptors on/in tumour cells (140). This workflow of researching what bacteria are present in a niche of interest and developing biological therapeutics for them to deliver is described in Figure 3.1.

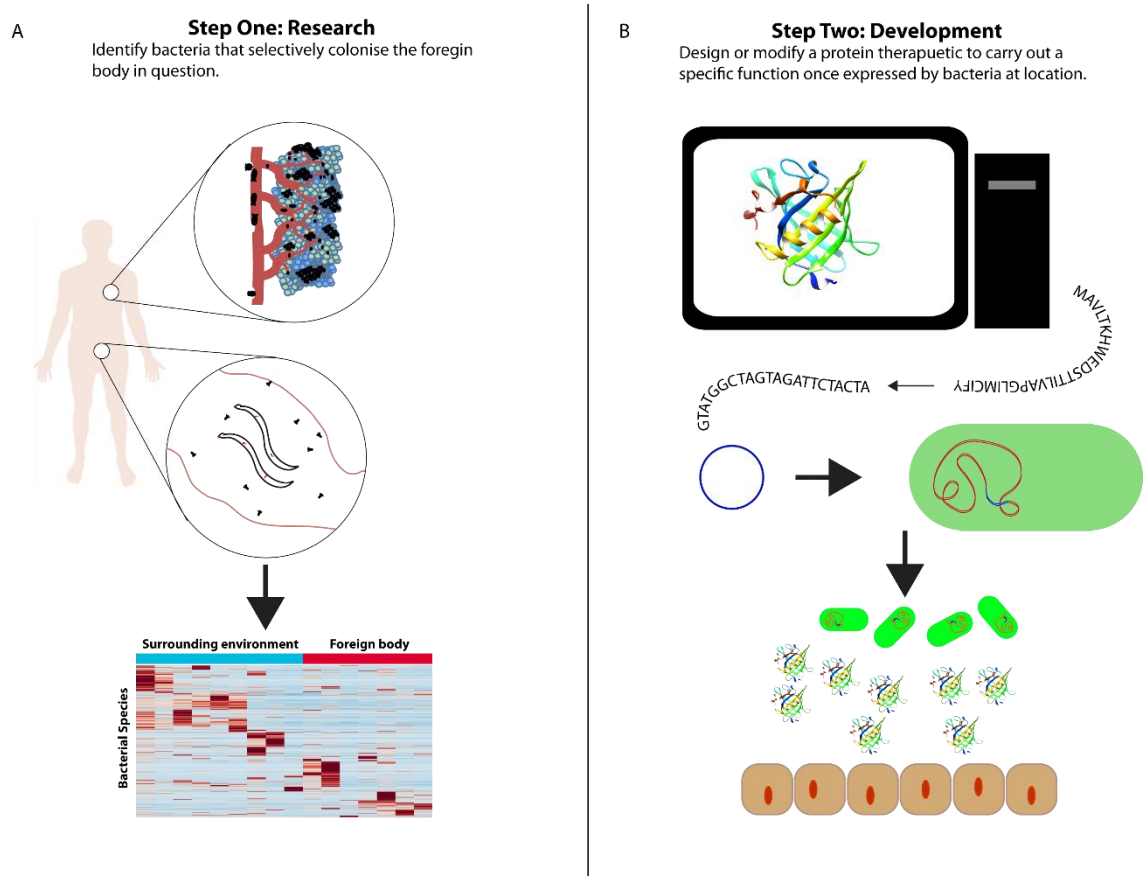


Figure 3.1: *The convergence of microbiome research and in silico protein design for R&D. (A) Shows the process of identifying bacteria that selectively colonise a foreign body of interest through bioinformatic analysis. (B) Shows a workflow for making use of the information learned in (A). A protein is designed for a specific purpose in silico, after which a bacterial candidate derived from A is genetically engineered to produce this protein, thus exploiting them as delivery vehicles for protein therapeutics.*

Examples of successful bacterial production of functional molecules *in situ*, but potentially distal to the site of action, include the production of cytokines, monoclonal antibodies and other molecules by *Lactococcus lactis* in the gastro-intestinal tract (141). A summary of recent developments in bacterial-mediated cancer therapies can be seen in the table below, adapted from Mansour Sedighi *et al* (142).

Table 3.1: *Summary of studies of bacterially mediated cancer therapy. (Adapted from (142))*

Treatment Strategy	Bacteria used and treatment approach	Outcome
Bacteria as immunotherapeutic agents	*Streptococcus pyogenes;* intentional infection of cancer patient with erysipelas.	Rapid tumour progression
	Attenuated *Salmonella Typhimurium;* vaccination of B16F10 tumour-bearing mice by derivatives of Salmonella Typhimurium (SL1344 InvA or SL3261AT InvA	Anti-Tumour effect
	Listeria monocytogenes; vaccination via recombinant *Listeria monocytogenes* (Lm-NP); breast, melanoma and cervical cancer.	Regression in growth of all types of tumours
	Clostridium spp; concurrent gas gangrene in patients with tumours	Tumour Regression
Bacteria as vehicles to produce tumouricidal agents	*Clostridium novyi;* IV injection of *C novyi* NT spores and a single IV dose of liposomal doxorubicin (Doxil) administered into mice bearing colorectal cancer	Elimination of tumours
	C. novyi NT and *C. sporogenes,* conjunction of pMTL-555-VHH	Rise of delivery of therapeutic agents

	construct of a VHH-AG2 expressing vector (an anti HIF-1a) into these bacteria	
	Bifidobacterium longum 105-A and 108-A, IV injection of the pBLES100 (constructed by cloning a *B. longum* plasmid and a gene encoding spectinomycin adenyltransferase AAD from *Enterococcus faecalis* into the *E. coli* vector pBR322) to B16-F10 melanoma tumour-bearing mice	Increase in specific gene delivery vectors in the tumour
Bacterially produced toxins/enzymes	*Salmonella enterica* Serovar Typhimurium, orally administered construction of Salmonella-based surviving vaccine into BALB/c, colon, DBT and GL261 glioblastoma-bearing mice	Vaccine as an adjuvant against different types of cancer
	Streptococci and *Serratia marcescens*, injection of bacterial concoction derived from heat-killed streptococcal and *Serratia marcescens* (Coley's Toxin) into body, sarcomas	A severe erysipelas infection led to the cure of cancer
	Clostridium perfringens, intratumoural injections of either 2, 10 ug of *Clostridium perfringes* entereotoxin (CPE) in xenografts of T47D breast cancer cells in mice.	Rapid and dose dependent cytolysis
	Pseudomonas aeruginosa, IV injection of the chimeric fusion protein	Significant antitumour

	interleukin-4-Pseudomonas exotoxin (IL4-PE) into GMB induced in nude mice and intratumour administration of IL4 PE in malignant astrocytoma in a phase I clinical trial	activity

A more precise approach of this kind has the potential, at a minimum, to limit side effects of traditional treatments that occur due to systemic administration, and to increase efficacy of treatments (140). Although considerable progress must be made before bacteria can be used clinically for cancer treatment and detection, it is hoped that the mainstream incorporation of microbiome research into research and development pipelines may accelerate this process.

In silico platforms for protein analysis and design

Many of the factors that attribute to the successful production/behaviour of a protein fall beyond the remit of computational biology, but some can be controlled for by bioinformatic analysis and prediction.

Foremost among these is the protein folding problem. The proteins used are rarely in their native state, and can at the very least expect to have additional functional 'parts', such as secretion sequences, detection tags etc., while at the other end of the spectrum, novel proteins are being developed with increasing regularity and confidence facilitated by *in silico* design tools. Predicting the expected 3D structure can inform the user as to something as simple as its predicted stability in nature, or whether a functional peptide 'part' is buried within the structure, rendering it non-functional. There are laboratory methods for characterising the 3D structure of a protein, such as Nuclear Magnetic Resonance spectroscopy, X-ray crystallography and Cryo-Electron Microscopy. In addition to the cost and expertise these methods require, they also need physical protein sample, which prevents their use as a design/screening tool prior to their build.

The prediction of the three dimensional structure of a test sequence in isolation would be of minimal experimental value without accompanying functional information. Fortunately, these two features are inherently related, as a protein's structure is the determining factor in its function. Once a model has been generated, predictions of features such as binding sites, interactions with other proteins, and transport machinery can be made, with the caveat that they are only as reliable as the underlying structural prediction.

Secondary to this initial question, other features that can be predicted or modified using *in silico* tools which can benefit research include:

- Sequence based parameters
- *De novo* Protein Design

Predicting protein 3D structure

The field of *in silico* protein structure prediction has expanded dramatically since 1994 when the first Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP) was held, but remains one of the more challenging approaches in the field of computational biology. Levinthal's paradox tells us that due to the large number of degrees of freedom in an as yet unfolded polypeptide chain, the number of possible conformation of this protein is enormous. The given example is that a polypeptide with 100 residues, and therefore 99 peptide bonds and 198 different phi and psi bond angles, would have 3e198 different conformations. The paradox is that although small proteins fold almost instantly, on a microsecond timescale, if a protein were to arrive at its correct fold structure by sequential sampling, this process would take longer than the age of the universe to complete (15). Computational methods provide a tentative solution to this problem, but major concerns over the reliability and accuracy of these methods remains, particularly when analysing larger proteins.

Nevertheless, several research groups have dedicated themselves to this problem and similar ones. The leading *in silico* protein prediction software currently available for academic use includes I-TASSER by the Zhang lab (143), the Rosetta Suite (144) and

SWISS-MODEL (145). An indication of the growth in awareness of the value and importance of this field is that Google entered the 2018 edition of CASP with A7D, a *de novo* structure prediction method with deep-learning based scoring (146). Protein structure prediction protocols can be broadly separated based on methodology, into homology modelling methods, and *ab initio* methods, although many leading tools combine the two approaches. Homology modelling, or comparative modelling, relies on sequence similarity between the test sequence and the sequences of already characterised proteins. This can be carried out either through global alignments of entire primary protein sequences or local alignments of smaller fragments in a process referred to as "threading." The fact that three dimensional structure is more conserved than amino acid composition amongst proteins makes this modelling method very effective for proteins that share medium to high levels of sequence similarity with those in the PDB database. If a test sequence has less than 20% identity with those in the PDB, this method becomes unreliable (147,148).

Predicting a protein's structure from its amino acid sequence alone, with no reference template structures, is still an unsolved problem despite any advancement over the previous ~50 years. This process is called *ab initio* folding. A simplification of this process is a search of possible conformations that the test sequence could take, which is supervised by an energy scoring function, usually related in some way to Gibbs free energy. A review by Lee et al states that there are three factors required for a successful Ab Initio prediction protocol (149): (i) An accurate energy function, where native protein structure correlates with thermodynamic stability, (ii) a computationally efficient method for conducting the conformational search and identifying low energy states, (iii) a protocol for identifying near-native models from a large number of conformers (149). To reiterate, both methods have strengths and weaknesses. Homology modelling requires sequence identity with already characterised proteins, and *ab initio* modelling is, at the time of writing, ineffective for larger constructs, although this is likely to change in the future as computational power increases. These limitations mean that the most successful modelling protocols must incorporate aspects from both *ab initio* and homology modelling into their process.

I-TASSER is primarily a homology based modelling tool, but does incorporate limited *ab initio* functionality, primarily to fill in gaps left by its threading tools in aligning sequences to reference databases. It works by initially searching for structural templates of fragments of the input sequence with a technique called threading or fold recognition. These are then assembled into full length models using replica exchange Monte Carlo simulations - this is the homology modelling method. Any unaligned regions of the test sequence are built by *ab initio* modelling. Further clustering and refinement steps result in five candidate models by default, each with a corresponding confidence (C-score) score (143). This tool is available as a web server, but can also be downloaded as a stand-alone tool.

Similarly to I-Tasser, the Rosetta Commons also maintains a web server for protein structural prediction called Robetta (150). Contrasting with I-TASSER, standalone tools within the Rosetta Suite provide many more functional options to the user when approaching the problem of protein structural prediction. *AbInitoRelax* provides a general framework for *ab initio* modelling of proteins, with different version available for membrane and metalloproteins. There are also facilities for homology modelling. *RosettaCM* allows the user to select templates themselves, either from the PDB database or previously *ab initio* modelled proteins. As with all standalone tools within the Rosetta suite, manual intervention is possible to tweak functionality to suit a particular target protein (151).

A comparison of workflows between I-TASSER, ostensibly a homology based algorithm, and and Rosetta, which is *ab initio* based, is shown below.

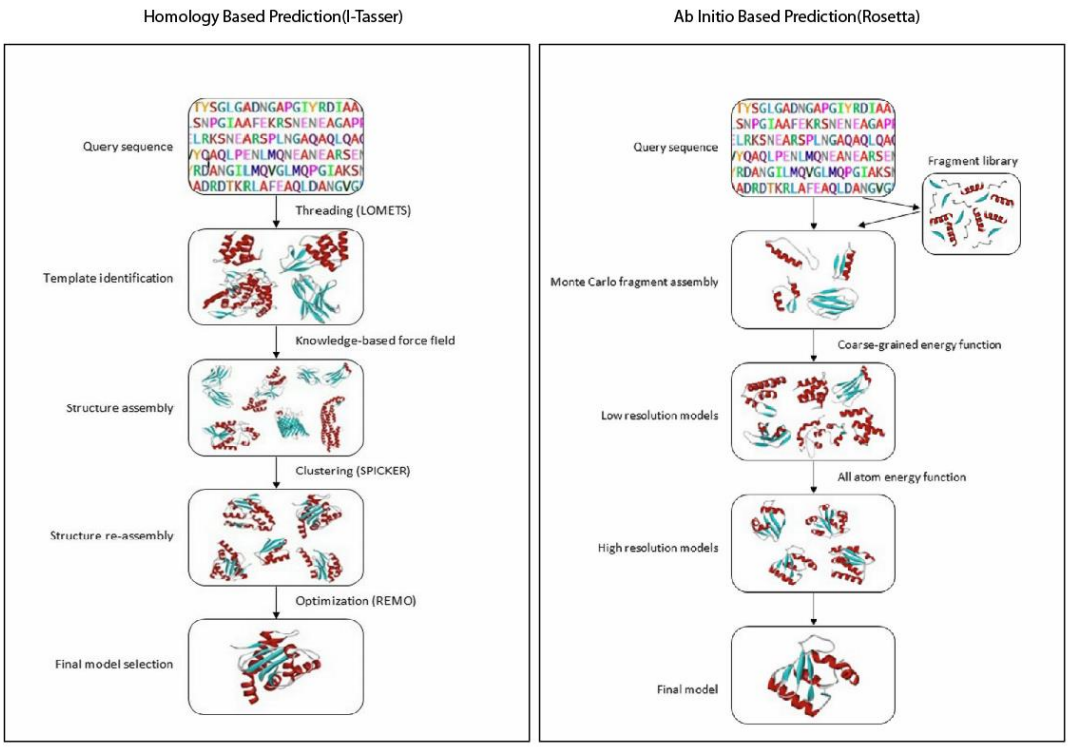


Figure 3.2: *Comparison of workflows between homology and ab initio protein prediction algorithms. Adapted from (152).*

A7D focusses exclusively on the challenging task of *ab initio* modelling, without the aid of any homologous template structures. This method of modelling is likely to increase in importance with the advent of *de novo* protein design exploring hitherto un-sampled protein fold space. The predictions are made by an automatic free modelling structure prediction system guided by a scoring system based on two different neural networks, both of which are deep convolutional neural networks. Convolutional neural networks are used predominantly for the processing of images. A deep residual convolutional neural network, trained on a non-redundant subset of the PDB database, builds a distribution of expected distances between the C-beta atoms of adjacent amino acids in proteins, and a second network trained to output a score as a function of multiple sequence alignments, and predictions from the first network (146).

A selection of the tools available for research involving the prediction of protein 3D structures is found below.

Table 3.2: Platforms for in silico protein modelling

*In practise I-TASSER and Rosetta incorporate both Homology and *ab initio* methods. **Prediction Tool**	**Platform**	**Method**
I-TASSER(143)*	Web-based and Standalone	Homology
Modeller(153)	Standalone	Homology
SWISS-MODEL(154)	Web-based	Homology
Phyre2(155)	Web-based	Homology
Fragfold(156)	Standalone	Ab-Initio
Rosetta(151)*	Web-based and Standalone	Ab-initio

Predicting protein function

Predicting the function of an *in silico* designed protein primarily relates to predicting its interactions with other proteins, ligands or other biomolecules and predicting the location of the active sites facilitating these interactions. As the majority of *in silico* designed proteins are either *de novo* or redesigned existing scaffolds, it is rarely necessary to investigate the overall function of a protein from first principles. When necessary, this can be done by aligning the amino acid sequence to annotated functional databases such as Swiss-Prot, or by comparing the three dimensional structure of the query protein to an annotated experimentally derived 3D protein structure database such as the PDB.

Predicting Protein active sites

Identification of protein active sites facilitating binding to targets is a crucial step in protein annotation or design. At present there is no one gold standard method predicting these sites, thus a common approach among the more successful strategies is to combine complementary prediction algorithms. As per the Continuous Automated Model EvaluatiOn community wide survey (CAMEO (157)), the top performing strategy for active site prediction is COACH (158), within the I-TASSER suite. COACH combines outputs from five different active site prediction algorithms including TM-SITE which employs reference based substructure comparison, S-SITE based on sequence alignments, and COFACTOR which threads sequence fragments through the BioLIP (158) protein functional database providing insights such as Gene Ontology and Enzyme Commission annotation in addition to binding site prediction.

An alternative strategy for binding site identification, if using a novel protein that makes database reliant methods ineffective, is to use global protein-protein docking tools which will be described later in this text. These can suggest likely interaction sites between a protein and its target based on protein conformation and an in-built energy function.

Predicting Protein-Protein interactions

Protein-Protein interactions, commonly referred to as ‘docking’, are the physical interactions that occur between two or more proteins. This physical contact should be specific, in that it involves active sites directed at each other, and care must be taken to eliminate chance interactions. This is one of the biggest challenges to the exploration of these inter-protein dynamics. While it is relatively easy to show *in silico* that two proteins have some affinity towards each other in certain conformations, it is much more difficult to show conclusively that two or more proteins do not interact (28).

Many tools, both standalone and web-based, exist for this purpose, but range from extremely basic tools that only give an overview, to more in-depth tools that have the ability to completely characterise the relationship between two proteins, but require considerable manual intervention. Broadly speaking, web-based servers can provide an overview of potential interactions between proteins. Web-based servers such as ClusPro yield a selection of potential interaction points between two structures, with associated

confidence scores (159). Other web based servers include HADDOCK (160) and SWARMDOCK (161). At the minimum level of user involvement, all three of these servers can take as input two pdb files and predict the interactions that occur between them. They do offer limited levels of advanced usage - for example, ClusPro allows for the removal of unstructured protein regions, and consideration of small angle X-ray scattering data among others. HADDOCK allows the user to provide interaction restraints which can guide the search, but if these are not provided, the accuracy of the algorithm regresses (159).

These web-based docking algorithms are convenient as they can give an outline of potential interactions with no requirements for expertise or computing power. If user expertise and computing power are available, standalone tools such as Autodock and Rosetta are considerably more powerful. Autodock is a suite of molecular modelling tools, initially designed to predict interactions between proteins and small molecules. Adaptations of these algorithms have led to their use in full protein-protein interaction prediction. Although still supported, Autodock has largely been superseded by Autodock vina, which delivers improvements both in accuracy and speed (162). Autodock vina still retains the focus of Autodock, which is the docking of proteins to small molecules, and although it can be used to predict interactions between two full proteins, it is a very slow process. The advances in speed that the heuristically modified QuickVina2 brings over Autodock Vina (163) allow for the prediction of interactions between full size proteins, provided that some information is known regarding binding sites.

A recurring feature in any review of tools for *in silico* protein analysis, Rosetta has an extensive range of bespoke tools for the analysis of protein interactions. The general protein-protein interaction prediction framework within the suite is RosettaDock (164), which also exists as a web-server. In addition, numerous tools for the prediction of more specific interactions exist:

- RosettaLigand (165) is the premier tool within the suite for prediction of interactions between proteins and small molecules.

- RosettaMP (166) is a tool specifically for the design of membrane-spanning proteins, and predicting their interactions.
- The RosettaScripts scripting language allows for the generation of job-specific docking pipelines and scoring functions (167).

Sequence-dependent information

When confidence in the model is high, the predicted protein structure is extremely useful when attempting to predict function. As mentioned earlier, it is not always possible to predict the protein structure with any such confidence when the protein is greater than 150AA in length. There are other *in silico* parameters available to help assess a test sequence. The ProtParam tool hosted by ExPASy/Swiss institute of Bioinformatics is extremely useful for providing sequence-dependent data, as opposed to model/structure prediction-dependent data such as that provided by I-Tasser or Rosetta. Examples of the information available include the "Instability Index" of a protein as defined by Guruprasad *et al* (168). This provides an estimate of the expected stability of a protein *in vitro*, based on correlations identified between specific dipeptides and either stability or instability. The formula takes as input an amino acid sequence, and gives a score between 0-100, with an instability index below 40 indicating a stable protein. The grand average hydropathicity (GRAVY) of a protein calculated on the Kyte-Doolittle scale is also offered on this web server, as well as several other descriptive features such as the Aliphatic Index and extinction coefficients (169). Given the huge amount of variables implicating the production of a protein and its subsequent structure and function, as many as possible should be controlled for.

Towards in silico Protein Design

The three dimensional structure of a protein determines its function in most cases, and this is a function of the primary amino acid sequence of a protein. When we consider that there are 20^{100} possible variations of a 100 Amino Acid long protein sequence, the scale of possibility in protein design becomes apparent. The figure below, adapted from work by Huang *et al*, demonstrates the considerable gap between protein conformations that exist in nature, and the total conformational space that is possible (170). Given the

extent of conformational and therefore functional potential still unexplored, a priority should be to design novel proteins to combat currently unsolved problems in medicine and human health.

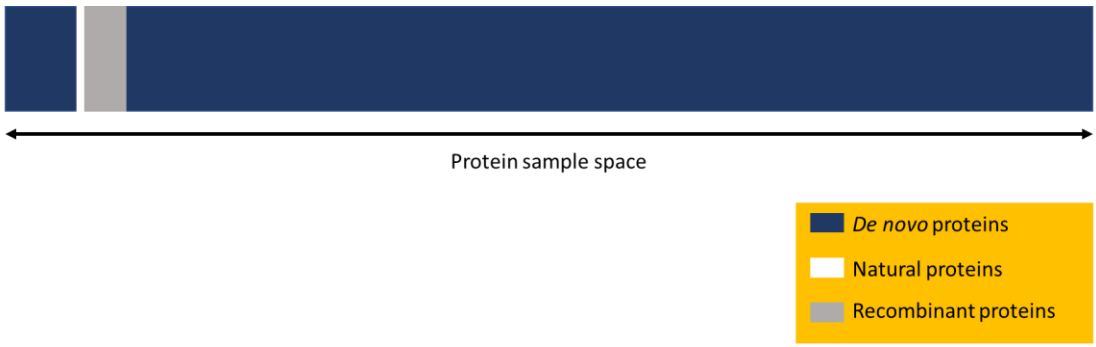


Figure 3.3: *The scale of possibility within protein conformational space. The small dots represent the fold space explored by native, naturally occurring proteins, the larger spots represent the conformational possibilities arising from directed evolution, and the blue background represents the entire conformational space.*

Despite the immense potential of rational *de novo* protein design, more than 95% of protein engineering is still carried out by inserting random mutations and selecting those which confer an advantage (171). The rational design of proteins falls into two categories, the redesign of existing proteins in a process analogous to directed evolution, and the *de novo* design of completely novel proteins. Protein redesign uses naturally occurring proteins as scaffolds, and then engineers them to introduce desired changes, such as increased stability or new functional properties (172). This will produce novel proteins, but their origins will be firmly based in the naturally occurring protein fold space. The majority of protein engineering to date has been of this nature. This method is convenient as it provides a protein backbone starting block, particularly if the desired effect represents a minor alteration in the protein's function. This becomes complicated when large numbers of amino acids are altered, since it becomes inevitable that the structure will also be altered. Native proteins are only marginally stable in many cases, so even small sequence changes can lead to dramatic changes such as aggregation or

unfolding (170). Major advances in medical science directly resulting from this degree of protein design include the humanisation of antibodies from other animal species, which entails modifying the wild type antibody to resemble human antibodies while retaining the original function. Two examples are Alemtuzumab (173) and Mepozulimab (174) for the treatment of multiple sclerosis and eosinophilic asthma respectively.

True d*e novo* protein design explores the entirety of protein sequence space, guided only by the physical interactions that control protein folding. The scale of possible protein conformations, once naturally occurring proteins are left behind is enormous. *De novo* protein design is based on the hypothesis that a protein will always fold into the shape associated with the lowest free energy state allowable by the amino acid sequence. Therefore, if an accurate method for measuring the energy of protein chains is available, in addition to a method to sample different structures and sequences it should be possible to identify sequences that fold into novel structures (170). Once the desired shape has been reached, the stability of the novel protein can be improved by making minor adjustments, maximising the difference in free energy between the desired conformation and alternatives.

There are few if any intuitive protocols for *de novo* protein design available, and generally speaking, expertise in computational biology and protein structural sciences is a minimum requirement before proceeding. Some programmes exist which make use of existing knowledge to create a framework within which non-expert design of *de novo* proteins is possible.

Intelligent System for Analysis, Model Building And Rational Design (ISAMBARD) is a suite of tools developed by Wood *et al* with the aim of facilitating the rational design of *de novo* proteins and structures, and subsequently assessing their viability. In the words of the authors, it provides "a starting point for going into the dark matter of protein fold space (172)." Geometrically regular protein structures such as α-helical coiled coils have been parameterised mathematically starting with Crick in 1953, and built on by several other groups in the following years. This allows for the parametric

design of repeat structures which can be used as scaffolds for further design in combination with other available tools and software.

The Rosetta Suite has vast capabilities for the experienced user in terms of *de novo* protein design. This includes accounting for both L and D amino acids (175), reliably designing both structure and function (176), and the design of protein switches where a *de novo* protein can change shape in response to external stimuli (177). For the non-expert user, some protocols exist for fragment-based design. This involves combining sections of several protein regions of known structure to form a new backbone. As this method uses already characterised proteins as building blocks, it is limited in the conformational space it can sample (172). This fragment-based design can be performed using tools such as RosettaRemodel (178).

Conclusion

The most apparent advantages of incorporating *in silico* analysis of protein structures into any synthetic biology pipeline are speed and cost. Thousands of potential constructs can be screened using combinations of the tools mentioned in this text, condensing possibilities down to a selection deemed most likely to be successful for the more expensive and time consuming laboratory work. In addition to this, the results of any laboratory work can be fed back into the *in silico* pipeline, thus improving any future simulations. The potential for expanding this *in silico* screening into the design of bespoke protein conformations tailored to a specific task has also been demonstrated and stands to revolutionise this and many other fields in the coming years.

References:

1. Turnbaugh, P.J., Ley, R.E., Hamady, M., Fraser-Liggett, C.M., Knight, R. and Gordon, J.I. (2007) The human microbiome project. *Nature*, **449**, 804-810.
2. Lamont, R.F., Sobel, J.D., Akins, R.A., Hassan, S.S., Chaiworapongsa, T., Kusanovic, J.P. and Romero, R. (2011) The vaginal microbiome: new information about genital tract flora using molecular based techniques. *BJOG*, **118**, 533-549.
3. Tighe, S., Afshinnekoo, E., Rock, T.M., McGrath, K., Alexander, N., McIntyre, A., Ahsanuddin, S., Bezdan, D., Green, S.J., Joye, S. *et al.* (2017) Genomic Methods and Microbiological Technologies for Profiling Novel and Extreme Environments for the Extreme Microbiome Project (XMP). *J Biomol Tech*, **28**, 31-39.
4. Ursell, L.K., Metcalf, J.L., Parfrey, L.W. and Knight, R. (2012) Defining the human microbiome. *Nutr Rev*, **70 Suppl 1**, S38-S44.
5. Lane, N. (2015) The unseen world: reflections on Leeuwenhoek (1677) 'Concerning little animals'. *Philos Trans R Soc Lond B Biol Sci*, **370**, 20140344.
6. Dewhirst, F.E., Chen, T., Izard, J., Paster, B.J., Tanner, A.C.R., Yu, W.-H., Lakshmanan, A. and Wade, W.G. (2010) The Human Oral Microbiome. *Journal of Bacteriology*, **192**, 5002.
7. Grice, E.A. and Segre, J.A. (2011) The skin microbiome. *Nat Rev Microbiol*, **9**, 244-253.
8. Fettweis, J.M., Serrano, M.G., Brooks, J.P., Edwards, D.J., Girerd, P.H., Parikh, H.I., Huang, B., Arodz, T.J., Edupuganti, L., Glascock, A.L. *et al.* (2019) The vaginal microbiome and preterm birth. *Nature Medicine*, **25**, 1012-1021.
9. Koskinen, K., Reichert, J.L., Hoier, S., Schachenreiter, J., Duller, S., Moissl-Eichinger, C. and Schöpf, V. (2018) The nasal microbiome mirrors and potentially shapes olfactory function. *Scientific Reports*, **8**, 1296.
10. O'Dwyer, D.N., Dickson, R.P. and Moore, B.B. (2016) The Lung Microbiome, Immunity, and the Pathogenesis of Chronic Lung Disease. *J Immunol*, **196**, 4839-4847.
11. Riquelme, E., Zhang, Y., Zhang, L., Montiel, M., Zoltan, M., Dong, W., Quesada, P., Sahin, I., Chandra, V., San Lucas, A. *et al.* (2019) Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. *Cell*, **178**, 795-806.e712.
12. Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J. and Walker, A.W. (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, **12**, 87.
13. Helmink, B.A., Khan, M.A.W., Hermann, A., Gopalakrishnan, V. and Wargo, J.A. (2019) The microbiome, cancer, and cancer therapy. *Nature Medicine*, **25**, 377-388.
14. Marshall, B. and Warren, J.R. (1984) UNIDENTIFIED CURVED BACILLI IN THE STOMACH OF PATIENTS WITH GASTRITIS AND PEPTIC ULCERATION. *The Lancet*, **323**, 1311-1315.
15. York, A. (2018) Fusobacterium persistence in colorectal cancer. *Nature Reviews Microbiology*, **16**, 2-2.
16. Kostic, A.D., Chun, E., Robertson, L., Glickman, J.N., Gallini, C.A., Michaud, M., Clancy, T.E., Chung, D.C., Lochhead, P., Hold, G.L. *et al.* (2013) Fusobacterium nucleatum potentiates intestinal tumorigenesis and modulates the tumor-immune microenvironment. *Cell host & microbe*, **14**, 207-215.
17. Di Domenico, E.G., Cavallo, I., Pontone, M., Toma, L. and Ensoli, F. (2017) Biofilm Producing Salmonella Typhi: Chronic Colonization and Development of Gallbladder Cancer. *Int J Mol Sci*, **18**, 1887.
18. Wang, Z., Shaheen, N.J., Whiteman, D.C., Anderson, L.A., Vaughan, T.L., Corley, D.A., El-Serag, H.B., Rubenstein, J.H. and Thrift, A.P. (2018) Helicobacter pylori

Infection Is Associated With Reduced Risk of Barrett's Esophagus: An Analysis of the Barrett's and Esophageal Adenocarcinoma Consortium. *The American journal of gastroenterology*, **113**, 1148-1155.
19. Urbaniak, C., Gloor, G.B., Brackstone, M., Scott, L., Tangney, M. and Reid, G. (2016) The Microbiota of Breast Tissue and Its Association with Breast Cancer. *Appl Environ Microbiol*, **82**, 5039-5048.
20. Zhou, B., Sun, C., Huang, J., Xia, M., Guo, E., Li, N., Lu, H., Shan, W., Wu, Y., Li, Y. *et al.* (2019) The biodiversity Composition of Microbiome in Ovarian Carcinoma Patients. *Scientific Reports*, **9**, 1691.
21. O'Connor, H., MacSharry, J., Bueso, Y.F., Lindsay, S., Kavanagh, E.L., Tangney, M., Clyne, M., Saldova, R. and McCann, A. (2018) Resident bacteria in breast cancer tissue: pathogenic agents or harmless commensals? *Discovery medicine*, **26**, 93-102.
22. Eisenhofer, R., Minich, J.J., Marotz, C., Cooper, A., Knight, R. and Weyrich, L.S. (2019) Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology*, **27**, 105-117.
23. de Goffau, M.C., Lager, S., Salter, S.J., Wagner, J., Kronbichler, A., Charnock-Jones, D.S., Peacock, S.J., Smith, G.C.S. and Parkhill, J. (2018) Recognizing the reagent microbiome. *Nature Microbiology*, **3**, 851-853.
24. Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock, G.M., Isaacs, F., Rozowsky, J. *et al.* (2016) The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology*, **17**, 53-53.
25. Escobar-Zepeda, A., Vera-Ponce de León, A. and Sanchez-Flores, A. (2015) The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. *Frontiers in Genetics*, **6**.
26. Utturkar, S.M., Klingeman, D.M., Land, M.L., Schadt, C.W., Doktycz, M.J., Pelletier, D.A. and Brown, S.D. (2014) Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics*, **30**, 2709-2716.
27. Ambardar, S., Gupta, R., Trakroo, D., Lal, R. and Vakhlu, J. (2016) High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian journal of microbiology*, **56**, 394-404.
28. Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J. and Segata, N. (2017) Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, **35**, 833.
29. Pollard, M.O., Gurdasani, D., Mentzer, A.J., Porter, T. and Sandhu, M.S. (2018) Long reads: their purpose and place. *Human Molecular Genetics*, **27**, R234-R241.
30. Bowden, R., Davies, R.W., Heger, A., Pagnamenta, A.T., de Cesare, M., Oikkonen, L.E., Parkes, D., Freeman, C., Dhalla, F., Patel, S.Y. *et al.* (2019) Sequencing of human genomes with nanopore technology. *Nature Communications*, **10**, 1869.
31. Pightling, A.W., Pettengill, J.B., Luo, Y., Baugher, J.D., Rand, H. and Strain, E. (2018) Interpreting Whole-Genome Sequence Analyses of Foodborne Bacteria for Regulatory Applications and Outbreak Investigations. *Front Microbiol*, **9**.
32. Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.-H., Whitman, W.B., Euzéby, J., Amann, R. and Rosselló-Móra, R. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, **12**, 635.
33. Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A. and Chen, W. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 6241-6246.
34. Woese, C.R. and Fox, G.E. (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5088-5090.

35. Janda, J.M. and Abbott, S.L. (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology*, **45**, 2761-2764.
36. Bukin, Y.S., Galachyants, Y.P., Morozov, I.V., Bukin, S.V., Zakharenko, A.S. and Zemskaya, T.I. (2019) The effect of 16S rRNA region choice on bacterial community metabarcoding results. *Scientific Data*, **6**, 190007.
37. Graspeuntner, S., Loeper, N., Künzel, S., Baines, J.F. and Rupp, J. (2018) Selection of validated hypervariable regions is crucial in 16S-based microbiota studies of the female genital tract. *Scientific Reports*, **8**, 9678.
38. Bodilis, J., Nsigue-Meilo, S., Besaury, L. and Quillet, L. (2012) Variable copy number, intra-genomic heterogeneities and lateral transfers of the 16S rRNA gene in Pseudomonas. *PLoS One*, **7**, e35647.
39. Fuks, G., Elgart, M., Amir, A., Zeisel, A., Turnbaugh, P.J., Soen, Y. and Shental, N. (2018) Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome*, **6**, 17.
40. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66-74.
41. Gevers, D., Kugathasan, S., Denson, L.A., Vazquez-Baeza, Y., Van Treuren, W., Ren, B., Schwager, E., Knights, D., Song, S.J., Yassour, M. *et al.* (2014) The treatment-naive microbiome in new-onset Crohn's disease. *Cell host & microbe*, **15**, 382-392.
42. Loman, N.J., Constantinidou, C., Christner, M., Rohde, H., Chan, J.Z., Quick, J., Weir, J.C., Quince, C., Smith, G.P., Betley, J.R. *et al.* (2013) A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic Escherichia coli O104:H4. *Jama*, **309**, 1502-1510.
43. Tan, G., Opitz, L., Schlapbach, R. and Rehrauer, H. (2019) Long fragments achieve lower base quality in Illumina paired-end sequencing. *Scientific Reports*, **9**, 2856.
44. Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, **30**, 2114-2120.
45. Schmieder, R. and Edwards, R. (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)*, **27**, 863-864.
46. Sokal, R.R. (1963) The Principles and Practice of Numerical Taxonomy. *Taxon*, **12**, 190-199.
47. Pollock, J., Glendinning, L., Wisedchanwet, T. and Watson, M. (2018) The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Applied and Environmental Microbiology*, **84**, e02627-02617.
48. Callahan, B.J., McMurdie, P.J. and Holmes, S.P. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J*, **11**, 2639-2643.
49. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J. and Holmes, S.P. (2016) DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods*, **13**, 581-583.
50. Fricker, A.M., Podlesny, D. and Fricke, W.F. (2019) What is new and relevant for sequencing-based microbiome research? A mini-review. *Journal of Advanced Research*, **19**, 105-112.
51. Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J. and Glöckner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*, **35**, 7188-7196.

52. Cole, J.R., Wang, Q., Fish, J.A., Chai, B., McGarrell, D.M., Sun, Y., Brown, C.T., Porras-Alfaro, A., Kuske, C.R. and Tiedje, J.M. (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res*, **42**, D633-D642.
53. McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R. and Hugenholtz, P. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J*, **6**, 610-618.
54. Yilmaz, P., Parfrey, L.W., Yarza, P., Gerken, J., Pruesse, E., Quast, C., Schweer, T., Peplies, J., Ludwig, W. and Glöckner, F.O. (2013) The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res*, **42**, D643-D648.
55. Nilsson, R.H., Larsson, K.-H., Taylor, A.F S., Bengtsson-Palme, J., Jeppesen, T.S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F.O., Tedersoo, L. *et al.* (2018) The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res*, **47**, D259-D264.
56. Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J. *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol*, **75**, 7537-7541.
57. Espejo, R.T. and Plaza, N. (2018) Multiple Ribosomal RNA Operons in Bacteria; Their Concerted Evolution and Potential Consequences on the Rate of Evolution of Their 16S rRNA. *Front Microbiol*, **9**, 1232-1232.
58. Allard, G., Ryan, F.J., Jeffery, I.B. and Claesson, M.J. (2015) SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics*, **16**, 324.
59. Vollmers, J., Wiegand, S. and Kaster, A.-K. (2017) Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! *PLoS One*, **12**, e0169662.
60. Lischer, H.E.L. and Shimizu, K.K. (2017) Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, **18**, 474-474.
61. Ghurye, J.S., Cepeda-Espinoza, V. and Pop, M. (2016) Metagenomic Assembly: Overview, Challenges and Applications. *Yale J Biol Med*, **89**, 353-362.
62. Peng, Y., Leung, H.C., Yiu, S.M. and Chin, F.Y. (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420-1428.
63. Nurk, S., Meleshko, D., Korobeynikov, A. and Pevzner, P.A. (2017) metaSPAdes: a new versatile metagenomic assembler. *Genome Res*, **27**, 824-834.
64. Lin, Y.-Y., Hsieh, C.-H., Chen, J.-H., Lu, X., Kao, J.-H., Chen, P.-J., Chen, D.-S. and Wang, H.-Y. (2017) De novo assembly of highly polymorphic metagenomic data using in situ generated reference sequences and a novel BLAST-based assembly pipeline. *BMC Bioinformatics*, **18**, 223-223.
65. Peabody, M.A., Van Rossum, T., Lo, R. and Brinkman, F.S.L. (2015) Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics*, **16**, 362.
66. Wood, D.E. and Salzberg, S.L. (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, **15**, R46.
67. Novoa, E.M., Jungreis, I., Jaillon, O. and Kellis, M. (2019) Elucidation of Codon Usage Signatures across the Domains of Life. *Molecular Biology and Evolution*.
68. Truong, D.T., Franzosa, E.A., Tickle, T.L., Scholz, M., Weingart, G., Pasolli, E., Tett, A., Huttenhower, C. and Segata, N. (2015) MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, **12**, 902.

69. Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. and Huttenhower, C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods*, **9**, 811-814.
70. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. and Morishima, K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*, **45**, D353-D361.
71. Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*, **28**, 33-36.
72. UniProt, C. (2008) The universal protein resource (UniProt). *Nucleic Acids Res*, **36**, D190-D195.
73. UniProt, C. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res*, **42**, D191-D198.
74. Franzosa, E.A., McIver, L.J., Rahnavard, G., Thompson, L.R., Schirmer, M., Weingart, G., Lipson, K.S., Knight, R., Caporaso, J.G., Segata, N. *et al.* (2018) Species-level functional profiling of metagenomes and metatranscriptomes. *Nature Methods*, **15**, 962-968.
75. Huson, D.H., Auch, A.F., Qi, J. and Schuster, S.C. (2007) MEGAN analysis of metagenomic data. *Genome Res*, **17**, 377-386.
76. Bao, E., Jiang, T., Kaloshian, I. and Girke, T. (2011) SEED: efficient clustering of next-generation sequences. *Bioinformatics*, **27**, 2502-2509.
77. Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, **28**, 27-30.
78. Burge, S., Kelly, E., Lonsdale, D., Mutowo-Muellenet, P., McAnulla, C., Mitchell, A., Sangrador-Vegas, A., Yong, S.Y., Mulder, N. and Hunter, S. (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database (Oxford)*, **2012**, bar068.
79. Yoshimura, D., Kajitani, R., Gotoh, Y., Katahira, K., Okuno, M., Ogura, Y., Hayashi, T. and Itoh, T. (2019) Evaluation of SNP calling methods for closely related bacterial isolates and a novel high-accuracy pipeline: BactSNP. *Microbial Genomics*, **5**.
80. Langmead, B. (2010) Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics*, **Chapter 11**, Unit-11.17.
81. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, **25**, 1754-1760.
82. Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford, England)*, **27**, 2987-2993.
83. Segata, N. (2018) On the Road to Strain-Resolved Comparative Metagenomics. *mSystems*, **3**, e00190-00117.
84. Gossman, W., Wasey, A. and Salen, P. (2019), *StatPearls*. StatPearls Publishing

StatPearls Publishing LLC., Treasure Island (FL).
85. Truong, D.T., Tett, A. and Pasolli, E. (2017) Microbial strain-level population structure and genetic diversity from metagenomes. **27**, 626-638.
86. Sloan, D.B., Broz, A.K., Sharbrough, J. and Wu, Z. (2018) Detecting Rare Mutations and DNA Damage with Sequencing-Based Methods. *Trends Biotechnol*, **36**, 729-740.
87. de la Cuesta-Zuluaga, J. and Escobar, J.S. (2016) Considerations For Optimizing Microbiome Analysis Using a Marker Gene. *Front Nutr*, **3**, 26.
88. Jari Oksanen, F.G.B., Michael Friendly. (2019) Vegan: Community Ecology Package.
89. McMurdie, P.J. and Holmes, S. (2013) phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One*, **8**, e61217.

90. Paradis, E., Claude, J. and Strimmer, K. (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**, 289-290.
91. Whittaker, R.H. (1972) EVOLUTION AND MEASUREMENT OF SPECIES DIVERSITY. *TAXON*, **21**, 213-251.
92. Morris, E.K., Caruso, T., Buscot, F., Fischer, M., Hancock, C., Maier, T.S., Meiners, T., Müller, C., Obermaier, E., Prati, D. *et al.* (2014) Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecol Evol*, **4**, 3514-3524.
93. Hughes, J.B., Hellmann, J.J., Ricketts, T.H. and Bohannan, B.J. (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and environmental microbiology*, **67**, 4399-4406.
94. Ricotta, C. and Podani, J. (2017) On some properties of the Bray-Curtis dissimilarity and their ecological meaning. *Ecological Complexity*, **31**, 201-205.
95. Prokopenko, D., Hecker, J., Silverman, E.K., Pagano, M., Nöthen, M.M., Dina, C., Lange, C. and Fier, H.L. (2016) Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 Genomes Project. *Bioinformatics (Oxford, England)*, **32**, 1366-1372.
96. Lozupone, C. and Knight, R. (2005) UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. *Applied and Environmental Microbiology*, **71**, 8228.
97. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, **15**, 550-550.
98. Paulson, J.N., Stine, O.C., Bravo, H.C. and Pop, M. (2013) Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, **10**, 1200.
99. Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R. and Gloor, G.B. (2014) 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.
100. De Filippis, F., Parente, E. and Ercolini, D. (2018) Recent Past, Present, and Future of the Food Microbiome. *Annual Review of Food Science and Technology*, **9**, 589-608.
101. Golob, J.L., Margolis, E., Hoffman, N.G. and Fredricks, D.N. (2017) Evaluating the accuracy of amplicon-based microbiome computational pipelines on simulated human gut microbial communities. *BMC Bioinformatics*, **18**, 283.
102. Richter, D.C., Ott, F., Auch, A.F., Schmid, R. and Huson, D.H. (2008) MetaSim: a sequencing simulator for genomics and metagenomics. *PLoS One*, **3**, e3373.
103. Leinonen, R., Sugawara, H., Shumway, M. and International Nucleotide Sequence Database, C. (2011) The sequence read archive. *Nucleic Acids Res*, **39**, D19-D21.
104. Keegan, K.P., Glass, E.M. and Meyer, F. (2016) MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. *Methods Mol Biol*, **1399**, 207-233.
105. Wilke, A., Bischof, J., Gerlach, W., Glass, E., Harrison, T., Keegan, K.P., Paczian, T., Trimble, W.L., Bagchi, S., Grama, A. *et al.* (2016) The MG-RAST metagenomics database and portal in 2015. *Nucleic Acids Res*, **44**, D590-D594.
106. Urbaniak, C., Cummins, J., Brackstone, M., Macklaim, J.M., Gloor, G.B., Baban, C.K., Scott, L., Hanlon, D.M., Burton, J.P., Francis, K.P. *et al.* (2014) Microbiota of Human Breast Tissue. *Applied and Environmental Microbiology*, **80**, 3007.
107. Xuan, C., Shamonki, J.M., Chung, A., DiNome, M.L., Chung, M., Sieling, P.A. and Lee, D.J. (2014) Microbial Dysbiosis Is Associated with Human Breast Cancer. *PLoS One*, **9**, e83744.

108. Lehouritis, P., Cummins, J., Stanton, M., Murphy, C.T., McCarthy, F.O., Reid, G., Urbaniak, C., Byrne, W.L. and Tangney, M. (2015) Local bacteria affect the efficacy of chemotherapeutic drugs. *Scientific Reports*, **5**, 14554.
109. Geller, L.T., Barzily-Rokni, M., Danino, T., Jonas, O.H., Shental, N., Nejman, D., Gavert, N., Zwang, Y., Cooper, Z.A., Shee, K. *et al.* (2017) Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science (New York, N.Y.)*, **357**, 1156-1160.
110. Nelson, M.T., Pope, C.E., Marsh, R.L., Wolter, D.J., Weiss, E.J., Hager, K.R., Vo, A.T., Brittnacher, M.J., Radey, M.C., Hayden, H.S. *et al.* (2019) Human and Extracellular DNA Depletion for Metagenomic Analysis of Complex Clinical Infection Samples Yields Optimized Viable Microbiome Profiles. *Cell Reports*, **26**, 2227-2240.e2225.
111. Wilkins, A., Chauhan, R., Rust, A., Pearson, A., Daley, F., Manodoro, F., Fenwick, K., Bliss, J., Yarnold, J. and Somaiah, N. (2018) FFPE breast tumour blocks provide reliable sources of both germline and malignant DNA for investigation of genetic determinants of individual tumour responses to treatment. *Breast cancer research and treatment*, **170**, 573-581.
112. Marotz, C.A., Sanders, J.G., Zuniga, C., Zaramela, L.S., Knight, R. and Zengler, K. (2018) Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome*, **6**, 42.
113. Wen, Y., Xiao, F., Wang, C. and Wang, Z. (2016) The impact of different methods of DNA extraction on microbial community measures of BALF samples based on metagenomic data. *American journal of translational research*, **8**, 1412-1425.
114. Correa, P. and Piazuelo, M.B. (2011) Helicobacter pylori Infection and Gastric Adenocarcinoma. *US gastroenterology & hepatology review*, **7**, 59-64.
115. Shang, F.-M. and Liu, H.-L. (2018) Fusobacterium nucleatum and colorectal cancer: A review. *World Journal of Gastrointestinal Oncology*, **10**, 71-81.
116. Jiang, S., Zhang, S., Langenfeld, J., Lo, S.-C. and Rogers, M.B. (2007) Mycoplasma infection transforms normal lung cells and induces bone morphogenetic protein 2 expression by post-transcriptional mechanisms. *Journal of Cellular Biochemistry*, **104**, 580-594.
117. Cummins, J. and Tangney, M. (2013) Bacteria and tumours: causative agents or opportunistic inhabitants? *Infectious Agents and Cancer*, **8**, 11.
118. Flores Bueso, Y., Lehouritis, P. and Tangney, M. (2018) In situ biomolecule production by bacteria; a synthetic biology approach to medicine. *Journal of Controlled Release*, **275**, 217-228.
119. Zheng, J.H., Nguyen, V.H., Jiang, S.-N., Park, S.-H., Tan, W., Hong, S.H., Shin, M.G., Chung, I.-J., Hong, Y., Bom, H.-S. *et al.* (2017) Two-step enhanced cancer immunotherapy with engineered Salmonella typhimurium secreting heterologous flagellin. *Sci Transl Med*, **9**, eaak9537.
120. Jervis-Bardy, J., Leong, L.E.X., Marri, S., Smith, R.J., Choo, J.M., Smith-Vaughan, H.C., Nosworthy, E., Morris, P.S., O'Leary, S., Rogers, G.B. *et al.* (2015) Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of Illumina MiSeq data. *Microbiome*, **3**, 19.
121. Strong, M.J., Xu, G., Morici, L., Splinter Bon-Durant, S., Baddoo, M., Lin, Z., Fewell, C., Taylor, C.M. and Flemington, E.K. (2014) Microbial contamination in next generation sequencing: implications for sequence-based analysis of clinical samples. *PLoS Pathog*, **10**, e1004437-e1004437.
122. Kulakov, L.A., McAlister, M.B., Ogden, K.L., Larkin, M.J. and O'Hanlon, J.F. (2002) Analysis of bacteria contaminating ultrapure water in industrial systems. *Applied and environmental microbiology*, **68**, 1548-1555.

123. Grizzle, W.E., Bell, W.C. and Sexton, K.C. (2010) Issues in collecting, processing and storing human tissues and associated information to support biomedical research. *Cancer biomarkers : section A of Disease markers*, **9**, 531-549.
124. Do, H. and Dobrovic, A. (2015) Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization. *Clinical Chemistry*, **61**, 64.
125. Chen, L., Liu, P., Evans, T.C. and Ettwiller, L.M. (2017) DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science*, **355**, 752.
126. Stewart, C.J., Fatemizadeh, R., Parsons, P., Lamb, C.A., Shady, D.A., Petrosino, J.F. and Hair, A.B. (2019) Using formalin fixed paraffin embedded tissue to characterize the preterm gut microbiota in necrotising enterocolitis and spontaneous isolated perforation using marginal and diseased tissue. *BMC Microbiology*, **19**, 52.
127. Racsa, L.D., DeLeon-Carnes, M., Hiskey, M. and Guarner, J. (2017) Identification of bacterial pathogens from formalin-fixed, paraffin-embedded tissues by using 16S sequencing: retrospective correlation of results to clinicians' responses. *Human Pathology*, **59**, 132-138.
128. Qiagen. (2012) QIAamp DNA FFPE Tissue Handbook.
129. Chakravorty, S., Helb, D., Burday, M., Connell, N. and Alland, D. (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of microbiological methods*, **69**, 330-339.
130. Ranjan, R., Rani, A., Metwally, A., McGee, H.S. and Perkins, D.L. (2016) Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem Biophys Res Commun*, **469**, 967-977.
131. Zhang, C., Cleveland, K., Schnoll-Sussman, F., McClure, B., Bigg, M., Thakkar, P., Schultz, N., Shah, M.A. and Betel, D. (2015) Identification of low abundance microbiome in clinical samples using whole genome sequencing. *Genome biology*, **16**, 265-265.
132. Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E. *et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res*, **21**, 494-504.
133. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C. and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, **27**, 2194-2200.
134. Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. and Kelley, S.T. (2011) Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*, **8**, 761.
135. Davis, N.M., Proctor, D., Holmes, S.P., Relman, D.A. and Callahan, B.J. (2017) Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *bioRxiv*.
136. Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I. *et al.* (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335.
137. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T. *et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*, **36**, 338.
138. Hogan, G., Walker, S., Turnbull, F., Curiao, T., Morrison, A.A., Flores, Y., Andrews, L., Claesson, M.J., Tangney, M. and Bartley, D.J. (2019) Microbiome analysis as a platform R&D tool for parasitic nematode disease management. *ISME J*.
139. Danino, T., Prindle, A., Kwong, G.A., Skalak, M., Li, H., Allen, K., Hasty, J. and Bhatia, S.N. (2015) Programmable probiotics for detection of cancer in urine. *Sci Transl Med*, **7**, 289ra284.

140. Van Dessel, N., Swofford, C.A. and Forbes, N.S. (2015) Potent and tumor specific: arming bacteria with therapeutic proteins. *Ther Deliv*, **6**, 385-399.
141. Steidler, L., Rottiers, P. and Coulie, B. (2009) Actobiotics as a novel method for cytokine delivery. *Annals of the New York Academy of Sciences*, **1182**, 135-145.
142. Sedighi, M., Zahedi Bialvaei, A., Hamblin, M.R., Ohadi, E., Asadi, A., Halajzadeh, M., Lohrasbi, V., Mohammadzadeh, N., Amiriani, T., Krutova, M. *et al.* (2019) Therapeutic bacteria to combat cancer; current advances, challenges, and opportunities. *Cancer Med*, **8**, 3167-3181.
143. Yang, J. and Zhang, Y. (2015) Protein Structure and Function Prediction Using I-TASSER. *Curr Protoc Bioinformatics*, **52**, 5.8.1-5.8.15.
144. Kaufmann, K.W., Lemmon, G.H., Deluca, S.L., Sheehan, J.H. and Meiler, J. (2010) Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry*, **49**, 2987-2998.
145. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L. *et al.* (2018) SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, **46**, W296-w303.
146. R.Evans. (2018) De novo structure prediction with deep-learning based scoring. *Thirteenth Critical Assessment of Techniques for Protein Structure Prediction*.
147. Chothia, C. and Lesk, A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J*, **5**, 823-826.
148. Marti-Renom, M.A., Stuart, A.C., Fiser, A., Sanchez, R., Melo, F. and Sali, A. (2000) Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure*, **29**, 291-325.
149. Lee J., F.P.L., Zhang Y. (2017) *Ab Initio Protein Structure Prediction*

150. Kim, D.E., Chivian, D. and Baker, D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic acids research*, **32**, W526-531.
151. Ovchinnikov, S., Park, H., Kim, D.E. and DiMaio, F. (2018) Protein structure prediction using Rosetta in CASP12. **86 Suppl 1**, 113-121.
152. Khor, B.Y., Tye, G.J., Lim, T.S. and Choong, Y.S. (2015) General overview on structure prediction of twilight-zone proteins. *Theor Biol Med Model*, **12**, 15-15.
153. Webb, B. and Sali, A. (2014) Protein structure modeling with MODELLER. *Methods in molecular biology (Clifton, N.J.)*, **1137**, 1-15.
154. Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C. (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucleic acids research*, **31**, 3381-3385.
155. Kelley, L.A., Mezulis, S., Yates, C.M., Wass, M.N. and Sternberg, M.J. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. **10**, 845-858.
156. Jones, D.T. (2001) Predicting novel protein folds by using FRAGFOLD. *Proteins*, **Suppl 5**, 127-132.
157. Haas, J., Barbato, A., Behringer, D., Studer, G., Roth, S., Bertoni, M., Mostaguir, K., Gumienny, R. and Schwede, T. (2018) Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12. *Proteins: Structure, Function, and Bioinformatics*, **86**, 387-398.
158. Yang, J., Roy, A. and Zhang, Y. (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics (Oxford, England)*, **29**, 2588-2595.
159. Kozakov, D., Hall, D.R., Xia, B., Porter, K.A., Padhorny, D., Yueh, C., Beglov, D. and Vajda, S. (2017) The ClusPro web server for protein-protein docking. *Nat Protoc*, **12**, 255-278.

160. de Vries, S.J., van Dijk, M. and Bonvin, A.M. (2010) The HADDOCK web server for data-driven biomolecular docking. *Nat Protoc*, **5**, 883-897.
161. Torchala, M., Moal, I.H., Chaleil, R.A., Fernandez-Recio, J. and Bates, P.A. (2013) SwarmDock: a server for flexible protein-protein docking. *Bioinformatics (Oxford, England)*, **29**, 807-809.
162. Trott, O. and Olson, A.J. (2010) AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, **31**, 455-461.
163. Alhossary, A., Handoko, S.D., Mu, Y. and Kwoh, C.K. (2015) Fast, accurate, and reliable molecular docking with QuickVina 2. *Bioinformatics (Oxford, England)*, **31**, 2214-2216.
164. Marze, N.A., Roy Burman, S.S., Sheffler, W. and Gray, J.J. (2018) Efficient flexible backbone protein–protein docking for challenging targets. *Bioinformatics (Oxford, England)*, **34**, 3461-3469.
165. Lemmon, G. and Meiler, J. (2012) Rosetta Ligand docking with flexible XML protocols. *Methods in molecular biology (Clifton, N.J.)*, **819**, 143-155.
166. Alford, R.F., Koehler Leman, J., Weitzner, B.D., Duran, A.M., Tilley, D.C., Elazar, A. and Gray, J.J. (2015) An Integrated Framework Advancing Membrane Protein Modeling and Design. *PLOS Computational Biology*, **11**, e1004398.
167. Fleishman, S.J., Leaver-Fay, A., Corn, J.E., Strauch, E.-M., Khare, S.D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G. *et al.* (2011) RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. *PLOS ONE*, **6**, e20161.
168. Guruprasad, K., Reddy, B.V. and Pandit, M.W. (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein engineering*, **4**, 155-161.
169. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D. and Bairoch, A. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic acids research*, **31**, 3784-3788.
170. Huang, P.-S., Boyken, S.E. and Baker, D. (2016) The coming of age of de novo protein design. *Nature*, **537**, 320.
171. Perkel, J.M. (2019) The computational protein designers. *Nature*, **571**, 585-587.
172. Wood, C.W., Heal, J.W., Thomson, A.R., Bartlett, G.J., Ibarra, A.A., Brady, R.L., Sessions, R.B. and Woolfson, D.N. (2017) ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design. *Bioinformatics*, **33**, 3043-3050.
173. Havrdova, E., Horakova, D. and Kovarova, I. (2015) Alemtuzumab in the treatment of multiple sclerosis: key clinical trial results and considerations for use. *Ther Adv Neurol Disord*, **8**, 31-45.
174. Poulakos, M.N., Cargill, S.M., Waineo, M.F. and Wolford, A.L., Jr. (2017) Mepolizumab for the treatment of severe eosinophilic asthma. *American journal of health-system pharmacy : AJHP : official journal of the American Society of Health-System Pharmacists*, **74**, 963-969.
175. Hosseinzadeh, P., Bhardwaj, G., Mulligan, V.K., Shortridge, M.D., Craven, T.W., Pardo-Avila, F., Rettie, S.A., Kim, D.E., Silva, D.-A., Ibrahim, Y.M. *et al.* (2017) Comprehensive computational design of ordered peptide macrocycles. *Science*, **358**, 1461.
176. Koepnick, B., Flatten, J., Husain, T., Ford, A., Silva, D.-A., Bick, M.J., Bauer, A., Liu, G., Ishida, Y., Boykov, A. *et al.* (2019) De novo protein design by citizen scientists. *Nature*, **570**, 390-394.

177. Langan, R.A., Boyken, S.E., Ng, A.H., Samson, J.A., Dods, G., Westbrook, A.M., Nguyen, T.H., Lajoie, M.J., Chen, Z., Berger, S. *et al.* (2019) De novo design of bioactive protein switches. *Nature*, **572**, 205-210.
178. Huang, P.-S., Ban, Y.-E.A., Richter, F., Andre, I., Vernon, R., Schief, W.R. and Baker, D. (2011) RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. *PLOS ONE*, **6**, e24109.

Chapter II

"Assessing bioinformatic amplicon sequencing contamination control strategies via mock bacterial communities"

-This manuscript is awaiting the publication of relevant Chapter IV data prior to submission.

Sidney Walker, Marcus J. Claesson, Mark Tangney

ABSTRACT

Background Alterations in the microbiological signal caused by the presence of contaminant DNA are a major issue in microbiome studies. Considerable effort has gone into developing strategies and tools to identify and bioinformatically stop environmental bacterial contamination from distorting biological signal.

Aim A consistently effective contamination control strategy incorporating biological and bioinformatic methods, and the ability to validate this method, particularly when sampling from new environments for the first time would be of considerable benefit to future microbiome research.

Methods This study compares options for the removal of contaminant DNA and proposes an optimal approach. The effect of these on the results of a sequencing study were validated through the use of a mock community. The effectiveness of contamination control at the extreme end of the spectrum is demonstrated, using samples featuring low levels of bacterial biomass, which are then formalin fixed and embedded in paraffin. This required the samples to be subjected to a number of different reagents during the DNA extraction and purification process, as is necessary for bacterial sequence analysis of FFPE samples, providing many potential sources of contamination.

Results Even in samples that consisted mainly of contaminant DNA, it was possible to reliably isolate the true sample DNA with a retroactive bioinformatics method based on the use of negative controls, to an extent where the effect on any downstream microbiome analysis would be negligible, as verified by the mock community.

Conclusions All labs carrying out sequencing experiments, but particularly those dealing with low biomass or otherwise challenging samples, should carry out a similar analysis validating their own biological and bioinformatic contamination control methods to gauge the degree to which environmental contaminants may affect future sequence based analysis.

INTRODUCTION

Increasingly affordable culture-independent microbial surveys have sparked a surge in research furthering the understanding of the relationships between bacteria, but also viruses and fungi, and their hosts. This can be done primarily through the sequencing of amplified marker genes such as the 16S rRNA gene region in bacteria and ITS region in fungi or by non-specific sequencing of all DNA in a community using whole genome sequencing (1,2). Despite the advances that these new methods have yielded, there has been a realization within the field of sequence-based microbiome analysis of the threat posed by environmental contaminant DNA to the accuracy and reproducibility of research in this area. This has progressed to the extent that the validity of many microbiome surveys of low biomass environments have been rightly questioned(2,3). In response several groups have published both wet- and dry-lab methodologies for mitigating the impact of this contaminant DNA. These range from suggested protocols for negative controls(4), to bespoke bioinformatics tools for the retrospective identification and removal of contaminant sequences.

Two advances in bacterial contamination removal are SourceTracker(5) and Decontam(6). SourceTracker predicts both the proportion of contamination and its origins, using a Bayesian approach combined with Gibbs sampling. The results of a sequencing experiment are divided into “sink” samples, and negative controls denoted as “source” samples. The algorithm divides “sink” samples into individual reads, each of which can be assigned to one of the “source” environments, or an unknown source if it is not predicted to have originated from a negative control sample. In summary, SourceTracker provides a clear picture of the extent to which negative controls have affected samples, but does not identify the taxa in question. The Decontam algorithm is more direct in its approach, and removes contaminant sequences based on two assumptions; (i)that sequences of contaminant origin are likely to inversely correlate with sample DNA concentration, and (ii) that contaminant DNA will have a higher prevalence in negative control samples(6).

DNA contamination typically arises from DNA extraction kits, PCR reagents and the general lab environment (7). While all types of samples can be affected, low biomass

samples are particularly susceptible. These can be contaminated to a degree where the true microbial composition of the sample is completely altered(2). In addition to contamination from extraction kits and enzymes used in PCR reactions, certain samples require the use of additional buffers and reagents in order to extract bacterial DNA suitable for sequencing experiments. Two examples that are becoming much more prevalent in microbiome analysis are:

- Samples with an overwhelming ratio of host DNA to bacterial DNA, such as non-tract biopsies. These need to be treated with bacterial enrichment solutions, many of which are not sterile(8).
- Formalin Fixed, Paraffin Embedded samples. The formalin fixing process damages and crosslinks the DNA, and several steps must be taken to account for this(9,10).

In both these cases enzymatic action is required, for purposes such as the depletion of host DNA, repair of DNA damaged by the formalin fixing process, or to lyse more resilient bacteria. Enzymatic action means that the reagents used cannot be autoclaved and therefore are not sterile.

Several groups have recommended a more all-encompassing negative control strategy, incorporating many possible sources of contamination as well as positive controls or standards(11). It may seem logical to sequence every possible source of contamination if the circumstances allow, to be as thorough as possible. In practice this can lead to additional problems if not combined with correct retroactive bioinformatics-based contamination removal. This is because cross contamination between samples may occur, when DNA originating from the sample environment is transferred between samples. There are several causes for this but excluding human error during pipetting these are often very difficult if not impossible to control for in situ. Bacteria can become aerosolized when samples are being loaded into wells in PCR plates, or when the cover is removed from the PCR plate following the PCR reaction(12). There is a phenomenon known as “Tag switching” where sample barcodes migrate between wells(13). Barcodes can also be mistaken between samples as a result of sequencing miscalls due to poor sequence quality(14), a phenomenon which is thought to occur in between 0.6 and 6%

of all reads in a sequencing run(15). Cross contamination has a negligible impact on samples unless they are of extraordinarily low biomass, but negative controls with low quantities of input DNA are particularly susceptible to artefacts of this nature as they may have very few microbes and so can appear to be dominated by an microbial sequence that is in reality just highly abundant in samples. This is of particular concern when carrying out the conservative contamination removal by subtraction method such as the one possible to implement in the QIIME pipeline(16). Other commonly used solutions include filtering out low abundance taxa below an arbitrary threshold(17), this method would run the risk of also removing rare genuine taxa from the dataset. More importantly, if a source of contamination was abundant enough to have a significant impact on downstream analysis, it would not be removed by this method(6).

Of the recent 16S rRNA gene sequencing surveys present in the literature, the most challenging sample type that stands out is formalin-fixed paraffin embedded (FFPE) tissue(18). Here we test if samples having both low biomass and numerous plausible sources of contamination, could be reliably and reproducibly explored. We were able to isolate the endogenous biological signal, using a variety of negative controls as suggested by Eisenhofer *et al* (12), combined with bioinformatic contamination elimination, and crucially were able to validate the effectiveness of our approach by using mock bacterial communities in FFPE. We show that when a robust negative control strategy is combined with an effective bioinformatic contamination removal strategy, the effect of contamination on the overall biological signal can be almost entirely eliminated. This opens the door for microbiome investigations into a wide range of samples types, not limited to FFPE, many of which are currently treated with some degree of scepticism due to their susceptibility to contamination.

MATERIALS AND METHODS

All laboratory work was carried out by other members of the Tangney lab. This study simulated a challenging sample condition by creating a bespoke mock sample of mouse tumor cells and 4 different bacterial taxa. These were then formalin-fixed and paraffin embedded in the same way as genuine patient samples. Following this they are treated with a number of reagents and solutions to decrosslink the DNA, enrich and repair bacterial DNA and ultimately prepare a 16S sequencing library (8,19,20).

Mock Community Design

Four known bacterial species were used in this mock community. They were *E. coli* – K12 MG1655, *Salmonella* Typhimurium 7207, *Staphylococcus aureus* newman, and *Streptococcus agalactiae* COH1. To replicate a clinical FFPE sample, which is typically a biopsy, as closely as possible, *Mus musculus* mammary gland cancer cells (4T1) were also added. These cells were pelleted and suspended in formalin, before being added to a sterile mould with an equal volume of sterile agar. This was then dehydrated and paraffin embedded as per the protocol outlined in Chapter IV.

FFPE sample preparation

The formalin fixed biological standards were then treated in the same way as FFPE samples and processed according to an in-house protocol (for further details, see Chapter IV)

V3-V4 16S rRNA sequencing

Genomic DNA was amplified using 16S rRNA gene amplicon polymerase chain reaction (PCR) primers targeting the hypervariable V3–V4 region of the 16S rRNA gene: V3–V4 forward, 5′-TCGTCGGCAGCGTCAGATGTGT ATAAGAGACAGCCTACGGGNGGCWGCAG-3′ and V3–V4 reverse, 5′-GTCTCGTGGGCTCGGAGATGTGTA TAAGAGACAGGACTACHVGGGTATCTAATCC-3' (Illumina 16S Metagenomic Sequencing Protocol, Illumina, CA, USA). A 35-µl PCR was performed for each sample per the following recipe: 3.5 µl of template DNA, 17.5 µl of KAPA HiFi HotStart ReadyMix (Roche), 0.7 µl of both primers (initial concentration, 10 pmol/µl), 0.1 µg/µl bovine serum albumin fraction V (Sigma), and 8 µl of 10 mM TrisCl (Qiagen). Thermal cycling was completed in an Eppendorf Mastercycler per the directions in the 'Amplicon PCR' section of the '16S Metagenomic Sequencing Library Preparation' protocol (Illumina). Amplification was confirmed by running 5 µl of PCR product on a 1.5% agarose gel at 70 volts for 80 min, followed by imaging on a Gel Doc EZ System (Bio-Rad). The product was ~450 base pairs (bp) in size. PCR-positive products were cleaned per the 'PCR CleanUp' section of the Illumina protocol, with the exception that drying times were reduced to half the prescribed duration to account for the additional drying that occurs in a laminar airflow hood. Sequencing libraries were then prepared using the Nextera XT Index Kit (Illumina) and cleaned per the Illumina protocol. Libraries were quantified using a Qubit fluorometer (Invitrogen) using the 'High Sensitivity' assay. Sample processing was subsequently completed at Genewiz inc. Samples were normalised, pooled and underwent a paired-end 300 bp run on the Illumina MiSeq platform.

Bioinformatics analysis

The quality of the paired-end sequence data was initially visualised using FastQC v0.11.6, and then filtered and trimmed using Trimmomatic v0.36 to ensure a minimum average quality of 25. The remaining high-quality reads were then imported into the R environment v3.4.4 for analysis with the DADA2 package v1.8.0. After further quality filtering, error correction and chimera removal, the raw reads generated by the sequencing process were refined into a table of Amplicon Sequence Variants (ASVs), which can be considered analogous to OTU's, and their distribution among the samples. The ASVs were classified using the classify.seqs function in Mothur and the RDP database.

Reference Data Generation

The sequences present in samples were blasted against a database of the 16S rRNA gene regions of the known input bacteria. Only reads with 97% sequence similarity were retained, this was to allow for sequencing miscalls, and DNA damage due to the formalin fixing process. The database was formatted using the *makeblastdb* facility and the search was carried out using *blastn,* both programmes are contained in the BLAST+ toolkit. This reference table will be referred to as the "Reference table". BLAST was used with the exact reference sequences as opposed to relying on the classification from Mothur, as most species classification algorithms have a degree of trade-off between accuracy and speed(21).

Statistical analysis

All statistical analysis was carried out in the R environment v3.6.0. The Vegan package v2.5.2, Phyloseq package v1.2.4 and the Ape package v5.1 were used to calculate beta diversity. All statistical comparisons between grouped samples were carried out using Wilcoxon signed-rank tests. Visualisation was performed using the Ggplot2 package v3.2.1.

Bioinformatic retrospective contamination removal

Three different approaches to contamination removal were compared in this study based on options often used in relevant sequencing studies.

- Contamination removal by subtraction:
  Negative controls and samples were divided into two separate count tables, and any sequence present in the negative controls is removed from the dataset.
- Decontam algorithm(6)
  Use of Decontam with default settings, using negative control samples as a guide.
- Combined guided approach
  Combination of two published bioinformatic contamination control tools, Decontam and SourceTracker(22). This is the recommended method and is described in greater detail in Figure 1.

These methods were validated biological standard mock community instead of patient samples, allowing for a quantifiable measure of their effectiveness by comparing them to the Reference table as shown in Figures 2-6.

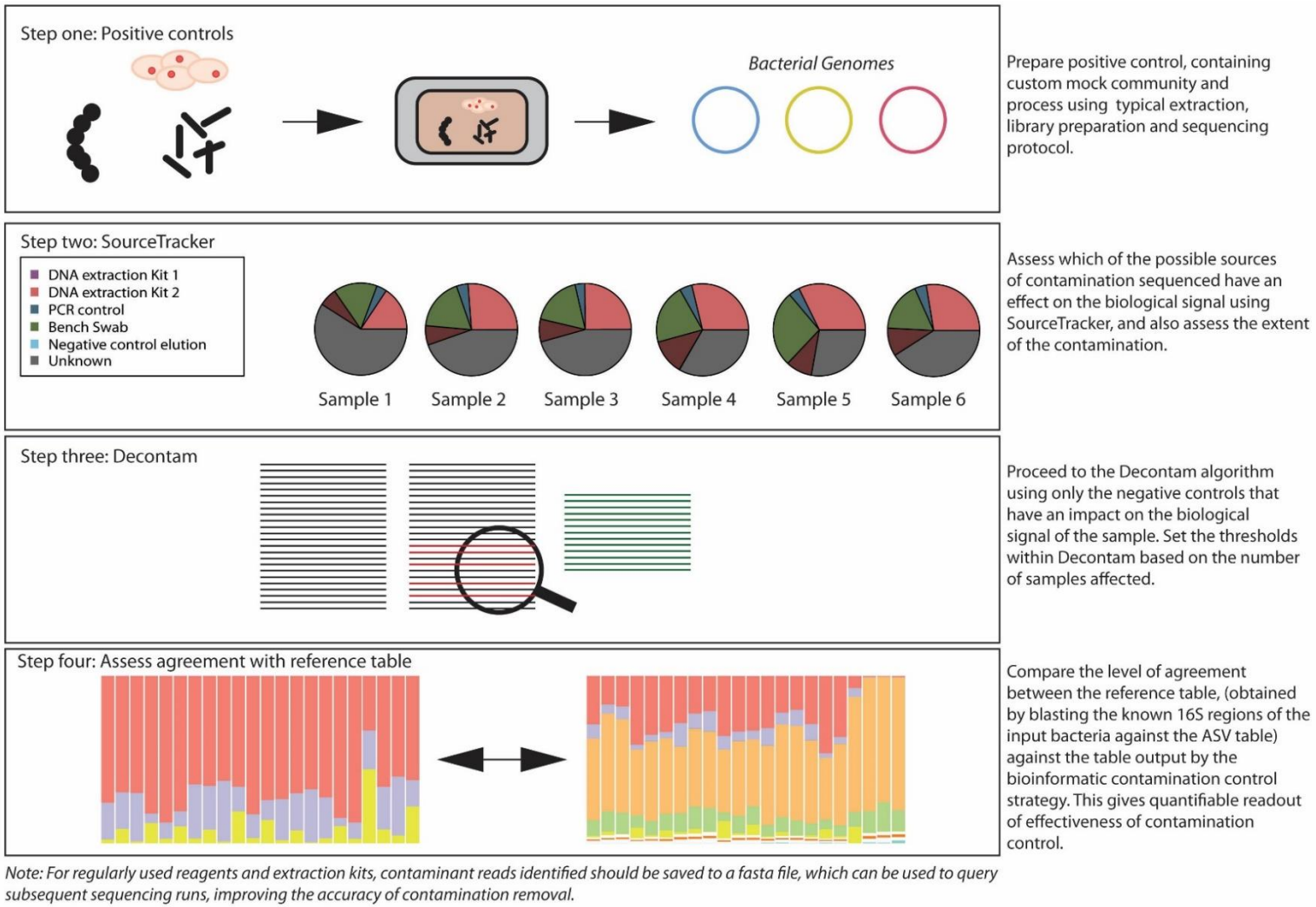


Figure 1: Proposed workflow for contamination removal and subsequent validation. *This involves preparing the positive controls in a manner consistent with the patient samples, extracting DNA, and sequencing. Resulting data is assessed for contamination, using negative controls and bioinformatic techniques. The entire process is then validated by comparison with a reference table containing only the known bacterial species used when preparing the standard community.*

RESULTS

Figure 2 shows a comparison between all bacterial sequences in the dataset, and the Reference table. It is clear that contaminant bacterial DNA has had a major influence on the true biological signal as there is considerable difference between the two groups in terms of sample composition. Many obvious contaminant families such as Xanthomonadaceae (members of which are typically environmental organisms(23)) are present in the "Test" group only. The three families found in the Reference table are all significantly decreased in the "Test" group in this instance ($p = < 0.001$). Finally, figure 2B models the effect this contamination would have on any downstream analysis by examining the distance between the paired samples on a PcOA plot.

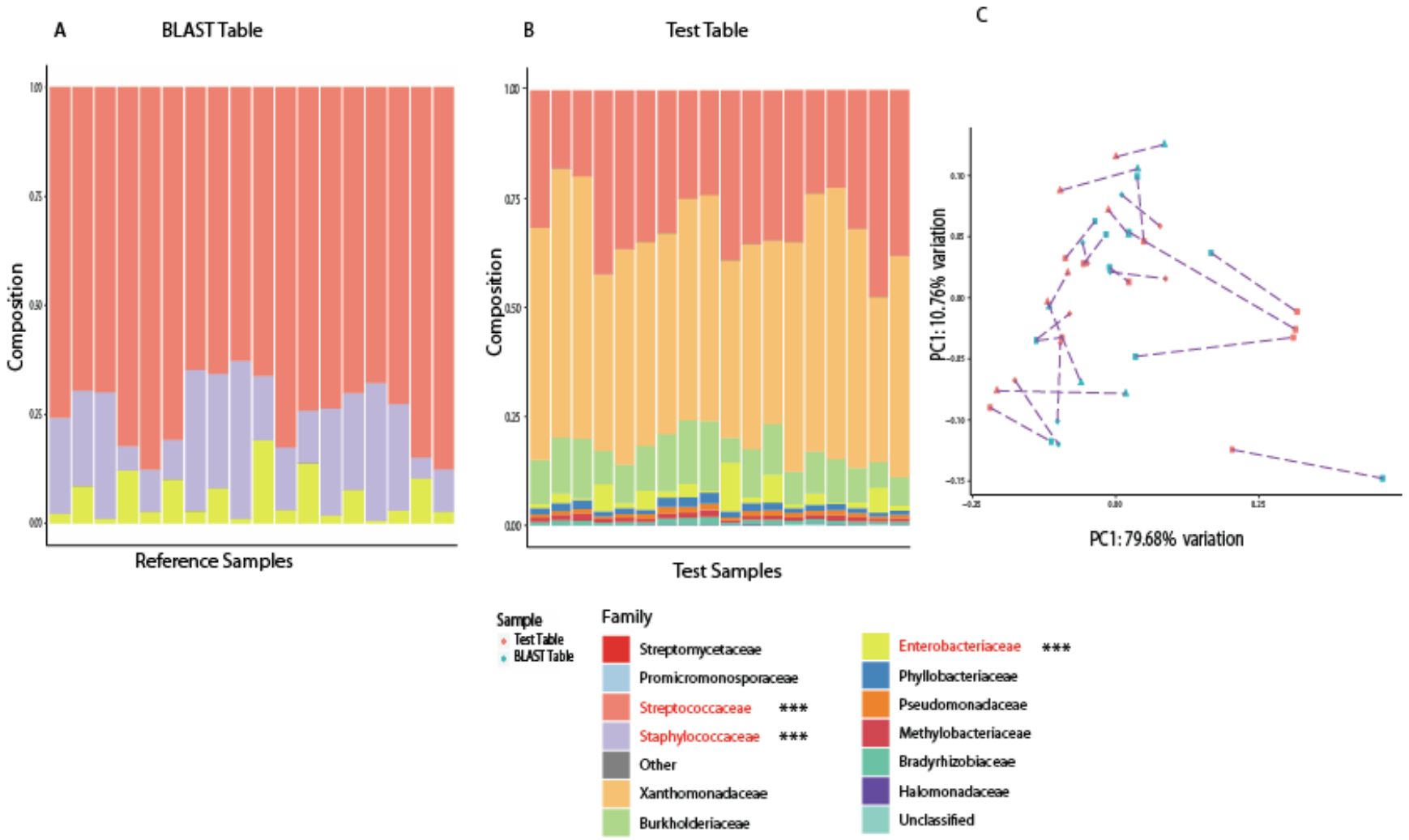


Figure 2: ***Comparison of (A) Reference table (obtained by BLASTing known input sequences against ASV table generated by DADA2) vs (BB) full count table with no contamination based modifications. The impact contamination has on sample beta diversity is shown in (CC), distance matrix was calculated using Bray-Curtis dissimilarity.***

The following figures highlight the effectiveness of different negative control strategies.

Figure 3 shows the results of a contamination removal by the commonly used subtraction approach, where any sequence found in the negative controls is removed from the count table, which is then compared to the Reference table. While the sample composition tables look more similar than in Figure 2, there are statistically significantly lower levels of *Enterobacteriaceae* (p= 7.2e-6) and higher levels of *Staphylococcaceae* (p = 0.013). Despite the apparent visual similarity seen in terms of sample composition, the euclidean distance between paired samples indicates that contamination in combination with the heavy-handed contamination removal strategy would impair the accuracy of any further analysis using this data.

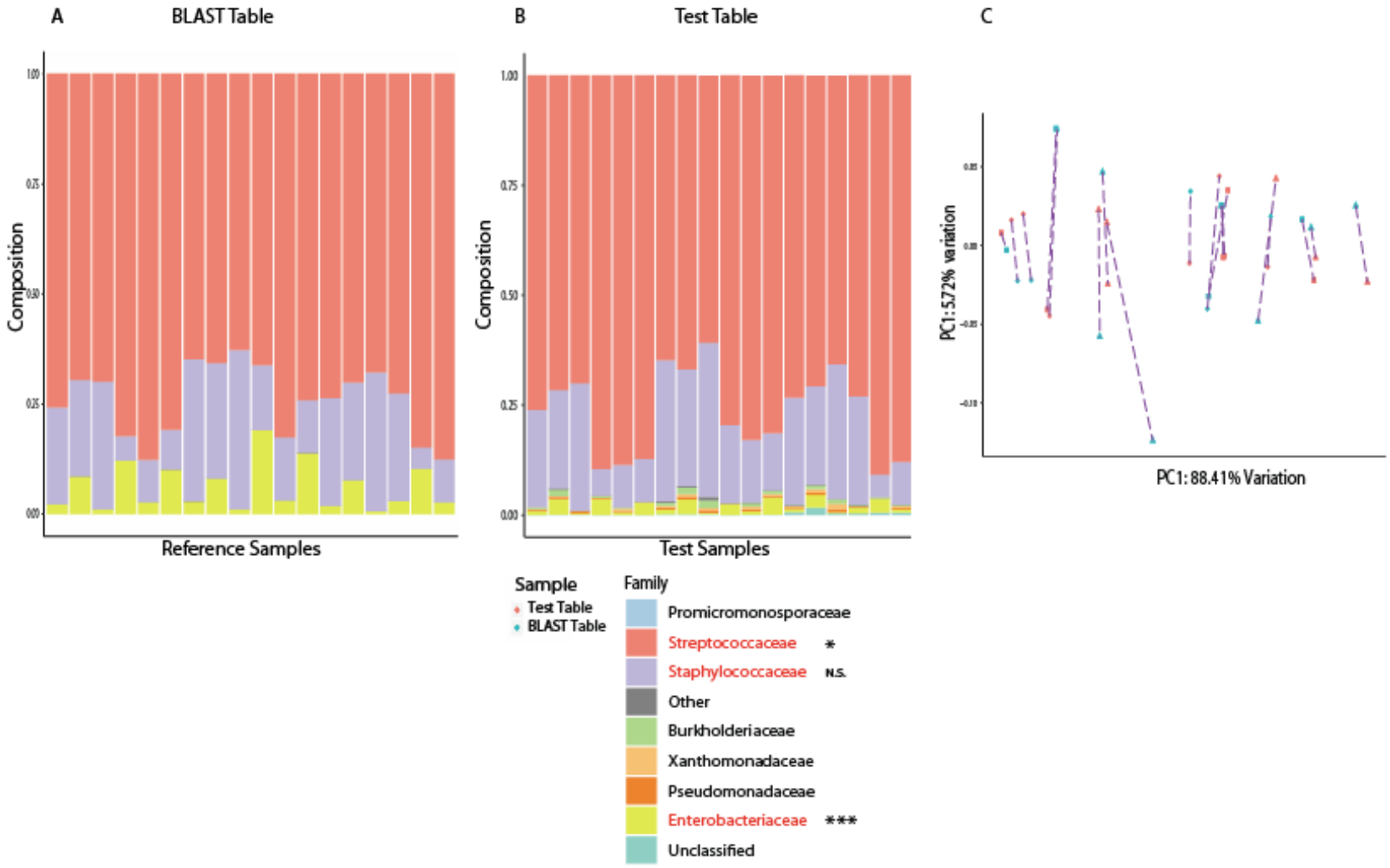


Figure 3: ***Comparison of (A) Reference Table vs (C) count table with contamination removed by subtraction.*** *All ASV's found in negative controls are removed from entire dataset.. The impact contamination has on sample beta diversity is shown in (B), with the distance matrix calculated using Bray-Curtis dissimilarity.*

Figure 4 shows the results of a generic implementation of the Decontam algorithm, using the negative control samples as a reference. In this instance the algorithm has no tangible effect, as we can see from the marked differences in sample composition at the family level with statistically significant reductions in the levels of *Enterobacteriaceae* (p = 7.2e-6), *Staphylococcaceae* (p = 7.6e-6), and *Streptococcaceae* (p = 7.6e-6). In addition and the euclidean distance between paired samples on the PcOA plot shows that any downstream analysis would be inaccurate.

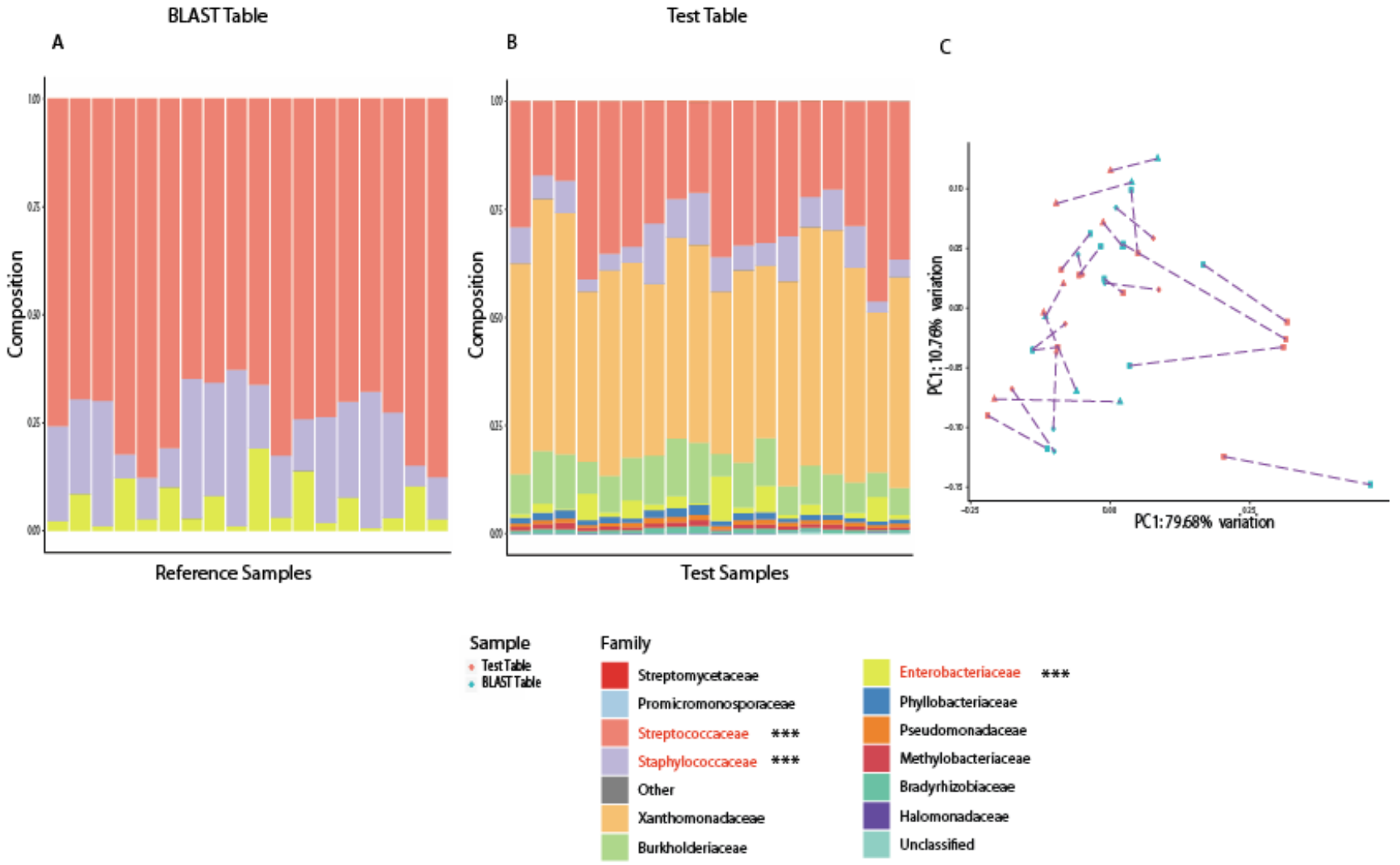


Figure 4: ***Comparison of (A) REFERENCE table vs (BB) count table with contamination removed by blind use of Decontam algorithm.*** *The impact contamination has on sample beta diversity is shown in (CC), with distance matrix calculated by Bray Curtis dissimilarity.*

Results of proposed contamination removal workflow

Figure 5 shows the output of the SourceTracker algorithm, which is used to assess the proportion of bacteria originating in the surrounding environment (i.e. negative controls) present in the samples when applied to the test data. The grey shaded region in the pie charts is the proportion bacteria not attributable to environmental contamination by source tracker. Thus contamination was ubiquitous among the samples analysed, with the highest contributions coming from "extraction negative" and "Negative control solution 3". This information was used to inform the next step in our contamination removal strategy.

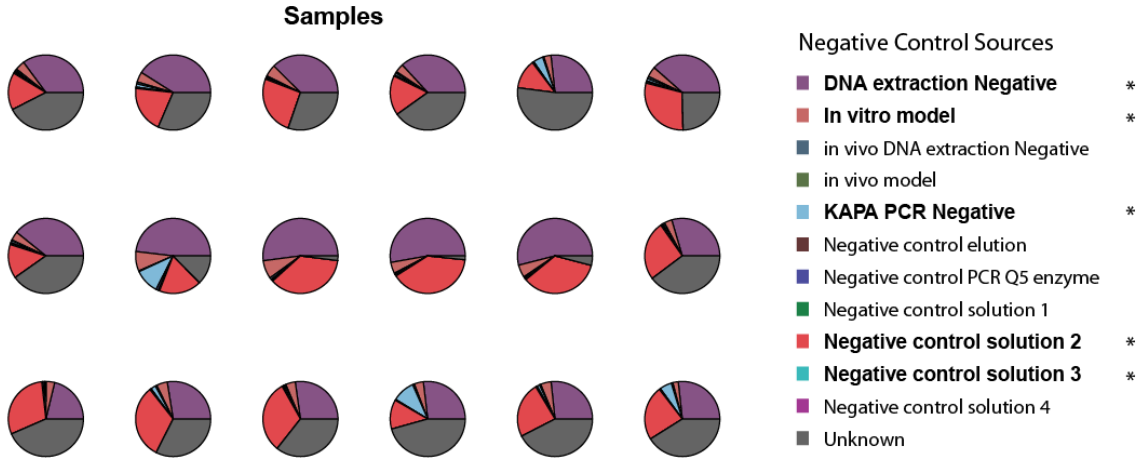


Figure 5: Sourcetracker output, mapping the sequences detected in samples to sources of contamination. *Grey shaded region indicates suspected genuine bacterial sequences of sample origin, other shaded regions correspond to different negative controls. The negative controls impacting on samples that should be used for contamination removal are identified in bold with stars.*

Figure 6 compares the REFERENCE table with the count table resulting from this contamination removal strategy, involving both SourceTracker and Decontam, as outlined in Figure 1. The sample composition plot in this case is very similar to the reference table, with the only difference being low levels of the common contaminant family *Pseudomondaceae*.

Comparisons of sample composition at the family level of taxonomy found no difference in the levels of any of the taxa originating in the samples. Here, unlike in the case of the contamination removal by subtraction, we see that there is no significant difference in beta diversity between paired samples, suggesting that contamination is unlikely to markedly affect any downstream microbiome analysis.

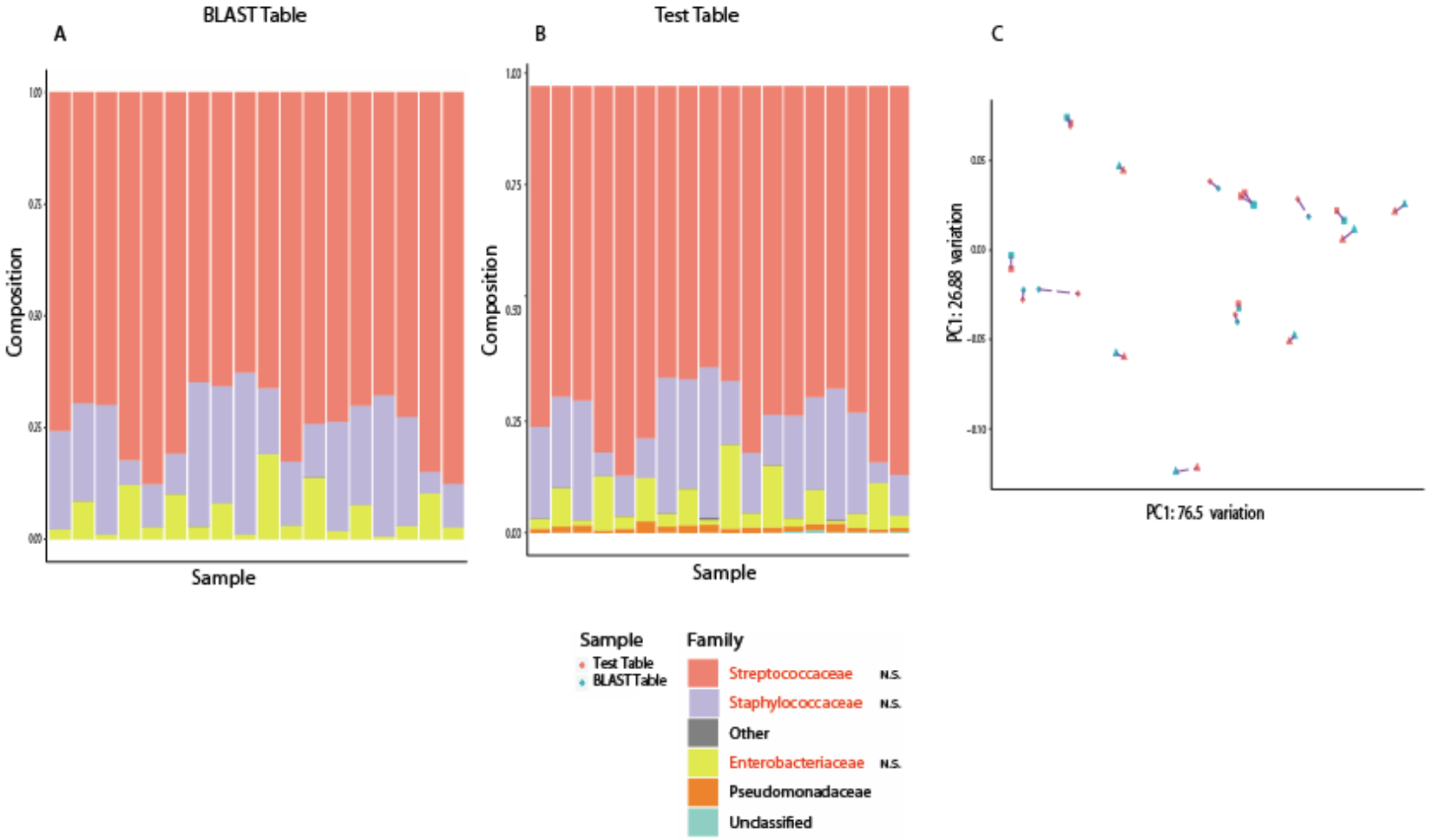


Figure 6: ***Comparison of (A) REFERENCE table vs (C) count table with contamination removed by Decontam, and guided by SourceTracker. The impact contamination has on sample beta diversity is shown in (B), with the distance matrix calculared using Bray-Curtis dissimilarity.***

The effectiveness of this strategy was validated by re-analysis with the SourceTracker algorithm (Figure 7). This time the algorithm shows that the level of contamination in the samples has decreased significantly. Most samples show only trace amounts of contamination, with only one sample still having a level of contamination comparable to pre-contamination removal levels. A second verification step is shown in the same figure. As contamination control inevitably involves the removal of reads, care should be taken to ensure that sufficient sampling depth remains to accurately characterise each sample, as evidenced by a plateauing of the rarefaction curve for a sample.

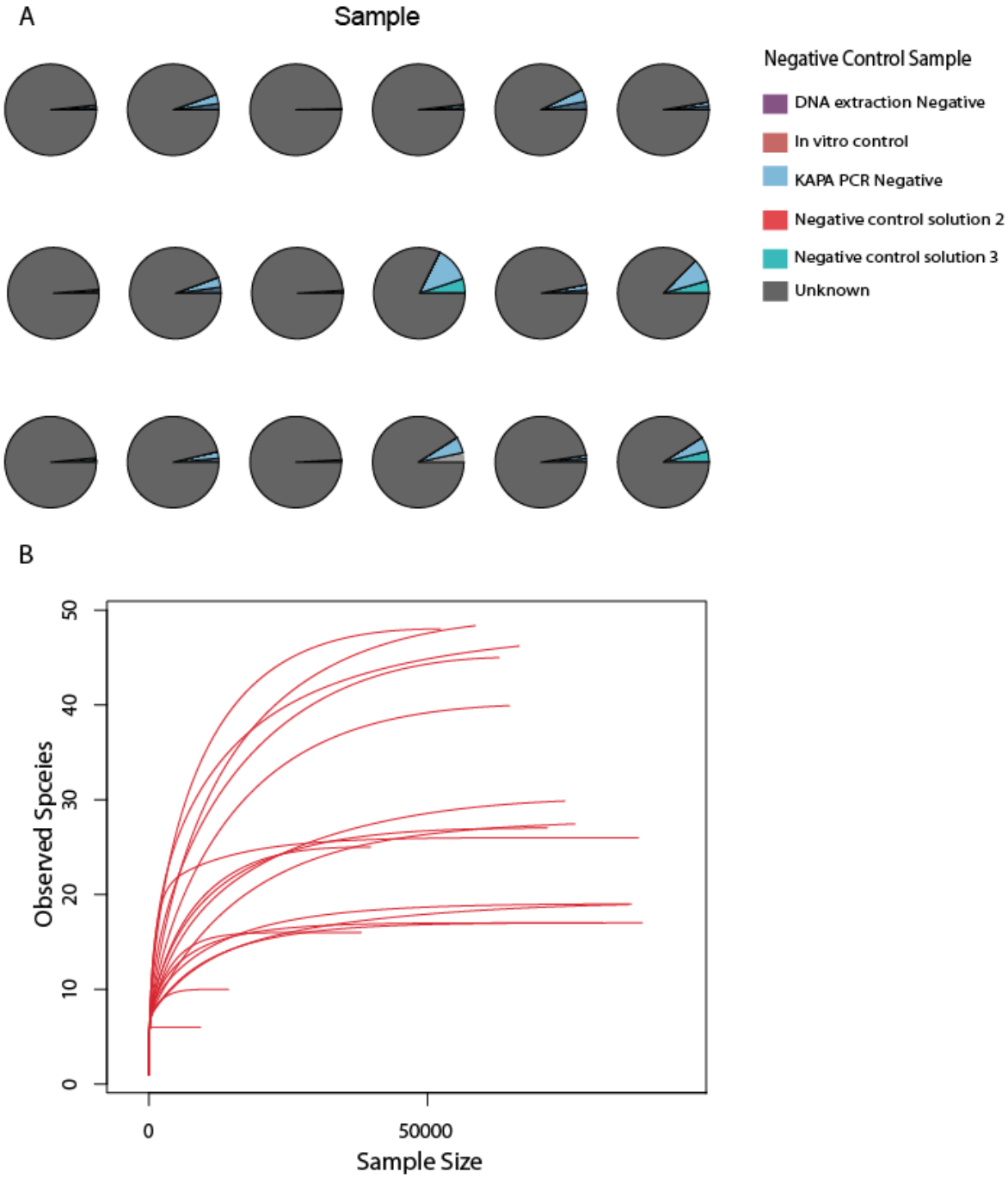


Figure 7: ***(A) SourceTracker algorithm run on new count table following contamination removal. Grey shaded region represents the true biological signal. (B) Rarefaction curve plotting observed species against number of reads examined.***

In summary, Table 2 below shows the mean Euclidean distance between matched samples following the different contamination removal strategies outlined up to this point to mathematically assess their effect on any further analysis. The manually supervised contamination removal approach incorporating both SourceTracker and Decontam significantly improves on all the other methods examined (*p = < 0.001)*.

Table 1: Table showing the mean Euclidean distance between paired samples for each ASV table vs the reference table, on PcOA plot. *Guided method is statistically significantly closer to reference method than any of the other groups. (p = < 0.001 in all cases)*

Strategy	**Mean Distance**
REFERENCE table vs No Contamination Control	0.070
REFERENCE table vs Contamination removal by Subtraction	0.062
REFERENCE table vs Generic implementation of Decontam	0.070
REFERENCE table vs Manually supervised contamination removal	0.009

DISCUSSION

The choice of approach to contamination removal was shown to have a marked effect on the composition of the samples analysed, and thus on any downstream analysis. As expected from samples of this nature, there was considerable contamination present. The initial comparison between the REFERENCE table and the untreated count table show that the sample composition was completely altered by environmental contaminants. That this would also affect any downstream diversity analysis was confirmed as the paired samples diverge significantly on a PcOA plot calculated with Bray-Curtis dissimilarity (Figure 2). It must be concluded that carrying out a microbiome survey samples of this kind, and potentially many low biomass or FFPE samples, without accounting for this contamination would have been untenable.

The problem of contamination control is not as simple as removing any sequence variant that appears in the negative controls. This approach is shown in Figure 3. Although the sample composition plots looked similar, the paired samples showed a high degree of between sample variation. The reason for the difference in beta diversity between the REFERENCE table and the untreated count table in this instance was that several high abundance sequences were erroneously removed from the dataset as they appeared in low quantities in some of the negative controls. This does not necessarily mean that they are contaminants, but rather could be artefacts of the sample preparation or sequencing process, as a result of cross contamination within a run as previously reported (14). So while a blanket removal of these potential contaminant sequences does show an improvement over making no intervention, it is not a perfect solution.

The previous example showed a contamination control method that was too conservative, with many falsely identified contaminants. The approach highlighted in Figure 4 was the opposite in that many true contaminants were allowed to remain in the ASV table. This strategy is included as a warning that simple “black-box” use of many bioinformatic tools, and Decontam in particular can lead to erroneous results negatively impacting further research. The inclusion of negative controls blindly without assessing their impact on the samples, or whether the number of reads present would leave them susceptible to being dominated by a cross-

contaminant has a significant impact on the accuracy of the Decontam algorithm. Equally important is setting the threshold to a level that matches the degree of contamination presentpresent in samples, to ensure that the contamination removal process is not too lax or too strict. SourceTracker can assist in this. If decontamination tools are run blindly, it is shown here that they have little if any beneficial effect on the accuracy of results, manual supervision of this process is necessary.

SourceTracker should be used to assess the relationship between samples and negative controls before attempting to remove contaminant sequences. This allows for contamination control to only be based on those negative controls that have a clear and significant effect on the samples. This lowers the possibility of false positives as seen in Figure 3, or false negatives as seen in Figure 4. While this manual approach does still show some contamination, indicating a higher false negative rate than the contamination removal by subtraction method, it has a considerably lower false positive rate in terms of contaminant identification. The effectiveness of this strategy is shown by the fact that there is minimal difference between paired samples on PcoA plot or sample composition plot. The method issignificantly more accurate than others tested ($p < 0.001$), and appears to negate the impact of contamination on downstream analysis.

We have shown that when properly combined, a robust negative control strategy along with manually supervised bioinformatic retrospective removal of known contaminants can limit contamination to an extent where the effect on any downstream microbial analysis is inconsequential. There is always room for improvement, and in a similar fashion to many recent publications that have published tables of bacterial taxa known to be environmental contaminants in sequencing experiments, labs should strive to develop in house databases of contaminant reads identified in commonly used reagents. These could be used as an initial screen of any count tables generated, working in a similar manner to the many reference-based chimera removal tools that exist today (24).

One final consideration when undertaking contamination removal is the fact that reads must be discarded during this process. This must be taken into consideration when designing the sequencing experiment. A typical Miseq sequencing run can be

expected to produce 13.2-15 million paired end reads, which are roughly evenly distributed among samples (25). The sequencing library should be generated with this in mind to ensure sufficient sampling depth remains after the contamination removal process, which can be simplified using the sequencing coverage calculator found on the Illumina website(26). This can be assessed after the fact using rarefaction curves, which the number of reads checked vs the number of new species identified and the curve is expected to plateau if sufficient sampling depth has been achieved (27).

CONCLUSION

When combined with positive and negative controls, it is shown that even samples with heavy contamination can be restored to a state where they can give accurate and reproducible information. The recommendation is that all future sequencing studies involving samples vulnerable to contamination be accompanied by both a wide variety of negative controls, and a number of positive controls that closely resemble the samples being analysed.

REFERENCES

1. Quince, C., Walker, A.W., Simpson, J.T., Loman, N.J. and Segata, N. (2017) Shotgun metagenomics, from sampling to analysis. *Nature Biotechnology*, **35**, 833.
2. Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J. and Walker, A.W. (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, **12**, 87.
3. de Goffau, M.C., Lager, S., Salter, S.J., Wagner, J., Kronbichler, A., Charnock-Jones, D.S., Peacock, S.J., Smith, G.C.S. and Parkhill, J. (2018) Recognizing the reagent microbiome. *Nature Microbiology*, **3**, 851-853.
4. Eisenhofer, R., Minich, J.J., Marotz, C., Cooper, A., Knight, R. and Weyrich, L.S. (2019) Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol*, **27**, 105-117.
5. Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. and Kelley, S.T. (2011) Bayesian community-wide culture-independent microbial source tracking. *Nat Methods*, **8**, 761-763.
6. Davis, N.M., Proctor, D.M., Holmes, S.P., Relman, D.A. and Callahan, B.J. (2018) Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, **6**, 226.
7. Pollock, J., Glendinning, L., Wisedchanwet, T. and Watson, M. (2018) The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. *Applied and Environmental Microbiology*, **84**, e02627-02617.
8. Marotz, C.A., Sanders, J.G., Zuniga, C., Zaramela, L.S., Knight, R. and Zengler, K. (2018) Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome*, **6**, 42.
9. Hosein, A.N., Song, S., McCart Reed, A.E., Jayanthan, J., Reid, L.E., Kutasovic, J.R., Cummings, M.C., Waddell, N., Lakhani, S.R., Chenevix-Trench, G. *et al.* (2013) Evaluating the repair of DNA derived from formalin-fixed paraffin-embedded tissues prior to genomic profiling by SNP-CGH analysis. *Laboratory investigation; a journal of technical methods and pathology*, **93**, 701-710.
10. Qiagen. (2-12) QIAamp DNA FFPE Tissue Handbook. *Sample and Assay Technologies*.
11. Hornung, B.V.H., Zwittink, R.D. and Kuijper, E.J. (2019) Issues and current standards of controls in microbiome research. *FEMS microbiology ecology*, **95**, fiz045.
12. Eisenhofer, R., Minich, J.J., Marotz, C., Cooper, A., Knight, R. and Weyrich, L.S. (2019) Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology*, **27**, 105-117.
13. Carlsen, T., Aas, A.B., Lindner, D., Vrålstad, T., Schumacher, T. and Kauserud, H. (2012) Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecology*, **5**, 747-749.
14. Sheik, C.S., Reese, B.K., Twing, K.I., Sylvan, J.B., Grim, S.L., Schrenk, M.O., Sogin, M.L. and Colwell, F.S. (2018) Identification and Removal of Contaminant Sequences From Ribosomal Gene Databases: Lessons From the Census of Deep Life. *Front Microbiol*, **9**, 840-840.
15. Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., Granger, B., Green, L., Howd, T., Mason, T. *et al.* (2018) Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics*, **19**, 332-332.
16. Kuczynski, J., Stombaugh, J., Walters, W.A., González, A., Caporaso, J.G. and Knight, R. (2011) Using QIIME to analyze 16S rRNA gene sequences from microbial communities. *Curr Protoc Bioinformatics*, **Chapter 10**, Unit10.17-10.17.

17. Bokulich, N.A., Subramanian, S., Faith, J.J., Gevers, D., Gordon, J.I., Knight, R., Mills, D.A. and Caporaso, J.G. (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature methods*, **10**, 57-59.
18. Stewart, C.J., Fatemizadeh, R., Parsons, P., Lamb, C.A., Shady, D.A., Petrosino, J.F. and Hair, A.B. (2019) Using formalin fixed paraffin embedded tissue to characterize the preterm gut microbiota in necrotising enterocolitis and spontaneous isolated perforation using marginal and diseased tissue. *BMC Microbiology*, **19**, 52.
19. Evans, T.C. and Nichols, N.M. (2008) DNA Repair Enzymes. *Current Protocols in Molecular Biology*, **84**, 3.9.1-3.9.12.
20. Do, H. and Dobrovic, A. (2015) Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization. *Clinical Chemistry*, **61**, 64.
21. Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, **26**, 2460-2461.
22. Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. and Kelley, S.T. (2011) Bayesian community-wide culture-independent microbial source tracking. *Nature Methods*, **8**, 761.
23. LaSala, P.R., Segal, J., Han, F.S., Tarrand, J.J. and Han, X.Y. (2007) First Reported Infections Caused by Three Newly Described Genera in the Family <em>Xanthomonadaceae</em&gt. *J Clin Microbiol*, **45**, 641.
24. Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S.K., Sodergren, E. *et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Res*, **21**, 494-504.
25. Ambardar, S., Gupta, R., Trakroo, D., Lal, R. and Vakhlu, J. (2016) High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian J Microbiol*, **56**, 394-404.
26. center, I.s. (2019).
27. Zaheer, R., Noyes, N., Ortega Polo, R., Cook, S.R., Marinier, E., Van Domselaar, G., Belk, K.E., Morley, P.S. and McAllister, T.A. (2018) Impact of sequencing depth on the characterization of the microbiome and resistome. *Scientific Reports*, **8**, 5890.

Chapter III

Bacteria in breast tumours

This manuscript is currently being prepared for submission in combination with another research project.

ABSTRACT

Background Although the existence of a bacterial community in both breast tumour tissue and healthy adjacent tissue has been reported by numerous groups since 2014, it remains a contentious issue. Tumour samples provide many obstacles to carrying out robust and reliable microbial surveys, primarily due to the anticipated low bacterial biomass of these samples. This feature of breast tumour samples has stifled research in this field as previous studies analysing low biomass data such as breast tissue have been cited for taking insufficient precautions in limiting the effect environmental contamination has on the results. While the debate continues over the presence or absence of bacteria in niches of low biomass, no further research can be conducted on how bacteria could be utilised for therapeutic purposes if they are indeed present.

Aim This study set out to definitively assess the presence of endogenous bacterial communities in breast tumours and the associated healthy adjacent tissue.

Methods The study incorporated a robust negative control strategy, makes use of the recent developments in bioinformatic contamination removal, and examines choice of primer site for amplification or presence of extracellular DNA playing a role in the outcome.

Results The presence of a detectable tumour microbiome was evident in the majority of tumour samples, and was similar in community structure to that of the skin and normal adjacent tissue with some statistically significant differences, amounting to a distinct microbial signature unlikely to be due to sample or kit contamination.

Conclusions This study indicates the presence of bacterial communities in both malignant and non-malignant breast tissue in the majority of cases, and that the two can be differentiated based on their bacterial composition.

INTRODUCTION

According to the World Health Organisation, breast cancer affects 2.1 million women worldwide annually, which resulted in 627,000 deaths in 2018, constituting 15% of all cancer deaths among women that year. While incidence rates increase 3.1% globally, they still vary considerably between high income countries such as the United States (92 per 100,000) and lower income regions such as eastern Asia (27 per 100,000) (1). This is reflective not only of the increased likelihood of risk factors such as hormonal contraception, lack of breastfeeding, obesity and alcohol use, but also of the lower rate of detection in poorer regions so these rates may be closer together than currently thought (1). Potential avenues towards the development of improved treatment and detection strategies come from a myriad of sources, and one that should not be discounted, despite early setbacks, is the possibility of utilising endogenous bacterial communities within breast tumours for therapeutic or diagnostic purposes. There are four key physiological features shared by solid tumours which theoretically should promote bacterial colonisation: (i) Leaky vasculature which could allow circulating bacteria to embed in tumour tissue, (ii) the immune privileged nature of tumours, (iii) solid tumours possess low oxygen regions suitable for the proliferation of facultative and anaerobic bacteria, and (iv) high turnover regions of tumours are nutrient rich, promoting bacterial growth (2).

Research has been conducted assessing the viability of using bacterial colonisation of tumour environments as a diagnostic or therapeutic tool (3). Further research would benefit from high quality reproducible work to definitively confirm the natural presence of bacteria in tumour tissues. Unfortunately, features inherent of tumours and adjacent tissue have hampered progress. Firstly, the quantities of bacteria present are often so low that it becomes difficult to differentiate between any bacteria genuinely originating in the sample, and those arising from environmental contamination during the extraction, or library preparation process. This issue has plagued numerous recent studies of low bacterial biomass environments, including work done by our own group. Related to this, human biopsies, particularly those from “non-tract” locations can be expected to have overwhelming ratios of host to bacterial DNA. In these circumstances, 16S rRNA gene-specific primers have been shown to amplify human reads in addition to bacterial reads, which further clouds the analysis of these environments (4). Since

the idea of bacteria living within healthy and malignant breast tissue was first theorised (5), tumours at a variety of other body sites have been examined in the hope of discovering a bacterial microenvironment. Many of these have been beleaguered by the problems outlined above, but recent studies (6,7) have shown that at least some tumour sites do appear to contain a detectable bacterial microenvironment and that the tools, both laboratory and bioinformatic, exist to reliably analyse these microenvironments.

Such samples have only become accessible to researchers due to the increased emphasis in recent years on reproducibility and quality control in microbiome research, primarily centred on contamination control. The fact that the majority of reagents and extraction kits used in sequencing library preparation are not sterile and can therefore alter the microbial profile of a sample was first brought to light by Salter *et al* in 2014 (8) and confirmed repeatedly by several high impact publications such as De Goffau *et al (9).* In response to this, guidelines have been published for effective negative control strategies to quantify the effect of these kit contaminants (10), and bioinformatics tools developed to retrospectively mitigate the effect they have on eventual analysis (11,12).

Here, we account for potential sources of error that have been highlighted in previous research of a similar nature. This begins with a robust contamination control strategy outlined here in considerable detail. Following this, the effect of the 16S rRNA gene region targeted, and the potential presence of extracellular DNA in samples is also investigated. There has been considerable debate over which hypervariable region to target, and two of the most common regions (V1-V2 and V3-V4) were compared to assess what effect, if any, choice of hypervariable region had on sequencing results. The phylogenetic variability within the V4 region shows the strongest correlation with the phylogenetic variability of the 16S rRNA gene fragment overall, and the combined length of the V3-V4 region of 439 bases yields a large region for discrimination between taxa while still allowing for trimming due to poor sequencing quality. Conversely, the V1-V2 region is only 298 bases in length in *E. coli*, but this allows for a near total overlap of forward and reverse reads, which ensures for considerable noise reduction from sequencing errors (13). The V1-V2 region of the 16S rRNA gene has been adopted by many groups studying low

biomass samples as this shorter length has been shown to allow more efficient amplification of low abundance template sequences (14).

Determining the presence of bacteria in tumours as distinct from environmental contamination is approached here by combining a robust negative control strategy with effective bioinformatic removal of contaminant reads. Two recent bioinformatic tools, Sourcetracker(11,12) and Decontom(11) were used to facilitate this. Sourcetracker uses Bayesian statistics to predict the proportions of bacteria in samples that may have originated from designated source environments, which in this case were negative controls. When used in this way, it works well as an initial screening tool to assess the degree of contamination present. Decontam removes contaminant reads, either by eliminating reads that have an inverse correlation with input DNA, or based on their abundance negative controls. In both cases Decontam requires that a threshold is set, and this can be dictated by the results of SourceTracker.

One of the limitations of DNA sequence analysis is that it gives no indication of whether the DNA present in a sample is contained within living bacteria, dead intact bacteria, or extracellular DNA originating from biofilms or dead bacteria. A pre-treatment step with DNAse, an enzyme that non-specifically cleaves DNA is often incorporated into metagenomics workflows to remove this extracellular DNA prior to amplification, ensuring that only intact bacteria contribute to the microbial profiling of an environment (15).

With these confounding factors controlled for, the ultimate aim of this study is to provide a highly reproducible and reliable survey of the bacterial communities present in breast tumours and their adjacent tissues.

METHODS

Sample Collection to reduce contamination

'Fresh' specimens were provided directly from the operating theatre, as opposed to sectioning in histology laboratory, in order to minimise exposure to environmental contaminants. All samples were provided by a single surgical team under a single consultant ensuring consistency. Tumour tissue was biopsied using a 14 French ACHIEVE^{TM} programmable automatic biopsy system. Additionally, a skin swab (SS), a normal adjacent (NA) sample were taken from each patient to complement the tumour sample (TS), to ensure that any variability found in the diversity or composition of different sample types was down to the niches themselves and not confounded by person to person variation in the microbiome.

Sequencing Library Preparation

DNA extraction and library preparation work was performed by other members of the Tangney lab.

Genomic DNA was amplified using 16S rRNA gene amplicon polymerase chain reaction (PCR) primers targeting the hypervariable V3–V4 region of the 16S rRNA gene: V3–V4 forward, 5′-TCGTCGGCAGCGTCAGATGTGT ATAAGAGACAGCCTACGGGNGGCWGCAG-3′ and V3–V4 reverse, 5′-GTCTCGTGGGCTCGGAGATGTGTA TAAGAGACAGGACTACHVGGGTATCTAATCC-3' (Illumina 16S Metagenomic Sequencing Protocol, Illumina, CA, USA).

A 35-µl PCR was performed for each sample per the following recipe: 3.5 µl of template DNA, 17.5 µl of KAPA HiFi HotStart ReadyMix (Roche), 0.7 µl of both primers (initial concentration, 10 pmol/µl), 0.1 µg/µl bovine serum albumin fraction V (Sigma), and 8 µl of 10 mM TrisCl (Qiagen). Thermal cycling was completed in an Eppendorf Mastercycler per the directions in the 'Amplicon PCR' section of the '16S Metagenomic Sequencing Library Preparation' protocol (Illumina). Amplification was confirmed by running 5 µl of PCR product on a 1.5% agarose gel at 70 volts for 80 min, followed by imaging on a Gel Doc EZ System (Bio-Rad). The product was ~450 base pairs (bp) in size. PCR-positive products were cleaned per

the ‘PCR CleanUp’ section of the Illumina protocol, with the exception that drying times were reduced to half the prescribed duration to account for the additional drying that occurs in a laminar airflow hood. Sequencing libraries were then prepared using the Nextera XT Index Kit (Illumina) and cleaned per the Illumina protocol. Libraries were quantified using a Qubit fluorometer (Invitrogen) using the ‘High Sensitivity’ assay. Sample processing was subsequently completed at Genewiz inc. Samples were normalised, pooled and underwent a paired-end 300 bp run on the Illumina MiSeq platform.

Bioinformatic analysis

The quality of the paired-end sequence data was visualised using FastQC v0.11.6, and then filtered and trimmed using Trimmomatic v0.36 to ensure a minimum average quality of 25. The remaining high-quality reads were then imported into the R environment v3.4.4 for analysis with the DADA2 package v1.8.0. After further quality filtering, error correction and chimera removal, the raw reads generated by the sequencing process were refined into a table of Amplicon Sequence Variants (ASVs) and their distribution among the samples. It is recommended that ASVs (formerly called ‘Ribosomal Sequence Variants’) are used in place of ‘operational taxonomic units’ (OTU). OTUs are clustered at a pre-determined threshold of similarity, typically 97%, which distances them from the ecological reality present in a sample. As the name suggests, an ASV represents an existing biological sequence variant found in the sample.

Alpha diversity calculated as Chao1 species richness, and Bray-Curtis distances, for analysis of beta diversity, were calculated using the PhyloSeq package v1.24, and the Vegan package v2.52. Beta diversity calculations produce distance matrices with as many columns and rows as there are samples; thus, beta diversity is often represented using some form of dimensionality reduction, in this case, using principal co-ordinates analysis (PCoA) with the Ape package v5.1. Hierarchical clustering, an unsupervised method that can reveal key taxa that distinguish their respective environments, was performed with the heatplot function in the made4 package v1.54. Differential abundance analysis was carried out using Deseq2 v1.2.0, which identifies differentially abundant features between two groups within the data.

Tests of means were performed using the Mann-Whitney *U* test unless otherwise stated, and correlations were calculated using Spearman's rank correlation coefficient. Where applicable, false positive rates were controlled below 5% using the FDR procedure. Random forest classification trees were built using the RandomForest(v4.6.15) and pROC(v1.15.3) packages in R. Developed in 2001(16), this tool is particularly suitable for classification based on sequence data as it is computationally efficient, gives an estimate of the importance of each predictive variable (in this case ASV's), and limits model overfitting by making decisions on splitting of samples at a particular node on a randomly sampled subset of the dataset, rather than all available sequences. As thousands of trees are created with each implementation of the algorithm there is no risk of loss of information(16).

Despite not identifying the contaminant taxa themselves, the source tracker utility is invaluable in estimating the proportion of a sample ("Sink") that may have originated in a negative control ("Source") Decontam can remove taxa, based on presence or absence in negative controls, or inverse correlations with input DNA but requires a threshold to be set, which can be dictated by SourceTracker. The effectiveness of this can then be confirmed by SourceTracker.

RESULTS

Table 1: ***Samples analysed. Multiple samples per patient in some cases to account for possible sources of error.***

**Indicates >500 reads remaining after removal of environmental contamination and non-microbial reads.*

	Tumour	Normal	Skin Swab	Total
Sequenced samples (including replicates)	*37*	*40*	*31*	*108*
Surviving Contamination Removal*	*29*	*40*	*30*	*99*
DNase – replicates (No DNAse treatment of samples)	2	10	3	15
V1-V2 Primer Pair replicates	8	0	0	8
Final analysis	19	30	27	76

1. Contamination control

The full breakdown of comparisons performed to control for sources of error, and as part of the final analysis can be seen in Table 1. Figure 1A shows the composition at family level of the samples, prior to any contamination removal. The samples are grouped by sample type (Ductal tissue, Normal Adjacent tissue, Skin Swab, Tumour tissue) and within this by DNAse status. Some sample were treated without DNAse to investigate the effect of DNAse treatment on the levels of environmental contamination. In addition, the family level taxonomic composition of each sample and the number of reads associated with each sample can be seen above the plot (i). Figure 1B shows a sample composition pie chart at a per sample level, showing the estimated proportion of reads within each sample originating in one of the negative controls. Grey shaded regions indicate the proportion of non-contaminant reads.

Some samples show contamination in excess of 50 %, but the majority show little to no effect by environmental sources, which is encouraging for downstream analysis.

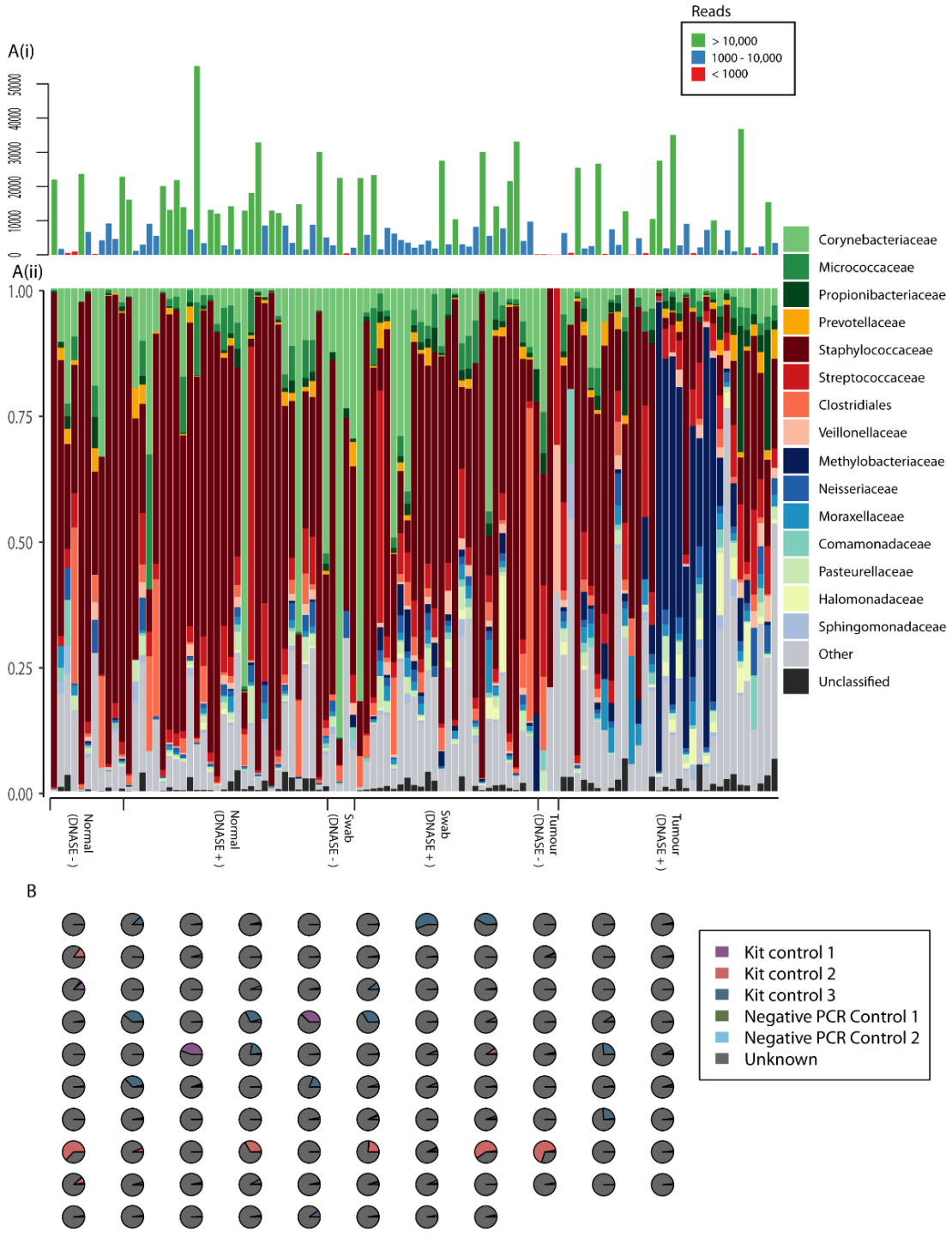


Figure 1: ***Sample overview prior to contamination removal.*** *(A) (i) Family level composition of patient samples, ordered by sample type and DNASE status consecutively. (ii) indicates the number of reads present in each sample. (B) SourceTracker output using the negative controls as "Source" samples, and the patient samples and "Sink" samples. Data shows light to moderate contamination*

among samples, with only 3 samples showing in excess of 50% contamination, as per SourceTracker.

The effectiveness of the contamination strategy can be seen by comparing the results shown in Figure 1 with those in Figure 2 below, which show the same samples in the same order, but following removal of reads identified as environmental contaminants by a combination of SourceTracker and Decontam. As the initial iteration of SourceTracker only implicated the kit control samples in introducing contamination to the samples, the PCR controls were dropped prior to contamination removal. The low number of reads in these samples also made them high risk samples for false positive identification of contaminants due to cross contamination between patient samples and negative controls, either during the library preparation or sequencing stage. As can be seen in Figure 2, the contamination removal strategy was effective, and all samples are now almost entirely free of detectable contamination.

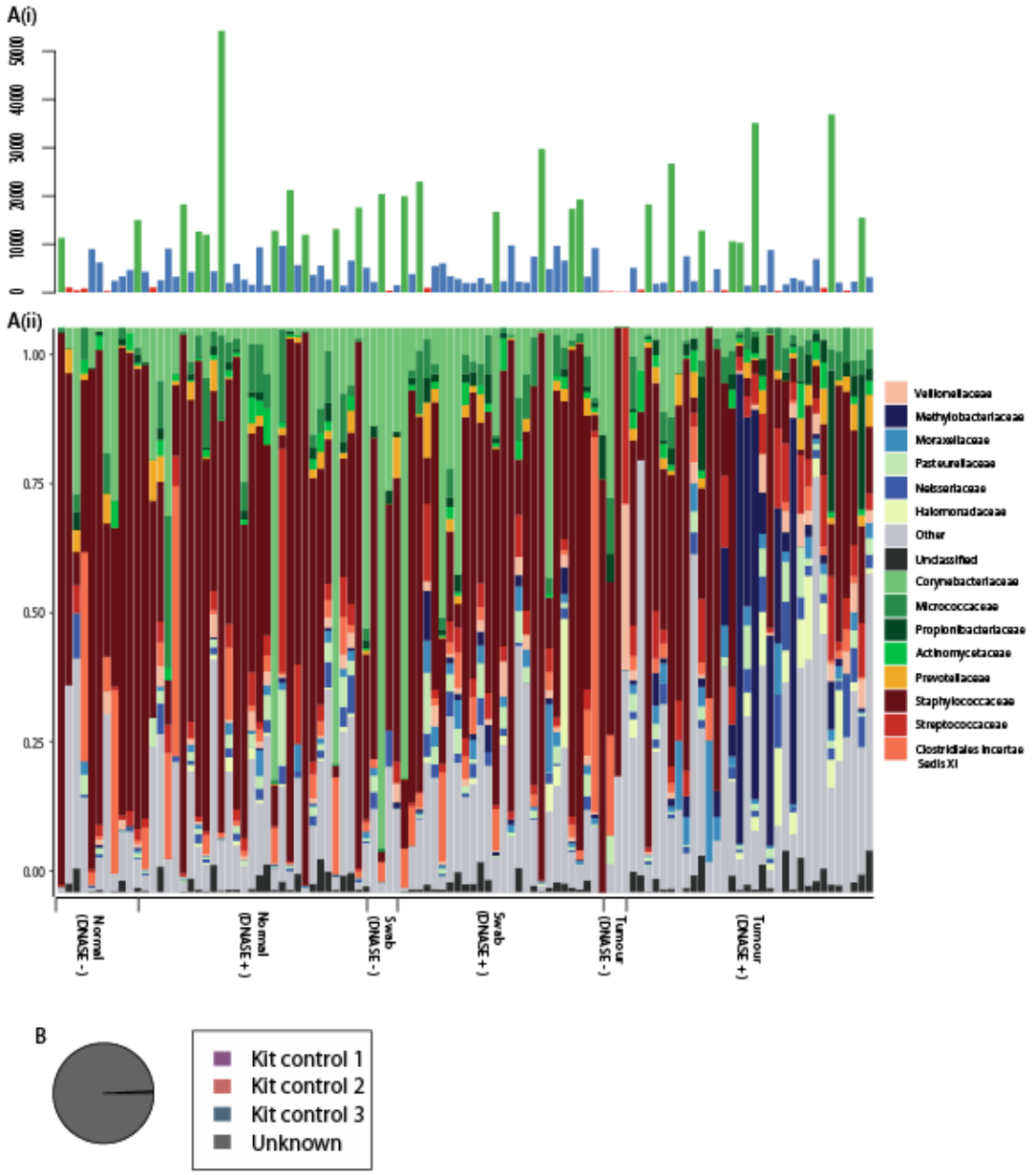


Figure 2: ***Sample overview following contamination removal.*** *(A) (i) Family level composition of patient samples, ordered by sample type and DNASE status consecutively. (ii) indicates the number of reads present in each sample;; (B) SourceTracker output using the negative controls as "Source" samples, and the patient samples and "Sink" samples. Consensus plot shows clearance of reads originating in negative controls as per SourceTracker.*

Contamination removal saw a reduction in the average reads per sample from 9084 to 6934, but the overall structure of the bacterial community remains unchanged. Although some trace amounts of contamination do remain in a few samples, it is safe to say that contamination does not remain in the levels necessary to alter any

biological signal present in the samples, and should not significantly impact any downstream analysis.

Due to the susceptibility of these samples to contamination, and the fact that trace amounts of contamination are still present as per SourceTracker, post retrospective contamination removal, the decision was made to employ the conservative strategy of removing any ASV appearing in the negative controls from the dataset.

2. Target region of choice

As mentioned earlier, several groups have suggested a shift to the V1-V2 hypervariable region of the 16S rRNA gene fragment, from the more widely used V3-V4 region. The effects of which target region is used are shown in Figure 3.

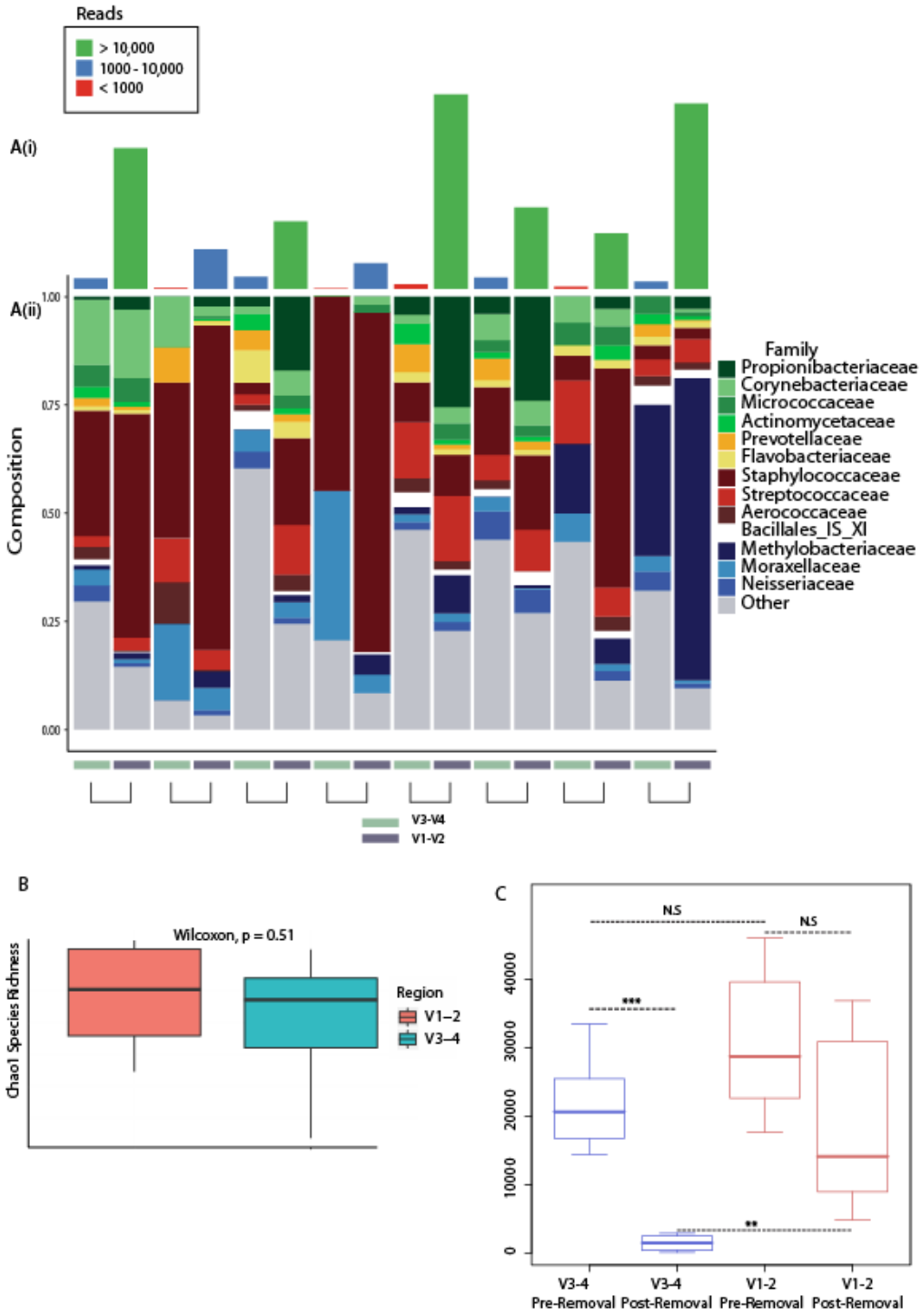


Figure 3: ***Pairwise comparison of samples using V1-V2 and V3-V4 primer pairs.*** *(A) (i) Reads per sample following the contamination removal outlined previously. (ii) Sample composition at the family level of paired samples. (B) Average Chao1 species richness between samples amplified using V1-V2 primers (red) and V3-V4 primers (blue). (C) Comparison reads per sample pre and post removal of contaminant and human aligning reads. In both (B) and (C) statistical testing is performed using Wilcoxon signed-rank test.*

Samples amplified using primers targeting the V1-V2 hypervariable region showed a consistently increased number of reads per sample after the removal of contaminant and non -bacterial reads (as per Mothur classifier) (Figure 3A) and also a decreased reduction in overall reads when comparing samples before and after this removal of reads. More than 90 % of the reads that did not classify as bacterial representative sequences were confirmed as human reads by BLAST. This tells us that the V1-V2 region is undoubtedly more suited to samples presenting with low biomass and an extremely high ratio of human reads. In the context of this study, the disparity in reads per sample did not have a negative impact on the diversity of the environments.

3. Effect of DNase treatment

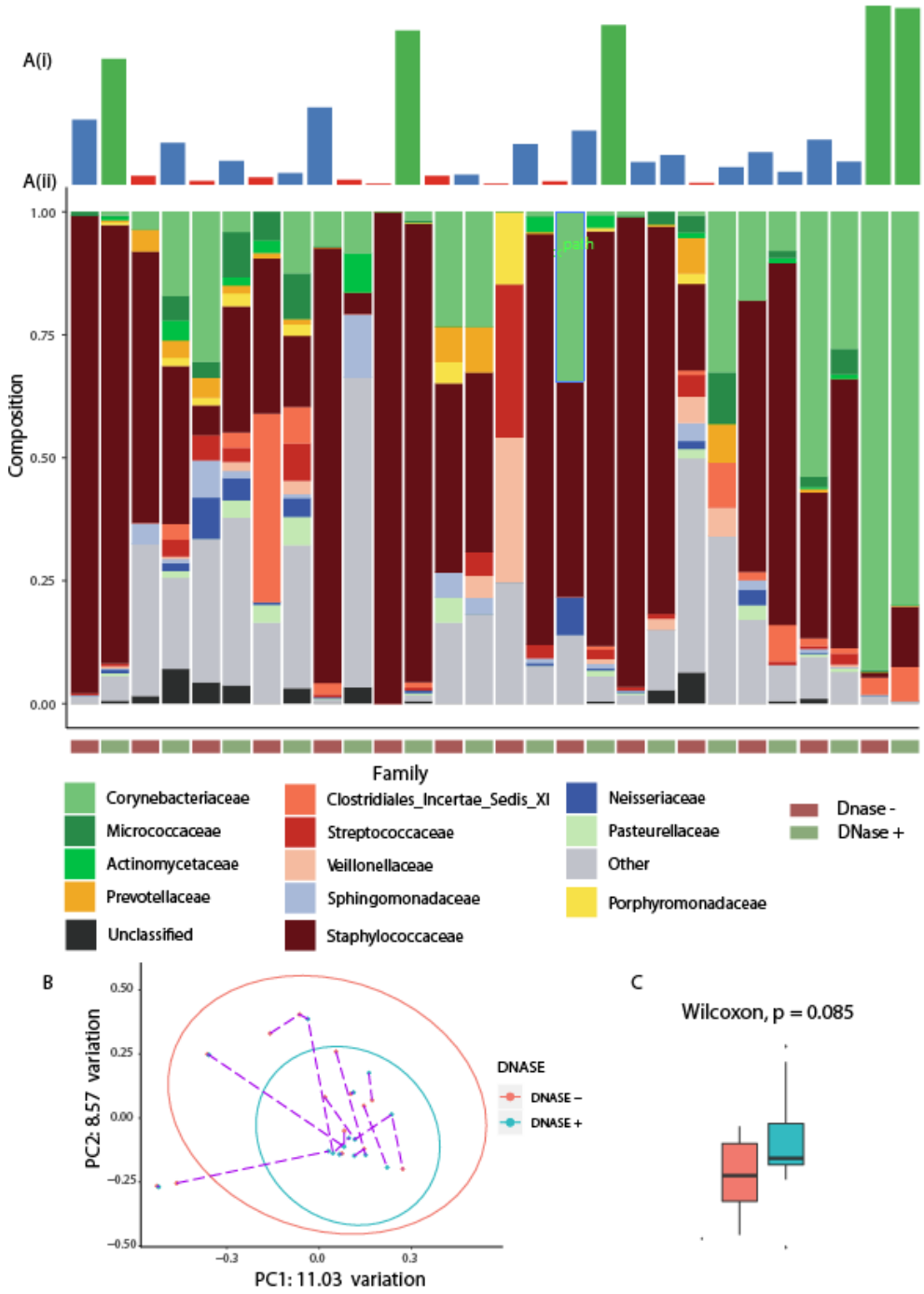


Figure 4: ***Pairwise comparison of samples using with and without DNase treatment.*** ***(A)*** *(i) Reads per sample following the contamination removal outlined previously. (ii) Sample composition at the family level of paired samples.* ***(B)*** *Bray-Curtis dissimilarity between Dnase + and – samples, showing no significant difference between groups as per PERMANOVA (p = 0.98).* ***(C)*** *Alpha diversity boxplots calculated using Chao1 species richness. Significance detected using wilxocon signed-rank test. The three panels of this figure show conclusively that no intra-sample variation due to DNAse treatment is detected.*

No significant difference was seen between samples with or without DNase treatment in terms of alpha diversity, calculated using Chao1 species richness. Beta diversity, calculated using Bray-Curtis dissimilarity and visualised on a PcOA plot showed no observable separation based on DNAse status, which was confirmed statistically using Permanova analysis. This is mirrored by the sample composition plot, where paired samples closely resemble each other.

4. Diversity Overview

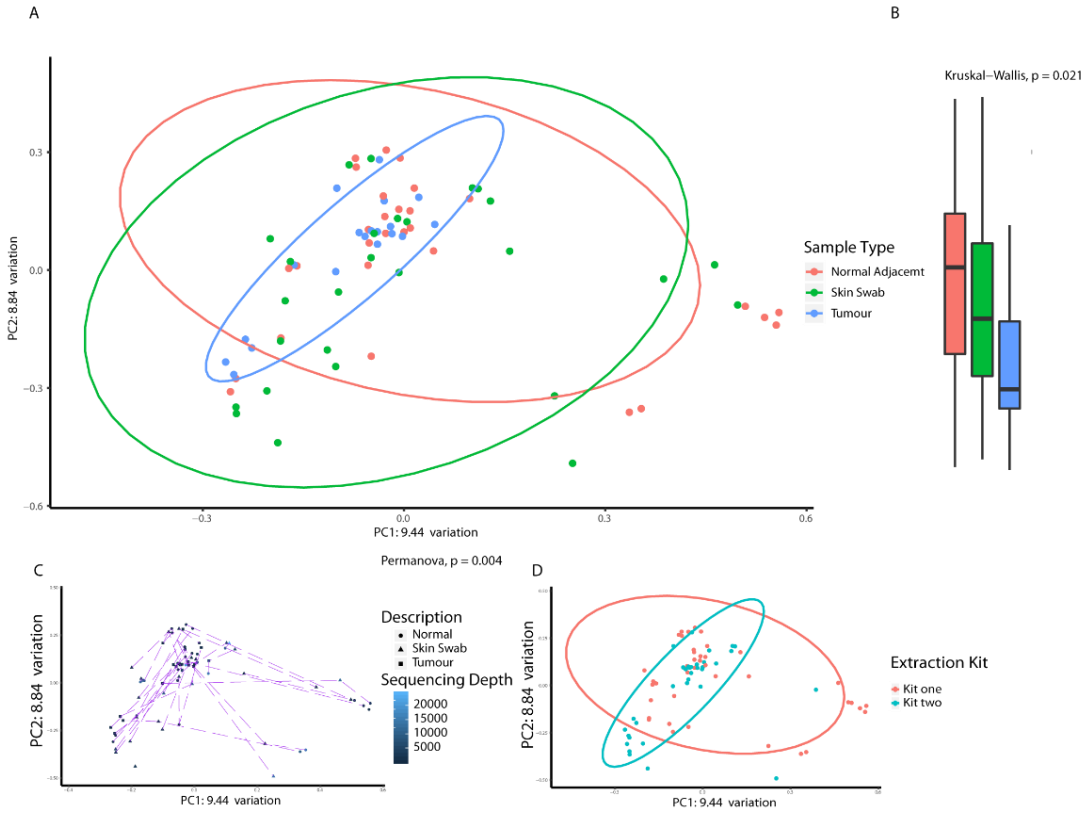


Figure 5: ***Diversity based comparison of the three patient environments.*** ***(A)*** *Bray-Curtis dissimilarity comparing Tumour, Normal Adjacent and Skin swab samples types. Ellipses represent 80 % confidence interval for sample type.* ***(B)*** *Alpha diversity boxplots visualising differences in Chao1 species richness between samples. Significant decrease in diversity as per Kruskal-Wallis test (p = 0.021).* ***(C)*** *Elimination of potential confounding factors. Same ordination, due to Bray-Curtis dissimilarity as* ***(A)*** *with points overlaid with colouring by reads per sample, with paired samples joined by line. Neither of these causes any significant clustering.* ***(D)*** *Same ordination, due to Bray-Curtis dissimilarity as* ***(A)*** *with points coloured by extraction kit used, shows that kit bias had minimal influence in this ecological survey.*

The three environments assessed were broadly similar, with the significance implied by Permanova testing likely to relate to the tighter clustering of tumour samples compared to normal adjacent tissue and skin swab samples. The fact that the tighter clustering of tumoural samples occurs within the broader confidence regions for the other two sample types suggests that the dissimilarity is caused by an absence of ASVs more than the presence of tumour unique ASVs. This is reinforced by alpha diversity analysis using Chao1 species richness, which shows tumour samples having a significantly lower alpha diversity. As always, but particularly in the case of low biomass samples, confounding factors affecting the accuracy of results must be ruled out. Figure 7C shows the same ordination plot but with samples from the same patient connected by dashed lines, and rules out a clustering of samples by patient origin. Figure 7D again shows the same ordination plot, but overlaid with the extraction kit used, showing that kit bias played a minimal role in the outcome of this survey.

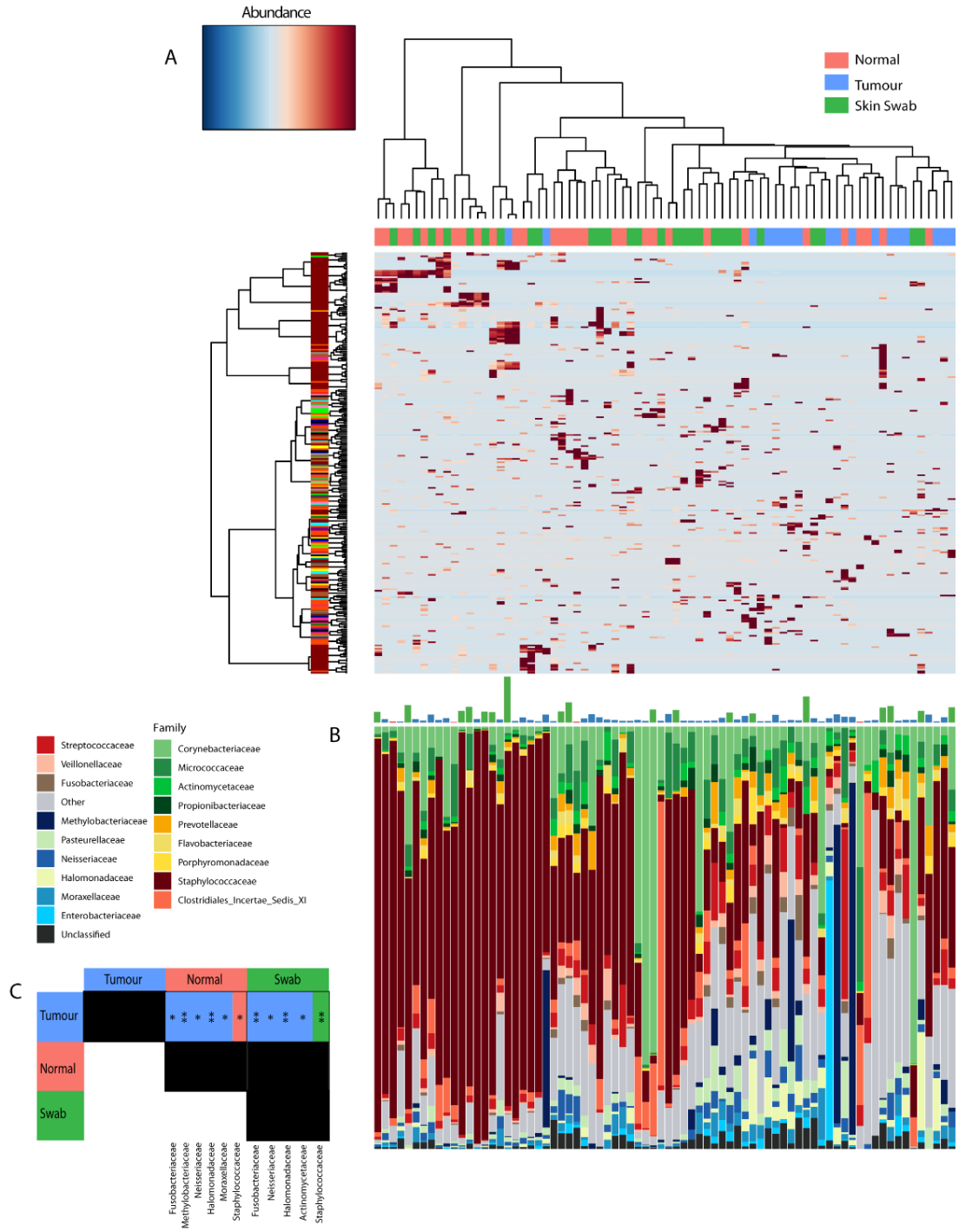


Figure 6: ***Clustering analysis of samples with corresponding sample composition.*** *Hierarchical clustering carried out using Ward's method, and 1-Pearson correlation distance used to calculate distance between rows and columns. The sample composition plot shows the proportions of different bacterial families within a given sample.*

Unsupervised hierarchical clustering was performed to detect discrete differences between samples that were not obvious using principal co-ordinates analysis. This is visualised in a heatmap where each column represents a sample and each row a unique ASV. In this instance, it shows that the majority of tumour samples appear in the right most cluster, while the skin swab and normal adjacent samples are more randomly dispersed throughout the heatmap. The basis for this clustering is evident when the composition plot is examined, with sixsix families having significantly different mean proportions between tumour and normal samples, and fivefive between tumour and skin swabs. There were no significantly different families when comparing normal and skin swab samples.

Following on from this, Deseq2 was used to detect differentially enriched taxa at the ASV level, between the different environments.

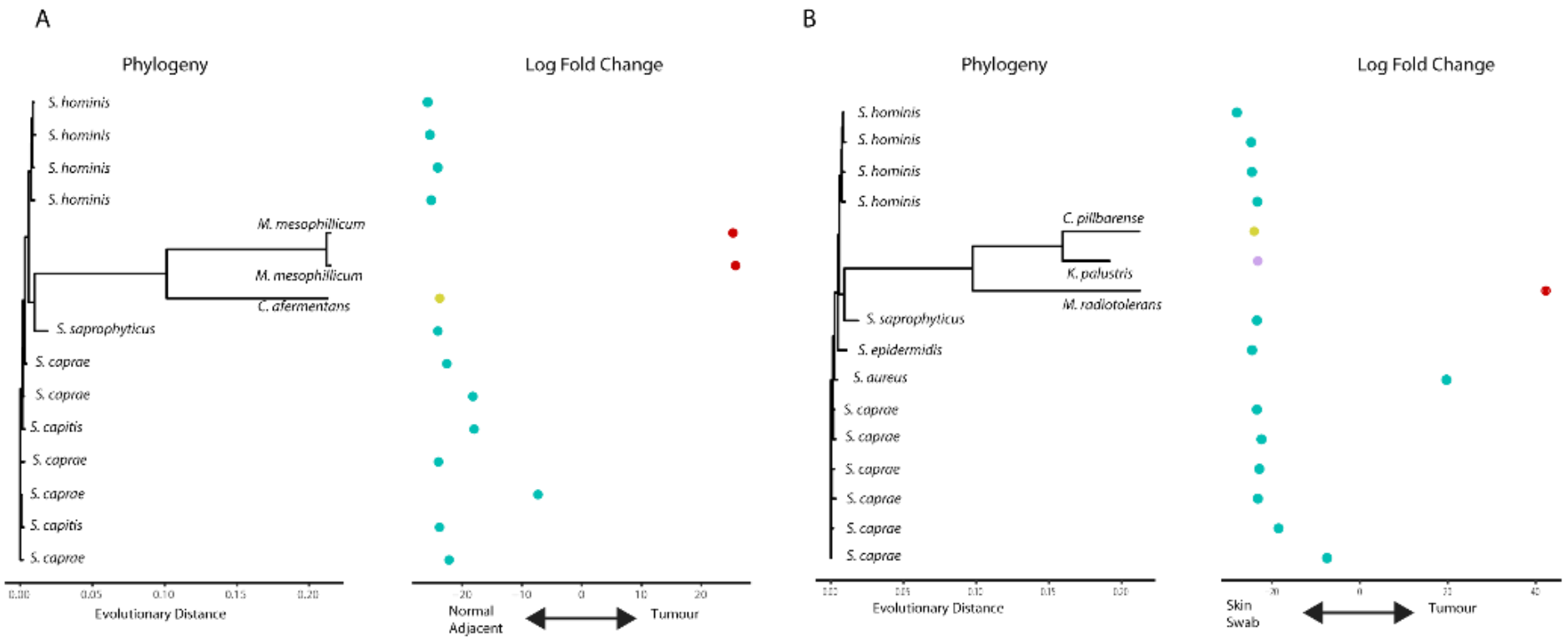


Figure 7*: Deseq2 based investigation of differentially enriched taxa between sample types. (A)* *Normal adjacent vs Tumour.* ***(B)*** *Skin swab vs Tumour. Both* ***(A)*** *and* ***(B)*** *show log fold change of taxa between the two environments, and phylogenetic relationship between them as per UPGMA phylogenetic tree. UPGMA branch labels are highest scoring blast hit for fasta sequence of ASV.*

Overall, 14 ASVs were found to be differentially abundant between tumour and normal adjacent samples, and 16 between tumour and skin swab samples, while there were no significantly different ASVs between Skin Swab and Normal Samples. Their evolutionary relatedness was explored phylogenetically, with each node labelled with its highest scoring BLAST hit in an attempt to gain more information about these closely related ASV's. The differences are predominantly in *Staphylococcus* spp. with *S. hominis, S. caprae and S. saprophyticus* all significantly elevated in skin swab and normal adjacent samples vs tumour samples. Tumour samples showed significant enrichment of *S. aureus, Methylobacterium spp., C. pillbarense* and *K. palustris.*

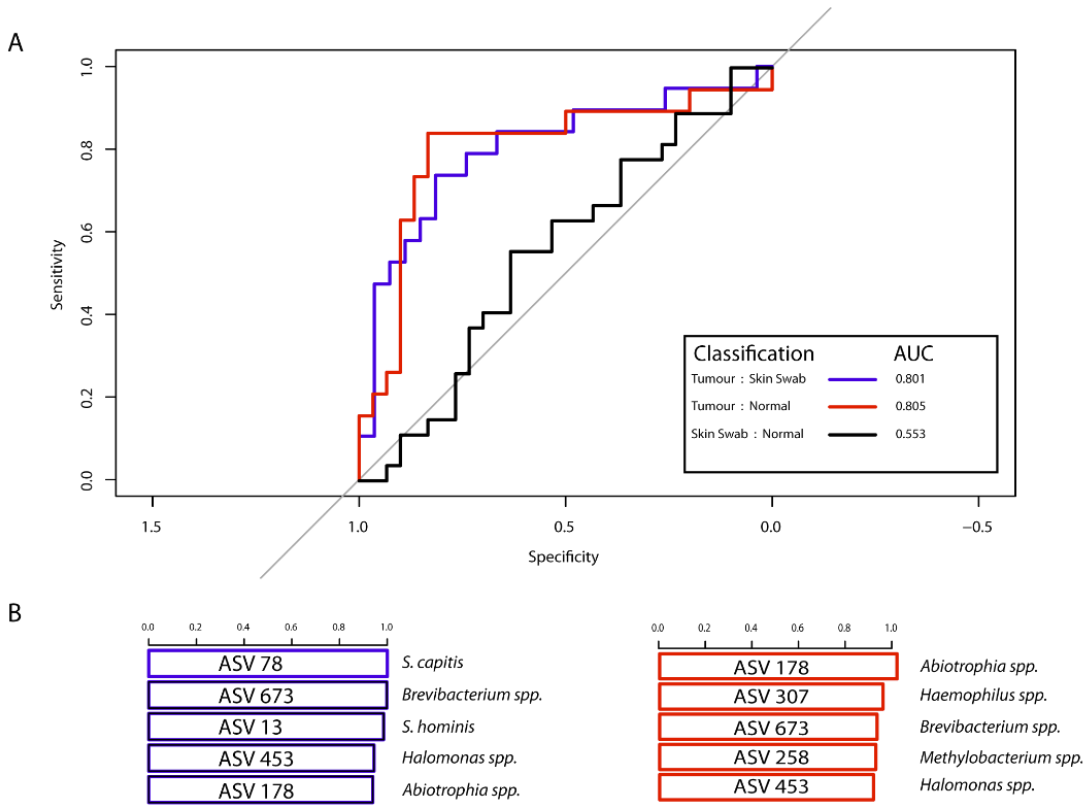


Figure 8: ***Random Forest-based classification of sample type****. **(A)** accuracy of classification represented as area under the curve of receiver operating characteristic curve. True positive rate is plotted on the X axis and false positives are plotted on the Y axis. **(B)** Mean decrease in accuracy of prediction when ASV is removed from feature table, scaled from 0-1, (top,t 5 results).*

Having detected significantly differentially abundant taxa at both the family and ASV level, the next step was to see if the ASV distribution between samples could be used to classify as either tumour, normal adjacent or skin swab. The RandomForest package in R was implemented for this purpose, using only the ASVs that appear in at least 5 % of samples. The algorithm was able to effectively differentiate between tumour and both normal and skin swab samples at a rate significantly higher than the random chance of correct classification, (AUC 0.801 and AUC 0.805 respectively) as seen in Figure 8. This was not the case when attempting to classify skin swab and normal adjacent samples, where the AUC was 0.553. As the receiver operating characteristic curve plots the true positive rate against the false positive rate, the worst possible score for a model is in fact 0.5 not 0, this means that the chances of the model correctly predicting whether a sample is normal adjacent or skin swab in origin is only marginally better than a blind guess.

Tumour histology analysis

The effect of any histological differences between tumours on their bacterial content was also examined (Table 2).

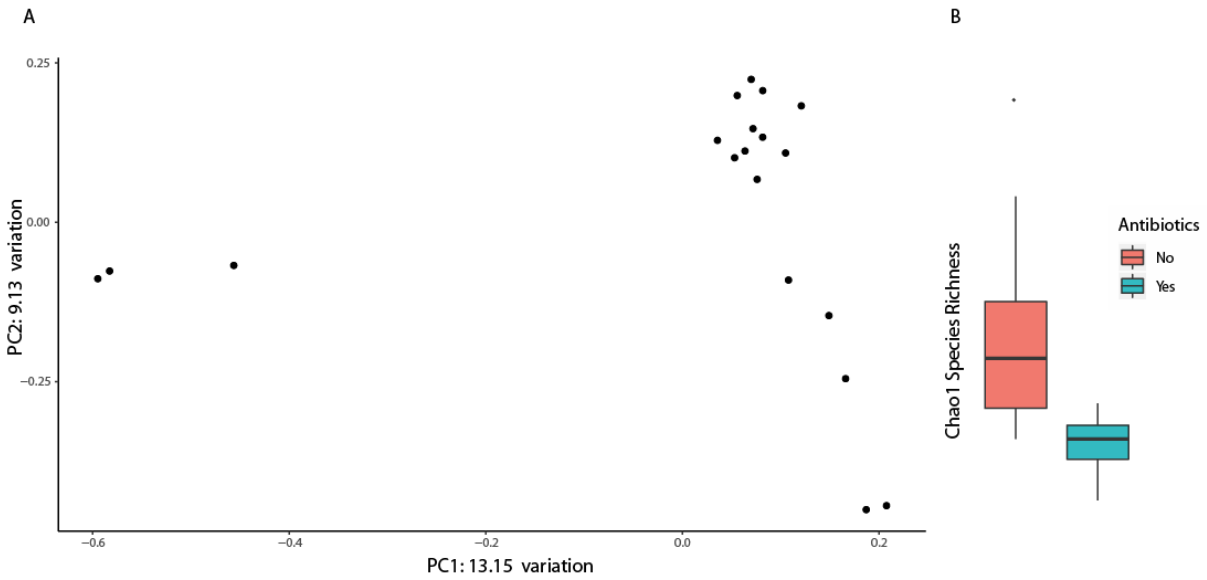


Figure 9: *PcOA comparison of tumour samples, via alpha and beta diversity.* ***(A)*** *Bray-Curtis dissimilarity between sample variations within the tumour samples.* ***(B)*** *Alpha diversity boxplot showing Chao1 species richness in tumour samples is significantly decreased in patients administered antibiotics.*

Table 2: ***Statistical analysis of the effect of histological features on alpha (Chao1) and beta (Bray Curtis) diversity.*** *Table shows associated p-values of tests performed.*

	Chao1	Bray Curtis
Surgery	**0.43**	**0.714**
Tumour Grade	**0.46**	**0.803**
Probiotic	**0.92**	**0.578**
Antibiotic	**0.019**	**0.507**
Necrosis	**0.68**	**0.683**
Metastases	**0.77**	**0.091**
Skin Involvement	**1**	**0.707**
Age	*0.805*	*0.988*
Tumour Size	*0.59*	*0.902*
Test Performed	*Wilcoxon Rank sum test* ***Spearman Correlation***	

When viewed in isolation in Figure 9, the tumour samples are not clustered as closely as when visualised in conjunction with normal adjacent tissue and although a considerable amount of metadata was available to attempt to explain this distribution, the only significant association detected was between patient antibiotic administration and alpha diversity.

DISCUSSION

Considerable recent research urging caution when undertaking sequence-based analysis of low biomass ecological niches particularly with regards to environmental contamination dictated the starting point of this study. An important aspect of any analysis into the bacterial composition of tumour tissue is to show that environmental or kit contamination is not the driving determinant of any bacterial community identified. Any microbiological survey is susceptible to contamination, but low biomass samples are disproportionately affected. In this instance, through the use of a robust negative control strategy and cutting edge bioinformatic tools for retrospective removal of contaminant sequencing reads, it can be stated that any potential contamination of samples used in this study has been contained, to the extent possible with bioinformatic software.

The community structure in samples amplified with V1-V2 primers was grossly similar to those amplified with V3-V4 primers upon visual inspection of sample composition plots, and there were no significant difference in terms of Chao1 species richness. This is reassuring, in that the 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 and bacterial contamination had been filtered out. As can be seen in Figure 3, samples amplified with primers targeting the V1-V2 region have a consistently and significantly higher number of reads. This is not the case at the end of the DADA2 pipeline, but once the unclassified reads are filtered out, most of which are shorter in length and align to the human genome, there is a significant reduction in the number of reads in the V3-V4 samples that is not seen in the V1-V2. This stems from an underreported problem in low biomass microbiome research, in that when the ratio of host DNA is overwhelming, human mitochondrial DNA can be amplified by primers targeting the 16S region. While human contamination is a very common problem in amplification-free WGS sequencing strategies (17), it is rarely reported as an issue in amplicon based sequencing strategies. In this instance, this has been particularly acute as the two most abundant ASVs in the entire dataset, when blasted, give as their top scoring hits the GenBank sequence MN516694.1 which is defined as "*Homo sapiens isolate S90_f1_ath haplogroup W1b1 mitochondrion.*" Both ASVs are absent from V1-V2 samples. An important lesson to

be learned from this is that while we were fortunate that our samples were of low complexity, relative to faecal samples for example, the impact of primer choice on eventual read depth must be considered if undertaking a survey of a “tract” biopsy such as the respiratory tract or digestive tract, which may require a greater sequencing depth to fully characterise (18).

The comparison of paired samples with and without DNAse treatment can be expected to reveal the extent to which extracellular DNA is distorting the detected bacterial community structure of an environment through sequencing. In this instance, as can be seen from Figure 6, DNase treatment had no effect on the composition of the samples. This can be seen visually in the sample composition plot, but also statistically in that there is no significant difference between the two groups in either alpha diversity or beta diversity. This informs that any bacterial DNA found in the samples originates either from live or dead but still intact bacteria.

The key aim of this study was to definitively confirm or deny the presence of bacteria within breast tumours, and surrounding tissue, and if present, hypothesise where these communities could have originated. Our initial alpha and beta diversity analysis, as shown in Figure 7, indicates that tumour samples have a significantly lower alpha diversity than their paired normal adjacent or skin swab samples, and while they samples do not cluster separately on a PcOA plot, the clustering is considerably more concentrated for tumour samples than the other two. When this information is combined with the hierarchical clustering and sample composition plots in Figure 8, we can see that the samples are all broadly similar in overall structure, with the skin-associated *Staphylococcaceae* (19) and *Corynebacteriaceae* (20), the dominant families present overall, providing a suggestion as to the origin of the bacteria found in these normal adjacent and breast tumour biopsies. This is unsurprising, as microbiome samples from closely proximal body sitescommonly share taxonomic traits (21). Despite this broad similarity, the clustering of samples does indicate that differences do exist particularly between the tumour samples and the non-tumour samples. This is of considerable interest due to the diagnostic and therapeutic potential that bacteria selectively colonising tumours over the surrounding tissue would have. Analysis of the mean proportion of a sample that a particular family occupied showed significant differences between the tumour and both non-tumour samples. The family level differences found were further explored

using Deseq2-based analysis of differentially enriched taxa at the ASV level. This further highlighted that differences exist between tumoural and non-tumoural samples.

At both family and ASV level, the skin swab and normal adjacent samples have no significantly enriched taxa, and both show increases in the classically skin-associated taxa of *Staphyloccus(aceae)* and *Corynebacterium(aceae)* which dominate the dataset, when compared with tumour samples. The tumour samples present with a more varied range of enriched taxa. Interestingly, the *Fusobacteriaceae* family, which has previously been implicated in a variety of human cancers, most notably colorectal cancer (22), was found to be elevated in tumour samples when compared with both non-tumour sample types. While at the ASV level it should be noted that while *S. capitis*, *S. hominis* and *S. caprae* are all elevated in skin swab and normal adjacent samples, and are either aerobic in the case of *S. capitis* or shown to considerable more suited to aerobic conditions than not, as is the case for *S. hominis and S. caprae*, it is the truly facultative *S. aureus* that is elevated in the tumour samples. Given that members of the *Fusobacteriaceae* family are all either facultative or anaerobic, this lends credence to the theory that the hypoxic regions known to be characteristic of tumours but absent in normal adjacent tissue or skin swabs, could have a determining impact on bacterial community composition (23,24). One caveat of the conclusions drawn, particularly from the different species of *Staphylococcus*, is that while they can be discriminated using regions of the 16S rRNA gene fragment (25), the differences are at most a few base changes over an entire variable region, meaning sequencing errors could lead to false speciation events. In this instance, the sequencing data were quality-filtered to ensure a minimum per base quality of 30 according to the Phred scoring system, which equates to 1 error in every 1000 bases meaning we can be confident of the accuracy of the results. That being said, a more comprehensive analysis of the complex staphylococcal community present in these samples using a different marker gene such as the *tuf* region could yield additional valuable information (26,27).

Finally, the Random Forest ensemble learning method was employed in an attempt to reveal unique microbial profiles between the sample groups undetected by standard multivariate techniques. A distinct microbial signature was detected for tumour samples, as opposed to either skin swab or normal adjacent samples. This

was evidenced by the area under the receiver operator curve being significantly higher when differentiating tumoural from non-tumoural samples (0.801 and 0.805) than when attempting to distinguish skin swab.

The origin of the bacteria in these tumours is of considerable interest. For this reason, it is important to investigate all possible relationships between groups, exploring whether bacterial profiles hold true within patients, within sample types, or within the culture status of the samples. Unfortunately, the only significant interaction detected was between antibiotic administration and alpha diversity, which is unsurprising. We suspect that due to the well-documented heterogeneity of tumours (28), there is no guarantee that the section of the tumour sent for sequencing matched any of the histological indications provided at the time of tumour removal and clinical assessment.

This study indicates the presence of endogenous bacterial communities in both malignant and non-malignant breast tissue in the majority of cases, and that the two can be differentiated based on their bacterial composition. Future work focusing on bacterial biomarker discovery would benefit from strain level analysis of these communities provided by whole genome sequencing methods.

References:

1. Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Ruddy, K., Tsang, J. and Cardoso, F. (2019) Breast cancer. *Nature Reviews Disease Primers*, **5**, 66.
2. Cummins, J. and Tangney, M. (2013) Bacteria and tumours: causative agents or opportunistic inhabitants? *Infectious Agents and Cancer*, **8**, 11.
3. Danino, T., Prindle, A., Kwong, G.A., Skalak, M., Li, H., Allen, K., Hasty, J. and Bhatia, S.N. (2015) Programmable probiotics for detection of cancer in urine. *Sci Transl Med*, **7**, 289ra284-289ra284.
4. Fricker, A.M., Podlesny, D. and Fricke, W.F. (2019) What is new and relevant for sequencing-based microbiome research? A mini-review. *Journal of Advanced Research*, **19**, 105-112.
5. Urbaniak, C., Cummins, J., Brackstone, M., Macklaim, J.M., Gloor, G.B., Baban, C.K., Scott, L., O'Hanlon, D.M., Burton, J.P., Francis, K.P. *et al.* (2014) Microbiota of human breast tissue. *Appl Environ Microbiol*, **80**, 3007-3014.
6. Riquelme, E., Zhang, Y., Zhang, L., Montiel, M., Zoltan, M., Dong, W., Quesada, P., Sahin, I., Chandra, V., San Lucas, A. *et al.* (2019) Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. *Cell*, **178**, 795-806.e712.
7. Flemer, B., Lynch, D.B., Brown, J.M., Jeffery, I.B., Ryan, F.J., Claesson, M.J., O'Riordain, M., Shanahan, F. and O'Toole, P.W. (2017) Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut*, **66**, 633-643.
8. Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J. and Walker, A.W. (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, **12**, 87.
9. de Goffau, M.C., Lager, S., Salter, S.J., Wagner, J., Kronbichler, A., Charnock-Jones, D.S., Peacock, S.J., Smith, G.C.S. and Parkhill, J. (2018) Recognizing the reagent microbiome. *Nature Microbiology*, **3**, 851-853.
10. Eisenhofer, R., Minich, J.J., Marotz, C., Cooper, A., Knight, R. and Weyrich, L.S. (2019) Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends in Microbiology*, **27**, 105-117.
11. Davis, N.M., Proctor, D.M., Holmes, S.P., Relman, D.A. and Callahan, B.J. (2018) Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, **6**, 226.
12. Knights, D., Kuczynski, J., Charlson, E.S., Zaneveld, J., Mozer, M.C., Collman, R.G., Bushman, F.D., Knight, R. and Kelley, S.T. (2011) Bayesian community-wide culture-independent microbial source tracking. *Nat Methods*, **8**, 761-763.
13. Yang, B., Wang, Y. and Qian, P.-Y. (2016) Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics*, **17**, 135.
14. Lauder, A.P., Roche, A.M., Sherrill-Mix, S., Bailey, A., Laughlin, A.L., Bittinger, K., Leite, R., Elovitz, M.A., Parry, S. and Bushman, F.D. (2016) Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome*, **4**, 29-29.
15. Nelson, M.T., Pope, C.E., Marsh, R.L., Wolter, D.J., Weiss, E.J., Hager, K.R., Vo, A.T., Brittnacher, M.J., Radey, M.C., Hayden, H.S. *et al.* (2019) Human and Extracellular DNA Depletion for Metagenomic Analysis of Complex Clinical Infection Samples Yields Optimized Viable Microbiome Profiles. *Cell Reports*, **26**, 2227-2240.e2225.
16. Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32.
17. Marotz, C.A., Sanders, J.G., Zuniga, C., Zaramela, L.S., Knight, R. and Zengler, K. (2018) Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome*, **6**, 42.

18. Zaheer, R., Noyes, N., Ortega Polo, R., Cook, S.R., Marinier, E., Van Domselaar, G., Belk, K.E., Morley, P.S. and McAllister, T.A. (2018) Impact of sequencing depth on the characterization of the microbiome and resistome. *Scientific Reports*, **8**, 5890.
19. Otto, M. (2010) Staphylococcus colonization of the skin and antimicrobial peptides. *Expert Rev Dermatol*, **5**, 183-195.
20. Oh, J., Conlan, S., Polley, E.C., Segre, J.A. and Kong, H.H. (2012) Shifts in human skin and nares microbiota of healthy children and adults. *Genome medicine*, **4**, 77.
21. Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J. and Huttenhower, C. (2012) Microbial Co-occurrence Relationships in the Human Microbiome. *PLOS Computational Biology*, **8**, e1002606.
22. Warren, R.L., Freeman, D.J., Pleasance, S., Watson, P., Moore, R.A., Cochrane, K., Allen-Vercoe, E. and Holt, R.A. (2013) Co-occurrence of anaerobic bacteria in colorectal carcinomas. *Microbiome*, **1**, 16-16.
23. Cummins, J. and Tangney, M. (2013) Bacteria and tumours: causative agents or opportunistic inhabitants? *Infect Agent Cancer*, **8**, 11.
24. Felfoul, O., Mohammadi, M., Taherkhani, S., de Lanauze, D., Zhong Xu, Y., Loghin, D., Essa, S., Jancik, S., Houle, D., Lafleur, M. *et al.* (2016) Magneto-aerotactic bacteria deliver drug-containing nanoliposomes to tumour hypoxic regions. *Nature Nanotechnology*, **11**, 941-947.
25. Mellmann, A., Becker, K., von Eiff, C., Keckevoet, U., Schumann, P. and Harmsen, D. (2006) Sequencing and staphylococci identification. *Emerg Infect Dis*, **12**, 333-336.
26. Hwang, S.M., Kim, M.S., Park, K.U., Song, J. and Kim, E.-C. (2011) Tuf gene sequence analysis has greater discriminatory power than 16S rRNA sequence analysis in identification of clinical isolates of coagulase-negative staphylococci. *Journal of clinical microbiology*, **49**, 4142-4149.
27. Li, X., Xing, J., Li, B., Wang, P. and Liu, J. (2012) Use of tuf as a target for sequence-based identification of Gram-positive cocci of the genus Enterococcus, Streptococcus, coagulase-negative Staphylococcus, and Lactococcus. *Ann Clin Microbiol Antimicrob*, **11**, 31-31.
28. Dagogo-Jack, I. and Shaw, A.T. (2018) Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology*, **15**, 81-94.

Chapter IV

Development of novel methodology for study of bacterial DNA from FFPE samples

This chapter is under review as*:*

"Protoblock - A biological standard for formalin fixed samples"

Yensi Flores Bueso, Sidney P Walker, Glenn Hogan, Marcus Claesson, Mark Tangney

Microbiome.

This chapter is under review as:

"Characterisation of FFPE-Induced Bacterial DNA Damage and Development of a DNA Repair Method for Metagenomics & Metataxonomics"

Yensi Flores Bueso*, Sidney P Walker*, Mark Tangney

Nucleic Acids Research.

*Authors contributed equally.

ABSTRACT

Background The role of the microbiome in health status is an expanding research area and in recent times, body sites classically considered sterile have been found to harbour an endogenous microbiome. One of the key rate limiting factors in progression of such research is difficulty in accessing sufficient tissue samples for statistically significant analysis to be carried out or to perform retrospective analyses. FFPE tissue represents the biggest repository of human tissue samples and could represent a vital resource for expanding microbiome research. Currently, there are several key features which limit bacteria related data generation from this material: i) DNA damage inherent to formalin fixation; ii) a high ratio of host to bacterial DNA, impairing sequence and PCR-based analyses; iii) inefficient DNA extraction methods, leading to poor sensitivity and data bias; and iv) vulnerability to contamination.

Aims We sought to develop a method for processing of FFPE samples to yield improved quantity and range of bacterial DNA present in samples than currently available methods, of the quality required for 16S sequencing and whole genome sequencing.

Methods A laboratory process was developed where samples undergo host DNA depletion, bacterial lysis, formalin crosslink digestion, DNA purification and DNA repair. The method was developed and validated using bespoke FFPE mock community models, FFPE murine samples, and clinical human tissue samples. DNA quantity and quality in terms of fragment length and sequence fidelity was assessed by qPCR and whole genome shotgun sequencing. The method was validates as a tool for microbiome research using 16S rRNA gene sequencing with the results compared against paired samples extracted with the current gold standard QIAGEN QIAamp FFPE tissue kit.

Results i) First, a mock community study model was developed and validated bioinformatically. This ‘Protoblock’ permitted a precise representation of biological material ‘before and after’ FFPE treatment, enabling the study to relate the outputs of laboratory analyses to reality. ii) This was used to characterise the nature and severity of FFPE-induced damage in bacterial DNA, followed by development of aneffective strategy for repairing it, based on the Base Excision Repair system.

Analyses of outputs from qPCR, high resolution melt analysis, Sanger Sequencing Shotgun Sequencing analysis were used to determine the most effective DNA repair strategy. iii) Bioinformatic validation of the combined method shows a significant improvement over the current gold standard QIAGEN QIAmp FFPE tissue kit, using both mock communities and FFPE murine faecal samples.

Conclusion This novel method may precipitate the reliable use of standard clinical FFPE tissue samples for modern bacterial sequencing studies.

INTRODUCTION

As DNA sequencing sensitivity and accuracy increases, sites previously considered sterile have also presented detectable microbial profiles. These discoveries have led to a greater demand for patient samples to undertake sequencing and other qualitative experiments such as qPCR for particular bacterial species. In an attempt to satisfy this increased demand, the use of formalin fixed paraffin embedded tissue for research involving bacterial DNA has been explored, with several recent studies published using FFPE samples as a starting point [1-6]. FFPE blocks are considered the gold standard for post-operative tissue storage in hospital settings and their reliable use for DNA analyses could open up a trove of potential samples for research. However, at present, no specific method exists for bacterial DNA in FFPE samples. There are a plethora confounding features present when carrying out sequence-based analysis of bacterial communities [7], and when coupled with the criticisms levelled at recent sequencing experiments targeting similarly challenging sample types (8) it is unlikely that large scale metagenomics studies using FFPE samples will remain tenable without the development of dedicated methodologies and biological standards. The key characteristics of FFPE samples that impair effective microbial analysis are:

- Formalin-derived crosslinks and damages to DNA present in the sample(9)
- A high ratio of host to bacterial DNA(10)
- All FFPE DNA extraction methods to date are optimised for human cells
- The extent of processing necessary leaves samples vulnerable to contamination
- No standards exist to validate the effects of the above on downstream analysis.

Many studies have characterised FFPE-induced damage in human DNA, yet the effects on bacterial DNA remain uncharacterised. The impact on data generation from FFPE-induced DNA damage is expected to be much more significant in bacterial studies when compared to human studies. Where human studies benefit from high DNA quantities in samples and a well-known reference genome, in metagenomics research, the DNA template is often minimum and concealed in a high human DNA background, and the sequences studied are not limited to one genome. FFPE-induced damage to bacterial DNA, and the effect on downstream

analysis, needs to be accurately characterised. This will inform on requirements for a method to repair DNA damage, improving the fidelity of future analyses.

To repair the damaged DNA, reconstitution of the intrinsic Base Excision Repair (BER) pathway in vitro shows considerable potential. This involves the excision of a damaged base by a DNA glycosylase enzyme, backbone incision facilitated by AP lyase, Ends processing by Polynucleotide Kinase, Gap filling by DNA Polymerase and nick ligation by DNA ligase.

In low bacterial biomass biopsy samples, such as most non tract human biopsies, host DNA constitutes in excess of 99 % of total DNA. This severely limits metagenomic studies, as the vast majority of sequencing reads available are invested by this background human DNA. This is of critical concern, particularly for whole genome shotgun (WGS) methods (11). It has been also shown to affect the outputs of 16S rRNA amplicon sequencing, since in reactions of low bacterial to human DNA ratios, human DNA can be annealed and amplified during 16S PCR (12). Furthermore, a reduction in bacterial diversity and particularly rare bacterial taxa can occur during dilutions made to avoid overloading DNA in PCR reactions [13]. For these regions, any reduction in the ratio of background mammalian to target bacterial DNA would improve readout. DNAse treatment can reduce the quantity of intact background DNA, if it's activity can be targeted to mammalian cells, e.g. by restricting access of the DNAse enzyme to only mammalian cells. Mammalian specific-membrane permeabilisation may achieve this.

Bacterial lysis is a critical step in sample processing for metagenomic analysis. It can be major source of bias in community composition, as lysis methods that favour particular taxa will cause overrepresentation in the final analysis [12, 14-18]. Many methods for unbiased bacterial lysis of non-fixed samples have been proposed and applied, including bead-beating, enzymatic lysis, detergents and denaturing agents [15, 16, 19]. Recently, several studies have agreed that bead-beating is the lysis method that yields higher uniformity of bacterial lysis and have shown that combining bead-beating with other methods shows further improvements in uniformity [19, 20].

FFPE samples are characterised by DNA damage that includes high levels of fragmentation and DNA damage reducing the recovery of PCR/sequencing readable

DNA [21, 22]. As previously mentioned, FFPE samples typically have low bacterial biomass concealed by large quantities of DNA from the larger human genome. Bead-beating decreases DNA yields by causing DNA fragmentation leading to the formation of chimeras during PCR [23-25], which would be particularly detrimental for FFPE samples. For this sample type, lysis must be performed under conditions that do not negatively affect the integrity of DNA, such as enzymatic lysis. Accordingly, the *Association of Biomolecular Resource Facilities Metagenomics Research Group* developed a mix of six lytic enzymes (achromopeptidase, chitinase, lyticase, lysostaphin, lysozyme, and mutanolysin) that target the cell wall of bacteria, yeast, and fungi, and is able to lyse recalcitrant endospores [26]. The incorporation of this enzyme, known as Metapolyzyme (Sigma-Aldrich), in sample preparation has been shown to increase the recovery of spheroplasts or protoplasts, and improve the overall DNA recovery across taxa in multiple sample types [25, 26]. Recently, a metagenomic study was performed on ancient DNA specimens (with similar levels of DNA damage as FFPE), validating the efficacy of Metapolyzyme over traditional bead beating methods in this sample type (27).

The issue of biological standards was largely overlooked in the early years of microbiome research. As 15 years have passed since the seminal work by Craig Venter in 2004 [28], several publications have taken stock of progress so far, and highlighted areas for improvement. A recurring theme within these publications has been the lack of standards available [29-32]. Given the numerous potential sources of error associated with FFPE samples outlined previously, more than perhaps any other sample type, FFPE tissue urgently requires the development of standards to ensure the validity of results and to promote reproducibility.

With this in mind, it was deemed appropriate to initiate this study with development of such an FFPE study model, to

- inform on the extent and nature of DNA damage due to FFPE, guiding the development of a DNA repair strategy
- inform on any possible bias arising from ineffective bacterial lysis for example
- inform on the incorporation of environmental contaminant bacteria into the DNA sequencing library due to the extensive processing of samples required.

Study Aims

Portions of this project fall beyond the remit of bioinformatics research. This chapter focuses on different approaches incorporating bacterial sequence analysis during the design and validation stages of both the Protoblock FFPE biological standard and the DNA repair strategy. Following this, an assessment of the final protocol as a tool for metagenomics/metataxanomic analysis was carried out. This was performed using the Protoblock biological standard and formalin fixed mouse faeces as a higher biomass sample. In these cases the newly designed protocol was compared with the current gold standard, the Qiagen QIAmp FFPE kit. Lastly, low biomass samples of malignant formalin fixed patient breast tissue were processed using the novel method, and compared with paired fresh frozen samples

METHODS

A full description of the novel protocol developed can be found in appendix 1 of this thesis. A summary is displayed in Figure 1. All lab work was performed by other members of the Tangney lab.

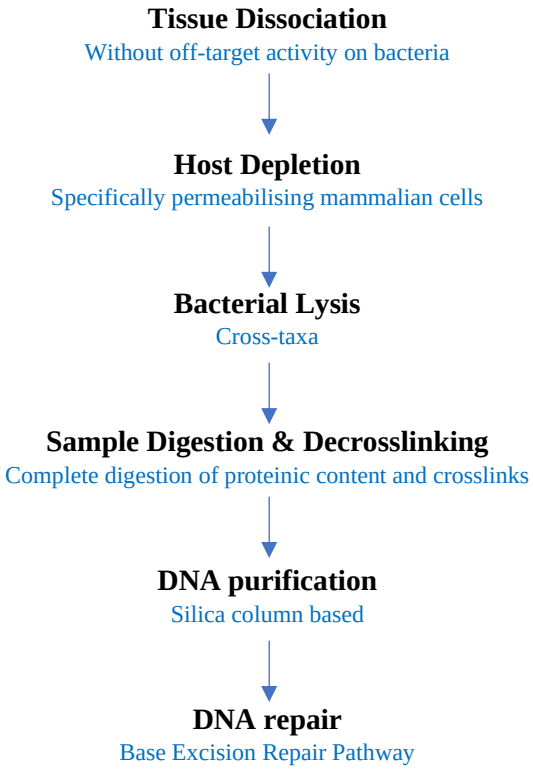


Figure 1: ***Full Protocol for bacterial DNA isolation from FFPE samples*** *– describing steps for process and in blue the requirements for the step*

Bioinformatic methods for data analysis

qPCR data analysis

Statistical analysis was performed in the base R environment (v3.6.1). Visualisations were carried out using the ggplot2 package (v3.2.1).

WGS sequence analysis

All metrics relating to sequence data were calculated in the Linux environment, and using the QUAST tool (v5.0.2) and statistical analysis performed in the base R environment (v3.6.1). Visualisations were carried out using the ggplot2 package.

Method for variant calling

Filtering HiSeq sequence data was quality filtered. Only very high quality bases were considered, to minimise the risk of sequencing errors causing false positive variants. Short fragments were also removed to reduce the likelihood of spurious alignments of regions from contaminant bacterial genomes. Trimmomatic (v0.38) was used to remove all reads shorter than 60bp in length, and to trim reads when the average per base quality in a sliding window of size 4 dropped below 30.

Alignment Of the three possible Burrows-Wheeler alignment tools, the BWA-mem aligner was used as the average read length was 150 bp, and BWA-mem (v0.7.17) is recommended when reads are over 70 bp in length. Default settings were used with the exception of allowing alignments with a minimum score of 0, rather than the default 30. Given the stringent parameters used for read length and quality filtering, relaxing the minimum alignment score gave the best possible chance of variant detection. All samples were aligned to the original reference genomes.

Variant Calling Variant calling was done with BCF tools, using the BCF call function. The variants were then filtered using the norm and filter functions within BCF tools. Filtering was done to remove variants when the read depth was below 10, the quality was below 40, or when the variant identified was not supported by both the forward and reverse read of a read pair. The number of variants identified was then normalised between samples based on the read coverage in the initial alignment BAM file.

Validation Using the Picard tool within the GATK suite, all samples were down-sampled to ensure SNP: Coverage ratio remained constant when coverage was reduced to lowest level present in samples.

16S sequence analysis

The quality of the paired-end sequence data was initially visualised using FastQC v0.11.6, and then filtered and trimmed using Trimmomatic v0.36 to ensure a minimum average quality of 25. The remaining high-quality reads were then imported into the R environment v3.4.4 for analysis with the DADA2 package v1.8.0. After further quality filtering, error correction and chimera removal, the raw reads generated by the sequencing process were refined into a table of Amplicon Sequence Variants (ASVs) and their distribution among the samples. It is recommended that ASVs (formerly called 'Ribosomal Sequence Variants') be used in place of 'operational taxonomic units' (OTU), in part because ASVs give better resolution than OTUs, which are clustered based on similarity.

The following statistical analyses were carried out in R: Shannon alpha diversity and Chao1 species richness metrics, and Bray-Curtis distances, for analysis of beta diversity, were calculated using the PhyloSeq package v1.24, and the Vegan package v2.52. Beta diversity calculations produce distance matrices with as many columns and rows as there are samples; thus, beta diversity is often represented using some form of dimensionality reduction, in this case, using principal co-ordinates analysis (PCoA) with the Ape package v5.1. Hierarchical clustering, an unsupervised method that can reveal key taxa that distinguish their respective environments, was performed with the heatplot function in the made4 package v1.54. Differential abundance analysis was carried out using Deseq2 v1.2.0, which identifies differentially abundant features between two groups within the data. Tests of means were performed using the Mann-Whitney *U* test unless otherwise stated, and correlations were calculated using Spearman's rank correlation coefficient. Where applicable, false positive rates were controlled below 5% using the FDR procedure..

Despite not identifying the contaminant taxa themselves, the source tracker utility is invaluable in estimating the proportion of a sample (“Sink”) that may have originated in a negative control (“Source”) Decontam can remove taxa, based on presence or absence in negative controls, or inverse correlations with input DNA. This tool requires a threshold to be set, which can be dictated by SourceTracker. The effectiveness of this can then be confirmed by SourceTracker.

RESULTS

Table 1: ***Bacterial load of protoblocks used for 16S and WGS sequencing***

	Counts in microscope / volume measured for each type of FFPE block [single strain to mixed strain]			**Calculations for DNA purified from blocks**			
Cell type	**Microscope Counts in block**	**Cells/µl in mixed block**	**Cells in 15 µm slide**	**Cells DNA extraction**	**Genomes in elution**	**Cells in 16S PCR**	**Ratio**
4T1 Mouse Tumour Cell Line	2.20E+07	8.85E+04	1.06E+06	1.28E+07	2.55E+05	**3.83E+06**	
E. coli	3.10E+07	1.25E+05	1.46E+06	1.75E+07	3.50E+05	**5.25E+06**	0.17
S. aureus	9.01E+06	3.63E+04	4.22E+05	5.06E+06	1.01E+05	**1.52E+06**	0.05
B. longum	8.50E+06	3.43E+04	4.01E+05	4.81E+06	9.62E+04	**1.44E+06**	0.05
L. amylophilus	8.12E+07	3.28E+05	3.74E+06	4.48E+07	8.97E+05	**1.35E+07**	0.43
B. thetamicroniota	5.82E+07	2.34E+05	2.71E+06	3.26E+07	6.51E+05	**9.77E+06**	0.31

Validation of the biological standard

DNA was obtained from Protoblocks containing a mix of the five bacterial strains specified in Table 1, and was purified with the QIAGEN QIAamp FFPE tissue kit as described the methods provided in Appendix 1. qPCR recovery was determined by quantifying long DNA fragments (450 bp) specific for each bacterial strain and normalised to a loaded concentration of 10^5 cells. Bacterial quantities in qPCR reactions are as specified in Table 1. The outputs of these reactions show a clear bias of sample prep toward Gram-negative bacteria, with an average recovery 2.5 log-fold higher ($p<0.001$) than Gram-positive bacteria. Recovery determined by qPCR was 0.17% for *Staphylococcus*, 1% for *Lactobacillus*, 1.5% for *Bifidobacterium*, 3.3% for *E. coli* and 15% for *Bacteroides* (Figure 2B (iii)). This was further confirmed with 16S rRNA gene sequencing (Figure 2B (iv)). As seen in Figure 2B, the composition plot shows a clear bias towards Gram-negative bacteria. For example, while equal quantities of *Bacteroides* (29%) and *Staphylococcus* (27%) cells were present in the Protoblocks, DNA analyses (16S rRNA gene sequencing and qPCR) reported *Bacteroides* represent more than 50% of the sequences recovered, while the more difficult to lyse Gram-positive *Staphylococcus* is almost

lost from the plot. Therefore, even in FFPE samples, a mechanism for the lysis of Gram-positive bacteria must be considered for metagenomics studies. Furthermore, the use of standards such as the Protoblock is essential when working with FFPE samples, and even more so when using protocols that are not designed for this sample type and study purpose.

The Protoblock is susceptible to contamination in a similar way to clinical FFPE samples. The priority of the fixing process is to preserve the tissue for later histological analysis, not to prepare a sample suitable for high throughput bacterial sequencing. In this instance, contamination was detected as shown by the number of reads in the negative controls (Figure 4B (v)). It is unlikely to have had a significant effect on the overall biological signal. Given that the bacterial reads detected and their taxonomic classifications differ completely from those of the protoblocks analysed. However, it remains a threat for low biomass samples.

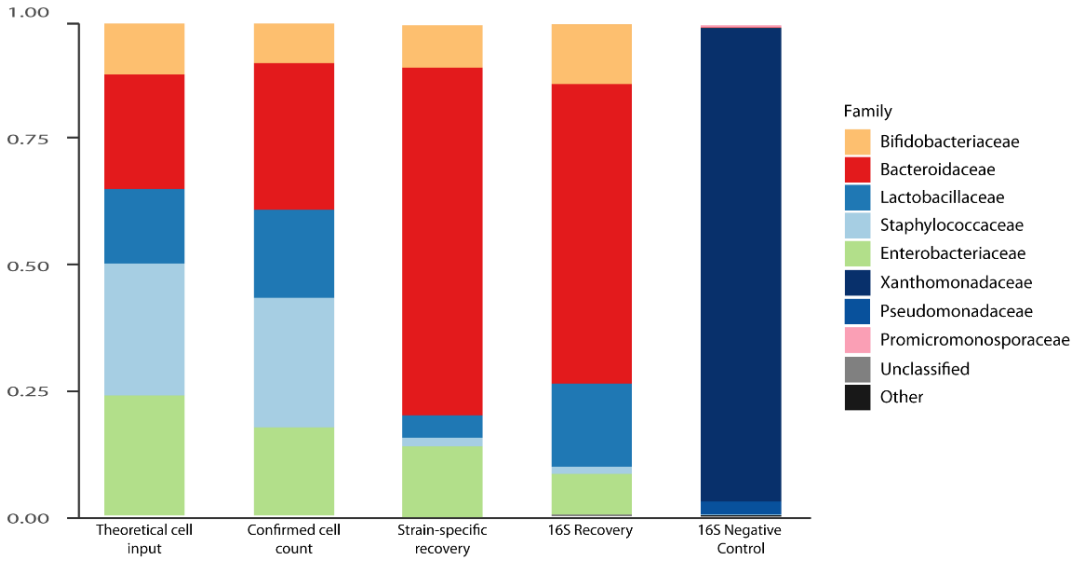


Figure 2: ***Evaluation of impacts on downstream analysis.*** *Measuring bias introduced by the DNA extraction process. Sample composition Bar plot of:* ***i)*** *Cells added to protoblock,* ***ii)*** *Confirmed bacterial counts per block,* ***iii)*** *Strain specific qPCR,* ***iv)*** *Pooled sample composition as per 16S rRNA gene sequence analysis,* ***v)*** *Composition of an empty protoblock (sterile agar only) processed in parallel with loaded protoblocks.*

The presence of DNA sequence artefacts as a result of FFPE was assessed in a simplified Protoblock model populated with *E. coli* K-12. DNA was extracted and its

concentration determined by qPCR and normalised to 10^6 genome copies. HRM was performed in three contiguous DNA fragments (length □ 100 bp) that make up a region of the InsH1 gene (See figure 3B (ii)). To determine the presence of any sequence aberrations in Protoblock FFPE DNA, their melting temperature (Tm) was compared with that of Non-fixed (NF) DNA and the differences measured. Figure 3B (i) shows the final Tm for each fragment investigated. Tm shifts with variable levels of significance were observed in all fragments. This is indicative of a change in the underlying DNA sequence, as would be expected in a clinical FFPE sample. To confirm these results, DNA from both samples was analysed by whole genome sequencing. Findings from the DNA melting temperature analysis correlated with the results of WGS, with statistically significant variations between the FFPE genome and the reference genome evident (see Figure 3C).

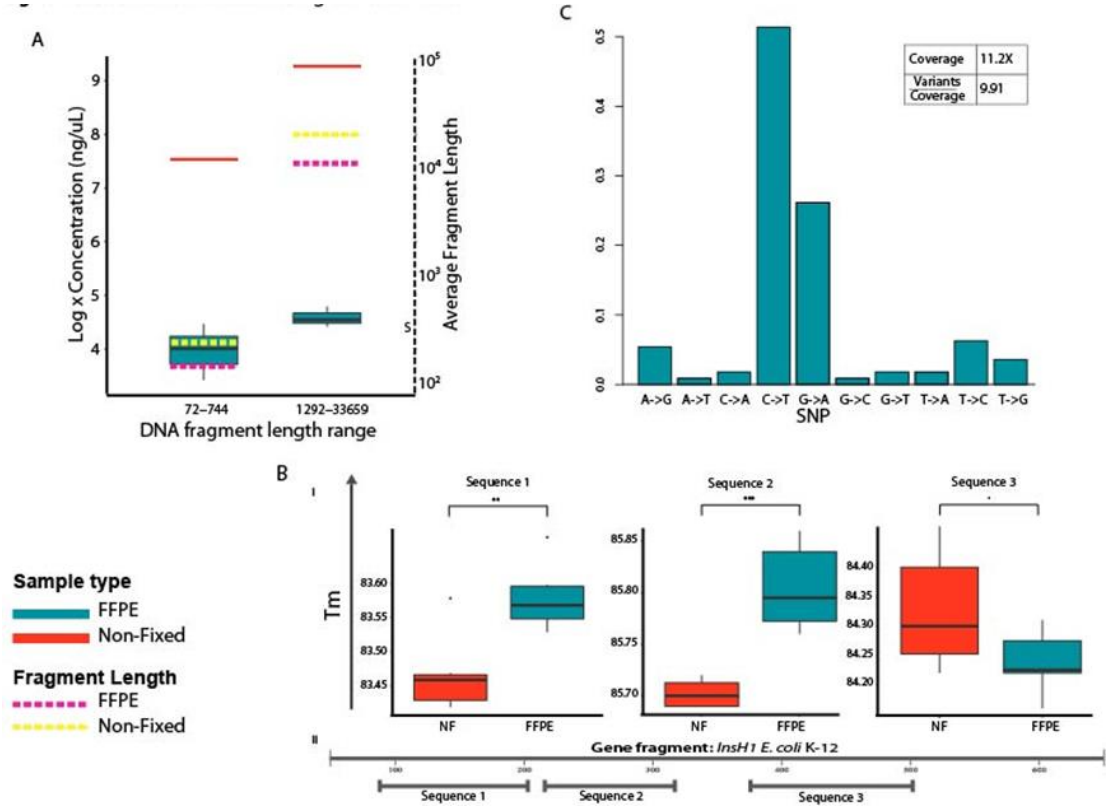


Figure 3: ***Assessment of DNA damage in Protoblocks. A) Evaluation of DNA integrity with Bioanalyser high-sensitivity tape station.*** *DNA concentration (boxes) is plotted on the y-axis and fragment length (dotted line) on the z-axis. Results were extracted from 2 peak-regions in the electropherograms (72-744 bp and >1,292). Here, the short fragment region FFPE samples had an average fragment length of 207 bp and an average concentration of 0.06 ng/µl, while its non-fixed counterpart had an average length of 462 bp and a concentration of 1.86 ng/µl. In the larger region fragment, FFPE samples had an average concentration of 0.099 ng/µl and an average fragment length of 13,119 bp, whereas NF sample had a concentration of 10.57 ng/µl and an average length of 31100 bp.* ***B) Evaluation of DNA sequence aberrations by high-resolution-melt analysis. i)*** *Box plots of normalised DNA quantities from protoblocks populated with E. coli (blue) and NF E. coli (red). Clear shifts in the melting temperatures in 2 of the 3 sequences were observed, with temperature shifts that were on average 0.1-0.5^{o}C apart from NF counterparts.* ***ii)*** *Schematic of sequences used for HRM analysis: 3 DNA fragments with an average length of 100 bp were analysed, for each test and each sample type, n = 6.* ***C) Confirmation of sequence alteration by WGS.*** *DNA from the same protoblocks as B was analysed by whole genome sequencing and subsequent variant calling against the reference genome E. coli K12 MG1655. Here, the rate of their occurrence is plotted against the y-axis. Variant calling, and level of coverage is measured using SAMTOOLS/BCFTOOLS.*

Validation of DNA repair strategy by WGS

The reconstitution of a BER system, targeting different types of DNA damage found on FFPE samples was addressed by mixing the pathways for the glycosylases treated in the system. Since FPG-BER (Figure 4a) yielded the best results for single glycosylase-BER reactions, this enzyme was combined with ENDO VIII and UDG and their efficiency in reducing sequence artefacts tested by HRM. As shown in Figure 4b, all combinations resulted in sequences with ΔTm lower than those of untreated FFPE DNA. The FPG + UDG mix showed the best performance at reducing the ΔTm (31 %), followed by FPG + Endo VIII (18 %). However, in terms of improving the PCR readability of a 500 bp fragment, FPG + Endo VIII (47% increase, $p < 0.01$) outperformed FPG + UDG (30% increase, $p < 0.01$), as measured by Taq qPCR. To confirm these results, normalised DNA concentration from 6 replicates for each BER mix and 6 unrepaired samples were pooled into one ($n = \Sigma 6$) and sent for analysis by WGS (Figure 4c). At this level of resolution, it is evident that the repair mix with FPG + Endo VIII offered the highest improvements in sequence quality in terms of providing *(i)* a coverage 4X higher than unrepaired, *(i)* 4X more total reads and quality filter (QF)-passed reads, and *(iii)* a 50% reduction in the number of variants detected per sequence coverage. This repair mix was thus selected as the best repair mix for bacterial FFPE DNA.

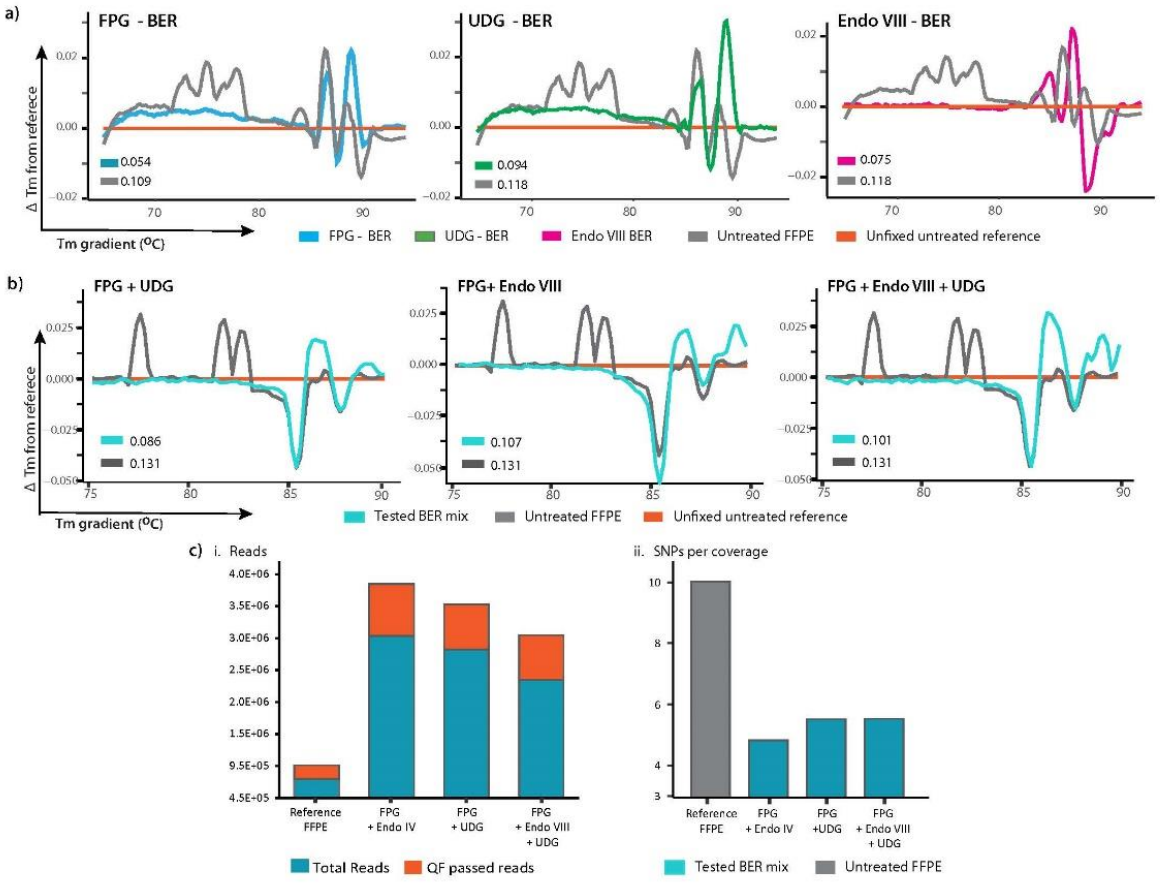


Figure 4: ***Reconstitution of BER pathway repairing FFPE DNA damage. a) Single glycosylase BER.*** *The BER pathway was reconstituted first as single pathways triggered by either UDG, FPG or Endo VIII. The efficiency of each system in correcting DNA damage was tested by HRM (n = 7 for each line). The more similar a DNA sequence is to the NF reference, the lower the difference in melting temperature (ΔTm closer to 0). FPG showed the highest efficiency in correcting FFFPE DNA damage as evidenced by the lowest ΔTm of 0.054.* ***b) Multiple glycosylase BER****. Mixes containing FPG show improved sequence quality as evidenced by reduced ΔTm vs untreated.* ***c) WGS****. To further confirm these results, six replicates treated with each mix were pooled (n = Σ6) and analysed by WGS. Data validated that all mixes improved the sequence (i) coverage, (ii) number of reads and QP reads and reduced the amount of SNPs (iii). The best performance in all cases was observed in the BER mix with FPG and Endo VIII.*

The sum of the above treatment strategies (decrosslinking and DNA repair) was tested by WGS in DNA sourced from Protoblocks containing the same 5 bacterial strains as previously described, fixed for 48 h and stored for 2 months. The dewaxed and lysed contents of the blocks were decrosslinked at 80 °C with a chaotrophe salt based buffer (Appendix 1). The purified DNA was repaired with the BER mix based on the FPG + Endo VIII repair pathway. Experimental replicates were pooled and

sent for WGS analysis. Results for this analysis are shown in Figure 5. The results obtained from exposing bacterial FFPE DNA to the proposed new protocol are compared with those from DNA from paired-samples treated with the reference Qiagen protocol (decrosslinking at 90 ^{o}C, without DNA repair), and paired NF DNA. These results indicate that bacterial FFPE DNA treated with the proposed method shows an improvement in integrity, readability, and sequence quality, as evidenced by: (i) Integrity [Average fragment length (a, b)]: Plotted in Figure 5a, are the average fragment lengths measured by bioanalyser. Fragment length of DNA treated with the new protocol (444 bp) is 3.3X longer than that treated with the reference protocol (136 bp). Importantly, this raises the average fragment length to that of fragments typically desired for 16S sequencing (460 bp). The same effect was observed in the length of fragments read by WGS, where fragment lengths were 2-3 bp longer on average (Figure 5b). (ii) Readability: With the new protocol, the number of Total Reads and (QF)-pass reads per layer of coverage were increased by 24 % and 34 % respectively, and the ratio of QF-passed to Total reads increased by 8.4 %. (iii) Sequence quality: This was measured in terms of number of sequence artefacts detected. The number of chimeric reads per coverage detected in samples treated with the new protocol was reduced by 57 % ($p = 0.37$) (Figure 5e). Similarly, the number of SNPs detected was reduced by 58% ($p = 0.41$) (Figure 5f). All of these findings are supported by results from quantitative PCR and Tm analysis. Although these improvements are not supported by statistical significance, given the considerable effect size, we are confident that this lack of significance is due to sample size alone. Altogether, the sum of strategies proposed here were thoroughly investigated by PCR/sequencing. These results consistently indicate an improvement in the sequence integrity, readability and quality of readable bacterial FFPE DNA.

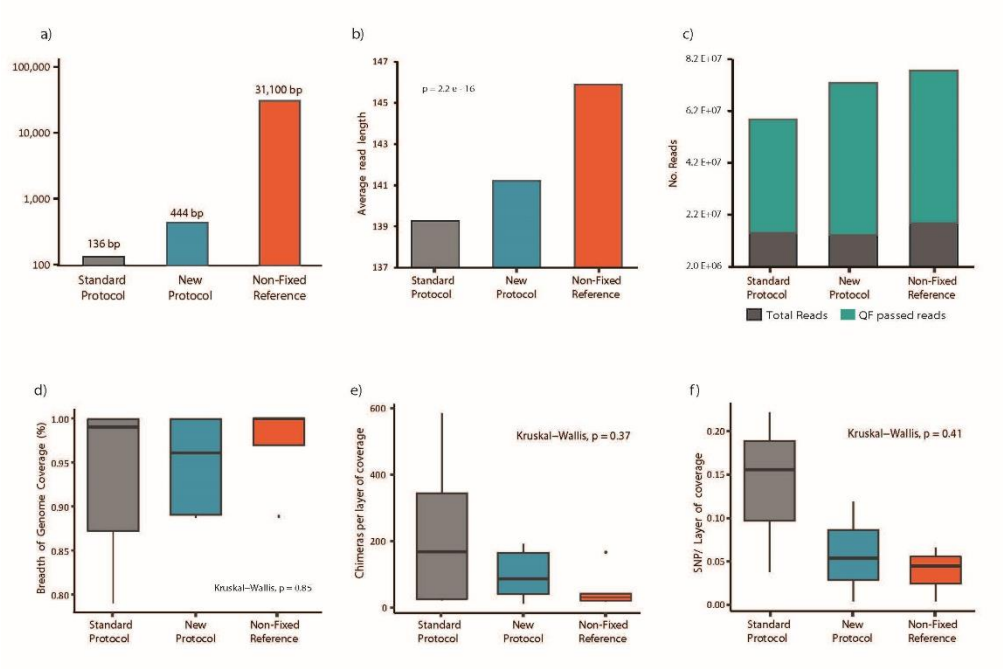


Figure 5: ***Combined protocol – bacterial DNA.*** *Outputs of Bioanalyser and whole genome sequencing for bacterial FFPE DNA exposed to the combined treatment (blue, labelled as New Protocol, Σn = 6). This was compared with that obtained from six pooled paired-samples decrosslinked with the reference protocol and unrepaired (grey, Labelled reference protocol, Σn = 6) and that from DNA obtained from NF samples with the same bacterial and DNA content (orange, Labelled NF, Σn = 3). Improvement in DNA readability, sequence quality and integrity was measured by:* <u>*Integrity*</u> *(fragment length):* ***(a)*** *bioanalyser* ***(b)*** *WGS.* <u>*Readability*</u>*:* ***(c)*** *Quantity of reads and filter pass reads per coverage.* ***(d)*** *% Breath of genome coverage.* <u>*Sequence quality*</u>*:* ***(e)*** *Number of chimeric reads per layer of coverage.* ***(f)*** *Number of SNPs per layer of coverage.*

Validation of the protocol by 16S sequencing

The level of processing required when creating a sequencing library from FFPE samples, coupled with the anticipated low biomass of the samples, makes them highly susceptible to contamination. Figure 6 shows a representative sample from each sample group, and each library preparation method. The "In House methods" are consistently more susceptible to contamination than the gold standard Qiagen method, and a controllable level of contamination is present in all sample types with the exception of FFPE breast samples, which are overwhelmingly contaminated. The output of the SourceTracker algorithm also indicates which negative controls were implicated in the contamination, and Figure 6A, shows the composition of these samples at the family level.

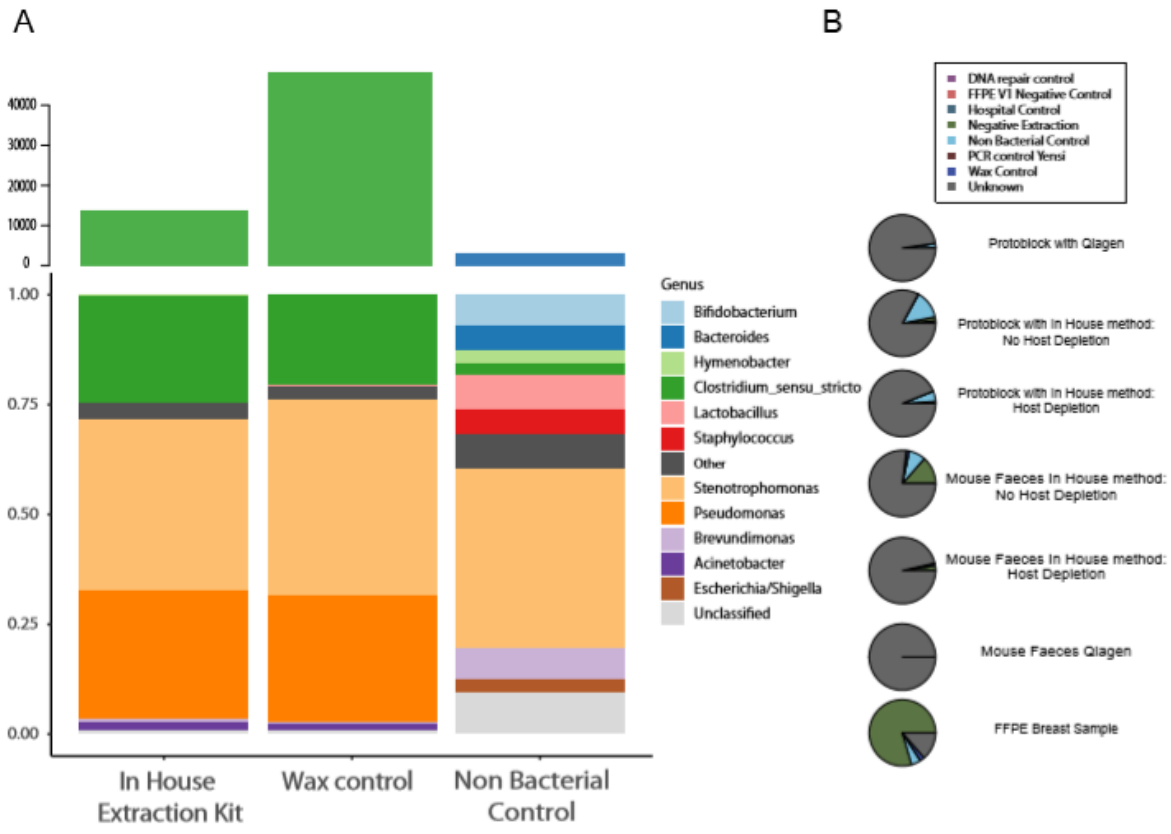


Figure 6: ***Summary of environmental contamination.*** **(B)** *Shows the output of the SourceTracker algorithm, with one representative pie chart per sample type indicating the degree of contamination present.* ***(A)*** *Shows the sample composition of the three negative controls implicated by the SourceTracker algorithm. Data indicate that although contamination is present in most samples only FFPE breast samples are overwhelmingly affected by environmental contaminants. In addition, only three of the eight negative controls are implicated.*

The use of Protoblocks with known bacterial composition allowed for accurate quantification of the number of sequencing reads lost due to contamination between the three different treatment groups. As seen in Figure 7, in both the Qiagen protocol (Q) and the *In house with host depletion* (HDP) protocols, the proportion of reads removed as part of the contamination control workflow was less than 5%. With the *in house without host depletion* (HDN) protocol almost a quarter of all reads obtained from these samples had to be removed.

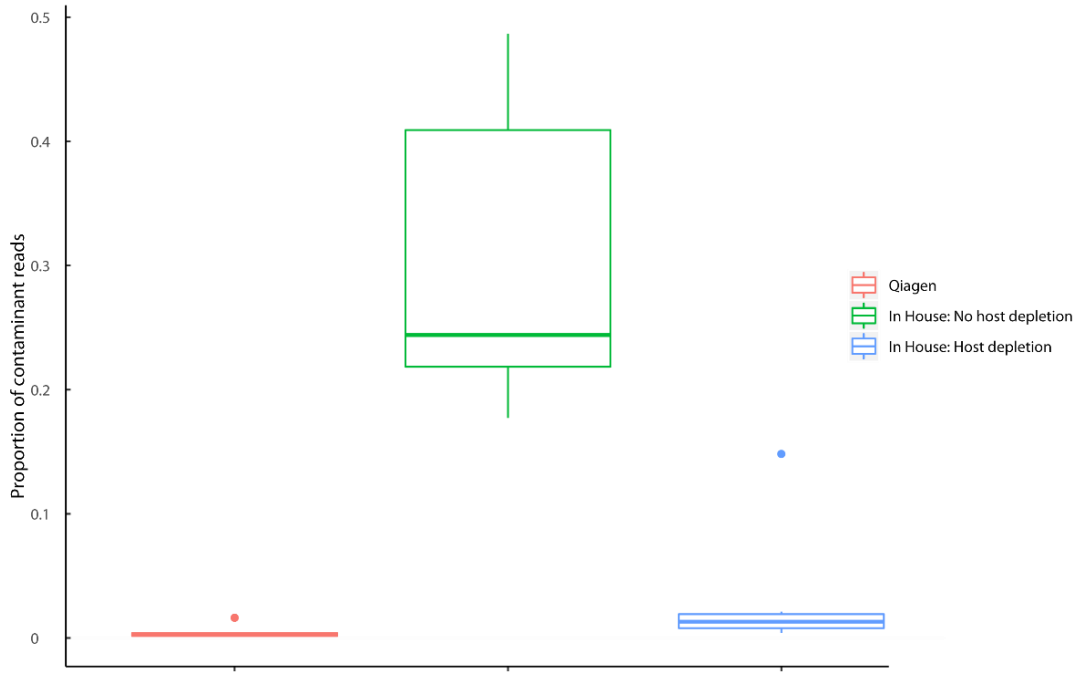


Figure 7: ***Proportion of reads lost due to environmental contamination introduced during processing.*** *The data indicates that while only a marginal percentage of reads are consumed by environmental contaminant DNA in the Qiagen and HDP samples, just under 30% of reads on average are lost in HDN samples.*

Samples labelled as HDP went through the DNA protocol (bacterial lysis, sample digestion, DNA purification and repair) plus a host depletion step, and HDN samples, did not include a host depletion step. The precise quantities of bacteria added to the FFPE mock communities can be seen in Table 1. This information allowed a robust analysis of methodological bias in terms of under or over-representation of different bacteria. As shown in Figure 8, the Q protocol, which is not optimised for bacterial DNA, showed statistically significant under or overrepresentations in all five genera present in the Protoblock, particularly in the case of *Bacteroides* and *Lactobacillus* which were over and underrepresented by

more than 20% respectively. In the HDN method, no significant bias was observed with lactobacillus, the deviation in *Bacteroides* was marginally significant, while all other genera were significantly under or overrepresented. The HDP method was the least susceptible to bias, with only the proportion of *E. coli* presenting as significantly different from what was theoretically present in the Protoblock.

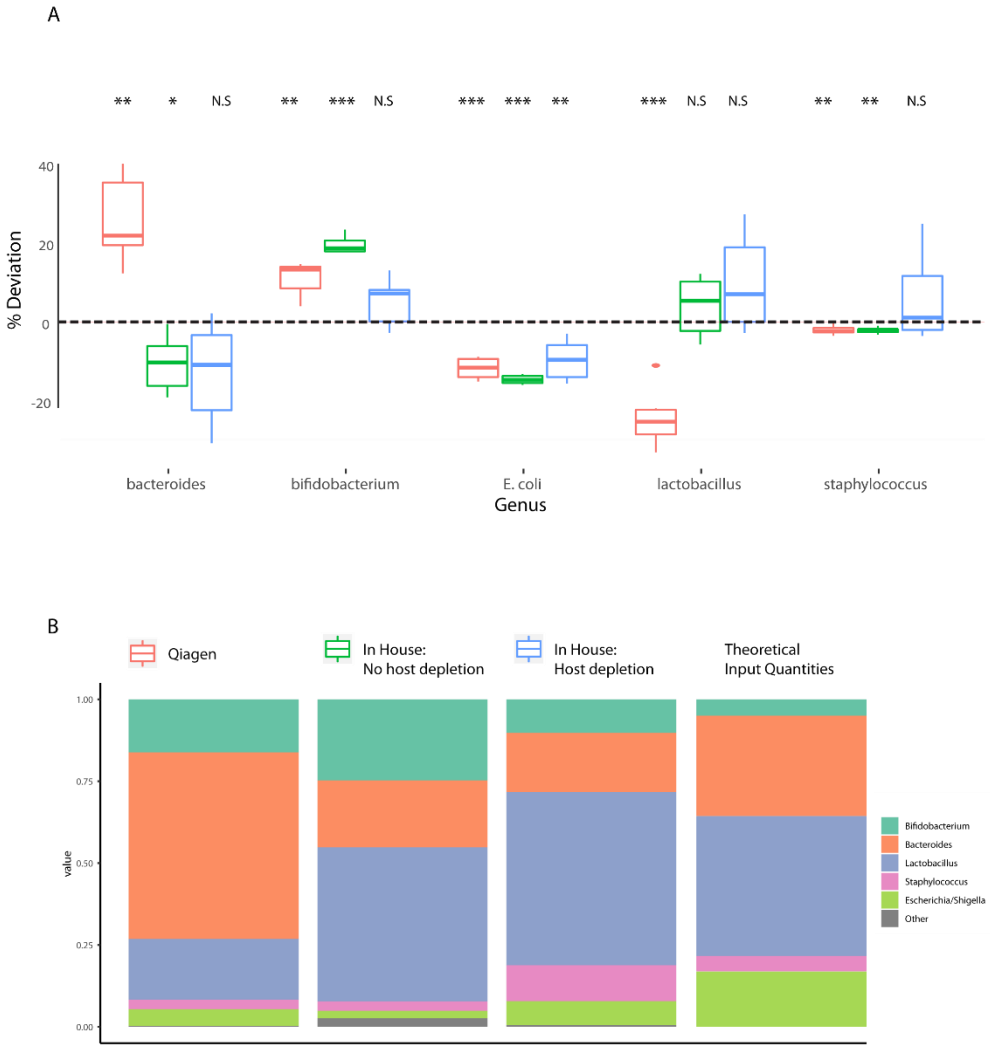


Figure 8: ***Assessment of bias in terms of bacterial community composition between methods****. **(A)** Shows the percentage deviation of bacterial composition per genera, per extraction method, from the original quantities input into the protoblock. **(B)** Shows sample composition of all samples merged by extraction kit, with the right most column representing the ideal proportions as dictated by the input quantities. Visually HDP has the least degree of bias over the five bacterial genera. This is confirmed statistically in* ***(A).***

Faecal samples

The comparisons facilitated by the protoblocks were complemented by mouse faecal samples, which were formalin fixed and paraffin embedded as described in methods (murine models) and their protocol included bacterial lysis, sample digestion, DNA purification and DNA repair, with host depletion + tissue dissociation (DT-P) or without any of these two treatments (DT-N). The community structure in these samples was considerably more complex than in the Protoblock.

Beta diversity analysis using Bray-Curtis dissimilarity shows no significant difference between the IHN and Qiagen methods. This can be seen visually as the samples cluster together, and is confirmed by PERMANOVA analysis, (p = 0.231). Both Qiagen and DT-N are significantly dissimilar to DT-P as per PERMANOVA, (p = <0.001) (Figure 8).

The driving factors behind the distinct clustering were assessed by searching for correlations between the dominant bacterial families seen in the samples, and either of the two principal coordinate axes. The correlations were carried out using Spearman's method, and multiple testing was controlled for using the FDR method. This was expanded upon in Figure 9, with a direct comparison of sample composition between FFPE vs Flash-frozen samples in the three treatment groups.

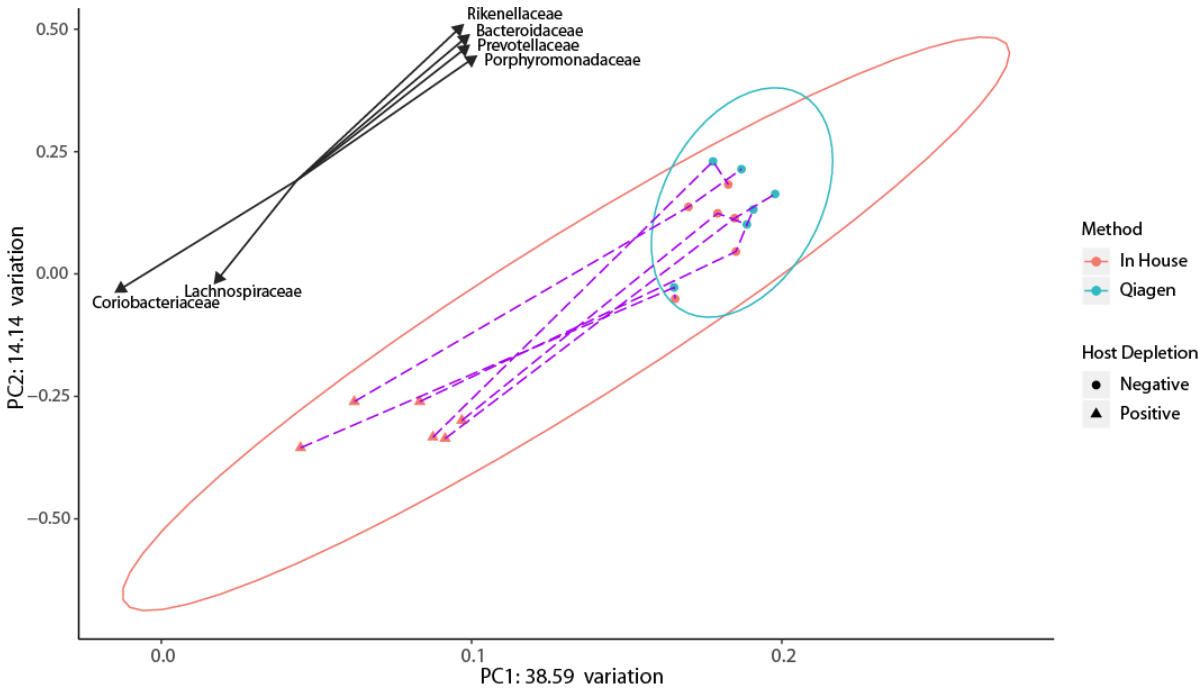


Figure 9: ***Principal coordinate analysis of matched murine samples.*** *Points coloured by extraction method, and shaped by host depletion status. PcOA plot supported by correlations of major bacterial families present in dataset with PC1 and PC2 values used to generate plot. Only significantly correlating families show, with significance tested for using Spearman's method. False discovery rate controlled for using FDR method. Data indicates that host depletion strategy has an effect on Gram negative bacteria.*

Figure 10A compares Q and DT-P paired samples. In this instance, the Gram positive *Coriobacteriaceae* and *Lactobacilliceae* were significantly elevated in terms of mean proportion in the DT-P samples, while the Gram negative *Porphyromonadaceae*, *Rickenellaceae*, *Prevotellaceae* and *Bacteroidacaeae* were elevated in samples treated with Q. Figure 10B compares the paired samples prepared using the Q and DT-N methods respectively. In this instance, the there was no significant difference in the Gram positive families, while the two previously indicated Gram negative families *Coriobacteriaceae* and *Lactobacilliceae* were elevated in the DT-N group. Also elevated were the *Pseudomonadaceae* and *Promicromonosporaceae* families, which are likely to be residual environmental contaminants missed by the retrospective bioinformatic contamination removal. Figure 10C compares the in house method with and without *host depletion + tissue*

dissociation, where the difference was in the Gram negative families, which were elevated in the DT-N samples.

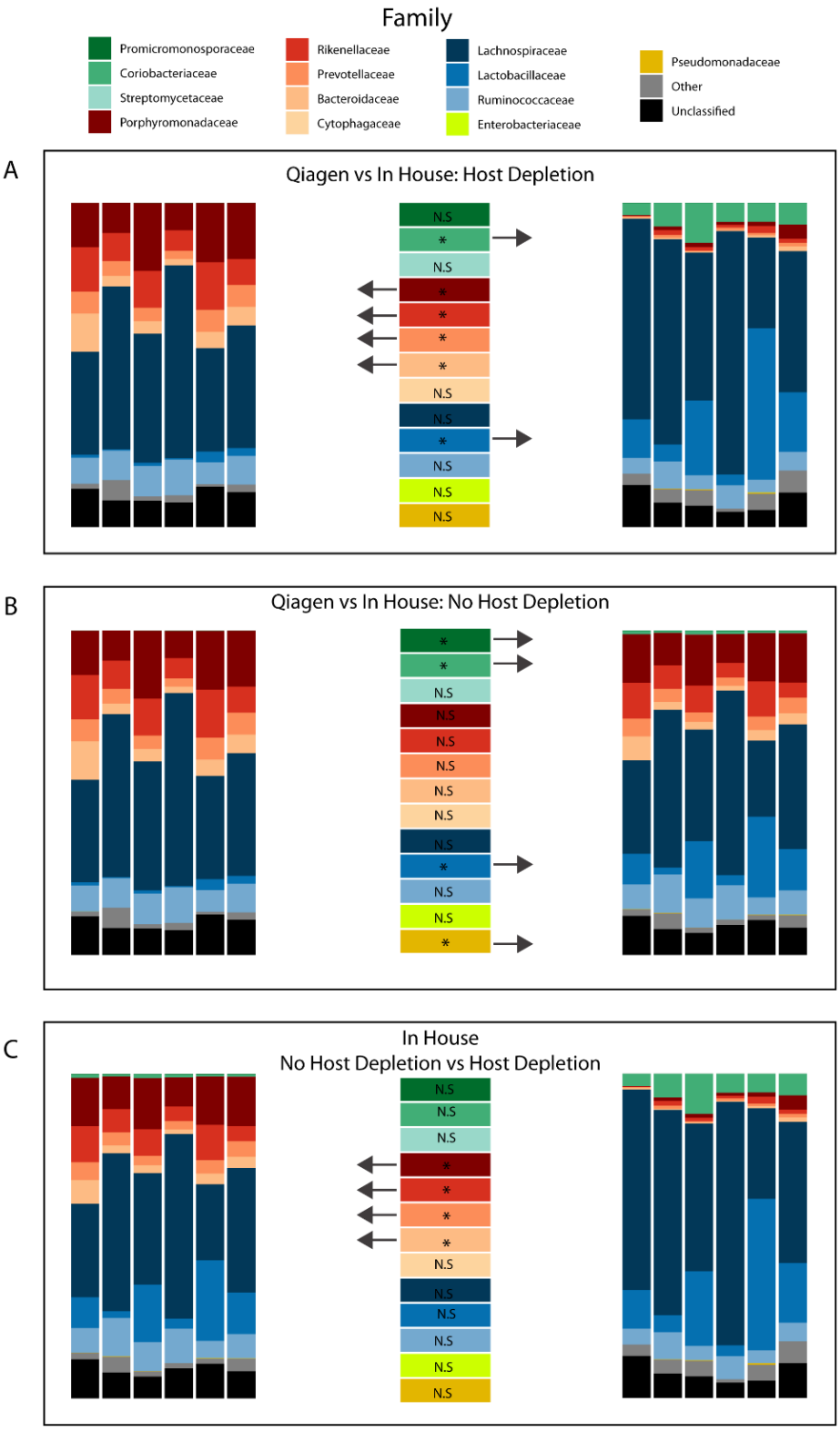


Figure 10: ***Mouse faecal sample composition comparison between methods.*** *Mean abundance of major families between groups tested using Wilcox signed rank test, with false discovery rate controlled for using the FDR method. The arrow indicates*

the direction if increase in cases of significant difference. ***(A)*** *Compares Qiagen with HDP.* ***(B)*** *Compares Qiagen with HDN.* ***(C)*** *Compares HDN with HDP.*

The final assessment of the method was the analysis of FFPE malignant breast tissue samples. The accuracy was verified by comparing the FFPE samples with their matched freshly frozen samples. As was suggested by the representative pie chart of the FFPE breast samples in Figure 6B, the quantity of environmental contamination was overwhelming, this was unsurprising given the low level of microbial biomass present in the samples. Even after contamination removal, leaving all other sample types with little to no contamination, the FFPE breast samples in Figure 11B are dominated by the *Pseudomonadaceae* and *Xanthomonadaceae* families seen in the negative control samples and bear little resemblance to their fresh frozen counterparts in Figure 11C. However, when Figure 11B is recreated in 9D, with all *Pseudomonadaceae* and *Xanthomonadaceae* associated sequences manually removed, there is resemblance between the two groups that begins to justify what the Venn diagram in Figure 11A indicates in terms of shared bacterial families. In total 24.6% of the total bacterial abundance in Figure 11D is accounted for by bacterial families also found in the fresh samples shown in Figure 11B.

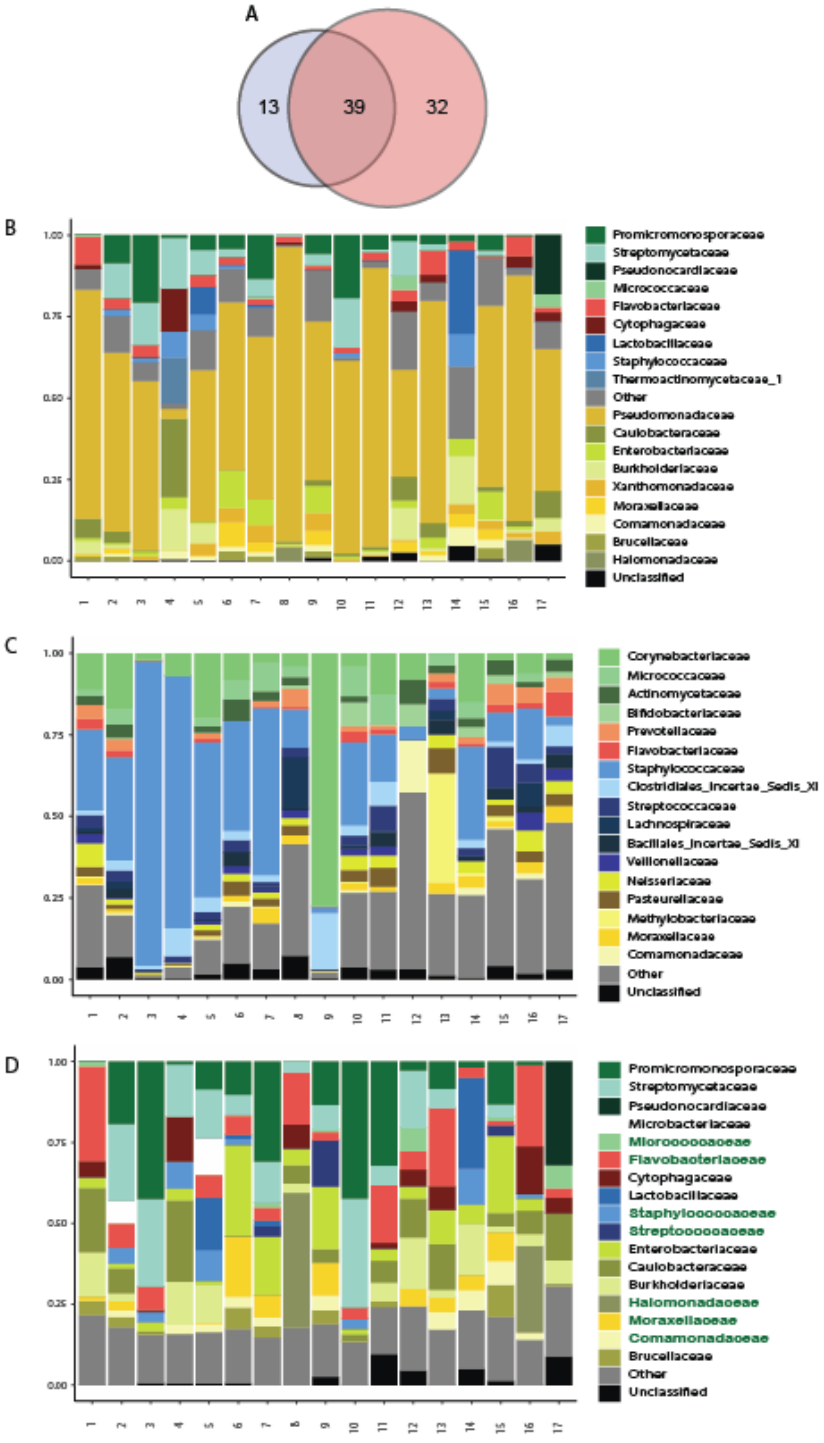


Figure 11: ***Sample composition comparison between matched patient samples. (A)*** *Venn diagram visualising the observed families in* ***(B)*** *FFPE breast tissue and* ***(C)*** *Matched fresh frozen breast samples processed using Molzym Ultra-Deep Microbiome kit.* ***(D)*** *FFPE samples with the two obvious contaminant families, Xanthomonadaceae and Pseudomonadaceae manually removed.*

DISCUSSION

Given the potentials that FFPE material could bring to the field of metagenomics, a method that allows access to this material is essential. Currently, there are no methods available to process this sample type for metagenomics. This research presents novel strategies to treat these samples in order to guarantee a truthful representation of the bacterial communities inhabiting tissues.

Protoblock

It is well-established that DNA from FFPE samples is damaged [33]. The extent of this damage can increase with the length of exposure to formalin, sample storage time and the pH of the formalin used for fixation (34,35). DNA damage found in FFPE samples spans from cross-links, fragmentation, loss of bases and point mutations [22, 33, 36, 37]. With this in mind, an adequate control for this sample type must undergo the same extent of damage for it to be representative of the samples being treated (38).

All Protoblock DNA sequences analysed by HRM exhibited a profile indicative of the presence of sequence artefacts, with aberrant profiles typical of heteroduplexes with deviations in Tm of up to 0.1^{o}C from that of non-fixed template. This correlates with previous studies where FFPE DNA displayed aberrant melting profiles which is indicative of low-level, non-identical changes randomly distributed in the template [39]. Given the random distribution of these artefacts, their presence in undetectable by Sanger sequencing as they are masked by the abundant correct sequences. However, the reduced number and lengths of fragments that can be sequenced are indicative of alterations that lead to sequencing failure [40]. The presence of “true mutations” (identical nucleotide changes at an exact position in >2 independent sequences) was investigated here by WGS. As seen in figure 3C, the majority of variants detected corresponded to C>T and G>A transitions. This again, is in line with findings in human tissue FFPE samples and in accordance with HRM profiles with low-level, heteroduplex sequence changes [39, 40].

The impact of bacterial lysis strategies in sample prep has been documented, but not fully characterised as a source of introduced bias. The use of standards has been suggested as a tool to account for this and other storage or sample prep bias [41, 42]. Use of the Protoblock effectively detects this bias and therefore is useful as a

standard for bacterial prep of FFPE samples in order to ensure the accuracy of a metagenomic project, especially in determining abundance. Findings shown in Figure 4B demonstrate that the QIAGEN QIAamp FFPE tissue kit (currently the gold standard for FFPE DNA isolation, developed for mammalian FFPE DNA) is biased towards Gram-negative bacteria. This is not surprising, since the method does not include a bacterial lysis step. Given the lack of a standardised method to process FFPE samples for metagenomic studies, the use of standards such as the Protoblock is essential to guarantee the accuracy, precision, and limit of detection of the sample processing workflow.

Contamination is a considerable threat to the accuracy of low biomass samples such as biopsies, even prior to any formalin fixing or eventual pre-processing of these fixed samples for high throughput sequencing. Steps such as deparaffination and DNA repair require the use of enzymatic solutions that are difficult to keep sterile, and contamination from these sources could easily obscure the true results in cases of low microbial load. Use of the Protoblock can inform users as to the level of contamination introduced by any processing of FFPE samples required in advance of a sequencing experiment.

DNA Repair

While the HRM melting curve analysis provided a valuable guide, confirmation was provided by qPCR and sequencing data. This shows significant improvements in the integrity, readability and sequence quality after treatment with the reconstituted BER reactions, thus confirming its efficacy. In this sample-type, it is evident that targeting DNA damage derived from oxidation with FPG and Endo VIII, within the enzyme mixes tested here, yields the most significant improvements. After exhaustive comparisons of different approaches to the problem, the strategy found to be most effective involves decrosslinking using a chaotrophic agent such as Guanidine hydrochloride, and a slightly reduced temperature of 80 ^{o}C. This is followed by removal of damaged bases by using a combination of Formamidopyrimidine DNA glycosylase and Endonuclease VIII, and repair through the short-patch BER repair pathway (triggered by both of this glycosylases) by combining T4-PNK, DNA polymerase and DNA ligase.

DNA extraction protocol

Most, if not all host depletion strategies report some off target effects on bacteria [43]. To fully explore the effects that this would have on downstream sequence analysis, paired protoblock samples treated with (HDP) and without (HDN) host depletion were analysed by 16S sequencing and compared to the gold standard Qiagen QIAamp FFPE tissue kit (Q). The results from this analysis indicate, that while there might be a loss of Gram negative bacteria, this does not significantly affect the outputs of 16S sequencing.

In this analysis the Q method, which is not optimised for bacterial lysis, showed statistically significant deviation from the input proportions across all five bacterial species present in the Protoblock. The HDN method showed improvement on the Q method, with the HDP method being the best performing approach in this instance. This improvement in performance is related to incorporation of a host depletion step, since it is the only variable tested here. It can be hypothesised that this may be due to (1) a reduction of contaminants (as shown in Figure 7) that improves the ratio of bacteria present in the samples being sequenced and (2) the reduction of mammalian DNA positively affects the PCR reaction, by improving the access to target sequences.

This was further explored in murine FFPE faecal samples were exposed or not to a combined treatment with tissue dissociation and host depletion. Based on the evidence from the Protoblock-based comparison of the three methods, the expectation would be for the DT-P (in house with host depletion and tissue dissociation) and DT-N (in house without host depletion and tissue dissociation) to cluster together on a PcOA. However, in this instance, it was the DT-N and Q methods that clustered, showing no statistically significant difference in terms of their Bray-Curtis dissimilarity. Both are significantly different to the samples processed using the DT-P method. Subsequent spearman correlation of the dominant bacterial families identified across the samples with the PC1 and PC2 axis reveals that this separation on the PcOA plot is driven by Gram status. Gram positive bacteria correlate significantly with the direction of the DT-P samples, and Gram negative samples correlate significantly with the two other groups (Figure 9). These findings are corroborated by results in Figure 10. Altogether, these results confirm a significant loss of G- bacteria after the combined treatment with tissue dissociation and host depletion strategies, indicating that Proteinase K debilitates the OM of G-

bacteria, exposing the phospholipid bilayer, which can be then accessed by Saponin, leading to G- bacteria loss. However, this is a necessary step in processing tissues, and thus a further optimisation of this step is necessary. This could be addressed by incorporating a short decrosslinking step that will allow tissue dissociation enzymes to be more effective, leading to a reduction on incubation times or enzyme units used in the reaction. This could lead to less off-target effects in G- bacteria.

By a process of elimination, the best net performing method in this instance appears to be the DT-N method. The DT-P method shows significantly increased Gram positive bacterial family abundance such as *Lactobacillaceae* and *Coriobacteriaceae* when compared with the Qiagen method; conversely the Qiagen method shows significantly more Gram negative bacteria such as *Prevotellaceae* and *Bacteroidaceae*. The DT-N method shows significantly more Gram negative bacterial families vs DT-P (Figure 10C), and significantly more Gram positive families such as *Coriobacteriaceae* vs Q, with no families significantly reduced in abundance vs either group. confirmingThis ed that the tissue dissociation strategy needs to be optimised.

Despite major efforts on maintaining an aseptic technique, there are still numerous potential sources of contamination, ranging from the wax used to embed samples, through all the DNA purification solutions and enzymes, which are unsuitable for sterilisation or could not be gamma irradiated at our facilities. Thus, it is unsurprising that there was a considerable amount of contamination present in the samples. The biomass in the Protoblock and murine faecal samples is sufficient to ensure that the majority of the reads are of sample origin according to the SourceTracker algorithm, but the FFPE breast samples appeared to consist almost entirely of bacterial reads attributed to one or more of the negative controls. The SourceTracker output in Figure 6B indicates that all contamination is attributable to three negative control samples, namely the Wax control, taken from the edges of the blocks of patient samples, the “In House method” negative control, and the non-bacterial control, which is an empty Protoblock FFPE processed at our facilities. The first two negative controls were dominated by the genera *Stenotrophomonas, Pseudomonas* and *Clostridium*, all of which count among the most abundant genera in the dataset. The presence of both high and low abundance environmental contaminants presents a problem for most bioinformatic contamination removal

methods, and highlights the value of using both positive and negative controls to assist in contamination removal [44]. In this instance, we are provided with a much clearer picture of the contamination induced during the process by the use of the Protoblock in conjunction with negative controls. This allows us to conclude that in the case of the Protoblocks and the mouse faecal samples, any contamination introduced by the method can be controlled to below any level where it would affect downstream analysis. Figure 6 also provided us with evidence of a phenomenon that is gaining more attention in microbiome research, cross contamination, which originates within the pool of samples(45). This phenomenon is known to affect lower biomass samples, and can be clearly seen in the non bacterial control where five of the common bacterial families across the dataset also appear in the negative controls. This is particularly dangerous when undertaking established, but conservative contamination removal by subtraction approaches.

Non gastro-intestinal tract biopsies are notoriously low in microbial biomass (46), a fact that is further compounded in analysis of FFPE biopsies by the fact that the formalin fixation process accounts for a log fold reduction in the quantity of recoverable DNA (47). These challenges clearly manifest in the comparison of paired fresh and FFPE breast samples. Once the major contaminant ASV's and those suspected of aligning to the human genome are removed, the FFPE breast samples are still dominated by known contaminant families, seen in the negative controls in Figure 6. Encouragingly, there are some common families to both the FFPE breast samples shown in Figure 11B and the fresh frozen breast sample shown in Figure 11C. As mentioned in the results, manual removal of the *Pseudomonadaceae* and *Xanthomonadaceae* families reveals a sample composition plot where 24.6% of the total bacterial abundance in FFPE breast tissue is accounted for by the bacterial families also present in the fresh frozen breast samples (Figure 11A).

The reason for Figure 11D is that it is a crude retrospective imitation of a potential improvement to make this method a viable option for low biomass FFPE studies. With the main contaminants inherent to the In House FFPE protocol now identified, these can be biologically removed from the sample by blocking their amplification from the 16S PCR pool. Numerous methods have been developed to achieve an asymmetric PCR reaction that will favour the amplification of certain target regions and avoid the amplification of other, which have been used extensively for SNP

detection or to reduce off-target capture during sequencing library enrichment. This is achieved by: (1) Blocking extension with DNA probe/oligo that has high affinity towards a specific DNA sequence (on either DNA strand) that includes a 3' end (i.e. phosphate, inverted dNTP). (2) Inhibiting primer annealing with a homologous peptide nucleic acid (PMA) or locked nucleic acids (LNAs), which have increased thermal or base stacking stability, respectively and will inhibit PCR [48-50].

Conclusion

Strategies for the unbiased treatment of FFPE samples for metagenomic analysis are presented in this work validated by a variety of approaches on mock bacterial communities, murine models and human breast tissue samples. The results shown here confirm that most of these strategies would have a positive effect in the treatment for metagenomics. However, key areas that need to be addressed are the optimisation of a tissue dissociation strategy that does not lead to G- bacterial loss and the biological decontamination of samples previous to the analysis.

References:

1. Riquelme, E., et al., *Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes.* Cell, 2019. **178**(4): p. 795-806.e12.
2. Stewart, C.J., et al., *Using formalin fixed paraffin embedded tissue to characterize the preterm gut microbiota in necrotising enterocolitis and spontaneous isolated perforation using marginal and diseased tissue.* BMC Microbiology, 2019. **19**(1): p. 52.
3. Racsa, L.D., et al., *Identification of bacterial pathogens from formalin-fixed, paraffin-embedded tissues by using 16S sequencing: retrospective correlation of results to clinicians' responses.* Human Pathology, 2017. **59**: p. 132-138.
4. Emery, D.C., et al., *16S rRNA Next Generation Sequencing Analysis Shows Bacteria in Alzheimer's Post-Mortem Brain.* Frontiers in Aging Neuroscience, 2017. **9**(195).
5. Hiskey, M., et al., *Analysis of Molecular Identification of Bacterial Pathogens on Formalin-Fixed Paraffin-Embedded Tissue.* American Journal of Clinical Pathology, 2015. **144**(suppl_2): p. A224-A224.
6. Banerjee, S., et al., *Distinct Microbial Signatures Associated With Different Breast Cancer Types.* Frontiers in Microbiology, 2018. **9**(951).
7. Clooney, A.G., et al., *Comparing Apples and Oranges?: Next Generation Sequencing and Its Impact on Microbiome Analysis.* PLoS One, 2016. **11**(2): p. e0148028.
8. Salter, S.J., et al., *Reagent and laboratory contamination can critically impact sequence-based microbiome analyses.* BMC Biology, 2014. **12**(1): p. 87.
9. Chen, L., et al., *DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification.* Science, 2017. **355**(6326): p. 752.
10. Marotz, C.A., et al., *Improving saliva shotgun metagenomics by chemical host DNA depletion.* Microbiome, 2018. **6**(1): p. 42.
11. 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, 2019. **10**(1277).
12. Fricker, A.M., D. Podlesny, and W.F. Fricke, *What is new and relevant for sequencing-based microbiome research? A mini-review.* Journal of Advanced Research, 2019. **19**: p. 105-112.
13. Wu, J.Y., et al., *Effects of polymerase, template dilution and cycle number on PCR based 16 S rRNA diversity analysis using the deep sequencing method.* BMC Microbiol, 2010. **10**: p. 255.
14. Teng, F., et al., *Impact of DNA extraction method and targeted 16S-rRNA hypervariable region on oral microbiota profiling.* Scientific Reports, 2018. **8**(1): p. 16321.
15. Abusleme, L., et al., *Influence of DNA extraction on oral microbial profiles obtained via 16S rRNA gene sequencing.* Journal of Oral Microbiology, 2014. **6**(1): p. 23990.
16. Fiedorová, K., et al., *The Impact of DNA Extraction Methods on Stool Bacterial and Fungal Microbiota Community Recovery.* Frontiers in microbiology, 2019. **10**: p. 821-821.
17. Knudsen, B.E., et al., *Impact of Sample Type and DNA Isolation Procedure on Genomic Inference of Microbiome Composition.* mSystems, 2016. **1**(5): p. e00095-16.
18. Methé, B.A., et al., *A framework for human microbiome research.* Nature, 2012. **486**(7402): p. 215-221.
19. Gill, C., et al., *Evaluation of Lysis Methods for the Extraction of Bacterial DNA for Analysis of the Vaginal Microbiota.* PLOS ONE, 2016. **11**(9): p. e0163148.
20. Bag, S., et al., *An Improved Method for High Quality Metagenomics DNA Extraction from Human and Environmental Samples.* Scientific Reports, 2016. **6**(1): p. 26775.

21. Einaga, N., et al., *Assessment of the quality of DNA from various formalin-fixed paraffin-embedded (FFPE) tissues and the use of this DNA for next-generation sequencing (NGS) with no artifactual mutation.* PloS one, 2017. **12**(5): p. e0176280-e0176280.
22. Zhang, P. and B.D. Lehmann, *The Utilization of Formalin Fixed-Paraffin-Embedded Specimens in High Throughput Genomic Studies.* 2017. **2017**: p. 1926304.
23. Yuan, S., et al., *Evaluation of Methods for the Extraction and Purification of DNA from the Human Microbiome.* PLOS ONE, 2012. **7**(3): p. e33865.
24. Liesack, W., H. Weyland, and E. Stackebrandt, *Potential risks of gene amplification by PCR as determined by 16S rDNA analysis of a mixed-culture of strict barophilic bacteria.* Microbial Ecology, 1991. **21**(1): p. 191-198.
25. Bøifot, K.O., et al., *Performance evaluation of a new custom, multi-component DNA isolation method optimized for use in shotgun metagenomic sequencing-based aerosol microbiome research.* bioRxiv, 2019: p. 744334.
26. Tighe, S., et al., *Genomic Methods and Microbiological Technologies for Profiling Novel and Extreme Environments for the Extreme Microbiome Project (XMP).* Journal of biomolecular techniques : JBT, 2017. **28**(1): p. 31-39.
27. Zaikova, E., et al., *Antarctic Relic Microbial Mat Community Revealed by Metagenomics and Metatranscriptomics.* Frontiers in Ecology and Evolution, 2019. **7**(1).
28. Venter, J.C., et al., *Environmental genome shotgun sequencing of the Sargasso Sea.* Science, 2004. **304**(5667): p. 66-74.
29. Costea, P.I., et al., *Enterotypes in the landscape of gut microbial community composition.* Nature Microbiology, 2018. **3**(1): p. 8-16.
30. Proctor, L., *Priorities for the next 10 years of human microbiome research.* Nature, 2019. **569**(7758): p. 623-625.
31. Costea, P.I., et al., *Towards standards for human fecal sample processing in metagenomic studies.* Nature Biotechnology, 2017. **35**: p. 1069.
32. Katsnelson, A., *Standards Seekers Put the Human Microbiome in Their Sights.* ACS Cent Sci, 2019. **5**(6): p. 929-932.
33. Robbe, P., et al., *Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project.* Genetics in Medicine, 2018. **20**(10): p. 1196-1205.
34. Hughes, S. and J. Lau, *A technique for fast and accurate measurement of hand volumes using Archimedes' principle.* Australasian Physics & Engineering Sciences in Medicine, 2008. **31**(1): p. 56.
35. Do, H. and A. Dobrovic, *Sequence Artifacts in DNA from Formalin-Fixed Tissues: Causes and Strategies for Minimization.* Clinical Chemistry, 2015. **61**(1): p. 64.
36. Hosein, A.N., et al., *Evaluating the repair of DNA derived from formalin-fixed paraffin-embedded tissues prior to genomic profiling by SNP-CGH analysis.* Lab Invest, 2013. **93**(6): p. 701-10.
37. Williams, C., et al., *A high frequency of sequence alterations is due to formalin fixation of archival specimens.* Am J Pathol, 1999. **155**(5): p. 1467-71.
38. Choo, J.M., L.E.X. Leong, and G.B. Rogers, *Sample storage conditions significantly influence faecal microbiome profiles.* Scientific Reports, 2015. **5**: p. 16350.
39. Do, H. and A. Dobrovic, *Limited copy number-high resolution melting (LCN-HRM) enables the detection and identification by sequencing of low level mutations in cancer biopsies.* Mol Cancer, 2009. **8**: p. 82.
40. Do, H. and A. Dobrovic, *Dramatic reduction of sequence artefacts from DNA isolated from formalin-fixed cancer biopsies by treatment with uracil- DNA glycosylase.* Oncotarget, 2012. **3**(5): p. 546-58.
41. Goldberg, B., et al., *Making the Leap from Research Laboratory to Clinic: Challenges and Opportunities for Next-Generation Sequencing in Infectious Disease Diagnostics.* mBio, 2015. **6**(6): p. e01888-15.

42. Hornung, B.V.H., R.D. Zwittink, and E.J. Kuijper, *Issues and current standards of controls in microbiome research.* FEMS Microbiology Ecology, 2019. **95**(5).
43. Marotz, C.A., et al., *Improving saliva shotgun metagenomics by chemical host DNA depletion.* Microbiome, 2018. **6**(1): p. 42-42.
44. Eisenhofer, R., et al., *Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations.* Trends Microbiol, 2019. **27**(2): p. 105-117.
45. Minich, J.J., et al., *Quantifying and Understanding Well-to-Well Contamination in Microbiome Research.* mSystems, 2019. **4**(4): p. e00186-19.
46. 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, 2015. **3**: p. 19-19.
47. Hykin, S.M., K. Bi, and J.A. McGuire, *Fixing Formalin: A Method to Recover Genomic-Scale DNA Sequence Data from Formalin-Fixed Museum Specimens Using High-Throughput Sequencing.* PloS one, 2015. **10**(10): p. e0141579-e0141579.
48. Vestheim, H., B.E. Deagle, and S.N. Jarman, *Application of blocking oligonucleotides to improve signal-to-noise ratio in a PCR.* Methods Mol Biol, 2011. **687**: p. 265-74.
49. Bender, M., et al., *Use of a PNA probe to block DNA-mediated PCR product formation in prokaryotic RT-PCR.* Biotechniques, 2007. **42**(5): p. 609-10, 612-4.
50. Wang, H., et al., *Allele-specific, non-extendable primer blocker PCR (AS-NEPB-PCR) for DNA mutation detection in cancer.* J Mol Diagn, 2013. **15**(1): p. 62-9.

Chapter V

Microbiome analysis as a platform R&D tool for parasitic nematode disease management

This chapter has been published as:

"Microbiome analysis as a platform R&D tool for parasitic nematode disease management"

Glenn Hogan*, **Sidney Walker***, Frank Turnbull, Tania Curiao, Alison A. Morrison, Yensi Flores, Leigh Andrews[e], Marcus J. Claesson, Mark Tangney**, Dave J. Bartley** *ISME J* **13**, 2664-2680 (2019). (Impact Factor 9.43).

* Contributed equally; ** Contributed equally

ABSTRACT

Background The relationship between bacterial communities and their host is being extensively investigated for the potential to improve the host's health. Little is known about the interplay between the microbiota of parasites and the health of the infected host.

Aim Using nematode co-infection of lambs as a proof-of-concept model, the aim of this study was to characterise the microbiomes of nematodes and that of their host, enabling identification of candidate nematode-specific microbiota member(s) that could be exploited as drug development tools or for targeted therapy.

Methods Deep sequencing techniques were used to elucidate the microbiomes of different life stages of two parasitic nematodes of ruminants, *Haemonchus contortus* and *Teladorsagia circumcincta*, as well as that of the co-infected ovine hosts, pre- and post-infection.

Conclusions Bioinformatic analyses demonstrated significant differences between the composition of the nematode and ovine microbiomes. The two nematode species also differed significantly. Data indicated a shift in the constitution of the larval nematode microbiome after exposure to the ovine microbiome, and in the ovine intestinal microbial community over time as a result of helminth co-infection. Several bacterial species were identified in nematodes that were absent from their surrounding abomasal environment, the most significant of which included *Escherichia coli/Shigella*. The ability to purposefully infect nematode species with engineered *E. coli* was demonstrated *in vitro*, validating the concept of using this bacterium as a nematode-specific drug development tool and/or drug delivery vehicle.

To our knowledge, this is the first description of the concept of exploiting a parasite's microbiome for drug development and treatment purposes.

INTRODUCTION

Nematode infection is of major concern to human health in middle and low-income countries, particularly in cases of foodborne disease (1). Additionally, animals infected by pathogenic nematodes are a serious health, welfare and economic burden for countries reliant on agriculture (2). Effective interventions are therefore necessary to promote human health, protect livestock, and ensure production efficiency. Current standard practices for eradicating helminthic disease focus on the routine and frequent administration of anthelmintics, small-molecule drugs, to infected hosts. However, as with many chemicals, the development of resistance means that these drugs' effectiveness is reducing (3), and alternative treatments are of paramount importance (4). Large numbers of new chemical drug classes are unlikely to be synthesised and licensed to combat growing drug resistance in nematodes in the near future, given the large time commitment required for drug research and development (5). Admittedly, a small number of compounds are at the early stage of investigation for controlling human whipworm infections (6,7). Yet, contingency strategies and tools to help expedite drug development are still desirable.

In parasitic disease, attempts have been made to characterise the interplay between helminths and the bacterial populations inhabiting the mammalian gut, elucidating the ways in which the activity of the parasite affects the constituency of the gut microbiota and vice versa (8-10). These studies have suggested that the co-evolution of these two communities has established a relationship wherein the survival of either population is impacted by the other. Susceptibility and resistance to helminth infection in humans have been linked with certain bacterial taxa, suggesting that there may exist an ideal host microbial profile that guards against such disease (11). In fact, it has recently been discovered that parasites themselves have a microbiome. The nematode microbiome has become an increasingly popular area of study and has seen considerable advancement over the past two years due to 16S rRNA gene sequencing accessibility: the microbiomes of *Caenorhabditis elegans* (12), the ruminant parasite *Haemonchus contortus* (13), the murine parasite *Trichuris muris* (9), soil and beetle-associated nematodes (14), the marine nematode *Litoditis marina* (15) and various other marine nematodes (16) have all been sequenced.

High-throughput technologies are ideally placed to examine the interplay between the microbial communities within nematodes and the microbial communities of the animals they infect. However, while big data have been utilised to expand our understanding of the nematode microbiome, less consideration has been given to how this information might be applied to the therapeutic benefit of parasite-infected organisms. Defining the microbial communities of nematodes and their host opens opportunities for exploiting differences for drug development and/or treatment purposes. Identifying bacterial communities that uniquely colonise the nematode presents an opportunity to investigate their use as oral agents that specifically target the parasite, leaving the host unaffected.

Exploitation of the host microbiota as a means of treating disease in the host is well studied across multiple species – from the use of faecal microbiota transplantation for inducing remission in ulcerative colitis in humans (17) to the treatment of laminitis in horses (18); however, exploitation of the parasite microbiome as an aid to drug development and treatment has not yet been described. We hypothesised that: i) nematode co-infection of the host would significantly alter the host microbiome over time; ii) the host microbiome would significantly alter the microbiome of the nematodes; and iii) despite interactions between host and parasite microbiota, key differences between the two would be apparent that would welcome their further investigation as aids to drug development and treatment.

In this study, the microbiomes of the ovine abomasum and intestines were characterised following co-infection of lambs with the pathogenic nematodes *H. contortus* and *Teladorsagia circumcincta*. The abomasum is one of four compartments of the ruminant stomach, in which *H. contortus* and *T. circumcincta* live (19), and of the four compartments bears the closest resemblance to the anatomy and functionality of the simple stomach of non-ruminants (20). The microbiomes of both nematodes were also characterised at both the infective larval (L_3) and adult stages of their development, marking this as the first report of the *T. circumcincta* microbiome and the first comparative study where different nematode genera are derived from the same host. The ovine model chosen is appropriate for a proof-of-concept study, and the blood-feeding parasite *H. contortus* is a good model system for blood-feeding nematodes. This study also offers insights into the effects of parasites on the host, and vice versa. The effects on the host are quantified by

monitoring changes in the ovine microbiome over the 28 days of parasitic co-infection. Effects on the parasite are examined by comparing the microbiomes of pre- and post-infection nematode larvae.

MATERIALS AND METHODS

All laboratory work was performed by other members of the Tangney lab.

Ovine and parasite samples were collected at various timepoints over a 28-day infection (Supplementary Figure 1).

Parasite material – adult nematodes

Four lambs were artificially co-infected with 15,000 infective larvae (L_3; 5000 *H. contortus* and 10,000 *T. circumcincta*). 28 days post-infection (i.e. at the point of culling), adult worms were collected from the abomasa of each lamb (21). The nematodes were sexed, staged, and species-identified using criteria described in the Ministry of Agriculture, Fisheries and Food document (22)[273][272][272][272][22]. Separate pools of 100 adult male and 100 adult female worms were species-identified, washed twice in sterile phosphate-buffered saline (PBS) to remove surface-adherent bacteria, snap frozen in liquid nitrogen, and transferred to -80 °C storage prior to deoxyribonucleic acid (DNA) extraction. Both worm species were processed separately.

Parasite material – pre-infection and post-infection larvae

To provide an indication of the microbial diversity present within the L_3 population that were used to generate the adult material, sub-samples of ~10,000 infective larvae used in the artificial challenge doses were snap frozen in liquid nitrogen on the day of challenge and stored -80°C storage prior to DNA extraction. Faecal material containing eggs (both *H. contortus* and *T. circumcincta*) from the patent parasite infections were collected from the infected donor lambs at post mortem (d28) and incubated at 22°C for 14 days. Infective larvae derived from the d28 faeces were extracted, enumerated and identified to species level, snap frozen in liquid nitrogen and stored at -80°C in pools of ~ 10,000 larvae. Figure 1 shows the nematode lifecycle, and its association with the ruminant digestive system.

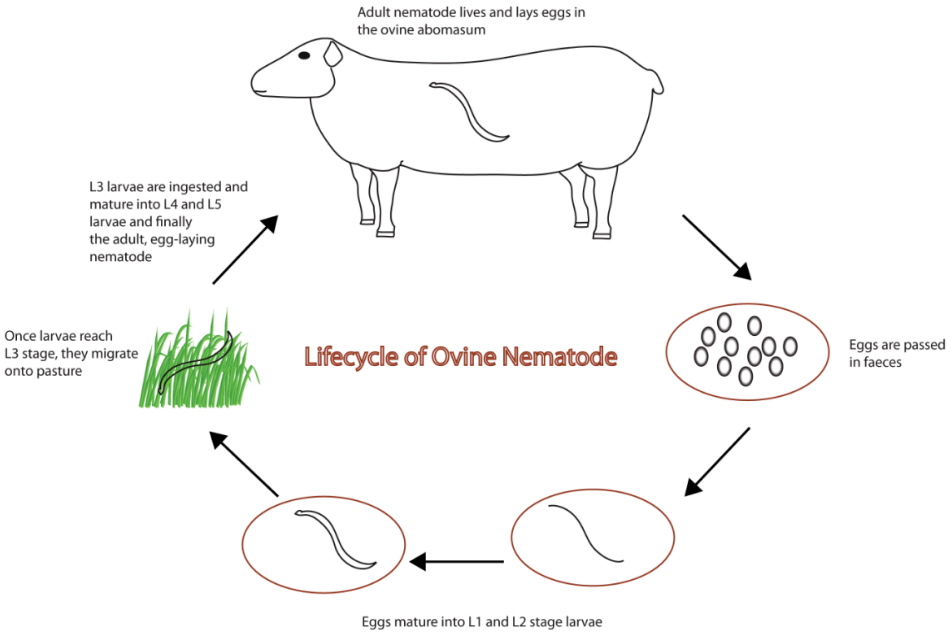


Figure 1: ***The nematode life cycle and its association with the ruminant digestive system.***

Ovine faecal and abomasal sample collection

Individual faecal samples were collected *per rectum* at days 0, 1, 2, 5, 7, 9, 14, 19, 21, and 28 post infection from all donor animals. Faecal samples were transferred to -80°C storage prior to DNA extraction. Sub-samples of abomasal contents were collected at *post-mortem* from each lamb donor.

Confirmation of bacterial presence within nematodes

To validate the presence of bacteria within ovine nematodes, wax sections from *H. contortus* adult worms were Gram-stained following standard procedures (23).

Genomic DNA extraction

The adult worms were transferred to 2 ml Lysing Matrix B tubes (MP Biomedicals) and were re-suspended in 500 µl sterile phosphate buffered saline (PBS). The larvae

were homongenised using a Precellys24 homogeniser (Bertin Technologies) at 6000 rpm for 30 sec for three cycles. The DNA extraction was conducted using the DNeasy Blood and Tissue Kit (Qiagen). To homogenate tubes, 500 µl ATL buffer supplemented with 12 mAU proteinase K (Promega) was added, followed by incubation at 56 °C for 2 h. To pellet the 0.1 mm glass beads, the Lysing Matrix B tubes were centrifuged at 15,000 x *g* for 5 min. The supernatant was transferred to a clean 2 ml microcentrifuge tube and this step was repeated to ensure no glass beads were transferred to the DNeasy Mini spin columns. The DNeasy Blood and Tissue Kit guidelines for Animal Tissues (Spin-Column Protocol) were followed, eluting the DNA in 100 µl of Buffer AE before DNA quantification using a NanoDrop ND1000 UV-Vis spectrophotometer (NanoDrop Technologies) and the tubes were stored at -80 °C.

Controls

Negative control tubes were included to account for environmental contaminants present throughout the processing of the samples. These consisted of 1 ml PBS that was exposed to the equipment used during the *post-mortem*, lab environment, DNeasy Blood and Tissue Kits (Qiagen), and Lysing Matrix B tubes (MP Biomedicals) as well as a DNA extraction conducted on the diluent Ultrapure water.

V3-V4 16S rRNA gene sequencing: PCR amplification

Genomic DNA was amplified using 16S rRNA gene amplicon polymerase chain reaction (PCR) primers targeting the hypervariable V3-V4 region of the 16S rRNA gene: V3-V4 forward, 5'TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGGGNGGCWG CAG3'; and V3-V4 reverse, 5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACTACHVGGGTAT CTAATCC3' (Illumina 16S Metagenomic Sequencing Protocol, Illumina, CA, USA). A 35-µl PCR was performed for each sample per the following recipe: 3.5 µl template DNA, 17.5 µl KAPA HiFi HotStart ReadyMix (Roche), 0.7 µl of both primers (initial concentration, 10 pmol/µl), 0.1 µg/µl bovine serum albumin fraction V (Sigma), and 8 µl 10 mM Tris-Cl (Qiagen). Thermal cycling was completed in an

Eppendorf Mastercycler per the directions in the 'Amplicon PCR' section of the '16S Metagenomic Sequencing Library Preparation' protocol (Illumina). Amplification was confirmed by running 5 µl of PCR product on a 1.5% agarose gel at 70 volts for 80 min, followed by imaging on a Gel Doc EZ System (Bio-Rad). The product was approximately 450 base pairs (bp) in size.

PCR-positive products were cleaned per the 'PCR Clean-Up' section of the Illumina protocol, with the exception that drying times were reduced to half the prescribed duration to account for the additional drying that occurs in a laminar airflow hood. Sequencing libraries were then prepared using the Nextera XT Index Kit (Illumina) and cleaned per the Illumina protocol. Libraries were quantified using a Qubit fluorometer (Invitrogen) using the 'High Sensitivity' assay. Sample processing was subsequently completed at Macrogen Inc., Seoul, South Korea. Samples were normalised, pooled, and underwent a paired-end 450 bp run on the Illumina MiSeq platform.

Bioinformatics analyses

The quality of the paired-end sequence data was initially visualised using FastQC v0.11.6, and then filtered and trimmed using Trimmomatic v0.36 to ensure a minimum average quality of 25. The remaining high-quality reads were then imported into the R environment v3.4.4 for analysis with the DADA2 package v1.8.0. After further quality filtering, error correction and chimera removal, the raw reads generated by the sequencing process were refined into a table of Amplicon Sequence Variants (ASVs) and their distribution among the samples. It is recommended that ASVs (formerly called 'Ribosomal Sequence Variants') are used in place of 'operational taxonomic units' (OTU), in part because ASVs give better resolution than OTUs, which are clustered based on similarity (24). ASVs were then exported back into Linux and a second stage of chimera removal was carried out using USEARCH v9 in conjunction with the ChimeraSlayer Gold database v6. The remaining ASVs were screened for contamination using the Decontam package in R v1.0.0. The ASVs were classified at genus level using the classify.seqs function in Mothur. Additional species-level classification was performed using SPINGO.

The following statistical analyses were carried out in R: Shannon alpha diversity and Chao1 species richness metrics, and Bray-Curtis distances, for analysis of beta diversity, were calculated using the PhyloSeq package v1.24, and the Vegan package v2.52. Beta diversity calculations produce distance matrices with as many columns and rows as there are samples; thus, beta diversity is often represented using some form of dimensionality reduction, in this case, using principal co-ordinates analysis (PCoA) with the Ape package v5.1. Hierarchical clustering, an unsupervised method that can reveal key taxa that distinguish their respective environments, was performed with the heat plot function in the made4 package v1.54. Differential abundance analysis was carried out using Deseq2 v1.2.0, which identifies differentially abundant features between two groups within the data (25). Tests of means were performed using the Mann-Whitney *U* test unless otherwise stated, and correlations were calculated using Spearman's rank correlation coefficient. Where applicable, false positive rates were controlled below 5% using the Bonferroni procedure.

The SourceTracker algorithm was implemented to ensure that any differences between pre- and post-infection nematode larvae were not due to the adherence of gut bacteria to the surface of the latter group, following their exposure to the ovine intestinal tract. The 15 larval nematode samples were treated as 'sink' samples and compared with five 'source' samples to investigate the level of contamination present, if any. SourceTracker v1.0 was implemented in the R environment.

Phylogenetic analyses were carried out by downloading genomic data for well-characterised laboratory and pathogenic bacterial strains from the SILVA database and creating multiple sequence alignments with our own relevant ASVs using the MUSCLE alignment tool, hosted by the European Bioinformatics Institute (EBI). The resulting alignment was then exported to PhyML, where a phylogenetic tree was constructed using the maximum likelihood method. Lastly, this tree was exported to the iTOL web server for visualisation.

E. coli larval feeding

Eggs of *H. contortus* MHco3(ISE) were purified and isolated from faecal samples derived from mono-specifically infected donor lambs using a saturated NaCl flotation method. The eggs were washed and re-suspended in water before being added to NGM agar plates supplemented with *E. coli* OP50-1:GFP (pFPV25.1) and incubated at 22°C for 48 h to allow hatching of first-stage larvae and subsequent development to second-stage larvae.

RESULTS

Bacterial presence within nematodes

Figure 2A and 2B show cross sectional images of *H. contortus* gut with Gram-positive bacteria visible throughout.

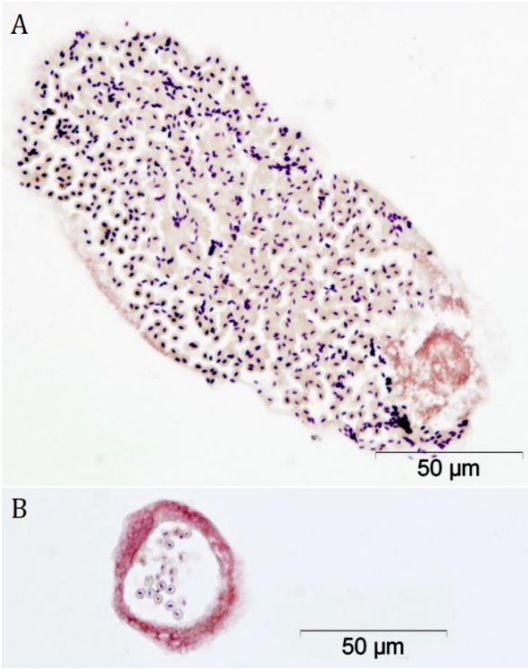


Figure 2: ***Stained sections through gut of an adult H. contortus.*** *Staining shows the presence of Gram-positive bacteria in cross-sections of the intestinal lumen of an adult H. contortus. Gram-positive organisms stain blue-black;~~,~~ Gram-negative organisms and nuclei stain red (images kindly generated and supplied by Jeanie Finlayson, Moredun Research Institute).*

Sample collection and processing

Several samples that proceeded to PCR were not sequenced (Supplementary Figure 2) because either the amplicon PCR failed to amplify the target gene, or the concentration of the sample fell below the 5 ng/µl threshold for sequencing following the second PCR clean-up, indicating either an imperfect DNA extraction or a low abundance of bacteria in these samples. No amplification was evident in the diluent Ultrapure water, nor in the PBS exposed to the *post-mortem* laboratory equipment, laboratory environment, Lysing Matrix B tubes, and run through the DNA extraction kits; however, control samples proceeded to sequencing regardless, as it is now recognised that sequencing of control samples should be standard practice in microbiome work, especially with low-biomass samples, in which low-level contamination may have a large impact on sample readout (26).

Cohort characteristics

Microbiome analysis was carried out on a total of 5,608,303 error-corrected, non-chimeric ASV reads over the entire dataset, with an average read depth of 89,021 reads per sample. This was broken down into a total of 14,351 unique ASVs identified across the four environments studied (Supplementary Figure 3). Of the four environments sequenced, the larval nematode microbiome was the most distinct, with 84.9% of the total ASVs detected belonging uniquely to the larvae, followed by the faecal microbiome with 73.4% unique ASVs. The mature nematode and abomasal microbiomes were considerably less distinct, with 38.2% and 30% unique ASVs, respectively. Six negative control samples were also sequenced: Ultrapure diluent water, lab environment PBS, *post-mortem* suite PBS and PBS run through two DNA extraction kits and lysing matrix tubes. Considerably fewer error-corrected, non-chimeric ASV reads were generated, with an average of 649. Deeper analysis of these samples showed that there was no crossover between ASVs present in the negative controls and experimental samples (Supplementary Figure 4). It was therefore concluded that the biological signal from the experimental samples was not influenced by contamination.

General population structure of the ovine and nematode microbiomes

The microbiomes of the four environments studied were initially classified at phylum level across all individual samples (Figure 3). Their average, grouped composition was as follows: The abomasum contained 49.5% Firmicutes, 36% Bacteroidetes, 2.9% Fibrobacteres, 1.2% Proteobacteria, 1.1% Actinobacteria, 1% Planctomycetes, 1% Candidatus Saccharibacteria, with the remaining fraction comprising either unclassified or negligible proportions. The lamb faecal microbiome contained 67% Firmicutes, 11% Bacteroidetes, 8.5% Candidatus Saccharibacteria, 3.4% Spirochetes, 2.9% Actinobacteria, 1.2% Verrucamicrobia, with the remaining fraction comprising either unclassified or negligible proportions. The larval nematode microbiome contained 67% Proteobacteria, 18% Bacteroidetes, 8% Actinobacteria, 1.6% Planctomycetes, and 1.5% Firmicutes, with the remaining fraction comprising either unclassified or negligible proportions. Finally, the microbiome of the adult nematodes contained 68% Firmicutes, 16% Bacteroidetes, 2.5% Actinobacteria, 2.5% Planctomycetes, 2.2% Candidatus Saccharibacteria, 1.6% Proteobacteria, and 1.1% Verrucomicrobia, with the remaining fraction comprising either unclassified or negligible proportions. The four environments are distinguishable even at phylum level. Nematode larvae have a microbiome dominated by Proteobacteria, a phylum that is not evident in the other environments. The microbiome of the mature nematode more closely resembles the two host sites sampled, suggesting that the host's environment may influence the microbial populations within the parasite. Despite the resemblance of the adult nematode to the faeces and abomasum of the lambs at this taxonomic level, there are still several phyla that are significantly different in terms of their proportions between these environments (Figure 3).

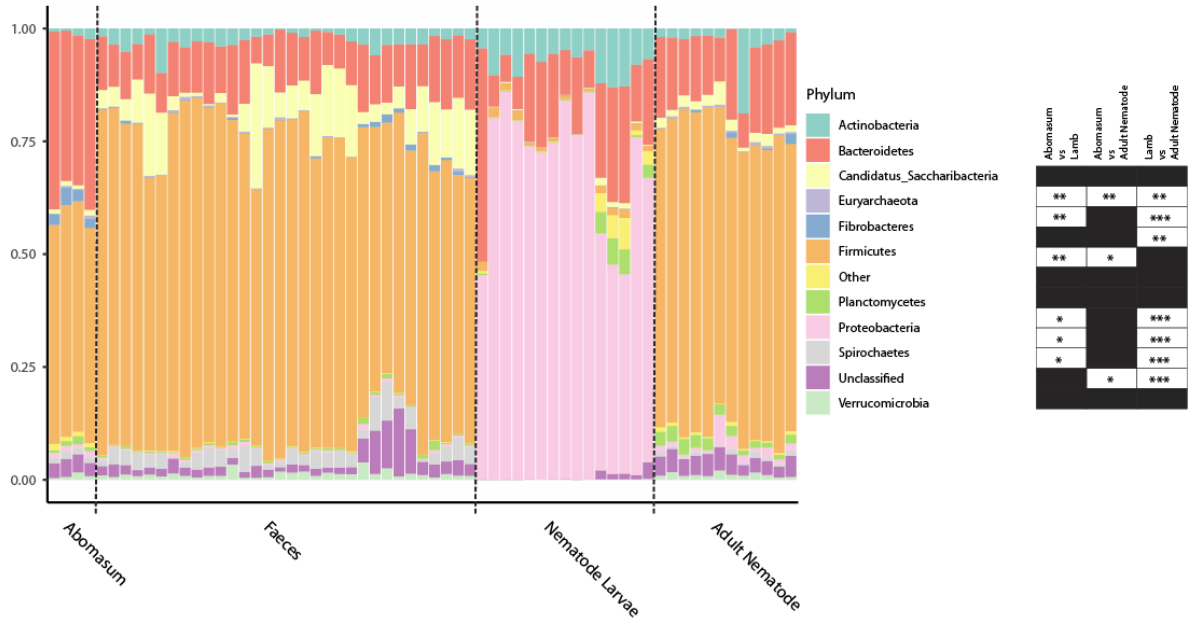


Figure 3: ***Composition at phylum level of the ovine microbiome (abomasal lumen contents and faeces) and nematode microbiome (larval and adult nematodes).*** *Each 'Nematode Larvae' sample contains ~10,000 pooled larvae, 5 of which are pre-infection larvae and 10 of which are post-infection larvae; each 'Adult Nematode' sample contains 100 pooled adult nematodes (five H. contortus (4 males, 1 mixed sex) and seven T. circumcincta (5 females, 1 male, 1 mixed sex) samples); each 'Abomasum' sample is derived from the abomasal washings of one of four lambs; and each 'Faeces' sample is derived from one of four lambs across 10 timepoints. Phyla constituting less than 1% of the total phylum distribution were labelled 'Other'. 'Nematode Larvae' were omitted from statistical testing due to their obvious distinctiveness from the other sample groups. The other three samples were compared for proportions of the different phyla identified - initially with a Kruskal-Wallis test, and then a Mann-Whitney U test, making individual comparisons if warranted. Critical values for significance were adjusted using the Bonferroni method.*

Diversity of the ovine and nematode microbiomes

Alpha diversity, measured using Chao1 species richness showed significant differences between all groups compared, excepting adult nematode and faecal samples, which were similar in terms of species richness (Figure 4). Larvae were the least diverse group, while the abomasum showed the highest diversity. Beta diversity using Bray-Curtis dissimilarity shows three clusters of samples: lamb faecal samples, nematode larvae, and one cluster comprising adult nematodes and lamb abomasa. Hierarchical clustering of the samples based on their composition at ASV level was also performed (Supplementary Figure 5). This was carried out using the Bray-Curtis distance matrix and the Ward-Linkage method. The Ward-Linkage method revealed the same patterns within the data as those observed in the dimensional reduction of the Bray-Curtis dissimilarity matrix, corroborating these findings. Despite apparent similarities at phylum level between the adult nematode and ovine faeces, when individual ASVs are compared, the adult nematode bears the closest resemblance to the ovine abomasum indicating that individual ASVs do not overlap as much as phylum-level annotations between the adult nematode and faeces.

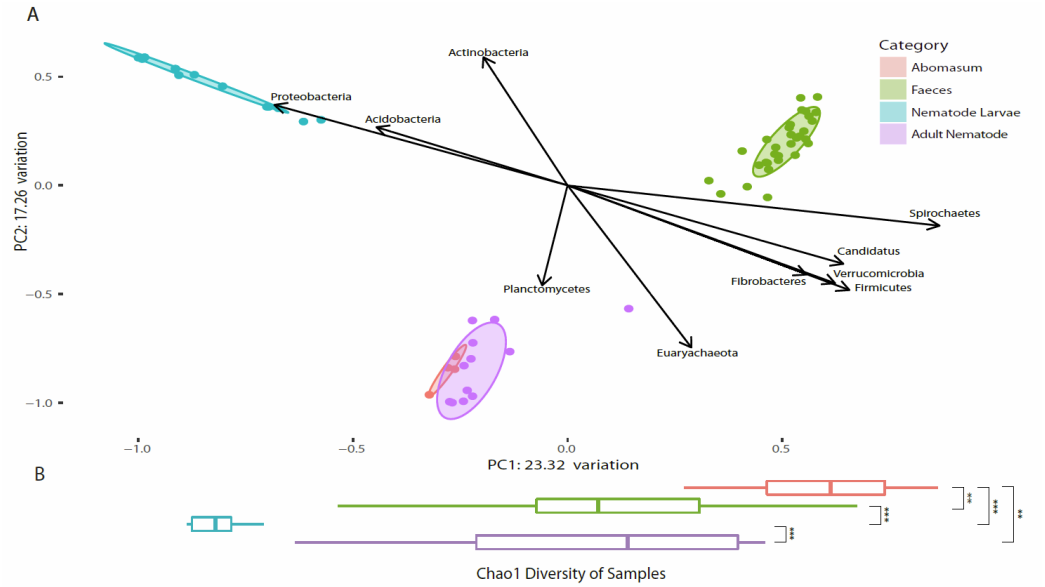


Figure 4: ***Bray-Curtis dissimilarity of the ovine microbiome (abomasal lumen contents and faeces) and nematode microbiome (larval and adult nematodes), correlated with phyla, and Chao1 species richness. (A)*** *Bray-Curtis dissimilarity of microbiomes studied. For the 'Abomasum' samples, each point on the plot is a sample derived from the abomasal washings from one of four lambs, collected 28 days post-infection. For the 'Faeces' samples, each point on the plot is a sample derived from a stool sample collected from one of four lambs from one of ten timepoints over a 28-day infection period. For the 'Nematode Larvae' samples, each point on the plot is a sample derived from a pooled mixture of ~10,000 larvae, and for the 'Adult Nematode' samples, each point on the plot is a sample derived from a pooled mixture of 100 nematodes (five H. contortus (four males, one mixed sex) and seven T. circumcincta (five females, one male, one mixed sex) samples). Ellipses show 80% confidence intervals for their respective groups. Of the 13 different phyla identified, 10 correlate significantly with one or both of the components of the PCoA based on Spearman's rank correlation coefficient. By superimposing this over the PCoA plot, the relationship between these phyla and their environments is visualised.* ***(B)*** *Horizontal alpha diversity boxplots of microbiomes studied are representative of Chao1 species richness. Significance was determined per the Mann-Whitney U test.*

Analysis of inter-sex and inter-species differences in the adult nematode microbiome

The nematode microbiomes were probed for variation resulting from differences in sex and species. Alpha and beta diversity between male, female, and mixed-sex pools of adult nematodes were examined (Supplementary Figure 6). No significant difference was found in terms of alpha diversity based on Chao1 species richness, using the Mann-Whitney U test (p = 0.546). When beta diversity was visualised using a PcOA plot samples clustered based on the sheep of origin and not based on gender.

The microbiomes of *H. contortus* and *T. circumcincta* adult worms were compared at family level (Figure 5). Due to the novel nature of the microbiomes of both *H. contortus* and *T. circumcincta*, 37.6% of ASVs present in *H. contortus* samples and 34.1% of ASVs present in *T. circumcincta* samples were not classified to family level. The microbiome of *H. contortus* comprised the following families: 36.2% Ruminococcaceae, 27.4% Lachnospiraceae, 11.4% Prevotellaceae, 5.7% Acidaminococcaceae, 4.2% Planctomycetaceae, 1.8 % Acetobacteraceae, 1.4% Spirochetaceae, 1.2% Veillonellaceae, with the remaining fraction comprising negligible proportions. The microbiome of *T. circumcincta* comprised the following families: 37% Lachnospiraceae, 26% Ruminococcaceae, 6.5% Prevotellaceae, 3.5% Planctomycetaceae, 3.3% Acidaminococcaceae, 3% Coriobacteriaceae, 2% Bifidobacteriaceae, with the remaining fraction comprising negligible proportions. Veillonellaceae and Acetobacteraceae were present in significantly higher numbers in *H. contortus* (p = 0.01 and p = 0.005, respectively), while Coriobacteriaceae was significantly more abundant in *T. circumcincta* (p = 0.005). Significance was determined per the Mann-Whitney *U* test.

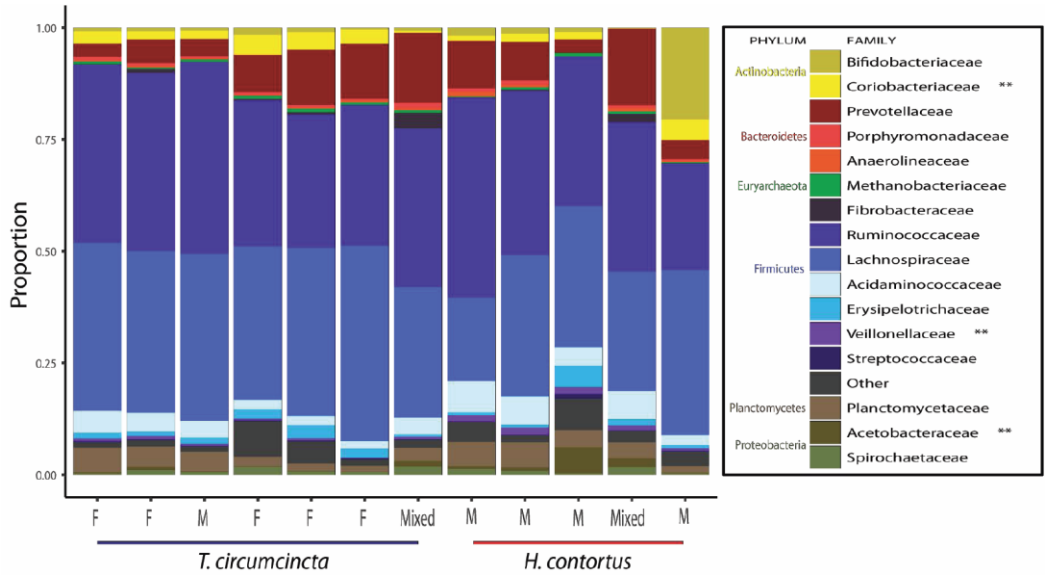


Figure 5: ***Adult nematode Microbiome composition at family level of H. contortus and T. circumcincta.*** *The extent to which various bacterial families contribute to the overall make-up of the microbiomes of H. contortus and T. circumcincta. Each column is derived from a pooled mixture of 100 nematodes (five H. contortus (four males, one mixed sex) and seven T. circumcincta (five females, one male, one mixed sex) samples). Nematodes were taken from the ovine abomasum at post-mortem, 28 days post-infection. Families constituting less than 1% of the total family distribution for a sample were labelled 'Other'.*

Alpha diversity in *H. contortus* was lower than in *T. circumcincta* (Supplementary Figure 7). However, the significance of this comparison between the two nematode microbiomes must be considered in the context of sample size (*H. contortus* n = 5 and *T. circumcincta* n = 7). Differential abundance analysis using Deseq2 revealed 18 ASVs significantly elevated in one nematode: 5 in *H. contortus*, and 13 in *T. circumcincta* (Supplementary Figure 8). Unlike the Mann-Whitney *U* test, this method is applied to individual ASVs. Ruminococcaceae/*Ruminococcus* and Clostridiales dominate the differentially elevated ASVs in *T. circumcincta* and are absent from the differentially elevated ASVs in *H. contortus.*

Effect of nematode infection on the faecal microbiome of the host over time

Changes in alpha and beta diversity of the faecal microbiome of infected lambs were examined over several time points between day 0 and day 28 of infection (Figure 6). Post-infection, there is a decrease in species richness within the faecal microbiome, and an increase in dissimilarity over time, compared with the faecal microbiome pre-infection. There is a significant negative Spearman correlation between alpha diversity and time (p = 0.03). Increasing dissimilarity over time is indicated by a strong positive correlation between principal component axis 1 and time. This same principal component, which explains the most variation in the PCoA, also has a statistically significant negative correlation with alpha diversity. This means that the more dissimilar the infected microbiome becomes compared with the pre-infected microbiome, the lower its alpha diversity becomes. Despite the positive correlation between beta diversity and time, when the mean beta diversity of samples at time points 0 and 28 were compared, there was no statistically significant difference (p = 0.89), although visually it appears to decrease slightly (Supplementary Figure 9).

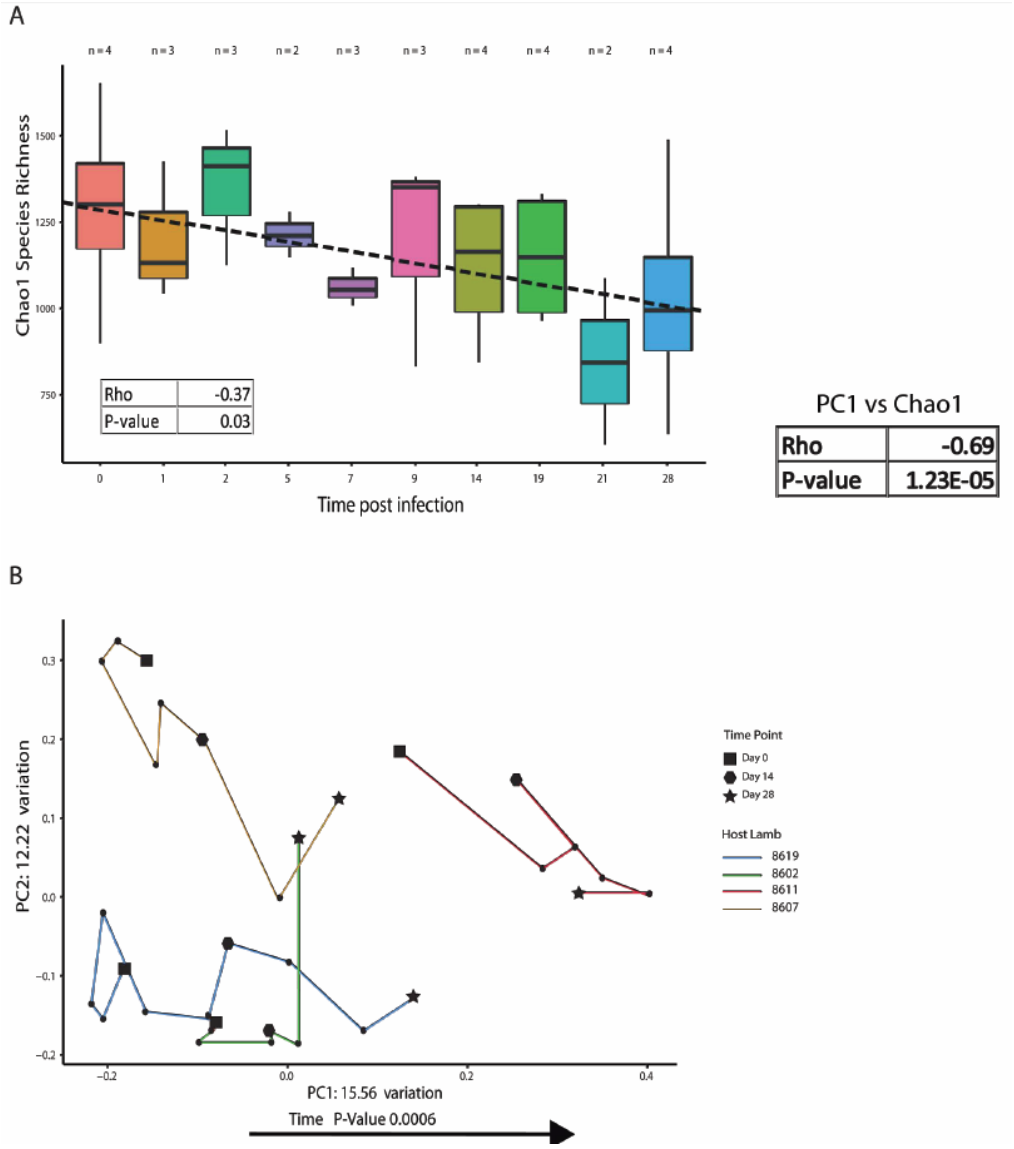


Figure 6*: Changes in alpha and beta diversity of the ovine faecal microbiome over time, post-infection.* *Faecal samples were obtained from two-to-four lambs at 10 timepoints over 28 days. All correlation tests used Spearman's rank correlation coefficient.* ***(A)*** *Changes in alpha diversity of the ovine faecal microbiome over time. There is a statistically significant decrease in Chao1 species richness from day 0 to day 28 of infection.* ***(B)*** *Changes in beta diversity of the ovine faecal microbiome over time. There is a trend in the movement of the lamb faecal microbiome along the x-axis in a positive direction over time, thus becoming more dissimilar to the uninfected lamb microbiome.*

These diversity metrics inform on changes in the overall relatedness of samples but give no information about the individual microbes implicated in the faecal microbiome dysbiosis. All ASVs detected were correlated against time using Spearman's rank correlation coefficient. There were 39 significant ASVs based on this test, of which 11 showed a positive linear relationship with time and 28 a negative one, post-infection (Supplementary Figure 10). The two most prevalent ASVs associated with time were classified as *Bifidobacterium* spp. and *Sharpea* spp., both of which show a negative relationship with time. When blasted against the *nr* database, these two sequences had 100% identity with *Bifidobacterium merycicum,* and *Sharpea azabuensis*. Seven statistically significant ASVs were classified as Ruminococcaceae. Other ASVs, such as the six identified as Candidatus Saccharibacteria, have an ambiguous relationship with time, post-infection, as four of these ASVs show positive correlations, and two negative.

Dialister spp. and *Clostridium* spp. have both been implicated in compromising the human host's ability to clear nematode infection [11]. Conversely, many other bacterial genera and families are suspected to 'immunise' the host against nematode infection (e.g. *Subdoligranulum* spp., *Acinetobacter* spp., *Paracoccus* spp., *Gemminger* spp., Peptococcaceae, Moraxellaceae, Corynebacteriaceae and Hyphomicrobiaceae). Of these bacteria, we observed only Hyphomicrobiaceae in our data, which was significantly elevated in pre-infection larvae over post-infection larvae ($p < 0.05$). Moreover, it is known that helminth infection in mice results in increased abundance of the Lactobacillaceae family, leading to the hypothesis that the anti-inflammatory activity of these bacteria may create permissive conditions for nematode survival in the gut (27). We found similar results with this family in our ovine model, in which a positive correlation with time was observed post-infection (rho = 0.43, $p = 0.01$).

Effect of the ovine microbiome on the nematode microbiome

In addition to defining the effect of nematode infection on the host, the effect of the host microbiome on the microbial composition of the nematode was also investigated by comparing the microbiomes of larval nematodes pre-infection and post-infection. The SourceTracker algorithm failed to detect contamination in the

larvae that may have arisen from the ovine intestinal tract. (Supplementary Figure 11).

There is a significant increase in alpha diversity in the pre-infection larvae compared with post-infection larvae as measured by Chao1 species richness (Figure 7C). The two groups of larvae were also clearly differentiated based on their dissimilarity in the PCoA plot (Figure 7A), with the clustering by group confirmed statistically by PERMANOVA analysis.

The families Planctomycetaceae and Hyphomicrobiaceae are significantly elevated in the pre-infection larvae, while Rhodocyclaceae and Methylobacteriaceae are elevated in post-infection larvae (Figure 7B). ASVs that were differentially abundant between the two groups were identified using DESeq2. 2037 unique ASVs were identified across all larval nematode samples, of which 97 were elevated in the pre-infection larvae, and 190 in the post-infection larvae. In all cases this was statistically significant after correcting for multiple testing. A volcano plot depicting this distribution, and a table of all ASVs identified (Supplementary Figure 12 and 13).

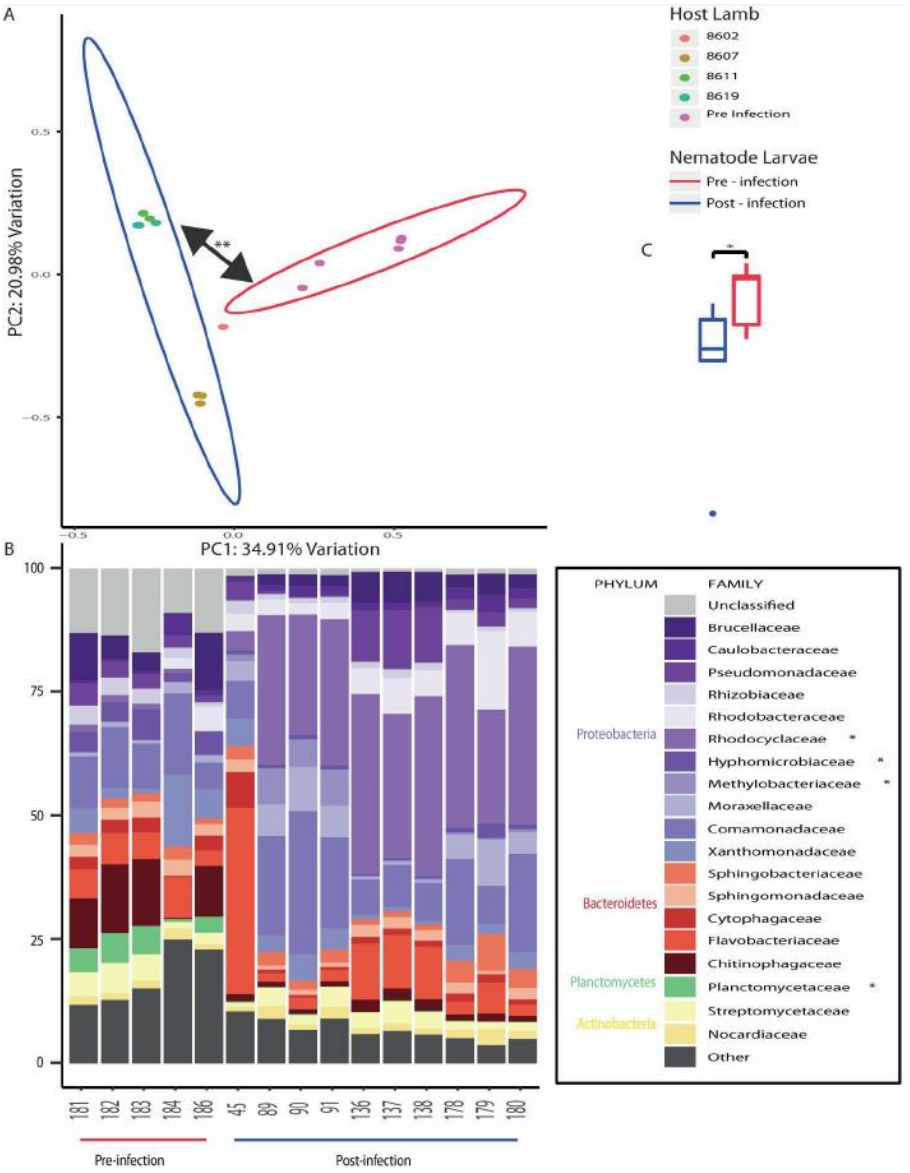


Figure 7: ***(A) Bray-Curtis dissimilarity between pre-infection and post-infection nematode larvae. (B) Microbiome composition at family level of pre-infection and post-infection nematode larvae. (C) Boxplot of Chao1 species richness of pre-infection and post-infection nematode larvae. (A)*** *Bray-Curtis dissimilarity between pre-infection and post-infection larvae. Each point on the plot is derived from a pooled mixture of ~10,000 larvae (5 pre-infection larvae and 10 post-infection larvae). Ellipses show 80% confidence intervals for their respective groups. The two groups separate based on the dissimilarity of their microbial composition. Statistical testing was performed by permutational multivariate analysis of variance.* ***(B)*** *Compositional boxplot of the 19 most-prevalent bacterial families. Each column is derived from a pooled mixture of ~10,000 larvae. Significance testing was performed by the Wilcoxon signed-rank test, with critical values adjusted for multiple comparisons using the Bonferroni method.* ***(C)*** *Boxplot comparing alpha diversity between the two groups as measured by Chao1 species richness. The pre-infection boxplot is derived from five pooled samples of ~10,000 larvae each. The post-infection boxplot is derived from 10 pooled samples of ~10,000 larvae each. Statistical testing was performed by the Wilcoxon signed-rank test.*

Comparison of the nematode and ovine microbiomes

We investigated the capacity for ovine-adapted bacterial taxa to persist in the nematode microbiome. Firstly, nematode larvae were compared with ovine faecal samples, and adult nematodes were compared with ovine abomasal washings on the basis that these samples originated from a common environment – i.e. the ovine gut and abomasum, respectively. Relatively little convergence was evident between the nematode larvae and ovine faecal samples, with only 227 shared ASVs of a possible 9422 unique ASVs identified across both groups (Supplementary Figure 13 and 14) Conversely, when comparing adult nematodes with ovine abomasal washings, 2494 shared ASVs of a possible 6936 unique ASVs were identified across both groups. Samples clustered definitively based on the host animal of origin.

Next, we reviewed several recent studies that have profiled the ovine microbiome at various sites in the digestive tract according to the abundances of endogenous bacteria present (28,29). We then examined our own nematode microbiome data for the presence of bacteria found in sheep in relatively high abundances. Virtually all taxa present in relatively high abundances in the ovine gut, such as *Ruminococcus* spp. and *Bacteroides* spp., were absent from the larvae; however the Peptostreptococcaceae family was identified in all 32 faecal samples and 14/15 larvae. Abomasum-adapted taxa such as *Oscillospira* spp., *Succinivibrio* spp. and *Bacteroides* spp. were not found in the adult nematodes, but *Prevotella* spp., one of the most abundant genera in the ovine abomasum, was found in every ovine abomasum and adult nematode sample, along with the abomasally-adapted *Fibrobacter* spp., which was also found in all abomasal samples, and 10/12 nematode samples (data not shown).

Also of interest were potential differences between the adult nematode and the ovine abomasum. The adult nematode and the abomasal lumen content microbiomes were compared using Deseq2. Twelve ASVs were significantly differentially abundant between the nematode microbiome and that of the ovine abomasum (Figure 8A). The most prevalent differentially abundant ASV was classified as *E. coli*/*Shigella* spp. (the taxonomic resolution necessary to distinguish these bacteria is impossible using 16S rRNA gene sequencing analysis (30)). Following this, ASVs classified as *E. coli*/*Shigella* were screened for in the dataset, resulting in the discovery of four in

total. At least one ASVs appeared in every larval sample, and in seven of the 12 adult nematode samples. ASV 75, the most abundant putative *E. coli*/*Shigella* ASV, was also present at low levels in some of the lamb faecal samples but all ASVs were absent from the ovine abomasum (Supplementary Figure 15A). Nematode colonisation by *E. coli*/*Shigella* did not appear to be specific for either species of nematode – the two ASVs 75 and 295 combined were found in 4/7 *T. circumcincta* samples and 3/5 *H. contortus* samples.

Phylogenetic analyses were carried out, comparing the four *E. coli/Shigella* ASVs found in the dataset with other well-characterised and clinically relevant strains to provide evolutionary context (Supplementary Figure 15B). The bootstrapping values were provided over 1000 iterations. The more distantly related *Klebsiella* spp. and *Salmonella* spp. formed the outgroups, as expected; however, the evolutionary distance between *E. coli/Shigella* genera was limited, as can be seen by the low bootstrapping values at many of the branch points. ASV_295 appears most distantly related to the remaining species, and therefore it is reasonable to suggest that ASV_6240 and *E. coli* MG1655 form a distinct separate clade, although it is not possible to confirm that evolutionary distance exists between ASV_75, ASV_7656, *E. coli* 0157:H7 and *Shigella* spp.

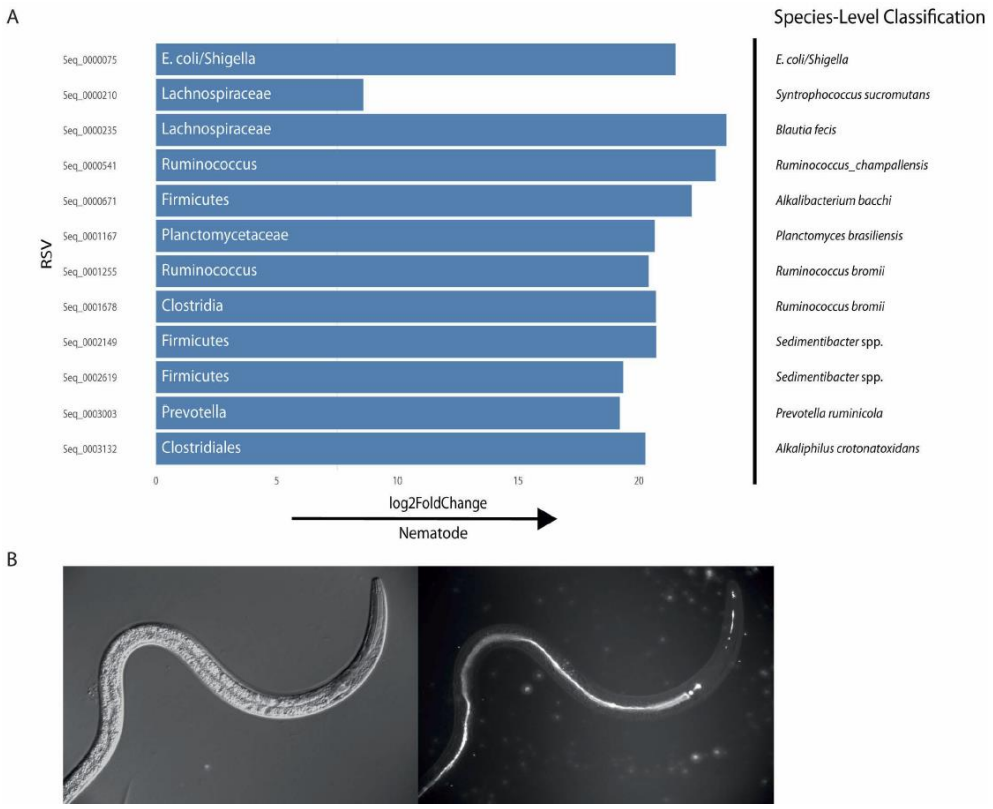


Figure 8: ***(A) Differentially abundant ASVs between the adult nematode and the ovine abomasum, showing level of fold change between either environment. (B) Oral ingestion of engineered E. coli by larvae in vitro. (A)*** *Metabarcoding data for the adult nematodes were derived from 12 pooled samples (five H. contortus (4 males, 1 mixed sex) and seven T. circumcincta (5 females, 1 male, 1 mixed sex) samples) of 100 nematodes each. Metabarcoding data for the abomasum were derived from the abomasal washings of four lambs. Bacteria are labelled with the most accurate taxonomic classification available for that ASV. Differential abundance was determined with Deseq2. Additional classification to species level with SPINGO is provided. This classification was performed with no confidence cut-offs; thus, it is revealing yet imperfect with respect to the identification of the bacteria.* ***(B)*** *Eggs of H. contortus MHco3(ISE) were hatched to first-stage larvae and developed to second-stage larvae on NGM agar supplemented with E. coli OP50-1:GFP (pFPV25.1). DIC image left, U.V. Image on right depicting ingestion of GFP labelled OP50 in pharynx and entire length of gut (Mag x250).*

Oral ingestion of engineered E. coli by larvae in vitro

In vitro oral ingestion of engineered *E. coli* was investigated to assess the potential for exogenous bacteria to reside within the guts of these nematodes, and to locally express heterologous genes. First stage nematode larvae were grown on a plate seeded with an *E. coli* strain, genetically modified to express green fluorescent protein (GFP). Fluorescence microscopy showed GFP fluorescence in the pharynx and the entire length of gut, specifically within GFP-expressing, *E. coli*-fed nematodes. Similar results are observed with *T. circumcincta* (data not shown).

DISCUSSION

The quality and depth of our sequencing analysis permits a thorough understanding of spatial, kinetic and organism-specific patterns of the microbiomes of helminth-infected hosts. This approach is potentially applicable to parasitic disease at large, including helminthic and ectoparasitic infections, on the condition that differences exist between the host and parasite microbiome. Due to the preferential colonisation of the abomasum by *H. contortus* and *T. circumcincta*, it was pertinent to compare these compartments for identification of bacteria that favour nematode cohabitation. and the same rationale was used in the comparison of nematode larvae and ovine faecal samples. The identification of differentially abundant taxa represents valuable knowledge to exploit in future research.

A past study of the *H. contortus* microbiome with primers targeting both the V3-V4 and V5-V7 regions of the 16S rRNA gene resulted in higher OTU capture using the former primer set, although the latter set contrastingly was capable of detecting the phylum Gemmatimonadetes, albeit in relatively low abundance (13). The V3-V4 region of the 16S rRNA gene was sequenced for all samples in this study, rather than the V5-V7 because, while targeting the V5-V7 region would be necessary for mapping comprehensively the microbiome of *H. contortus* by facilitating identification of its less abundant taxa, here our objective was to identify nematode-specific bacteria that are present in relatively high abundance, because these bacteria would be more amenable to concentrating within a nematode, were they administered exogenously. However, there are ways in which less abundant taxa may have important applications for treatment of parasitic disease. For example, there is evidence that bacteria can influence their environment considerably even if their abundance is low (31). Furthermore, it is known that some bacteria, such as *Wolbachia* spp., are essential for the development of filarial nematodes, and that antibiotics targeting *Wolbachia* spp. have filaricidal activity (32). Thus, the use of antibiotics to target nematode-essential bacteria present either in low or high abundance is a valid treatment strategy. An alternative method could involve feeding the infected host a modified diet that would deprive the bacteria in question of essential nutrients.

The commonalities and differences we observed between the ovine and nematode microbiomes (Figures 3 and 4, and Supplementary Figure 5) are interesting because, in the former case, it presents the possibility that the microbiota of either organism may be influencing that of the other and, in the latter case, it means that the differences between parasite and host could be exploited to the benefit of the infected animal. The abomasum and adult nematode microbiomes are by far the most closely related environments (Figure 4 and Supplementary Figure 5). This could be considered unsurprising because these environments are in intimate contact with one another; yet, nematode larvae and host faeces, from which the larvae derive, separate into two distinct clusters despite their proximity. We reasoned that there may exist differences between host and parasite amenable to exploitation despite their gross similarity.

H. contortus and *T. circumcincta* have contrasting lifestyles, the former being a blood feeder and the latter a mucosal grazer (33). Thus, characterising both species simultaneously in a co-infection model could illuminate the effects of alternate feeding habits on the nematode and ovine microbiome. Analysing different species in isolation across separate studies could complicate the identification of the source of any variation, as inter-study differences in soil composition, animal feed, age and immune status of host and living conditions, for example, could affect the ovine microbiota and therefore the microbiota of the nematode. To our knowledge, this is the first report of a parasite-host microbiome study in ruminant livestock that incorporates a co-infection model. It is also the first characterisation of the microbiome of *T. circumcincta*.

Co-infection models are important because it is accepted that different parasites co-habiting the same host can affect each other profoundly in ways that would not occur were they infecting the host as lone pathogens (34). This can result in one parasite creating a permissive environment for the other parasite or, conversely, one parasite negatively affecting the other parasite's growth. In some cases, parasitic cohabiters can have more influence on their host than on each other (35). Additionally, multiple studies claim that co-infection of humans and livestock with nematodes is common (36,37), meaning that more microbiome studies of host and parasite should incorporate co-infection models. Admittedly, this study does not examine the parasite-host microbiome interrelationship in a single-infection model. Therefore, the

effects of *H. contortus* or *T. circumcincta* alone on the ovine microbiome may be different than what is observed here. In response to a critical lack of information regarding the effects of co-infection on cohabiting parasites, a recent study has successfully employed methodology to predict how two nematodes will influence each other in terms of survival, even when they are examined in different host species (34). Future research would benefit this field by attempting to predict how host co-infection influences the microbiome compared with single-strain infections.

We discovered that the two species of nematode contain microbiomes that are in many ways comparable. This is not unexpected, given the finding that marine nematodes deriving even from different parts of the planet contain similar microbiomes (16). However, there are statistically significant differences that are worth noting, namely that the families Veillonellaceae and Acetobacteraceae are both elevated in *H. contortus*, and Coriobacteriaceae is elevated in *T. circumcincta* ($p > 0.01$) (Figure 5). The fact that different species of nematode living in the same host have quantifiable differences in their microbiomes suggests that the contrasting lifestyles between the two species may be directly responsible for significant changes in microbiome constitution.

Microbiomes associated with improved host health are noted for having high levels of microbial diversity. As such, if parasitic nematode infections were to alter the host's microbiome, they may have more a profound effect on the health of the host than what is currently appreciated. Infection with multiple parasitic species is a natural phenomenon and is underlined as a more crucial determinant of the effects of infection on host health than host-specific and environmental factors (38); thus, the effects of co-infection on the microbiome could be just as pronounced. We detected an obvious decrease in alpha diversity 21 days post-infection. *H. contortus* and *T. circumcincta* pre-patent periods are both approximately three weeks (39,40), suggesting that nematode infection has a lesser impact on the microbiome of the host in the initial stages of the nematode life cycle, and only begins to have a noticeable effect once the parasites mature and move into the abomasal lumen rather than residing within the tissue. However, the dose administered to the lambs in this study was sub-clinical, which also may explain why the decrease in alpha diversity was not observed until the latter part of the life cycle. It is possible that the effects on

microbiome diversity could become magnified and/or occur earlier if infections were more acute.

Notably, previous work, albeit within goats, showed that *H. contortus* infection did not result in a shift in abomasal microbiome diversity; however, an effect was seen on the abundances of several bacterial species (41). Contrastingly, infection of lambs with *H. contortus* alone was found to increase microbiome diversity in the abomasum (42). Differences observed may be attributable to inter-species differences and/or inter-study differences. For example, although both studies administered the same dose of *H.* contortus, the latter study involved pre-treatment of its animals with the anthelmintics ivermectin and levamisole, which may have removed pre-existing infection that otherwise may have affected study outcome. A study of humans, many of whom were infected with multiple nematodes (most commonly *Trichuris* spp., followed by *Ascaris* spp., followed by hookworm), concluded that helminth infection resulted in an increase in diversity of the faecal microbiome (37). It could be the case that the effect of nematode infection on microbiome diversity within the host may be microbiome-specific (i.e. abomasal vs. faecal), and/or species-specific (i.e. ovine vs. caprine vs. human). It is perhaps relevant that *Trichuris* spp., *Ascaris* spp. and hookworm are each intestinal helminths, while *H. contortus* and *T. circumcincta* are abomasal helminths. It is reasonable to postulate that parasites will have varying impacts on body sites with which they are directly in contact, than if they were persisting remotely. Furthermore, changes that occur as a result of abomasal colonisation may have dramatically different effects on microbial viability and composition in other, downstream *in vivo* compartments (e.g. the intestines) that would not occur were the intestines colonised. For example, there is evidence that colonisation with *H. contortus* decreases the acidity of the ruminant stomach (42), potentially altering microbial growth patterns here and other areas of the gut. Further study is required to fully understand the extent to which parasite lifestyle and host-specific factors come to bear on microbiome diversity.

In addition to a quantifiable decrease in diversity, the quality of the shift is also noteworthy. *Bifidobacterium merycicum* and *Sharpea azabuensis*, both of which become reduced over time, would be considered typical constituents of a healthy ruminant microbiome (43,44). Similarly, Ruminococcaceae can be considered a

dominant ruminant bacterial family (45) and again, all associated ASVs show a negative correlation with time. Unlike the dominant ruminant bacteria which are clearly affected by nematode infection of the host, some other changes in the host microbiome not directly related to parasitic infection are inevitable due to interactions between bacteria. Bacterial species compete for resources in various ecological niches within the host, produce antibiotics, and often rely on syntrophy for their survival (46). Thus, it is cautioned that the results of microbiome studies must be considered against a potential background of inter-bacteria interactions that may confound precise interpretation of changes observed.

Taxa that have suggested involvement in either maintenance or clearance of human nematode infection, such as *Dialister* spp. and *Lactovum* spp. (11), were largely unfound in the ovine microbiome in the present study, with the exception of the Hyphomicrobiaceae family, which was elevated in pre-infection nematode larvae over post-infection larvae. Thus, while these bacteria may have an important role to play in human infection, it is improbable that they are fundamental to the establishment or curtailment of nematode colonisation of the ruminant host, and at the very least might only facilitate the establishment or removal of infection. An increase in the level of anti-inflammatory Lactobacillaceae in murine models of others studies (10), and in the present ovine study, is suggestive of a symbiotic relationship between bacteria and parasite, wherein Lactobacillaceae thrive in the presence of nematode infection, while nematode infection is sustained by the dampened immune response effected by this altered microbial signature.

The degree of overlap observed in this study between host and parasite microbiomes occupying the same environment within the host provides insight into the origination of the nematode microbiome and is suggestive of the ability of ruminant-adapted taxa to invade a new niche within the host. The data present a strong case for the mature nematode either feeding on or being passively colonised by constituent bacteria of the ovine abomasum. While many taxa associated with the abomasum are absent from the adult nematode microbiome, there is a significant degree of overlap between the two groups at an ASV level, especially by the highly abundant, abomasally-adapted genera *Prevotella* spp. and *Fibrobacter* spp. All adult nematodes cluster definitively by host organism (Supplementary Figure 14),

suggesting that these common taxa were indeed acquired by the nematode upon reaching the abomasum.

The identification of differentially abundant taxa presents future opportunities for use as research tools, or indeed therapeutic approaches. While invaluable in combatting helminthic disease, anthelmintic drugs have been the victims of their own success. Frequent and routine use of anthelmintic has led to the prevalence of anthelmintic resistance increasing globally, with multiple class anthelmintic resistance being commonplace in *H. contortus* and *T. circumcincta* globally (47). The development of anthelmintic resistance and consumer concerns over chemical residues in the milk and meat products of treated animals (48) are potentially limiting factors in the deployment of these drugs in the future.

Our metabarcoding data suggest that the microbiomes of *H. contortus* and *T. circumcincta* are significantly different from their ovine environment most notably with respect to *E. coli/Shigella* spp. *E. coli* may be a much more natural coloniser of nematodes than of animals, and there are several pieces of clinical evidence that support this. Firstly, it is known in human subjects that *E. coli* is not among the most abundant species found in the gastrointestinal tract and that its numbers may in fact be quite low (49). Moreover, probiotic strains of *E. coli*, such as *E. coli* Nissle 1917, are frequently unsuccessful colonisers of the human gut even when administered in relatively high doses (50), and once colonised often do not persist for long in the gut once the dose is stopped (51). Thus, naturally low levels of *E. coli* in animals may be sufficient to ensure its selective compartmentalisation in nematodes. Alternatively, it is possible that *E. coli* is vertically transmitted in nematodes and that migration from the host either does not take place or has a lesser impact than vertical transmission.

This study provides a rationale for the study and use of parasite-specific bacteria in drug development practices. The successful feeding of infective nematodes with a genetically modified bacterium could be exploited in several ways. An example is a bacterial assay formatted to assess the efficacy of anthelminthic drugs. Bacteria have recently been engineered to ‘sense’ molecules that cannot be quantified by non-invasive methods (52,53). These bacteria can detect exposure to a drug, and record this exposure using a memory circuit. This could create a platform through which pharmacokinetic studies on anti-parasitic drugs could be easily and non-invasively

performed – both on market-approved compounds and drugs still undergoing clinical testing. Alternatively, bacteria could be used as vehicles for drug delivery, which has many advantages beyond conventional chemical medicines, not least of which is the targeted delivery of therapeutics (52).

E. coli is an ideal candidate for bacteria-mediated drug delivery. It is readily engineered and highly flexible as a drug testing platform and various strains of this species have attracted interest for their probiotic properties (54). Its preclinical validation in various drug delivery modalities is also a reassuring aspect of this bacterium (53,55-59). Thus, the selective colonisation of the nematode microbiome by *E. coli/Shigella* is encouraging and invites further investigation of bacteria as orally administrable, target-specific agents.

In summary, this study highlights the potential value in exploitation of nematode microbiota in progression of novel treatments for parasitic diseases affecting both animals and humans.

ACKNOWLEDGMENTS

We are grateful to the Bioservices Division, Moredun Research Institute, for provision of samples, and acknowledge the provision of *in vitro* larval images with kind permission from Prof Antony Page, University of Glasgow.

ETHICAL STATEMENT

All experimental procedures described here were approved by the Moredun Research Animal Welfare and Ethical Review Body and were conducted under the legislation of a UK Home Office License (reference P95890EC1) in accordance with the Animals (Scientific Procedures) Act of 1986.

References

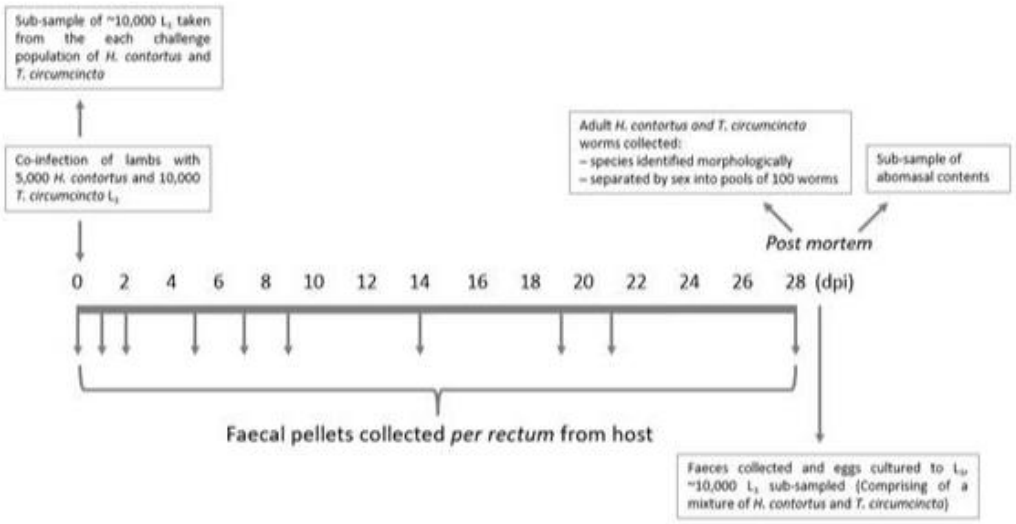
1. Torgerson, P.R., Devleesschauwer, B., Praet, N., Speybroeck, N., Willingham, A.L., Kasuga, F., Rokni, M.B., Zhou, X.N., Fevre, E.M., Sripa, B. *et al.* (2015) World Health Organization Estimates of the Global and Regional Disease Burden of 11 Foodborne Parasitic Diseases, 2010: A Data Synthesis. *PLoS medicine*, **12**, e1001920.
2. Kenyon, F., Hutchings, F., Morgan-Davies, C., van Dijk, J. and Bartley, D.J. (2017) Worm Control in Livestock: Bringing Science to the Field. *Trends in parasitology*, **33**, 669-677.
3. Rose, H., Rinaldi, L., Bosco, A., Mavrot, F., de Waal, T., Skuce, P., Charlier, J., Torgerson, P.R., Hertzberg, H., Hendrickx, G. *et al.* (2015) Widespread anthelmintic resistance in European farmed ruminants: a systematic review. *The Veterinary record*, **176**, 546.
4. Peachey, L.E., Pinchbeck, G.L., Matthews, J.B., Burden, F.A., Behnke, J.M. and Hodgkinson, J.E. (2016) Papaya latex supernatant has a potent effect on the free-living stages of equid cyathostomins in vitro. *Veterinary Parasitology*, **228**, 23-29.
5. Hogan, G. and Tangney, M. (2018) The Who, What, and Why of Drug Discovery and Development. *Trends in pharmacological sciences*, **39**, 848-852.
6. Partridge, F.A., Murphy, E.A., Willis, N.J., Bataille, C.J., Forman, R., Heyer-Chauhan, N., Marinic, B., Sowood, D.J., Wynne, G.M., Else, K.J. *et al.* (2017) Dihydrobenz[e][1,4]oxazepin-2(3H)-ones, a new anthelmintic chemotype immobilising whipworm and reducing infectivity in vivo. *PLoS neglected tropical diseases*, **11**, e0005359.
7. Partridge, F.A., Forman, R., Willis, N.J., Bataille, C.J.R., Murphy, E.A., Brown, A.E., Heyer-Chauhan, N., Marinic, B., Sowood, D.J.C., Wynne, G.M. *et al.* (2018) 2,4-Diaminothieno[3,2-d]pyrimidines, a new class of anthelmintic with activity against adult and egg stages of whipworm. *PLoS neglected tropical diseases*, **12**, e0006487.
8. Zaiss, M.M. and Harris, N.L. (2016) Interactions between the intestinal microbiome and helminth parasites. *Parasite Immunology*, **38**, 5-11.
9. White, E.C., Houlden, A., Bancroft, A.J., Hayes, K.S., Goldrick, M., Grencis, R.K. and Roberts, I.S. (2018) Manipulation of host and parasite microbiotas: Survival strategies during chronic nematode infection. *Science advances*, **4**, eaap7399.
10. Glendinning, L., Nausch, N., Free, A., W Taylor, D. and Mutapi, F. (2014) *The microbiota and helminths: Sharing the same niche in the human host*.
11. Rosa, B.A., Supali, T., Gankpala, L., Djuardi, Y., Sartono, E., Zhou, Y., Fischer, K., Martin, J., Tyagi, R., Bolay, F.K. *et al.* (2018) Differential human gut microbiome assemblages during soil-transmitted helminth infections in Indonesia and Liberia. *Microbiome*, **6**, 33.
12. Dirksen, P., Marsh, S.A., Braker, I., Heitland, N., Wagner, S., Nakad, R., Mader, S., Petersen, C., Kowallik, V., Rosenstiel, P. *et al.* (2016) The native microbiome of the nematode Caenorhabditis elegans: gateway to a new host-microbiome model. *BMC biology*, **14**, 38.
13. El-Ashram, S. and Suo, X. (2017) Exploring the microbial community (microflora) associated with ovine Haemonchus contortus (macroflora) field strains. *Scientific reports*, **7**, 70.
14. Meyer, J.M., Baskaran, P., Quast, C., Susoy, V., Rodelsperger, C., Glockner, F.O. and Sommer, R.J. (2017) Succession and dynamics of Pristionchus nematodes and their microbiome during decomposition of Oryctes borbonicus on La Reunion Island. *Environmental microbiology*, **19**, 1476-1489.
15. Derycke, S., De Meester, N., Rigaux, A., Creer, S., Bik, H., Thomas, W.K. and Moens, T. (2016) Coexisting cryptic species of the Litoditis marina complex

(Nematoda) show differential resource use and have distinct microbiomes with high intraspecific variability. *Molecular ecology*, **25**, 2093-2110.
16. Schuelke, T., Pereira, T.J., Hardy, S.M. and Bik, H.M. (2018) Nematode-associated microbial taxa do not correlate with host phylogeny, geographic region or feeding morphology in marine sediment habitats. *Molecular ecology*, **27**, 1930-1951.
17. Paramsothy, S., Kamm, M.A., Kaakoush, N.O., Walsh, A.J., van den Bogaerde, J., Samuel, D., Leong, R.W.L., Connor, S., Ng, W., Paramsothy, R. *et al.* (2017) Multidonor intensive faecal microbiota transplantation for active ulcerative colitis: a randomised placebo-controlled trial. *Lancet*, **389**, 1218-1228.
18. Biddle, A.S. (2013) An In Vitro Model of the Horse Gut Microbiome Enables Identification of Lactate-Utilizing Bacteria That Differentially Respond to Starch Induction. **8**.
19. Schallig, H.D. (2000) Immunological responses of sheep to Haemonchus contortus. *Parasitology*, **120 Suppl**, S63-72.
20. Harfoot, C.G. (1978) Anatomy, physiology and microbiology of the ruminant digestive tract. *Progress in lipid research*, **17**, 1-19.
21. Patterson, D.M., Jackson, F., Huntley, J.F., Stevenson, L.M., Jones, D.G., Jackson, E. and Russel, A.J. (1996) Studies on caprine responsiveness to nematodiasis: segregation of male goats into responders and non-responders. *International journal for parasitology*, **26**, 187-194.
22. Marze, N.A., Roy Burman, S.S., Sheffler, W. and Gray, J.J. (2018) Efficient flexible backbone protein–protein docking for challenging targets. *Bioinformatics (Oxford, England)*, **34**, 3461-3469.
23. Bancroft, J.D. and Gamble, M. (2008) *Theory and Practice of Histological Techniques*. Churchill Livingstone.
24. Callahan, B.J., McMurdie, P.J. and Holmes, S.P. (2017) Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, **11**, 2639-2643.
25. Love, M.I., Huber, W. and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, **15**, 550-550.
26. Eisenhofer, R., Minich, J.J., Marotz, C., Cooper, A., Knight, R. and Weyrich, L.S. (2019) Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. *Trends Microbiol*, **27**, 105-117.
27. Glendinning, L., Nausch, N., Free, A., Taylor, D.W. and Mutapi, F. (2014) The microbiota and helminths: sharing the same niche in the human host. *Parasitology*, **141**, 1255-1271.
28. Wang, J., Fan, H., Han, Y., Zhao, J. and Zhou, Z. (2017) Characterization of the microbial communities along the gastrointestinal tract of sheep by 454 pyrosequencing analysis. *Asian-Australasian journal of animal sciences*, **30**, 100-110.
29. Zeng, Y., Zeng, D., Ni, X., Zhu, H., Jian, P., Zhou, Y., Xu, S., Lin, Y., Li, Y., Yin, Z. *et al.* (2017) Microbial community compositions in the gastrointestinal tract of Chinese Mongolian sheep using Illumina MiSeq sequencing revealed high microbial diversity. *AMB Express*, **7**, 75.
30. Chen, L., Cai, Y., Zhou, G., Shi, X., Su, J., Chen, G. and Lin, K. (2014) Rapid Sanger sequencing of the 16S rRNA gene for identification of some common pathogens. *PloS one*, **9**, e88886.
31. Pester, M., Bittner, N., Deevong, P., Wagner, M. and Loy, A. (2010) A 'rare biosphere' microorganism contributes to sulfate reduction in a peatland. *The ISME journal*, **4**, 1591-1602.
32. Slatko, B.E., Luck, A.N., Dobson, S.L. and Foster, J.M. (2014) Wolbachia endosymbionts and human disease control. *Molecular and biochemical parasitology*, **195**, 88-95.

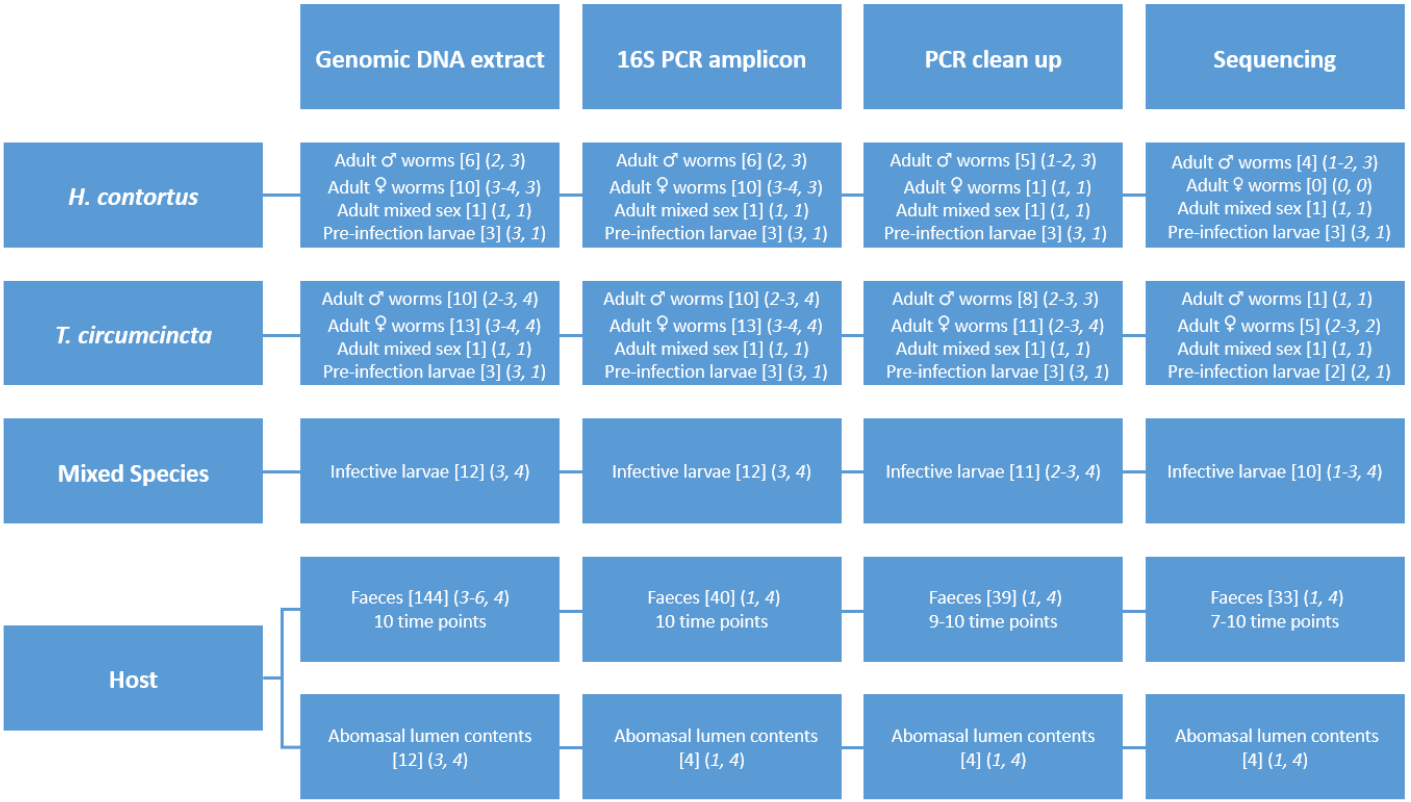
33. Murray, J. and Smith, W.D. (1994) Ingestion of host immunoglobulin by three non-blood-feeding nematode parasites of ruminants. *Research in veterinary science*, **57**, 387-389.
34. Lello, J., McClure, S.J., Tyrrell, K. and Viney, M.E. (2018) Predicting the effects of parasite co-infection across species boundaries. *Proceedings. Biological sciences*, **285**.
35. Murphy, L., Pathak, A.K. and Cattadori, I.M. (2013) A co-infection with two gastrointestinal nematodes alters host immune responses and only partially parasite dynamics. *Parasite immunology*, **35**, 421-432.
36. Almeida, F.A., Bassetto, C.C., Amarante, M.R.V., Albuquerque, A.C.A., Starling, R.Z.C. and Amarante, A. (2018) Helminth infections and hybridization between Haemonchus contortus and Haemonchus placei in sheep from Santana do Livramento, Brazil. *Revista brasileira de parasitologia veterinaria = Brazilian journal of veterinary parasitology : Orgao Oficial do Colegio Brasileiro de Parasitologia Veterinaria*, **27**, 280-288.
37. Lee, S.C., Tang, M.S., Lim, Y.A., Choy, S.H., Kurtz, Z.D., Cox, L.M., Gundra, U.M., Cho, I., Bonneau, R., Blaser, M.J. *et al.* (2014) Helminth colonization is associated with increased diversity of the gut microbiota. *PLoS neglected tropical diseases*, **8**, e2880.
38. Telfer, S., Lambin, X., Birtles, R., Beldomenico, P., Burthe, S., Paterson, S. and Begon, M. (2010) Species interactions in a parasite community drive infection risk in a wildlife population. *Science*, **330**, 243-246.
39. Goossens, B., Osaer, S., Kora, S., Jaitner, J., Ndao, M. and Geerts, S. (1997) The interaction of Trypanosoma congolense and Haemonchus contortus in Djallonke sheep. *International journal for parasitology*, **27**, 1579-1584.
40. Kenyon, F., Sargison, N.D., Skuce, P.J. and Jackson, F. (2009) Sheep helminth parasitic disease in south eastern Scotland arising as a possible consequence of climate change. *Veterinary parasitology*, **163**, 293-297.
41. Li, R.W., Li, W., Sun, J., Yu, P., Baldwin, R.L. and Urban, J.F. (2016) The effect of helminth infection on the microbial composition and structure of the caprine abomasal microbiome. *Scientific reports*, **6**, 20606.
42. El-Ashram, S., Al Nasr, I., Abouhajer, F., El-Kemary, M., Huang, G., Dincel, G., Mehmood, R., Hu, M. and Suo, X. (2017) Microbial community and ovine host response varies with early and late stages of Haemonchus contortus infection. *Veterinary research communications*, **41**, 263-277.
43. Kamke, J., Kittelmann, S., Soni, P., Li, Y., Tavendale, M., Ganesh, S., Janssen, P.H., Shi, W., Froula, J., Rubin, E.M. *et al.* (2016) Rumen metagenome and metatranscriptome analyses of low methane yield sheep reveals a Sharpea-enriched microbiome characterised by lactic acid formation and utilisation. *Microbiome*, **4**, 56.
44. Biavati, B. and Mattarelli, P. (1991) Bifidobacterium ruminantium sp. nov. and Bifidobacterium merycicum sp. nov. from the rumens of cattle. *International journal of systematic bacteriology*, **41**, 163-168.
45. Henderson, G., Cox, F., Ganesh, S., Jonker, A., Young, W., Global Rumen Census, C. and Janssen, P.H. (2015) Rumen microbial community composition varies with diet and host, but a core microbiome is found across a wide geographical range. *Scientific Reports*, **5**, 14567.
46. Menon, R., Ramanan, V. and Korolev, K.S. (2018) Interactions between species introduce spurious associations in microbiome studies. *PLOS Computational Biology*, **14**, e1005939.
47. Kaplan, R.M. (2004) Drug resistance in nematodes of veterinary importance: a status report. *Trends in parasitology*, **20**, 477-481.
48. Fernandes, M.A.M., Gilaverte, S., Bianchi, M.D., da Silva, C.J.A., Molento, M.B., Reyes, F.G.R. and Monteiro, A.L.G. (2017) Moxidectin residues in tissues of lambs

submitted to three endoparasite control programs. *Research in veterinary science*, **114**, 406-411.
49. Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T. *et al.* (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59-65.
50. Prilassnig, M., Wenisch, C., Daxboeck, F. and Feierl, G. (2007) Are probiotics detectable in human feces after oral uptake by healthy volunteers? *Wiener klinische Wochenschrift*, **119**, 456-462.
51. Joeres-Nguyen-Xuan, T.H., Boehm, S.K., Joeres, L., Schulze, J. and Kruis, W. (2010) Survival of the probiotic Escherichia coli Nissle 1917 (EcN) in the gastrointestinal tract given in combination with oral mesalamine to healthy volunteers. *Inflammatory bowel diseases*, **16**, 256-262.
52. Riglar, D.T. and Silver, P.A. (2018) Engineering bacteria for diagnostic and therapeutic applications. *Nature Reviews Microbiology*, **16**, 214.
53. Flores Bueso, Y., Lehouritis, P. and Tangney, M. (2018) In situ biomolecule production by bacteria; a synthetic biology approach to medicine. *J Control Release*, **275**, 217-228.
54. Wassenaar, T.M. (2016) Insights from 100 Years of Research with Probiotic E. Coli. *European journal of microbiology & immunology*, **6**, 147-161.
55. Murphy, C., Rettedal, E., Lehouritis, P., Devoy, C. and Tangney, M. (2017) Intratumoural production of TNFalpha by bacteria mediates cancer therapy. *PLoS One*, **12**, e0180034.
56. Lehouritis, P., Stanton, M., McCarthy, F.O., Jeavons, M. and Tangney, M. (2016) Activation of multiple chemotherapeutic prodrugs by the natural enzymolome of tumour-localised probiotic bacteria. *Journal of controlled release : official journal of the Controlled Release Society*, **222**, 9-17.
57. Cronin, M., Le Boeuf, F., Murphy, C., Roy, D.G., Falls, T., Bell, J.C. and Tangney, M. (2014) Bacterial-mediated knockdown of tumor resistance to an oncolytic virus enhances therapy. *Molecular therapy : the journal of the American Society of Gene Therapy*, **22**, 1188-1197.
58. Byrne, W.L., Murphy, C.T., Cronin, M., Wirth, T. and Tangney, M. (2014) Bacterial-mediated DNA delivery to tumour associated phagocytic cells. *Journal of controlled release : official journal of the Controlled Release Society*, **196**, 384-393.
59. Lehouritis, P., Hogan, G. and Tangney, M. (2017) Designer bacteria as intratumoural enzyme biofactories. *Advanced drug delivery reviews*, **118**, 8-23.

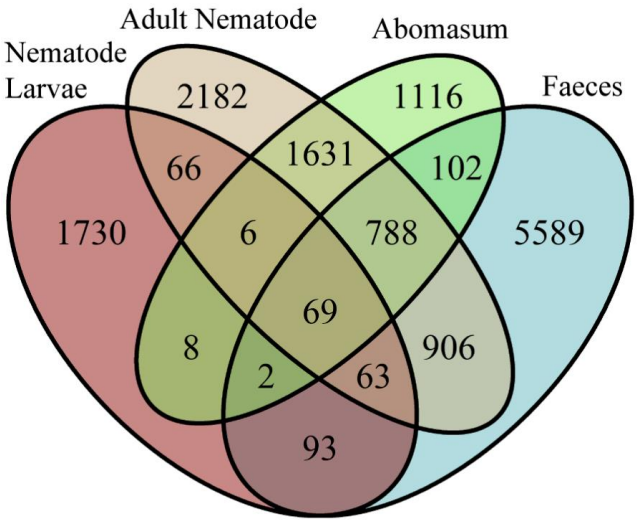
Supplementary Figures



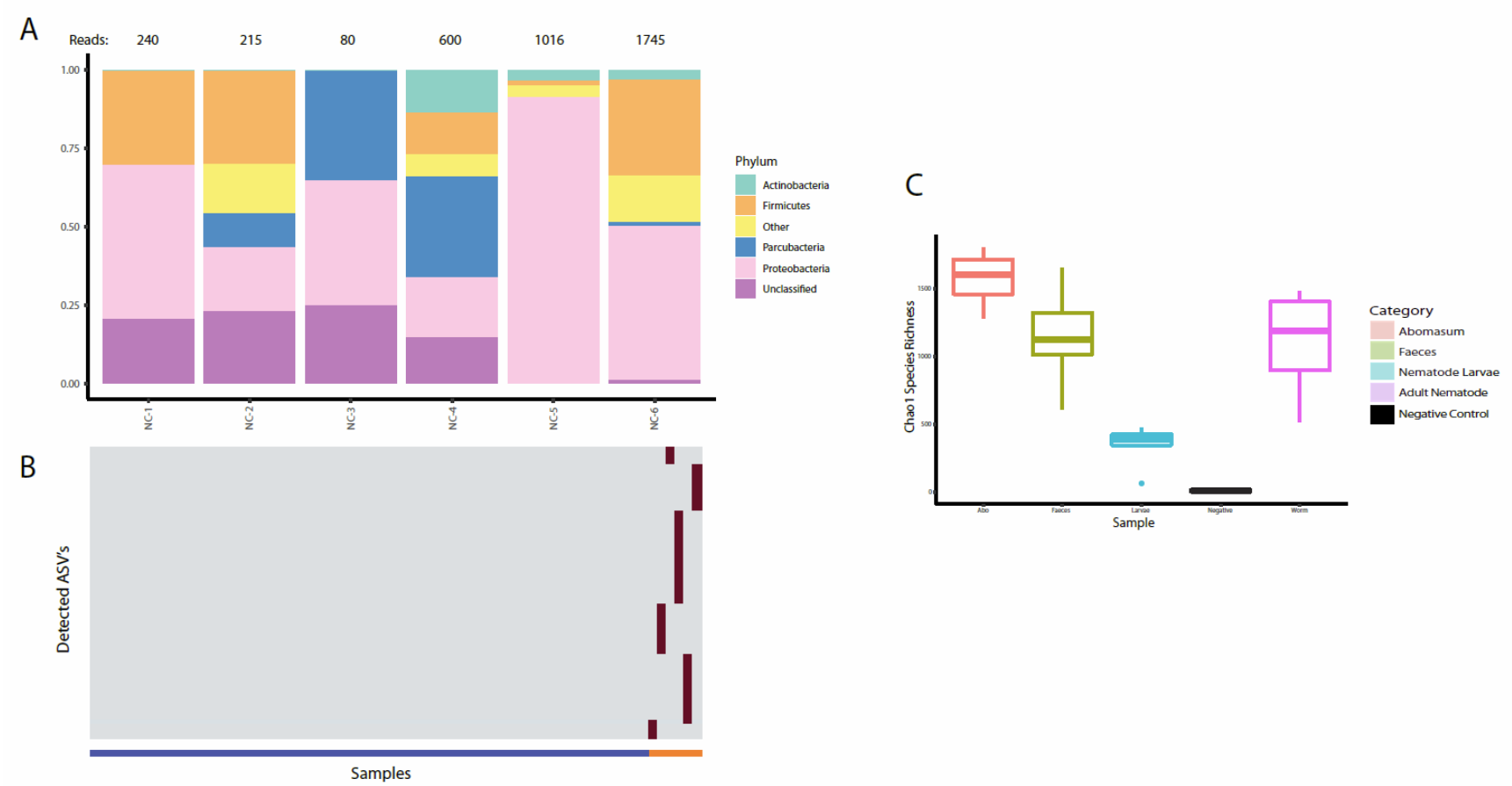
Supplementary Figure 1*: Overview of study timeline.* *The points at which host and parasite samples were collected across a 28-day nematode co-infection in lambs.*



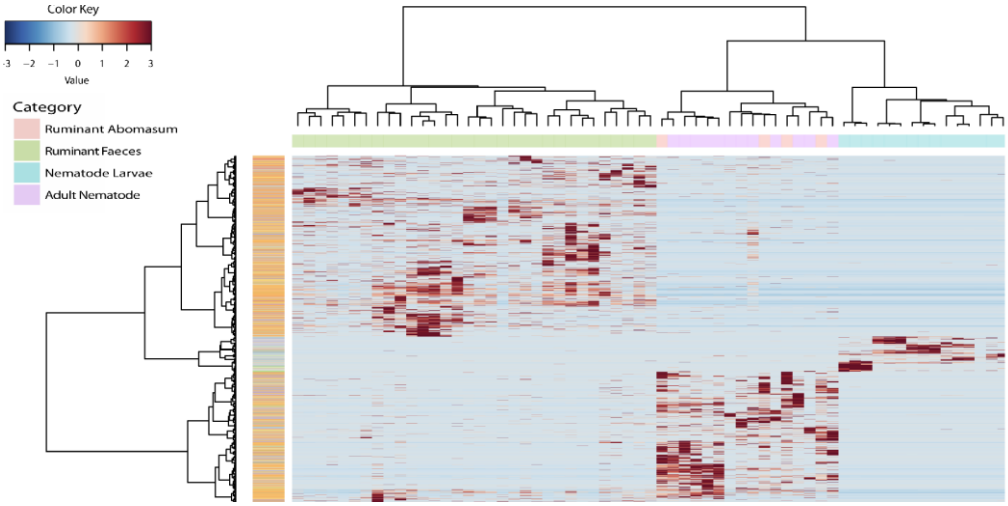
Supplementary Figure 2*: Flow diagram outlining the number of samples at each stage of the process, from genomic DNA extraction to the final sequencing of bacterial 16S rRNA gene amplicons.* *Numbers of samples are indicated for nematodes (pre-infection larvae and sheep-derived larval and adult H. contortus and T. circumcincta) and lambs (abomasal and faecal samples). The total number of samples processed is shown in square brackets and the round brackets show (number of replicates, number of animals). In total, 215 genomic DNA extractions were processed, excluding negative controls.*



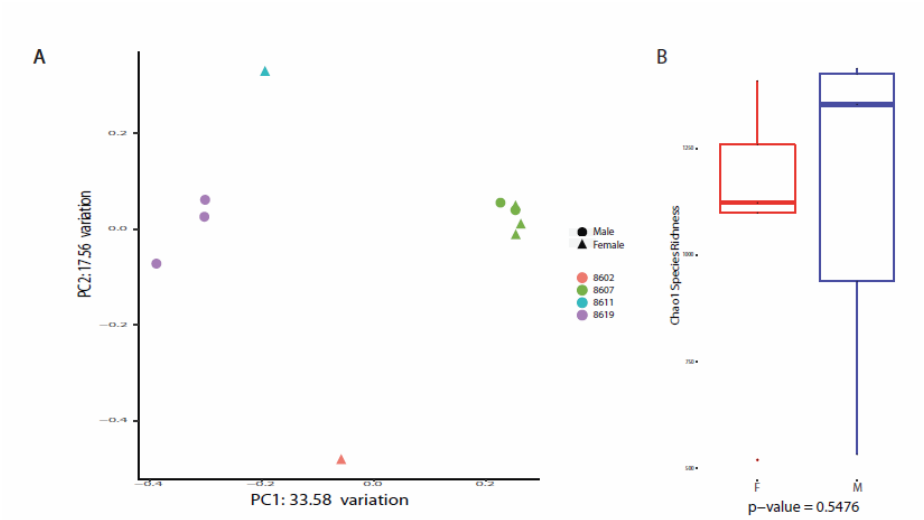
Supplementary Figure 3: ***Distribution of ASVs among the ovine microbiome (abomasal lumen contents and faeces) and nematode microbiome (larval and adult nematodes).*** *ASVs for 'Nematode Larvae' were derived from 15 pooled samples of ~10,000 larvae each (5 pre-infection larvae and 10 post-infection larvae); ASVs for the 'Adult Nematode' were derived from 12 pooled samples of 100 nematodes each (five H. contortus (four males, 1 mixed sex) and seven T. circumcincta (5 females, 1 male, 1 mixed sex) samples); ASVs for the 'Abomasum' were derived from the abomasal washings of four lambs; and ASVs for the 'Faeces' were derived from 33 faecal samples taken from the four lambs across 10 timepoints. ASVs unique to any of the four environments are indicated by numbers in non-overlapping sections of the diagram. ASVs shared between two or more environments are indicated by numbers in overlapping sections.*



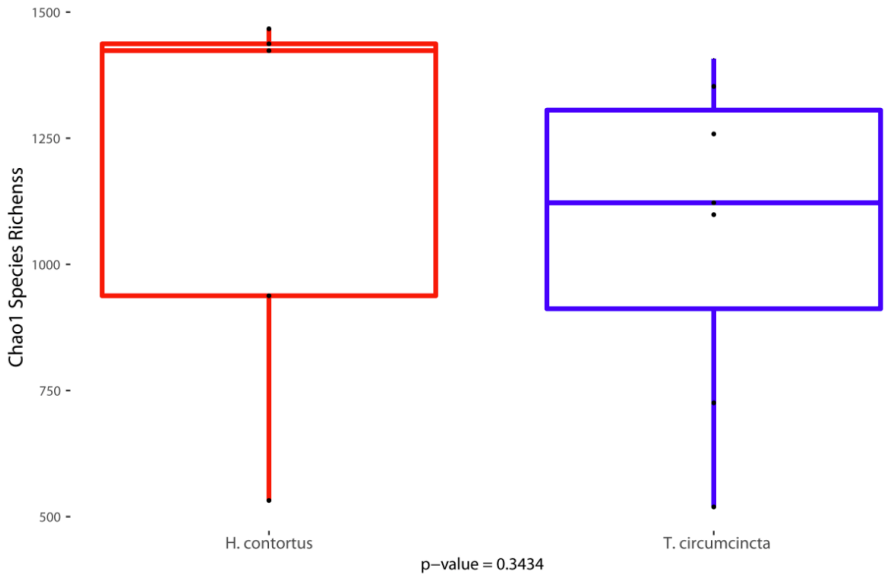
Supplementary Figure 4: Bioinformatic analyses of negative control samples to probe for potential sample contamination. *(A) Sample composition boxplot, at phylum-level, of the 6 negative control samples sequenced. Numbers of high-quality error-free reads obtained from each sample are shown above each column. (B) Heatplot of 75 unique ASVs identified in the negative control samples, illustrating their absence in the experimental samples. (C) Comparison of average Chao1 species richness between different grouped sample types.*



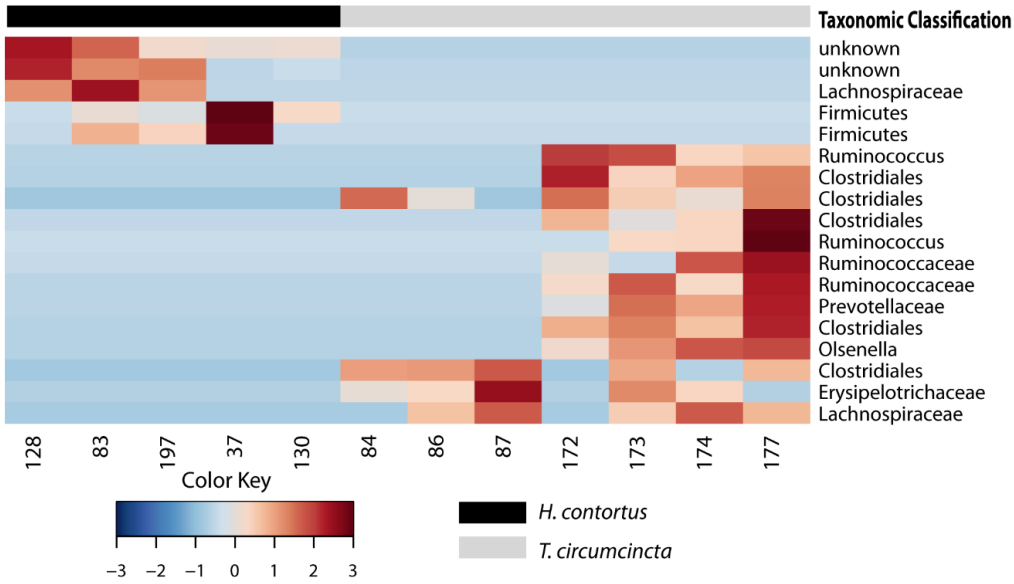
Supplementary Figure 5*: Hierarchical clustering of ASVs, filtered for 5% presence across the ovine microbiome (abomasal lumen contents and faeces) and nematode microbiome (larval and adult nematodes).* *Data for 'Nematode Larvae' were derived from 15 pooled samples of ~10,000 larvae each (5 pre-infection larvae and 10 post-infection larvae); data for the 'Adult Nematode' were derived from a pooled mixture of 100 nematodes (five H. contortus (4 males, 1 mixed sex) and seven T. circumcincta (5 females, 1 male, 1 mixed sex) samples); data for the 'Abomasum' were derived from the abomasal washings of four lambs; and data for the 'Faeces' were derived from 33 faecal samples taken from the four lambs across 10 timepoints. Hierarchical clustering was carried out using the Ward-linkage procedure. Each row represents one ASV, with the relevant phylum indicated by the coloured bar to the left of the plot.*



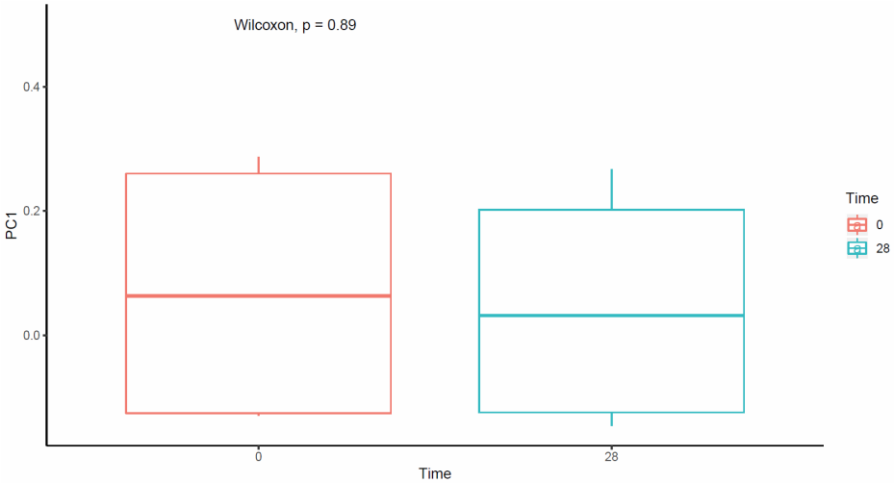
Supplementary Figure 6: ***Comparison of the microbiomes of male and female nematodes.*** *(A) Beta diversity, visualised using Bray-Curtis dissimilarity, is represented in two dimensions using a PCoA plot. Samples are coloured according to the host lamb of origin, and shaped according to gender. Data were derived from five pooled H. contortus (4 male, 1 mixed sex) samples of 100 nematodes each and seven pooled T. circumcincta (5 females, 1 male, 1 mixed sex) samples of 100 nematodes each. (B) Boxplot comparing average Chao1 Species richness between male and female adult nematodes.*



Supplementary Figure 7*: Comparison of alpha diversity between the adult nematodes H. contortus and T. circumcincta.* *Data were derived from five pooled H. contortus (4 male, 1 mixed sex) samples of 100 nematodes each and seven pooled T. circumcincta (5 females, 1 male, 1 mixed sex) samples of 100 nematodes each. Alpha diversity measured as Chao1 species richness. Statistical testing was performed by the Wilcoxon rank sum test.*



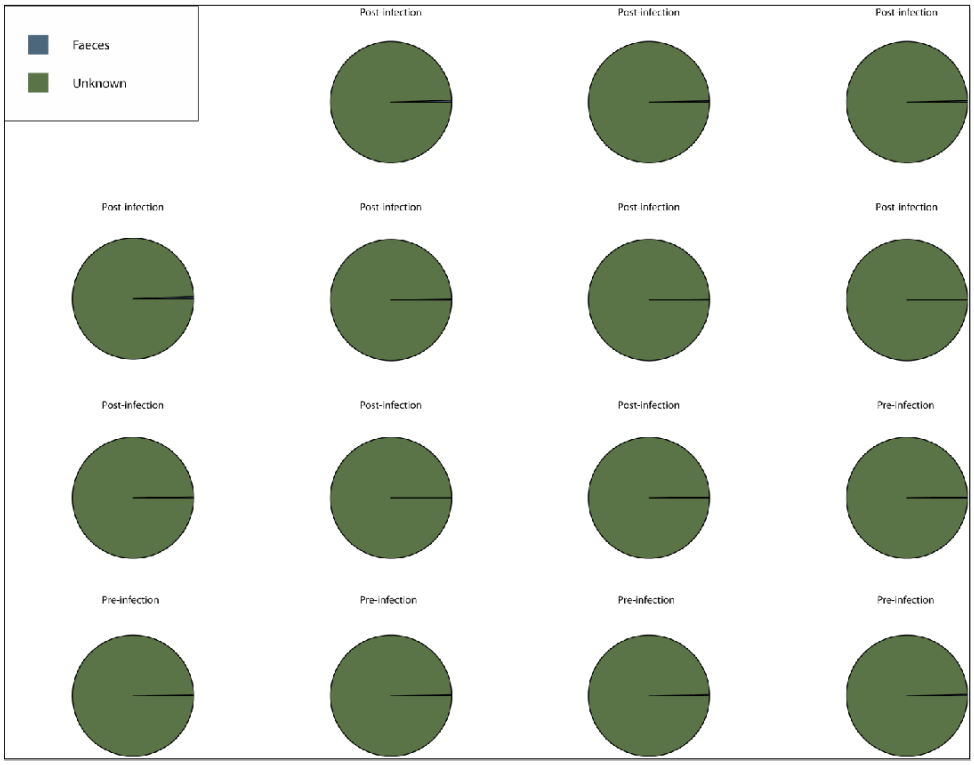
Supplementary Figure 8*: Differential abundance of ASVs between the adult nematodes H. contortus and T. circumcincta.* *ASVs for H. contortus were derived from five pooled samples of 100 nematodes each (4 male, 1 mixed sex) and ASVs for T. circumcincta were derived from seven pooled samples of 100 nematodes each (5 females, 1 male, 1 mixed sex). Differential abundance was calculated based on log fold change, as calculated by the Deseq2 algorithm.*



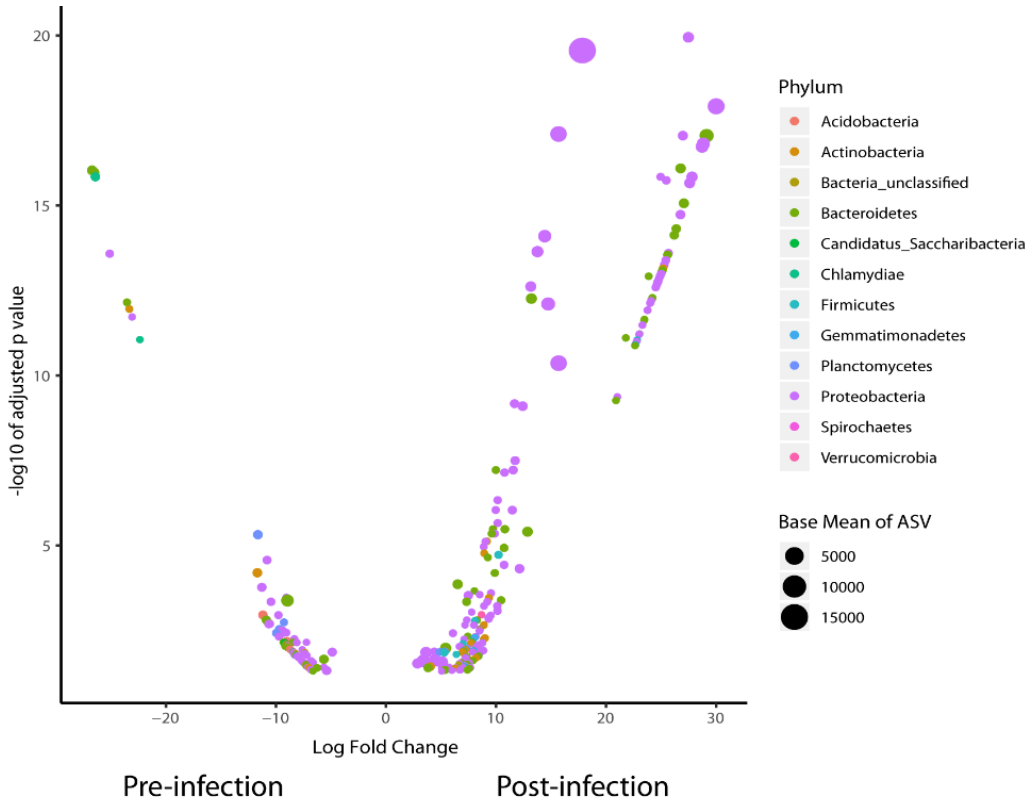
Supplementary Figure 9*: Comparison of beta diversity of the ovine intestinal microbiome, measured by Bray-Curtis dissimilarity vs time.*

RSV	Rho	P-Value	Taxonomic Classification
Seq_0000012	-0.612822668	0.041512087	Bifidobacterium
Seq_0000164	-0.63529918	0.028053108	Sharpea
Seq_0000165	0.672763897	0.013049163	Candidatus_Saccharibacteria
Seq_0000233	-0.750100801	0.001467203	Alistipes
Seq_0000285	0.716508809	0.00431197	Clostridium_XlVa
Seq_0000293	-0.608394632	0.045367563	Ruminococcaceae
Seq_0000333	-0.604781366	0.048434515	Oscillibacter
Seq_0000344	-0.618958441	0.037830096	Firmicutes
Seq_0000376	-0.612193612	0.041512087	Clostridiales
Seq_0000405	0.604485117	0.048434515	Coprococcus
Seq_0000422	0.671515292	0.013049163	Bacteroidetes
Seq_0000437	0.62254652	0.036104329	Methanobrevibacter
Seq_0000439	-0.726809941	0.00313266	Phascolarctobacterium
Seq_0000466	0.621438138	0.036104329	Clostridiales
Seq_0000475	-0.755543316	0.001467203	Eubacterium
Seq_0000513	-0.707314361	0.005075083	Coprococcus
Seq_0000592	-0.778678407	0.001170284	Firmicutes
Seq_0000633	0.633160551	0.028348035	Candidatus_Saccharibacteria
Seq_0000675	0.623367405	0.036104329	Candidatus_Saccharibacteria
Seq_0000765	-0.754736167	0.001467203	Clostridiales
Seq_0000817	-0.702479012	0.005119297	Bacteroidetes
Seq_0000995	0.648093682	0.024258278	Candidatus_Saccharibacteria
Seq_0001009	-0.616117602	0.040100468	Ruminococcaceae
Seq_0001160	-0.635736011	0.028053108	Candidatus_Saccharibacteria
Seq_0001321	-0.612739102	0.041512087	Ruminococcaceae
Seq_0001323	0.637714342	0.028053108	Lachnospiraceae
Seq_0001725	-0.702594839	0.005119297	Ruminococcaceae
Seq_0001746	-0.638452433	0.028053108	Firmicutes
Seq_0001748	0.709341508	0.005075083	Erysipelotrichaceae
Seq_0001952	-0.6457356	0.025011923	Oscillibacter
Seq_0002026	-0.73999151	0.001974592	Ruminococcaceae
Seq_0002167	-0.684390866	0.009720514	Clostridiales
Seq_0002216	-0.658066155	0.020216589	Firmicutes
Seq_0002652	-0.675537108	0.012956815	Clostridiales
Seq_0002679	-0.648216405	0.024258278	Clostridiales
Seq_0002857	-0.621972625	0.036104329	Candidatus_Saccharibacteria
Seq_0002920	-0.637523731	0.028053108	Firmicutes
Seq_0004953	-0.634513897	0.028053108	Ruminococcaceae
Seq_0005973	-0.653430855	0.022467297	Ruminococcaceae

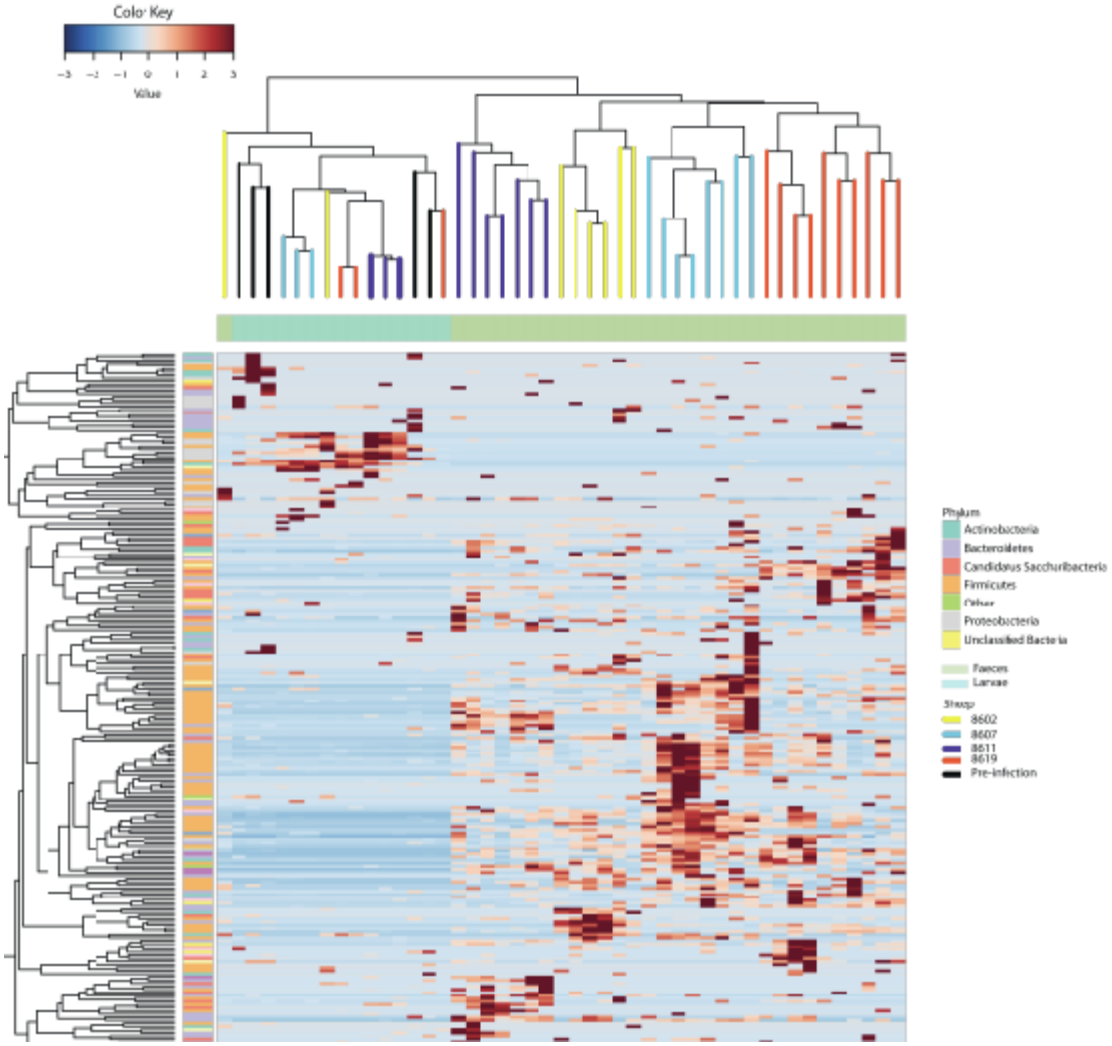
Supplementary Figure 10*: Statistically significant ASVs in the ovine faecal microbiome when correlated against time.* *ASVs were derived from 33 faecal samples taken from four lambs across 10 timepoints. ASVs present in lamb faecal samples were correlated against time using the Spearman correlation method. Results were corrected for multiple testing using the Bonferroni method.*



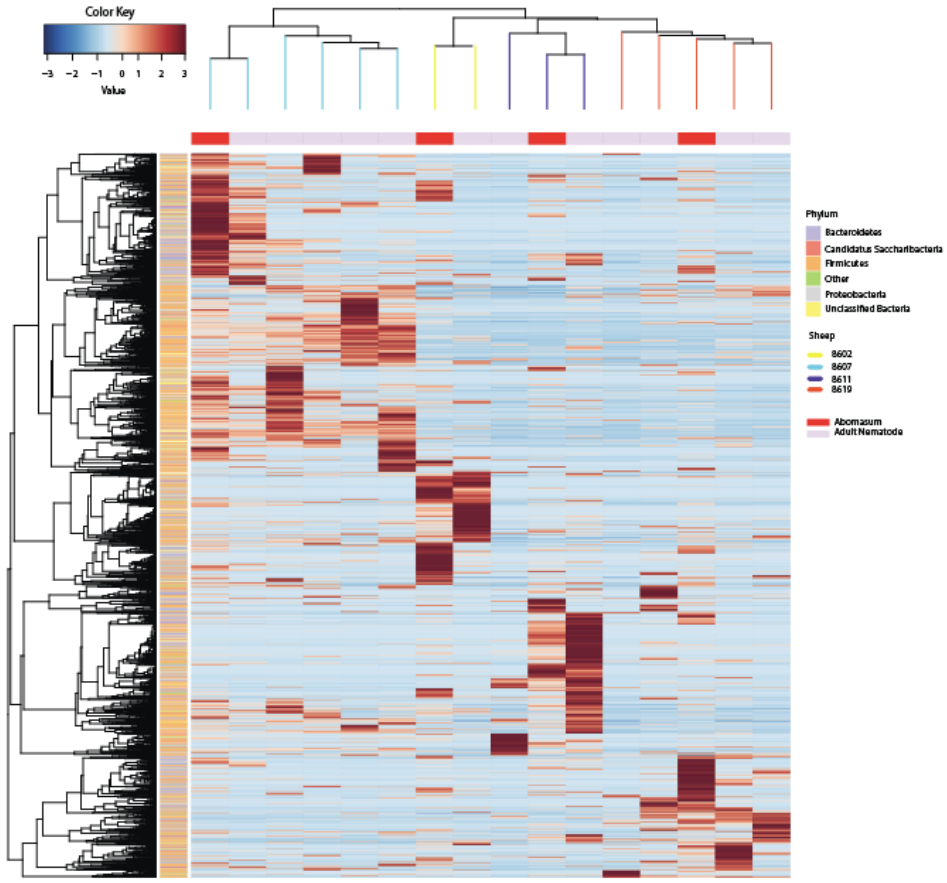
Supplementary Figure 11*: Proportion of bacteria in the larval nematode microbiome potentially originating from host (ovine) contamination.* *Analyses were carried out using the SourceTracker algorithm.*



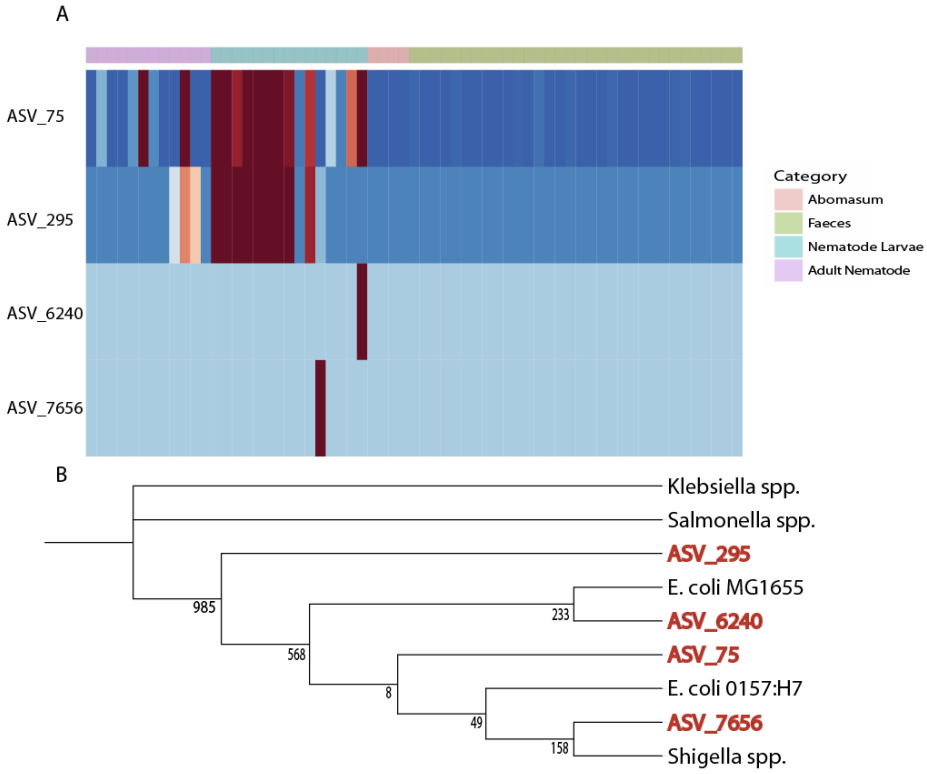
Supplementary Figure 12: ***Differential abundance of ASVs between pre- and post-infection nematode larvae.*** Each point on the plot is derived from a pooled mixture of ~10,000 nematode larvae. *Differential abundance was calculated using Deseq2; points are coloured according to phylum, and size is scaled to the base mean of ASVs.*



Supplementary Figure 13*: Hierarchical clustering of ASVs present both in ovine faeces and nematode larvae.*



Supplementary Figure 14*: Hierarchical clustering of ASVs present both in the ovine abomasum and the adult nematode.*



Supplementary Figure 16*: Differential abundance of E. coli/Shigella* spp. *between the adult nematode and ovine abomasal lumen microbiomes and phylogenetic analysis.* *(A) ASVs for the 'Adult Nematode' were derived from from 12 pooled samples of 100 nematodes each (five H. contortus (4 male, 1 mixed sex) and seven T. circumcincta (5 females, 1 male, 1 mixed sex) samples); ASVs for the 'Abomasum' were derived from the abomasal washings of four lambs. Each row is individually scaled from dark-blue to red, with dark-blue indicating that the ASV in question is absent, and red indicating the most abundant sample for that ASV. (B) Dendogram of ASVs compared with V3-V4 regions of significant strains of E. coli, Shigella, Klebsiella and Salmonella sequences to provide evolutionary context.*

Chapter VI

Discussion and future prospects

This research consists of novel approaches and protocols to broaden the scope of research material for microbiome research, and aid in reproducibility. This is followed by two approaches attempting to progress the translational aspects of this research.

Given the potential impact of environmental contamination on sequencing projects such as those seeking to characterise tumour-related microbiota, a robust contamination control strategy with accompanying validation such as that shown in Chapter 2 is the minimum requirement before proceeding any further with research of this kind. This methodology proved to be effective when used on samples showing mild to moderate levels of contamination, but was unable to completely differentiate low abundance bacterial reads from low abundance contaminant reads in the presence of overwhelming levels of contamination as in the FFPE breast tumour samples of Chapter 4. It is clear that there is room for improvement for retrospective removal of environmental contamination using bioinformatic methods. Future work to achieve this could involve the use of an ensemble classification method such as RandomForest to accurately identify contaminant reads from within datasets with the assistance of biological standards and use them as a training set with which to query newly generated data for contamination.

Unfortunately, the advancement of retrospective contamination removal does not address the larger issue of the inefficiency caused by environmental contamination, in terms of discarded sequencing reads. The cost of sequencing is constantly be sinking, but it remains an expensive proposition for many labs globally. Given the extent of contamination observed in some samples across this research thesis, the practical effect of this is a doubling of the cost of sequencing at a minimum. The results of either an ensemble based classification method, or a biological standard such as the protoblock described in Chapter 4 to identify contamination can be used to inform biologists of contaminant sequences which could then be targeted by amplification blocking oligonucleotides (1), effectively removing them from the PCR pool, as discussed earlier.

Justified criticisms of previous strategies for characterising the microbial communities within tumours employed by a number of groups, have provided a reasonable doubt regarding the presence of endogenous bacteria within tumours. A

new survey of the potential breast tumour microbiome was required, taking into account the various concerns raised in recent times, particularly the presence of contamination. This re-affirmed the suggested presence of a bacterial community in both malignant and non-malignant breast samples, possibly due to migration of bacteria inhabiting the skin. This is evidenced by the similarity between skin swabs, healthy adjacent samples and tumour samples. Additionally, a distinct microbial signature within tumour samples was detected using a variety of statistical methods.

In addition to the strain level analysis of bacterial communities required for biomarker discovery discussed earlier, which has the potential to progress the effectiveness of bacterially administered therapeutics, future work could focus on sequencing a matching host genomic and transcriptomic profile to complement the metataxonomic profile acquired here. This would allow the integration of information about possible biotransformation of therapeutics by bacteria (2), effect of bacterial community on host gene expression (3) and variation in reponse to treatment due to genetic profile (4) into one consolidated profile, progressing personalised cancer treatment.

Chapter 4 describes a multifaceted approach to address two significant issues which are arresting the progress of microbiome research, particularly into non-GIT based ecological niches where stool cannot be used as a proxy;

- Access to samples, particularly healthy controls is restricted by the invasive nature of the sampling process. This makes it difficult to obtain a sufficient sample size for statistical significance.
- Lack of representative biological standards to validate experimental accuracy.

This chapter describes the first method for the extraction of DNA from FFPE samples that is tailored to the unique characteristics of bacterial DNA, and while the method does require further optimisation, the initial results are promising. Many of these results are derived using a novel biological standard to complement FFPE samples. This “protoblock” was designed with the increasing requirements of validation and reproducibility within microbiome research in mind and offers significant extensibility over the current industry leading Zymo mock communities,

when applied to FFPE samples. Two examples of this are the inclusion of host DNA, and formalin fixing to mirror the conditions faced by DNA in the samples.

Future work from a bioinformatic perspective must be to provide a definitive answer on the applicability of this method to low biomass samples such as tumour tissue samples. Given the log fold decrease in DNA quantity expected after formalin fixation, and the low levels of bacteria expected in tumour biopsies (particularly in relation to the levels of host DNA), it is entirely possible that no endogenous bacterial community remains in levels high enough to be differentiated from contamination. Recent studies successfully characterising bacteria in several other tumour sites from FFPE samples using the gold standard Qiagen kit dictate that this method should also be examined with the FFPE samples used here. The results of this could be compared with improvements made to our own method outlined previously.

Despite appearing to be a thematic outlier, the aim in Chapter 5 of characterising and exploiting the bacterial community within a foreign body for drug development or treatment purposes is as relevant to human tumours as it is to parasitic nematode infection, and can be seen as a proof of concept. The result, a bacterial taxonon only found in nematodes and absent from the surrounding host, with deliberate colonisation validated *ex vivo* would represent the ideal result in any metagenomics survey of a human tumour. That being said, the work is significant in its own right as a potential counter measure to rising anthelminthic resistance.

Without ignoring the effect nematode infection has on livestock globally, future work should focus on determining whether the results of this study are replicable in a human model. Nematode infections affect up to 50 % of the global population (5), of which 450 million are seriously ill as a result. Only 125,000 deaths each year are directly attributable to nematode infection, compared to 3 million caused by malaria which affects a similar demographic, but this a function of the reporting methods as nematode infection related morbidity is not directly fatal. A more accurate metric is disease affected life years (DALYs). 39 million DALYs are lost each year from nematode infection compared to 35.7 million from malaria (6). The practise of only reporting deaths directly attributable to nematode infection, and the fact that almost all cases are in the developing world, has led to nematode infections being termed

neglected diseases from an awareness, research and funding perspective. As in livestock, anthelminthic resistance is developing in human nematode infection (5) and new treatment strategies are essential.

Work thus far has centred on the accurate characterisation of ecological niches, with the hope of identifying bacterial biomarkers. If a candidate is found, a possible therapeutic intervention is to engineer the candidate bacterium to produce a therapeutically useful biomolecule such as a protein at the site of required intervention. These proteins are rarely in their native form, rather an aggregation of several functional subunits, and as most proteins only have marginal stability, these modifications can destroy their function. There are numerous *in silico* options for the prediction of protein function, as has been discussed previously, and the crux of Appendix I was the development of a method for their integration to provide a unified score that could be associated with the expected performance of a candidate protein. This in turn would be expected to significantly streamline the design and testing stages of biomolecular design. At present the tool is limited by the relatively small number of predictive features used, and the relatively narrow scope of the experimental validation. Future work must try and broaden the applicability by validating the tool in different experimental conditions, which could be accelerated through the construction of a database for community-based reporting of experimental results following publication.

On a more general note, the field of bioinformatics is facing a potential critical juncture. The length of the individual reads being sequenced by single cell methods such as Oxford Nanopore are constantly increasing, while the total reads yielded by the newest Illumina ultra-high-throughput methods are also increasing rapidly. The result is data generation exceeding the rate at which the processors needed to analyse them are advancing. This comes at a time when bioinformaticians are incorporating more computationally intensive machine learning algorithms into the analysis of sequencing data, and biological research in general begins to incorporate more bioinformatics in the scientific method. This machine learning driven research in particular favours scalability, and already breakthroughs have been facilitated by large technology corporations who have both the machine learning expertise and processing power in abundance. A prime example of this being the unprecedented improvement of Googles' "alpha-fold" protein modelling algorithm when compared

to the best efforts of academic institutions and in a fraction of the time. The involvement of companies such as Amazon and Google has precipitated scientific breakthroughs that may have taken years otherwise, but steps must be taken to preserve not for profit university research in the future.

Given that one of the aims of this research project is to advance the cause of precision medicine, this provides a fitting example. A recently published review in Nature states that although people of European descent account for only 16% percent of the global population, they represent almost 80% of the sequenced genetic information available. It is safe to assume the same holds true for metagenomic information (7). This is unsurprising, as it presumably roughly mirrors the funding for scientific research. Although this imbalance is beginning to be adjusted by projects such as the GenomeAsia 100K project (8), it is difficult to see how more private involvement in academic research through institutions such as Google will help to address this issue rather than exacerbate it.

References

1. Vestheim, H., Deagle, B.E. and Jarman, S.N. (2011) Application of blocking oligonucleotides to improve signal-to-noise ratio in a PCR. *Methods Mol Biol*, **687**, 265-274.
2. Lehouritis, P., Cummins, J., Stanton, M., Murphy, C.T., McCarthy, F.O., Reid, G., Urbaniak, C., Byrne, W.L. and Tangney, M. (2015) Local bacteria affect the efficacy of chemotherapeutic drugs. *Scientific Reports*, **5**, 14554.
3. Nichols, R.G., Peters, J.M. and Patterson, A.D. (2019) Interplay Between the Host, the Human Microbiome, and Drug Metabolism. *Human Genomics*, **13**, 27.
4. Roell, K.R., Havener, T.M., Reif, D.M., Jack, J., McLeod, H.L., Wiltshire, T. and Motsinger-Reif, A.A. (2019) Synergistic Chemotherapy Drug Response Is a Genetic Trait in Lymphoblastoid Cell Lines. *Frontiers in Genetics*, **10**.
5. Stepek, G., Buttle, D.J., Duce, I.R. and Behnke, J.M. (2006) Human gastrointestinal nematode infections: are new control methods required? *Int J Exp Pathol*, **87**, 325-341.
6. Chan, M.S. (1997) The global burden of intestinal nematode infections--fifty years on. *Parasitology today (Personal ed.)*, **13**, 438-443.
7. Genetics, N. (2019) Genetics for all. *Nature Genetics*, **51**, 579-579.
8. Wall, J.D., Stawiski, E.W., Ratan, A., Kim, H.L., Kim, C., Gupta, R., Suryamohan, K., Gusareva, E.S., Purbojati, R.W., Bhangale, T. *et al.* (2019) The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*, **576**, 106-111.

Appendix I

Function2Form Bridge – Towards synthetic protein holistic performance-prediction

A version of this chapter has been published as:

Function2Form Bridge – Towards synthetic protein holistic performance-prediction

VVB Yallapragada[*], **Sidney P Walker**[*], Ciaran Devoy, Stephen Buckley, Yensi Flores, Mark Tangney; *Proteins: Structure, Function, Bioinformatics (2019);* PMID: 31589780

**Contributed equally*

ABSTRACT

Background Protein engineering and synthetic biology stand to benefit immensely from recent advances in *in silico* tools for structural and functional analyses of proteins. In the context of designing novel proteins, current *in silico* tools inform the user on individual parameters of a query protein, with output scores/metrics unique to each parameter. In reality, proteins feature multiple 'parts'/functions, and modification of a protein aimed at altering a given part, typically has collateral impact on other protein parts. A system for prediction of the combined effect of design parameters on the overall performance of the final protein does not exist.

Aim Function2Form Bridge (F2F-Bridge), attempts to address this by combining the scores of different design parameters pertaining to the protein being analysed into a single easily interpreted output describing overall performance.

Methods The strategy comprises 1. A mathematical strategy combining data from a myriad of *in silico* tools into an OP-score (a singularsingle score informing on a user-defined overall performance); 2. The F2F-Plot, a graphical means of informing the laboratory biologist holistically on designed construct suitability in the context of multiple parameters, highlighting scope for improvement.

Conclusion F2F predictive output was compared with laboratory data from a range of synthetic proteins designed, built and tested for this study. Statistical/machine learning approaches for predicting overall performance, for use alongside the F2F plot, were also examined. Comparisons between laboratory performance and F2F predictions demonstrated close and reliable correlations.

This user-friendly strategy represents a pivotal enabler in increasing accessibility of synthetic protein building and *de novo* protein design.

INTRODUCTION

Proteins are large biomolecules, which perform various fundamental functions of life. The structure and function of these biomolecules is defined by a sequence of amino acids, which begin to fold into a 3D structure even during the protein's synthesis by a ribosome [1]. Our understanding of this process, and consequent ability to modify and engineer proteins, has progressed dramatically in recent times. This has gone hand in hand with the development and improvement of computational tools designed to predict how proteins will behave. Tools exist allowing the user to predict:

- the three dimensional structure of a protein [2]
- its physical and chemical properties [3]
- how it interacts with other proteins [4]
- which active sites facilitate these interactions [5]
- with more expert use, tools also exist to modify these proteins if any of these parameters do not match what is desired in the rapidly expanding field of *de novo* protein design [6].

These tools offer a considerable advantage over the traditional structural exploratory techniques of NMR and CryoEM in terms of cost and ease of use, and the gap in terms of accuracy between the gold standard and *in silico* approaches is shrinking.

The design of a protein involves defining the overall desired function, and associating this with a 3D structure. This is in turn coded into an amino acid sequence. In many cases, this overall function is achieved by fusing different sub-functional protein components (parts) together. In addition to *de novo* protein design, a simple example is the conjugation of protein therapeutics with delivery factors, such as cell penetrating peptides to enhance their efficiency. Once the construct has been defined, the typical process of protein modification or design for therapeutic use entails designing thousands of variant structures in order to find the small minority of these proteins that will be (i) expressed by the bacterial cellular machinery and (ii) in the correct conformation to carry out the desired function. Following this, optimal candidates are selected and validated in a wet-lab setting against the pre-defined overall function.

Proteins, whether natural or designed, have broad applications across many fields of science, but medical research in particular has benefitted from our increase in understanding and the research and development of protein based therapeutics has been a clear beneficiary of this. The scope of protein-based biopharmaceuticals is broad, but it is dominated by humanised monoclonal antibodies, which made up 48 % of the therapeutic proteins market in 2010. A fundamental problem with these antibodies has been their size - at 150 KDa on average, they are often too large to bind to the desired active sites, or efficiently penetrate into host tissue targets such as tumours. With the advances in *in silico* capacity, groups are now isolating only the active site from these antibodies and fusing them with much smaller and more stable backbones, or in other cases dispensing with naturally occurring antibodies entirely and simply reverse engineering *de novo* antibodies based on the requirements of the active site [7].

A key problem with these *in silico* mediated advances in protein analysis and design, is that they remain simulations of how the protein will fold, or bind to an active site for example. The overall performance prediction problem is how to weigh the positive and negative effects of modifications to a synthetic protein in such a way that the effectiveness of the construct in experimental conditions can be predicted. Advances in the *in silico* prediction of overall protein performance have to the potential to yield considerable savings both in time and money by reducing the amount of wet-lab testing and validation that must go in to the production of a novel protein for the first time. Given all the disparate data now available through *in silico* protein analysis, the potential for a big data approach to attempt to predict the function of these proteins warrants attention.

In this work, a novel mathematical strategy (F2F-Bridge) aimed at predicting the overall performance of a synthetic protein is proposed. Several test sequences were designed for a defined overall function. The individual scores for all the different design parameters pertaining to each test sequence are condensed into a graphical output. The result is a visual and numerical evaluation of the test sequence. The graphical output (F2F-Plot) and the numerical evaluation (OP-score) together form a novel mathematical strategy (F2F-bridge) that scores, ranks and predicts the overall

performance of the given set of test sequences. This method combines user input with *in silico* data to give insights into the predicted overall performance of a test sequence. With view to eventually developing a robust tool for protein performance prediction, relationships between *in silico* and laboratory data for test proteins were also examined using two different strategies for feature selection and predictive model building: LASSO and regression-based decision trees implemented with the RandomForest algorithm.

MATERIALS AND METHODS

Laboratory work was performed by other members of the Tangney lab.

In silico design of test sequences:

Luminescence proteins: Each construct was designed to have a luminescent domain, a binding domain, a solubility tag and a secretion signal. All parts are linked in all possible permutations using different rigid and flexible linker sequences[8] (Figure 1)**.** Variable heavy and light chain AA sequences from different antibodies were used as the binding domains, from an antibody targeting either cell surface associated epithelial mucin 1 (MUC1; mammalian antigen) or Clumping factor A (ClfA) of *Staphylococcus aureus* (bacterial antigen). Test sequences were designed to bind to their respective target and present luminescence as a readout (bound protein luminescence). Fluorescence proteins (used for further validation) are described in Supplementary Text 3.

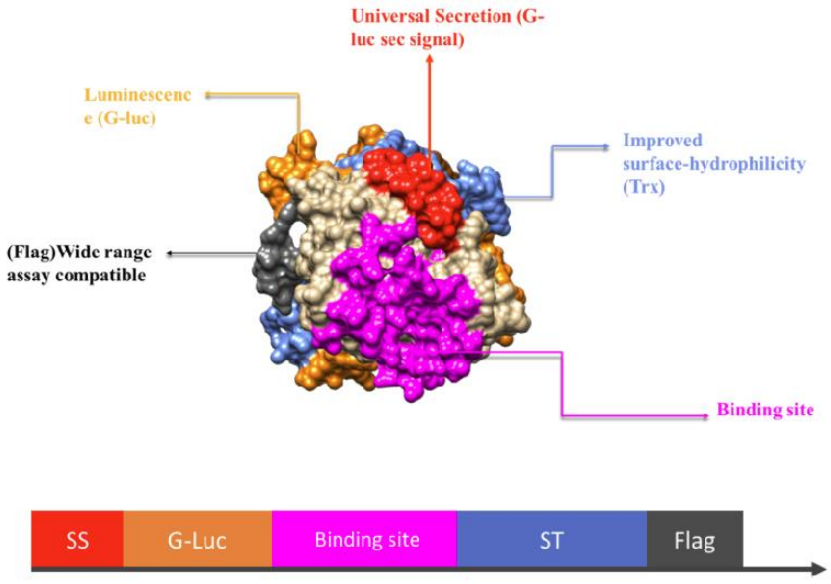


Figure 1: *3D structure of the designed luminescence test sequences examined showing various sub-function parts for a defined overall function.*

Data Generation

The different *in silico* features analysed in relation to the overall performance of the test protein, and how they are generated is outlined in Table 1, with more detailed instructions found in Supplementary Text 1.

Table 1: ***Common in silico tools, and the purpose they serve***

Design Parameter	**Metric and Scale range**	**Metric description**	**Effect on overall performance**	**Generated:**
3D structure	C-score 2 : -5	Confidence in the model and folds predicted	Higher confidence reflects more reliable model	I-Tasser [9]
Docking	ΔG kcal/mol	Gibbs free energy released by reaction	Protein-protein interactions at the active site predicted	Autodock Vina*[10]
Quality of model	RC – score	Proportion of amino acids in different regions based on steric hindrance	High agreement with stereochemistry and free energy reflects stability of structure	Saves server[11]
Active site solvent accessibility	0-9	A measure of the exposure of A residue or group of residues	Depending upon the function, the active site could be exposed to the solvent or 'hidden' inside the core	R (Using I-TASSER output)/GET AREA server[12]
Surface Hydrophobicity	-4.5 to +4.5	Each amino acid has a hydrophobicity score between -4.5 and +4.5 as per Kyte Doolittle scale.	Ensuring ideal surface hydrophobicity aids solubility	R (Using I-TASSER output)
Size	kDa	Total weight of the protein	Size forms an important factor if the protein is required to cross/penetrate membranes and biological barriers	ProtParam Hosted by Expasy[13]
Isoelectric point	pH 0 to 14	Point at which molecule carries no	The integrity of the structure of the	ProtParam Hosted by

		net charge	protein in a setting is influenced by the isoelectric point	Expasy[13]
Potential active sites	0 to n	Number of potential active sites	Predicting the potential active sites on the designed protein informs on potential off-target effects	Coach Server
Instability	0 to 100	Half life of protein *in vitro*	Gives an indication of the viability of the protein	ProtParam Hosted by Expasy[13]

Function2Form Bridge

The data described in the data generation stage were taken as input into the Function2Form function, written in the R programming language by the author. The F2F function takes as input a data frame with all proteins to be screened as rows, and the different *in silico* observations as columns. The first row of the table contains a set of user desired input values. Some of these are based on the benchmarks provided by *in silico* programmes such as RC score and Instability, while in other cases the user can specify the ideal needs for the protein (e.g. hydrophobic and < 30 kDa). The features used in this study are detailed in Table 1, but the F2F-Bridge function is not limited to these and can be easily expanded or condensed depending on individual user requirements.

The test sequences are then screened by the function and the output is a data frame of the test sequences and their respective scores, along with a radar plot for each test sequence, highlighting the differences between each sequence and the input parameters as can be seen in Figure 5. Areas where the candidate protein does not meet the preset requirements will appear outside the coloured region of the reference values. As well as this visual analysis of the suitability of the protein, a score indicating overall function is provided. The OP score is the grand average of absolute distance between each particular feature of the protein, and the user-specified/program specified reference range. For cases of high throughput *in silico* screening, an additional function allows the user to extract the *n* top scoring test sequences from the overall selection.

F2F R function

Generating the OP score

(i) Where possible, convert different *in silico* observations to same scale

$$i = \left(\frac{(O - O_{min})}{(O_{max} - O_{min})}\right) * (N_{max} - N_{min}) + N_{min}$$

Where O is the old range and N is the new range, which in the case of the F2F function is always 0-100.

(ii) F2F function then iteratively scores each test sequence supplied in input table

$$OP\ Score\ of\ x = \frac{\Sigma|xi - yi|}{n}$$

Where *x* is the test sequence to be scored, *y* is the set of reference values, *i* refers to the *ith* observation within the *in silico* data table supplied to the algorithm, and *n* is the total number of observations *i.*

Generating the F2F plot

Figure 3 shows the graphical output of the F2F function. Each sequence tested is assigned an OP score as discussed above, and a radar plot is generated. For every *in silico* observation provided in the input data, an axis is created on the radar plot. This enables the user to see which specific *in silico* feature or features are making the test sequence unfit for purpose.

Statistical and Machine learning methods

Two other methods of transforming the *in silico* data into a prediction of overall performance were assessed. The aim was to examine the data for relationships of any kind between the predictive features and the laboratory output with the view to either design a system of weights for the predictive features to improve the F2F plot, or to create a new predictive tool to be used in conjunction with the F2F Bridge.

i) *LASSO* Feature selection and subsequent generation of predictive models can be used in conjunction with the F2F-Bridge programme. LASSO regression puts a constraint on the sum of the absolute values of the model parameters, which must be less than a fixed upper limit. It does this by applying a regularisation process (shrinkage) where it penalizes regression coefficients and shrinks a selection to zero. Variables that still have a non-zero coefficient after the shrinking process are selected to be part of the final model [14] .

ii) *TREE BASED METHOD* The second method was to use a regression tree as per the random Forest package in R to generate a variable importance plot, again using the *in silico* parameters as input, and laboratory detected luminescence as the indicator of overall performance. An outline of how random forest works in generating these regression trees is as follows. A predefined number of bootstrapping samples are drawn from the original data. For each of these samples, an "un-pruned" regression tree is grown. Traditionally, the best split at each node to differentiate all predictors would be used, but in this instance the best split is found amongst a random subset of the predictors. Following this, predictions are made by aggregating the predictions of the pre-defined number of trees and taking the average value[15]. The quality of the model was ensured by finding the optimal number of features to randomly sample at each split, and to ensure that enough iterations of the model are run to ensure that the out of bag error has stabilised.

Laboratory validation

The luminescence test synthetic proteins examined are outlined in Figure 3. Two biological facets were used to assess the effectiveness of the functional prediction strategies – i) binding; ii) secretion. Sub-function parts on the test sequences include: (i) **Active site:** Heavy and light chains of anti-MUC1 antibody (C595) and anti-ClfA antibody were fused with EAAAK (rigid) and GGGGS (flexible) linkers to obtain Monospecific bivalent diabodies and Monovalent ScFVs (monobodies), (ii) **Secretion signal:** Gaussia luciferase's native secretion signal, (iii) **Solubility enhancer:** SUMO tag, (iv) **Reporter:** Truncated version of Gaussia luciferase was used as a luminescence reporter. (v) **Detection tag:** Flag peptide was used as a detection tag for downstream assays.

Presence or absence of certain sub-function parts or their design orientation has a significant effect on the overall performance of the protein and should be accounted carefully in the design phase. In our case, over 50 different amino acid sequences were designed against each target. Of these, eight variants per target were synthesised for testing in the laboratory. These test sequences vary in (a) (+/-) solubility enhancer, (b) (+/-) and positioning of Active site and (c) the type/format of Active site. All these test sequences were tested for their overall performance. Laboratory data was used to validate and improve the results from the F2F-Bridge. An outline of the laboratory workflow can be seen in Figure 2, and a more detailed description on synthesis and build of 'test sequences' can be found in Supplementary Text 3.

Data generation from laboratory experiments with luminescence proteins

Binding assays: 10^8 *Staphylococcus aureus* TCH959 (naturally bearing *clfA)* or 10^6 MCF7 cells (naturally bearing MUC1) were blocked with 5% BSA for 2 h followed by incubation with supernatant containing each test construct. Cells were washed 3 times and resuspended in PBS. Luminescence was measured using Promega GloMax® 96 luminometer. In our case, since bound luminescence is the overall function, the luminescence readings corresponding to each test sequence are recorded and used for validating and improving F2F Bridge.

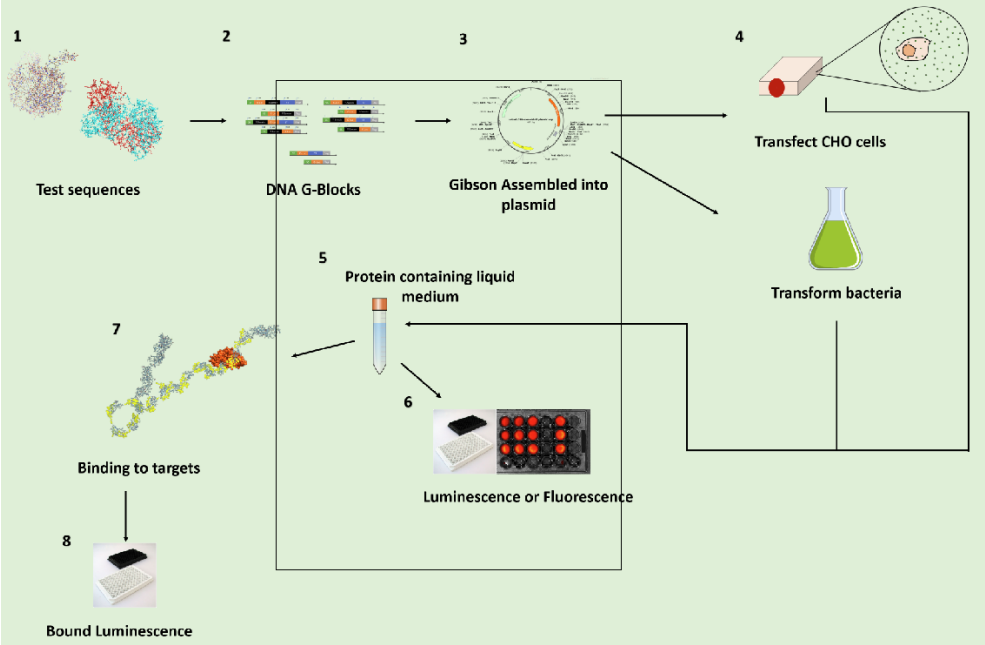


Figure 2: ***Workflow of laboratory validation of test sequences.***

Statistical analysis

All statistical testing, unless otherwise stated, was performed in the base R environment v3.4.3 [16]. The LASSO regression feature selection method was implemented using the Glmnet library v2.0-16 [17], and the Random Forest regression tree analysis was performed using the RandomForest library v4.6-14 [18]. The radar plot within the F2F-bridge function was implemented with the fmsb library, v0.6.3 [19]. Visualisation was carried out using the ggplot2 package,v3.1.1 [20].

RESULTS

Overall Biological Performance - Bound luminescence

16 test amino acids sequences were scored using F2F-plot. The result of F2F-plot analysis and associated scores for all test sequences can be seen in Figure 3 and Supplementary Figure 1, and a detailed workflow of the strategy can be found in Supplementary Text 2, additionally the raw data can be found in Supplementary Table 1. It can be seen from the output of the F2F-bridge that the *ClfA Monobody 2* test sequence (represented by the pink shaded region) has an instability index score that is far higher than the user required level (blue shaded region), but there is minimal difference between the different shaded regions across the other axes, which contributes to *ClfA Monobody 2* having a low and therefore good score. The test sequence predicted to have the poorest overall performance, *ClfA Diabody 2*, has levels of solvent accessibility and docking affinity that are much lower than required, as well as an increased instability index. These factors combine to give this test sequence the highest and therefore worst score.

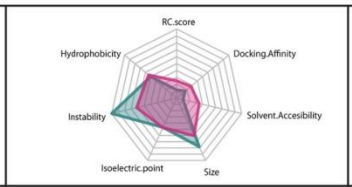
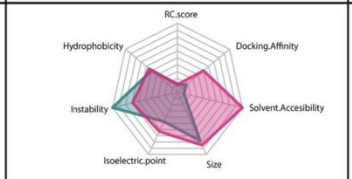
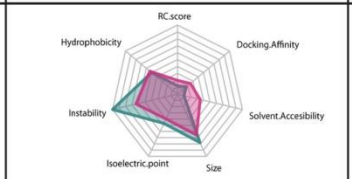
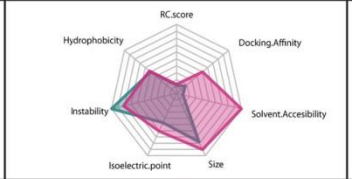
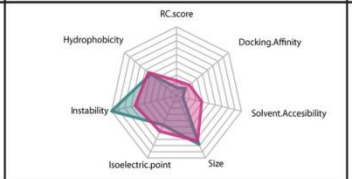
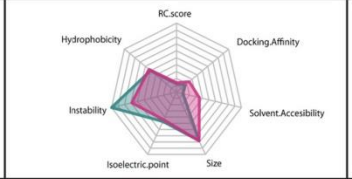
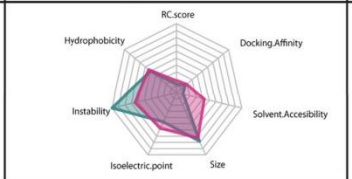
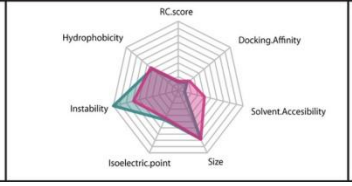
Test Sequence	F2F Result	F2F Score
ClfA Diabody 1		6.88
ClfA Diabody 2		9.48
ClfA Diabody 3		6.03
ClfA Diabody 6		9.27
ClfA Monobody 1		5.62
ClfA Monobody 2		3.62
ClfA Monobody 3		5.29
ClfA Monobody 4		3.8

Figure 3: ***F2F-bridge output for ClfA test sequences.*** *Areas where the candidate protein (pink) does not meet the preset requirements is highlighted by contrasting with the coloured region of the reference values (blue). The plot is generated using a bespoke function written in R, and detailed instructions on its use can be found in supplementary materials, with a link to the github repository containing the code.*

These sequences were used to generate the corresponding proteins as outlined in the Methods section. Biological performance of these proteins was assessed in the laboratory using binding assays with luminescence as the readout, corresponding to Overall Performance.

Table 2 shows the results (laboratory – luminescence units; F2F – OP score) of all test sequences whose biological performance was predicted with F2F-Bridge. The accuracy of F2F prediction was assessed by comparing the F2F-prediction of overall biological performance with the laboratory luminescence data.

Table 2: ***Agreement between experimental results and F2F plot - Luminescence.***

A

WetLab output	F2F-Plot Score	Test Sequence
2788	4.69	Muc1 Monobody 2
96232	5.04	Muc1 Monobody 4
106712	5.73	Muc1 Monobody 3
4914	5.79	Muc1 Monobody 1
2156	6.68	Muc1 Diabody 1
2436	6.9	Muc1 Diabody 3
1597	9.27	Muc1 Diabody 6
30	9.48	Muc1 Diabody 2

B

WetLab output	F2F-Plot Score	Test Sequence
9302	3.62	ClfA Monobody 2
15179	3.8	ClfA Monobody 4
26519	5.29	ClfA Monobody3
90901	5.62	ClfA Monobody 1
45542	6.03	ClfA Diabody 3
9507	6.88	ClfA Diabody 1
209	9.27	ClfA Diabody 6
1110	9.48	ClfA Diabody 2

In both antiMuc1(A) and Anti-ClfA(B), the binding affinity of each protein is measured by luminescence output. Proteins are ranked by their OP score in both tables, and coloured from green (best performing protein), through yellow, to red (worst performing protein) for both luminescence and OP score (Lowest number = best OP score.)

Overall, the correlation pattern showed the F2F-Bridge method providing a general guide for how the test sequence can be expected to perform. In a database of this limited size, there was no statistically significant correlation between the OP-score and the 'bound protein luminescence' evident. This was repeated with 'luminescence' as the laboratory output, but again there was no relationship present (Data not shown).

Investigation of alternative methods to complement F2F plot

1) LASSO Regression

'Bound luminescence' as overall performance

The output of Lasso analysis of the test sequences for bound protein luminescence can be seen in Figure 4. In the case of test sequences against MUC1, the features deemed to have the most effect on bound protein luminescence were Docking Affinity, Hydrophobicity, Solvent accessibility and isoelectric point. However, when these predictive features were input into a linear model, no linear relationship was detected. The same was found when analysing the potential relationship between the test sequences against ClfA and their eventual bound protein luminescence. In this case, Hydrophobicity and Instability were the features selected, and again, no linear relationship was found.

Despite the LASSO method detecting possible relationships between the predictive features and the test variable (bound protein luminescence), no predictive linear model could be constructed (Table 3). We speculate that there may be too many variables present from a laboratory perspective for us to accurately predict bound luminescence with a database of this size.

Table 3: *Results of multiple regression analysis of features selected by Lasso regression analysis against experimentally determined luminescence.*

Test sequences	***p*-value**	**Adjusted R Squared**
antiClfA	0.507	-0.06
antiMUC1	0.867	-0.6

'Secreted luminescence' as overall performance

To counter this, we also examined 'luminescence' only resulting from protein secretion as a measure of overall biological performance. As we are just measuring the degree of luminescence of the test sequences, and not their binding to a target, the two groups (antiMUC1 and antiClfA) can also be directly compared, doubling our sample size. Figure 4 uses the LASSO method to investigate the relationship between *in silico* observations and luminescence due to secretion of anti–Clfa test sequences. *In silico* observations implicated in dictating the level of secretion of a test sequence were identified by LASSO. These were then used to generate a linear model to examine the degree to which they explained the luminescence due to secretion of the test sequences. This gave a linear model that explained 84.6% of the variability in luminescence of all test sequences, with an associated p-value of 0.004. This led us to explore the utility of a LASSO dictated linear model as a predictive tool. The luminescence levels predicted by the model were correlated with the experimentally determined luminescence levels. This, and also correlation coefficients of individual test sequence groups against their luminescence, are shown in Figure 6, and numerically in Table 4. As would be expected, antiClfA test sequences which are the training set, show stronger correlation (Rho 0.93), but that of the antiMUC1 test sequences was also significant.

2) Alternative methods for prediction of test sequence performance; Random Forest Regression Tree analysis

The same methodology was used for a regression tree implemented within random Forest. The test sequences against ClfA were used as the training data set, and the test sequences against MUC1 test sequences were used as the test set. The regression model derived from the random forest algorithm was able to explain 41% of the variability in the luminescence of the training set data test sequences (Figure 6). The values predicted by random forest for luminescence for the individual proteins were then correlated with their experimentally derived levels of luminescence (Table 5). This method showed a significant correlation between the luminescence values predicted by the random forest algorithm, and the experimental values. Random

forest regression analysis was carried out with bound protein luminescence as the laboratory output, but no significant results were found (data not shown).

Table 5: *Results of various predictive models of secreted luminescence generated using RandomForest regression trees, when correlated with the lab-generated values.*

	antiClfA	**antiMuc1**	**Total**
p value	0.002	0.05	1e-5
Rho	0.92	0.71	0.87

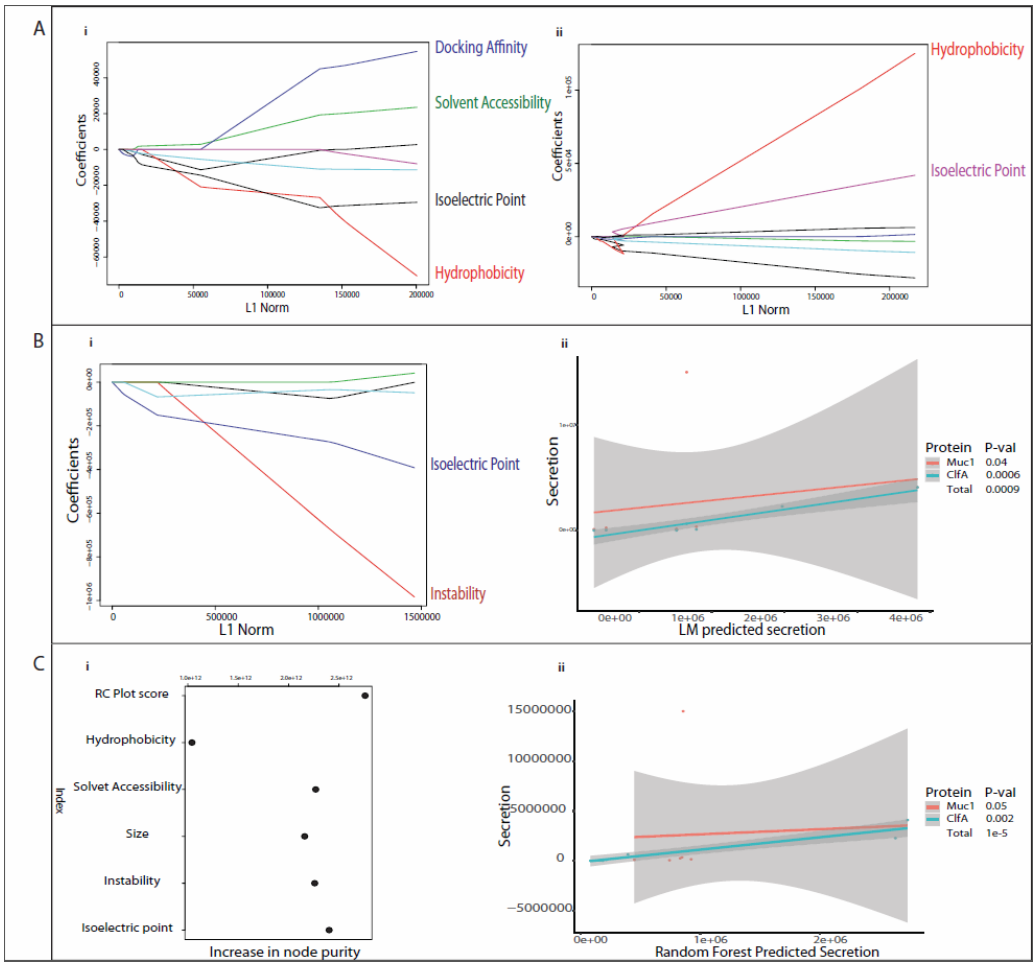


Figure 4: ***Output of F2F Post Hoc Analysis***

(A) *shows the graphical output of LASSO features selection for both* ***(i)*** *antiMuc1 and* ***(ii)*** *antiClfA test sequences for predicting* ***bound protein luminescence****. Each coloured line corresponds to a predictive feature used in the F2F-Bridge function. The lines plot the path of the variables coefficient against the L1-norm, of the whole coefficient vector as lambda varies. Both show that Lasso regression analysis was capable of identifying relationships between the examined predictive features, and experimentally determined luminescence.*

(B)(i) *shows the same LASSO based feature selection for antiClfA test sequences, using* ***secreted luminescence*** *as output. In this instance, Isoelectric point enters the model first, and Instability appears to have the most pronounced effect on the test variable. The relationship between these and the experimentally derived luminescence was tested with a multiple linear regression. Both features were found*

to be significant in the model, which explained 84.61% of the variability in luminescence, with an associated p-value of 0.004. This allowed us to build a predictive model using antiClfA test sequences as the training set. ***(ii)*** *Shows a correlation plot of luminescence values for test sequences predicted by a LASSO directed linear model, vs experimentally derived secreted luminescence values. The training set (antiClfA test sequences) is coloured blue, and the test set (antiMUC1) is coloured red.*

(C) *summarises the Random Forest regression tree analysis. As with* ***(B)*** *the successful model was trained on the antiClfA data, and tested on the antiMUC1 data.* ***(i)*** *shows mean node purity for each predictive feature. The lower this value, the more important it is to the model. The model, trained on the antiClfA test sequences was able to explain 41% of the variability in experimentally determined secreted luminescence. The model was then used to predict secreted luminescence values of the training set (antiClfA) and the test set (antiMUC1), and these predicted values were correlated with experimentally derived luminescence values in figure* ***(ii)****. The overall correlation coefficient was 0.87, with an associated p-value of 1e-05, indicating the value of the predictive model generated.*

We postulate that with an increased amount of data available for training the models, this accuracy can only increase. A summary of all the tests performed and their outcomes is shown in Table 6.

Table 6: ***Summary of statistical tests performed.***

In silico **Test Performed**	**Biological Performance Tested**	**Test Sequence**	**Result**
"Blind" F2F Bridge	Binding and Luminescence	antiClfA and antiMUC1	The F2F plot was able to provide a guide for the expected performance of the test sequence when the test sequences were ranked by OP score and by laboratory output, and the accompanying plot was able to inform on how to improve the test sequence. No statistically significant relationship between OP score and laboratory output could be found.
LASSO feature selection and linear model **building**	Binding	antiClfA and antiMUC1	LASSO regression analysis was able to detect discrete patterns in the data, showing Hydrophobicity and Isoelectric point both to have a positive relationship with bound luminescence in antiClfA. In the case of antiMUC1 Docking Affinity and Solvent accessibility were shown to have a positive effect, Isoelectric point and Hydrophobicity a negative one.
Using LASSO regression analysis dictated linear model as a **predictive tool**	Binding	antiClfA and antiMUC1	The models predicted in the above analysis were unable to exaplain any of the variability in the bound luminescence of antiMUC1 or ClfA test sequences.
LASSO feature selection and linear model **building**	Luminescence	antiClfA	LASSO regression analysis was able to detect discrete patterns in the data, a linear regression with solvent accessibility and instability was able to explain 86.4% of the variability in luminescence in the antiClfA samples.
Using LASSO regression analysis dictated linear model as a **predictive tool**	Luminescence	antiClfA and antiMUC1	The model created in the above test was used to predict luminescence values for both antiClfA and antiMUC1. In both cases these predictions showed strong positive correlations with the experimental luminescence

			values which were statistically significant.
Random Forest regression tree model **building**	Luminescence	antiClfA	A regression tree implemented with randomForest was able to explain ~41% of the variability in the luminescence of antiClfA test sequences.
Using Random Forest regression tree as a **predictive tool**	Luminescence	antiClfA and antiMUC1	The model created in the above test was used to predict luminescence values for antiClfA and antiMUC1 test sequences. In both cases, these predictions showed strong positive correlations with the experimental luminescence values that were statistically significant.

Further Validation – Fluorescence Proteins

The F2F plot was further validated on a second dataset of 8 test sequences and resulting laboratory data (fluorescence readings from proteins). In this instance there was no need for the secondary functionality of feature selection and model building based on a subset of the data, as the OP score predicted showed a statistically significant inverse correlation with overall performance of the test sequence. These results are presented in Figure 5 and Table 7, and the accompanying F2F plots, scores and raw data are shown in supplementary Figure 2 and Table 2. In this case, fluorescence was the overall function of the 8 test sequences to be predicted. Unlike the dataset discussed previously, in this case, the F2F-plot predicted scores showed a strong inverse correlation with the overall performance of the proteins, which is the desired result.

WetLab Output	F2F-Plot Score	Test Sequence
1.52E+09	4.92	Fluorescence Construct 5
1.05E+09	5.61	Fluorescence Construct 2
3.73E+07	6.43	Fluorescence Construct 6
2.91E+07	6.73	Fluorescence Construct 3
3.06E+08	8.67	Fluorescence Construct 1
2.85E+08	10.68	Fluorescence Construct 8
2.52E+07	12.58	Fluorescence Construct 4
1.88E+07	14.35	Fluorescence Construct 7

Best Performing — Worst Performing

Table 7*: Agreement between experimental results and F2F plot - Fluorescence.*

Proteins are ranked by their OP score, and coloured from green (best performing protein), through yellow, to red (worst performing protein) for both Fluorescence and OP score. (Lowest number = best OP score.)

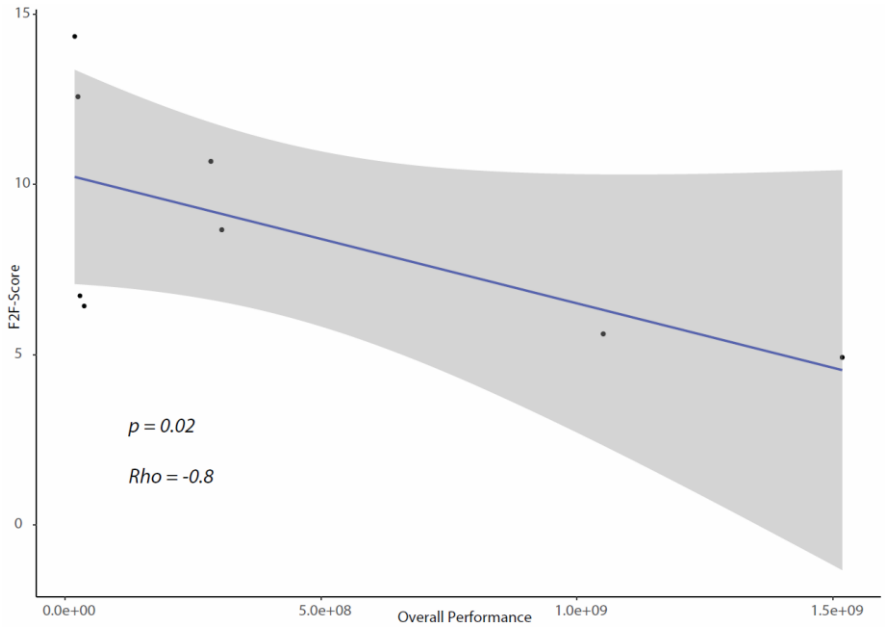


Figure 5*: Correlation plot of OP-Score vs Overall Biological Performance.*

Overall biological performance was scored for fluorescence. Correlation test carried out using Spearman's method.

DISCUSSION

In this study, we have shown how F2F plot could help to visualize the overall performance of a test sequence, particularly if complemented by the statistical methods examined. For each sequence, the OP score predicts the expected efficacy, and the F2F plot provides a graphical overview of predicted strengths and weaknesses. Unlike the pre-existing *in silico* tools that inform the quality of individual design parameter, F2F bridge takes a top down approach on overall performance by predicting the collective influence of all the design parameters on given test sequence. Such a holistic outlook on the overall performance holds a key for informed protein design. F2F bridge could be used either for low throughput design (see Table 8), accounting for the 'pitfalls and merits', in the design corresponding to a particular test sequence, or for high throughput *in silico* screening by comparing and ranking a set of test sequences.

Table 8*: The two associated workflows of the F2F Bridge*.

Scale of use	Outcome
High Throughput	A database of test sequences or extant proteins of known sequence can be queried with the F2F-bridge scoring each test sequence and identifying those most suitable.
Low Throughput	On a protein by protein basis the F2F-bridge provides a graphical overview of the relationship between the features of the test sequence and the optimal values specified by the user, informing the user on how to improve the test sequence.

By informing the end user (laboratory biologist) with performance predictions, F2F plot and OP score together aim to ultimately bridge the gap between easily generated, seemingly abstract *in silico* values, and experimental results.

When combined with downstream testing, we have shown that patterns exist in the easily generated *in silico* data that can be used to generate predictive models.

F2F Bridge An unweighted and unsupervised combination of features deemed likely to have an effect on biological performance showed promising results. The F2F-Bridge is able to give an early indication of the expected performance of a test sequence. Given the ease of implementation, in comparison with laboratory experiments, any information provided about candidate test sequences prior to synthesis is extremely valuable. As well as the OP-score provided, the accompanying radar plot can also highlight any design aspects of the test sequence that diverge from what is required as per the user input parameters. We expect a considerable improvement in performance of both the F2F-Bridge and associated models, with an expanded dataset, in the meantime further work to refine or build on the method was carried out.

Feature selection driven linear models with LASSO

LASSO regression was used to search for patterns in the data that could help predict the level of luminescence due to a test sequence binding to its target. Features that have an effect on bound luminescence were identified. When viewing the plots of normalised Lambda1 vs coefficients, what is important is the point at which the predictive feature enters the model, and the effect it has on the dependent variable. In this case, it was impossible to incorporate them into a statistically significant linear model to predict any of the variability in bound protein luminescence. We speculate that, with a database of this small size, there were too many different processes involved from a laboratory perspective for the strategy to accurately predict the final outcome (steps in cell expression of a given test sequence, protein binding to target, luminescence production). For this reason, and the opportunity to increase the sample size, we also analysed the secreted luminescence data. This was much more successful in terms of improving on the F2F-Bridge.

LASSO regression-based feature selection was used again, to look for patterns between *in silico* observations of the antiClfA test sequences and their experimentally determined secreted luminescence [14] . This was more effective. We have shown that a linear model can be predicted using LASSO feature selection that can predict values for luminescence that correlate strongly with experimentally determined values. This functioned retrospectively to find patterns between *in silico* features of antiClfA test sequences and their eventual levels of secreted luminescence, it can also function prospectively. Once the linear relationship between a subset of the *in silico* features and luminescence has been established for antiClfA, this was used to successfully predict luminescence values in a test set of antiMUC1 proteins which correlated with the experimental values. The antiMUC1 data originated from a different experiment, and a different class of proteins to the antiClfA test sequence training set, so the fact that the linear model still has predictive power is extremely encouraging.

Random Forest regression trees

The random forest regression tree model, given the same *in silico* features as the LASSO regression, was able to explain ~41.01% of the variability in secreted luminescence within the antiClfA test sequence dataset. The predicted secreted luminescence values generated by the regression tree model significantly correlated with the experimentally derived secreted luminescence values. On a group by group basis, it is extremely encouraging that, as with the LASSO based method previously, a random forest regression model trained on the antiClfA test sequences was then able to predict luminescence for the test set (antiMUC1 test sequences) that correlated significantly with the values derived experimentally [15]. Future work, with an expanded database would involve assessing whether these associations identified with the two above methods become stronger as the database size increases, and also, if this method of predicting overall biological performance holds true for other tasks, opening up the possibility of the design of an accurate prognostic tool for test sequence performance.

Relevance to the laboratory scientist

It is important to frame these results in the context of the difference in cost between *in silico*- and laboratory-based screening. *In silico* screening requires a fraction of the time, money or expertise of laboratory-based screening, so maximising the value

of this data can lead to considerable operational savings. Laboratory experimentation is often a prolonged process in biological research. In most cases, the data from laboratory assays has to be processed/filtered to observe the intended correlations between the experimental aims.

We have shown that *in silico* predicted protein attributes can play a significant role in optimising the design and production of protein constructs. With a larger sample size, we expect the aforementioned methods to become more accurate, and therefore call for the establishment of a community wide database for sharing both *in silico* and experimental data so that this can be incorporated into a much larger training data set applicable to a variety of biological functions. Inspiration came, in part, from the SourceTracker algorithm used in metagenomic studies to track possible sources of contamination in HTS studies. This involved establishing a database of known contaminants, used by the algorithm to refine the search for contaminant bacteria[21] . We hope that the establishment of a database of potential test sequences used in other research laboratories, combined with their in silico parameters and in the case of those that are synthesised, their biological output, would similarly improve the accuracy of the predictions of the F2F-Bridge. To expedite the process of database formation we plan to launch a server in the coming months which will take as input an amino acid sequence, and perform all of the necessary calculations, and will also accept one or more measures of overall biological function, in the hope that fellow researchers will submit their published data. Both linear models (selected with lasso regression) and tree based methods (random forest regression trees) can further assist in the prediction of "overall biological performance". With larger datasets, we hope to develop one or more of the following strategies:

- A method of applying weights to the features in the F2F-Bridge based on the outputs of the two models previously mentioned
- Develop a distinct predictive tool based purely on one or both of these methods
- Design a protocol suited to large scale projects whereby an initial subset of the test sequences are generated and screen experimentally and *in silico*, by the F2F-Bridge itself, or with a combination of the three methods outlines in this paper. The resulting information could then be fed back into the design protocol to refine this process, increasing the success rate of constructs.

Conclusions

The design-build-test-learn approach of synthetic biology stands to benefit immensely from this method. laboratory assays to test multiple test sequences demand a huge amount of resources, time and human effort. In such a situation, Function2Form becomes an indispensable strategy for a biologist to visualise and improve a given test sequence or to triage potential best performers by scoring and ranking the test sequences. Integrating F2F-Bridge into the ‘learn’ step aids user empowerment by providing a laboratory biologist with a holistic readout on the overall performance of the protein. With a community-based data reporting system and larger data sets, the accuracy of the F2F-Bridge could be tuned to Pareto optimality.

References:

1. Anfinsen, C.B., *The formation and stabilization of protein structure.* The Biochemical journal, 1972. **128**(4): p. 737-749.
2. Yang, J. and Y. Zhang, *Protein Structure and Function Prediction Using I-TASSER.* Current protocols in bioinformatics, 2015. **52**: p. 5.8.1-5.8.15.
3. Gasteiger, E., et al., *ExPASy: The proteomics server for in-depth protein knowledge and analysis.* Nucleic acids research, 2003. **31**(13): p. 3784-3788.
4. Trott, O. and A.J. Olson, *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* Journal of computational chemistry, 2010. **31**(2): p. 455-461.
5. Sloan, D.B., et al., *Detecting Rare Mutations and DNA Damage with Sequencing-Based Methods.* Trends in biotechnology, 2018. **36**(7): p. 729-740.
6. Kaufmann, K.W., et al., *Practically useful: what the Rosetta protein modeling suite can do for you.* Biochemistry, 2010. **49**(14): p. 2987-98.
7. D.S, D., *Therapeutic Proteins*, in *Therapeutic Proteins. Methods in Molecular Biology (Methods and Protocols)*. 2012: Humana Press, Totowa, NJ.
8. Chen, X., J.L. Zaro, and W.C. Shen, *Fusion protein linkers: property, design and functionality.* Adv Drug Deliv Rev, 2013. **65**(10): p. 1357-69.
9. Yang, J., et al., *The I-TASSER Suite: protein structure and function prediction.* Nat Methods, 2015. **12**(1): p. 7-8.
10. Alhossary, A., et al., *Fast, accurate, and reliable molecular docking with QuickVina 2.* Bioinformatics, 2015. **31**(13): p. 2214-6.
11. UCLA. *SAVES server*. 2019; Available from: http://servicesn.mbi.ucla.edu/SAVES/.
12. Fraczkiewicz, R. and W. Braun, *Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules.* Journal of Computational Chemistry, 1998. **19**(3): p. 319-333.
13. Wilkins, M.R., et al., *Protein identification and analysis tools in the ExPASy server*. Methods Mol Biol, 1999. **112**: p. 531-52.
14. Tibshirani, R., *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society, 2996. **58**(1).
15. Liaw, A. and M. Wiener, *Classification and Regression by RandomForest*. Vol. 23. 2001.
16. Team, R.C., *R: A language and environment for statistical computing.* . R Foundation for Statistical Computing, 2017.
17. Jerome Friedman, T.H., Robert Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 2010. **33**(1): p. 1-22.
18. M.Wiener, A.L.a., *Classification and Regression by randomForest.* R News, 2002. **2**(3): p. 18-22.
19. Nakazawa, M., *Functions for Medical Statistics Book with some Demographic Data*. Pearson Education Japan. 2007.
20. Wickham, H., *ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York, 2009.
21. Knights, D., et al., *Bayesian community-wide culture-independent microbial source tracking.* Nature Methods, 2011. **8**: p. 761.

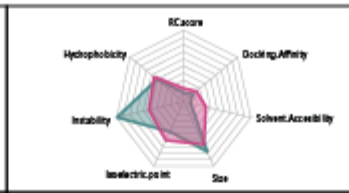
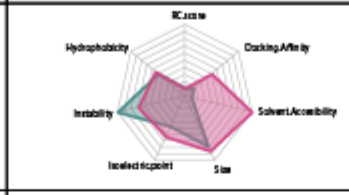
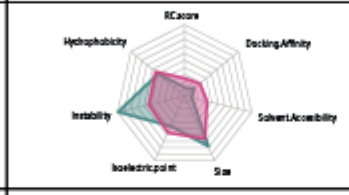
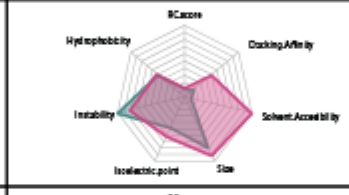
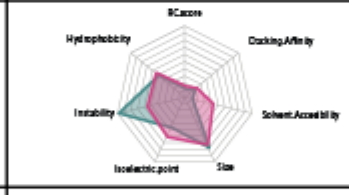
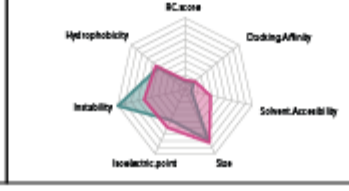
Supplementary Material

Supplementary Text 1:

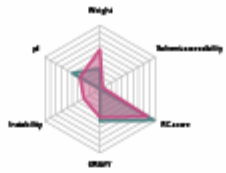
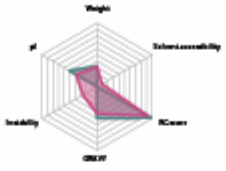
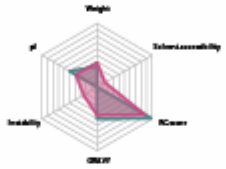
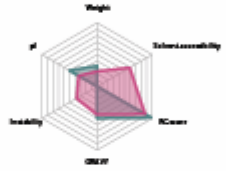
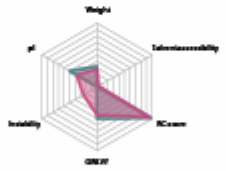
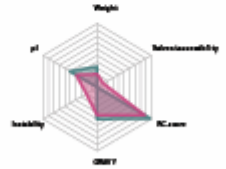
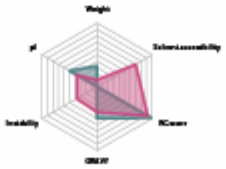
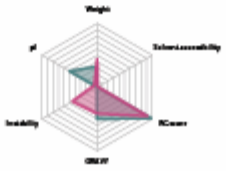
Calculation of *in silico* parameters

Many *in silico* features, including those of *Molecular Weight*, *Theoretical pI*, and *Instability Index* used in this study, are calculated using the ProtParam facility, hosted by expasy. This web-server takes as input only the amino acid sequence and does not require any further user engagement. As these are all calculated based on the amino acid sequence they can all alternatively be calculated with simple scripts in R or Python, as has been done with *Grand Average of Hydropathicity* (see R script in Github repository). The protein tertiary structure prediction was generated by the I-TASSER suite (v5.1), this tool also provides a file detailing the per residue *solvent accessibility*. This can be subset in R to find the accessibility of the active site. If I-TASSER is not used, online tools for solvent accessibility of particular residues exist, such as the GETAREA tool, hosted by the Sealy Center for Structural Biology. The Ramachandran plot is generated on the Saves Server, using the Verify3D utility.

The protein-protein interaction or "Docking" was modelled using a heuristic implementation of the Autodock Vina algorithm, Qvina2, within the MGLTools/Autodock Tools (v1.5.6) interface. The number of potential active sites within a test sequence was calculated using COACH, which is another algorithm within the ITASSER suite.

Test Sequence	F2F Result	F2F Score
Muc1 Diabody 1		6.68
Muc1 Diabody 2		9.48
Muc1 Diabody 3		6.9
Muc1 Diabody 6		9.27
Muc1 Monobody 1		5.79
Muc1 Monobody 2		4.69
Muc1 Monobody 3		5.73
Muc1 Monobody 4		5.04

Supplementary Figure 1*: F2F-bridge output for MUC1 test sequences*

Test Sequence	F2F Result	F2F Score
Construct 1		8.67
Construct 2		5.61
Construct 3		6.73
Construct 4		12.58
Construct 5		4.92
Construct 6		6.43
Construct 7		14.35
Construct 8		10.68

Supplementary Figure 2: ***F2F-bridge output for fluorescence test sequences***

Supplementary Text 2

General description of F2F bridge workflow.

1) **Collection of data**

The protein scientist must identify the features considered important for the analysis, from the literature or from experience. The predictive features to be included in the analysis must be converted to the same scale.

2) **Preparation of data**

The data must be stored in a table in the following format:

-Columns must be the predictive features selected for the experiment

-Rows 4 to end must be the unique names of the test sequences to be analysed

-Rows 1 and 2 must be the minimum and maximum values (Should be scaled 1:100 unless impossible)

-Row 3 should contain the user supplied values either taken from the literature or suited to the experimental conditions

3) **Running the programme and creating the data**

The F2F function takes as input a table prepared in the manner described in step 2 and produces both a plot for each sequence and a data frame containing all sequences and their associated F2F-plot score. The script can be called from the linux command line, or executed within R. For high throughput analysis, the option of generating a plot can be disabled. The data frame of scores will be saved to the current working directory.

4) **Database free mode**

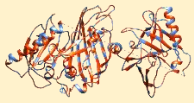
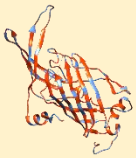
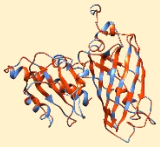
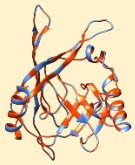
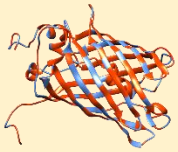
As a database of protein test sequences, their OP-scores, and their overall biological performance is ideally required to apply a system of weights to the predictive features used in the plot, an alternative is provided until such a database can be established. A function for feature selection with LASSO is provided, and can be used to detect relationships between the input *in silico* data and the overall performance on a subset of the experimental data, and the resulting model can then be applied to the remaining data. The user is not restricted to the LASSO function provided, a variety of tools for feature selection and subsequent model building exist, such as RandomForest which was also implemented in the main manuscript.

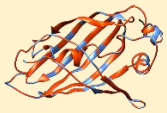
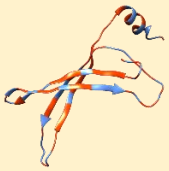
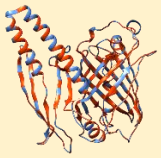
Comprehensive annotated code for F2F bridge can be found at https://github.com/Sidneyw91/F2F-Bridge

Supplementary Text 3:

Fluorescence proteins

Eight synthetic proteins based on the mCerulean Fluorescent Protein, or parts thereof, were generated, corresponding to Supplementary Figure 3.

Fluorescence construct number	3D Model	Construct Description	Wetlab Fluorescence
1		Split mCerulean with modification 1 (docked)	3.06E+08
2		Split mCerulean (docked)	1.05E+09
3		Split mCerulean with chromophore and modification 1	2.91E+07
4		Split mCerulean without chromophore and with modification 1	2.52E+07
5		mCerulean	1.52E+09

6		Split mCerulean with chromophore	3.73E+07
7		Split mCerulean without chromophore	1.88E+07
8		Split mCerulean with modification 2 (docked)	2.85E+08

Protein-related Laboratory Methods

DNA construct design and build: DNA sequences were obtained by reverse translating the amino acid sequences using EMBOSS Backtranseq (https://www.ebi.ac.uk/Tools/st/emboss_backtranseq/). The DNA sequences were codon optimized using IDT codon optimisation tool (https://eu.idtdna.com/codonopt). Each DNA construct was designed with a FLAG-tag and homology arms which were verified for upstream experiments using SnapGene's Gibson Assembly simulator (SnapGene.com).

Gene Block synthesis: Gene blocks for the test constructs were sourced from IDT (Integrated DNA Technologies, Inc) and amplified using corresponding PCR primers. The amplicons were verified using gel electrophoresis (1.5 % agarose) and ImageLab 5.2.1, (Bio Rad Inc) was used for band visualisation.

Primer Design: Primers were designed using Benchling (Benchling.com) to determine appropriate regions for construct amplification, followed by the use of

Primer3Plus to test the primer suitability in terms of appropriate Tm as well as the presence of G-C clamps. NEBuilder assembly tool (www.nebuilder.neb.com) was used to design assembly primers for the purpose of facilitating construct insertion into the plasmid during Gibson Assembly. The finalised primers were obtained from IDT.

Competent E. coli: *E.coli* cells were made competent following the protocol described in Cohen et al. 1972. All cells were stored at -80 ˚C and thawed at room temperature. OG176 (Oxford genetics, mammalian expression vector) was used for amplification and expression of the test sequences with luminescence as the overall function and RSFDuet-1 (Novagen, bacterial expression vector) was used for amplification and expression of the test sequences with fluorescence as the overall function. Both the expression plasmids included Kanamycin resistance gene (Kn^R). For plasmid amplification, the plasmids were transformed into *E. coli* BL21 by mixing 100 ng plasmid DNA into 30 µL of competent cells. The cells were incubated on ice for 20 min and heat shocked by placing at 42˚C for 45 sec. The cells were then placed on ice for a further two min. The cells are then suspended into 500 µL of LB, 100 µL transformed cells were cultured on LB agar supplemented with 50 µg/mL kanamycin and incubated O/N at 37 ˚C. Select colonies were then grown in 20 mL liquid LB with 30 ng/mL kanamycin O/N.

Plasmid Extraction: After suspension in liquid LB supplemented with 30 ng/mL kanamycin O/N, transformed cells were pelleted by centrifugation at 4000 rpm (2500 x g) for 10 min. Following the instructions of the Monarch Plasmid miniprep kit (New England Biolabs) plasmid DNA was extracted, eluted in 15 µL EB and DNA concentration was quantified with a Nanodrop. The eluted samples were stored at -20 ˚C until further processing.

Restriction Digestion: The plasmids were digested by appropriate restriction enzymes (*NcoI, AflII, NdeI, and AvrII*) with the addition of CutSmart reaction buffer (New England Biolabs) and dH2O, for a total reaction volume 50 µL. The sample

was then incubated at 37 ˚C for 1 h, after which time the digestion was confirmed by gel electrophoresis on a 1.5 % agarose gel at 80 V for 90 min. Plasmid DNA was purified using a PCR purification kit (Qiagen) and eluted in 15 μL EB.

Gibson Assembly: The assembly master mix was made up in accordance to the protocols and reagents described by DG Gibson et al 2009. The gene blocks were combined with the plasmid in a DNA concentration ratio of 3:1 in which 72 ng/μL plasmid DNA was incubated in a Gibson Assembly master mix with 225 ng/μL of construct DNA. The mixture was incubated for 1h at 50 ºC followed by transformed into *E. coli* BL21 cells.

Colony PCR: Colony PCR was used to determine the success of the Gibson Assembly and evaluate the transformation of the construct into bacterial cells. In this case, select colonies were added to a PCR master mix containing; 25 μL Q5 polymerase (NEB), 2.5 μL of forward and reverse primers and 20 μL milliQ. Sanger sequencing was then carried out by GATC's light-run service and was verified by aligning with a reference sequence.

Mammalian cell transfection (Luminescence proteins): CHO-K1 (ATCC® CCL-61™) cells were used for luminescence protein production. Turbofect transfection reagent (Cat No: R0532) was used for *in vitro* transfection. Transfection was carried out using manufacturer's protocol and supernatant containing protein collected after 48 h.

Binding assays: 10^8 *Staphylococcus aureus* TCH959 (naturally bearing *clfA)* or 10^6 MCF7 cells (naturally bearing MUC1) were blocked with 5% BSA for 2 h followed by incubation with supernatant containing each test construct. Cells were washed 3 times and resuspended in PBS. Luminescence was measured using Promega GloMax® 96 luminometer.

Fluorescence protein production and bacteria harvesting: Samples were grown overnight in liquid LB with 30 ug/mL kanamycin. 100 ml fresh LB was inoculated with 5 ml of overnight culture. Bacteria were induced with 1 mM Isopropyl ß-D-thiogalactoside at 0.5-0.6. OD. Bacteria were harvested when they reached an OD 0.8. Bacteria were washed and pelleted by centrifugation at 2,500 x g for 10 min. BugBuster lysing buffer supplemented with cOmplete protease inhibitor (Roche) and Lysonase reagent used for bacterial cell lysis according to the manufacturer's protocols. Protein production was confirmed by running an SDS page.

Fluorescence assays: Fluorescence was measured using an Omega Plate Reader (BMG LabTech) and IVIS Lumina II imaging system (Perkin Elmer). Samples were diluted in PBS and transferred to a 96 well plate to measure fluorescence.

Acknowledgements:

I was fortunate to have two supervisors for my PhD, Dr Mark Tangney and Dr Marcus Claesson. I would like to express my gratitude to both of them for their support, guidance and patience over the last three and a bit years. I am also grateful to APC Microbiome for providing the funding that allowed me to undertake this research.

There are many friends and colleagues throughout both APC Microbiome Ireland and Cancer Research @ UCC to whom I am grateful. I would like to thank the members of the Tangney lab who generated all the laboratory data that allowed me to do a PhD in the first place. First and foremost, Yensi Flores, but also Glenn, Vamsi, Stephen, Ciaran and all the other members past and present. I would also like to thank my early bioinformatics mentors when I was still finding my feet, Feargal, Adam and Hugh. They blazed an early path which made it easier for us to follow. As well as these three, I also thank David, Maurice, Tom and Mrinmoy who were a constant source of advice and more importantly distraction over the course of my studies along with all the people whom I shared office 4.11 with over the years, currently Jamie, Shriram, Julia, Jill, Paddy, Rachel and Anna.

To my partner Ruth, who figured out how to get me to apply myself to my studies very early on in our relationship, and was a constant source of advice for the duration, I will always be thankful.

Finally, my thanks go to my family. Their endless support, encouragement and at times even genuine interest, made it all worthwhile, with a special mention to both of my grandmothers, sadly only one of whom will get a chance to read this.