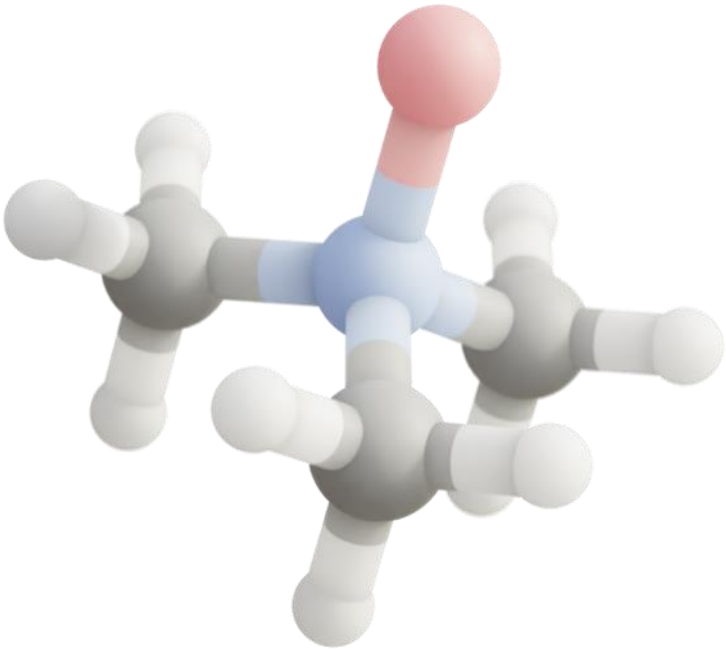
Understanding the metabolism of trimethylamine *N*-oxide in human gut bacteria

Samuel Dawson

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

September 2023



Copyright statement

Abstract

Trimethylamine *N*-oxide (TMAO) is a dietary methylamine that has been implicated in the development of cardiovascular disease and metabolic disease, but also in protective effects on the liver and blood-brain barrier. While some TMAO is directly ingested in foods such as fish and dairy products, most TMAO in the human body is the product of microbial metabolism. In the human gut microbes breakdown TMAO to trimethylamine, which is transported to the liver where it is converted back to TMAO by hepatic enzymes and excreted in the urine. Previous work reported the protein TorA, encoded by *Escherichia* and *Klebsiella* spp., as being the main source of TMAO to trimethylamine metabolism in the human gut. A thorough literature review found several non-TorA TMAO metabolism pathways that could be utilised by gut bacteria. A bioinformatics analysis of ~36,000 genomes found that TorA was not as prevalent as previously thought, with less than 1 % of *Klebsiella* genomes encoding TorA. A caecal *K. pneumoniae* isolate, L4-FAA5, was fully characterised to examine its TMAO metabolism activity. L4-FAA5 was found to carry non-TorA TMAO metabolism proteins and showed significantly increased growth in the presence of TMAO in anaerobic conditions. The molybdoenzyme BisC was found to be highly prevalent in *Klebsiella* spp. in the bioinformatic analysis. BisC was found to have high (63.3 %) amino acid sequence identity with the TMAO reductase TorZ and so work was carried out to characterise its TMAO reductase ability. Bioinformatic analysis showed a high (>90 %) structural similarity between BisC carried by L4-FAA5 and known TMAO reductases. This led to work to clone, express, and purify L4-FAA5 BisC to assess its TMAO reductase ability, with preliminary results suggesting that BisC can reduce TMAO. This work used extensive bioinformatic analysis to inform laboratory work that increases the understanding of microbial TMAO metabolism in the human gut.

Acknowledgements

First of all, I would like to thank my PhD supervisor Professor Lesley Hoyles for her invaluable support and guidance over the last four years. I would also like to thank Dr Jack C. Leo for his assistance with my enzymatic work, and Dr Anne L. McCartney for the collection of growth data and samples in chapter 2. Many thanks as well to the Nottingham Trent University School of Science and Technology for providing me with additional funding so that I could complete my PhD.

I would like to apologise to my friends and family who have had to listen to me complain about this, especially for the last few months! Thank you all for your support throughout my PhD! Extra thanks to Cameron “Operon Art” Baines for sharing in the joys of being a PhD student with me and being of great help when discussing lab work, as well as the creation of the title TMAO image and the addition of annotations to figure 3.2. A great thanks to my parents. Without your care and support throughout everything I wouldn’t have made it this far.

Un beidzot, liels paldies Elīzai par tavu mīlestību un atbalstu. Es to nevarētu paveikt bez tevis.

Contents

Table of Figures

Table of Tables

Abbreviations

ANI, average nucleotide identity

BCFA, branched-chain fatty acid

BisC, biotin sulfoxide reductase

BSH, bile salt hydrolase

CARD, The Comprehensive Antibiotic Resistance Database

CDS, coding sequence

DMSO, dimethyl sulfoxide

FFAR, free fatty acid receptor

FMO, flavin-containing monooxygenase

FPLC, fast protein liquid chromatography

FXR, farnesoid X receptor

GTDB, Genome Taxonomy Database

IBD, inflammatory bowel disease

LPS, lipopolysaccharide

MDP, molybdopterin biosynthesis pathway

MAG, metagenome-assembled genome

MGD, molybdenum guanine dinucleotide

Moco, molybdenum cofactor

MsrP, methionine sulfoxide reductase

MsrQ, inner membrane dehydrogenase

PC, phosphatidylcholine

PCR, polymerase chain reaction

PDB, protein database

SCFA, short-chain fatty acid

SOC, super optimal broth with catabolite repression

TMA, trimethylamine

TMAO, trimethylamine *N*-oxide

TR1, TMAO reductase 1

TR2, TMAO reductase 2

VFDB, Virulence Factor Database

Chapter 1 Introduction

1.1 The human gut microbiota

The human gut is home to a large and varied population of microbes, estimated at over 10^{14} cells and whose metabolic capabilities far exceed those of its host [1,2]. These microbes consist of organisms from all 3 domains of life, as well as a large number of both eukaryotic and bacterial viruses [3]. Viruses are the most prevalent component of the microbiota, followed by bacteria, with fungi, protists, and archaea all accounting for a relatively small portion of the microbiota [3]. The population structure of the gastrointestinal microbiota varies across the human gut because of changing environmental conditions from oesophagus to the anus, with conditions such as pH, oxygen levels, transit time and nutrient availability contributing to microbial diversity (Figure 1.1).

1.2 The stomach

Despite the harsh environment of the stomach actively functioning as a measure to prevent bacterial colonisation, some genera of bacteria are still often present in low numbers, between 10^1 and 10^3 CFU/mL [4–6]. In the stomach pH is between 0.3 and 6.7, depending on how recently food has been consumed and the presence of health conditions such as atrophic gastritis [4,7]. The transit time of stomach contents can also be altered by these two factors, being between 0 and 6 hours [7]. The stomach is dominated by bacteria of the phyla *Pseudomonadota* (formerly *Proteobacteria*) and *Bacillota* (formerly *Firmicutes*). The most notable bacterial inhabitants of the stomach include *Helicobacter pylori*, *Streptococcus* spp., and *Staphylococcus* spp., as well as *Lactobacillus* spp. [8,9]. However, the presence of some non-*Helicobacter* genera may be attributed to their levels in fermented foods (e.g. *Lactobacillus* spp.) and these populations may be transient [9]. Higher levels of microbial diversity also seem to occur only when *H. pylori* is absent or present in low levels in the stomach [9,10]. The reason for this *H. pylori*-mediated decrease in biodiversity is currently unclear, although it has been suggested to be caused by the induction of host antimicrobial peptides by *H. pylori* [11]. In patients with gastric cancer the genera *Lactobacillus*, *Prevotella*, *Streptococcus*, and *Veillonella* were found to dominate the stomach microbiota with *H. pylori* appearing in a relatively low abundance [12]. It is unknown what effect these bacteria

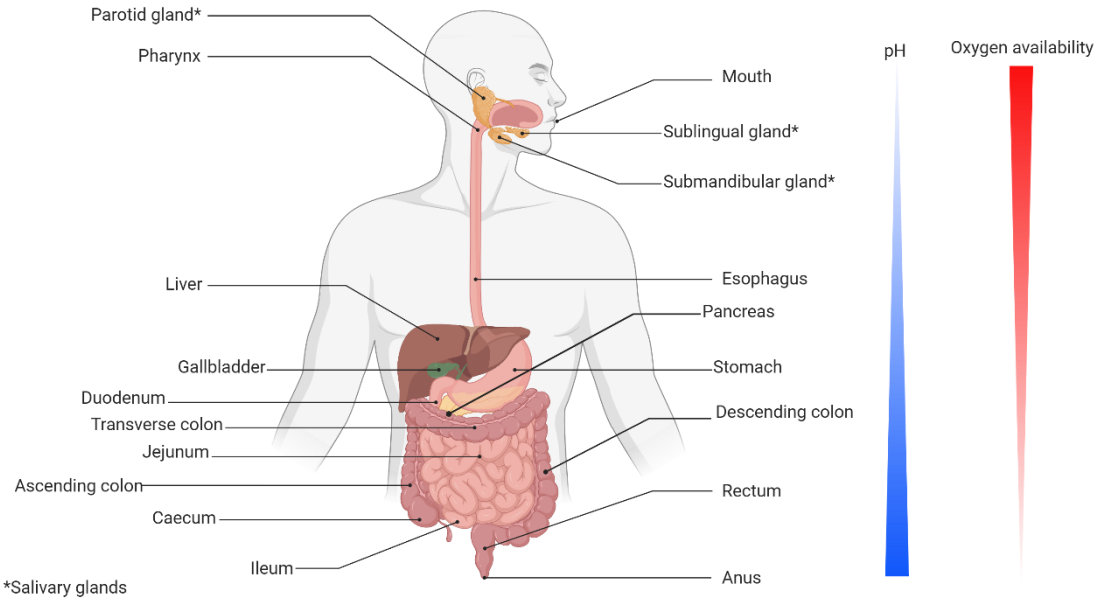


Figure 1.1. The human digestive tract.

From the stomach to the rectum, pH increases through the digestive tract and oxygen availability decreases.

have on the development and progression of gastric cancer. *Streptococcus* spp., *Staphylococcus* spp., and *Lactobacillus* spp. appear to dominate the healthy stomach [9]. Other bacteria that have been cultured from gastric samples include *Bifidobacterium* spp., *Enterococcus* spp., *Propionibacterium* spp., and *Pseudomonas* spp. [5].

H. pylori has been associated with several human health conditions, namely chronic gastritis, peptic ulcers, and gastric cancer. *H. pylori* survives the acidic conditions of the stomach via the production of ammonia from urea, raising the pH surrounding the cell and enabling it to reach the mucus layer closer to the wall of the stomach where the pH is higher [11]. This colonisation of the space near the stomach wall can lead to inflammation, increasing the risk of gastric cancer developing [12]. The presence of *H. pylori* may also have some positive effects on human health however, reducing chances of diarrhoeal disease and gastroesophageal reflux, which can lead to oesophageal cancer [13,14].

Lactobacillus and related spp. are also able to colonise the gastric mucus, producing lactic acid from lactose and lowering the pH, thereby neutralising the ammonia produced by *H. pylori*. The species *Limosilactobacillus fermentum*, *Lactobacillus acidophilus* and *Lacticaseibacillus casei* have been shown to have negative effects on the growth of *H. pylori* both *in vivo* and *in vitro* [11].

1.3 Small intestine

The small intestine is divided into three sections, each with different bacterial populations and environmental conditions [15,16]. These sections are the duodenum, jejunum, and ileum; these areas are difficult to retrieve samples from and so the populations of bacteria in the small intestine are not as well characterised as other sections of the human gut. In the duodenum, conditions remain acidic with a pH between 4.8 and 7 and numbers of bacteria are between 10^2 and 10^4 CFU/mL [7,15]. In the duodenum and the jejunum, the bacterial population is dominated by *Lactobacillus* spp., *Streptococcus* spp., *Staphylococcus* spp., *Veillonella* spp., and *Acinetobacter* spp. [15,17,18]. The ileum has a higher pH, between 6.4 and 8.2, and a lower flow rate with a transit time of 2-8 hours, allowing higher numbers of bacteria to grow, with bacterial loads being between 10^6 and 10^8 CFU/mL [7,15]. The ileum also has a lower oxygen concentration compared to the duodenum and jejunum, allowing for the growth of facultative and obligate anaerobes [18]. This change in conditions causes

groups such as clostridia, *Bifidobacterium* spp., *Bacteroides* spp., and *Fusobacterium* spp. to become more prevalent. Coliforms, such as *Escherichia coli* and *Klebsiella pneumoniae*, also become more prevalent [17–19].

1.4 Large intestine

The large intestine has the greatest number of bacterial cells, between 10^{10} and 10^{11} cells per gram of gut contents, due to the much lower flow rate compared to other parts of the gastrointestinal tract, pH between 5.3 and 8, and high nutrient availability [7,15,20]. This, along with the anaerobic environment of the caecum and colon, leads to the bacterial populations here being largely made up of obligate anaerobes [17,21]. These consist of the clostridia families *Lachnospiraceae* and *Ruminococcaceae* and the *Bacteroidota* (formerly *Bacteroidetes*) families *Bacteroidaceae, Prevotellaceae* and *Rikenellaceae* [17,21,22]. While these obligate anaerobes dominate the microbiota in the large intestine the environment is extremely diverse, with an estimated 500 different species of bacteria present [20].

1.5 Microbiota-host co-metabolism

Bacterial populations in the human gut are highly diverse, as are their genes and metabolic capabilities. The number of microbial genes encoded in the gut microbiota has been found to be over 22 million, dwarfing the 23,000 genes present in the human genome [23]. This massive number of microbial genes means that the metabolic capability of the bacteria in the microbiota far exceeds that of its host, with potentially hundreds of different metabolites in the human body being influenced by the activity of gut bacteria [24]. The complex ways in which the microbiota and its host interact through the production and utilisation of these metabolites is referred to as co-metabolism, with both parties affecting the phenome (i.e. all traits expressed by a cell, tissue, organ, organism, or species) of the other.

The presence of bacteria in rodents has been shown to influence their metabolome (i.e. all metabolites present within an organism, cell, or tissue). Over 100 different metabolites in rats have been shown to have their levels in urine and faeces altered after antibiotic-induced disruption, with those levels recovering alongside the microbiota post-treatment [24]. Key examples of these metabolites are the short-chain fatty acids (SCFAs), which have their levels in rat faeces reduced post antibiotic treatment, with butyrate being reduced

the most. Levels of indoles are also reduced in the faeces and urine of these rats, with the level of their precursor tryptophan being increased. In mice the production of trimethylamine (TMA) from trimethylamine *N*-oxide (TMAO) was found to be greatly reduced when the mice were treated with antibiotics [25]. This effect of the microbiota has also been demonstrated in germ-free mice. When microbes are introduced to germ-free mice an effect can been seen on their metabolomes in as little as 5 days, with these effects increasing over time and the presence of only a single species of bacteria being enough to cause metabolomic differences such as a build-up of dietary glycans and absence of secondary bile acids in germ-free mice [26–28]. Germ-free mice colonised by human microbiota bacteria also see differences in their metabolomes when compared to conventional mice, showing in increase of tryptamine and indoxyl glucuronide in the faeces of colonised mice and a decrease in creatine, creatinine, and trisaccharide when compared to germ-free mice [27].

In humans a reduction of microbiota diversity induced by the food additive carboxymethylcellulose has been shown to have effects on the metabolome, reducing the levels of amino acids detected in faeces[29]. It is not currently understood how carboxymethylcellulose induces these changes. The *FUT2* genotype in humans has also been shown to have effects on bacteria in the gut, altering the species present and the metabolites they produce [30,31]. The *FUT2* gene encodes a fucosyltransferase that fucosylates host glycans present in gut mucus. Individuals lacking a functional *FUT2* gene were found to have a reduced presence of microbial metabolic pathways related to amino acid, cofactor, and vitamin metabolism [30]. Numbers of the SCFA-producing bacteria *Roseburia* and *Faecalibacterium* were found to be reduced in individuals without a functional *FUT2* gene, alongside an increase in numbers of proteobacteria [30]. Abundance of the genera *Parabacteroides*, *Eubacterium*, *Parasutterella*, *Bacteroides*, and the family *Lachnospiraceae* have been found to be raised in *FUT2*-negative mice [31]. There was also an increase in pathways related to carbohydrate and lipid metabolism, and glycan biosynthesis. Changes to the metabolism of the species *Bacteroides thetaiotaomicron* have also been observed in mice that are *FUT2*-negative, with pathways related to fucose catabolism being downregulated [31].

1.5.1 SCFAs

SCFAs are the most common microbial metabolites produced by the gut microbiota from the fermentation of dietary fibres, which are indigestible carbohydrates [32]. This fermentation largely takes place in the proximal side of the large intestine where anaerobic bacteria dominate, with the main SCFAs produced being acetate, butyrate, and propionate [32]. In the gut, bacteria have been found to produce acetate via three different pathways. One of these is the Wood-Ljungdahl pathway, which is utilised by *Blautia hydrogenotrophica*, *Clostridium* spp., and *Streptococcus* spp. to convert formate to acetate [33]. Other bacteria such as *Bacteroides* spp., *Prevotella* spp., *Ruminococcus* spp., and *Akkermansia muciniphila* utilise pathways that convert pyruvate to acetate with acetyl-CoA as an intermediate [33]. Unique to *Bifidobacterium* spp. is the bifid shunt, which is a different acetate production pathway with converts fructose-6-phosphate to acetate [34]. Propionate can be produced by many different species of bacteria that utilise one of three different pathways. *Coprococcus catus* and *Megasphaera elsdenii* convert lactate to propionate via the acrylate pathway; *Phascolarctobacterium succinatutens*, *Dialister* spp., *Bacteroides* spp., and *Veillonella* spp. convert succinate to propionate via the succinate pathway; and *Roseburia inulinivorans, Blautia obeum* and *Salmonella* spp., convert propanediol to propionate via the propanediol pathway [33]. There are two pathways for butyrate production utilised by human gut bacteria. One of these pathways converts acetyl-CoA to butyrate and is present in *Coprococcus comes* and *Coprococcus eutactus* [33]. The second pathway converts acetate or lactate to butyrate. Organisms that are able to utilise lactate to produce butyrate are *M*. *elsdenii*, *Veillonella* spp., *Anaerobutyricum hallii*, *Anaerostipes caccae*, and a potential *Clostridium* species [35,36]. Organisms that utilise acetate in butyrate production are *Anaerostipes* spp., *C. catus*, *Eubacterium rectale*, *Anaerobutyricum hallii*, *Faecalibacterium prausnitzii*, and *Roseburia* spp. [33].

SCFA precursors can be produced by many species of gut-associated bacteria. For example, *Bifidobacterium* spp. produce lactate, acetate, formate, and succinate in their bifid shunt pathway, while *Lactobacillus* and *Enterococcus* spp. produce lactate [34,37,38]. Succinate is also produced by members of the phylum *Bacteroidota*, such as *Prevotella copri* and *Bacteroides* spp. [39]. This usage of microbial metabolites by other bacteria is referred to as

cross-feeding. The utilisation of microbial acetate, lactate, and succinate to produce propionate and butyrate is a key part of SCFA production in the human gut [37,38,40].

The highest concentration of SCFAs in the gut is found in the caecum at 131 mmol/kg, with this concentration steadily decreasing along the gut [41]. These metabolites lower the pH of the intestine and are readily absorbed by the host [15]. Once absorbed by the host SCFAs are either utilised in ATP production in colonocytes (butyrate) or they enter the blood (propionate, acetate) where they can have a variety of different effects on human health [33,42]. SCFAs can be detected in the blood of healthy individuals at concentrations between 0.3 and 1.5 µM (butyrate), 0.9 and 1.2 µM (propionate), and 22 and 42 µM (acetate); but typically are almost absent in faeces and urine [43]. Butyrate has been implicated in the prevention of colon cancer, healthy colonic cell proliferation and reductions in colonic inflammation [33,44]. SCFAs also interact with the human free fatty acid receptors FFAR2 and FFAR3, which can influence host metabolism [32,33,42]. These receptors are present in many different areas of the body; with FFAR2 being expressed in enteroendocrine L and immune cells, fungiform papillae, and the circulatory system; and FFAR3 being expressed in the colon, kidneys, sympathetic nervous and circulatory systems, and endothelial cells lining the blood-brain barrier [42,43,45]. FFAR2 preferentially binds acetate and propionate, while FFAR3 binds butyrate and propionate [46]. Stimulation of *FFAR2* gene expression with acetate in mice has been shown to improve glucose homeostasis and lipid metabolism, while *FFAR2*-negative mice are more prone to obesity compared with *FFAR2*-postive mice [33]. *FFAR2*-negative mice have also been shown to become obese when fed a normal diet, with a suppression of insulin signalling in their adipocytes [45]. These effects of *FFAR2* deletion could be caused by a reduction in pancreatic beta cell mass seen in *FFAR2*-negative mice, which leads to the development of type 2 diabetes and obesity when mice are fed a normal diet [47]. This effect may not extend to humans however, due to work reported that the deletion of *FFAR2* in a human pancreatic cell line did not see this reduction in cell mass, and instead saw an increase in insulin synthesis [47]. Despite the negative effects of *FFAR2* deletion being shown in some work, *FFAR2*-negative mice have also been shown to be protected from fat gain when fed a high-fat diet and show improved glucose homeostasis when compared to *FFAR2*-postive mice [45]. It has been suggested that the differences in results of these studies may be due to microbiota differences between mice, which could

lead to altered SCFA production [45]. *FFAR3*-positive mice have been found to have higher rates of obesity compared with *FFAR3*-negative mice, but FFAR3 activation with butyrate has been shown to increase human beta cell mitochondrial respiration which could alleviate issues caused by damaged beta cells that are characteristic in type 2 diabetes [33,47]. FFAR2 and FFAR3 also influence the production of the gut-secreted hormones GLP-1 and PYY, both of which have anti-obesity and anti-diabetic effects [47,48]. These are both produced by enteroendocrine L cells in an SCFA-dependent manner, shown by *FFAR2*-negative mice not producing GLP-1 and activated FFAR2 increasing levels of both PYY and GLP-1 [47,48]. FFAR2 activation has been linked to lipolysis and the inhibition of adipogenesis, with pertussis toxin-mediated FFAR2 disruption preventing lipolysis in adipose tissue and mice overexpressing FFAR2 being found to be protected from obesity when on a high-fat diet [48] A microbiota-related increase of SCFAs in wild-type mice has also been shown to improve metabolic homeostasis with an increase in PYY being linked to increased FFAR2 activation, leading to better body weight control and glucose homeostasis [48]. FFAR2 and FFAR3 also contribute to immune response regulation in an SCFA-dependent manner. The levels of SCFAs at the site of infections, especially acetate, are elevated, and the activation of FFAR2 with acetate has been shown to improve the outcomes on infection in mice [46]. *FFAR2*-negative mice have been shown to be more susceptible to infection with *Citrobacter rodentium* and *K. pneumoniae*, which could be due to FFAR2 being important to the antibody response from dendritic and B cells and FFAR2 aiding in the recruitment of neutrophils, as well as other innate immune response regulation [46,48]. Both FFAR2 and FFAR3 also play a role in the regulation of inflammation within the immune response. *FFAR2*-negative mice have been shown to exhibit a more severe inflammatory response in colitis, arthritis, and asthma. SCFA activation of FFAR2 has also been shown to reduce inflammation in a murine model of colitis [47]. SCFA activation of FFAR3 has been shown to reduce the production of proinflammatory cytokines and increase the production of anti-inflammatory cytokines, as well as enhancing the adaptive immune response [47]. *FFAR3*-negative mice have also been found to have a reduced immune response to *C. rodentium* infection [47]. However, some work suggests that FFAR3 may promote inflammation, with SCFA-mediated activation of FFAR3 being linked to an increase in cytokine and chemokine production, leading to caecal inflammation in domestic goats [47].

SCFAs have also been shown to be able to pass the blood-brain barrier [42]. This has led to work showing that propionate may have activity on the sympathetic nervous system in mice via *FFAR3*, as well as a protective effect on the blood-brain barrier in a model using a human cell line [42,43]. It has also been shown using stable-isotope-labelled substrates in mice that gut-derived acetate accumulates in the hypothalamus, leading to a decrease in appetite via neurotransmitter interactions [49].

1.5.2 Bile acids

Bile acids are important metabolites modified by the gut microbiota, with up to 66 bile acids being detected in blood and 55 being detected in urine [50]. In humans, bile acids are synthesised from cholesterol before being conjugated with either glycine or taurine before being stored in the gallbladder to be excreted in bile [51]. The primary bile acids synthesised by humans are cholic and chenodeoxycholic acid, along with their glycine and taurine conjugates [52]. These bile acids are secreted into the duodenum in bile, where they play a role in the absorption of lipids and aid in the innate immune system via the disruption of microbial cell membranes [51,52]. Excreted bile acids are then reabsorbed in the ileum, often after being altered by the gut microbiota, before being transported back to the liver and converted to primary bile acids before being excreted again [52]. This process is referred to as enterohepatic recirculation (Figure 1.2).

There are four different microbial metabolism pathways related to the alteration of bile acids in the human gut. These are deconjugation, dehydroxylation, oxidation, and epimerization [52]. The deconjugation of bile acids is carried out by bacteria that encode bile salt hydrolases (BSH) [53]. Commensal microbes that have been found to encode these include *Lactobacillus*, *Bifidobacterium*, *Enterococcus*, *Clostridium,* and *Bacteroides* spp., as well as the archaea *Methanobrevibacter smithii* and *Methanosphaera stadtmanae* [51,54]. The effects of microbes encoding BSHs include a lowering of serum cholesterol levels that has been seen in both humans and mice [51]. Weight gain also appears to be affected by the presence of BSHs, with the expression of a cloned BSH being expressed in the guts of mice increasing weight gain and a study in humans correlating a decrease in BSH abundance to an increase in weight gain [51,55]. Conversely a reduction in BSH activity in mice and reduction in BSH carrying *Lactobacillus* and *Clostridium* species was found to lead to a reduction

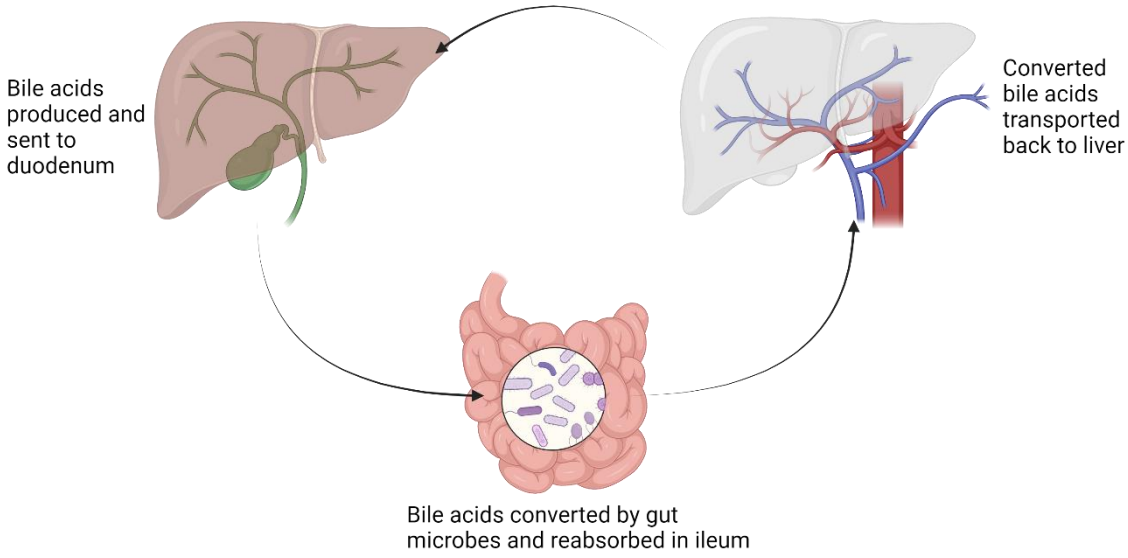


Figure 1.2. The process of enterohepatic recirculation.

Bile acids produced from cholesterol are conjugated in the liver and stored in the gallbladder in bile. This bile is then excreted into the duodenum where the bile acids are altered by microbes. These new bile acids are then reabsorbed and transported back to the liver where they are converted back to their original forms.

in weight gain that is potentially mediated by the bile-acid-activated farnesoid X receptor (FXR) [51]. Dehydroxylation of cholic acid to deoxycholic acid and chenodeoxycholic or ursodeoxycholic acid to lithocholic acid can be carried out by microbes encoding the *bai* operon [54]. This operon has been found in genomes belonging to *Clostridioides* spp., *Eggerthella* spp., and *Lachnosporidium* spp. [54]. The epimerisation of the primary bile acids is a two-step process, with the same hydroxyl group being oxidised and then reduced [52]. As cholic acid has three hydroxyl groups there are three epimers, alongside an intermediate bile acid for each of these epimers. Chenodeoxycholic acid has two hydroxyl groups and so can only be converted into two epimers, with two intermediates. The oxidation and reduction steps of bile acid epimerisation can be carried out by separate species that possess position-specific enzymes to each hydroxyl group [52]. Separate to these four catabolic pathways, an anabolic interaction between gut microbes and bile acids exists. Cholic acid can be conjugated into phenylalanocholic acid, tyrosocholic acid, and leucocholic acid in mice, potentially by *Clostridium* spp. [56].

Alongside their role in lipid digestion, bile acids have been found to have an important role in human metabolism [57]. The receptors TGR5 and FXR have been found to interact with bile acids, with differing effects [56,58,59]. Both of these receptors play a role in the modulation of bile acids by inhibiting their production, preventing bile acid concentrations from reaching cytotoxic levels [59,60]. The concentration of bile acid required for a cytotoxic effect varies depending on the cell being exposed and the acid it is being exposed to. An example of bile acid cytotoxic concentrations is lithocholic acid having an IC_{50} of 28 µM against a rat kidney cell line, while deoxycholic acid had an IC_{50} of 185 µM [61]. In human plasma the bile acid glycochenodeoxycholic acid has been measured as the most concentrated at 13.9 µM [50]. Bile acids have also been found to play a role in glucose homeostasis, fat homeostasis and immune cell function; leading them to be implicated in the development of type 2 diabetes, non-alcoholic fatty liver disease and obesity [51,53,57–59,62].

1.5.3 Protein-related metabolism

Bacteria residing in the gut can utilise dietary and endogenous proteins, peptides, and amino acids (Figure 1.3). Between 5 and 10 % of dietary protein is available for use by the gut microbiota, where microbes can utilise proteins for energy production and peptides for

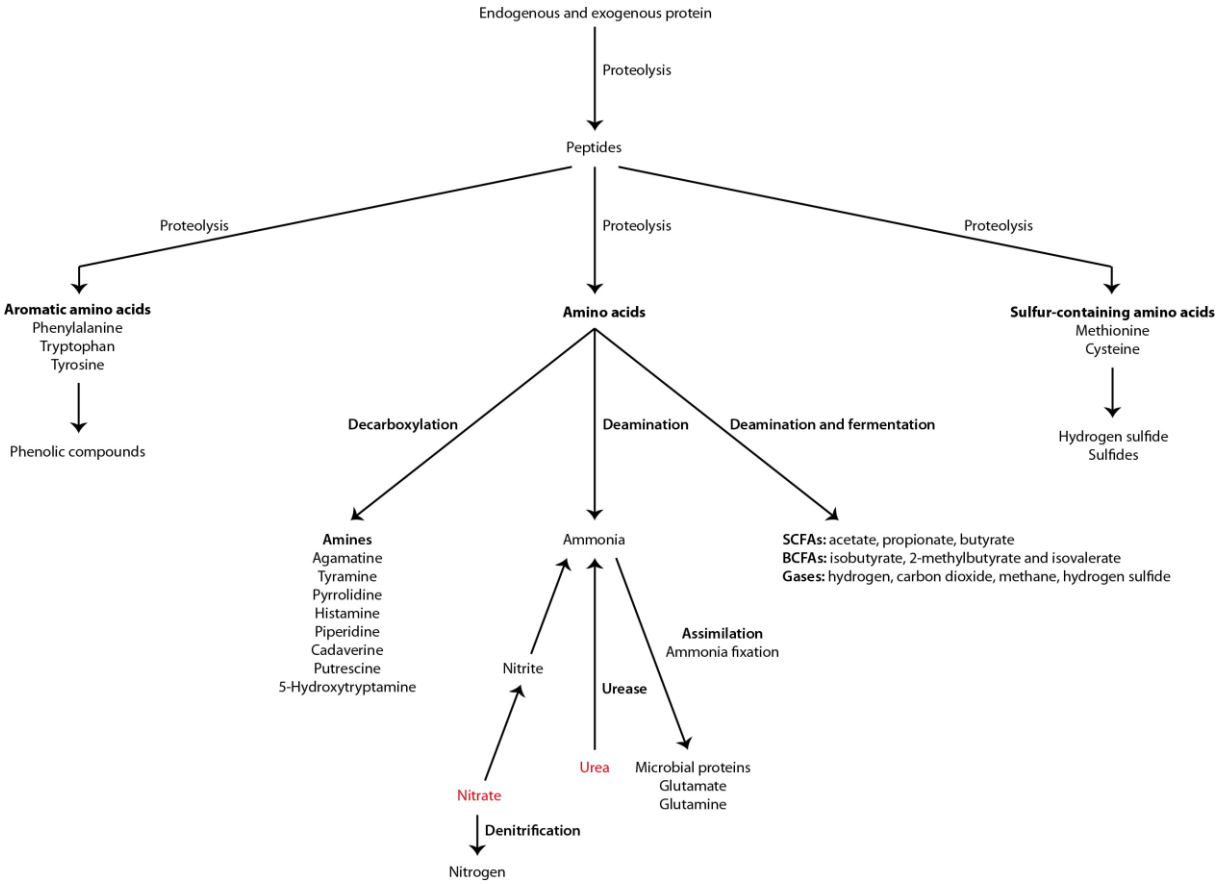


Figure 1.3. Microbial use of proteins, peptides, and amino acids in the large intestine.

Amino-acid-utilizing bacteria are more dominant than peptide-utilizers in the human faecal microbiota [63]. Compared with carbohydrate fermentation, protein and amino acid use by the gut microbiota results in a wider range of metabolites [64]. Non-protein sources of ammonia are shown in red. Image used with permission of L. Hoyles (unpublished). SCFAs, short-chain fatty acids; BCFAs, branched-chain fatty acids.

their own protein production [23]. The products of microbial protein metabolism include small amounts of SCFAs and branched-chain fatty acids such as isovalerate, isobutyrate, and 2-methylbutyrate, alongside compounds such as ammonia, hydrogen sulfide, indoles, and phenols [23,65].

An example of an amino acid derivative influencing human health is imidazole propionate. Histidine can be metabolized directly to imidazole propionate by *Adlercreutzia equolifaciens*, *Shewanella oneidensis*, and *Brevibacillus laterosporus*, or via the intermediate urocanate by bacteria encoding urocanate reductase (*Anaerococcus*, *Aerococcus*, *Streptococcus*, *Adlercreutzia*, *Eggerthella*, *Lactobacillus*, *Shewanella*, and *Brevibacillus* spp.). Imidazole propionate can impair insulin signalling and glucose tolerance, contributing to the development of type 2 diabetes [66]. In European patients with type 2 diabetes levels of imidazole propionate were found to be higher when compared to their healthy counterparts, with imidazole propionate being associated with inflammatory bowel disease (IBD) and a low microbial diversity in the gut [67]. Microbial metabolism of the amino acid phenylalanine to phenylacetate has also been linked to hepatic lipid accumulation in non-diabetic obese women, contributing to the development of hepatic steatosis in this group [68].

Another amino-acid-derived metabolite that can have effects on mammalian health is indole. Indole is generated in the gut via the microbial breakdown of tryptophan, with levels of tryptophan in germ-free mice being 1.5 times higher than normal mice [23,69]. Species implicated in the production of indole from tryptophan are *Achromobacter liquefaciens*, *Bacteroides ovatus*, *B. thetaiotaomicron*, *E. coli*, *Paracidobacterium coliforme* and *Proteus vulgaris* [23]. Indole has been shown to reduce inflammation of mouse liver when challenged with the endotoxin lipopolysaccharide (LPS), as well as reducing liver inflammation and preventing an alteration in hepatic cholesterol metabolism *in vivo* when healthy mice were fed indole prior to an LPS challenge [70]. Other indole-containing compounds have also been found to affect mammalian health. The metabolite indoxyl sulfate is a nephrotoxin that has been identified in the serum of normal mice, but not in germ-free mice, that has been found to accumulate in patients with kidney failure [69]. Indoxyl sulfate is not directly produced by gut microbiota and is instead generated in the liver from indole by cytochrome P450 and sulfotransferase [71,72]. The compound indole-3-

propionic acid is created in the gut from the microbial metabolism of indole, and it has been found to be an antioxidant with neuroprotective properties [73]. Indole-3-propionic acid is produced by the *Clostridium* species *C. botulinum*, *C. caloritolerans*, *C. paraputrificum*, *C. sporogenes*, and *C. cadaveris*; as well as the *Peptostreptococcus* species *P. asaccharolyticus*, *P. russellii*, *P. anaerobius*, and *P. stomatis* [71].

Ammonia in the gut is generated from the deamination of amino acids or from the metabolism of urea [23,32]. An example of this is the deamination of histidine in the same pathway in which imidazole propionate is produced [66]. Bacteria in the gut can also utilise ammonia as a nitrogen source in the production of proteins. Most ammonia generated however is transported from the gut to the liver via the hepatic portal vein and converted back to urea in a form of metabolic retroconversion, before being excreted in urine [32]. This process can potentially cause damage to the liver or negatively affect gut permeability, contributing to the development of fatty liver disease [74].

1.5.4 Methylamines and TMAO

TMAO is a small organic compound with the formula C_3H_9NO (Figure 1.4). At room temperature, TMAO exists as a colourless crystalline solid that is soluble in water. TMAO can be found naturally in many organisms, especially in fish which have a tissue concentration of 8–789 mg/100 g depending on species, where it acts as an osmolyte. TMAO has been found to protect proteins from the destabilising effects of urea, as well as the destabilising effects of osmotic and hydrostatic pressures [75].

While some of the TMAO in humans is ingested from consumption of foods such as fish, a much larger amount is produced in the liver from products of microbial metabolism [25,76]. TMAO and other dietary methylamines such as choline, phosphatidylcholine (PC) and carnitine are converted to TMA by gut microbes in the small intestine [25,77,78]. This TMA then enters the bloodstream via the small intestine and the hepatic portal vein where it is transported to the liver and converted back to TMAO by hepatic flavin-containing monooxygenases (FMOs) [25,79]. This TMAO is then transported to the kidneys where it is excreted in the urine (Figure 1.5) [76]. In metabolically healthy humans, baseline serum TMAO levels can be around 3.5 µM with levels as high as 99.9 µM seen in patients suffering from renal failure and who required dialysis [80,81]. The average baseline level of TMA in humans

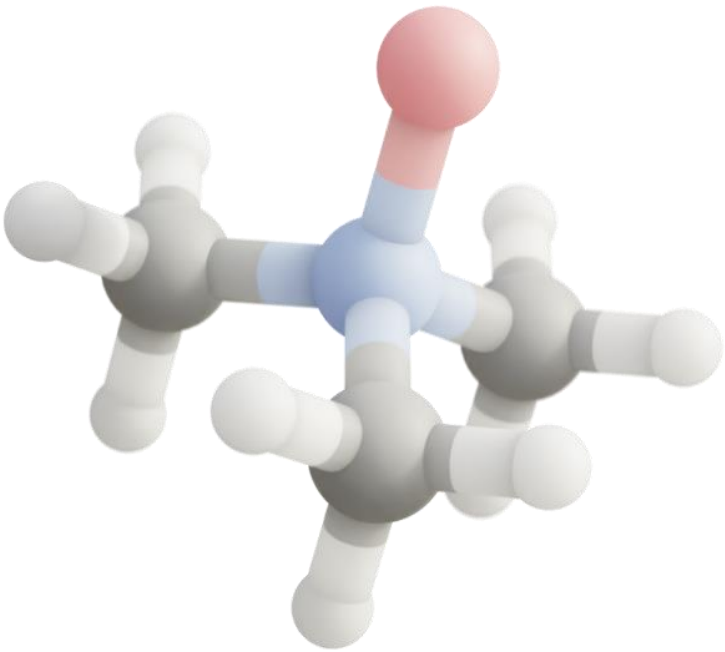


Figure 1.4. Molecular structure of TMAO.

White = H, Black = C, Blue = N, Red = O. TMAO is a polar molecule, with the oxygen atom being the negative end and the nitrogen atom being positive.

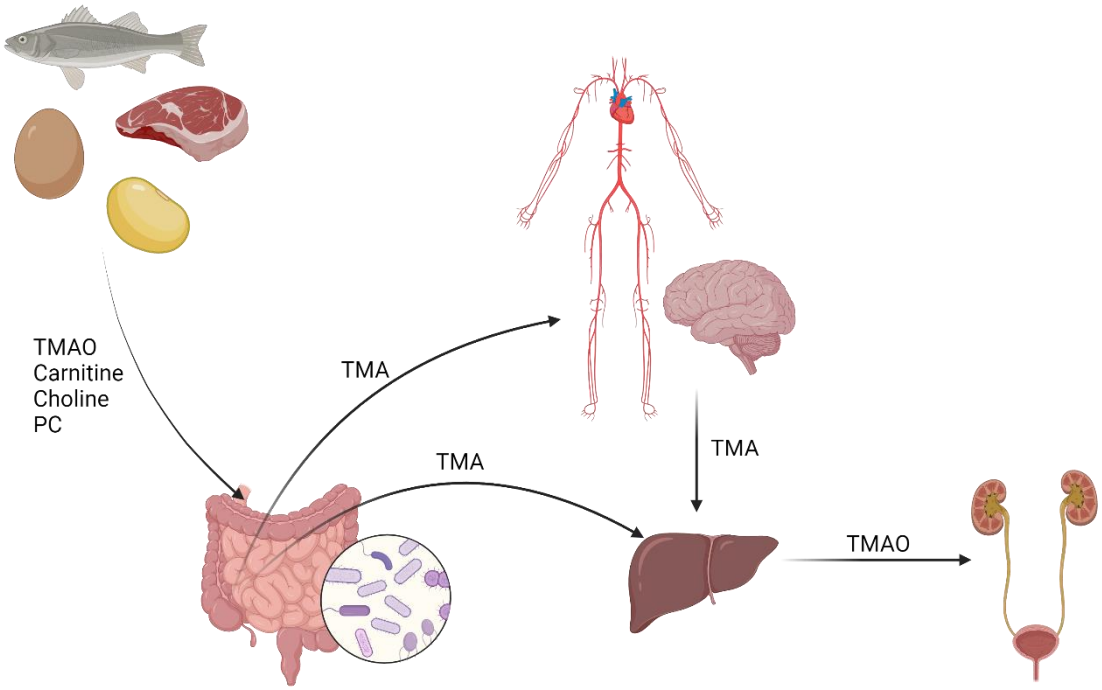


Figure 1.5. Dietary sources of circulating TMAO.

TMAO and other methylamines (choline, carnitine, PC) are ingested and then metabolised into TMA by microbes in the gut. This TMA then enters the blood where it interacts with other systems in the body and the liver. In the liver TMA is converted to TMAO by hepatic flavin monooxygenases, predominantly FMO3, before it is transported to the kidneys and excreted in urine. The conversion of ingested TMAO to TMA by gut microbes then back to TMAO by hepatic enzymes is a form of metabolic retroconversion.

is 0.418 µM with elevated levels averaging 1.39 µM in patients requiring dialysis [81]. While not as extreme as the patients suffering from severe renal failure, an elevated baseline level of TMAO has been associated with a higher risk of cardiovascular disease in humans [80,82,83]. This elevated level can be up to 8.8 µM with a median value of 5.0 µM, compared to an upper concentration of 5.9 µM and median value of 3.5 µM [80]. Koeth *et al*. [82] and Tang *et al*. [80] examined the levels of TMAO in plasma and their correlation with patients experiencing major adverse cardiac events (death, myocardial infarction, or stroke) over a 3-year period, while Wang *et al*. [83] found a dose-dependent association between plasma levels of TMAO and the presence of cardiovascular disease. These elevated serum TMAO levels have also been found in patients with type 2 diabetes, along with higher levels of TMAO in diabetic patients being associated with a greater risk of cardiovascular disease [84].

Mice are often used as a model to examine the effects of TMAO on mammalian systems and elevated serum TMAO levels have been associated with a higher risk of cardiovascular disease in mice [82,83,85]. Both studies used *ApoE*$^{-/-}$ mice which are predisposed to the development of atherosclerosis. The background mouse strain C57BL/6J is also prone to developing type 2 diabetes, atherosclerosis, and diet-induced obesity. Elevated levels of TMAO have also been found to have negative effects on glucose tolerance in non-obese, high-fat diet C57BL/6J mice [86]. There may also be issues with the usage of murine models in TMAO research as mice naturally have a higher concentration of circulating TMA. Male mice have been found to have a urinary TMA concentration of 5 mM, compared to a high of 4.39 µM in a healthy human [87,88]. In addition to having a naturally higher circulating TMA concentrations mice are also more sensitive to TMA, with the amine receptor TAAR5 being around 200 times more sensitive to TMA in mice than in humans [89]. TMAO has also been linked to an increase in risk of developing thrombosis in humans, potentially via TMAO interacting with platelets and FMO3 mediating platelet response time, with FMO3 supressed mice having a decreased platelet response [90].

Claims of TMAO being detrimental to cardiometabolic health are being called into question. In a long-term human study examining serum levels of TMAO, betaine, and choline no link was found between elevated serum TMAO levels and risk of cardiovascular disease [91]. The participants in this study were all undergoing screening for potential coronary artery disease, but no link was found between those with higher TMAO levels and those with

coronary artery disease. Mueller *et al*. [91] also found that renal function was influential on serum levels of TMAO, with lower renal function, defined by a reduction in estimated glomerular filtration rate, being correlated with higher serum TMAO levels. Other work has also made this association between chronic kidney disease and elevated serum TMAO levels in humans [92–95]. The inverse of this association has also been shown, with TMAO negatively effecting kidney function and reducing the estimated globular flow rate [95]. This is by TMAO contributing to renal fibrosis, which is a sign of kidney damage that leads to estimated globular flow rate decline. This work has also shown that a decline in circulating TMAO levels is associated with the administration of reno-protective drugs to patients with type 2 diabetes [95]. This points to elevated TMAO levels being more important when related to the health of the renal system and overall metabolic health, as opposed to solely the cardiovascular system. In addition to this a study examining the health of Western and Japanese patients found that elevated serum TMAO levels could only be linked to poor health in Western patients [96]. People living in Japan tend to consume more TMAO-rich food (i.e. fish), suggesting that in a Western diet other dietary components or factors that can influence serum levels of TMAO may be the cause of the link between elevated serum TMAO levels and poor cardiac health [96]. The high levels of carnitine consumed from red meat in Western diets may contribute to this, with the ingestion of carnitine being shown to increase serum levels of TMAO in humans via microbial metabolism and promote atherosclerosis in mice [82,97,98].

Several studies have been conducted where serum concentrations of TMAO, carnitine, gamma-butyrobetaine, and crotonobetaine have been linked to risk of cardiovascular disease, atherosclerosis, and death in humans [99–103]. However, in some of these studies the elevated methylamine serum concentrations were also correlated with reduced kidney function, with lower kidney function raising the risk of mortality and cardiovascular disease [99,101,102]. Consumption of red meat high in carnitine was linked to the development of atherosclerosis in humans, suggested to be due to the high serum levels of gamma-butyrobetaine, crotonobetaine, and TMAO created from microbial metabolism [100]. This link between food, atherosclerosis, and methylamine concentrations was not seen with fish or eggs however, despite fish's high TMAO levels and eggs' high PC levels. In individuals with type 2 diabetes and albuminuria TMAO-related metabolites in serum were linked to a

decrease in renal function but were not associated with an increased risk of developing cardiovascular disease [104]. The methylamine trimethyllysine has also been linked to the development of cardiovascular disease in humans [105]. Trimethyllysine has been found to be converted to TMA via an unknown pathway in anaerobic conditions by undefined microbes isolated from the caeca and colons of mice [105]. Despite these findings trimethyllysine was not found to affect the plasma concentration of TMAO in mice. Another study looking at dietary methylamines and atherosclerosis in the *Ldlr*$^{-/-}$ and *ApoE*$^{-/-}$ atherosclerotic murine models did not see any links [106]. In this study both mouse models were fed a diet supplemented with choline, betaine, or TMAO but neither model saw an increase in atherosclerosis development even when serum TMAO levels increased.

There is also more evidence pointing to TMAO potentially being beneficial to health as TMAO has been found to have protective effects on cardiac, hepatic, and cerebral health in murine models [107–112].

Regarding cardiac health, TMAO was found to decrease the severity of issues in three different murine models of cardiac disease. Elevated serum TMAO was found to decrease the number of aortic lesions found in *ApoE* $^{-/-}$ mice [108]. In a murine model of hypertension, chronic exposure to a low level of TMAO was found to reduce diastolic dysfunction and elevated levels of TMAO were not found to have any adverse effects on the cardiovascular system [110]. Exposure to TMAO was also found to have a protective effect on cardiac muscle, via the protection of mitochondrial activity, in a murine model of right ventricle failure [112].

TMAO has been shown to improve hepatic health through protection from non-alcoholic steatohepatitis. This was shown in rats fed a high-fat-high-cholesterol diet that were given a TMAO treatment of 120 mg/kg/day via oral gavage, leading to a reduction in hepatic endoplasmic reticulum stress and hepatic cell death [109]. This was potentially due to a TMAO-mediated reduction in cholesterol absorption leading to a decrease in hepatic cholesterol levels [109]. Another study also found that TMAO can reduce endoplasmic reticulum stress in mice, leading to improved glucose homeostasis and increased insulin secretion [111].

Several microbial metabolites have been shown to have protective effects on the blood-brain barrier [43,107,113]. TMAO was found to improve blood-brain barrier integrity in mice at

physiologically relevant concentrations, as well as improving their cognitive function [107]. Levels of TMAO, choline, and carnitine have been found to not be linked to cognitive decline and the development of Alzheimer's disease in humans, but the associated metabolite crotonobetaine has been linked to a higher risk of these health problems, while gamma-butyrobetaine has been linked to a reduction in risk [114]. In rats carnitine and the associated metabolite acetyl-L-carnitine have been shown to have neuroprotective affects [114].

TMAO has also been shown to slow tumour growth in murine model of pancreatic ductal adenocarcinoma, alongside improving the anti-cancer immune response by increasing the activation of tumour-associated macrophages, dendritic cells, and $CD8^+$ T cells [115]. TMAO directly injected into mice and TMAO generated endogenously by gut bacteria had this tumour-suppressing effect. TMAO was also found to oppose the upregulation of immunosuppressive molecules that are produced by pancreatic ductal adenocarcinoma tumours [115].

1.6 Interactions between the human gut microbiota and dietary methylamines

Current knowledge suggests that the conversion of TMAO to TMA is largely carried out by members of the family *Enterobacteriaceae* that encode the TorA protein [116,117]. This breakdown is part of an anaerobic respiration pathway that uses TMAO as the terminal electron acceptor and the TorA, TorZ or DmsA protein (Figure 1.6) [116,118,119]. Other microbial sources of TMA in the gut come from the breakdown of dietary methylamines such as choline, carnitine, and PC [25]. Despite it being known that microbes are responsible for the production of TMAO precursors in the human gut, it has been shown to not be possible to predict host TMAO production/circulating levels based on analysis of human faecal metagenomic datasets [120].

$$C_3H_9NO + 2H^+ + 2e^- \rightarrow C_3H_9N + H_2O$$

Figure 1.6. Chemical formula of the conversion of TMAO to TMA carried out by microbes.

Electrons are supplied by an electron transport chain quinol which is converted to a quinone via the loss of electrons.

Choline is an important part of a healthy human diet being found in meat, eggs, and soybeans. It contributes to cell membrane formation and neurotransmitter production, with choline-deficient diets leading to fatty liver disease in healthy men [121]. In the gut, bacteria catabolise choline into TMA and acetaldehyde, with the excretion of TMA in urine having been directly linked to the ingestion of choline [77,122]. Currently microbial conversion of choline to TMA is thought to be carried out by bacteria that encode the CutC protein [117,122]. The *cutC* gene can be found in two different gene clusters, with cluster 1 being found in *Bacillota*, *Deltaproteobacteria*, and *Actinomycetota* (formerly *Actinobacteria*), and cluster 2 being found in *Gammaproteobacteria* [123].

PC, found in high concentrations in meat and eggs [121], also contributes to this route of TMA formation by its breakdown into choline [80]. Which species of bacteria are responsible for the catabolism of PC to choline is not fully known, although it is suggested that the *Bacteroides* species *B. fragilis* and *B. thetaiotaomicron* carry out this process using phospholipases in either the phospholipase D or patatin-like phospholipase family [79].

Carnitine is also a common part of the human diet, being ingested primarily in red meat [82], although some carnitine is also produced endogenously [124]. Carnitine's function in humans is in the transfer of long-chain fatty acids to mitochondria and the binding of acyl residues from amino acids metabolism [124]. Carnitine that is not absorbed by the human gut is almost completely broken down by microbes, into TMA, gamma-butyrobetaine, and crotonobetaine [85,100,125]. The direct breakdown of carnitine to TMA is an aerobic process carried out by bacteria that encode the CntA and CntB proteins [125]. The genera of bacteria that have been suggested to encode these proteins are the gammaproteobacteria *Klebsiella*, *Escherichia*, *Citrobacter*, *Enterobacter, Shigella* and *Acinetobacter*, as well as the betaproteobacterium *Achromobacter piechaudii* and the firmicute *Sporosarcina newyorkensis* [117,125]. The indirect breakdown of carnitine into gamma-butyrobetaine into TMA is an anaerobic process carried out by multiple bacteria. The step from carnitine to gamma-butyrobetaine is thought to be facilitated by genes in the *caiTABCDE* operon [97]. The genes *gbuA, gbuB, gbuC,* and *gbuE* have been shown to be essential to the conversion of gamma-butyrobetaine to TMA [103,126]. Production of TMA from gamma-butyrobetaine has been attributed to "*Emergencia timonesis"* while the production of gamma-butryobetaine from carnitine was attributed to "*Edwardsiella lenta*", *Edwardsiella tarda, Escherichia*

fergusonii and *Proteus penneri* [97]. This *caiTABDE*-mediated breakdown of carnitine is also responsible for the production of the intermediate crotonobetaine [97].

Currently not much is known about interactions between microbes that live in the human gut and TMA. Lactic acid bacteria have been shown to exhibit changes to their growth and metabolism when exposed to TMAO with members of the families *Streptococcaceae* and *Enterococcaceae* producing more lactic acid when grown in the presence of TMAO, although no TMA was produced [25]. Some *Enterococcus* spp. have been shown to degrade TMA, which may have been why TMA had not been detected previously after lactic acid bacteria were exposed to TMAO [127].

1.7 Aims of this work

In this introduction the vast number of microorganisms in the human digestive system has been discussed, with a further examination of how the metabolism of these microbes can influence the metabolism of their host. While it is clear that knowledge in this area is growing there is still much to learn, especially with regard to the metabolism of dietary methylamines. Currently much work relating to TMAO focuses on how the presence of TMAO and TMA can influence the health of the host, with comparatively little work examining how TMAO and TMA can influence the bacteria of the human gut. This work aims to explore the metabolic capabilities of human gut bacteria with respect to their utilisation of TMAO, with a particular focus on *Klebsiella pneumoniae*. This will involve a large-scale bioinformatic analysis of human gut-associated bacterial genomes to search for TMAO metabolism genes, full characterisation of a caecal isolate of *K. pneumoniae* with analysis of the effect of TMAO on its metabolome and transcriptome, and characterisation of the poorly characterised molybdoenzyme and potential TMAO reductase BisC. Extra work has also been undertaken to bioinformatically characterise novel isolates of *Klebsiella*-infecting bacteriophage and depolymerases carried by them.

Chapter 2 Bioinformatic exploration of TMAO metabolism in human gut-associated bacteria

2.1 Main pathways involved in TMAO metabolism

TMAO is an osmolyte commonly found in fish, as well as in trace quantities in vegetables, meat, and dairy products. Dietary precursors to TMAO, such as choline, PC, and carnitine, are also present in these foods further contributing to levels of TMAO in the human gut. An extensive review of the available literature has identified six different pathways associated with microbial TMAO metabolism (Figure 2.1).

2.1.1 *torCAD*

The *torCAD* operon is responsible for the creation and utilisation of TMAO reductase 1 (TR1), a molybdoenzyme that facilitates electron transport in a form of anaerobic respiration [116]. This anaerobic respiration reaction reduces TMAO along with hydrogen ions to create TMA and water. While TorA is the periplasmic cytochrome responsible for the reduction of TMAO [116], TorC acts as a chaperone as it shuttles electrons to TorA to facilitate the reduction of TMAO [128]. TorD is not directly involved in the electron transport but instead acts as another chaperone for TorA, making it able to receive the molybdenum cofactor critical for its function [129]. This system has been suggested to be the most prevalent and important pathway for the reduction of TMAO within the human gut, with the *torA* gene being examined extensively.

When examining the TorA protein, TR1 has been most extensively examined in *Escherichia coli* [116,130,131], along with work studying TMAO reduction in *Salmonella* [132,133]. Other work focussing on TR1 largely involves marine bacteria such as *Shewanella oneidensis* and *Rhodopseudomonas capsulatus* [130,134].

In a 2016 study by Jameson *et al*. [117] the prevalence of TorA was examined in marine and human gut populations, via metagenome mining. This study suggested that the TorA protein, and by extension TR1, is the most important contributor to TMAO breakdown to TMA in the human gut. The abundance of different genera that were found to carry TorA was also presented. In healthy humans, *Escherichia*, *Klebsiella*, *Citrobacter*, *Eggerthella* and *Sutterella* were suggested to be the most abundant genera that carried TorA. However,

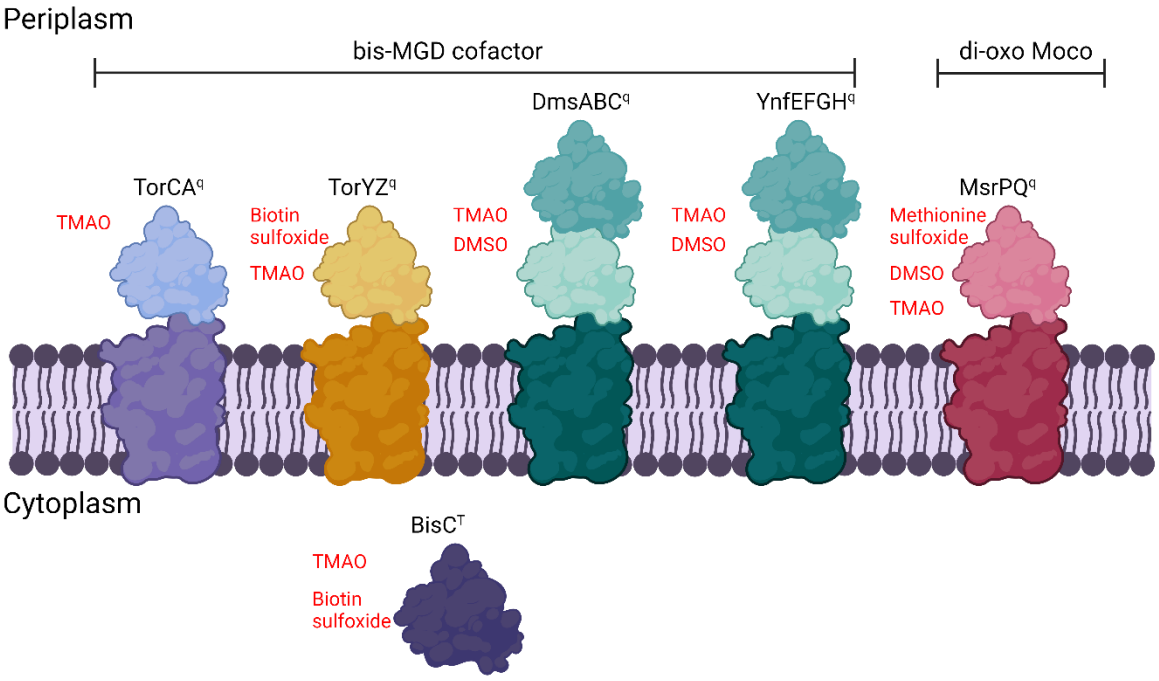


Figure 2.1. Enzymes related to TMAO metabolism in human gut-associated bacteria. Six different enzymes have been identified as potentially being relevant to TMAO metabolism in gut bacteria. Five of these have also been shown to have activity on other non-TMAO substrates. Enzymes marked by a ^{q} receive electrons from a quinone and enzymes marked by a ^{T} receive electrons from thioredoxin. *Enzyme substrates are marked in red.* TorCA, TorYZ, DmsABC, YnfEFGH, and BisC carry a bis-MGD molybdenum cofactor, while MsrPQ carries a di-oxo molybdenum cofactor. YnfEFGH is a coloured the same as DmsABC as it is a paralogue of it.

due to the way that this data has been presented how the abundance of these genera compare to non-TorA-encoding members of the gut microbiota cannot be determined. This has caused issues when comparing this study to other pieces of work as it is not known what percentage of metagenomes from a particular genus provided positive hits for TorA, preventing a clear picture of the prevalence of TorA in the gut from being formed. The methods provided are also unclear, with only brief mentions of the programs used and no clear indication of what the reference sequences used were. Following on from the reference sequence it appears that only a single representative protein sequence was used in the database. This causes issues as several different pathways have been collapsed into one by Jameson *et al*. under the title of TorA, including TorA and TorZ from marine species such as *Shewanella massilia* and *Photobacterium profundum*, along with DorA from *Rhodobacter capsulatus* and TorZ from *E. coli*.

The *torCAD* operon is also controlled by the TorT/S/R system, which is, in turn, controlled by the global repressor IscR [135]. These two systems working in tandem cause a high level of cell-to-cell variability in *torCAD* transcription levels in aerobic environments [135]. This functions by TorT detecting TMAO in the periplasm and interacting with TorS, causing TorS to phosphorylate TorR. The phosphorylated TorR then activates transcription of the *torCAD* operon [136]. IscR represses *torT* and *torS* with levels of IscR being higher in aerobic conditions, leading to higher levels of *torT* and *torS* repression in aerobic environments [135]. The opposite is true in anaerobic conditions, with higher levels of *torT* and *torS* transcription suggesting TorT and TorS will be found in higher abundance [135].

2.1.2 *torYZ*

The *torYZ* pathway is responsible for the production of TMAO reductase 2 (TR2). This is another molybdoenzyme that functions similarly to TR1, but TR2 is able to reduce biotin sulfoxide as well as TMAO [137]. In this pathway, TorY acts as the cytochrome while TorZ is the oxidoreductase [118]. While this pathway may seem similar to the *torCAD* operon the TorZ protein is more similar to another molybdoenzyme (BisC) than it is to TorA, sharing 63 % and 43 % identity, respectively. TorY also only shares a 34 % identity with TorC, and the system does not appear to have a TorD equivalent. The role of TorD is potentially taken up by another enzyme in the molybdopterin biosynthesis pathway (MDP).

This pathway is constitutively expressed, with levels of TMAO not appearing to have any effect on the levels of gene expression and has been described in *E. coli* under anaerobic conditions [118].

2.1.3 *dmsABC*

The *dmsABC* operon is responsible for producing dimethyl sulfoxide (DMSO) reductase, another molybdoenzyme that can reduce DMSO as well as TMAO. This is also done as part of the electron transport chain in anaerobic respiration, forming dimethyl sulfide and water from DMSO and hydrogen ions. DmsA forms the catalytic subunit of the DMSO reductase as it carries the molybdenum cofactor, while DmsB transfers the electrons [138]. DmsC acts as a membrane anchor subunit [139]. The protein DmsD is also involved in the biogenesis of DMSO reductase, acting as a chaperone to both DMSO reductase and TMAO reductase via twin-arginine translocation under anaerobic conditions [140,141]. DmsD has also been shown to act on YnfE and YnfF, which are paralogues of DmsA [142].

This operon is largely controlled by the FNR global repressor, which causes an increase in expression in anaerobic environments [143].

This operon was initially described in *E. coli* [138] and much of the work involving human-associated bacteria and the *dmsABC* operon also focuses on *E. coli*. The *dmsABC* operon in *Rhodobacter* spp. has also been studied extensively, although this has much more relevance to marine systems as these bacteria are not part of the human gut microbiota.

2.2 Other pathways involved in TMAO metabolism

2.2.1 *ynfEFGHI*

The *ynfEFGHI* operon is a paralogue of the *dmsABC* operon that has been shown to facilitate anaerobic respiration of DMSO, while also potentially being a selenate reductase in *Salmonella enterica* and *E. coli* [144]. In this operon, YnfE and YnfF are paralogues of DmsA, while YnfG and YnfH are paralogues of DmsB and DmsC, respectively [142]. YnfI has been renamed to DmsD, linking the two operons further [140]. The expression of *ynfE* has been seen to inhibit the expression of *ynfFGH* in *E. coli* although why this occurs is currently not known [142].

2.2.2 *msrPQ*

The MsrP (methionine sulfoxide reductase) and MsrQ (inner membrane dehydrogenase) proteins function together to protect periplasmic proteins from oxidative stress [145]. MsrPQ (previously known as YedYZ) uses electrons from the electron transport chain for its redox reactions [145]. While MsrQ potentially anchors MsrP to the cell membrane, purified MsrP carries a molybdopterin cofactor and can reduce TMAO and several other sulfoxides [146]. This system is under the control of the YedVW two-component system, which has been shown to trigger gene expression in the presence of hypochlorous acid in *E. coli* [145].

2.2.3 *bisC*

The *bisC* gene does not appear to be part of any of the full pathways detailed above, instead it functions with the MDP [147]. BisC (biotin sulfoxide reductase) is a cytoplasmic enzyme used mainly to reduce biotin sulfoxide, although it has also been shown to have activity on methionine *S*-oxide in *E. coli* [148]. While BisC has not been shown to reduce TMAO in *E. coli* or other human gut species, the *Rhodobacter sphaeroides* BisC enzyme has been shown to reduce TMAO when expressed in *E. coli* [149].

2.3 Aims of this work

As detailed above, most work investigating the ability of bacteria to reduce TMAO has only looked at the *torA* gene of the *torCAD* pathway. This has led to the TMAO-reducing ability of gut bacteria potentially being misunderstood, both in terms of levels of gene carriage as well as which groups of bacteria may contribute to TMAO reduction in the human gut. A detailed review of the literature has identified other TMAO pathways as potentially being relevant to the discussion of TMAO metabolism in the human gut. Each one of these pathways, excluding potentially *msrPQ*, require either other genes in their pathway or external pathways to function correctly. This means that to properly evaluate the prevalence of TMAO respiration in bacterial populations it is necessary to examine each gene in each pathway, as incomplete pathways have a high chance of being non-functional. This work aims to further examine which proteins involved in TMAO reduction are present in a wide range of gut bacteria, looking at a wider selection of pathways and proteins than considered previously.

2.4 Methods

2.4.1 Scripts

All scripts associated with this work are available from GitHub: https://github.com/samjtd/TMAO

2.4.2 Sequence sources

Publicly available draft genome sequences were downloaded from NCBI RefSeq between 01/04/2020 and 01/06/2020. Metagenome-assembled genome (MAG) representative sequences were downloaded from repositories found in work by Almeida *et al* [1]. Protein sequences used to construct BLASTp databases and alignments were downloaded from UniProt (Table 2.1).

2.4.3 Alignments of TMAO protein sequences

Protein sequences from different species to be aligned were identified based on Jameson *et al* [117]. The sequences for these proteins were downloaded from UniProt [150] and aligned using the online tool Clustal Omega EMBOSS Needle (https://www.ebi.ac.uk/Tools/psa/emboss_needle/) (v1.2.4) [151]. The alignment of TMAO pathway protein sequences was carried out the same way. The protein sequences used in this can be found in Table 2.1. *E. coli* sequences were selected on the basis of using the lab strain K12, non-*E. coli* sequences were selected as they were the only sequences available for the target proteins. The *Mannheimia varigena* sequence was selected at random.

2.4.4 FastANI analysis of RefSeq genomes

FastANI v1.33 analysis [152] was carried out on all *E. coli* (n = 18870) and *Klebsiella* (n = 9913) spp genomes. Average nucleotide identity (ANI) values of >95 % [153] against type strains of species within the genera were used to assign genomes to correct species (Table 2.2). Genomes that did not meet the >95 % ANI cutoff were identified using PubMLST [154].

2.4.5 BLAST searches of genomes and filtering of results

A BLASTp database [155] was constructed using *E. coli* protein sequences downloaded from UniProt (Table 2.1). Sequences were selected as they were all from the lab strain *E. coli* K12 or the only results for the desired species. Publicly available protein sequences from

NCBI RefSeq [156] had BLASTp run on them against this database. For species that were neither *E. coli* nor *Klebsiella* spp., which genomes were used was determined by their taxonomy in the Genome Taxonomy Database (GTDB) (v89) [157]. These genomes were downloaded from either NCBI RefSeq or GenBank [158]. Full numbers of genomes examined can be found in Table 2.3. The results of these BLAST alignments were then filtered using R, looking for hits with >90 % coverage and >70 % identity. Other genomes examined can be found in work by Ravcheev and Thiele [159].

2.4.6 BLAST searches on a publicly available metagenome reference dataset and filtering of results

The same BLASTp database as shown in Table 2.1 was used for these searches. BLASTp was used to search 4644 metagenome reference genomes [1]. These BLAST results were then filtered for hits with >90 % coverage and >70 % identity using R.

Prodigal (v.2.6.3) [160] annotations were used to identify which of the proteins identified were contiguous and formed full operons. Manual annotation of the DmsC protein in several genomes was also carried out, using the NCBI BLAST online tool and the UniProt database to check for correctness [150,155].

Table 2.1. UniProt protein sequences used in this study.

Protein	**UniProt entry**
E. coli TorA	P33225
E. coli TorC	P33226
E. coli TorD	P36662
E. coli TorY	P52005
E. coli TorZ	P46923
E. coli TorS	P39453
E. coli TorT	P38683
E. coli TorR	P38684
E. coli BisC	P20099
E. coli DmsA	P18775
E. coli DmsB	P18776
E. coli DmsC	P18777
E. coli MsrP	P76342
E. coli MsrQ	P76343
E. coli YnfE	P77374
E. coli YnfF	P77783
E. coli YnfG	P0AAJ1
E. coli YnfI	P76173
E. coli DmsD	P69853
Shewanella massilia TorA	O87948
Rhodobacter capsulatus DorA	Q52675
Rhodobacter capsulatus DmsA	Q57366
Photobacterium damselae TorZ	A0A4S2JIA1
Mannheimia varigena TorA	W0Q8Z5

Table 2.2. *Escherichia* and *Klebsiella* spp. type strain genomes included in this study. ^{T} signifies the genome belongs to a type strain.

Species and type strain	**Assembly accession**
E. coli ATCC 11775^{T}	GCF_003697165
K. indica	GCA_005860775
K. huaxiensis WCHKl090001^{T}	GCA_003261575
K. grimontii 06D021^{T}	GCA_900200035
K. oxytoca ATCC 13182^{T}	GCA_900977765
K. michiganensis SB4934^{T}	GCA_901556995
K. pasteurii SB3355^{T}	GCA_901563825
K. spallanzanii SB3356^{T}	GCA_901563875
K. aerogenes ATCC 13048^{T}	GCA_003417445
'*K. quasivariicola*' 10982	GCA_000523395
K. pneumoniae subsp. *pneumoniae* ATCC 13883^{T}	GCA_000742135
K. quasipneumoniae subsp. *quasipneumoniae* 01A030^{T}	GCA_000751755
K. variicola subsp. *variicola* SB1^{T}	GCA_900977835
K. quasipneumoniae subsp. *similipneumoniae* SB4697^{T}	GCA_900978135
K. variicola subsp. *tropica* SB5531^{T}	GCA_900978675
K. africana SB5857^{T}	GCA_900978845

Table 2.3. Number of genomes examined per taxon

Taxon	**Number of genomes**
Escherichia	18847
Klebsiella	9896
Burkholderia	2086
Burkholderiaceae	3
Citrobacter	245
Desulfitobacterium	12
Dorea	24
Eggerthella	12
Eggerthellaceae	20
Enterorhabdus	2
Gordonibacter	6
Lachnoclostridium	107
Parasutterella	4
Raoultella	55
Ruegeria	27
Sutterella	14
Sutterellaceae	10

2.5 Results

2.5.1 TMAO metabolism proteins are not highly conserved between families and different pathways

In previous work [117] a representative sequence was used in the analysis that appeared to have been generated from several different sequences. As some of the protein sequences that may have been used were for different proteins and from different species of bacteria, sequences were aligned to see if these sequences were comparable (Figure 2.2). There was a low level of homology for most sequences, especially between the marine species and *E. coli*. The highest level of homology between species was 64 % identity between *Mannheimia varigena* TorA and *S. massilia* TorA. On average there was a 44.9 % identity between all protein sequences.

Protein sequences from identified pathways were aligned against each other to show that it is important to examine each protein and pathways individually (Figure 2.3). Most proteins had a very low percentage identity, averaging at 22.15 % across the proteins examined. However, there are a few notable exceptions to this. Due to the *ynf* operon being paralogous to the *dms* operon there were high levels of homology between proteins that are a part of these pathways [142]. YnfE and YnfF are homologous with DmsA, having a 67 % and 66.5 % identity respectively and a 70.3 % identity with each other. YnfG has a 94.1 % identity with DmsB and YnfH has a 57 % identity with DmsC. YnfI and DmsD are the same

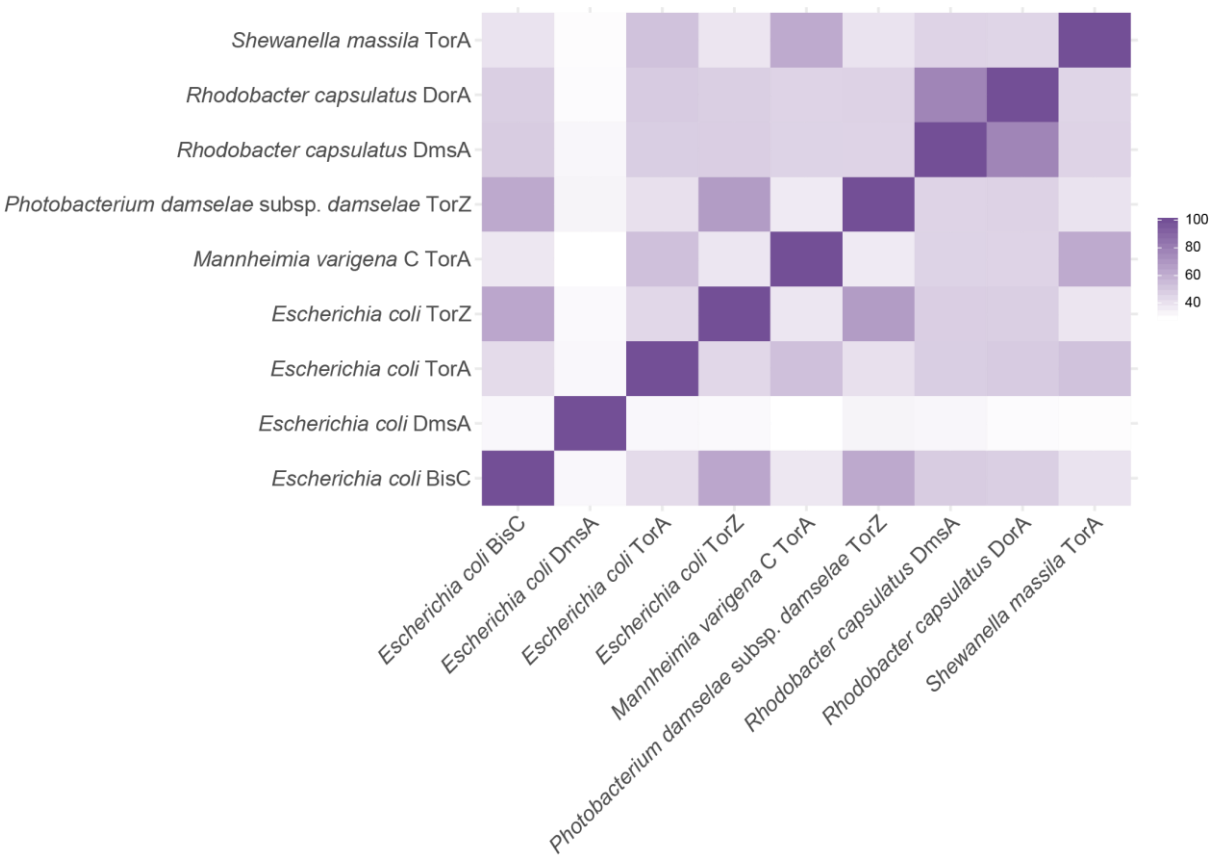


Figure 2.2. Comparison of different TMAO metabolism protein sequences.

The percentage identity between different proteins was low overall, averaging at 44.9 %. Proteins were selected based on the organisms described in work carried out by Jameson *et al* [117].

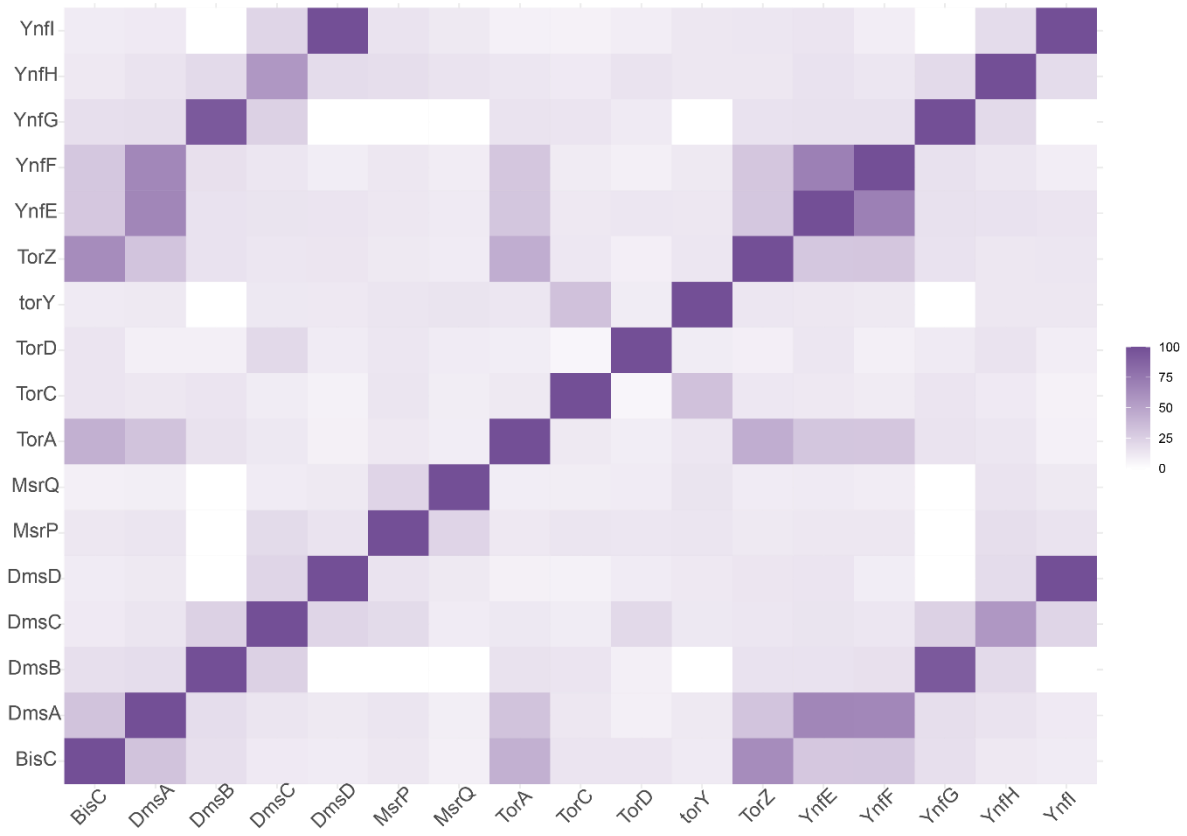


Figure 2.3. Comparison of different *E. coli* TMAO metabolism proteins.

The amino acid identity for these sequences was very low, averaging at 22.15 %. This low identity highlights the need to search for each of these proteins individually rather than via the creation of a consensus sequence. Average homology is higher between Ynf and Dms proteins as the *ynf* operon is homologous with the *dms* operon [142].

protein that can be found under either of these two labels. Two sequences used here were labelled differently but still had a 100 % identity with each other. Also of note is the 63.3 % identity shared between TorZ and BisC, which is higher than the 44.5 % identity shared between TorA and TorZ.

2.5.2 Most strains of *E. coli* encode multiple TMAO metabolism pathways while *Klebsiella* spp. lack *torCAD* and *torYZ*

E. coli (n=18870) and *Klebsiella* spp. (n=9913) genomes from RefSeq were analysed with FastANI to check that they were the correct species, based on a >95 % ANI cutoff [152,153]. Twenty-three of the *E. coli* genomes and 17 of the *Klebsiella* spp. genomes did not meet the >95 % ANI cutoff and were subsequently not included in this work. Suggested identities for these genomes can be found in Table 2.4 and were generated using PubMLST [154]. This left 18847 *E. coli* genomes (Figure 2.4) and 9898 *Klebsiella* spp. genomes (Figure 2.5) for analysis.

The majority of *E. coli* genomes examined encoded all of the TMAO metabolism proteins in the database. This was expected as all of these proteins have been described in strains of *E. coli* and the sequences used in the database are from *E. coli*. While this analysis was done to ensure that the database and script were working as intended, it is interesting to note that TorZ appears to be slightly less prevalent than other proteins.

The genus *Klebsiella* has previously been suggested to be a key organism in human gut TMAO metabolism, based on it being predicted to be one of the more prevalent carriers of TorA [117]. The current analysis has found that less than 1 % of publicly available *Klebsiella* genomes encode the TorA protein, along with other proteins responsible for the production of TR1 and TR2 (Figure 2.5). Most *Klebsiella* strains appear to be able to create DMSO reductase, although there is a notable lack of the DmsD protein across the genomes examined. BisC, MsrP and MsrQ are also encoded by the majority of *Klebsiella* strains. The Ynf proteins were not as prevalent as Dms proteins, despite their homology. Notably, YnfH is encoded by less than 1 % of the genomes examined, despite its homolog DmsC being encoded by 98.5 % of genomes.

Table 2.4. MLST identities of genomes below the >95 % ANI cutoff

Assembly accession	NCBI RefSeq identity	PubMLST identity
GCA_900083925	*K. quasipneumoniae* subsp. *quasipneumoniae*	*Serratia marcescens*
GCA_000963575	*K. pneumoniae*	*Cedecea neteri*
GCA_001187865	*'K. quasivariicola'*	*Superficierbacter electus*
GCA_000333535	*K. michiganensis*	*Kosakonia cowanii*
GCA_900083935	*K. pneumoniae*	*E. coli*
GCA_900493335	*K. pneumoniae*	*Enterobacter cloacae*
GCA_900076565	*K. quasipneumoniae* subsp. *similipneumoniae*	*Enterobacter hormaechei*
GCA_900076075	*K. grimontii*	*Enterobacter hormaechei*
GCA_900493505	*K. quasipneumoniae* subsp. *quasipneumoniae*	*Enterobacter hormaechei*
GCA_900076735	*K. quasipneumoniae* subsp. *quasipneumoniae*	*Enterobacter hormaechei*
GCA_900083845	*K. quasipneumoniae* subsp. *quasipneumoniae*	*Enterobacter hormaechei*
GCA_001312905	*K. michiganensis*	*Raoultella ornithinolytica*
GCA_900083685	*K. grimontii*	*Raoultella ornithinolytica*
GCA_900083755	*K. grimontii*	*Raoultella planticola*
GCA_005860775	*K. indica*	*K. indica*
GCA_900198535	*K. indica*	*K. indica*
GCA_009707385	*K. grimontii*	Inconclusive
GCF_000944855	*E. coli*	*E. coli*
GCF_000936005	*E. coli*	*E. coli*
GCF_000529815	*E. coli*	*K. pneumoniae*
GCF_003363055	*E. coli*	*Phytobacter ursingii*
GCF_003176195	*E. coli*	*K. quasipneumoniae*
GCF_003145355	*E. coli*	*Escherichia ruysiae*
GCF_000398885	*E. coli*	*E. coli*
GCF_011008595	*E. coli*	*Kluyvera genomosp*
GCF_900536495	*E. coli*	*Enterobacter hormaechei*
GCF_003175335	*E. coli*	*K. variicola*
GCF_001286085	*E. coli*	*Escherichia albertii*
GCF_002110245	*E. coli*	*E. coli*
GCF_900499915	*E. coli*	*Escherichia marmotae*
GCF_000713035	*E. coli*	*E. coli*
GCF_000601195	*E. coli*	*E. coli*
GCF_004745245	*E. coli*	*E. coli*

Assembly accession	NCBI RefSeq identity	PubMLST identity
GCF_005400045	*E. coli*	*E. coli*
GCF_900093145	*E. coli*	*K. pneumoniae*
GCF_001630835	*E. coli*	*E. coli*
GCF_003301495	*E. coli*	*K. aerogenes*
GCF_000459855	*E. coli*	*E. coli*
GCF_900448175	*E. coli*	*Citrobacter gillenii*
GCF_900497475	*E. coli*	*K. pneumoniae*

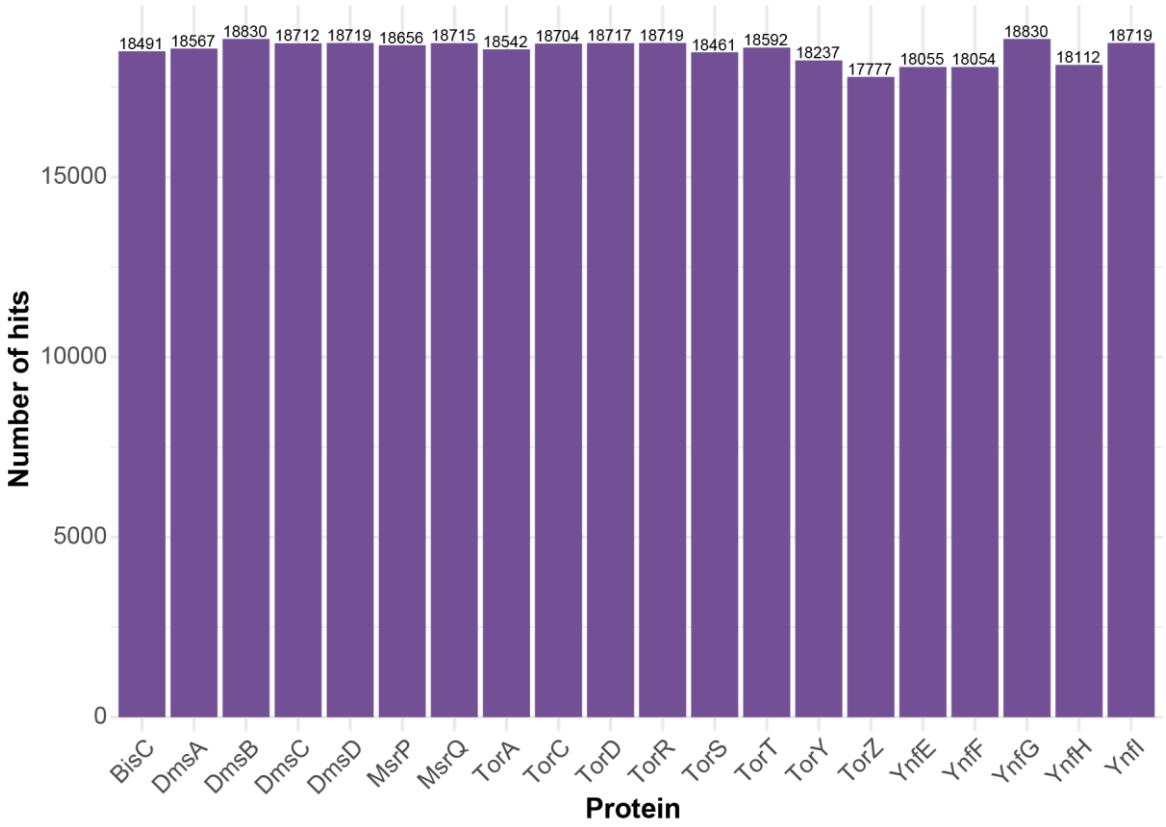


Figure 2.4. BLASTp results of 18847 publicly available *E. coli* genomes vs a TMAO metabolism protein database (hits with ≥90 % coverage and ≥70 % identity).

A ≥70 % identity cut-off was chosen as this helped prevent multiple proteins within the genomes giving hits on the same query sequence.

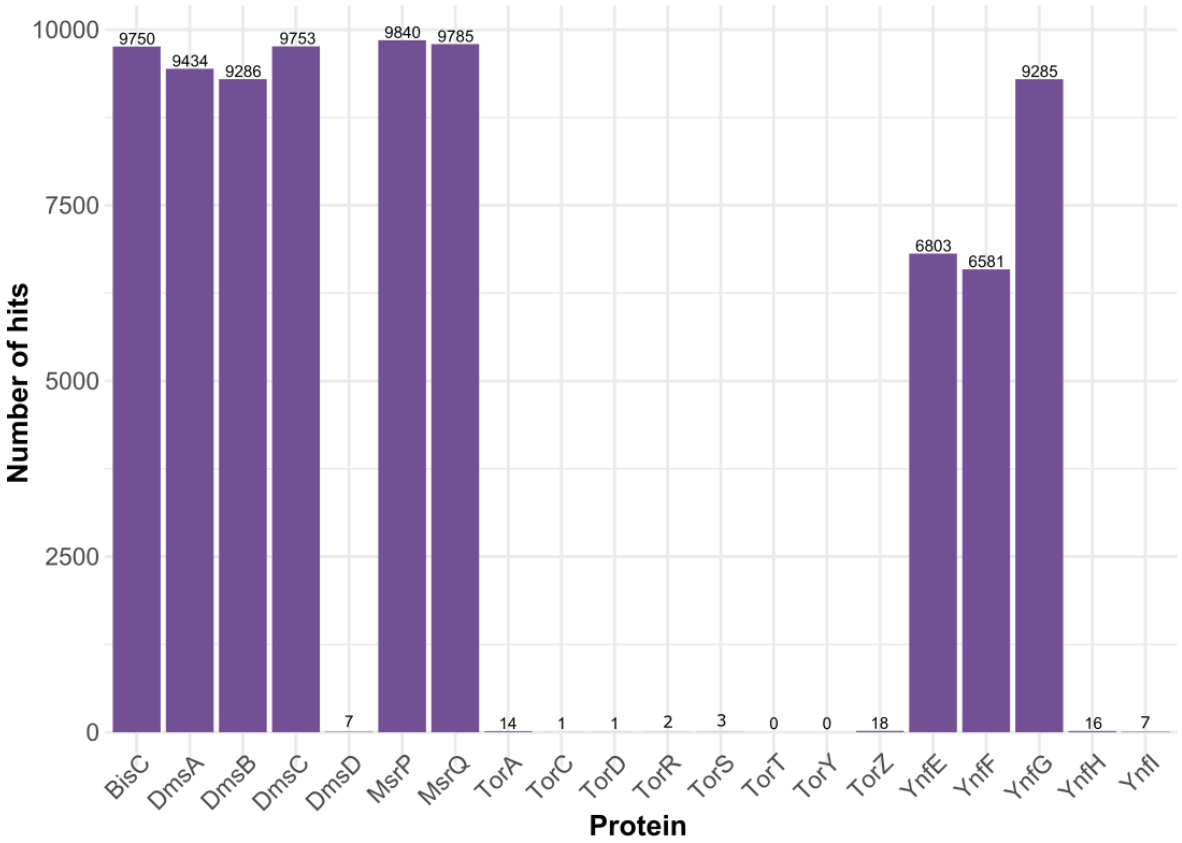


Figure 2.5. BLASTp results of 9898 publicly available *Klebsiella* spp. genomes vs a TMAO metabolism protein database (hits with ≥90 % coverage and ≥70 % identity).

All of the *Klebsiella* genomes available from RefSeq were checked using FastANI and were found to be predominated by *K. pneumoniae* (86 %) with relatively few representatives from other species (Figure 2.6). *K. variicola* subsp. *variicola* was the second most identified species but represented only 3.5 % of the genomes examined.

2.5.3 Neither TR1 nor TR2 is prevalent in bacterial species found in the human gut

Work had previously been done examining the aerobic and anaerobic respiration capabilities of human gut bacteria [159]. Bacterial genomes that were found to encode one or more of the proteins being examined in this study were ran through BLASTp vs the TMAO metabolism database. The only species of bacteria found to consistently encode for TR1 was *E. coli*, with *Salmonella* and *Citrobacter* also being found to encode TR1 (Figure 2.7). The *Helicobacter*, *Gordonibacter*, *Eggerthella, Clostridium* and *Clostridiales* genomes were previously found to encode TMAO metabolism proteins [159] but in this work were not found to encode any of the examined proteins. The work carried out by Ravcheev and Thiele [159] used an e-value cutoff as opposed to the coverage and identity percentage used in this work.

Bacterial genomes that were identified as being important for TMAO metabolism in humans based on the presence of the TorA protein [117] were obtained using entries from the GTDB and ran through BLAST (Figure 2.8). The only genera that were seen to encode TMAO metabolism proteins were *Burkholderia*, *Citrobacter* and *Raoultella*; with only some *Citrobacter* species giving a positive match for TorA (Figure 2.9). *Escherichia* and *Klebsiella* (covered above) along with *Marinobacter* and *Candidatus* spp. were not included.

A large number (n = 4644) of human gut MAG reference genomes were then examined for the presence of TMAO metabolism proteins [1]. Of these 4644 genomes, 122 were found to encode at least one TMAO metabolism protein (Figure 2.10). Thirty of these 122 were found to encode entire operons that would suggest a functional TMAO metabolism pathway (Figure 2.11), with an additional 13 being found to encode a complete DMSO reductase after additional analysis (Table 2.5).

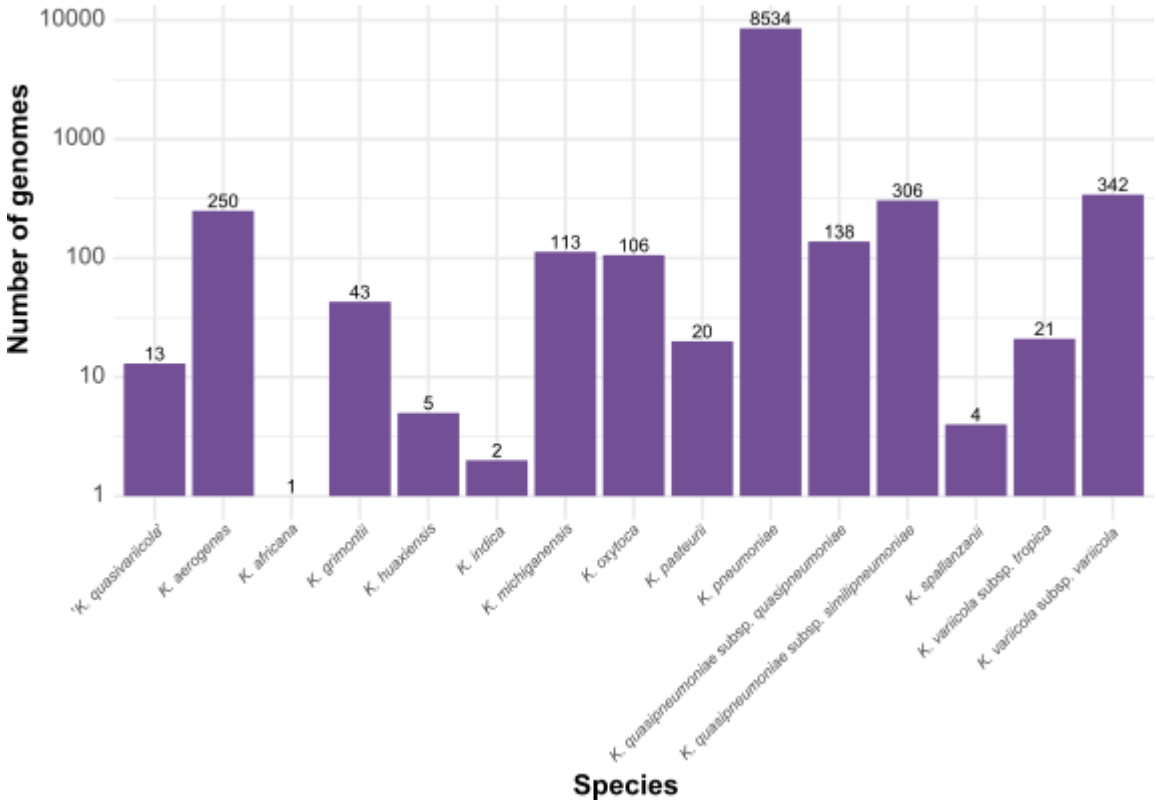


Figure 2.6. A breakdown of the number of different *Klebsiella* species genomes examined in this work.

The species of each genome was determined using FastANI with a >95 % cutoff for the ANI value [152,153]. Most (86 %) of the genomes were K. pneumoniae with the remaining Klebsiella species comprising no more than 3.5 % of the genomes each.

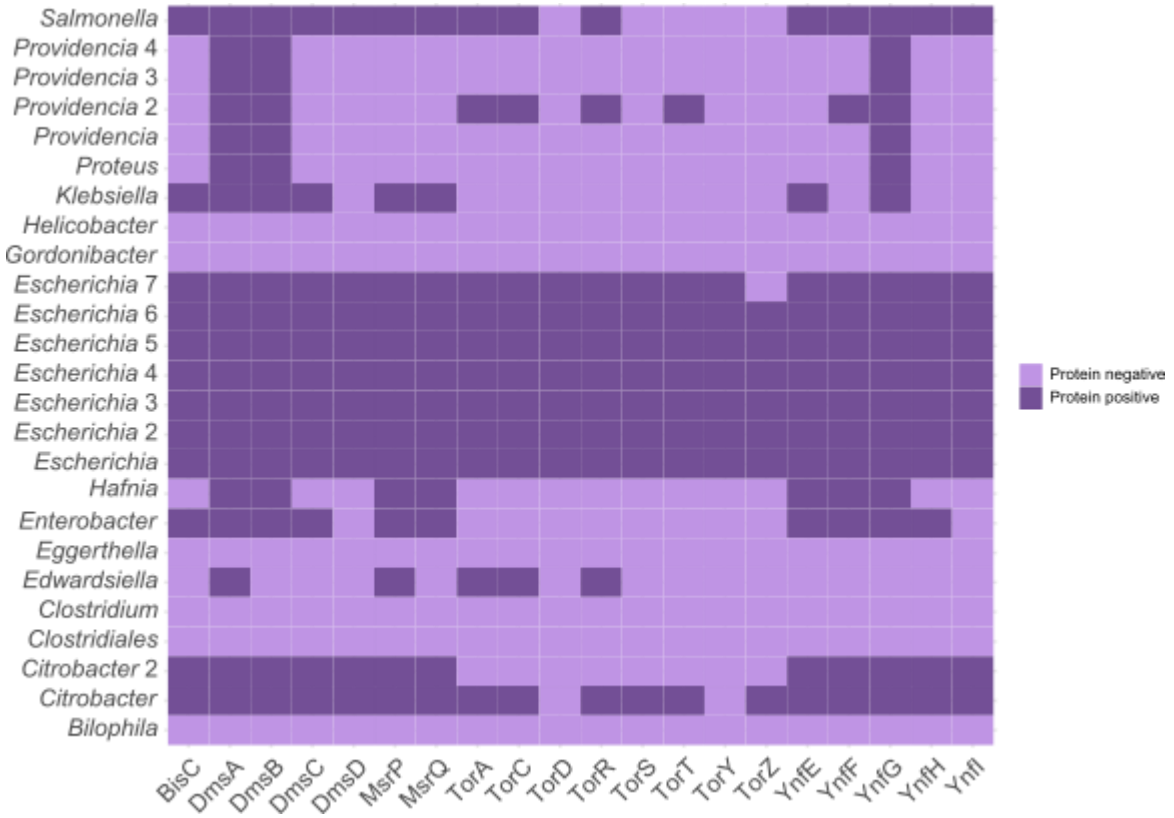


Figure 2.7. Results of BLASTp analysis of genomes previously found to carry TMAO metabolism proteins [159] (hits with ≥90 % coverage and ≥70 % identity).

E. coli strains that were examined were found to encode all proteins, except for *E. coli* 7 which did not encode TorZ.Four other genomes were found to encode Tor proteins. TorZ was only encoded by Citrobacter 1, excluding the E. coli.

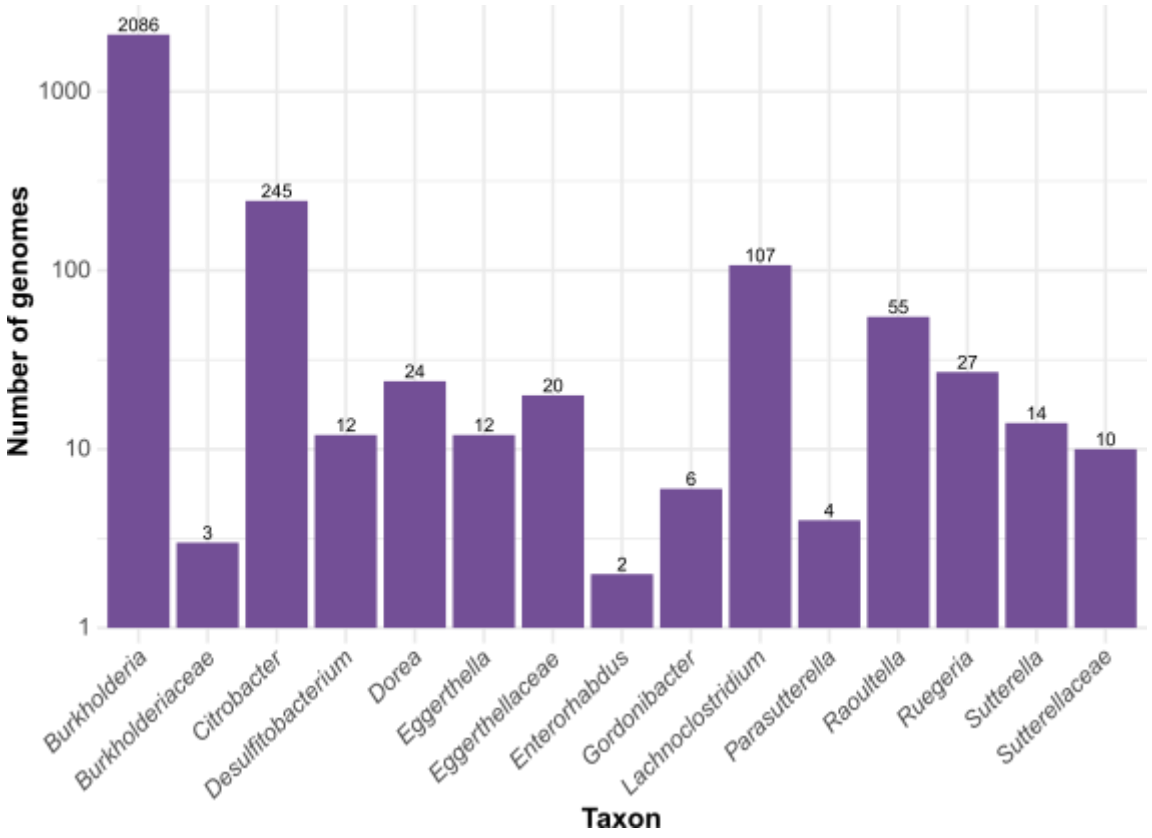


Figure 2.8. Number of genomes taken from GTDB.

The taxa examined were chosen based on their encoding of TorA as described in Jameson *et al* [117].

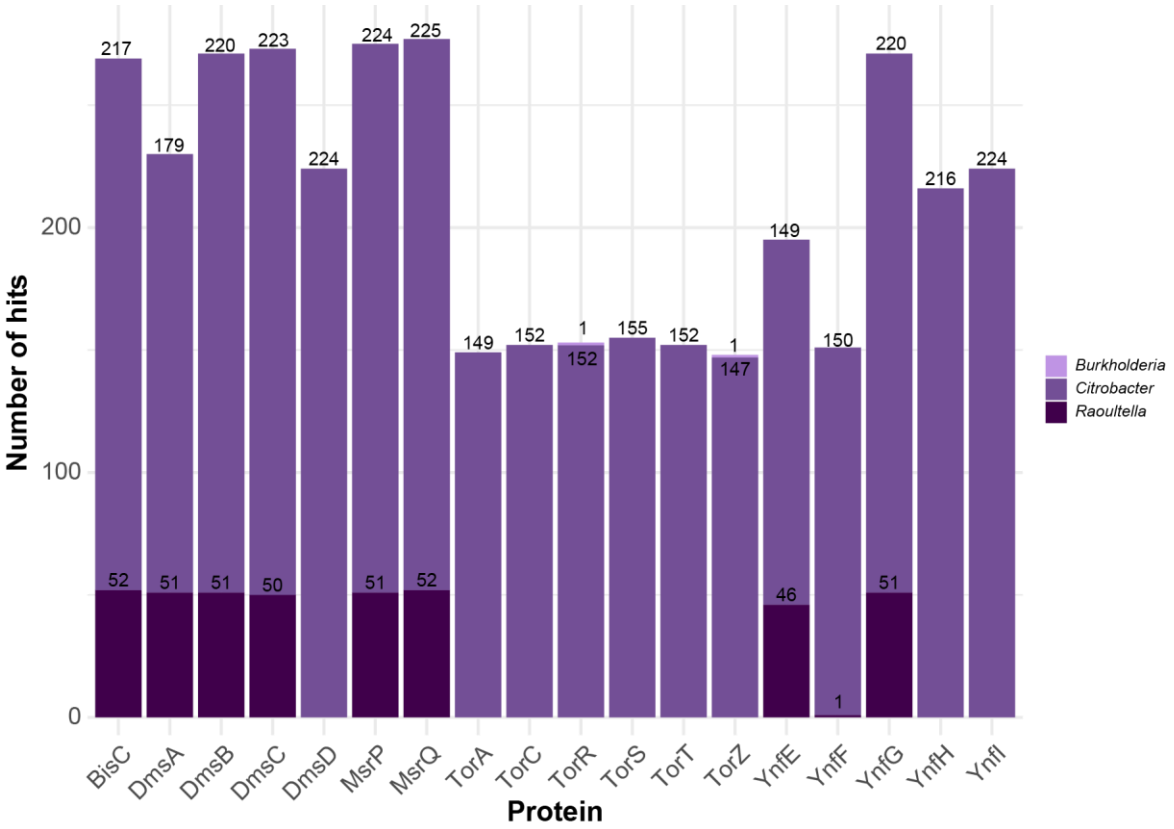


Figure 2.9. BLASTp results of GTDB 2627 bacterial genomes vs a TMAO metabolism protein database.

The genomes used were from taxa identified as being important for TMAO metabolism in humans by Jameson *et al* [117]. Citrobacter gave the highest number of hits, with some from Raoultella and 2 hits from Burkholderia.

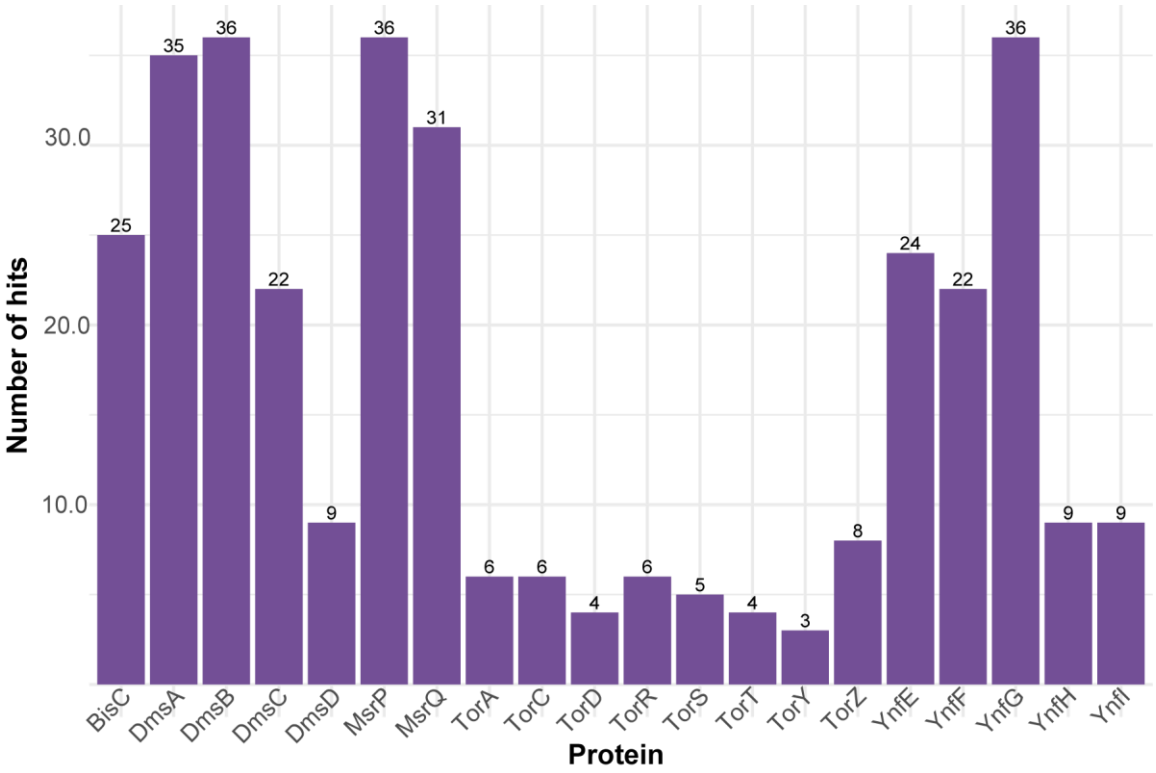


Figure 2.10. BLASTp results of 4644 human gut MAG reference genomes [1] vs a TMAO metabolism database.

The numbers of Tor proteins are much lower compared to other pathways, notably dms and msr. This is likely because of the often incomplete nature of MAGs. Details of the MAG reference genomes can be found in the original publication [1].

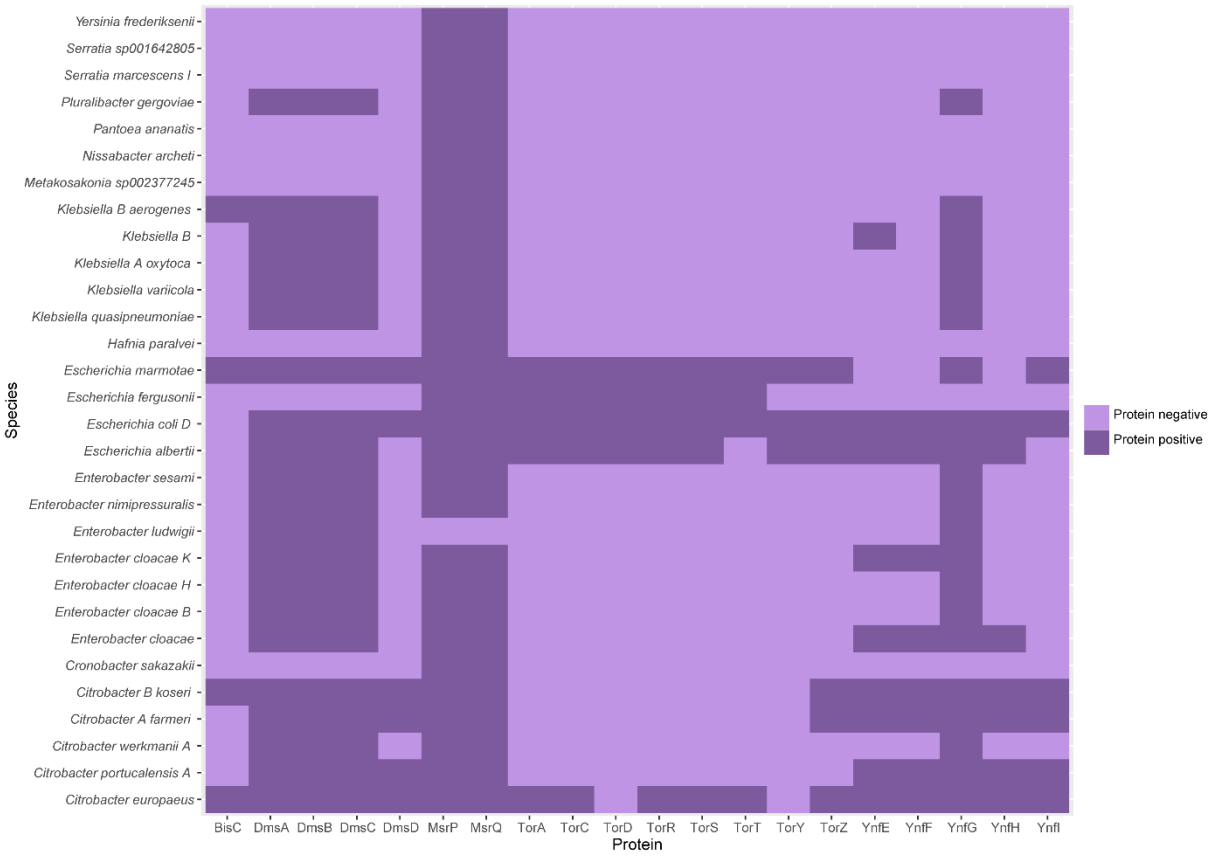


Figure 2.11. BLASTp results from Figure 2.10 when filtered to only show hits that are part of a contiguous operon, suggesting that these operons could be functional.

Thirty genomes of the 4644 MAGs met this criterion.

Table 2.5. Species of MAG that were found to carry DmsC along with DmsA and DmsB after additional analysis

Species	**Other pathways present**
Hafnia paralvei	Yes
Providencia alcalifaciens A	No
Cronobacter sakazakii	Yes
Yersinia kristensenii	No
Yersinia massiliensis	No
Yersinia frederiksenii	Yes
Morganella morganii	No
Morganella morganii A	No
Proteus mirabilis	No
Yersinia bercovieri	No
Metakosakonia sp002377245	Yes
Nissabacter archeti	No
Haemophilus influenzae	No

2.6 Discussion

TMAO is a common metabolite in the human body that can have its levels within systemic circulation modulated by bacteria living in the gut [25,161]. Much of the research examining the microbial metabolism of TMAO focuses on the *torCAD* operon and its associated protein TR1, leading to the prevalence of the TorA protein being used to determine which bacteria are responsible for TMAO metabolism in the human gut [117]. The genera that were largely attributed with encoding for TorA in the human gut were *Escherichia, Klebsiella, Citrobacter, Sutterella, Gordonibacter* and *Eggerthella* [117]. These genera, along with specific strains of bacteria that had been shown to encode for a combination of TR1, TR2 and DMSO reductase [159], were examined in the present study.

Using the GTDB in tandem with NCBI GenBank and RefSeq 31370 genomes were downloaded and analysed using BLASTp vs an in-house curated TMAO metabolism database built with reference to the literature[155–158]. *E. coli* genomes (n = 18847) were examined to ensure that the methods of generating and filtering results were working as intended. The majority of genomes examined were found to encode all TMAO metabolism proteins (Figure 2.4), which was expected as all of the reference proteins have been described in *E. coli* and the sequences used in the database were from *E. coli*. *Klebsiella* genomes (n = 9896) were examined, with this genus being more thoroughly examined due to the available number of genomes available and their relevance to other work currently being undertaken within the laboratory. The results here show that less than 1 % of publicly available *Klebsiella* genomes encode for either TR1 or TR2, although many strains still encode for DMSO reductase (Figure 2.5). This lack of both TR1 and TR2 was still seen even when identity thresholds were lowered to 50 % (Appendix A). This suggests that *Klebsiella* may play an important role in human gut TMAO metabolism, despite its lack of Tor proteins, as it may metabolise TMAO to TMA via DMSO reductase instead. Across the other 15 taxa that had genomes examined in this work, only *Citrobacter* spp. gave any hits for TorA and *Raoultella* spp. were the only other taxa that displayed potential for TMAO metabolism, being shown to encode DMSO reductase (Figure 2.9). The 15 taxa chosen here were picked based on their relevance as suggested by Jameson *et al*. [117].

In a 2021 paper by Almeida *et al*. [1]. a dataset of 204,938 reference genomes from 4644 different prokaryote species was compiled from publicly available genomes. A dataset of

4644 genomes was then created using the highest quality genomes from each species, referred to as the UHGG catalogue. This dataset was used in the work presented herein to examine the prevalence of TMAO metabolism proteins in human gut bacteria. This dataset was used as previous work had been completed using MAGs from three different studies [162–164] that have all since been incorporated into the UHGG catalogue, along with a large number of other publicly available genomes. This allowed for a much more comprehensive look into TMAO metabolism in human gut-associated bacteria.

The results presented here show that TorA may not be as prevalent in human gut bacteria as previously thought (Figure 2.10, Figure 2.11), and that even in bacteria that appear to encode TorA TR1 may not be produced due to a lack of both TorC and TorD. The proteins required to create DMSO reductase appear to be much more prevalent in human gut bacteria, which could mean that this enzyme and bacteria that encode it may be much more important to TMAO metabolism in the human gut than previously thought.

The TorZ protein appears to have been used in the generation of the consensus sequences used by Jameson *et al*. [117]. This protein however also does not seem to be as prevalent as has previously been shown, suggesting that TR2 may also not be as important to TMAO metabolism as DMSO reductase.

During the analysis of the results from the UHGG catalogue [1] it was noted that several species had positive hits for DmsA and DmsB, but not DmsC (Figure 2.10, Figure 2.11). This prompted further analysis that found the DmsC protein in each of the genomes that presented a positive hit for DmsA and DmsB. DmsC is the anchor protein for DMSO reductase and so the lack of similarity between the DmsC proteins found (Figure 2.12) may show a difference in the inner membrane proteins of some bacterial species.

Of particular interest is the BisC protein, biotin sulfoxide reductase. This protein is predicted to be encoded by a larger number of the genomes examined than either TR1 or TR2 (Figure 2.5, Figure 2.9, Figure 2.10) but has not been studied in any great detail. Interest in this protein comes from its similarities to TR2, as well as TR2's ability to reduce biotin sulfoxide. BisC from *R. sphaeroides* has been shown to reduce TMAO when expressed in *E. coli* [149], suggesting that the native enzyme may also have this ability.

The high presence of both MsrP and MsrQ is also noteworthy as MsrP has been shown to reduce TMAO [146]. MsrPQ appears to be present in most *Enterobacteriaceae*, suggesting that this family of bacteria may be the most important group of bacteria with respect to

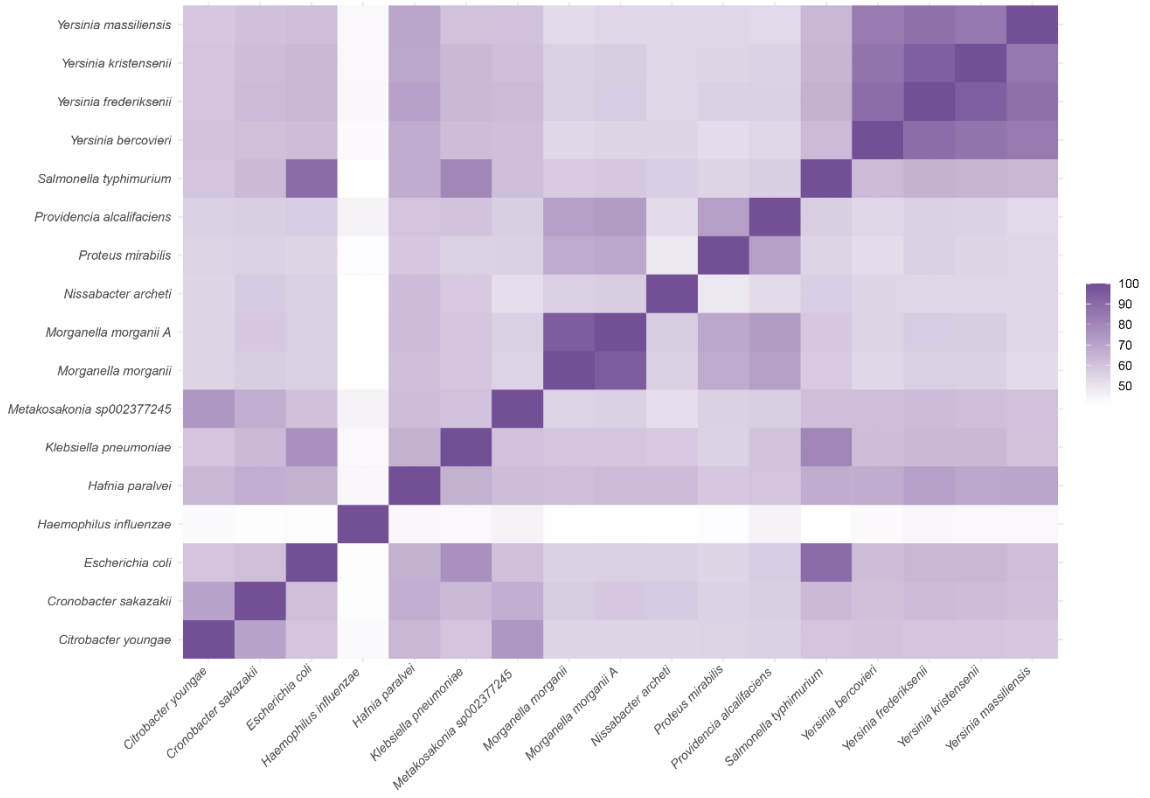


Figure 2.12. Comparison of different DmsC proteins found in genomes that were found to carry both DmsA and DmsB.

Enterobacteriaceae members were added for an extra comparison. Scale is % identity between amino acids sequences.

TMAO metabolism in the human gut. If MsrPQ is found to be an important part of TMAO metabolism it would be interesting as MsrPQ does not use TMAO to generate energy, instead it uses electrons from the electron transport chain in the repair of the cell membrane [145].

While nothing here can be claimed as concrete proof of function, what the results presented here do show is that while bioinformatic exploration of metabolism can be useful more physical laboratory work must be done to truly understand the systems at play in microbial TMAO metabolism.

Chapter 3 Characterisation of L4-FAA5, a caecal isolate of *Klebsiella pneumoniae* subsp. *pneumoniae*

3.1 Introduction

The genus *Klebsiella* is part of the family *Enterobacteriaceae* and comprises 19 different species [165]. *Klebsiella* spp. are ubiquitous, being found in soil, sewage, and plants, as well as being commensal organisms on the mucosal surfaces of mammals [166]. *Klebsiella pneumoniae* is a medically relevant species responsible for large numbers of multidrug-resistant nosocomial and community-acquired infections [167]. *K. pneumoniae* is a commensal organism and an opportunistic pathogen infecting the mucosal tissues of the respiratory and urinary tracts, and wounds [167].

In the human gut *K. pneumoniae* is present in the colon or stool in between 5 and 35 % of healthy individuals in Western countries, with these numbers increasing to between 18.8 and 87.7 % in Asian countries [168]. However, *K. pneumoniae* has been shown to colonise only the small intestine in mice, as well as being one of the most prevalent bacteria in human ileostomy samples when compared to its presence in human colostomy bag samples [169,170]. Some strains of *K. pneumoniae* have been found to directly impact human health in non-infectious ways. It is known that microbes in the human gut can utilise glucose in a fermentation pathway (Figure 3.1) that creates ethanol in high enough levels that it can begin to impact health in a minority of individuals [171]. *K. pneumoniae* has been found to be one of these microbes, with high ethanol-producing strains being isolated from a Chinese patient suffering from non-alcoholic steatohepatitis and gut fermentation syndrome [172]. Mice were found to develop hepatic steatosis when colonised by these high ethanol-producing *K. pneumoniae* in a manner similar to mice that were directly fed ethanol. Mice colonised by these bacteria were also found to have an increased blood alcohol content when they were fed glucose. This presented as an increase in serum levels of aspartate transaminase and alanine transamidase (indicative of liver damage), and liver levels of triglycerides and thiobarbituric acid reactive substances, which are all markers of

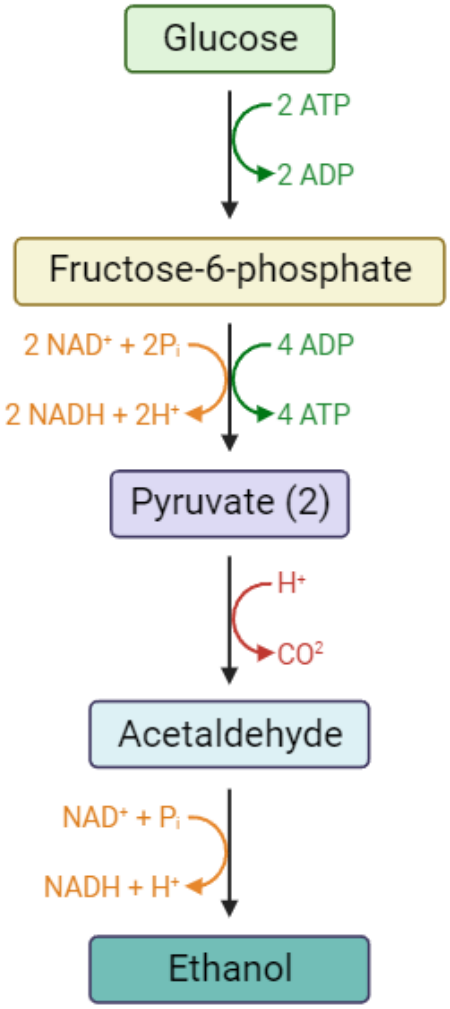


Figure 3.1. Simplified pathway to produce ethanol from glucose.

Glycolysis converts glucose to pyruvate. This pyruvate is then converted to acetaldehyde by pyruvate decarboxylase, which is then converted to ethanol by alcohol dehydrogenase [171].

non-alcoholic fatty liver disease [172]. The colonised mice were also found to have a higher level of immune cell activity in their liver, potentially contributing to the development of steatohepatitis (i.e. deposition of lipid droplets in liver tissue). Further work showed that *Klebsiella*-specific bacteriophages could be used to eliminate these high ethanol-producing strains and relieve steatohepatitis [173].

K. pneumoniae strains of the sequence type ST323 have been implicated in IBD, with these strains being highly abundant in the faeces of IBD patients experiencing a disease flare-up [174]. Clinical ST323 isolates from Israeli patients have been found to induce a pro-inflammatory response in the colon of germ-free mice, as well as in non-germ-free mice that were colonised with an IBD-associated *K. pneumoniae* isolate [174]. In the non-germ-free mice inflammation was reduced upon the removal of the IBD-associated *K. pneumoniae* via bacteriophage treatment. Genomic analysis of the ST323 *K. pneumoniae* shows an increase in metabolic pathways related to amino acid, nucleotide, mannose, and fructose metabolism in strains of this sequence type [174].

K. pneumoniae subsp. *pneumoniae* L4-FAA5 was originally isolated from the caecal effluent of a healthy woman [175]. A caecal isolate of *K. pneumoniae* was selected over a faecal isolate for this work as *Enterobacteriaceae* from the caecum have previously been shown to produce more TMA from TMAO, when compared to faecal isolates [25]. Previous work completed using L4-FAA5 has involved generating the draft sequence of its genome, and examination of changes in its growth and metabolism in the presence of TMAO [175,176]. This previous work had shown that in anaerobic conditions L4-FAA5 exhibited a statistically significant ($p < 0.05$) increased growth rate when in the presence of 10 mM TMAO, when compared to lower concentrations of TMAO, as well as producing TMA from TMAO. This work also attempted to use qPCR to show the effects of TMAO on gene expression, but due to issues with primer design and gene target selection this goal was not met. In the original work the genes *torR* and *torZ* were targeted, and these genes have since been found to not be present in the L4-FAA5 genome. Also the genes *fdoG*, *fdoH*, and *fdoI* were targeted as genes that should be upregulated during anaerobic growth, but these genes are expressed in both aerobic and anaerobic environments [176].

The work presented here aims to further this work by improving upon the experimental design via the collection of more growth data, transcriptomic analysis and qPCR with

properly designed primers, and quantification of metabolites using NMR. Alongside this, this work aims to characterise the complete genome sequence of L4-FAA5, enabling it to be used as a model for studying *Klebsiella* TMAO metabolism in the human gut.

3.2 Methods

3.2.1 Sequencing and annotation of the L4-FAA5 genome

DNA extraction, Illumina sequencing, and Nanopore sequencing were carried out by MicrobesNG (Birmingham, UK) as described previously [177]. The returned assembly was then annotated using Bakta Web v1.7.0 with database v5.0 [178]. The start gene for the chromosome was set as *dnaA*, *repB* for plasmid pSidero, and *repA* for plasmid pTra using Geneious Prime 2023.0.1. The assembly was then visualised using GenoVi v0.2.16 [179]. CheckM (v1.1.6) was used to check the quality and completeness of the L4-FAA5 genome [180].

3.2.2 Genotyping of the L4-FAA5 genome

The tool FastANI v1.33 was used to confirm the identity of L4-FAA5, based on a >95 % average nucleotide identity (ANI) cutoff [152,153]. Thirteen different *Klebsiella* type strain genomes were used as references in this analysis (Table 3.2) [180]. The sequence type of L4-FAA5 was found using PubMLST (accessed 22/05/23) [154]. The K type and O antigen identities were found using the tool Kaptive v2.0.7 [181]. Characterisation of the plasmids was carried out using the online tool PlasmidFinder v2.0.1 [182]. gutSMASH v1.0.0 was used to predict the metabolic capabilities of L4-FAA5 with respect to its ability to create and utilise primary metabolites that are relevant to the human gut in anaerobic conditions [183]. Further metabolic characterisation was carried out using KEGG and its associated ghostKOALA tool [184]. Antibiotic resistance genes were predicted using The Comprehensive Antibiotic Resistance Database (CARD) resistance gene identifier v6.0.2 with the database v3.2.7 [185]. Virulence genes in the L4-FAA5 genome were predicted using the Virulence Factor Database (VFDB) tool VFanalyzer (accessed 23/05/23) [186]. NCBI BLASTp+ v2.5.0 was used to predict TMAO metabolism (as described in Section 2.4 Methods, Chapter 2) [155].

3.2.3 Strain

K. pneumoniae L4-FAA5 [175] was resuscitated from a frozen stock on nutrient agar and incubated aerobically at 37 °C overnight. The strain was sub-cultured onto fresh nutrient agar, and again incubated aerobically overnight at 37 °C. Nutrient agar was made by adding 1.5 % bacteriological agar (Oxoid, UK) to nutrient broth (Sigma-Aldrich, NutriSelect Basic cat. no. N7519-250G).

3.2.4 Growth in presence and absence of TMAO

Anaerobic broth was created by boiling nutrient broth (Oxoid) supplemented with 4 mL of resazurin (Sigma-Aldrich) per litre (25 mg/100 mL stock solution) to remove oxygen. 0.5 g per litre of L-cysteine hydrochloride (Sigma-Aldrich) was then added to reduce residual oxygen, before transferring the broth to an anaerobic cabinet (Don Whitley A35 anaerobic workstation; 10 % CO_2 10 % H_2 80 % N_2, BOC), where it was dispensed into 16 mm Hungate tubes (Sciquip), which were sealed in the cabinet and then autoclaved at 121 °C, 15 p.p.i. for 15 min. Each Hungate tube contained 10 mL of broth.

For each of three biological replicates, a Hungate tube containing anaerobic nutrient broth was inoculated with a single colony of L4-FAA5 taken from the nutrient agar. The Hungate tubes were incubated without shaking, for 16 h at 37 °C in a water bath. The OD_{600} values of the cultures were measured using an Ultraspec 10 cell density metre (Amersham Biosciences), and the CFU/mL for each determined so that the starting inoculum for each biological replicate could be determined.

Hungate tubes (10× A–C) were inoculated (25G Terumo needle, 1 mL syringe) with 100 µL of sterile water, while another set of tubes (10× D–F) was inoculated with 100 µL of 1 M TMAO (Sigma-Aldrich) made up in sterile water. The tubes were left on the bench for 15 min after inoculation to ensure any oxygen introduced into the tubes was reduced by the L-cysteine hydrochloride present in the medium. All media remained straw-coloured (indicative of anaerobiosis). Tubes A and D were each inoculated with 100 µL from one of the overnight anaerobic cultures, tubes B and E were each inoculated with 100 µL from another overnight anaerobic culture, and tubes C and F were each inoculated with 100 µL from the final overnight anaerobic culture. The tubes were placed in a shaking (200 left-to-right shakes per minute; Clifton) water bath and incubated at 37 °C. An uninoculated

Hungate tube containing anaerobic nutrient broth was used as a negative control (and blank for OD_{600} readings). OD_{600} values were read every 30 min, in triplicate for each biological replicate.

Samples for processing (cultivation, metabolomics, transcriptomics and assessment of cell size) were taken every 90 min. At t90, t180, t270, etc., a single tube from each biological replicate was taken. An aliquot (500 µL) was used to prepare a dilution series (10^{-1} to 10^{-6}) in nutrient broth for determination of cful/mL; 5 µL was applied to a microscope slide and heat-fixed for further examination. The remainder of the cultures were centrifuged at 4696 ***g*** (Heraeus Megafuge 16R; Thermo Scientific) for 10 min, after which the supernatants were filter-sterilized and stored at –80 °C for future metabolomic work. The cell pellets were stored at –80 °C for future transcriptomic work.

3.2.5 Analyses of growth curve data

Data was imported into RStudio 2022.02.2+485 and analysed using GrowthCurver v0.3.1 [187]. The script used to analyse and visualise data is presented in the Appendix B, along with the data imported into R. A 2-way ANOVA test with a Tukey's post-hoc test using the Benjamini-Hochberg correction was used to determine the statistical significance of differences between control and TMAO-grown samples**.**

3.2.6 qPCR primer creation and validation

Primers were created from the annotated genome data for L4-FAA5 with the use of the Primer3 online tool, with extra checks carried out using the program SnapGene Viewer (www.snapgene.com) [188]. Primers used can be found in Table 3.1. Primers were ordered from Macrogen (South Korea) and validated with PCR using Mango Mix (Meridian Biosciences, UK). Primers were validated with a 60 °C annealing temperature and a 30 second extension step.

Table 3.1. Details of PCR primers designed for this study from genome of L4-FAA5

Primer target	**Sequence 5' – 3'**
dmsA forward	AACCAAAACCATCTTAACCGCC
dmsA reverse	TTTTCCCCGTCCGTTTCCAC
bisC forward	AACCAAAACCATCTTAACCGCC
bisC reverse	TTTTCCCCGTCCGTTTCCAC
msrP forward	GAAGATAAGGTCGCCGGTTACA
msrP reverse	ATGGTCAAGGGTCAGAGGTTTC
fdnG forward	AGAGGTGGCAAAAGAGAACAAC
fdnG reverse	GCCATTTTGATCGTAGAGATCC
fndH forward	CCCAAGACATTATTAAACGCTCCG
fdnH reverse	AGGAGACATCGATAAGCTTGGC
FNR forward	ATTCAGTCTGGCGGTTGTGC
FNR reverse	TCGTTCAGAGTAAAGGGGATGC
phoE forward	ATTAATGATGATGGGCTTTGTGG
phoE reverse	TTGATCTTGCCGTACACATCCA
infB forward	TGGCCTCAGAGATTCAGACCTC
infB reverse	TTGCGAGGTCACAGAATCATCA
pgi forward	AAAACATCAACCCAACGCAGAC
pgi reverse	AAAACGGTCGCTATCTTTGGC

3.2.7 Analysis of the L4-FAA5 transcriptome

Genomic DNA for standard curves was extracted from L4-FAA5 using the Qiagen (UK) Puregene Yeast/Bacteria kit B as per the manufacturer's instructions for Gram-negative bacteria. A 96-well 0.2 mL QuantStudio 3 system (Applied Biosystems) was used for all qPCR. PowerUp™ SYBR™ green master mix (Applied Biosystems) was used for all qPCRs, which were carried out in MicroAmp optical 96-well qPCR plates with MicroAmp optical qPCR strip caps (Applied Biosystems). The qPCR protocol consisted of a 2 minute 50 °C step followed by a 2 minute 95 °C. There were then 40 cycles of 95 °C for 15 seconds and 60 °C for 30 seconds, during which fluorescence data was collected. Finally, there was a melt curve step of 60 °C for 1 minute, followed by 95 °C for 15 seconds when data was collected.

Two RNA extraction methods have been trialled. Throughout both methods RNaseZap (Invitrogen) was used to eliminate environmental RNases. The first method utilises TRIzol (Invitrogen). Cell pellets stored at -80 °C were defrosted on ice before being resuspended in 1 mL of ice-cold TRIzol and incubated at room temperature for 5 minutes. An aliquot (500 µL) of TRIzol suspension was transferred to a 5PRIME heavy phase lock gel tube (Quantabio) and 100 µL of chloroform (Sigma-Aldrich) was added. The tubes were then shaken vigorously for 20 seconds, or until suspension had gone opaque. Tubes were then incubated at room temperature for 3 minutes before being centrifuged at 16,000 ***g*** for 10 minutes. The top aqueous layer was then removed via pipetting and transferred to a tube containing 250 µL isopropanol (Sigma-Aldrich) and incubated for 10 minutes before being centrifuged at 16,000 ***g*** for 10 minutes. Supernatant was then removed, and the pellet was washed with 500 µL of 75 % ethanol before being centrifuged at 16,000 ***g*** for 5 minutes. Ethanol was then removed via pipetting and the pellet was left to air-dry for 15-30 minutes until no residue was visible. The pellet was then resuspended in 30 µL of nuclease-free water and incubated at 37 °C for 30 minutes. The second method used a Thermo Scientific GeneJet RNA purification kit as per the manufacturer's instructions.

Once RNA had been extracted using either method the quantity was measured using a Qubit 4 Flurometer (Invitrogen) with the RNA BR assay kit (Invitrogen). Integrity was measured via gel electrophoresis using a 3 % agarose gel, ran for 30 minutes at 100 V. DNA was removed from samples using DNase I (Thermo Scientific) by adding 1 µL DNase I per 10 µL of sample, along with 1 µL of DNase I buffer (Thermo Scientific) per 10 µL of sample,

and incubating at 37 °C for 30 minutes. DNase was inactivated by adding 50 mM EDTA (Thermo Scientific) to a final concentration of 5 mM and incubating at 65 °C for 10 minutes.

3.3 Results

3.3.1 The genome of L4-FAA5 was completed and annotated

To fully characterise the genetic and metabolic potential of L4-FAA5 a complete genome sequence was required. A MinION/Illumina hybrid assembly was returned to us by MicrobesNG. Within this data a full chromosomal sequence was assembled, alongside two complete, circular plasmids (Figure 3.2). CheckM showed that the full genome of L4-FAA5 is 99.43 % complete with 0.43 % contamination. The L4-FAA5 chromosome is ~5.3 Mb in size with a 57.55 % GC content; 4826 coding sequences (CDS), 86 tRNAs, and 25 rRNAs were identified within the chromosome. The genome is of high quality according to the criteria of Bowers *et al*. [189].

The larger plasmid, referred to as pSidero in this thesis, is ~162 kb in size with a GC content of 50.23 %. pSidero encodes 179 CDS and no tRNAs or rRNAs. The second plasmid, referred to as pTra in this thesis, is ~95 kb in size and has a 49.97 % GC content. 115 CDS were identified within pTra but there were no tRNAs or rRNAs.

3.3.2 Genotyping of L4-FAA5 and its plasmids

The highest ANI was 99.02 % with the genome of *K. pneumoniae* subsp. *pneumoniae* ATCC 13883^{T}, confirming the identity of L4-FAA5 as a strain of this species (Table 3.2). L4-FAA5 was found to be of the ST380 sequence type, K2 capsule type, and O1 antigen type. Characterisation of the plasmids was also carried out using the online tool PlasmidFinder [182]. pSidero was identified as an IncFIB plasmid and pTra as an IncFIA plasmid.

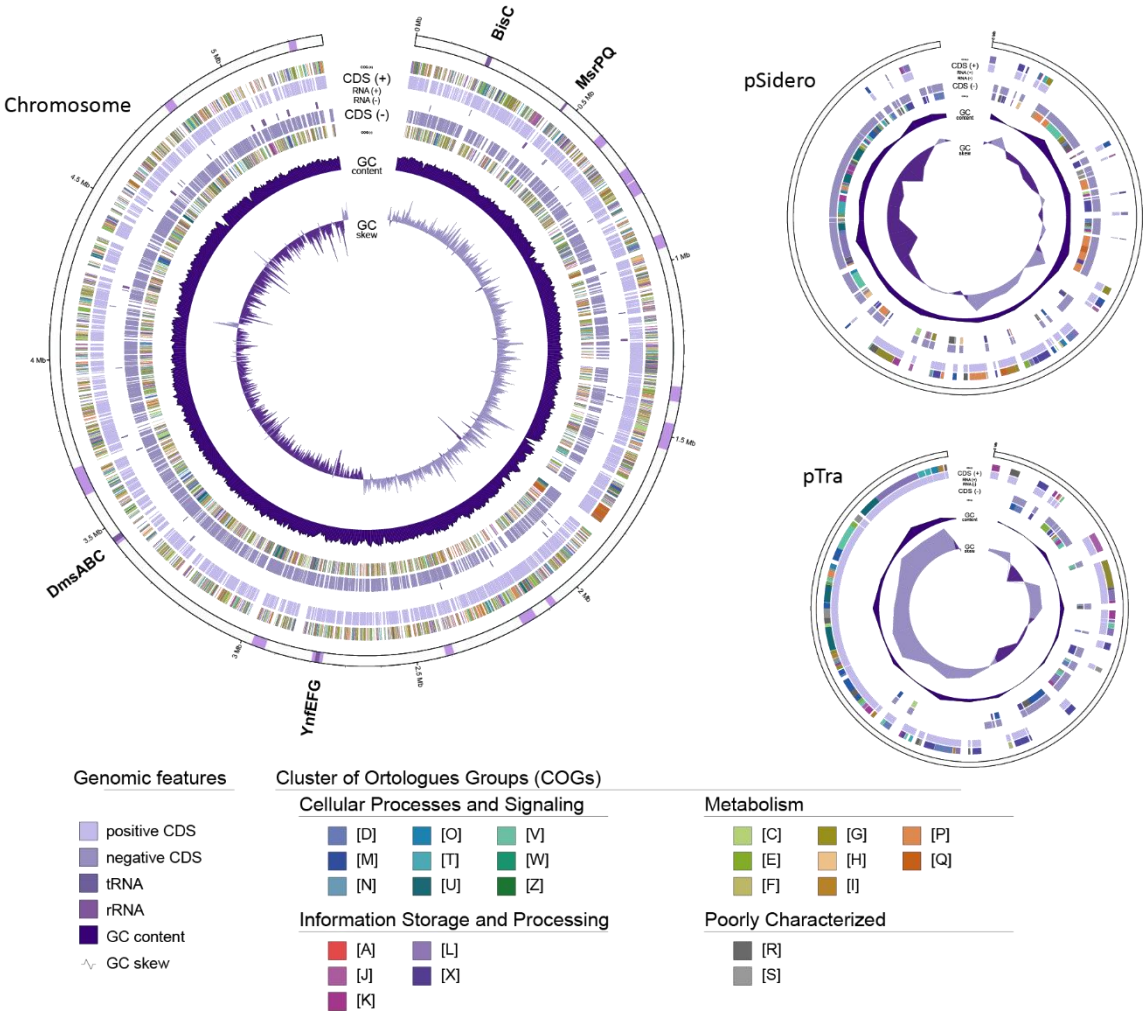


Figure 3.2. Genovi visualisation of the L4-FAA5 genome (sizes not to scale).

COG functional categories are as follows: **D** – Cell cycle control, division, chromosome partitioning. **M** – Cell wall/membrane/envelope biogenesis. **N** – Cell motility. **O** – Post-translational modification, protein turnover, chaperones. **T** – Signal transduction mechanism. **U** – Intracellular trafficking, secretion, and vesicular transport. **V** – Defence mechanism. **W** – Extracellular structures. **Z** – Cytoskeleton. **A** – RNA processing and modification. **J** – Translation, ribosomal structure, and biogenesis. **K** – Transcription. **L** – Replication, recombination, and repair. **X** – Prophages and transposons. **C** – Energy production and conversion. **E** – Amino acid transport and metabolism. **F** – Nucleotide transport and metabolism. **G** – Carbohydrate transport and metabolism. **H** – Coenzyme transport and metabolism. **I** – Lipid transport and metabolism. **P** – Inorganic ion transport and metabolism. **Q** – Secondary metabolite biosynthesis, transport, and metabolism. **R** – General function prediction. **S** – Function unknown.

TMAO metabolism proteins are labelled with text and marked in dark purple. Metabolic pathways predicted by gutSMASH are highlighted in light purple.

3.3.3 *In silico* prediction of the metabolic capabilities of L4-FAA5

The online tool gutSMASH was used to predict which metabolites L4-FAA5 can utilise and produce in anaerobic environments (Table 3.3) [183]. Two different gene clusters were identified as DMSO-TMAO reductases. These were further identified as the *dms* and *ynf* operons using NCBI BLASTp. The cluster starting at ~2.7 Mb of the chromosome is *ynf* and the cluster starting at ~3.4 Mb as *dms*. NADH dehydrogenase 1 and glycerol-3-phosphate dehydrogenase were predicted to be encoded by the L4-FAA5 chromosome, with these enzymes allowing NADH and glycerol-3-phospate to be used as electron donors in aerobic respiration [190]. Other clusters predicted to be involved in the electron transport chain are a nitrate reductase, formate dehydrogenase, and the *rnf* complex which is involved in anaerobic electron transfer [191]. The ability to cleave glycine to produce ammonia, CO_2, and electrons used to convert NAD+ to NADH was also predicted to be present in L4-FAA5 [192]. The PFOR II pathway was also predicted to be present in L4-FAA5, conferring the ability to convert pyruvate to acetate [193]. Other gene clusters related to SCFAs were two different clusters predicted to convert fumarate to succinate, a cluster predicted to breakdown threonine to propionate, and a cluster predicted to breakdown propanediol to propionate [194]. L4-FAA5 was also predicted to be able to breakdown ethanolamine to ethanol and acetyl-CoA, hydroxybenzoate to phenol, arginine to putrescine, and histidine to glutamate.

L4-FAA5 was also found to encode the TMAO metabolism proteins BisC, MsrP, MsrQ, DmsA, DmsB, DmsC, YnfE, YnfF, and YnfG (Table 3.4). Genes from the same metabolic pathway were found to be contiguous within the genome.

These pathways, except MsrPQ, were also identified using KEGG's ghostKOALA tool (Figure 3.3) [184]. Within KEGG none of these pathways were identified as being relevant for TMAO metabolism and the *torCAD* operon was not found. According to KEGG alone L4-FAA5 has no ability to reduce TMAO to TMA or other metabolites.

Table 3.2. Results of FastANI analysis of L4-FAA5 genome vs 13 *Klebsiella* reference genomes

Species and type strain identifier	ANI (%)	Assembly accession
K. pneumoniae subsp. *pneumoniae* ATCC 13883^T	99.02	GCA_000742135
K. africana SB5857^T	95.31	GCA_900978845
K. variicola subsp. *variicola* SB1^T	94.75	GCA_900977835
K. variicola subsp. *tropica* SB5531^T	94.32	GCA_900978675
K. quasipneumoniae subsp. *similipneumoniae* SB4697^T	93.90	GCA_900978135
K. quasipneumoniae subsp. *quasipneumoniae* 01A030^T	93.82	GCA_000751755
K. aerogenes ATCC 13048^T	86.10	GCA_003417445
K. michiganensis SB4934^T	84.54	GCA_901556995
K. grimontii 06D021^T	84.16	GCA_900200035
K. pasteurii SB3355^T	84.05	GCA_901563825
K. oxytoca ATCC 13182^T	83.90	GCA_900977765
K. spallanzanii SB3356^T	83.62	GCA_901563875
K. huaxiensis WCHKl090001^T	83.52	GCA_003261575

Table 3.3. Results of gutSMASH analysis of the L4-FAA5 genome

Most similar gutSMASH cluster	**Similarity (%)**	**L4-FAA5 genome coordinates**
DMSO-TMAO reductase *E. coli*	100	2,772,377 .. 2,795,018
DMSO-TMAO reductase *E. coli*	100	3,452,818 .. 3,481,549
Hydroxybenzoate to phenol *K. pneumoniae*	100	1,044,553 .. 1,066,573
NADH dehydrogenase 1 *E. coli*	100	1,479,667 .. 1,535,353
PFOR II pathway *B. thetaiotaomicron*	100	2,930,862 .. 2,969,799
Histidine to glutamate *K. pneumoniae*	100	3,612,856 .. 3,692,209
Fumarate to succinate *E. coli* SucDH	100	3,672,017 .. 3,692,569
Fumarate to succinate *E. coli* fum red	100	4,813,637 .. 4,836,154
Formate dehydrogenase *E. coli*	100	5,201,221 .. 5,225,818
Propanediol degredation *Salmonella enterica*	95	933,444 .. 968,042
Ethanolamine degredation *Salmonella typhimurium*	94	1,364,398 .. 1,400,679
Rnf complex *Clostridium porogenes*	83	2,395,497 .. 2,421,091
Glycine cleavage *Acetoanaerobium sticklandii*	80	756,601 .. 790,353
Nitrate reductase *E. coli*	80	2,148,653 .. 2,182,230
Threonine to propionate *E. coli*	50	2,071,542 .. 2,097,447
Arginine to putrescine *E. coli*	50	718,180 .. 755,452
Propanediol degradation *S. enterica*	29	594,777 .. 628,501

Table 3.4. BLASTp results of TMAO metabolism proteins encoded in the L4-FAA5 genome

Comparison with TMAO protein database outlined in Chapter 2.

BLAST hit	Identity (%)	Coverage (%)	L4-FAA5 genome coordinates
BisC	83.01	99.87	217,688 .. 220,018
MsrQ	78.71	99.50	460,851 .. 461,459
MsrP	86.75	99.40	461,459 .. 462,460
YnfE	85.70	99.88	2,780,452 .. 2,782,887
YnfF	70.77	98.77	2,780,452 .. 2,782,887
DmsB	90.73	99.51	2,782,898 .. 2,783,515
YnfG	88.29	99.51	2,782,898 .. 2,783,515
DmsC	77.35	99.65	3,467,618 .. 3,468,481
DmsB	90.73	99.51	3,468,483 .. 3,469,100
YnfG	88.29	99.51	3,468,483 .. 3,469,100
DmsA	89.80	99.88	3,469,111 .. 3,471,549

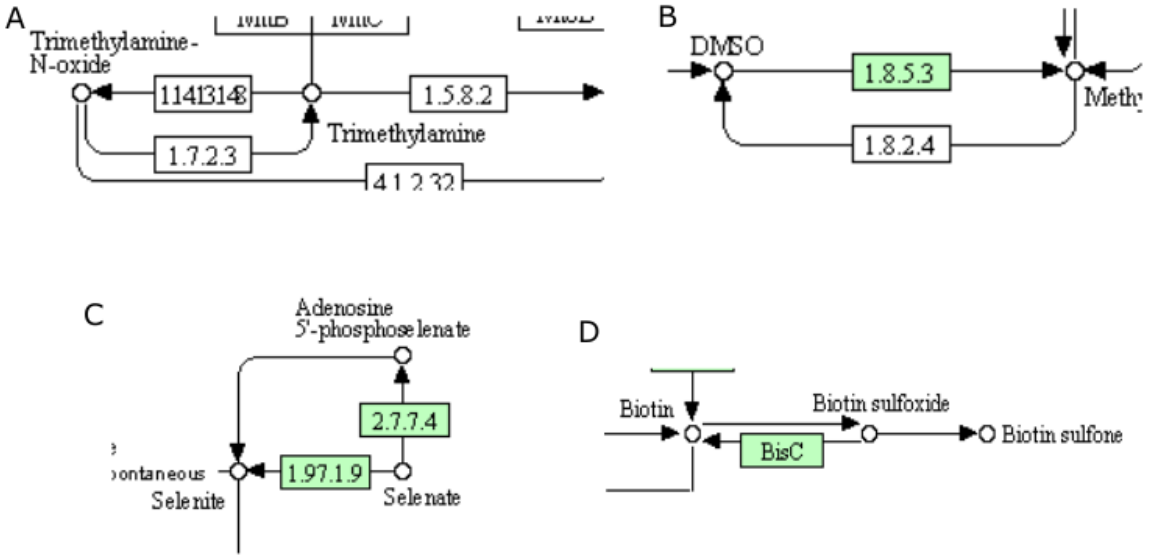


Figure 3.3. KEGG mapper results from the L4-FAA5 genome [184].

A) No *torCAD* genes were found by KEGG. **B)** The *dmsABC* genes were found by KEGG. **C)** The *ynfEFGH* genes were present in L4-FAA5. **D)** *bisC* is present in L4-FAA5. Within KEGG only the *tor* genes are directly linked to TMAO metabolism, despite other present operons being confirmed (*dms*, *ynf*) or predicted (*bisC*) to reduce TMAO.

3.3.4 *In silico* prediction of virulence genes and antibiotic resistance of L4-FAA5

The online tool VFAnalyzer of the VFDB was used to find virulence genes encoded by L4-FAA5 (Figure 3.4) [195,196]. L4-FAA5 was predicted to produce type 1 and type 3 fimbriae, as it encodes both *mrk* and *fim* genes. VFAnalyzer also predicted that L4-FAA5 would be able to create a capsule. These capsule production genes are not named by VFAnalyzer, but most are annotated as uncharacterised proteins by BAKTA. Named proteins include GalU, a phosphatase PAP2 family protein, and a Wzi family capsule assembly protein. Several siderophores were also predicted to be encoded by L4-FAA5, with enterobactin, yersiniabactin, and *E. coli*/*Shigella*-associated iron/manganese transport genes being chromosomally encoded and salmochelin and aerobactin being carried on pSidero. Capsule and siderophore regulatory genes *rcsA* and *rcsB* were also identified. A type 6 secretion system was predicted to be able to be produced due to the presence of several structural genes. L4-FAA5 was also found to encode genes required to produce the toxin colibactin. Despite the *rmpA* gene being present in VFDB it was not detected by VFAnalyzer, even though previous work had identified it in L4-FAA5 by PCR [175]. Primers used to target *rmpA* were analysed using SnapGene Viewer, and found to bind to a section of pSidero with the coordinates 8826 .. 9356 [197]. BLASTx was then used on this DNA sequence, which identified it as RmpA, annotated by BAKTA as 'Regulator of mucoid phenotype'. A copy of *rmpA* was also found on pTra annotated by BAKTA as 'RmpA2 protein'.

Antibiotic resistance genes were also identified in the genome of L4-FAA5 by the online BLAST database CARD [185]. L4-FAA5 was predicted to be resistant to a wide range of antibiotics, utilising several methods of resistance (Figure 3.5). All predicted antibiotic resistance genes were found in the chromosome, with none being present on either plasmid. Previous phenotypic work had found L4-FAA5 to be sensitive to ampicillin, cefotaxime, ceftazidime, amoxicillin–clavulanic acid, ertapenem, meropenem, ciprofloxacin, amikacin, gentamicin, tobramycin and colistin according to EUCAST

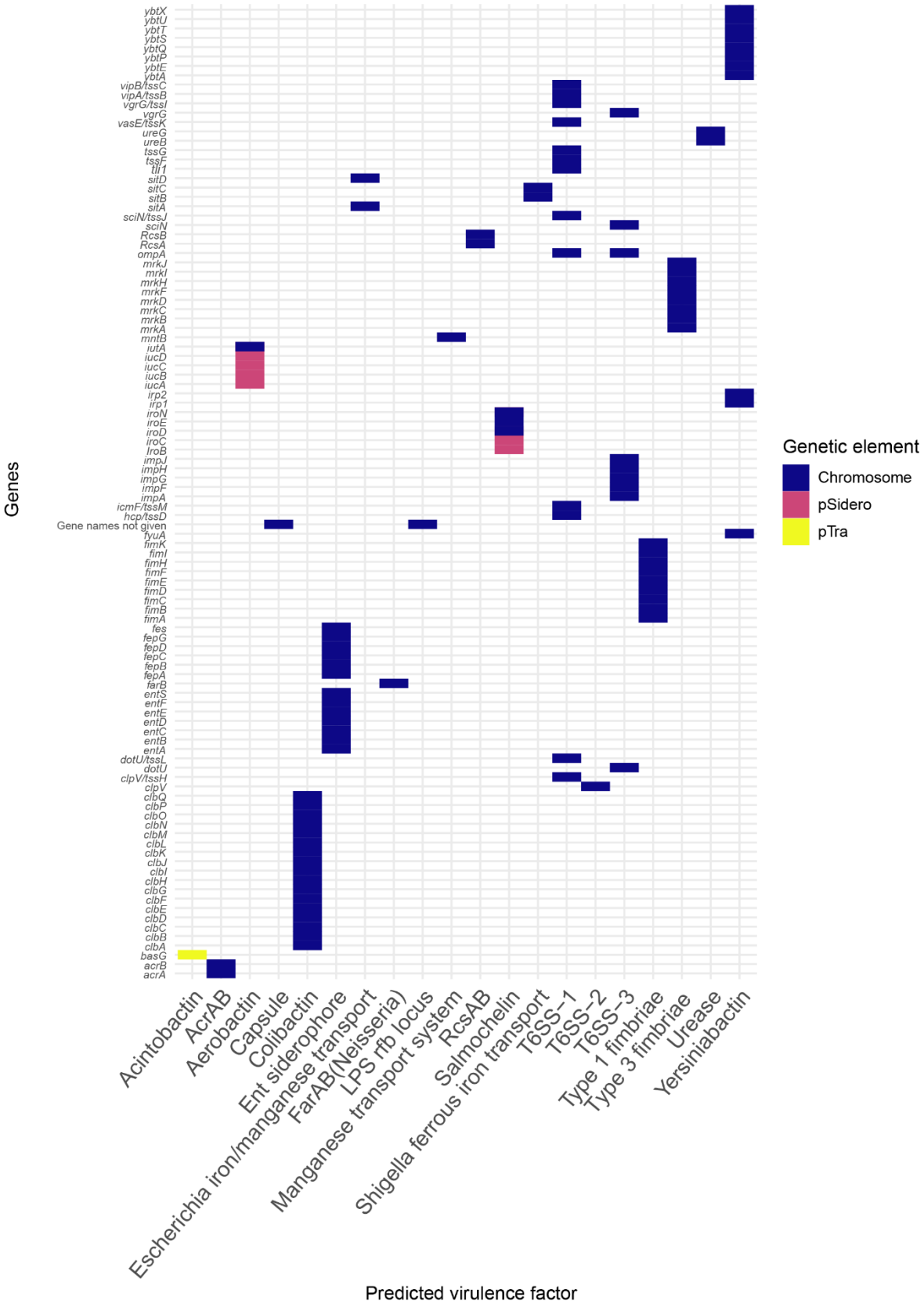


Figure 3.4. VFAnalyzer results from the L4-FAA5 genome [196].

The only virulence genes detected on the plasmids were the siderophores aerobactin on pSidero and acintobactin on pTra.

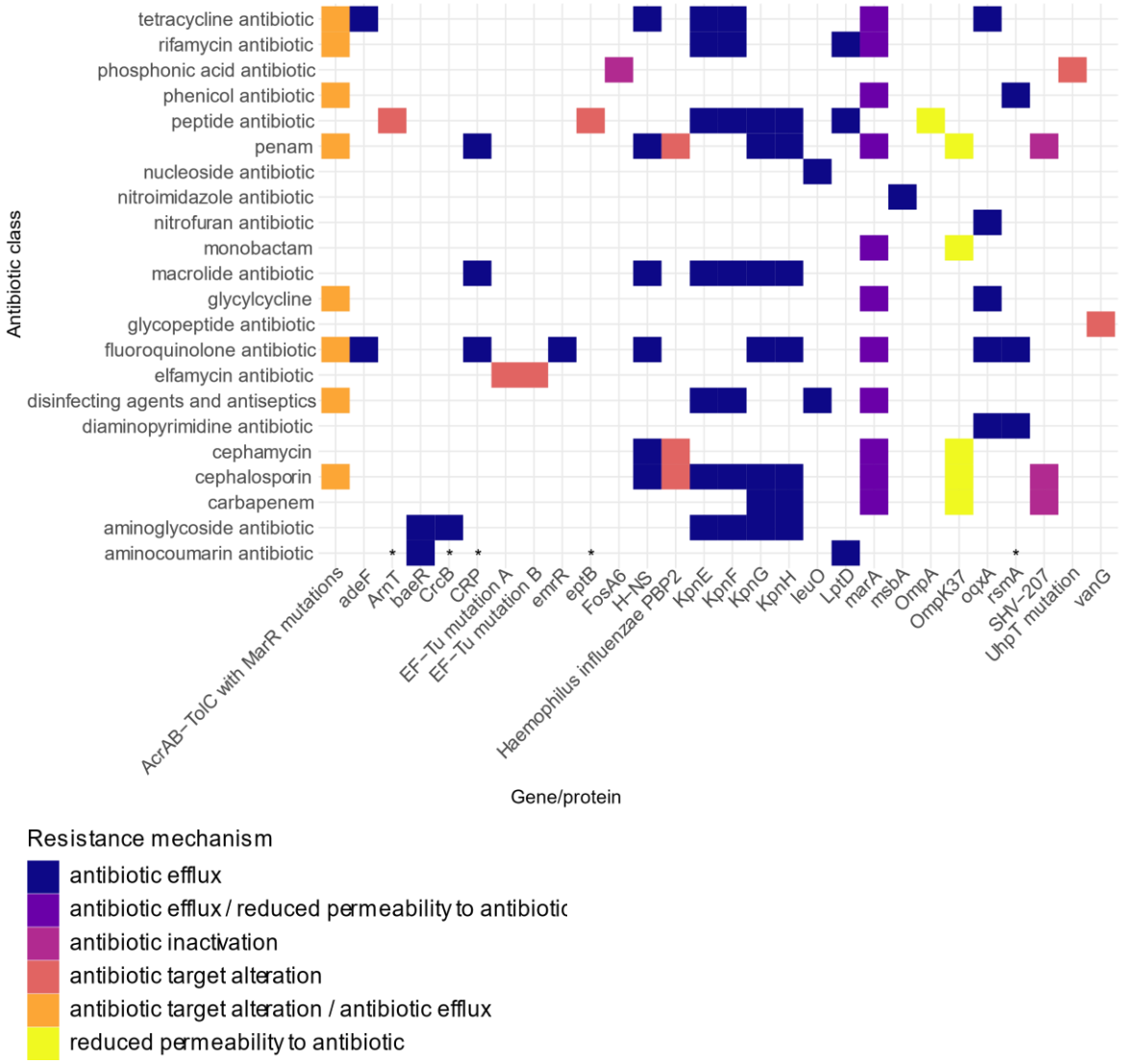


Figure 3.5. Antibiotic resistance genes and proteins detected in the L4-FAA5 genome by CARD.

Genes/proteins marked with * were labelled as 'perfect' hits by CARD' online RGI tool. Others were marked as 'strict'. Perfect hits are exact matches for resistance sequences/mutations in CARD, while strict hits are predicted based on CARD's obfuscated bit-score cut-off values [185].

guidelines; while being resistant to piperacillin–tazobactam, temocillin, cefoxitin, cefuroxime and aztreonam [198]. The genes identified by CARD as 'strict' hits conferred potential resistance to peptide, aminoglycoside, fluoroquinolone, macrolide, penem, diaminopyrimidine, fluoroquinolone, and phenicol antibiotics. CARD predictions do not appear to reflect the phenotypic resistance as L4-FAA5 was found to be susceptible to all tested classes of antibiotic with predicted resistance. For example, five genes conferring resistance to aminoglycosides were predicted to be present by CARD, with one of these hits being marked as 'strict'. However, the three aminoglycoside antibiotics tested (amikacin, gentamycin, and tobramycin) were found to be active against L4-FAA5.

3.3.5 Effect of TMAO on the anaerobic growth on L4-FAA5

There were statistically significant differences ($p < 0.05$) in the anaerobic growth of strain L4-FAA5 in the presence and absence of 10 mM TMAO (Figure 3.6). The empirical area under the curve was significantly different between the control and TMAO-grown cultures (Figure 3.6a, c): 87.1 ± 1.99 and 121.0 ± 2.91 relative units, respectively. TMAO significantly increased the growth of L4-FAA5. This was reflected in the cfu/mL counts determined for each time point sampled (Figure 3.6b; Table 3.5). Growth in TMAO also significantly reduced the doubling time of L4-FAA5 (Figure 3.6d): control 61.6 ± 3.78 min compared with TMAO 45.4 ± 1.93 min. In summary, under anaerobic conditions L4-FAA5 grew more prolifically and more quickly when in the presence of 10 mM TMAO.

3.3.6 Effects of TMAO on gene expression

Primers to be used in qPCR were first validated via traditional PCR and were shown to create single products (Figure 3.7). An extra band at over 1 kb was seen for FNR, but this was disregarded as the qPCR extension step was assumed to be short enough that this product would be unable to be created. After primer validation standard curves were generated to test primer efficiency (Figure 3.8). Melt curves showed that each pair of primers only generated a single product.

Table 3.5. Comparison (2-way ANOVA test with a Tukey's post-hoc test using the Benjamini-Hochberg correction) of cfu/mL data for control and TMAO-grown samples at each time point sampled.

Time (Mins)	p value	Adjusted p value
0	1.000	1.000
90	0.015	0.030
180	0.007	0.030
270	0.037	0.059
360	0.008	0.030
450	0.081	0.093
540	0.064	0.085
630	0.015	0.030

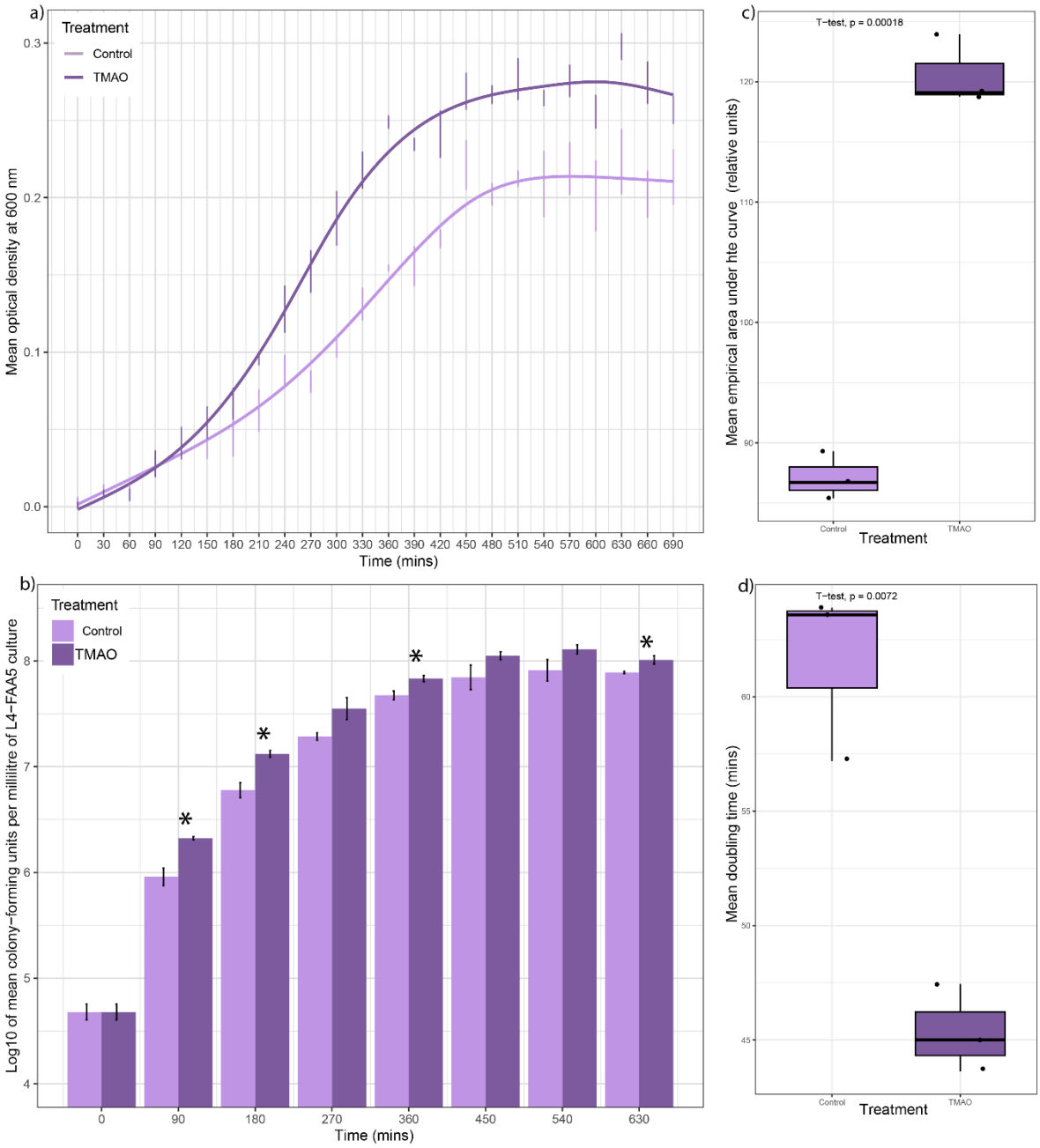


Figure 3.6. Summary of growth curve data for L4-FAA5 grown anaerobically in the presence and absence of 10 mM TMAO.

(a) Growth curves for L4-FAA5. The line of best fit was generated using the gam function of geom_smooth(); error bars are presented ± standard deviation. (b) Colony count data generated for samples taken every 90 min during the growth curve experiment. Data within each time point was compared: *, adjusted p value < 0.05. (c) Comparison of empirical area under the curve generated for the two growth conditions (auc_e generated using using GrowthCurver). (d) Comparison of doubling times (t_gen data generated using GrowthCurver). All data presented in (a–d) are shown based on three biological replicates (three technical replicates each). A 2-way ANOVA test with a Tukey's post-hoc test using the Benjamini-Hochberg correction was carried out to determine the significance of the colony count data (b).

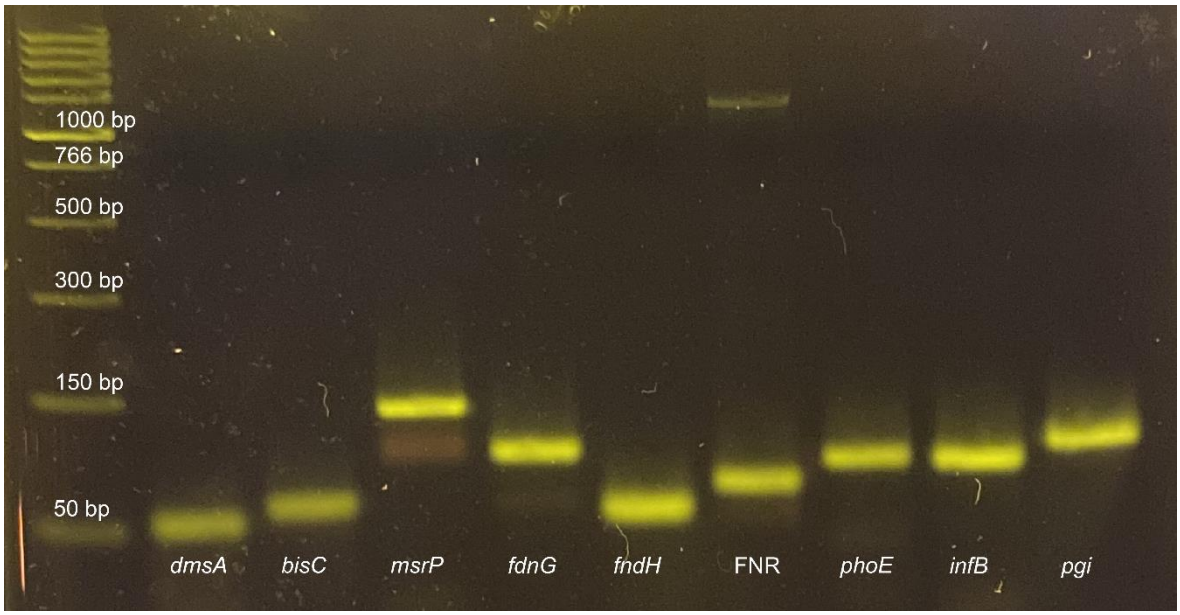


Figure 3.7. Results of traditional PCR used to validate primers to be used for qPCR. FNR appeared to generate a large PCR product. This was disregarded as the extension step in the qPCR protocol would be short enough to prevent this product from being produced.

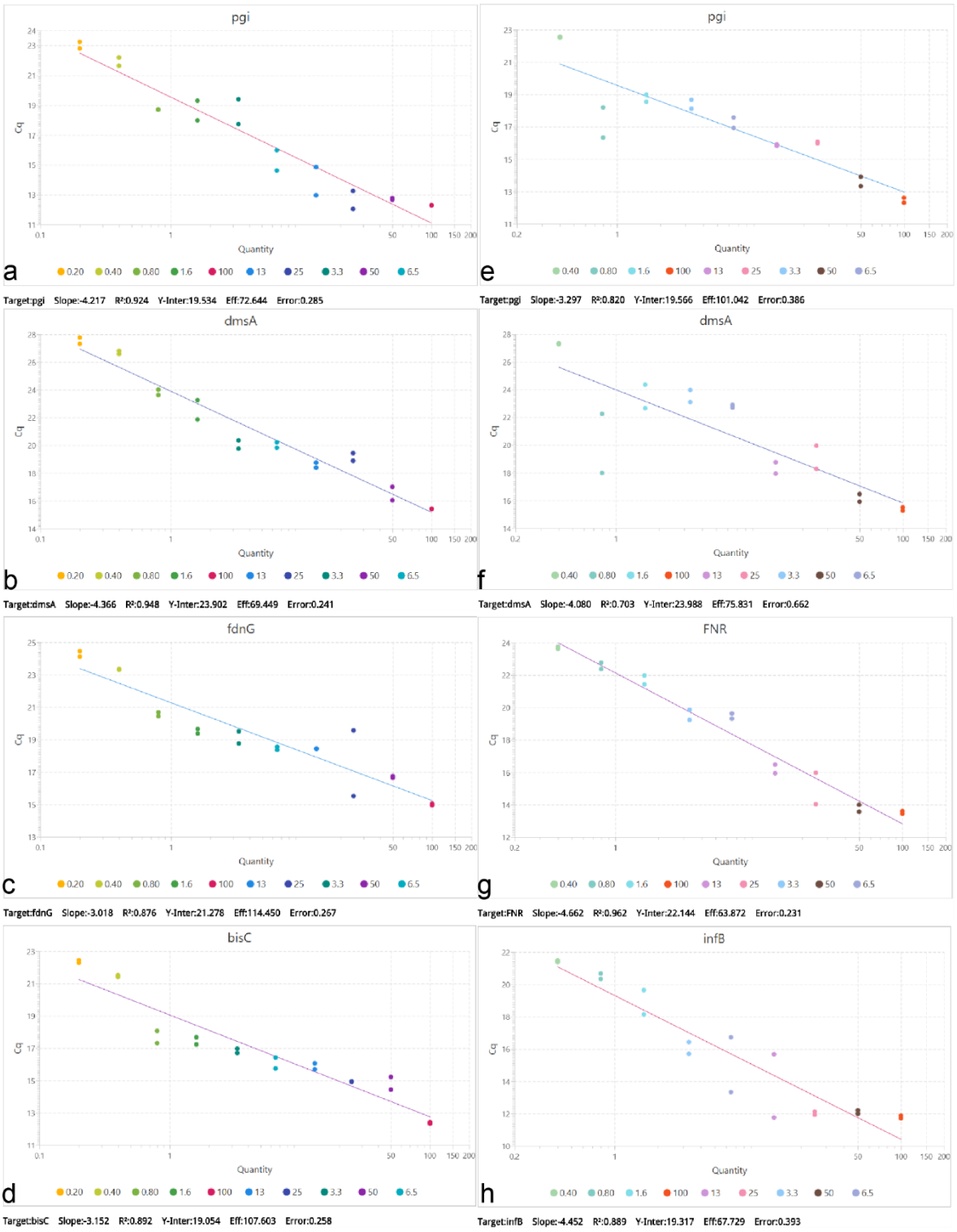


Figure 3.8. Results from the qPCR standard curve plates. Graphs a-d are from the first plate and graphs e-h are from the second. The x axis units are ng. Curves a, b, f, g, and h have low efficiencies (<90 %), which suggests an issue with primer design or standard preparation.

3.4 Discussion

Klebsiella spp. are ubiquitous in the environment, with *K. pneumoniae* important as both a nosocomial pathogen and a commensal organism [166,167]. *K. pneumoniae* L4-FAA5 had previously been isolated from the caecal effluent of a healthy woman and had been shown to exhibit an increased growth rate in the presence of TMAO in anaerobic conditions [175,176].

While the L4-FAA5 genome had been sequenced previously presented here is the complete assembly of the genome into three complete contigs, the chromosome and two plasmids. This allowed for a full characterisation of the L4-FAA5 genome. ANI analysis confirmed that L4-FAA5 is a *K. pneumoniae* subsp. *pneumoniae* isolate.

The first characterisation carried out based on the new genome data was prediction of metabolic potential of L4-FAA5 using the tools gutSMASH and ghostKOALA[183,184]. gutSMASH predicts the presence of gene clusters that are responsible for the anaerobic production of metabolites relevant to the human gut, while ghostKOALA predicts which genes can be assigned to KEGG pathways. gutSMASH predicted the presence of two TMAO reductase pathways, which were found to be the *dms* and *ynf* operons after further analysis using NCBI's BLASTp. Both pathways were found alongside *bisC* and *msrPQ* which were found using the methods (Section 2.4) described in Chapter 2. GhostKOALA/KEGG also identified these genes, excluding *msrPQ*. Within KEGG however, none of these pathways were linked to TMAO metabolism. This highlights the focus on *torA* as being the only pathway relevant for TMAO metabolism, as well as how this focus may cause issues with characterisation by ignoring alternative TMAO metabolism pathways. Other characterisation work involved using CARD and VFanalyzer to determine the antibiotic resistance and virulence genes, respectively, carried by L4-FAA5 [185,186]. VFanalyzer predicted the presence of capsule production genes in the L4-FAA5 chromosome, with the *rcsAB* capsule regulatory genes being encoded. While not found using VFanalyzer *rmpA* capsule regulation genes were also found on both pSidero and pTra. These regulatory genes usually confer a hypermucoviscous phenotype in *K. pneumoniae* [199], although this is not seen in L4-FAA5. The *rmpA* genes were not found by VFanalyzer but instead by secondary analysis using NCBI BLASTp, showing that relying on a single genome annotation tool can lead to an incomplete characterisation. L4-FAA5 was also found to carry several siderophores, which are crucial for host infection [199], both chromosomally and on pSidero.

Both type I and III fimbriae were predicted to be created by L4-FAA5 by VFanalyzer. These virulence factors play a role in adherence and biofilm formation, with type I potentially playing a larger role in the host and type II aiding adherence and biofilm formation on abiotic surfaces [199,200]. Type VI secretion system genes were also predicted to be carried by L4-FAA5 by VFanalyzer, which can aid virulence by delivering effector proteins directly into host cells, or by aiding colonisation via the killing of other bacterial cells to outcompete them [199]. The toxin colibactin is also worth noting as DNA damage caused by colibactins have been linked to the development of colorectal cancer [201].

By the results of this work CARD appears to be an unreliable way of predicting antibiotic resistance. L4-FAA5 was predicted to be resistant to 23 groups of antibiotics, but when compared to previously generated phenotypic data none of these resistances were seen.

L4-FAA5 has been shown to grow significantly faster in the presence of TMAO than without in anaerobic conditions. This supports the idea that TMAO is used in a form of anaerobic respiration by this strain, as anaerobic respiration is more efficient than fermentation and so would reduce the generation time of the cells. Due to the protein-stabilising effects of TMAO [75] the presence of it may also affect cell morphology, with future work aiming to examine this potential difference by using microscopy to measure the cell sizes of L4-FAA5 cells grown in the presence and absence of TMAO. Any other potential morphological changes will also be able to be examined using this method.

Work had been planned to examine the effects of TMAO on the transcriptome and metabolome of L4-FAA5. Gene targets for qPCR were chosen based the results of the genome annotation, as well as comments made in previous work [176]. The expression of the *msrP*, *fndH*, and *phoE* genes was planned to be included in the qPCR work, but it was decided to remove these genes from the analysis due to limits on time and resources. Other reasons for excluding these genes from this work are that the expression of *msrP* was not expected to be affected by TMAO as the MsrPQ proteins are involved in membrane repair and not energy production, with expression of MsrPQ being shown to be induced by the presence of hypochlorous acid as well as copper [145,202]. The expression of *fdnH* was expected to be the same as *fdnG* as they encode different subunits of the same protein, and as a control gene no change was expected from *phoE* and there were already two other control genes chosen. Standard curves were generated on the remaining targets although

there has been some issues with the reactions. Efficiencies for most targets are lower than desired, at approximately 60 %. Also, the efficiency for *pgi* changed greatly across two different plates. It is not clear what the issue with these standard curves is, but the low efficiencies may be due to primer design issues or pipetting errors with the standards used. Also causing problems with this work is the ability to extract high quality RNA for usage in cDNA synthesis for qPCR. At time of writing both a TRIzol-based method using phase-lock tubes and a kit-based method have been used, both leading to high quantities of RNA being extracted but the RNA being degraded and of low quality. RNA degradation during extractions usually comes from the presence of RNases in the environment. Precautions have been taken during extractions, but the issue is persisting. Current precautions involve usage of RNaseZap to eliminate environmental RNases, carrying out work in a laminar flow hood when possible, usage of PPE to protect samples from skin-borne RNase contamination, and usage of filtered pipette tips that have been certified as RNase and DNase free. Future work here revolves around ensuring that high-quality RNA can be extracted from samples and that primer design is not what is affecting the qPCR efficiencies.

Metabolomic work is still underway at time of writing. Currently method testing is being carried out with the aim of using NMR to quantify the amounts of TMAO in spent media from the growth experiments. Samples have been run but technical issues are currently holding this work back. Future work will involve getting the stored samples run through NMR to quantify the amounts of TMAO present once a method has been confirmed to work and all equipment is back in working order.

Overall, the main aims of this work, namely the effects of TMAO on gene expression and the effects of TMAO on the metabolome, have not been met. While samples for this work have been generated and stored issues with methodology have led to a delay in the collection of data. However, a thorough characterisation of L4-FAA5 has shown that multiple tools are necessary when attempting to predict an isolate's metabolic capability *in silico*. While tools like gutSMASH and ghostKOALA are very useful they may both miss key genes. Therefore, it is important to investigate genes of interest thoroughly using more manual methods and test predictions phenotypically in the laboratory where possible.

Chapter 4 BisC: a novel pathway for TMAO metabolism in *Klebsiella pneumoniae*

4.1 Introduction

The *Escherichia coli* gene *bisC* was first identified as being relevant in the reduction of biotin sulfoxide in 1973, before being identified as the structural gene for the enzyme biotin sulfoxide reductase BisC in 1981 [203,204]. BisC cloned from *Rhodobacter sphaeroides* and expressed in *E. coli* was then shown to reduce biotin sulfoxide to biotin, as well as several other substrates, including TMAO [149]. The *E. coli* BisC has been shown to reduce methionine sulfoxide, but no work has been done showing its activity against TMAO [148]. This reduction of cytoplasmic biotin sulfoxide and methionine sulfoxide has been theorised to aid in the protection of the cell against oxidative damage and in the scavenging of oxidised biotin [148,205]. This scavenging of oxidised biotin is particularly important as biotin is an important cofactor in a number of enzymes in *E. coli* [206]. These enzymes include the fatty acid biosynthesis enzyme acetyl CoA carboxylase, the Kreb's cycle enzyme pyruvate carboxylase, and the amino acid/fatty acid metabolism enzyme propionyl CoA carboxylase [206]. Overall BisC in species relevant to the human gut remains as a poorly characterised molybdoenzyme that may have potential TMAO reductase activity.

As a molybdoenzyme, BisC requires molybdenum to function in the form of a molybdenum cofactor (Moco) (Figure 4.1) [204]. In *E. coli* molybdoenzymes can be split into three families based on the structure of the Moco that they utilise, with BisC being in the DMSO reductase family [207]. The Moco that the DMSO reductase family utilises is the molybdenum guanine dinucleotide (MGD) co-factor bis-MGD. This family includes Dimethyl sulfoxide reductase, Trimethylamine reductase 1 (TR1), and Trimethylamine reductase 2 (TR2), which are the three enzymes in *E. coli* that are currently known to reduce TMAO to TMA [207]. Based on this and the high level of sequence similarity between BisC and TorZ, it may be possible for BisC to reduce TMAO. This would present a new TMAO utilisation pathway that would need to be considered when examining the metabolic capability of gut bacteria. Other molybdoenzymes in the DMSO reductase family are the membrane-associated enzymes nitrate reductase A (subunits NarG, H, and I), nitrate reductase Z (subunits NarZ, Y, and V),

and the periplasmic Nap nitrate reductase (subunits NapA, B, C, G, H)[207]. These three enzymes reduce nitrate and are produced in three different environmental conditions. Nitrate reductase A is produced in anaerobic conditions in the presence of high concentrations of nitrate, whereas Nap is produced in the presence of a low nitrate concentration [207,208]. Nitrate reductase Z is expressed during early stationary phase in aerobic conditions, whether nitrate is present or not [207,208]. The formate dehydrogenase Fdn (subunits G, H, and I) is a periplasmic enzyme that functions with nitrate reductase A to couple the reaction between the oxidation of formate and the reduction of nitrate in anaerobic conditions. The formate dehydrogenase Fdo (subunits G, H, and I) serves the same function but coupled with nitrate reductase Z in aerobic conditions [207,208]. Proteins encoded by the *ynfEFGH* operon mentioned in Chapter 2 are also part of the DMSO reductase family of molybdoenzymes.

Other Mocos are the sulfurated molybdenum cytosine dinucleotide cofactor, the di-oxo Moco, and base form of the cofactor which is referred to as Mo-MPT. The xanthine oxidase family utilises the sulfurated Moco and includes the xanthine dehydrogenase XdhABC, the aldehyde oxidoreductase PaoABC, and the xanthine dehydrogenase homologue XdhD [207]. The sulfite oxidase family uses the di-oxo Moco and only contains the MsrP protein, described in chapter 2 [207,208].

This work aims to further characterise the BisC enzyme in *Klebsiella pneumoniae* L4-FAA5, about which little is known but is hypothesised to be involved in TMAO metabolism, and show that there is another TMAO metabolism pathway that must be considered when exploring the influence of microbes on TMAO metabolism in the human gut.

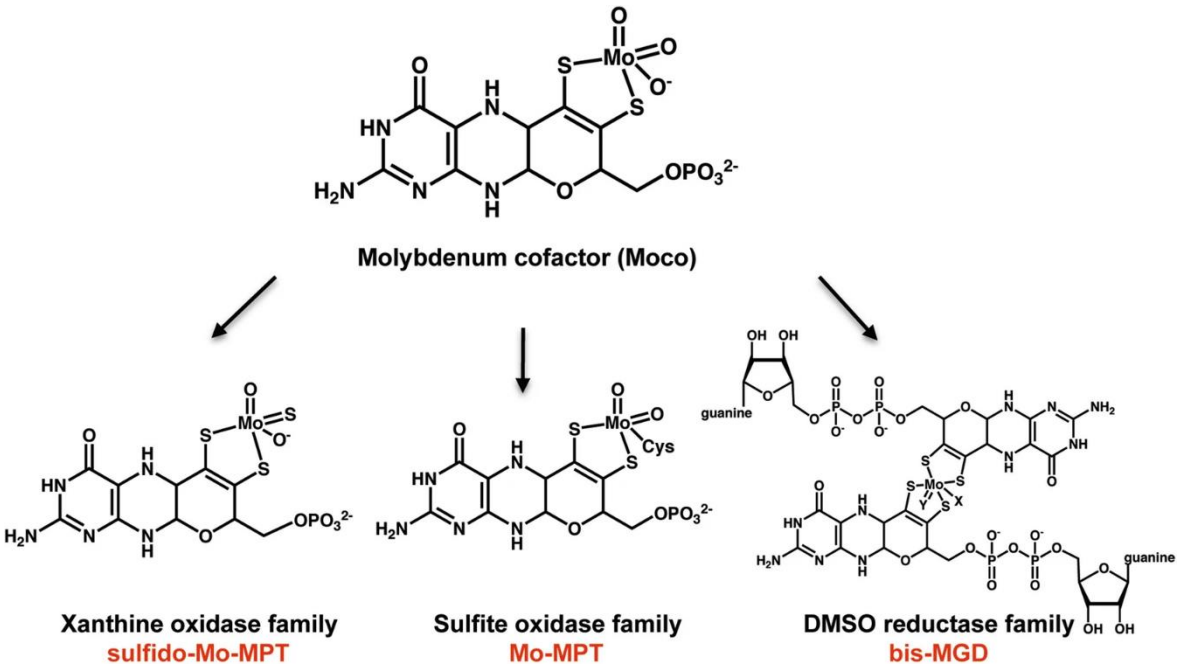


Figure 4.1. Structures of the molybdenum cofactors found in *E. coli*.

As a part of the DMSO reductase family BisC carries the bis-MGD cofactor. Figure taken from Iobbi-Nivol and Leimkühler, 2013 [207] and used under license granted by Elsevier.

4.2 Methods

4.2.1 Bioinformatic analysis of the BisC sequence and structure

Analysis of the BisC sequence is described in Chapter 2. Structural predictions for and comparisons of BisC encoded within L4-FAA5 were carried out using the I-TASSER server [209]. Further structural predictions of BisC were carried out using AlphaFold v2.3.0, hosted within Google Colab [210]. Presence of the bis-MGD cofactor was predicted using the online tool AlphaFill v2.1.0 [211]. Visualisation of structures was carried out using ChimeraX v1.7.1 [212], with the Matchmaker function being used for structural comparison [213]. The structure with the PDB ID 1DMR was used for structural comparison.

4.2.2 pBisK creation and transformation

Primers were created manually with the use of SnapGene Viewer (www.snapgene.com) and the NEBuilder online tool (www.nebuilder.neb.com). All primers were ordered from Macrogen (South Korea) and their details can be found in Table 4.1. *bisC* from *K. pneumoniae* L4-FAA5 was cloned from genomic DNA and inserted into a pET28b+ vector. A 6x-His tag was added to the C-terminus of BisC. Assembly of plasmid fragments was performed using NEBuilder HiFi DNA Assembly (New England Biolabs). Q5 polymerase master mix (New England Biolabs) was used for cloning PCRs.

pBisK was transformed into *E. coli* DH5α for storage. Competent DH5α cells were prepared by suspending cells in chilled TSS buffer before storage at -80 °C. TSS buffer was created by combining 25 mL LB medium (VWR, UK), 2.5 g PEG 8k (Signma-Aldrich), 1.23 mL DMSO (Fischer Chemical), and 0.75 mL 1 M $MgCl_2$. Transformation of pBisK involved addition of 1 µl of the Gibson assembly mix to thawed competent cells and incubating on ice for 60 minutes. After this, cells were heat shocked at 42 °C for 45 seconds before being placed back on ice. 1 mL of super optimal broth with catabolite repression (SOC) outgrowth medium (New England Biolabs) was added to the cells, which were then incubated at 37 °C for 1 hour before plating onto nutrient agar containing 100 µg/ml kanamycin (Sigma-Aldrich). Plates were incubated overnight at 37 °C. Colonies were checked for the presence of pBisK using colony PCR using primers bisK_fwd and bisK_rvs (Table 4.1). PCR was carried out using Mango Mix (Meridian Bioscience, UK).

pBisK was extracted from the storage strain using a Monarch MiniPrep kit (New England Biolabs). The same transformation method was used to insert pBisK into *E. coli* BL21 (DE3), replacing the Gibson assembly mix with 1 µl of purified pBisK.

4.2.3 Induction of pBisK and protein purification

Autoinduction medium was created via 5 mL of a mixed sugar solution (2.5 % D-glucose, 25 % glycerol, 10 % lactose), 12.5 mL of a mixed phosphate solution (0.5 M ammonium sulfate, 1 M sodium phosphate, 1 M potassium phosphate), and 250 µL of 1 M magnesium sulfate to 250 mL of nutrient broth. Cells carrying pBisK were grown in nutrient broth with 100 µg/mL of kanamycin for ~4 hours. One mL of culture was then added to the autoinduction medium with 1 mL of trace metal mix A5 with Co (Sigma-Aldrich) and incubated for 24 hours at 30 °C. Cells were harvested via centrifugation and resuspended in 10 mL of LEW buffer (50 mM NaH_2PO_4, 300 Mm NaCl, pH 8.0). One µL of DNase, lysozyme, and protease inhibitor cocktail each were added, alongside magnesium chloride to 1 mM final concentration. The cells were then sonicated on ice for 3 minutes, five times. This lysate was then purified as per the instructions for the Protino Ni-TED column. An SDS-PAGE gel (12 % Invitrogen) was ran to check for the presence and purity of protein samples before clean samples were concentrated. SDS sample buffer was from ThermoFisher Scientific, and the protein ladder used was PageRuler Plus Prestained Protein Ladder (ThermoFisher Scientific). The concentration of samples was checked using absorbance at 280 nm on a Nanodrop machine before being aliquoted and stored at -80 °C.

Table 4.1. Primers used in the creation of pBisK

Primer	**Use**	**Sequence 5’ – 3’**
bisK_fwd	Cloning *bisC* from L4-FAA5	ggctttgttagtgatggtgatggtgatgtccgcctgcgttggccggcggatc
bisK_rvs	Cloning *bisC* from L4-FAA5	atataccatgttgccaacctcatctgcaac
bkbone_fwd	Creating the pET28 backbone for pBisK	tgcagatgaggttggcaaCATGGTATATCTCCTTCTTAAAGTTAAAC
bkbone_rvs	Creating the pET28 backbone for pBisK	catcaccatcaccatcactaacaaagcccgaaaggaagctgag

4.2.4 Benzyl viologen assay

The benzyl viologen assay was carried out in an anaerobic environment created by cabinet (Don Whitley A35 anaerobic workstation; 10 % CO_2 10 % H_2 80 % N_2, BOC). All equipment and reagents, excluding the sodium dithionite (Sigma-Aldrich) which was made fresh, were kept in anaerobic conditions for a minimum of 12 hours prior to the start of the assay. The phosphate buffer was degassed using nitrogen. A reaction mix was prepared in a 100 mM 6.5 pH phosphate buffer, containing benzyl viologen (Sigma-Aldrich) to a concentration of 872 µM, sodium dithionite to a concentration of 1308 µM, and BisC to a concentration of 9 nM. An aliquot (295 µL) of this reaction mix was added to the wells of a 96-well plate. Five µL of TMAO stock of varying concentration was then added to the wells simultaneously and the plate was then placed inside a Cerillo Stratus plate reader overnight. Colour change of the reaction was measured at 600 nm. Final concentrations of TMAO in µM were: 11.9, 19.0, 47.5, 95.0, 950.0, 1900.0, and 2800.0. Results of the assay (not presented in this thesis) were analysed using the R package renz v0.1.1 [214].

4.3 Results

4.3.1 Functional and structural predictions of *Klebsiella pneumoniae* L4-FAA5 BisC

The online tool I-TASSER was used to predict the structure and function of BisC (Figure 4.2). I-TASSER found that BisC is structurally similar to other molybdoenzymes, despite sharing low sequence identity (Table 4.2) [209]. Amino acids critical to the carriage of the MGD ligand are present.

I-TASSER also predicted the presence of an MGD ligand with a 0.74 confidence score, and TMAO reductase activity with a 0.599 confidence score and 0.954 TM-score. The TMAO reductase activity prediction also predicted the presence of the bis-MGD co-factor.

AlphaFold [210,211] was also used to predict the structure of BisC, with AlphaFill [211] being used to predict the position of the bis-MGD cofactor (Figure 4.3). The Matchmaker tool in ChimeraX [212,213] was used to compare the structures of BisC and the *R. sphaeroides* DMSO reductase identified as being similar to BisC by I-TASSER (Figure 4.4). This gave an RMSD value of 0.895 Å.

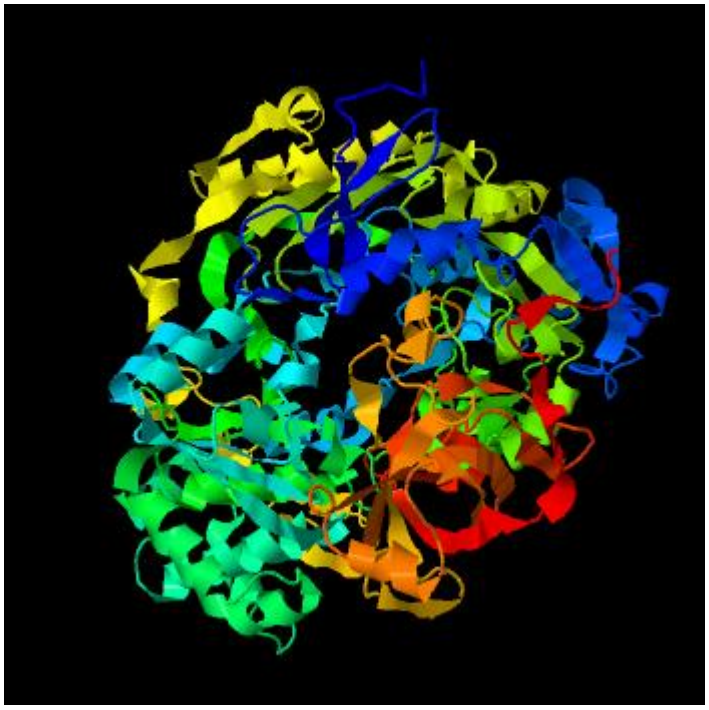


Figure 4.2. Structure predicted by I-TASSER based on the BisC amino acid sequence carried by *K. pneumoniae* L4-FAA5.

Table 4.2. The 10 most structurally similar proteins in PDB when compared to BisC from *Klebsiella pneumoniae* L4-FAA5

PDB ID	Enzyme	Organism	TM-Score (structural similarity)	% Identity	% Coverage
1dmr	DMSO reductase	*Rhodobacter capsulatus*	0.975	48.9	98.6
7l5i	MtsZ methionine sulfoxide reductase	*Haemophilus influenzae*	0.971	53.0	97.9
1eu1	DMSO reductase	*Rhodobacter sphaeroides*	0.969	48.0	97.8
1tmo	TMAO reductase	*Shewanella massilia*	0.954	39.8	98.1
1ti6	Pyrogallol-phloroglucinol transhydroxylase	*Pelobacter acidigallici*	0.912	24.7	95.6
3ir7	NarGHI mutant NarG-R94S	*Escherichia coli*	0.832	19.0	93.9
2ivf	Ethylbenzene dehydrogenase	*Aromatoleum aromaticum*	0.809	16.6	89.8
4ydd	PcrAB Perchlorate reductase	*Azospira suillum*	0.808	19.7	90.5
6cz7	Arsenate respiratory reductase	*Shewanella* sp. ANA-3	0.790	15.6	88.5
2vpw	Polysulfide reductase	*Thermus thermophilus*	0.771	16.6	87.1

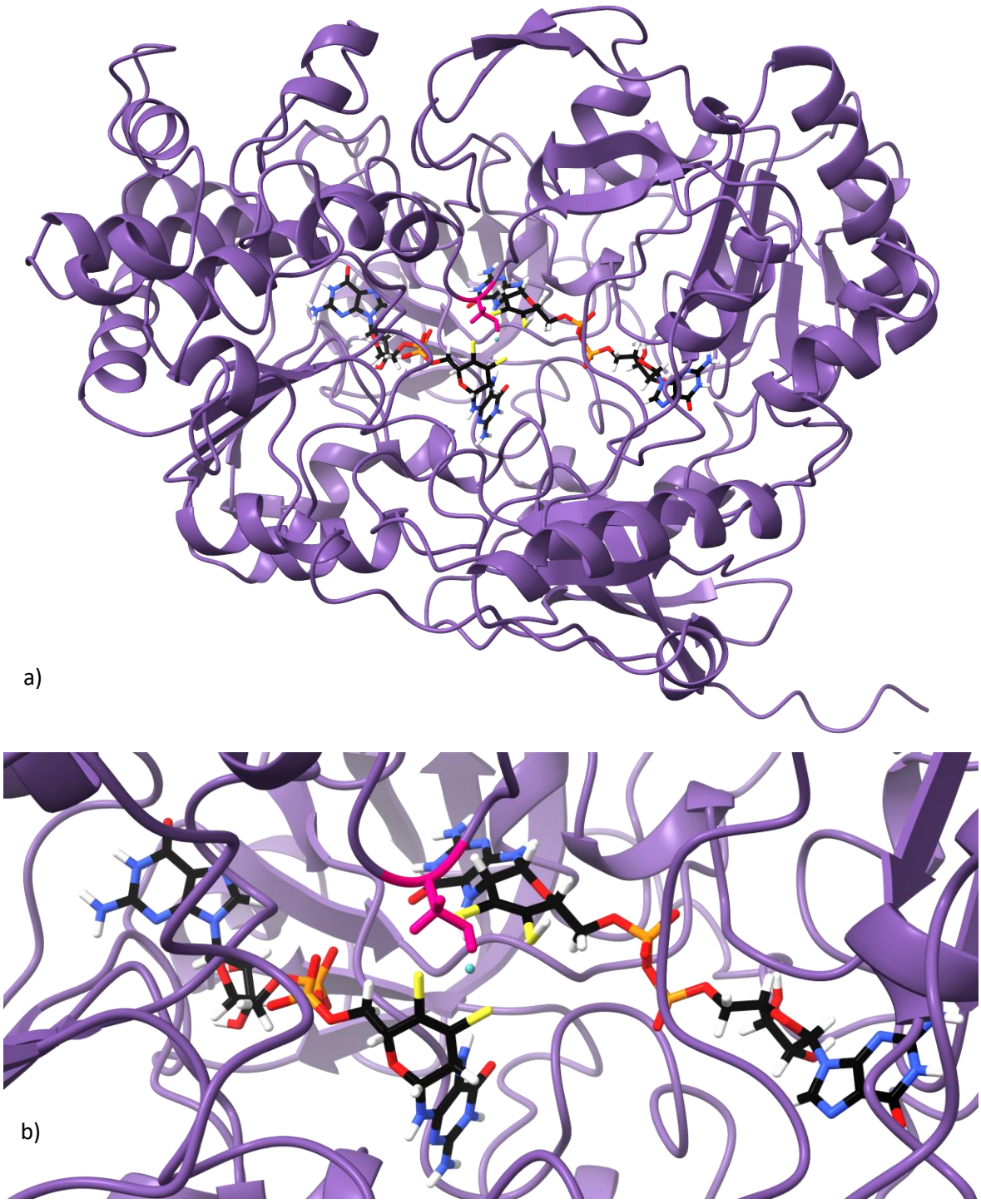


Figure 4.3. Structural predictions for BisC of *K. pneumoniae* L4-FAA5.

(a) Structure predicted by AlphafFold and AlphaFill based on the BisC amino acid sequence carried by *K. pneumoniae* L4-FAA5. (b) Close up of bis-MGD cofactor and enzyme active site. The residue highlighted in pink is the serine active site.

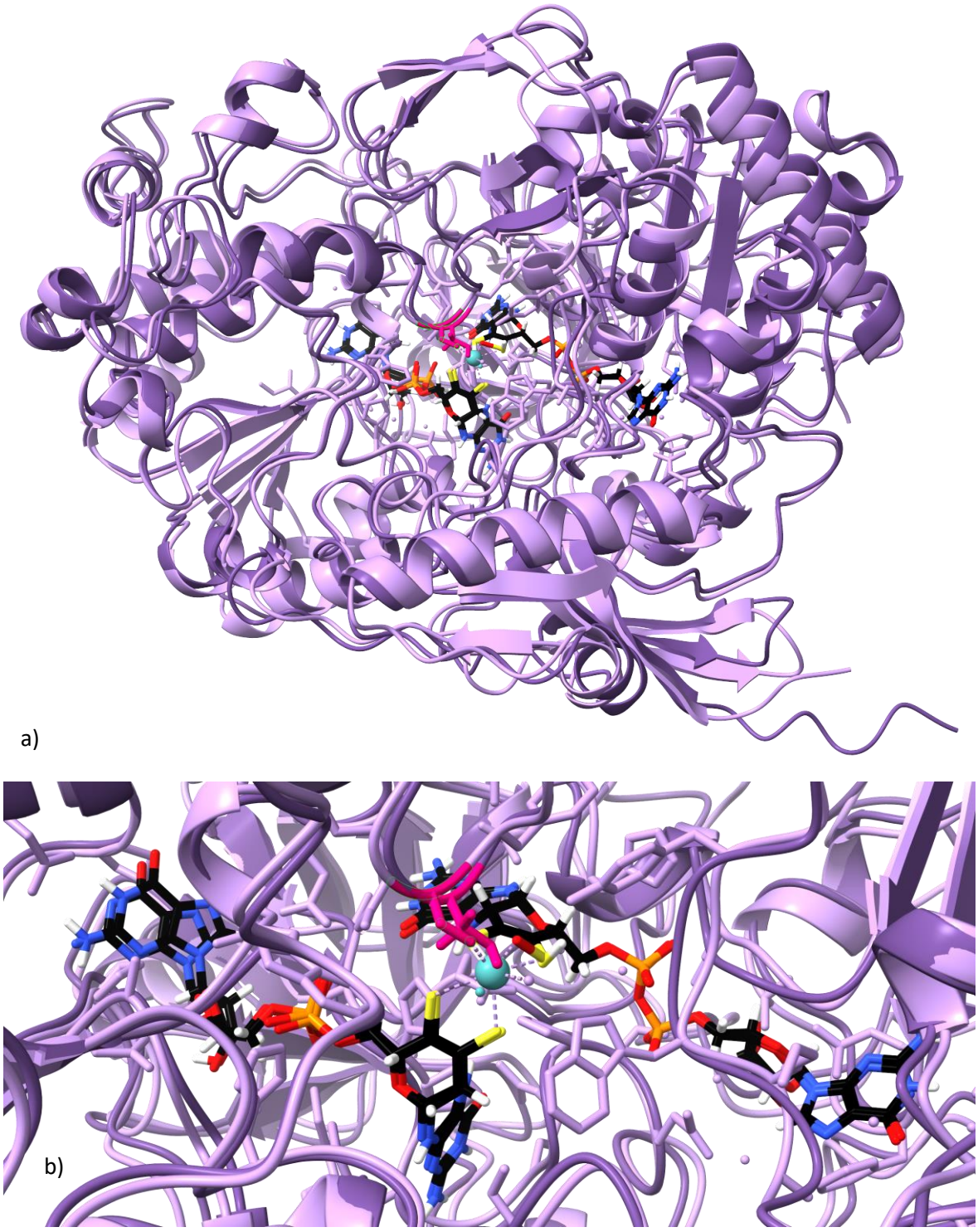


Figure 4.4. Comparison of BisC structural predictions.

(a) Overlay of the AlphaFold/AlphaFill predicted *K. pneumoniae* BisC structure (dark purple) and *R. sphaeroides* DMSO reductase (light purple). (b) Close up of bis-MGD cofactor and enzyme active site. The residue highlighted in pink is the serine active site.

4.3.2 The gene *bisC* was cloned from *Klebsiella pneumoniae* L4-FAA5, inserted into an expression vector, and transformed into storage and expression strains

To purify BisC for use in activity assays the gene *bisC* was cloned from genomic DNA extracted from *K. pneumoniae* L4-FAA5. A 6x-His tag was added to the cloned gene to be present at the C-terminus of the expressed protein. The *bisC* fragment was inserted into the expression vector pET28 b+. This plasmid was then named pBisK (Figure 4.5). pBisK was then transformed into *E. coli* DH5α for storage and *E. coli* BL21 DE3 for expression. Successful transformation was confirmed using colony PCR (Figure 4.6).

4.3.3 pBisK was induced and BisC was purified

E. coli BL21 (DE3) carrying pBisK were induced and cells were harvested. BisC was purified from harvested cells using a nickel column. Cell extracts and purified results were ran on an SDS-PAGE gel to check for expression of BisC and the purity of the sample (Figure 4.7). The band of the purified BisC was visible at around 85 kDa.

4.4 Discussion

The enzyme biotin sulfoxide reductase, or BisC, is a molybdoenzyme that has been shown to be reduce biotin sulfoxide and methionine sulfoxide in *E. coli* [148,205]. This functionality aids the cell in the scavenging of biotin and the assimilation of oxidised methionines [148,205]. As presented in Chapter 2 the amino acid sequence of BisC shares 66.3 % homology with the TMAO reductase TorZ, which is higher than the homology between TorA and TorZ. This, along with its prevalence in human gut-associated bacteria being higher than TorA/TR1, prompted a further investigation of BisC. This further investigation attempted to see if the BisC enzyme present in *K. pneumoniae* L4-FAA5 was able to reduce TMAO, which would show the presence of a novel pathway for TMAO reduction in the human gut and affect how research into this process should be approached in the future. The BisC encoded by *K. pneumoniae* L4-FAA5 was used in this work due to the usage of L4-FAA5 as a caecal TMAO metabolism model throughout this work. Also, due to the high prevalence of BisC in *K. pneumoniae* genomes, understanding the activity of its BisC enzyme is crucial in understanding the contribution of *K. pneumoniae* to TMAO metabolism in the human gut.

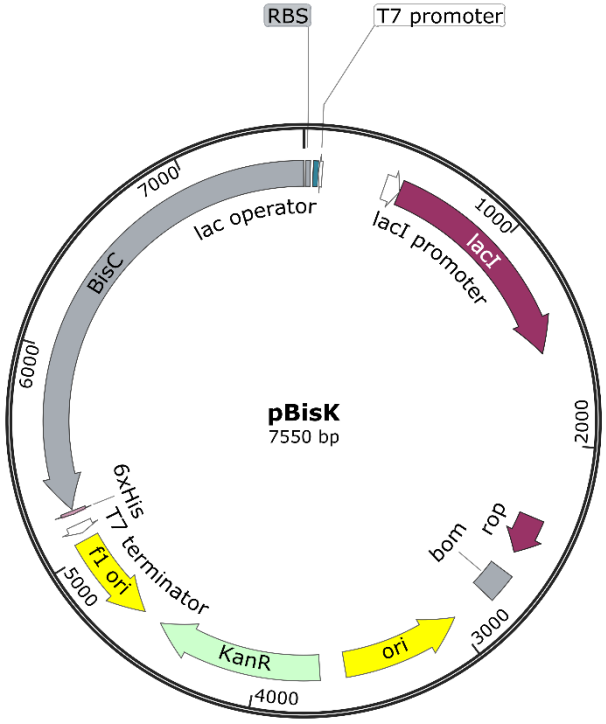


Figure 4.5. The pBisK plasmid used to express BisC.

Expression of BisC was under the control of a T7 promoter, allowing for expression using the *E. coli* strain BL21 (DE3). Kanamycin resistance is conferred by the expression of KanR which was used to select for colonies carry pBisK, as well as ensuring plasmid maintenance.

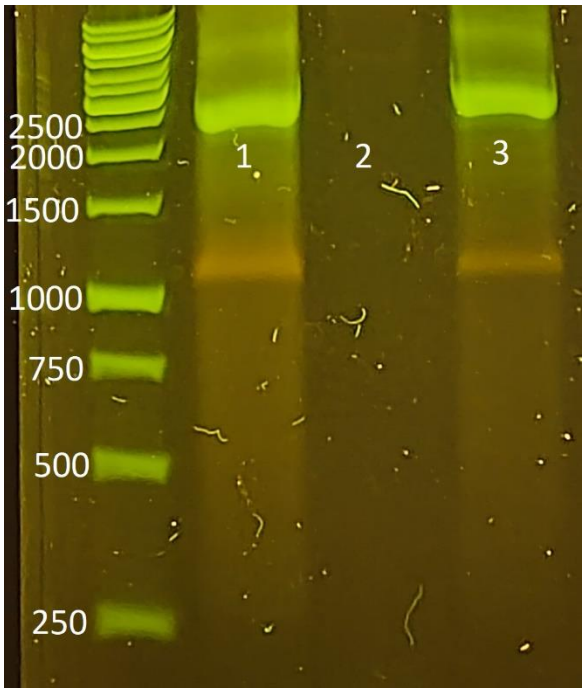


Figure 4.6. Gel electrophoresis of PCR used to check for the presence of pBisK in *E. coli* BL21 (DE3).

Well 1 contained products from colony PCR of BL21 (DE3) carrying pBisK. Well 2 contained products from colony PCR of BL21 (DE3) with no plasmids. Well 3 contained products from PCR using purified pBisK as a template.

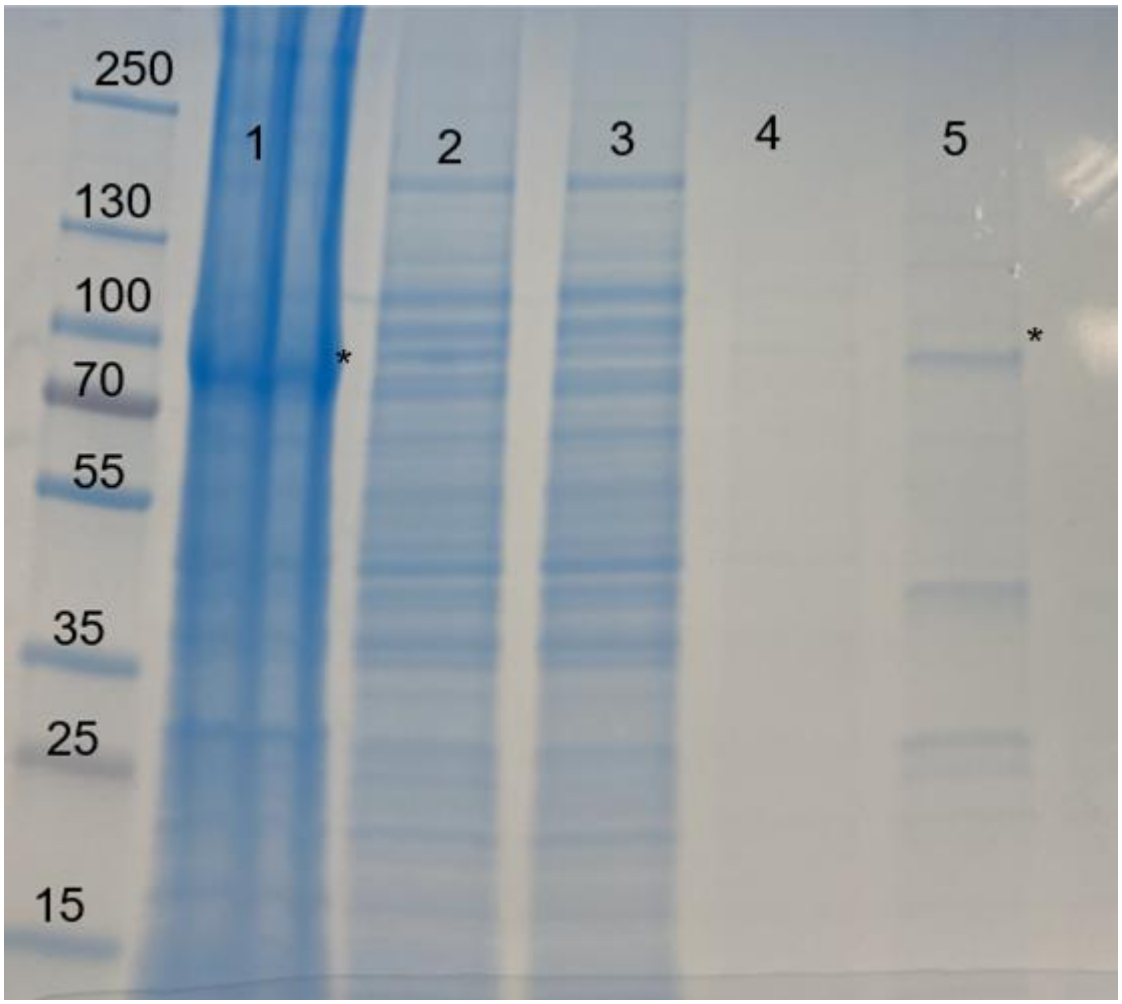


Figure 4.7. SDS-PAGE analysis of fractions taken during BisC purification.

Well 1 contained a whole cell lysate, well 2 contained clarified cell lysate, well 3 contained flowthrough from the binding step of purification, well 4 contained flowthroughs from the wash steps, and step 5 contained the eluted fractions. Protein was seen between the 100 kDa and 70 kDa bands on the ladder, approximately 85 kDa. BisC bands are marked with an asterisk (*).

After the amino acid sequence identity (%) between BisC and TorZ was seen to be high relative to other TMAO-related enzymes further bioinformatic analysis was carried out to see if these similarities could be found in other enzymes. The online tool I-TASSER [209] was used to predict the structure and function of BisC, as well as find other enzymes that were structurally similar to BisC. This analysis identified several TMAO reductases present in the protein database (PDB) with a high structural similarity to BisC. The top four most structurally similar enzymes (TM-Score >95 %) have all been shown to exhibit TMAO reductase activity [215–218], although sequence identity between these enzymes and BisC is relatively low with an average sequence identity of 47.43 %. I-TASSER also predicted TMAO reductase activity as the most likely function for BisC, although the confidence score for this was only 59.9 %. This high level of structural similarity was also seen when comparing an Alphafold generated model of L4-FAA5 BisC with the structure of DMSO reductase from *R. sphaeroides*, which has been shown to reduce TMAO. These results led to the progression of this work as it seemed possible that BisC may have TMAO reductase activity. The first stage of this was the creation of the *K. pneumoniae* L4-FAA5 BisC expression plasmid pBisK, along with its transformation into the expression strain *E. coli* BL21 (DE3).

Once pBisK had been inserted into *E. coli* BL21 (DE3) BisC could be expressed. This was first done as described in Methods, but at 37 °C instead of 30 °C with an overnight incubation as opposed to 24 hours, as well as an absence of the trace metal mix. As each enzyme produced required a molybdenum ion the trace metal mix aided BisC production by providing an excess of molybdate. Reducing the incubation temperature to 30 °C also appeared to increase BisC production, seen as a more intense band in the SDS-PAGE gel. Decreases in temperature aiding recombinant protein production has been mentioned in the literature, although there is not currently an understanding of why this aids protein production [219,220].

Currently there appear to be two smaller proteins present at ~25 and ~45 kDa in purified samples. The usage of a fast protein liquid chromatography (FPLC) system would have been used to try and separate these smaller proteins from BisC. This is because FPLC systems generate multiple smaller elution fractions, rather than the three elution fractions gained from a gravity column-based method. However, due to issues with the expression of BisC the usage of FPLC was not able to be fully tested. One of these issues was cells carrying

pBisK being unable to grow in liquid medium containing kanamycin, although this issue was resolved after finding that a change in growth medium from nutrient broth 3 to nutrient broth 1 was causing issues which may have been due to absence of meat extract in nutrient broth 1. From this point forward LB broth was used as it is a rich medium comparable to nutrient broth 3. Current issues are with the expression of BisC not occurring. Different growth temperatures (30 °C, 37 °C), adding extra molybdate (1 mM Na_2MoO_4), and using IPTG (1 mM) instead of lactose for induction are all method changes that have been trialled with no apparent effect on expression. The issue with BisC expression however may be due to the use of *E. coli* BL21 (DE3) as the expression strain. While *E. coli* BL21 (DE3) is able to produce molybdoenzymes it is unable to uptake molybdenum as it does not encode the ModABC molybdate transporter [221]. As each BisC protein requires a molybdate ion the expression strain being unable to acquire molybdate is an oversight that severely limits how much BisC can be produced. Current plans involve the creation of a new expression vector using the plasmid pISK-IBA3 to be transformed into the *E. coli* strain TOP10 which has been checked for the presence of the molybdate transporter and molybdoenzyme production using KEGG and ghostKOALA [184].

The benzyl viologen assay used to assess the activity of BisC requires optimisation. This assay utilises reduced benzyl viologen to act as an electron donor for BisC as it reduces TMAO. This means that the assay must be carried out in anaerobic conditions to avoid atmospheric or dissolved oxygen from oxidising the reduced benzyl viologen. Also, since the benzyl viologen must be reduced a reducing agent must be used, in this case sodium dithionite (also known as sodium hydrosulfite). The addition of a reducing agent can also aid the removal of any dissolved oxygen from the assay reagents, which is why sodium dithionite was added in excess. The removal of oxygen from the assay appears to have been the biggest issue when trying to obtain reproducible results in this work. During the optimisation process the most common issue was false positives caused by the oxidation of the benzyl viologen, independent of any BisC TMAO reductase activity. Another issue encountered was that molybdoenzymes are often unstable in storage, with stocks only being viable at -80 °C for around three weeks. This led to more BisC needing to be purified during the optimisation process, which stalled progress further when efforts to induce BisC production failed. At time of writing only a single benzyl viologen assay was completed

successfully. While this assay does appear to show that BisC has TMAO reductase ability, the data is unreliable and does not appear to fully follow the expected curve for a Michaelis-Menten plot as data points appear to cluster at lower TMAO concentration before *v* increases in a linear manner that does not follow the shape of the curve. This data is also the result of only one biological repeat (triplicate technical repeats), and it is also lacking the presence of the methionine sulfoxide positive control. The positive control was not being used during assay optimisation as false positives were the main issue being faced and the amount of methionine sulfoxide available was limited. Due to these reasons this data has not been presented in this thesis.

Overall, the aims of this work still have not been met fully. While the results of the bioinformatic analysis and the benzyl viologen assay seem to suggest that BisC has TMAO reductase ability more work needs to be done. First the current issues with the expression of BisC need to be resolved before work on assay optimisation can continue. This optimisation will focus on the false-positives, largely by altering the amount of sodium dithionite being added to ensure that dissolved oxygen has been removed while ensuring that the excess is not so great that is causes false-negative result.

Chapter 5 Characterisation of novel weberviruses and examination of their depolymerases

5.1 Introduction

Bacteriophages, also known as phages, are viruses that infect bacteria. These phages are thought to be the most abundant biological entities on Earth, with an estimated 10^{31} present, and with humans carrying an estimated 10^{13} viruses per individual [222]. In the human gut up to 10^{8} virus particles have been detected in healthy gut mucosa biopsies, with up to $4x10^{9}$ being detected in biopsies from patients with Crohn's disease, and another study reporting 10^{9} virus particles per gram of human faeces [222,223]. As phages can kill their hosts or persist within them, potentially altering their genotype and phenotype [224], knowing how phages interact with the gut microbiota is key to understanding the effects of the gut microbiota on human health. Examples of changes in the human gut virome are a decrease in viral diversity in type 1 diabetes patients, increased viral diversity in colorectal cancer patients, and increased abundance of *Caudoviricetes* phages in patients suffering from Crohn's disease and ulcerative colitis [222].

The first member of the genus *Webervirus* to be isolated was *Webervirus KP36*, which was isolated from treated sewage from a communal wastewater plant in Poland [225]. *Webervirus KP36* was found to be a lytic phage that infected ~12 % of clinical *K. pneumoniae* strains that it was tested against, as well as exhibiting depolymerase activity on its hosts [225,226]. Current virus taxonomy places the genus *Webervirus* within the family *Drexlerviridae,* class *Caudovirales*, phylum *Uroviricota*, kingdom *Heunggongvirae*, and realm *Duplodnaviria* [227]. Thirty-two species of *Webervirus* are currently listed within the International Committee on Taxonomy of Viruses database [227].

Depolymerases are proteins encoded by phages that are present within the tail fibres or tail spikes of the virus particle itself [228]. These depolymerases work to break down the bacterial capsule, also known was the K antigen, which allows the phage to infect the cell [228]. The phages *Webervirus KP36* and *Webervirus KLPN1* have been shown to exhibit depolymerase activity, with the depolymerase present in *Webervirus KP36* being isolated and shown to be active against the *K. pneumoniae* isolates with the K36 capsule type [175,226].

These depolymerases are notable as isolates that have been decapsulated by purified depolymerases have been found to be less resistant to the innate immune response in *Galleria mellonella*, more susceptible to phagocytosis by macrophages *in vitro*, and have less resistance to complement-mediated killing when exposed to human serum [226,228,229].

Webervirus KLPN1 had previously been isolated from the caecal effluent of a healthy human female on *K. pneumoniae* L4-FAA5 [175]. Work conducted by previous PhD students (Thomas Brook, Preetha Shibu; both University of Westminster) and masters students (Tobi Tijani, University of Westminster; Sara Garnett, Nottingham Trent University) isolated seven novel *Klebsiella* infecting weberviruses from sewage samples. These novel phages had also been sequenced and imaged (Figure 5.1), but no further work had been done on their characterisation. This work aims to characterise the genomes and depolymerases of these phages, as well as exploring the diversity of depolymerases encoded by weberviruses.

Note: this work was started during the COVID lockdown when I was unable to carry out laboratory work and had finished work described in Chapter 2, and continued while unable to do laboratory work for 6 months during the third year of my PhD after an accident.

5.2 Methods

5.2.1 Assembly and annotation of novel bacteriophage genomes

Phage sequence data was assembled using SPAdes (v3.13.0) [230] and visualised using Genovi (v0.4.3) [179]. CheckV (v0.8.1) [231] was used to determine quality and completeness of assembled genomes. Phages genomes were annotated using Prokka (v1.14.6) [232] with the PHROGs (v3) [233] phage-specific hidden Markov model.

5.2.2 Comparison of webervirus genomes

ViPTree (v1.9.1) [234] was used to determine the relationship between the seven novel phage genomes and other previously described dsDNA viruses. Publicly available *Webervirus* genomes were downloaded from NCBI GenBank [158] (Table 5.1) for further genome comparison. All webervirus genomes were uploaded to VIRIDIC web (http://rhea.icbm.uni-oldenburg.de/VIRIDIC/) [235] for further comparison. The output of this analysis was visualised using heatmap.2 in the R package ggplot (v3.1.1). vConTACT (v2.0) [236] was used

to create a gene-sharing network between all webervirus genomes, with the output being visualised using Gephi v0.9.20 (https://gephi.org/).

Large-subunit terminase genes were extracted from genomes using Biostrings (v2.62.0) [237]. A multiple-sequence alignment was created for the genes (Clustal Omega 1.2.3 implemented in Geneious Prime v2020.0.5; options – group sequences by similarity, 5 representative iterations). These alignments were used to create bootstrapped maximum-likelihood trees, to confirm relationships of phages seen in the VIRIDIC- and vConTACT-based analyses. Phangorn (v2.7.1) and Ape (v5.5) [238,239] were used to generate the trees using the LG model [240].

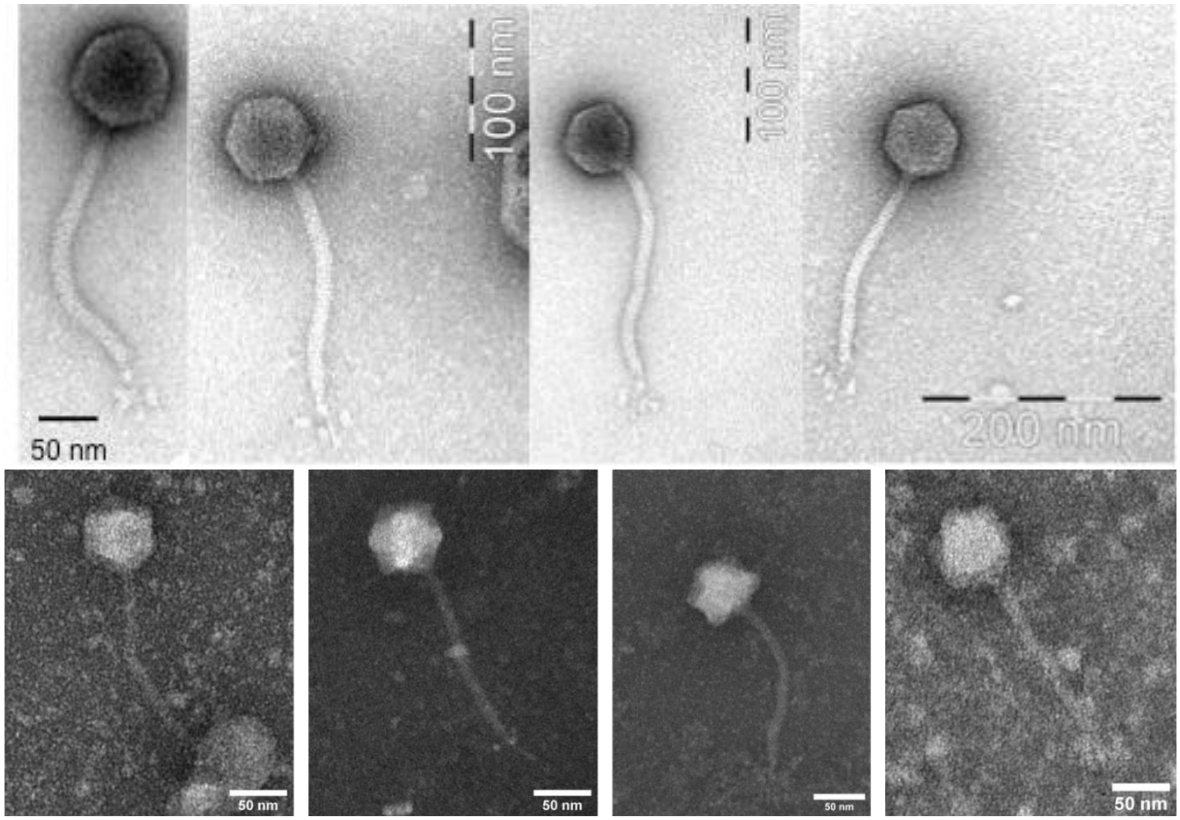


Figure 5.1. Transmission electron microscopy images of previously isolated weberviruses. Top row from left to right: Webervirus KLPN1, Webervirus KLPN2, Webervirus KLPN3, Webervirus KLPN4. The images were generated by Dr Horst Neve, Max Rubner-Institute, Kiel, Germany. Bottom row for left to right: Webervirus KLPN5, Webervirus KLPN6, Webervirus KLPN7, Webervirus KLPN8. The images were generated by Dr David Negus, Nottingham Trent University.

Table 5.1. Webervirus genomes included in this work

Phage/species	Accession	Isolated from	Location	Reference
Webervirus KLPN2	OM065837	Sewage	UK	This study
Webervirus KLPN3	OM065838	Sewage	UK	This study
Webervirus KLPN4	OM065839	Sewage	UK	This study
Webervirus KLPN5	OM065840	Sewage	UK	This study
Webervirus KLPN6	OM065841	Sewage	UK	This study
Webervirus KLPN7	OM065842	Sewage	UK	This study
Webervirus KLPN8	OM065843	Sewage	UK	This study
Klebsiella phage 066039 (P39)	MW042802.1	Sewage	China	(Fang *et al*., 2022)
Klebsiella phage ABTNL-2	MZ221764.1	Sewage	China	–
Klebsiella phage B1	MW672037.1	Sewage	Hungary	(Pertics *et al*., 2021)
Klebsiella phage IME268	MZ398242.1	Unknown	China	(Nazir *et al*., 2022)
Klebsiella phage LF20	MW417503.1	Hospital wastewater	China	–
Klebsiella phage MMBB	MT894005.1	Merri Creek watercourse	Australia	(Thung *et al*., 2021)
Klebsiella phage NJR15	NC_048044.1	Sewage	China	–
Klebsiella phage NJS2	NC_048043.1	Sewage	China	–
Klebsiella phage NJS3	MH633486.1	Hospital sewage	China	–
Klebsiella phage P1	MZ598515.1	Water	China	–
Klebsiella phage P528	MW021764.1	Sewage	USA	(Cranston *et al*., 2022)
Klebsiella phage PhiKpNIH-10	MN395285.1	Wastewater treatment plant	USA	(Hesse *et al*., 2020)
Klebsiella phage RAD2	NC_055956.1	Hospital wastewater	UK	(Dunstan *et al*., 2021)
Klebsiella phage Sanco	MK618657.1	Wastewater treatment plant	USA	(Richardson *et al*., 2019)
Klebsiella phage SH-Kp 160016	KY575286.1	Sewage	China	–
Klebsiella phage Solomon	MT701592.1	Wastewater treatment plant	USA	(Hudson *et al*., 2021)
Klebsiella phage vB_KpnS_2811	LR757892.1	Unknown	UK	–
Klebsiella phage vB_KpnS_KingDDD	MN013078.1	Sewage	USA	(Thurgood *et al*., 2020)
Klebsiella phage vB_KpnS_Penguinator	MN013087.1	Sewage	USA	(Thurgood *et al*., 2020)
Klebsiella phage vB_KpnS_ZX2	MW722081.1	Faeces	China	–
Klebsiella phage vB_kpnS-VAC10	MZ428227.1	Sewage	Spain	(Bleriot *et al*., 2021)
Klebsiella phage vB_KpnS-VAC11	MZ428228.1	Sewage	Spain	(Bleriot *et al*., 2021)
Klebsiella phage vB_kpnS-VAC2	MZ428221.1	Sewage	Spain	(Bleriot *et al*., 2021)
Klebsiella phage vB_KpnS-VAC4	MZ428222.1	Sewage	Spain	(Bleriot *et al*., 2021)
Klebsiella phage vB_KpnS-VAC5	MZ428223.1	Sewage	Spain	(Bleriot *et al*., 2021)
Klebsiella phage vB_KpnS-VAC6	MZ428224.1	Sewage	Spain	(Bleriot *et al*., 2021)
Klebsiella phage vB_KpnS-VAC7	MZ428225.1	Sewage	Spain	(Bleriot *et al*., 2021)
Klebsiella phage vB_KpnS-VAC8	MZ428226.1	Sewage	Spain	(Bleriot *et al*., 2021)
Webervirus alina	NC_049837.1	Sewage	USA	(Thurgood *et al*., 2020)
Webervirus call	NC_049836.1	Sewage	USA	(Thurgood *et al*., 2020)

Phage/species	Accession	Isolated from	Location	Reference
Webervirus domnhall	NC_049835.1	Sewage	USA	(Thurgood *et al*., 2020)
Webervirus F20	NC_043469.1	Lake water	South Korea	(Mishra *et al*., 2012)
Webervirus FZ10	NC_049840.1	Sewage	Russia	(Zurabov and Zhilenkov, 2019)
Webervirus GHK3	NC_048162.1	Sewage	China	(Cai *et al*., 2019)
Webervirus IMGroot	NC_049834.1	Sewage	USA	(Thurgood *et al*., 2020)
Webervirus JY917	NC_049843.1	Unknown	China	–
Webervirus KL	NC_049838.1	Unknown	Unknown	–
Webervirus KLPN1	NC_028760.1	Human caecal effluent	UK	(Hoyles *et al*., 2015)
Webervirus KLPPOU149	NC_049842.1	Unknown	Cote d'Ivoire	–
Webervirus KOX1	NC_047825.1	Wastewater	Australia	(Brown *et al*., 2017)
Webervirus KP1801	NC_049848.1	Hospital wastewater	Thailand	(Wintachai *et al*., 2020)
Webervirus KP36	NC_029099.1	Wastewater	Poland	(Majkowska-Skrobek *et al*., 2016)
Webervirus KpCol1	NC_047907.1	Wastewater	Turkey	–
Webervirus KpKT21phi1	NC_048143.1	Unknown	USA	(Nir-Paz *et al*., 2019)
Webervirus KpNIH2	NC_049845.1	Wastewater	USA	(Hesse *et al*., 2020)
Webervirus KpV522	NC_047784.1	Sewage	Russia	–
Webervirus mezzogao	NC_047850.1	Wastewater	USA	(Gao *et al*., 2017)
Webervirus N141	NC_047841.1	Unknown	South Korea	–
Webervirus NJS1	NC_048024.1	Sewage	China	–
Webervirus PKP126	NC_031053.1	Sewage	South Korea	(Park *et al*., 2017)
Webervirus segescirculi	NC_049833.1	Sewage	USA	(Thurgood *et al*., 2020)
Webervirus shelby	NC_049846.1	Pond water	USA	(Saldana *et al*., 2019)
Webervirus sin4	NC_049847.1	Wastewater	USA	(Castillo *et al*., 2019)
Webervirus skenny	NC_049841.1	Activated sludge	USA	(Gramer *et al*., 2019)
Webervirus sushi	NC_028774.1	Sewage	USA	(Nguyen *et al*., 2015)
Webervirus sweeny	NC_049839.1	Wastewater	USA	(Martinez *et al*., 2019)
Webervirus TAH8	NC_048042.1	Hospital sewage	China	–
Webervirus TSK1	NC_048126.1	Sewage	Pakistan	(Tabassum *et al*., 2018)
Webervirus wv13	NC_049844.1	Wastewater	Hungary	(Horváth *et al*., 2020)
Webervirus wv1513	NC_028786.1	Sewage	China	(Cao *et al*., 2015)

5.2.3 Identification and analysis of weberviruses in metagenomic datasets

PhageClouds [241] was used to identify phages related to known weberviruses in MAG datasets. The *Webervirus KLPN1* genome was searched against the PhageClouds database with a 0.15 threshold. Nucleotide sequences from MAGs identified by PhageClouds were retrieved from their associated databases [242–244]. CheckV (v0.8.1) [231] was used to check the contamination and completeness of the MAGs. ViPTree (Online v2.0) [234] was used to check their relationship to other weberviruses and they were also included in a vConTACT analysis (v2.0) [236] with the other previously identified webervirus genomes. Attempts were made to predict the MAG hosts using the CRISPR Spacer Database and Exploration Tool [245] and HostPhinder (v1.1) [246].

5.2.4 Global distribution of weberviruses

Distribution of weberviruses was determined by identifying the source and location information for the 67 GenBank genomes and 60 MAGs. Data was aggregated based on isolation source or geographical location. The geographical data was visualised using the R package rworldmap (v1.3.6) [247].

5.2.5 Analysis of depolymerases in weberviruses

A database of depolymerase protein sequences was created using known depolymerase sequences from *Webervirus KP36*. ORF34 from *Webervirus KP36* was used for a BLASTp search in UniProt (release 2021_03) [150], with those sequences labelled as being from *Klebsiella* phages and above 40 % identity added to the database. Other sequences were taken from publicly available genomes and identified using literature reporting the depolymerase sequences and their functionality. A full list of sequence sources can be found in Table 5.2. A BLAST (v2.12.0) [155] database was constructed from these sequences and a BLASTp search was ran versus all webervirus genomes. The results from this were then filtered in R using Tidyverse (v1.3.1) at 90 % coverage and 70 % identity cut-off values.

Biostrings (v2.62.0) [237] was used to extract the depolymerase sequences from the webervirus genomes. MUSCLE alignments were created in R using msa v1.26.0 [248] and a maximum-likelihood bootstrapped tree was created using Ape v5.6.2 [239] and Phangorn (v2.8.1) [238] and the WAG model [249].

Table 5.2. UniProt sequences used in the curated depolymerase database

Phage/species	Sequence ID	Cluster	Publication	Experimentally confirmed	Repository
Webervirus KLPN1	AKS10673.1		Hoyles *et al.*, 2015 [175]	No	NCBI GenBank
Webervirus KLPN1	AKS10674.1		Hoyles *et al.*, 2015 [175]	No	NCBI GenBank
Klebsiella phage B1	QTP95996.1	1	Pertics *et al.*, 2021 [250]	Yes	NCBI GenBank
Klebsiella phage RAD2	QUU29414.1	1	Dunstan *et al.*, 2021 [251]	Yes	NCBI GenBank
Klebsiella phage vB_kpnS_VAC2	QZE50483.1		Bleriot *et al.*, 2021 [251]	No	NCBI GenBank
Klebsiella phage vB_kpnS_VAC2	QZE50484.1		Bleriot *et al.*, 2021 [251]	No	NCBI GenBank
Webervirus KP36	YP_009226011.1	0	Majkowska-Skrobek *et al.*, 2016 [226]	Yes	NCBI GenBank
Webervirus GHK3	YP_009820105.1	1	Cai *et al.*, 2019 [252]	No	NCBI GenBank
Webervirus KL	A0A5P8FRW6		N/A	No	UniProt
Klebsiella phage vB_KpnS_Call	A0A5B9NNA5		N/A	No	UniProt
Webervirus KOX1	A0A1W6JSV6	3	N/A	No	UniProt
Webervirus KL	A0A5P8FS53		N/A	No	UniProt
Klebsiella phage vB_KpnS_Call	A0A5B9NE15		N/A	No	UniProt
Klebsiella phage vB_KpnS_FZ10	A0A4D6T4W2	2	N/A	No	UniProt
Webervirus wv1513	A0A0C5AJW8	3	N/A	No	UniProt
Klebsiella phage KMI8	A0A5B9NHK6		N/A	No	UniProt

5.2.6 Structural predictions of depolymerases

Structural predictions of six depolymerases were carried out using AlphaFold v2.3.0, hosted within Google Colab [210]. Visualisation of structures was carried out using ChimeraX v1.7.1 [212]. Depolymerases were chosen for this analysis based on if they had been experimentally confirmed and their relevance to the newly characterised phages. The online tool Foldseek v8-ef4e960 was also used to find other predicted depolymerase structures [253].

5.3 Results

5.3.1 Novel phages were successfully assembled and annotated

The genomes of the seven siphovirus-like (Figure 5.1) phages were assembled. Details of the chromosome sizes, as well as the predicted number of CDSs encoded by each phage genome can be found in Table 5.3. Each genome was found to be circular via analysis with Bandage (not shown) and visualised using Genovi [179] (Figure 5.2). CheckV [231] found no evidence of contamination in any of the genomes.

5.3.2 Phages were shown to be part of the genus *Webervirus*

ViPTree online showed that the novel genomes were from the genus *Webervirus* (not shown). An additional 59 webervirus genomes were then downloaded from NCBI GenBank to create a curated database of 67 genomes (Table 5.1). These genomes were included in a ViPTree analysis (Figure 5.3), and also used to create a gene-sharing network using vConTACT (v2.0) [236]. This network had 2,619 nodes and 51,589 edges, and was filtered based on the giant component topology which removed all nodes not connected to the main cluster (Figure 5.4). After filtering there were 1,523 nodes and 33,573 edges. This analysis confirmed that the novel phages are related to the family *Drexlerviridae* (Figure 5.4). The monophyletic nature of the genus *Webervirus,* including the seven novel genomes, was shown by analysis of their large-subunit terminase genes (Figure 5.5).

5.3.3 Metagenome-assembled phage genomes found to belong to the genus *Webervirus*

The online tool PhageClouds was used to identify potential members of the genus *Webervirus* in metagenomic datasets, with a total number of genomes around ~640,000 available to be searched. Forty-seven of the 67 publicly available webervirus genomes used above were found in the PhageClouds output. Five hits from PhageClouds from the PIGEON

dataset [254] represented phages already present in the curated webervirus database and so were excluded from further analyses. Fifty-four PhageClouds hits were MAGs from the Gut Phage Database [244], six from the Cenote Human Virome Database [242], and two from the Gut Virome Database [243]. Duplicates from these databases were removed leaving 60 MAGs to be used for further analysis (Table 5.4).

Table 5.3. Genome information for the seven novel weberviruses

Phage	Genome size (bp)	Number of CDSs	Isolated on
Webervirus KLPN2	51,264	80	*K. pneumoniae* L4-FAA5
Webervirus KLPN3	47,277	70	*K. pneumoniae* L4-FAA5
Webervirus KLPN4	50,522	78	*K. pneumoniae* L4-FAA5
Webervirus KLPN5	49,851	76	*K. pneumoniae* PS_misc6
Webervirus KLPN6	51,570	82	*K. variicola* PS_misc5
Webervirus KLPN7	49,712	76	*K. variicola* PS_misc5
Webervirus KLPN8	48,785	77	*K. variicola* PS_misc5

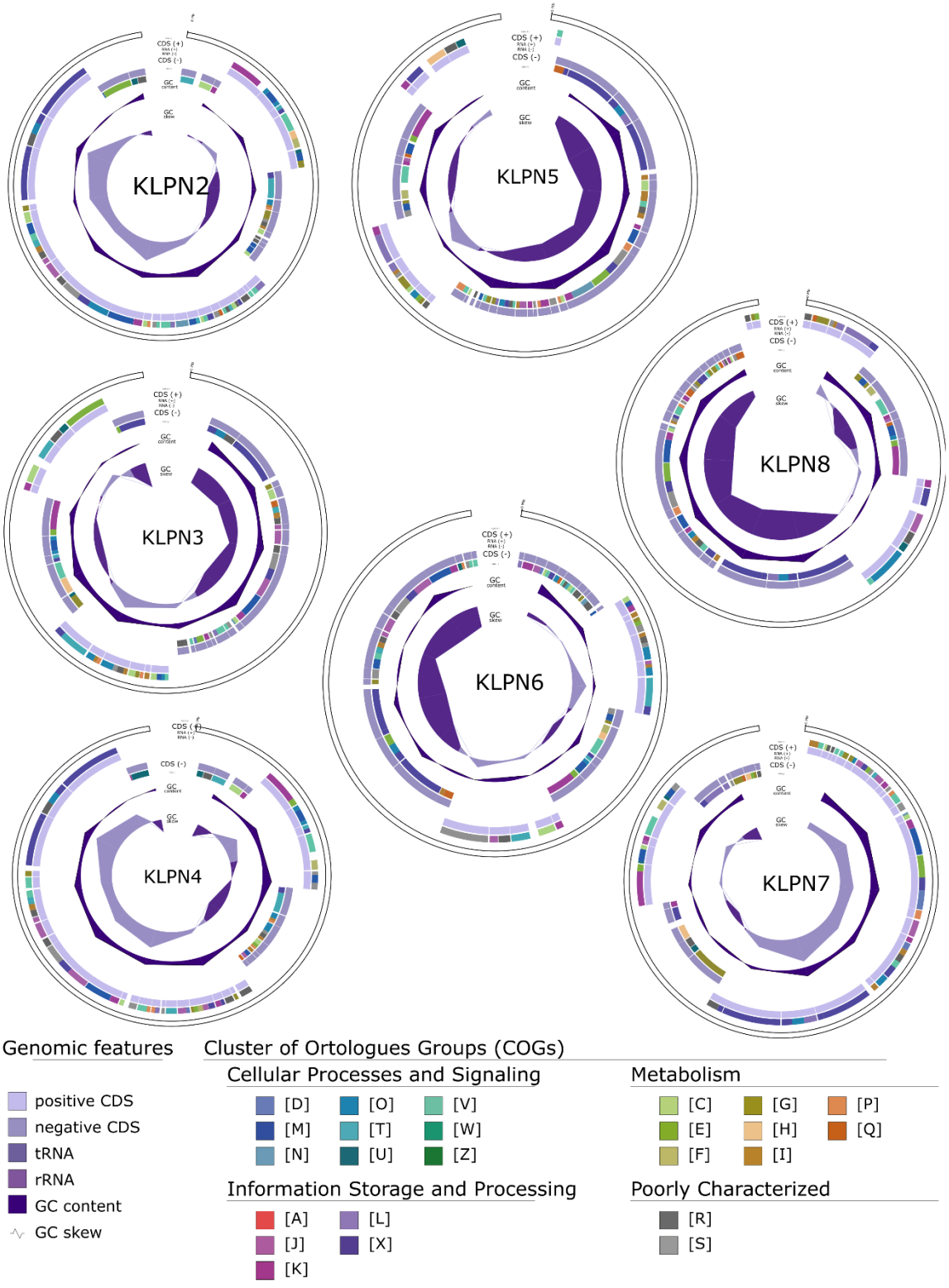


Figure 5.2. Genovi plots for each of the novel webervirus genomes.

Sizes not to scale. COG definitions can be found in legend to Figure 3.2.

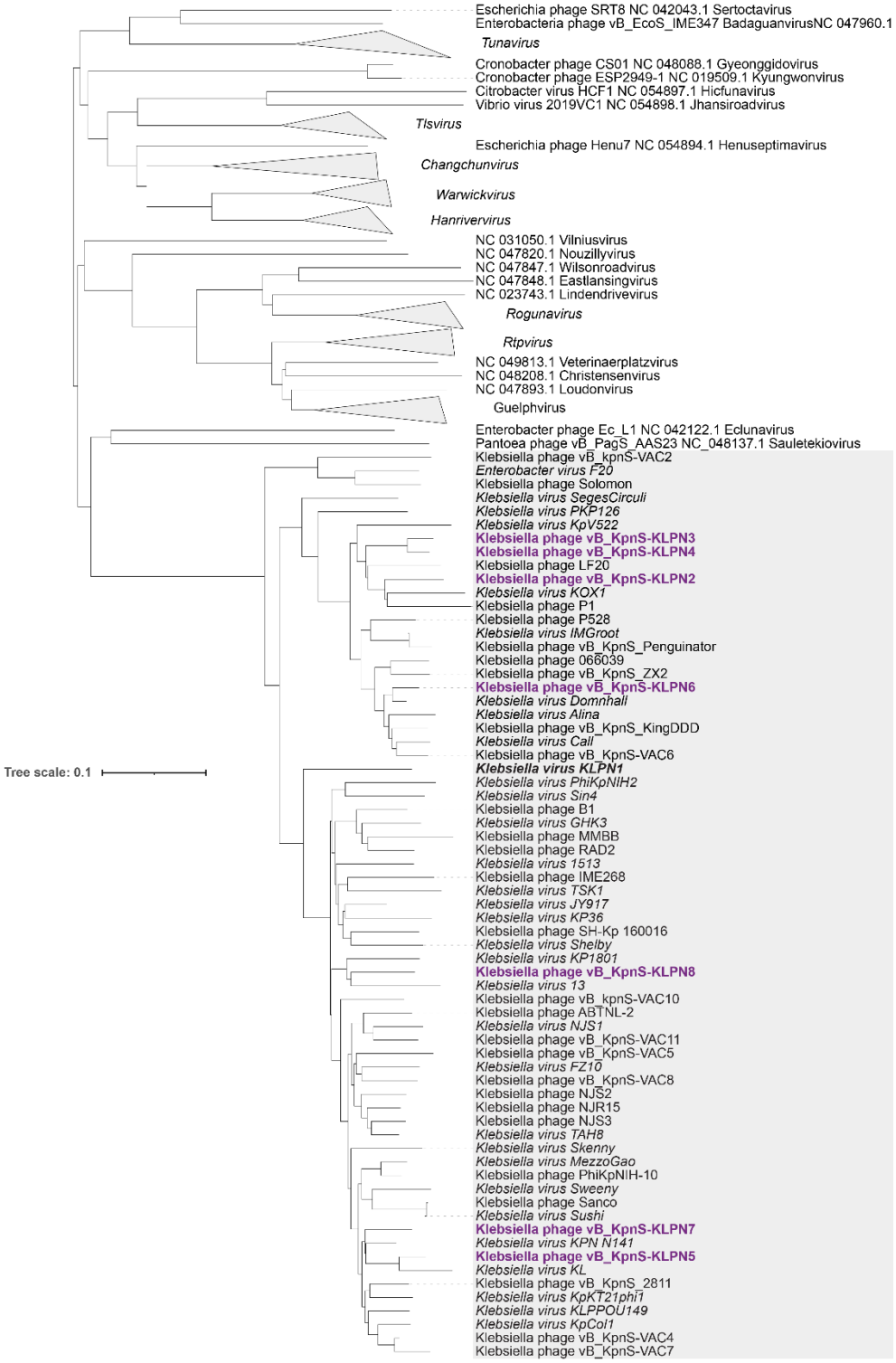


Figure 5.3. ViPTree analysis of publicly available *Drexlerviridae* genomes and genome data for the seven novel phages.

Webervirus genomes are highlighted grey. The seven novel genomes are shown in purple. The analysis is based on proteome data encoded within each of the phage genomes.

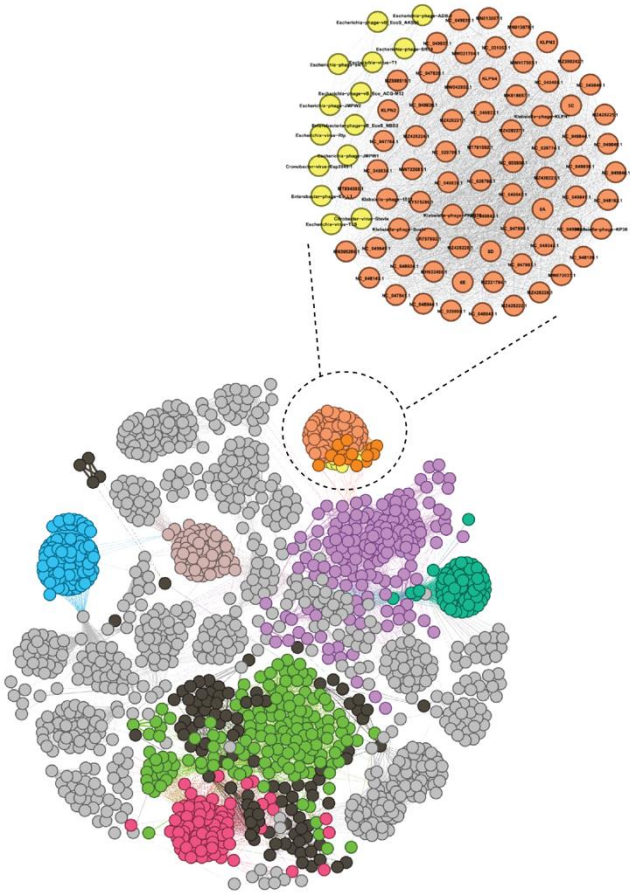


Figure 5.4. vConTACT filtered gene-sharing network in which only nodes connected to the main cluster are shown.

Filtered based on giant component; Fruchterman Reingold layout. The modularity of the network was determined (Gephi options selected: randomness, use weights, resolution 1.0). The network comprised 33 modules. Only modules representing >4 % of all nodes in the network are coloured. The circled module contained all 67 webervirus genomes (orange nodes) plus some other members of the family Drexlerviridae (yellow nodes). All the Klebsiella phage genomes (i.e. weberviruses) clustered together.

Figure 5.5. Phylogenetic tree (maximum likelihood) showing the relationship between members of the family Drexlerviridae based on the large-subunit terminase amino acid sequences encoded in genomes.

Bootstrap values are a percentage of 100 replicates. The tree is rooted at the midpoint. Phages highlighted in purple are the novel phages described in this work.

Table 5.4. Source information for MAGs included in this study

Phage	Accession/BioProject	Isolated from	Location	No. of samples found in
SAMEA2737751_a1_ct5309	PRJEB6997	Faeces (adult)	China	1
SAMEA2737768_a1_ct34917	PRJEB6997	Faeces (adult)	China	1
SAMN00792055_a1_ct11403	PRJNA422434	Faeces (adult)	China	1
SAMN05826713_a1_ct12717_vs1	PRJNA290380	Faeces (infant <1 year)	Russia	1
SAMN05826713_a1_ct6131_vs1	PRJNA290380	Faeces (infant <1 year)	Russia	1
SAMN10080877_a1_ct19236_vs1	PRJNA491626	Faeces	Cameroon	1
uvig_130754	PRJEB6997	Faeces	China	1
uvig_132550	PRJEB6997	Faeces	China	1
uvig_141073	PRJEB6997	Faeces	China	1
uvig_145376	PRJEB6997, PRJNA356102	Faeces	China	2
uvig_215036	PRJEB10878	Faeces	China	1
uvig_219619	PRJEB10878	Faeces	China	1
uvig_223573	PRJEB10878	Faeces	China	2
uvig_223847	PRJEB10878	Faeces	China	1
uvig_224277	PRJEB12123	Faeces	China	1
uvig_227178	PRJEB12123	Faeces	China	1
uvig_234015	PRJEB12123	Faeces	China	1
uvig_239791	PRJEB12123	Faeces	China	1
uvig_243694	PRJEB12124	Faeces	China	1
uvig_255004	PRJEB12124	Faeces	China	1

Phage	Accession/BioProject	Isolated from	Location	No. of samples found in
uvig_278768	PRJEB12947	Faeces	Denmark	2
uvig_279208	PRJEB12947	Faeces	Denmark	1
uvig_283917	PRJEB15111	Faeces	China	1
uvig_284377	PRJEB15111	Faeces	China	1
uvig_285149	PRJEB15111,PRJNA356102	Faeces	China	2
uvig_287240	PRJEB15111	Faeces	China	1
uvig_288431	PRJEB15111	Faeces	China	1
uvig_288643	PRJEB15111	Faeces	China	1
uvig_293010	PRJEB15371	Faeces	China	2
uvig_311634	PRJEB18755	Faeces	China	1
uvig_314355	PRJNA63661	Faeces (infant <1 year)	USA	4
uvig_323103	PRJNA422434	Faeces	China	1
uvig_326277	PRJNA422434	Faeces	China	1
uvig_327471	PRJNA422434	Faeces	China	1
uvig_328591	PRJNA422434	Faeces	China	1
uvig_329390	PRJNA422434	Faeces	China	1
uvig_330395	PRJNA422434	Faeces	China	1
uvig_331247	PRJNA422434	Faeces	China	1
uvig_334911	PRJNA422434	Faeces	China	1
uvig_334913	PRJNA422434	Faeces	China	1
uvig_338855	PRJNA177201	Faeces	Sweden	2
uvig_340901	PRJNA217052	Faeces	Fiji	16
uvig_346479	PRJNA217052	Faeces	Fiji	16

Phage	Accession/BioProject	Isolated from	Location	No. of samples found in
uvig_347013	PRJNA217052	Faeces	Fiji	16
uvig_348444	PRJNA217052	Faeces	Fiji	16
uvig_354241	PRJNA217052	Faeces	Fiji	16
uvig_369684	PRJNA272371	Faeces ((infant)	Singapore	2
uvig_376089	PRJNA283642	Faeces	USA	2
uvig_394929	PRJNA301903	Infant (< 1 year)	USA	1
uvig_437383	PRJNA375935, PRJNA353560	Faeces	China	2
uvig_464779	PRJNA356102	Faeces	China	2
uvig_467799	PRJEB6997, PRJNA356102	Faeces	China	2
uvig_474523	PRJNA356225	Faeces	China	1
uvig_535962	PRJNA354235	Faeces	USA	8
uvig_536741	PRJNA354235	Faeces	USA	8
uvig_574399	PRJNA397112	Faeces	India	1
uvig_574762	PRJNA397112	Faeces	India	1
uvig_63295	PRJEB5224	Faeces	Denmark	1
uvig_63387	PRJEB5224	Faeces	Denmark	1
Zuo_2017*	PRJNA353598	Faeces	China	1

*Full name of Zuo_2017 is GVDv1_Zuo_2017_SRR5677819_NODE_6_length_55276_cov_92.170424.

The MAGs were between 10,230 and 55,276 bp in length, with an average genome size of 41,737 bp. Thirteen of the 67 curated webervirus genomes were found to be complete and of high quality by CheckV and these genomes were added to the CheckV database (Appendix C). The updated CheckV database was used to assess the quality of the MAGs, and this analysis found that 27 of the MAGs were complete/high quality, 15 were high quality/high quality, eight were genome fragments/medium quality, and 10 were genome fragments/low quality.

The MAGs were uploaded to ViPTree online and were found to cluster within the genus *Webervirus* (not shown). Analysis with vConTACT confirmed this relationship (Figure 5.6a).

5.3.4 Host prediction for novel phages and MAGs

The seven novel webervirus genomes were uploaded to the CRISPR Spacer Database and Exploration tool [245], but this was unable to predict any host range for these phages. HostPhinder (v1.1) [246] was then used to attempt to predict the phage hosts. This tool predicted that the host of the novel phages is *K. pneumoniae*, with the same host being predicted for the 60 MAGs (Table 5.5).

5.3.5 Global distribution of weberviruses

The database of weberviruses used in this analysis contained 67 isolated phages with sequences from GenBank, and 60 MAGs identified from PhageClouds. Weberviruses have been recovered from several different sample types. Of the 127 weberviruses, 61 came from human faeces, 39 came from sewage, and 14 from wastewater. The sources activated sludge, human caecal effluent, and assorted water types had one genome associated with each of them. Seven weberviruses came from unknown sources (Figure 5.6b).

One-hundred-and-twenty-six of the genomes had a known geographical origin (Figure 5.6c). Fifty-four genomes came from China, 24 from the USA, 10 from the UK, eight from Spain, five from Fiji, four each from Denmark and Russia, three from South Korea, two each from Austria, Hungary, and India. Cameroon, Cote d'Ivoire, Pakistan, Poland, Singapore, Sweden, Thailand, and Turkey each had one genome associated with them. One-hundred-and-eight of the genomes had only been detected once. However, for MAGs with the uvig prefix from the Gut Phage Database [244] information was available with regard to the number of metagenomic samples each genome was recovered from. Eleven had been

Table 5.5. HostPhinder predictions for the 60 webervirus MAGs

Phage/MAG	*Klebsiella pneumoniae*	*Cronobacter sakazakii*	*Escherichia coli*	*Salmonella enterica*	*Shigella flexneri*	*Xanthomonas vesicatoria*
Klebsiella virus KLPN1	0.38	0.0027	0.0034	0.0018	0.0037	–
KLPN2 (T2)	0.12	0.0031	0.0042	0.0014	0.0038	0.0012
KLPN3 (T3)	0.12	0.0028	0.0038	0.0013	0.0041	–
KLPN4 (T5)	0.12	0.0036	0.0044	0.0017	0.0045	0.0012
KLPN5 (5A)	0.45	0.0032	0.0031	0.0016	0.0039	–
KLPN6 (5C)	0.11	0.0026	0.0037	0.0014	0.0044	0.0012
KLPN7 (5D)	0.47	0.0026	0.0032	0.0022	0.0042	–
KLPN8 (6E)	0.53	0.0026	0.0035	0.0017	0.0041	–
SAMEA2737751_a1_ct5309	0.41	0.003	0.0031	0.002	0.0045	–
SAMEA2737768_a1_ct34917	0.38	0.0028	0.0026	0.0014	0.0041	–
SAMN00792055_a1_ct11403	0.18	0.0035	0.0039	0.0016	0.0043	–
SAMN05826713_a1_ct12717_vs1	0.17	0.001	0.0015	–	0.0019	–
SAMN05826713_a1_ct6131_vs1	0.13	0.0012	–	–	0.00091	–
SAMN10080877_a1_ct19236_vs1	0.28	0.0022	0.003	0.0012	0.0036	–
uvig_130754	0.47	0.0027	0.0033	0.0016	0.0043	–
uvig_132550	0.38	0.0028	0.0026	0.0014	0.0041	–
uvig_141073	0.22	0.0019	0.0019	0.0013	0.0025	–
uvig_145376	0.11	0.0033	0.0038	0.0013	0.0028	–
uvig_215036	0.15	0.0033	0.0043	0.0016	0.0038	–
uvig_219619	0.48	0.0023	0.003	–	0.0041	–
uvig_223573	0.41	0.003	0.0031	0.0019	0.0044	–
uvig_223847	0.075	0.0027	0.0037	0.0011	0.0034	0.00086
uvig_224277	0.44	0.0029	0.004	0.0021	0.0047	–
uvig_227178	0.37	0.0023	0.0035	0.0016	0.0041	–
uvig_234015	0.24	0.0029	0.0025	0.0013	0.0033	–
uvig_239791	0.12	0.0024	0.0039	0.0014	0.0042	–
uvig_243694	0.12	0.0033	0.004	0.0013	0.0038	–
uvig_255004	0.17	0.0029	0.0034	0.0013	0.004	–
uvig_278768	0.5	0.0028	0.003	0.0015	0.0044	–
uvig_279208	0.5	0.0027	0.0028	0.0015	0.0042	–
uvig_283917	0.12	0.003	0.0039	0.0015	0.0044	0.0011
uvig_284377	0.12	0.0032	0.0039	0.0013	0.0037	0.001
uvig_285149	0.51	0.0027	0.0028	0.0017	0.0038	–
uvig_287240	0.24	0.0033	0.0036	0.0014	0.0034	–
uvig_288431	0.21	0.0032	0.0038	0.0011	0.004	–
uvig_288643	0.11	0.0028	0.0041	0.0015	0.004	–
uvig_293010	0.089	0.0014	0.0024	0.00076	0.0026	–
uvig_311634	0.16	0.001	0.0013	–	0.0014	–
uvig_314355	0.44	0.0029	0.0032	0.002	0.0046	–
uvig_323103	0.22	0.0018	0.0013	–	0.002	–

Phage/MAG	*Klebsiella pneumoniae*	*Cronobacter sakazakii*	*Escherichia coli*	*Salmonella enterica*	*Shigella flexneri*	*Xanthomonas vesicatoria*
uvig_326277	0.17	0.0025	0.0038	0.0015	0.004	0.0012
uvig_327471	0.12	0.0033	0.0048	0.0016	0.0048	0.0012
uvig_328591	0.12	0.0028	0.0037	0.0014	0.0038	0.00096
uvig_329390	0.21	0.003	0.0028	0.0013	0.0032	–
uvig_330395	0.18	0.0035	0.0039	0.0016	0.0043	–
uvig_331247	0.13	0.0029	0.0041	0.0013	0.0044	0.0012
uvig_334911	0.36	0.0025	0.0022	0.0012	0.0032	–
uvig_334913	0.12	0.00094	0.00086	0.00072	0.0012	–
uvig_338855	0.12	0.0011	0.0015	0.00089	0.0016	–
uvig_340901	0.12	–	0.00088	–	0.001	–
uvig_346479	0.14	–	0.0017	0.00095	0.0016	–
uvig_347013	0.1	0.0032	0.004	0.0015	0.0042	0.0011
uvig_348444	0.1	0.0032	0.004	0.0015	0.0042	0.0011
uvig_354241	0.1	0.0032	0.004	0.0015	0.0042	0.0011
uvig_369684	0.47	0.0028	0.0031	0.0018	0.0037	–
uvig_376089	0.43	0.0029	0.0035	0.0015	0.0046	–
uvig_394929	0.37	0.0026	0.003	0.0018	0.0042	–
uvig_437383	0.44	0.0033	0.0034	–	0.0044	–
uvig_464779	0.094	0.0019	0.0029	–	0.0029	–
uvig_467799	0.11	0.0033	0.0042	0.0013	0.0034	0.00086
uvig_474523	0.14	0.001	0.0022	0.00093	0.0027	–
uvig_535962	0.46	0.0026	0.0026	0.0016	0.0038	–
uvig_536741	0.46	0.0026	0.0026	0.0016	0.0038	–
uvig_574399	0.5	0.003	0.0029	0.0013	0.0039	–
uvig_574762	0.53	0.0029	0.0039	0.0021	0.0046	–
uvig_63295	0.11	–	0.002	–	0.0022	–
uvig_63387	0.1	0.00067	0.00082	–	0.0011	–
Zuo_2017	0.47	0.0035	0.0031	0.0016	0.0049	–

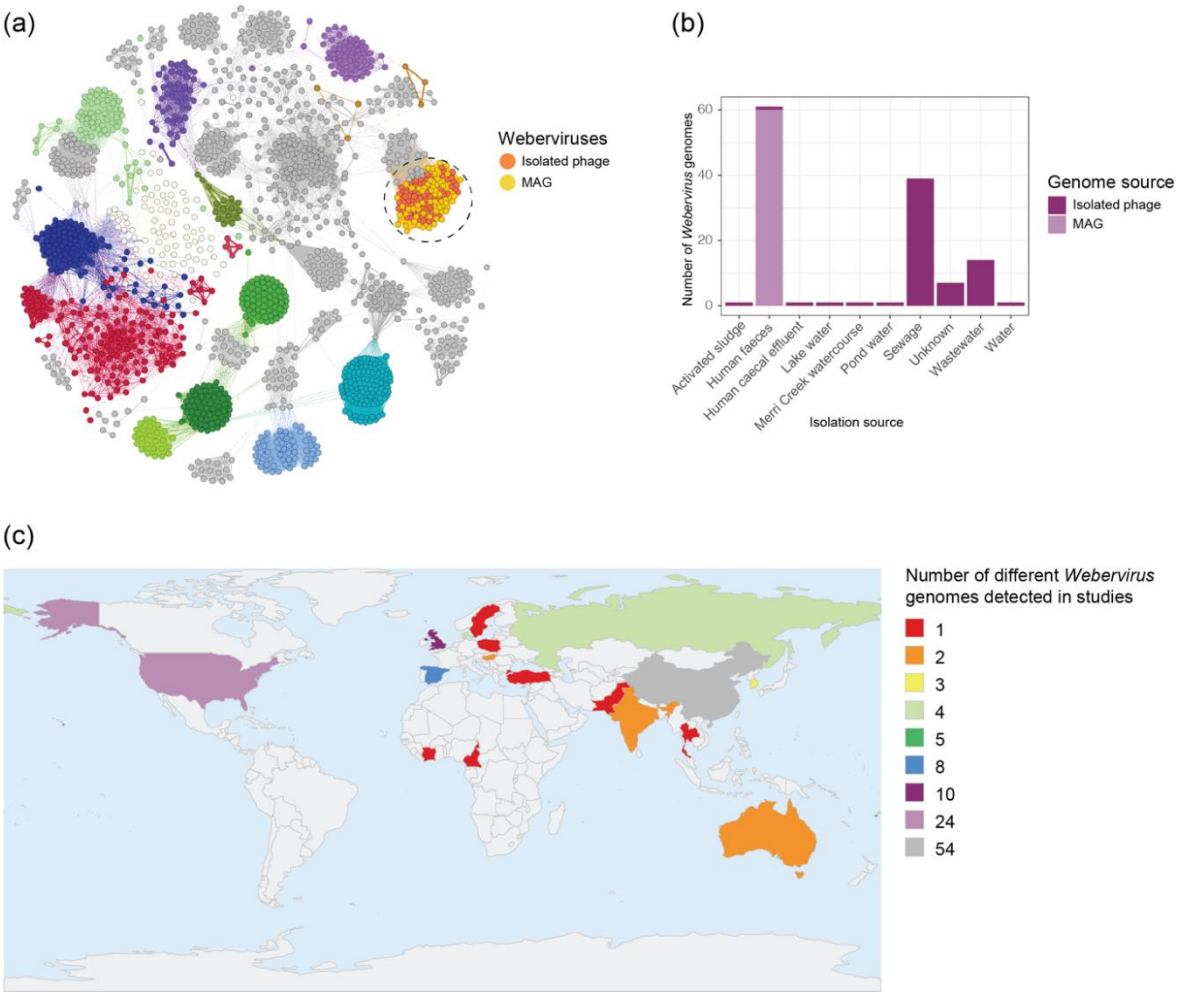


Figure 5.6. Global distribution of known weberviruses.

(a) vConTACT filtered gene-sharing network in which only nodes connected to the main cluster are shown (Fruchterman Reingold layout; filtered based on giant component; Gephi modularity options selected – randomness, use weights, resolution 1.0). The network comprises 34 modules. Only modules representing >3 % of all nodes in the network are coloured. Module 0 (represented by the circled orange and yellow nodes in the image) contained only the 127 webervirus genomes (n=67 isolated phages; n=60 MAGs). (b) Stacked bar graph showing the sources of the 127 webervirus genomes (n=67 isolated phages; n=60 MAGs). (c) Geographical distribution of 126 of the webervirus genomes included in this study (the location information was not available for one isolated phage, namely Webervirus KL).

recovered from two samples each, one recovered from four, two recovered from eight and five recovered from 16. This recovery of genomes form multiple samples was country-specific (Figure 5.8).

5.3.6 A range of depolymerases were identified within the genus *Webervirus*

The seven novel phages had all been previously shown to exhibit depolymerase activity (i.e. they had produced haloes around phage plaques on their host strains; Shibu, Brook, Garnett, Tijani, Negus & Hoyles, unpublished), prompting further analysis of depolymerases encoded by weberviruses. All 107 high-quality phages were found to encode at least one depolymerase, with 143 depolymerases being identified (Figure 5.9a). The depolymerase sequences were then used in phylogenetic analysis, which showed that the depolymerases grouped in four distinct clusters, supported by bootstrap values of 100 % (Figure 5.9b, Figure 5.10). A cluster 3 depolymerase was present in each phage (Figure 5.11), with the phage Zuo_2017 encoding two cluster 3 depolymerases. *Webervirus KP36* was found to encode a depolymerase that did not group with any of the other three clusters (Figure 5.11).

5.3.7 The structure of 6 different *Webervirus* depolymerases was predicted

3 experimentally confirmed depolymerases and 3 predicted depolymerases using Alphafold v2.3.0 and visualised using ChimeraX v1.7.1 (Figure 5.6) [210,212]. The experimentally confirmed depolymerases were from clusters 0 and 1 and appear to be structurally similar (Figure 5.7a, b, c). The predicted depolymerases were from clusters 2, 3, and 4 and appear structurally similar (Figure 5.7d, e, f). Foldseek v8-ef4e960 was used to try and compare the depolymerases with known proteins, but searches only returned hits with other phage proteins labelled as either tail or unknown proteins [253].

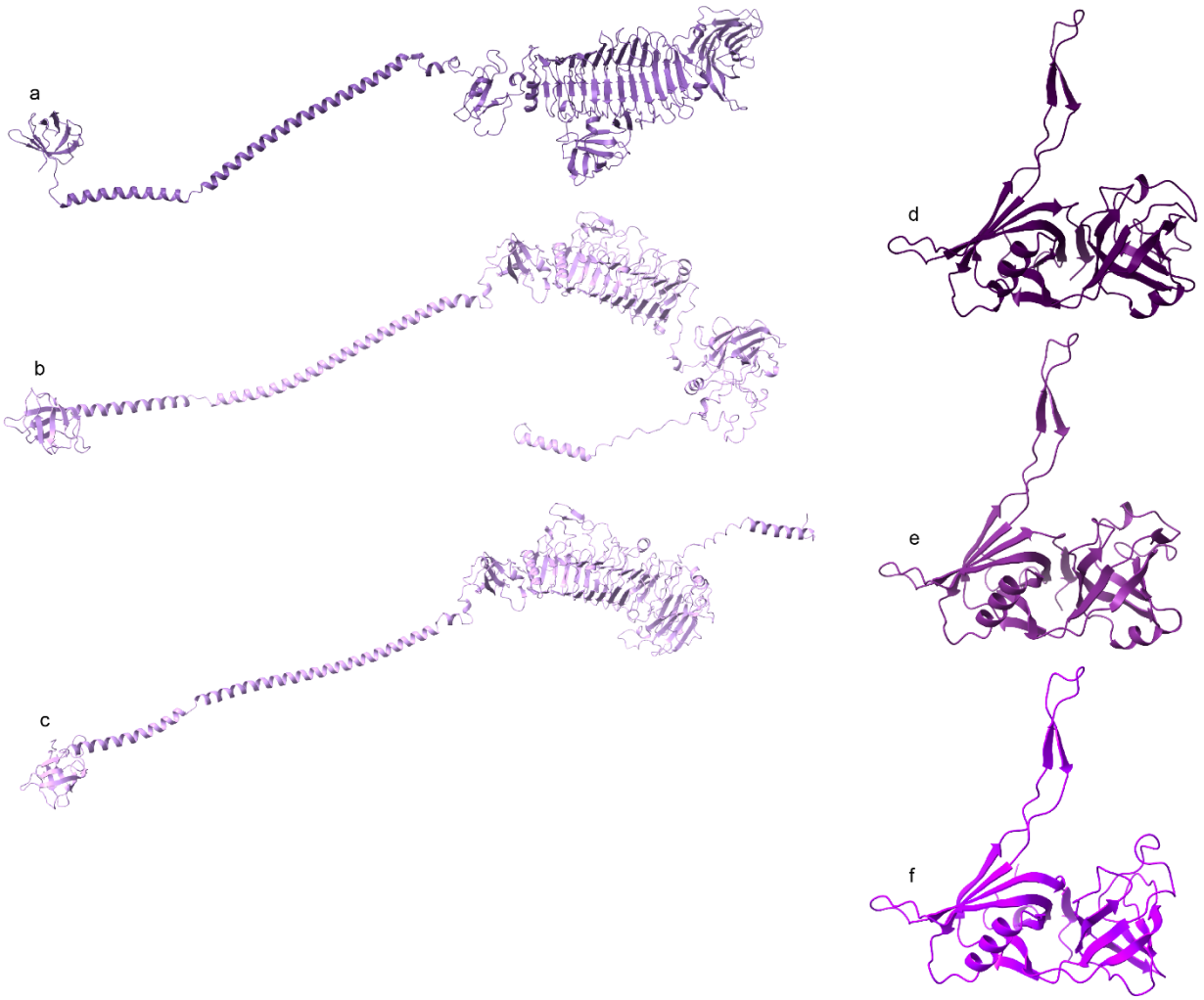


Figure 5.7. Structural predictions of Webervirus depolymerases.

Structures in figures a, b, and c have been experimentally confirmed. a) is the cluster 0 *Webervirus KP36* depolymerase. b) is the cluster 1 Klebsiella phage B1 depolymerase. c) is the cluster 1 Klebsiella phage RAD2 depolymerase.

Structures in figures d, e, and f have not been experimentally confirmed. d) is the cluster 2 Klebsiella phage vB_KpnS_FZ10 depolymerase. e) is the cluster 3 *Webervirus KLPN1* depolymerase. f) is the cluster 4 *Webervirus KP36* depolymerase.

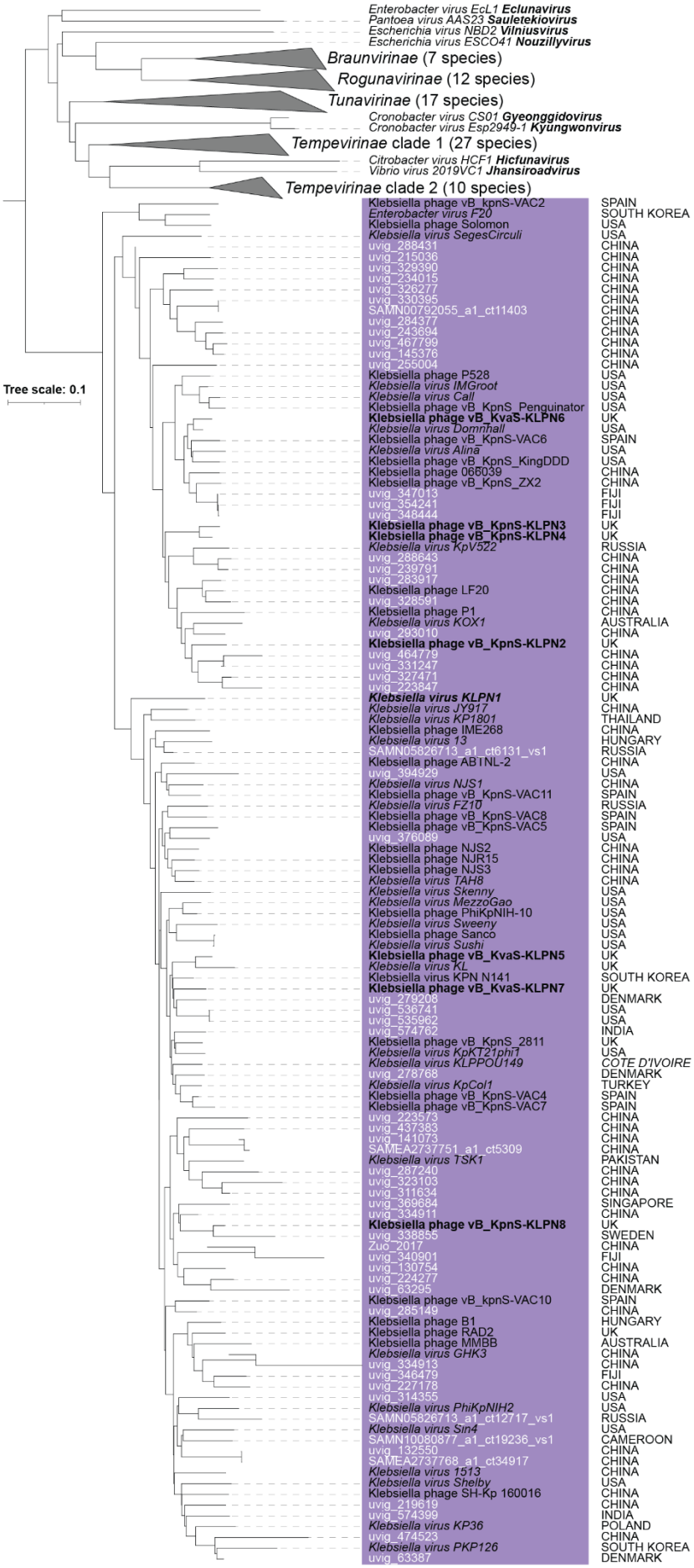


Figure 5.8. ViPTree analysis of 127 webervirus genomes based on country of isolation.

The analysis is based on proteome data encoded within each of the phage genomes. Webervirus genomes are highlighted purple. MAGs are shown in white text.

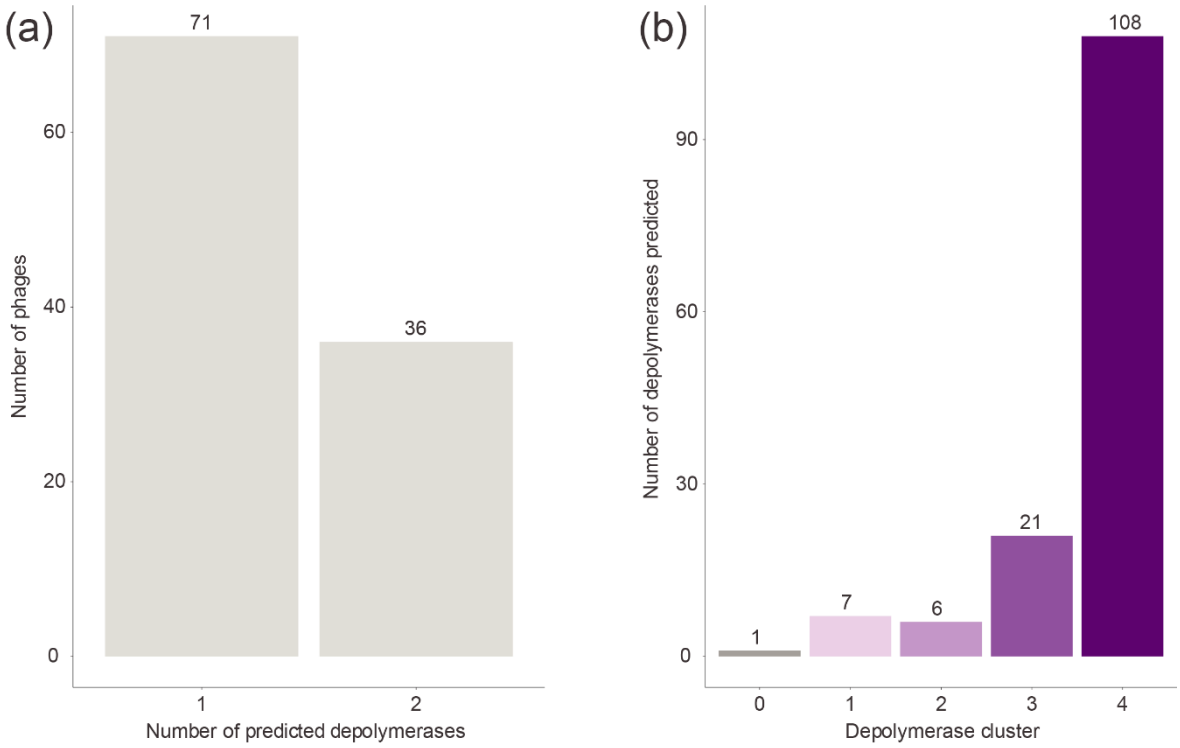


Figure 5.9. Depolymerases predicted to be encoded by weberviruses.

(a) Number of different depolymerases predicted in 107 phage genomes. (b) Number of phages encoding depolymerases belonging to the different clusters.

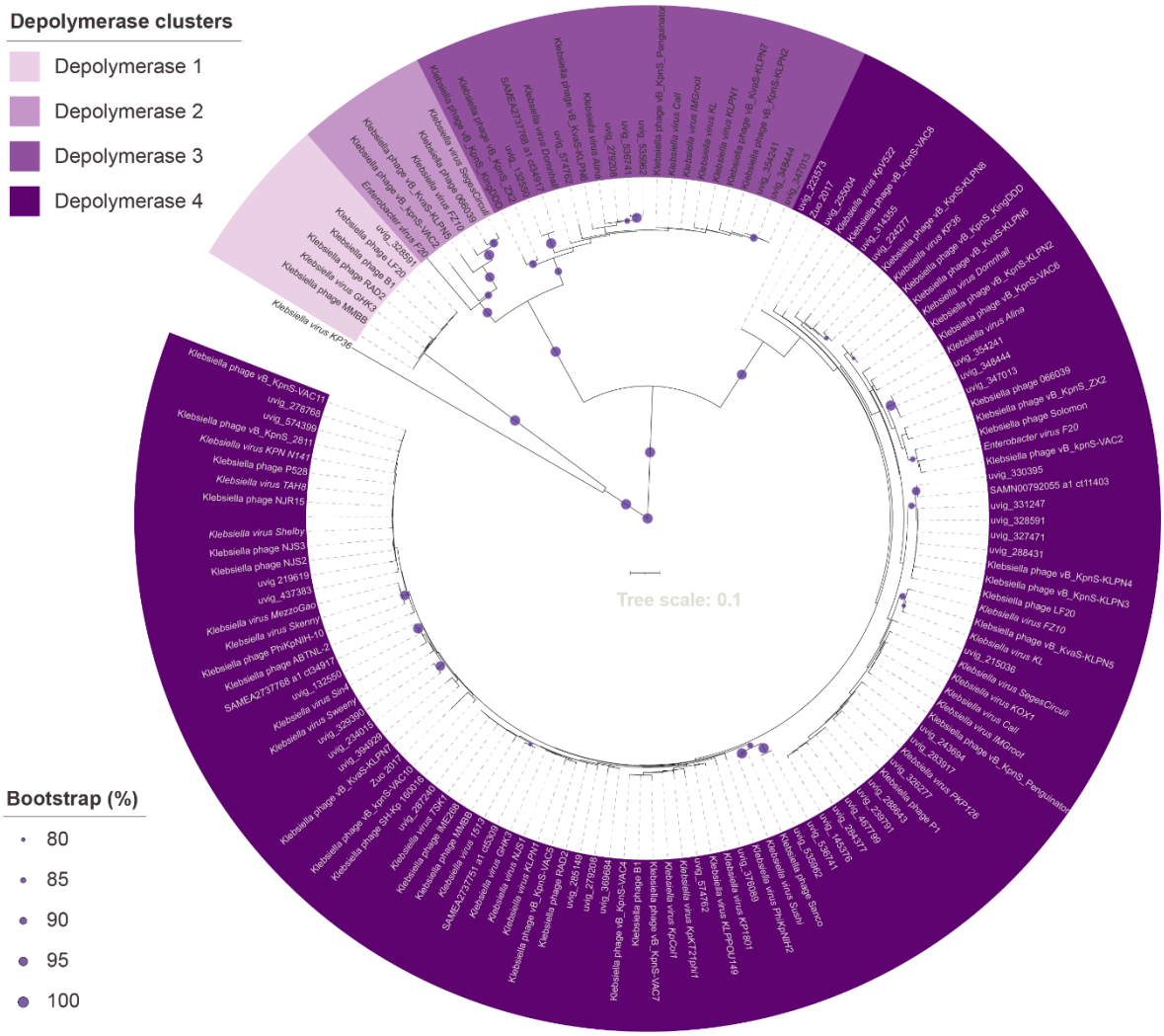


Figure 5.10. Phylogenetic analysis of protein sequences of depolymerases detected in the 107 genomes.

Bootstrap values are expressed as a percentage of 100 replications; scale bar, mean number of amino acid substitutions per position; the tree is rooted at the midpoint.

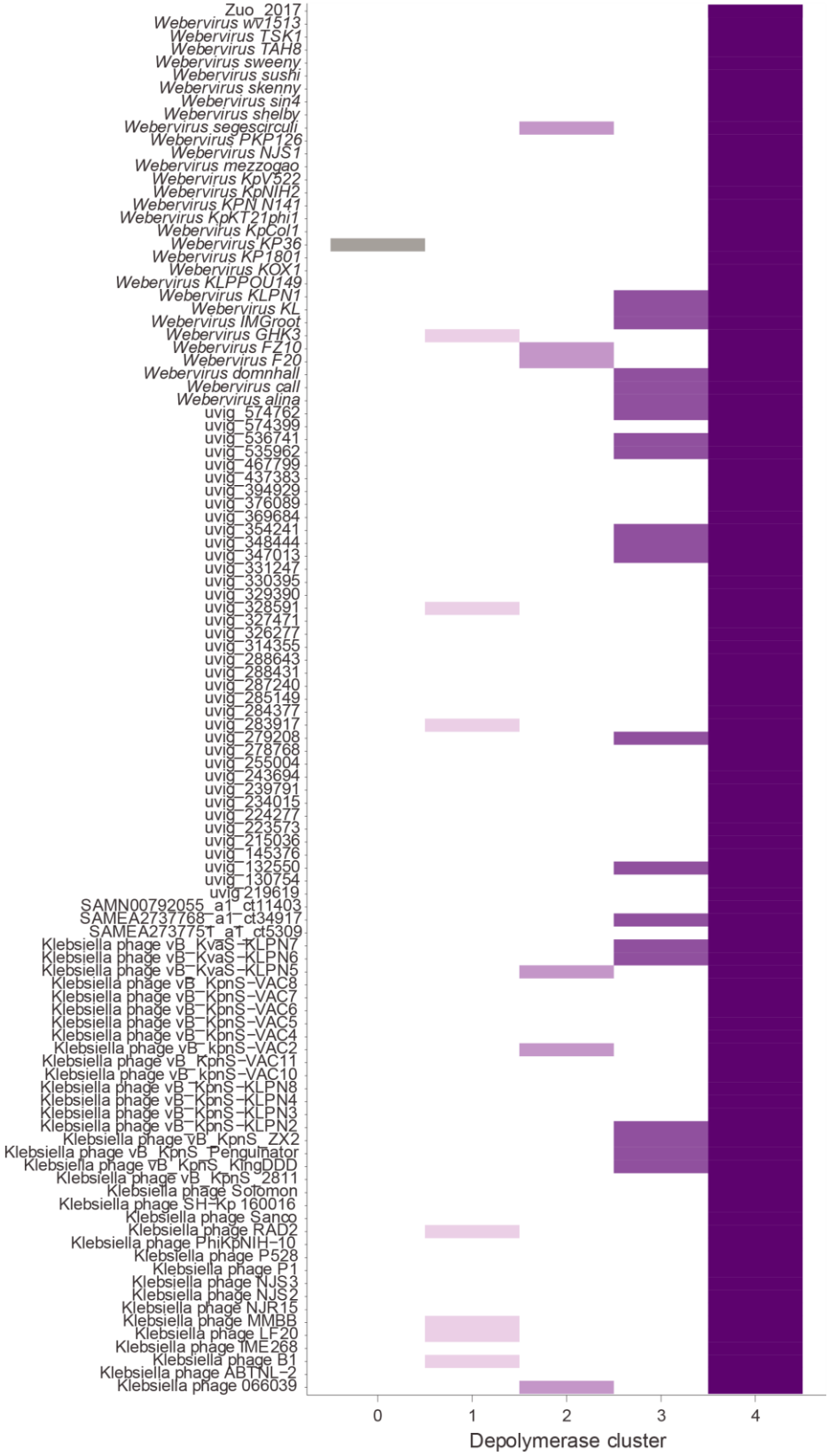


Figure 5.11. Overview of depolymerases predicted to be encoded by each phage.

Each phage examined was found to encode a cluster 4 depolymerase.

5.4 Discussion

As arguably the most abundant biological entities on Earth it is important to gain knowledge of bacteriophages and how they can interact with life. A complicated part of this understanding is viral taxonomy, which is becoming more important as phages begin to be discovered not just from isolation work but also by metagenomic analysis. Bioinformatic analysis of these phage genomes is key to their classification as viral taxonomy is now based phylogenetics and genome similarity, as opposed to the old system of classification by morphology [255,256]. This change to genomics-based classification led to the creation of the genus *Webervirus*, a genus of *Klebsiella*-infecting phages of which the first member, *Webervirus KP36*, was isolated in 2016 [225]. In this work seven novel phages were characterised with bioinformatic techniques to show that they are a part of the genus *Webervirus*, alongside an analysis of the depolymerases carried by all weberviruses.

The seven novel phages characterised in this work had been isolated and sequenced by others, but had had no further work done on them with respect to their genomes. Genomes were assembled *de novo* in this study (Figure 5.2). Annotation of the phages was carried out using the command-line tool Prokka [232] but with the usage of a different hidden-Markov model than the default, namely the PHROG database [233]. This database is optimised for the annotation of phage genomes, as opposed to prokaryotes which Prokka was originally designed for. This allows for a lesser number of annotations labelled as hypothetical proteins and more annotations being given functions. This annotation allowed for the identification of the large subunit amino acid sequences from the phages' terminase proteins. Terminases are a crucial protein in DNA viruses as they are responsible for the packaging of DNA into the protein capsid of the phage as a stage of virus particle assembly [257]. As these proteins play such a crucial role in DNA phages, terminases are almost ubiquitous among them and so can reliably be used in phylogenetic analyses. First the novel phage genomes were analysed using the tools ViPTree [234] and vConTACT [236] which compared the genomes with their own databases (Figure 5.3, 5.4). This showed a relationship between the novel phages and the genus *Webervirus*, allowing for more focussed analysis to take place. This further analysis used the large terminase subunit amino acid sequences of the novel phages, along with every publicly available genome listed as belonging to a webervirus in NCBI GenBank (Figure 5.5). This showed that the

genus *Webervirus* is monophyletic and confirmed that the novel phages belong to that genus.

As mentioned previously more phages are being discovered in metagenomic datasets instead of being isolated using classical microbiology techniques. This prompted work to see if it was possible to locate webervirus genomes in publicly available metagenome datasets. Using the genome of *Webervirus KLPN1* the tool PhageClouds [241] was used to search phage/MAG genome databases, leading to the identification of 60 MAGs that were shown to be weberviruses via vConTACT analysis. These MAG weberviruses were combined with the isolated webervirus genomes to create a curated database of 127 webervirus genomes, which the tool HostPhinder [246] had predicted were *Klebsiella*-infecting phages, that could be used in further analysis. Through examination of the sources of these phages it was seen that they have a worldwide spread (Figure 5.6).

The depolymerases of these phages are interesting as they are responsible for the phages' ability to infect their *Klebsiella* hosts as the depolymerases degrade the polysaccharide capsule of these bacteria, allowing them to inject their genetic material [228]. The degradation of this capsule by phage-derived depolymerases has been shown to make *Klebsiella* isolates more susceptible to host immune defences [228,229]. A curated database of webervirus depolymerase amino acid sequences was created through a literature review in which confirmed depolymerases were found, and usage of BLAST within the UniProt database to identify potential depolymerases within that database. This led to the creation of a BLASTp database of 16 depolymerases that was used to identify 143 depolymerases within the 127 webervirus genomes. These 143 depolymerase sequences were then subject to phylogenetic analysis, which grouped the depolymerases into four distinct clusters (Figure 5.9, 5.10, 5.11). Structural predictions were made of 3 experimentally confirmed depolymerases used in the depolymerases BLAST database and 3 predicted depolymerases to allow for structural comparisons (Figure 5.6). The cluster 0 and 1 depolymerases had been experimentally confirmed and appeared to be similar in structure, while the predicted cluster 2, 3, and 4 depolymerases were similar in structure to each other. Reasons for these similarities are unclear, and due to the limited number of experimentally confirmed structures conclusions cannot be drawn as to the effect of this on the activity of these depolymerases. Future work is planned to express and purify these

predicted depolymerases. First, this will show that depolymerases can be accurately predicted within the genomes of MAGs, and second it will examine whether the cluster of a depolymerase can be used to predict which capsule types of *Klebsiella* the depolymerase will be active against. This is possible with the usage of a company called Twist Bioscience that can create expression plasmids without needing template DNA to clone genes from.

This work has successfully characterised the novel phages that had been isolated, placing them in the genus *Webervirus*. It has also demonstrated the usage of tools that allow for the expansion of curated databases that can be used for further analysis, in this case expanding the number of *Webervirus* genomes available to be searched for depolymerases. The ongoing analysis of the depolymerases is interesting as it may allow for more accurate prediction of the host-range of weberviruses, as well as seeing if a particular group of depolymerases may be worth investigating further to explore their potential applications in the attenuation of *Klebsiella* infections via the degradation of the bacterium's protective capsule.

Chapter 6 General discussion

The relationship between human health and microbial metabolites is extremely complex, with hundreds of different microbes producing hundreds of different compounds that can interact with the human body. These interactions can be positive, like butyrate reducing the risk of colon cancer and inflammation, indole reducing liver inflammation, TMAO's protective effect on the blood brain barrier, and the SCFA-based stimulation of FFAR2 improving glucose homeostasis [33,43,44,70]. Equally some metabolites have been linked to negative effects such as imidazole propionate impairing insulin signalling, ammonia causing liver damage, TMAO's links to the development of atherosclerosis, and the presence of microbial BSHs being linked to obesity [51,53,68,74,93].

As one of the metabolites that are important to human health it is necessary to understand the microbial metabolism of TMAO in the human gut. Previous work had suggested that the protein TorA is the most prevalent and most important for TMAO metabolism in the human gut [117]. Due to the unclear methodology of this work and an absence of TorA that was seen in the *K. pneumoniae* isolate L4-FAA5, despite work [117] showing that *Klebsiella* spp. were important carriers of TorA in the gut, it was decided to explore the prevalence of TMAO metabolism proteins in human gut-associated bacteria with a new methodology.

The bioinformatic exploration of TMAO metabolism in human gut-associated bacteria started with a thorough literature review to identify what proteins may be important to TMAO metabolism. This found six different pathways, three confirmed to be TMAO reductase pathways and three other pathways that could potentially be relevant (Figure 2.1). The relevant pathways are *torCAD* [116] or TR1 which is the most studied, *torYZ* [137] or TR2, and *dmsABC* [138] which is a dimethyl sulfoxide reductase that also has TMAO reducing capabilities. The potential pathways involve *bisC* [205] which is a biotin sulfoxide reductase with a high structural similarity to TMAO reductases and high sequence identity to TR2, *msrPQ* [146] which is a membrane repair system that has been shown to reduce TMAO, and *ynfEFGH* [142] which is a homologue of *dmsABC*. A BLASTp database was built using *E. coli* sequences from UniProt (Table 2.1), which was then used to search a total of 36,064 publicly available genomes. This analysis revealed that the TorA pathway is only prevalent in a small number of genomes and is largely limited to *Citrobacter* spp. and *E. coli* (Figure

2.4, Figure 2.9). Notable for this thesis is that less than 1 % of *Klebsiella* genomes examined were found to encode any Tor proteins, despite how *Klebsiella* were predicted to be high carriers of TorA in previous work [117] (Figure 2.5). Instead, the DmsABC proteins were found to be much more prevalent in human gut-associated bacteria, meaning that this pathway may be much more important for TMAO metabolism in the human gut than TorCAD (Figure 2.10). The BisC protein was also found to be more prevalent in *Klebsiella* spp. than TorA, leading to the development of the work presented in Chapter 4 (Figure 2.10). A limitation of this work is that it only utilised protein sequences from *E. coli* and not other species. This may have caused issues with missing some hits in genomes from other species, especially from genomes that came from non-*Enterobacteriaceae* species. To improve this the TorA sequences from other species could be taken and added to a BLASTp database to see if this has any effects on the prevalence of TorA. More important though would be to focus more on characterising the activity of DmsABC vs TMAO in the human gut as this pathway appears to be much more prevalent in human gut-associated bacteria. Future work could also attempt to use the TMAO BLASTp database to attempt to predict the TMAO reductase ability of gut isolates, before testing this phenotypically to see how accurate this method of TMAO reductase activity prediction is.

L4-FAA5 is an isolate of *K. pneumoniae* subsp. *pneumoniae* that was isolated from the caecal effluent of a healthy woman [175]. This isolate had previously been sequenced and had some basic phenotypic characterisation carried out, including showing that in anaerobic conditions L4-FAA5's growth rate increased in the presence of TMAO [176]. In the work presented here L4-FAA5 was sequenced again, allowing for the full assembly of the chromosome and two plasmids (Figure 3.2). Having the full genome assembly of L4-FAA5 allowed for a much more in-depth bioinformatic characterisation of the isolate, with work being done to predict metabolic pathways, virulence, and antibiotic resistance (Table 3.3, 4, Figure 3.4, 3.5). This predicted that L4-FAA5 has the ability to utilise several different energy capture pathways, most importantly for this work being the *dmsABC*, *bisC*, and *ynfEFGH* pathways that may reduce TMAO. The commonly used pathway prediction tool ghostKOALA [184], however, did not predict any TMAO reductase ability in the genome of L4-FAA5, despite predicting the presence of all three of the pathways previously mentioned (Figure 3.3). L4-FAA5 was also predicted to carry several important virulence factors by

VFanalyzer [196]. These were capsule production genes and associated *rcsAB* and *rmpA* genes that should confer a hypermucoviscous phenotype, several different siderophores, a type 6 secretion system, and the toxin colibactin. It is worth noting that despite the prediction of a hypermucoviscous phenotype L4-FAA5 does not present this when grown on LB agar. Many different antibiotic resistance genes were also predicted using CARD [185]. These CARD resistance predictions did not match up with phenotypic data that had been gathered previously. L4-FAA5 had previously been found to be susceptible to each class of antibiotic that was predicted as 'strict' hit by card. This failure of CARD to accurately predict the antibiotic resistance phenotype of L4-FAA5 also casts doubt on other bioinformatic tools that were used to predict the metabolic capabilities of L4-FAA5. While the bioinformatic tools used in this work for this purpose may allow for prediction of metabolic capabilities, these results should be confirmed in a laboratory setting to assess their validity.

The growth of L4-FAA5 in the presence and absence of TMAO was also examined. L4-FAA5 was found to grow significantly faster in the presence of 10 mM TMAO than without, based on doubling time and CFU/mL counts (Figure 3.6). During the collection of this growth data samples were also taken for gene expression and metabolomic analysis. This analysis was planned to be done using qPCR and NMR. At time of writing qPCR primers have been validated and used tested in standard curves (Figure 3.7, 3.8). This showed that the qPCR primers form a single product, however their efficiencies in qPCR are too low (>90 %). This can cause issues in qPCR analysis by underestimating the expression of a gene, giving unreliable results. This can be resolved by a redesign of the primers to determine whether poor primer design is causing a low efficiency. Another problem may be a poor preparation of the standards used in the qPCR, which can lead to either the efficiency being too high or too low. This could be resolved by the repurifying of genomic DNA for usage as a standard to ensure that it is of high quality and then a repeat of the standard curves, but with extra attention being paid to the preparation of the standards. NMR analysis has not taken place as of writing. Method development is underway, with a preliminary spectrum being generated with an accurate quantification of TMAO in a sample of spent bacterial medium. Current issues with equipment have unfortunately led to the exclusion of metabolomic analysis from this work. Future work would involve this metabolic analysis being completed to fully examine how much TMA L4-FAA5 produces, as well as a completion of the gene

expression analysis. As well as the qPCR work, full transcriptomic analysis would be carried out to get a full picture of the effects of TMAO on the gene expression of L4-FAA5 in anaerobic conditions. This could be done using a technique such as RNA-seq.

The enzyme BisC is a poorly characterised cytoplasmic molybdoenzyme that has previously been shown to reduce biotin sulfoxide and methionine sulfoxide in *E. coli* [148,205]. BisC requires the carriage of a Moco to function, and as it contains a bis-MGD Moco it gets placed into the DMSO reductase family of molybdoenzymes, along with the previously discussed TorCAD, TorYZ and DmsABC [207]. The amino acid sequence of BisC was found to be more similar to TorZ than TorZ was to TorY, suggesting that BisC could potentially reduce TMAO. This led to a more in-depth bioinformatic analysis of the BisC carried by L4-FAA5 using the tool I-TASSER (Figure 4.1). I-TASSER predicted a >90 % structural similarity with TMAO reductases in its database, while also predicting a functional similarity with three other TMAO reductases (Table 4.2). In an attempt to confirm these predictions a plasmid, pBisK, was constructed to express BisC cloned from L4-FAA5 (Figure 4.2). Initial expressions showed promise, with BisC being successfully purified (Figure 4.4). This purified BisC was then used in a benzyl viologen assay. This assay functions by coupling reduced benzyl viologen to the reduction of TMAO, using the benzyl viologen as a source of electrons. This assay is extremely oxygen sensitive, and this caused many issues with this work. Most time with this assay was spent attempting to optimise it so that false positives caused by oxygen contamination could be eliminated. Only a single preliminary assay was able to be completed, showing that BisC may be able to reduce TMAO (Figure 4.5). Work was underway to replicate this assay and further optimise it, but issues were encountered when attempting to repurify BisC. At time of writing, the expression strain carrying pBisK appears to no longer be expressing BisC, despite being shown via PCR to still be carrying the plasmid. This was likely due to the expression strain *E. coli* BL21 (DE3) being unable to uptake molybdenum, which is necessary for the production of molybdoenzymes [221]. It was also planned to change the protein purification methods from using gravity-based column chromatography to an FPLC system, with the hopes that this could improve purity to eliminate the two extra bands seen in the SDS-PAGE gel image presented earlier (Figure 4.4). There were also plans to carry out identical work but on a BisC enzyme encoded by a

caecal isolate of *E. coli*, but these plans were put aside early on due to difficulties encountered during the L4-FAA5 BisC work.

Phages are likely the most abundant biological entities on Earth, and so it is key to understand them and how they interact with life. In the human gut phages are a major presence with an estimated 10^{13} carried in the average person [222]. The genus *Webervirus* is a genus of *Klebsiella*-infecting phages, with the first member first being isolated in 2016 and named *Webervirus KP36* [225]. Seven novel *Klebsiella*-infecting phages had previously been isolated and sequenced, but their genomes had not been assembled or characterised (Figure 5.1). In this work these phage genomes were fully assembled and annotated (Figure 5.2), before being characterised *in silico* to determine their genus. Depolymerases are notable proteins as they degrade the bacterial capsule which allows the phage to infect the decapsulated cell, as well as removing any other benefits of the capsule, such as protection from innate immune defences [226,228,229]. As part of this work the depolymerases of the genus *Webervirus* were analysed in preparation for future work to examine their effectiveness. Through the usage of the online tool ViPTree it was determined that the seven novel phage genomes belonged to the genus *Webervirus*, prompting the acquisition of all publicly available *Webervirus* sequences and the creation of a curated database of 67 *Webervirus* genomes. This database was then used with the command-line version of ViPTree [234] to again confirm the novel phages did group with other weberviruses (Figure 5.3). Analysis with vConTACT [236] also showed that the novel phages were part of the genus *Webervirus* by creating a gene-sharing network that clustered all weberviruses together (Figure 5.4). All genomes in the database were then annotated with the PHROG phage-specific hidden-Markov model using Prokka to allow for the extraction of large-subunit terminase protein sequences [232,233]. These sequences were aligned to show that the genus *Webervirus* is monophyletic (Figure 5.5). Using the *Webervirus KLPN1* genome, the online database PhageClouds [241] was searched for MAGs that could be weberviruses. This identified 60 MAGs that were analysed and shown to be a part of the genus *Webervirus*, before being combined with the 67 previously mentioned genome to create a database of 127 webervirus genomes (Figure 5.8). Using a curated depolymerase database these 127 genomes were searched for depolymerases using BLASTp. This identified 143 depolymerases, which could be grouped into four distinct clusters (Figure 5.8). Future work

plans to express and purify these depolymerases using printed DNA sequences. This would allow for proteins predicted from MAGs to be expressed without needing to isolate the specific phage, as well as allowing for expression of proteins from isolated phages that are not in a laboratory's isolate collection. These purified depolymerases would be used to examine if the cluster of a depolymerase could be used to predict which capsule types it is active against.

TMAO is thought to be used by gut bacteria, such as *Klebsiella* spp., as a terminal electron acceptor in anaerobic respiration [116,118,119]. This would allow *Klebsiella* spp. to grow faster in the anaerobic conditions of the human gut when TMAO is present, giving it an advantage against species that cannot utilise TMAO. This would also potentially give *Klebsiella* spp. an advantage if TMAO was present during an infection. TMAO also has a stabilising effect on proteins, which could help protect cells from immune-related damage in an infection [75]. Also, as TMAO can be used as a substrate by MsrPQ and potentially BisC its presence could help cells recover from damage caused by immune-related oxidative stress [145,148].

The interactions between microbial metabolism and human health is a complex subject. This work aimed to explore a small section of this field by focussing on the metabolite TMAO and how gut bacteria can interact with it. Bioinformatics proved to be a powerful tool when trying to reach these aims by allowing for many genomes to be searched for TMAO metabolic pathways, although these computational results still need validation via traditional lab-based methods as tools such as CARD show that bioinformatics can give unreliable results. Characterisation of *K. pneumoniae* L4-FAA5 also helped this work reach these aims by showing that growth rate of bacteria can be increased by the presence of TMAO, despite the isolate lacking any Tor proteins. The TMAO reductase ability of BisC has still not been fully ascertained but if this enzyme is found to be a novel TMAO reductase, it will greatly aid the understanding of which bacteria are responsible for TMAO metabolism in the human gut. Overall, this work has met its aim to contribute to the understanding of TMAO metabolism in the human gut.

COVID impact statement

Over the course of my PhD, I encountered several issues that hindered my ability to fully complete my laboratory work. Between March 2020 and January 2021, I was unable to access a lab space due the COVID lockdowns between March and July 2020, which was followed by a refurbishment of the lab where I work. This prevented me from having lab access until January 2021 as work on the lab refurbishment got delayed several times. During this time, I focused on my bioinformatics work presented in chapter 2, although personal issues hindered this as well. Unfortunately, in March 2022 I broke my leg and was unable to carry out lab work for another three months. During this time, I completed the work presented in chapter 5. Between COVID and breaking my leg I was unable to carry out practical for around 14 months.

Chapter 7 References

1. Almeida, A. *et al.* A unified catalog of 204,938 reference genomes from the human gut microbiome. *Nat Biotechnol* **39**, 105–114 (2021).

2. Thursby, E. & Juge, N. Introduction to the human gut microbiota. *Biochemical Journal* vol. 474 1823–1836 Preprint at https://doi.org/10.1042/BCJ20160510 (2017).

3. Matijašić, M. *et al.* Gut microbiota beyond bacteria-mycobiome, virome, archaeome, and eukaryotic parasites in IBD. *International Journal of Molecular Sciences* vol. 21 Preprint at https://doi.org/10.3390/ijms21082668 (2020).

4. Martinsen, T. C., Bergh, K. & Waldum, H. L. Gastric Juice: A Barrier Against Infectious Diseases. *Basic Clin Pharmacol Toxicol* **96**, 94–102 (2005).

5. Wu, W. M., Sheng, Y., Li, Y. & & Peng, H. Invited review Microbiota in the stomach: New insights. (2013) doi:10.1111/1751-2980.12116.

6. O'may, G. A., Reynolds, N., Smith, A. R., Kennedy, A. & Macfarlane, G. T. Effect of pH and Antibiotics on Microbial Overgrowth in the Stomachs and Duodena of Patients Undergoing Percutaneous Endoscopic Gastrostomy Feeding. *J Clin Microbiol* **43**, 3059–3065 (2005).

7. Procházková, N. *et al.* Advancing human gut microbiota research by considering gut transit time. *Gut* **72**, 180–191 (2023).

8. Stearns, J. C. *et al.* Bacterial biogeography of the human digestive tract. (2011) doi:10.1038/srep00170.

9. Delgado, S., Cabrera-Rubio, R., Mira, A., Suárez, A. & Mayo, B. Microbiological Survey of the Human Gastric Ecosystem Using Culturing and Pyrosequencing Methods. *Microb Ecol* **65**, 763–772 (2013).

10. Andersson, A. F., Lindberg, M., Jakobsson, H., Bäckhed, F. & Nyrén, P. Comparative Analysis of Human Gut Microbiota by Barcoded Pyrosequencing. *PLoS One* **3**, 2836 (2008).

11. Hunt, R. H. & Yaghoobi, M. The Esophageal and Gastric Microbiome in Health and Disease. *Gastroenterol Clin North Am* **46**, 121–141 (2017).

12. Dicksved, J. *et al.* Molecular characterization of the stomach microbiota in patients with gastric cancer and in controls. *J Med Microbiol* **58**, 509–516 (2009).

13. Rothenbacher, D., Blaser, M. J., Nter Bode, G. & Brenner, H. *Inverse Relationship between Gastric Colonization of Helicobacter Pylori and Diarrheal Illnesses in Children: Results of a Population-Based Cross-Sectional Study*. *The Journal of Infectious Diseases* vol. 182 (2000).

14. Atherton, J. C. & Blaser, M. J. Coadaptation of Helicobacter pylori and humans: ancient history, modern implications. *Journal of Clinical Investigation* **119**, 2475–2487 (2009).

15. Roberfroid, M. *et al.* Prebiotic effects: Metabolic and health benefits. *British Journal of Nutrition* **104**, (2010).

16. Zoetendal, E. G. *et al.* The human small intestinal microbiota is driven by rapid uptake and conversion of simple carbohydrates. *ISME J* **6**, 1415–1426 (2012).

17. Donaldson, G. P., Lee, S. M. & Mazmanian, S. K. Gut biogeography of the bacterial microbiota. *Nat Rev Microbiol* **14**, 20–32 (2015).

18. Kerckhoffs, A. P. M. *et al.* Sampling microbiota in the human gastrointestinal tract. *Gastrointestinal Microbiology* 25–50 (2006) doi:10.3109/9781420014952-3.

19. Savage, D. C. Microbial ecology of the gastrointestinal tract. *Annu Rev Microbiol* **31**, 107–133 (1977).

20. McCartney, A. L. & Gibson, G. R. The normal microbiota of the human gastrointestinal tract: History of analysis, succession, and dietary influences. *Gastrointestinal Microbiology* 51–74 (2006).

21. Wang, X., Heazlewood, S. P., Krause, D. O. & Florin, T. H. J. Molecular characterization of the microbial species that colonize human ileal and colonic mucosa by using 16S rDNA sequence analysis. *J Appl Microbiol* **95**, 508–520 (2003).

22. Eckburg, P. B. *et al.* Diversity of the Human Intestinal Microbial Flora. *Science (1979)* **308**, 1635–1638 (2005).

23. Krautkramer, K. A., Fan, J. & Bäckhed, F. Gut microbial metabolites as multi-kingdom intermediates. *Nat Rev Microbiol* **19**, (2021).

24. Zheng, X. *et al.* The footprints of gut microbial-mammalian co-metabolism. *J Proteome Res* **10**, 5512–5522 (2011).

25. Hoyles, L. *et al.* Metabolic retroconversion of trimethylamine N-oxide and the gut microbiota. *Microbiome* **6**, 73 (2018).

26. Marcobal, A. *et al.* Metabolome progression during early gut microbial colonization of gnotobiotic mice. *Sci Rep* **5**, (2015).

27. Marcobal, A. *et al.* A metabolomic view of how the human gut microbiota impacts the host metabolome using humanized and gnotobiotic mice. *ISME Journal* **7**, 1933–1943 (2013).

28. Rath, C. M. *et al.* Molecular analysis of model gut microbiotas by imaging mass spectrometry and nanodesorption electrospray ionization reveals dietary metabolite transformations. *Anal Chem* **84**, 9259–9267 (2012).

29. Chassaing, B. *et al.* Randomized Controlled-Feeding Study of Dietary Emulsifier Carboxymethylcellulose Reveals Detrimental Impacts on the Gut Microbiota and Metabolome. *Gastroenterology* **162**, 743–756 (2022).

30. Tong, M. *et al.* Reprograming of gut microbiome energy metabolism by the FUT2 Crohn's disease risk polymorphism. *ISME Journal* **8**, 2193–2206 (2014).

31. Kashyap, P. C. *et al.* Genetically dictated change in host mucus carbohydrate landscape exerts a diet-dependent effect on the gut microbiota. *Proc Natl Acad Sci U S A* **110**, 17059–17064 (2013).

32. Hoyles, L. & Swann, J. Influence of the human gut microbiome on the metabolic phenotype. in *The Handbook of Metabolic Phenotyping* 535–560 (Elsevier, 2018). doi:10.1016/B978-0-12-812293-8.00018-9.

33. Koh, A., De Vadder, F., Kovatcheva-Datchary, P. & Bäckhed, F. From dietary fiber to host physiology: Short-chain fatty acids as key bacterial metabolites. *Cell* **165**, 1332–1345 (2016).

34. Pokusaeva, K., Fitzgerald, G. F. & Van Sinderen, D. Carbohydrate metabolism in Bifidobacteria. *Genes and Nutrition* vol. 6 285–306 Preprint at https://doi.org/10.1007/s12263-010-0206-6 (2011).

35. Duncan, S. H., Louis, P. & Flint, H. J. Lactate-utilizing bacteria, isolated from human feces, that produce butyrate as a major fermentation product. *Appl Environ Microbiol* **70**, 5810–5817 (2004).

36. Tao, Y. *et al.* Production of Butyrate from Lactate by a Newly Isolated Clostridium sp. BPY5. *Appl Biochem Biotechnol* **179**, 361–374 (2016).

37. Belenguer, A. *et al.* Impact of pH on lactate formation and utilization by human fecal microbial communities. *Appl Environ Microbiol* **73**, 6526–6533 (2007).

38. Falony, G., Vlachou, A., Verbrugghe, K. & De Vuyst, L. Cross-feeding between Bifidobacterium longum BB536 and acetate-converting, butyrate-producing colon bacteria during growth on oligofructose. *Appl Environ Microbiol* **72**, 7835–7841 (2006).

39. Louis, P. & Flint, H. J. Formation of propionate and butyrate by the human colonic microbiota. *Environmental Microbiology* vol. 19 29–41 Preprint at https://doi.org/10.1111/1462-2920.13589 (2017).

40. Reichardt, N. *et al.* Phylogenetic distribution of three pathways for propionate production within the human gut microbiota. *ISME Journal* **8**, 1323–1335 (2014).

41. Hoyles, L. & Wallace, R. J. Gastrointestinal tract: intestinal fatty acid metabolism and implications for health. In Handbook of Hydrocarbon and Lipid Microbiology. 3119–3132 Preprint at (2010).

42. Dalile, B., Van Oudenhove, L., Vervliet, B. & Verbeke, K. The role of short-chain fatty acids in microbiota–gut–brain communication. *Nat Rev Gastroenterol Hepatol* **16**, 461–478 (2019).

43. Hoyles, L. *et al.* Microbiome–host systems interactions: Protective effects of propionate upon the blood–brain barrier. *Microbiome* **6**, (2018).

44. Zeng, H., Umar, S., Rust, B., Lazarova, D. & Bordonaro, M. Molecular Sciences Secondary Bile Acids and Short Chain Fatty Acids in the Colon: A Focus on Colonic Microbiome, Cell Proliferation, Inflammation, and Cancer. (2019) doi:10.3390/ijms20051214.

45. Ekechukwu, O. N. & Christian, M. Metabolic responses of light and taste receptors – unexpected actions of GPCRs in adipocytes. *Reviews in Endocrine and Metabolic Disorders* vol. 23 111–120 Preprint at https://doi.org/10.1007/s11154-021-09667-9 (2022).

46. Schlatterer, K., Peschel, A. & Kretschmer, D. Short-Chain Fatty Acid and FFAR2 Activation – A New Option for Treating Infections? *Frontiers in Cellular and Infection Microbiology* vol. 11 Preprint at https://doi.org/10.3389/fcimb.2021.785833 (2021).

47. Mishra, S. P., Karunakar, P., Taraphder, S. & Yadav, H. Free fatty acid receptors 2 and 3 as microbial metabolite sensors to shape host health: Pharmacophysiological view. *Biomedicines* vol. 8 Preprint at https://doi.org/10.3390/BIOMEDICINES8060154 (2020).

48. Grundmann, M., Bender, E., Schamberger, J. & Eitner, F. Pharmacology of free fatty acid receptors and their allosteric modulators. *International Journal of Molecular Sciences* vol. 22 1–38 Preprint at https://doi.org/10.3390/ijms22041763 (2021).

49. Frost, G. *et al.* ARTICLE The short-chain fatty acid acetate reduces appetite via a central homeostatic mechanism. *Nat Commun* (2014) doi:10.1038/ncomms4611.

50. Sarafian, M. H. *et al.* Bile Acid Profiling and Quantification in Biofluids Using Ultra-Performance Liquid Chromatography Tandem Mass Spectrometry. *Anal Chem* **87**, 9662–9670 (2015).

51. Long, S. L., Gahan, C. G. M. & Joyce, S. A. Interactions between gut bacteria and bile in health and disease. *Mol Aspects Med* **56**, 54–65 (2017).

52. Guzior, D. V. & Quinn, R. A. Review: microbial transformations of human bile acids. *Microbiome* vol. 9 Preprint at https://doi.org/10.1186/s40168-021-01101-1 (2021).

53. Joyce, S. A. & Gahan, C. G. M. Bile Acid Modifications at the Microbe-Host Interface: Potential for Nutraceutical and Pharmaceutical Interventions in Host Health. *Annu Rev Food Sci Technol* **7**, 313–333 (2016).

54. Heinken, A. *et al.* Systematic assessment of secondary bile acid metabolism in gut microbes reveals distinct metabolic capabilities in inflammatory bowel disease. *Microbiome* **7**, (2019).

55. Joyce, S. A., Shanahan, F., Hill, C. & Gahan, C. G. M. Bacterial bile salt hydrolase in host metabolism: Potential for influencing gastrointestinal microbe-host crosstalk. *Gut Microbes* **5**, 669–674 (2015).

56. Quinn, R. A. *et al.* Global chemical effects of the microbiome include new bile-acid conjugations. *Nature* **579**, (2020).

57. Fiorucci, S. & Distrutti, E. Bile Acid-Activated Receptors, Intestinal Microbiota, and the Treatment of Metabolic Disorders. *Trends Mol Med* **21**, (2015).

58. Hollman, D. A. A., Milona, A., Van Erpecum, K. J. & Van Mil, S. W. C. Anti-inflammatory and metabolic actions of FXR: Insights into molecular mechanisms. *Biochim Biophys Acta Mol Cell Biol Lipids* **1821**, 1443–1452 (2012).

59. Trabelsi, M. S., Lestavel, S., Staels, B. & Collet, X. Intestinal bile acid receptors are key regulators of glucose homeostasis. *Proceedings of the Nutrition Society* **76**, 192–202 (2017).

60. Kawamata, Y. *et al.* A G protein-coupled receptor responsive to bile acids. *Journal of Biological Chemistry* **278**, 9435–9440 (2003).

61. Morgan, W. A., NK, T. & Ding, Y. The use of High Performance Thin-Layer Chromatography to determine the role of membrane lipid composition in bile salt-induced kidney cell damage. *J Pharmacol Toxicol Methods* **57**, 70–73 (2008).

62. Paik, D. *et al.* Human gut bacteria produce T H 17-modulating bile acid metabolites. *Nature* **603**, 907 (2022).

63. Richardson, A. J., McKain, N. & Wallace, R. J. Ammonia production by human faecal bacteria, and the enumeration, isolation and characterization of bacteria capable of growth on peptides and amino acids. *BMC Microbiol* **13**, (2013).

64. Mafra, D., Barros, A. F. & Fouque, D. Dietary protein metabolism by gut microbiota and its consequences for chronic kidney disease patients. *Future Microbiol* **8**, 1317–1323 (2013).

65. Zhao, J., Zhang, X., Liu, H., Brown, M. A. & Qiao, S. Dietary Protein and Gut Microbiota Composition and Function. *Curr Protein Pept Sci* **20**, 145–154 (2018).

66. Koh, A. *et al.* Microbially Produced Imidazole Propionate Impairs Insulin Signaling through mTORC1. *Cell* **175**, 947-961.e17 (2018).

67. Molinaro, A. *et al.* Imidazole propionate is increased in diabetes and associated with dietary patterns and altered microbial ecology. *Nat Commun* **11**, (2020).

68. Hoyles, L. *et al.* Molecular phenomics and metagenomics of hepatic steatosis in non-diabetic obese women. *Nat Med* **24**, 1070–1080 (2018).

69. Wikoff, W. R. *et al. Metabolomics Analysis Reveals Large Effects of Gut Microflora on Mammalian Blood Metabolites*. www.pnas.org/cgi/content/full/ (2008).

70. Beaumont, M. *et al.* The gut microbiota metabolite indole alleviates liver inflammation in mice. *FASEB Journal* **32**, 6681–6693 (2018).

71. Roager, H. M. & Licht, T. R. Microbial tryptophan catabolites in health and disease. *Nature Communications* vol. 9 Preprint at https://doi.org/10.1038/s41467-018-05470-4 (2018).

72. Teunis, C., Nieuwdorp, M. & Hanssen, N. Interactions between Tryptophan Metabolism, the Gut Microbiome and the Immune System as Potential Drivers of Non-Alcohol Liver Disease (NAFLD) and Metabolic Diseases. *Metabolites* vol. 12 Preprint at https://doi.org/10.3390/metabo12060514 (2022).

73. Bendheim, P. E. *et al.* Development of indole-3-propionic acid (OXIGON™) for alzheimer's disease. *Journal of Molecular Neuroscience* **19**, 213–217 (2002).

74. Delgado, T. C., de las Heras, J. & Martínez-Chantar, M. L. Understanding gut-liver axis nitrogen metabolism in Fatty Liver Disease. *Frontiers in Endocrinology* vol. 13 Preprint at https://doi.org/10.3389/fendo.2022.1058101 (2022).

75. Ufnal, M., Zadlo, A. & Ostaszewski, R. TMAO: A small molecule of great expectations. *Nutrition* vol. 31 1317–1323 Preprint at https://doi.org/10.1016/j.nut.2015.05.006 (2015).

76. Taesuwan, S. *et al.* The metabolic fate of isotopically labeled trimethylamine-N-oxide (TMAO) in humans. *Journal of Nutritional Biochemistry* **45**, 77–82 (2017).

77. Zeisel SH, Wishnok JS, B. JK. Formation of Methylamines from Ingested Choline and Lecithin. *Journal of pharmacology and experimental therapies* 320–324 (1983).

78. Al-Waiz, M., Mikov, M., Mitchell, S. C. & Smith, R. L. The exogenous origin of trimethylamine in the mouse. *Metabolism* **41**, 135–136 (1992).

79. Stremmel, W. *et al.* Blood trimethylamine-n-oxide originates from microbiota mediated breakdown of phosphatidylcholine and absorption from small intestine. *PLoS One* **12**, (2017).

80. Tang, W. H. W. *et al.* Intestinal microbial metabolism of phosphatidylcholine and cardiovascular risk. *New England Journal of Medicine* **368**, 1575–1584 (2013).

81. Bain, M. A., Faull, R., Fornasini, G., Milne, R. W. & Evans, A. M. Accumulation of trimethylamine and trimethylamine-N-oxide in end-stage renal disease patients undergoing haemodialysis. *Nephrol Dial Transplant* **21**, 1300–1304 (2006).

82. Koeth, R. A. *et al.* Intestinal microbiota metabolism of l-carnitine, a nutrient in red meat, promotes atherosclerosis. *Nat Med* **19**, 576–585 (2013).

83. Wang, Z. *et al.* Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature* **472**, 57–65 (2011).

84. Lever, M. *et al.* Betaine and trimethylamine-N-oxide as predictors of cardiovascular outcomes show different patterns in diabetes mellitus: An observational study. *PLoS One* **9**, (2014).

85. Koeth, R. A. *et al.* γ-butyrobetaine is a proatherogenic intermediate in gut microbial metabolism of L-carnitine to TMAO. *Cell Metab* **20**, 799–812 (2014).

86. Gao, X. *et al.* Dietary trimethylamine N-oxide exacerbates impaired glucose tolerance in mice fed a high fat diet. *J Biosci Bioeng* **118**, 476–481 (2014).

87. Li, Q. *et al.* Synchronous evolution of an odor biosynthesis pathway and behavioral response. (2012) doi:10.1016/j.cub.2012.10.047.

88. Jia, X., Osborn, L. J. & Wang, Z. Simultaneous measurement of urinary trimethylamine (TMA) and trimethylamine N-oxide (TMAO) by liquid chromatography–mass spectrometry. *Molecules* **25**, (2020).

89. Liberles, S. D. Trace amine-associated receptors: Ligands, neural circuits, and behaviors. *Current Opinion in Neurobiology* vol. 34 1–7 Preprint at https://doi.org/10.1016/j.conb.2015.01.001 (2015).

90. Zhu, W. *et al.* Flavin monooxygenase 3, the host hepatic enzyme in the metaorganismal trimethylamine N-oxide-generating pathway, modulates platelet responsiveness and thrombosis risk. *Journal of Thrombosis and Haemostasis* **16**, 1857–1872 (2018).

91. Mueller, D. M. *et al.* Plasma levels of trimethylamine-N-oxide are confounded by impaired kidney function and poor metabolic control. *Atherosclerosis* **243**, 638–644 (2015).

92. Rhee, E. P. *et al.* A combined epidemiologic and metabolomic approach improves CKD prediction. *Journal of the American Society of Nephrology* **24**, 1330–1338 (2013).

93. Stubbs, J. R. *et al.* Serum Trimethylamine-N-Oxide is Elevated in CKD and Correlates with Coronary Atherosclerosis Burden. *Journal of the American Society of Nephrology* **27**, 305–313 (2016).

94. Tang, W. H. W. *et al.* Gut microbiota-dependent trimethylamine N-oxide (TMAO) pathway contributes to both development of renal insufficiency and mortality risk in chronic kidney disease. *Circ Res* **116**, 448–455 (2014).

95. Andrikopoulos, P. *et al.* Evidence of a causal and modifiable relationship between kidney function and circulating trimethylamine N-oxide. *Nat Commun* **14**, 5843 (2023).

96. Gibson, R. *et al.* The association of fish consumption and its urinary metabolites with cardiovascular risk factors: the International Study of Macro-/Micronutrients and Blood Pressure (INTERMAP). *Am J Clin Nutr* **111**, 280–290 (2020).

97. Koeth, R. A. *et al.* L-Carnitine in omnivorous diets induces an atherogenic gut microbial pathway in humans. *Journal of Clinical Investigation* **129**, 373–387 (2019).

98. Vallance, H. D. *et al.* Marked elevation in plasma trimethylamine-N-oxide (TMAO) in patients with mitochondrial disorders treated with oral L-carnitine. *Mol Genet Metab Rep* **15**, 130–133 (2018).

99. Fretts, A. M. *et al.* Association of Trimethylamine N -Oxide and Metabolites with Mortality in Older Adults. *JAMA Netw Open* **5**, E2213242 (2022).

100. Wang, M. *et al.* Dietary Meat, Trimethylamine N-Oxide-Related Metabolites, and Incident Cardiovascular Disease Among Older Adults: The Cardiovascular Health Study. *Arterioscler Thromb Vasc Biol* **42**, E273–E288 (2022).

101. Lee, Y. *et al.* Longitudinal plasma measures of trimethylamine N-oxide and risk of atherosclerotic cardiovascular disease events in community-based older adults. *J Am Heart Assoc* **10**, (2021).

102. Amrein, M. *et al.* Gut microbiota-dependent metabolite trimethylamine N-oxide (TMAO) and cardiovascular risk in patients with suspected functionally relevant coronary artery disease (fCAD). *Clinical Research in Cardiology* **111**, 692–704 (2022).

103. Buffa, J. A. *et al.* The microbial gbu gene cluster links cardiovascular disease risk associated with red meat consumption to microbiota l-carnitine catabolism. *Nat Microbiol* **7**, 73–86 (2022).

104. Winther, S. A. *et al.* Plasma trimethylamine N-oxide and its metabolic precursors and risk of mortality, cardiovascular and renal disease in individuals with type 2-diabetes and albuminuria. *PLoS One* **16**, (2021).

105. Li, X. S. *et al.* Untargeted metabolomics identifies trimethyllysine, a TMAO-producing nutrient precursor, as a predictor of incident cardiovascular disease risk. *JCI Insight* **3**, (2018).

106. Aldana-Hernández, P. *et al.* Dietary choline or trimethylamine N-oxide supplementation does not influence atherosclerosis development in Ldlr−/− and Apoe−/− male mice. *Journal of Nutrition* **150**, 249–255 (2020).

107. Hoyles, L. *et al.* Regulation of blood–brain barrier integrity by microbiome-associated methylamines and cognition by trimethylamine N-oxide. *Microbiome* **9**, 235 (2021).

108. Collins, H. L. *et al.* L-Carnitine intake and high trimethylamine N-oxide plasma levels correlate with low aortic lesions in ApoE-/- transgenic mice expressing CETP. *Atherosclerosis* **244**, 29–37 (2016).

109. Zhao, Z. H. *et al.* Trimethylamine N-oxide attenuates high-fat high-cholesterol diet-induced steatohepatitis by reducing hepatic cholesterol overload in rats. *World J Gastroenterol* **25**, 2450–2462 (2019).

110. Huc, T. *et al.* Chronic, low-dose TMAO treatment reduces diastolic dysfunction and heart fibrosis in hypertensive rats. *Am J Physiol Heart Circ Physiol* **315**, (2018).

111. Dumas, M. E. *et al.* Microbial-Host Co-metabolites Are Prodromal Markers Predicting Phenotypic Heterogeneity in Behavior, Obesity, and Impaired Glucose Tolerance. *Cell Rep* **20**, 136–148 (2017).

112. Videja, M. *et al.* Microbiota-Derived Metabolite Trimethylamine N-Oxide Protects Mitochondrial Energy Metabolism and Cardiac Functionality in a Rat Model of Right Ventricle Heart Failure. *Front Cell Dev Biol* **8**, 622741 (2021).

113. Stachulski, A. V., Knausenberger, T. B.-A., Shah, S. N., Hoyles, L. & McArthur, S. A host–gut microbial co-metabolite of aromatic amino acids, p-cresol glucuronide, promotes blood–brain barrier integrity in vivo. *bioRxiv* 2022.01.11.475932 (2022).

114. De Oliveira Otto, M. C. *et al.* Longitudinal Associations of Plasma TMAO and Related Metabolites with Cognitive Impairment and Dementia in Older Adults: The Cardiovascular Health Study. *Journal of Alzheimer's Disease* **89**, 1439–1452 (2022).

115. Zarour, H. M. Microbiome-derived metabolites counteract tumor-induced immunosuppression and boost immune checkpoint blockade. *Cell Metab* **34**, 1903–1905 (2022).

116. Méjean, V. *et al.* TMAO anaerobic respiration in Escherichia coli: involvement of the tor operon. *Mol Microbiol* **11**, 1169–1179 (1994).

117. Jameson, E. *et al.* Metagenomic data-mining reveals contrasting microbial populations responsible for trimethylamine formation in human gut and marine ecosystems. *Microb Genom* **2**, e000080 (2016).

118. Gon, S., Patte, J. C., Mejean, V. & Iobbi-Nivol, C. The torYZ (yecK bisZ) operon encodes a third respiratory trimethylamine N-oxide reductase in Escherichia coli. *J Bacteriol* **182**, 5779–5786 (2000).

119. Weiner, J. H., Rothery, A., Sambasivarao, D. & Trieber, C. A. *Molecular Analysis of Dimethylsulfoxide Reductase: A Complex Iroo-Sulfur Molybdoenzyme of Escherichia Co!I*. *Biochimica et Biophysica Acta* vol. 1102 (1992).

120. Ferrell, M. *et al.* Fecal microbiome composition does not predict diet-induced tmao production in healthy adults. *J Am Heart Assoc* **10**, (2021).

121. Zeisel, S. H., Mar, M.-H., Howe, J. C. & Holden, J. M. Concentrations of Choline-Containing Compounds and Betaine in Common Foods. *J Nutr* **133**, 1302–1307 (2003).

122. Craciun, S. & Balskus, E. P. Microbial conversion of choline to trimethylamine requires a glycyl radical enzyme. *Proc Natl Acad Sci U S A* **109**, 21307–21312 (2012).

123. Martínez-Del Campo, A. *et al.* Characterization and Detection of a Widely Distributed Gene Cluster That Predicts Anaerobic Choline Utilization by Human Gut Bacteria. (2015) doi:10.1128/mBio.00042-15.

124. Longo, N., Frigeni, M. & Pasquali, M. Carnitine transport and fatty acid oxidation. *Biochim Biophys Acta Mol Cell Res* **1863**, 2422–2435 (2016).

125. Zhu, Y. *et al.* Carnitine metabolism to trimethylamine by an unusual Rieske-type oxygenase from human microbiota. *Proc Natl Acad Sci U S A* **111**, 4268–4273 (2014).

126. Rajakovich, L. J., Fu, B., Bollenbach, M. & Balskus, E. P. Elucidation of an anaerobic pathway for metabolism of L-carnitine-derived γ-butyrobetaine to trimethylamine in human gut bacteria. (2021) doi:10.1073/pnas.2101498118/-/DCSupplemental.

127. Chen, Y. ran *et al.* Degradation of trimethylamine in vitro and in vivo by Enterococcus faecalis isolated from healthy human gut. *Int Biodeterior Biodegradation* **135**, 24–32 (2018).

128. Gon, S., Giudici-Orticoni, M. T., Méjean, V. & Iobbi-Nivol, C. Electron Transfer and Binding of the c-Type Cytochrome TorC to the Trimethylamine N-Oxide Reductase in Escherichia coli. *Journal of Biological Chemistry* **276**, 11545–11551 (2001).

129. Ilbert, M., Méjean, V., Giudici-Orticoni, M. T., Samama, J. P. & Iobbi-Nivol, C. Involvement of a mate chaperone (TorD) in the maturation pathway of molybdoenzyme TorA. *Journal of Biological Chemistry* **278**, 28787–28792 (2003).

130. Cox, J. C. & Knight, R. Trimethylamine N-oxide (TMAO) reductase activity in chlorate-resistant or respiration-deficient mutants of Escherichia coli. *FEMS Microbiol Lett* **12**, 249–252 (1981).

131. Ishimoto, M. & Shimokawa, O. Reduction of trimethylamine N-oxide by Escherichia coli as anaerobic respiration. *Z Allg Mikrobiol* **18**, 173–181 (1978).

132. Jourlin, C., Simon, G., Lepelletier, M., Chippaux, M. & Méjean, V. Conservation of cis-acting elements within the tor regulatory region among different Enterobacteriaceae. *Gene* **152**, 53–57 (1995).

133. Eun Kim, K. & Chang, G. W. Trimethylamine oxide reduction by Salmonella. *Can J Microbiol* **20**, 1745–1748 (1974).

134. Lemaire, O. N. *et al.* Efficient respiration on TMAO requires TorD and TorE auxiliary proteins in Shewanella oneidensis. *Res Microbiol* (2016) doi:10.1016/j.resmic.2016.05.004.

135. Carey, J. N. *et al.* Regulated Stochasticity in a Bacterial Signaling Network Permits Tolerance to a Rapid Environmental Change. *Cell* **173**, 196-207.e14 (2018).

136. Carey, J. N. & Goulian, M. A bacterial signaling system regulates noise to enable bet hedging. *Curr Genet* **65**, 65–70 (2018).

137. Del Campillo Campbell, A. & Campbell, A. Alternative gene for biotin sulfoxide reduction in Escherichia coli K-12. *J Mol Evol* **42**, 85–90 (1996).

138. Bilous, P. T., Cole, S. T., Anderson, W. F. & Weiner, J. H. Nucleotide sequence of the dmsABC operon encoding the anaerobic dimethylsulphoxide reductase of Escherichia coli. *Mol Microbiol* **2**, 785–795 (1988).

139. Weiner, J. H., Shaw, G., Turner, R. J. & Trieber, C. A. The topology of the anchor subunit of dimethyl sulfoxide reductase of Escherichia coli. *Journal of Biological Chemistry* **268**, 3238–3244 (1993).

140. Oresnik, I. J., Ladner, C. L. & Turner, R. J. Identification of a twin-arginine leader-binding protein. **40**, (2001).

141. Ray, N., Oates, J., Turner, R. J. & Robinson, C. DmsD is required for the biogenesis of DMSO reductase in Escherichia coli but not for the interaction of the DmsA signal peptide with the Tat apparatus. *FEBS Lett* **534**, 156–160 (2003).

142. Lubitz, S. P. & Weiner, J. H. The Escherichia coli ynfEFGHI operon encodes polypeptides which are paralogues of dimethyl sulfoxide reductase (DmsABC). *Arch Biochem Biophys* **418**, 205–216 (2003).

143. Bearson, S., Albrecht, J. A. & Gunsalus, R. P. Oxygen and nitrate-dependent regulation of dmsABC operon expression inEscherichia coli: Sites for Fnr and NarL protein interactions. *BMC Microbiol* **2**, 1–10 (2002).

144. Guymer, D., Maillard, J. & Sargent, F. A genetic analysis of in vivo selenate reduction by Salmonella enterica serovar Typhimurium LT2 and Escherichia coli K12. *Arch Microbiol* **191**, 519–528 (2009).

145. Gennaris, A. *et al.* Repairing oxidized proteins in the bacterial envelope using respiratory chain electrons. *Nature* **528**, 409–412 (2015).

146. Loschi, L. *et al.* Structural and biochemical identification of a novel bacterial oxidoreductase. *Journal of Biological Chemistry* **279**, 50391–50400 (2004).

147. Sargent, F. Constructing the wonders of the bacterial world: Biosynthesis of complex enzymes. *Microbiology (N Y)* **153**, 633–651 (2007).

148. Ezraty, B., Bos, J., Barras, F. & Aussel, L. Methionine sulfoxide reduction and assimilation in Escherichia coli: New role for the biotin sulfoxide reductase BisC. *J Bacteriol* **187**, 231–237 (2005).

149. Pollock, V. V. & Barber, M. J. Biotin Sulfoxide Reductase. *Journal of Biological Chemistry* **272**, 3355–3362 (1997).

150. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506–D515 (2019).

151. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, (2011).

152. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 1–8 (2018).

153. Chun, J. *et al.* Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol* **68**, 461–466 (2018).

154. Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open Peer Review Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications [version 1; peer review: 2 approved]. (2018) doi:10.12688/wellcomeopenres.14826.1.

155. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).

156. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).

157. Parks, D. H. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol* (2020) doi:10.1038/s41587-020-0501-8.

158. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res* **44**, D67–D72 (2016).

159. Ravcheev, D. A. & Thiele, I. Systematic genomic analysis reveals the complementary aerobic and anaerobic respiration capacities of the human gut microbiota. *Front Microbiol* **5**, 1–14 (2014).

160. Hyatt, D. *et al.* Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

161. Al-Waiz, M., Ayesh, R., Mitchell, S. C., Idle, J. R. & Smith, R. L. Disclosure of the metabolic retroversion of trimethylamine N-oxide in humans: A pharmacogenetic approach. *Clin Pharmacol Ther* **42**, 608–612 (1987).

162. Almeida, A. *et al.* A new genomic blueprint of the human gut microbiota. *Nature* **568**, 499–504 (2019).

163. Pasolli, E. *et al.* Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649-662.e20 (2019).

164. Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505–510 (2019).

165. Schoch, C. L. *et al.* NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database* vol. 2020 Preprint at https://doi.org/10.1093/database/baaa062 (2020).

166. Podschun, R. & Ullman, U. Klebsiella pneumoniae as nosocomial pathogens: epidemiology and resistance. *Clin Microbiol Rev* **11**, 589–603 (1998).

167. Herridge, W. P., Shibu, P., O'shea, J., Brook, T. C. & Hoyles, L. Bacteriophages of Klebsiella spp., their diversity and potential therapeutic uses. *J Med Microbiol* **69**, 176–194 (2020).

168. Russo, T. A. & Marr, C. M. *Hypervirulent Klebsiella Pneumoniae*. https://journals.asm.org/journal/cmr (2019).

169. Calderon-Gonzalez, R. *et al.* Modelling the Gastrointestinal Carriage of Klebsiella pneumoniae Infections. *mBio* **14**, (2023).

170. Yilmaz, B. *et al.* Plasticity of the adult human small intestinal stoma microbiota. *Cell Host Microbe* **30**, 1773-1787.e6 (2022).

171. Bayoumy, A. B., Mulder, C. J. J., Mol, J. J. & Tushuizen, M. E. Gut fermentation syndrome: A systematic review of case reports. *United European Gastroenterol J* **9**, 332–342 (2021).

172. Yuan, J. *et al.* Fatty Liver Disease Caused by High-Alcohol-Producing Klebsiella pneumoniae. *Cell Metab* **30**, 675-688.e7 (2019).

173. Gan, L. *et al.* Bacteriophage targeting microbiota alleviates non-alcoholic fatty liver disease induced by high alcohol-producing Klebsiella pneumoniae. *Nat Commun* **14**, (2023).

174. Federici, S. *et al.* Targeted suppression of human IBD-associated gut microbiota commensals by phage consortia for treatment of intestinal inflammation. *Cell* **185**, 2879-2898.e24 (2022).

175. Hoyles, L. *et al.* Klebsiella pneumoniae subsp. pneumoniae-bacteriophage combination fromthe caecal effluent of a healthy woman. *PeerJ* **2015**, (2015).

176. Chen. Influence of trimethylamine N -oxide on metabolism and gene expression of Klebsiella pneumoniae L4-FAA5 isolated from the human caecum. *MRes thesis for Imperial College London* 1–63 (2017).

177. Newberry, F. *et al.* Lytic bacteriophage vB_KmiS-Kmi2C disrupts biofilms formed by members of the Klebsiella oxytoca complex, and represents a novel virus family and genus. *J Appl Microbiol* **134**, (2023).

178. Schwengers, O. *et al.* Bakta: Rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microb Genom* **7**, (2021).

179. Cumsille, A. *et al.* GenoVi, an open-source automated circular genome visualizer for bacteria and archaea. *PLoS Comput Biol* **19**, (2023).

180. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043–1055 (2015).

181. Lam, M. M. C., Wick, R. R., Judd, L. M., Holt, K. E. & Wyres, K. L. Kaptive 2.0: updated capsule and lipopolysaccharide locus typing for the Klebsiella pneumoniae species complex. *Microb Genom* **8**, (2022).

182. Carattoli, A. *et al.* In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother* **58**, 3895–3903 (2014).

183. Pascal Andreu, V., Roel-Touris, J., Dodd, D., Fischbach, M. A. & Medema, M. H. The gutSMASH web server: Automated identification of primary metabolic gene clusters from the gut microbiota. *Nucleic Acids Res* **49**, W263–W270 (2021).

184. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* **428**, 726–731 (2016).

185. Alcock, B. P. *et al.* CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res* **51**, D690–D699 (2023).

186. Liu, B., Zheng, D., Zhou, S., Chen, L. & Yang, J. VFDB 2022: A general classification scheme for bacterial virulence factors. *Nucleic Acids Res* **50**, D912–D917 (2022).

187. Sprouffske, K. & Wagner, A. Growthcurver: an R package for obtaining interpretable metrics from microbial growth curves. *BMC Bioinformatics* **17**, 172 (2016).

188. Untergasser, A. *et al.* Primer3-new capabilities and interfaces. *Nucleic Acids Res* **40**, (2012).

189. Bowers, R. M. *et al.* Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* **35**, 725–731 (2017).

190. Unden, G. & Dünnwald, P. The Aerobic and Anaerobic Respiratory Chain of Escherichia coli and Salmonella enterica : Enzymes and Energetics . *EcoSal Plus* **3**, (2008).

191. Tremblay, P. L., Zhang, T., Dar, S. A., Leang, C. & Lovley, D. R. The Rnf complex of Clostridium ljungdahlii is a proton-translocating ferredoxin: NAD+ oxidoreductase essential for autotrophic growth. *mBio* **4**, (2013).

192. Okamura-Ikeda, K., Ohmura, Y., Fujiwara, K. & Motokawa, Y. Cloning and nucleotide sequence of the gcv operon encoding the Escherichia coli glycine-cleavage system. *Eur J Biochem* **216**, 539–548 (1993).

193. Khademian, M. & Imlay, J. A. Do reactive oxygen species or does oxygen itself confer obligate anaerobiosis? The case of Bacteroides thetaiotaomicron. *Mol Microbiol* **114**, 333–347 (2020).

194. Engels, C., Ruscheweyh, H. J., Beerenwinkel, N., Lacroix, C. & Schwab, C. The common gut microbe Eubacterium hallii also contributes to intestinal propionate formation. *Front Microbiol* **7**, (2016).

195. Chen, L. *et al.* VFDB: A reference database for bacterial virulence factors. *Nucleic Acids Res* **33**, (2005).

196. Liu, B., Zheng, D., Jin, Q., Chen, L. & Yang, J. VFDB 2019: A comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res* **47**, D687–D692 (2019).

197. Turton, J. F., Perry, C., Elgohari, S. & Hampton, C. V. PCR characterization and typing of Klebsiella pneumoniae using capsular type-specific, variable number tandem repeat and virulence gene targets. *J Med Microbiol* **59**, 541–547 (2010).

198. Shibu, P. & Hoyles, L. L4-FAA5 characterisation.

199. Zhu, J., Wang, T., Chen, L. & Du, H. Virulence Factors in Hypervirulent Klebsiella pneumoniae. *Frontiers in Microbiology* vol. 12 Preprint at https://doi.org/10.3389/fmicb.2021.642484 (2021).

200. Schroll, C., Barken, K. B., Krogfelt, K. A. & Struve, C. *Role of Type 1 and Type 3 Fimbriae in Klebsiella Pneumoniae Biofilm Formation*. *BMC Microbiology* vol. 10 http://www.biomedcentral.com/1471-2180/10/179 (2010).

201. Mousa, W. K. The microbiome-product colibactin hits unique cellular targets mediating host–microbe interaction. *Frontiers in Pharmacology* vol. 13 Preprint at https://doi.org/10.3389/fphar.2022.958012 (2022).

202. Vergnes, A. *et al.* Periplasmic oxidized-protein repair during copper stress in E. coli: A focus on the metallochaperone CusF. *PLoS Genet* **18**, (2022).

203. Dykhuizen', D. *Genetic Analysis of the System That Reduces Biotin-d-Sulfoxide in Escherichia Coli*. *JOURNAL OF BACTERIOLOGY* vol. 115 https://journals.asm.org/journal/jb (1973).

204. Campillo-Campbell, A. del & Campbell, A. *Molybdenum Cofactor Requirement for Biotin Sulfoxide Reduction in Escherichia Coli*. *JOURNAL OF BACTERIOLOGY* vol. 149 https://journals.asm.org/journal/jb (1982).

205. Pierson, D. E. & Campbell, A. *Cloning and Nucleotide Sequence of BisC, the Structural Gene for Biotin Sulfoxide Reductase in Escherichia Coli*. *JOURNAL OF BACTERIOLOGY* vol. 172 (1990).

206. Satiaputra, J., Shearwin, K. E., Booker, G. W. & Polyak, S. W. Mechanisms of biotin-regulated gene expression in microbes. *Synth Syst Biotechnol* **1**, 17–24 (2016).

207. Iobbi-Nivol, C. & Leimkühler, S. Molybdenum enzymes, their maturation and molybdenum cofactor biosynthesis in Escherichia coli ☆. *BBA - Bioenergetics* **1827**, 1086–1101 (2013).

208. Leimkühler, S. The biosynthesis of the molybdenum cofactors in Escherichia coli. (2020) doi:10.1111/1462-2920.15003.

209. Yang, J. & Zhang, Y. I-TASSER server: New development for protein structure and function predictions. *Nucleic Acids Res* **43**, W174–W181 (2015).

210. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

211. Hekkelman, M. L., de Vries, I., Joosten, R. P. & Perrakis, A. AlphaFill: enriching AlphaFold models with ligands and cofactors. *Nat Methods* **20**, 205–213 (2023).

212. Meng, E. C. *et al.* UCSF ChimeraX: Tools for structure building and analysis. *Protein Science* **32**, (2023).

213. Meng, E. C., Pettersen, E. F., Couch, G. S., Huang, C. C. & Ferrin, T. E. Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics* **7**, 339 (2006).

214. Aledo, J. C. renz: An R package for the analysis of enzyme kinetic data. *BMC Bioinformatics* **23**, (2022).

215. Czjzek, M. *et al. Crystal Structure of Oxidized Trimethylamine N-Oxide Reductase from Shewanella Massilia at 2.5 A Ê Resolution*. (1998).

216. Mcalpine, A. S., Mcewan, A. G., Shaw, A. L. & Bailey, S. *Molybdenum Active Centre of DMSO Reductase from Rhodobacter Capsulatus : Crystal Structure of the Oxidised Enzyme at 1.82-Å Resolution and the Dithionite-Reduced Enzyme at 2.8-Å Resolution*. *JBIC* vol. 2 (1997).

217. Struwe, M. A. *et al.* Active site architecture reveals coordination sphere flexibility and specificity determinants in a group of closely related molybdoenzymes. *Journal of Biological Chemistry* **296**, (2021).

218. Li, H. K., Temple, C., Rajagopalan, K. V. & Schindelin, H. The 1.3 Å Crystal structure of rhodobacter sphaeroides dimethyl sulfoxide reductase reveals two distinct molybdenum coordination environments. *J Am Chem Soc* **122**, 7673–7680 (2000).

219. Sahdev, S., Khattar, S. K. & Saini, K. S. Production of active eukaryotic proteins through bacterial expression systems: A review of the existing biotechnology strategies. *Molecular and Cellular Biochemistry* vol. 307 249–264 Preprint at https://doi.org/10.1007/s11010-007-9603-6 (2008).

220. San-Miguel, T., Pérez-Bermúdez, P. & Gavidia, I. Production of soluble eukaryotic recombinant proteins in E. coli is favoured in early log-phase cultures induced at low temperature. *Springerplus* **2**, 1–4 (2013).

221. Pinske, C., Bönn, M., Krüger, S., Lindenstrauß, U. & Sawers, R. G. Metabolic Deficiences Revealed in the Biotechnologically Important Model Bacterium Escherichia coli BL21(DE3). *PLoS One* **6**, e22830 (2011).

222. Liang, G. & Bushman, F. D. The human virome: assembly, composition and host interactions. *Nature Reviews Microbiology* vol. 19 514–527 Preprint at https://doi.org/10.1038/s41579-021-00536-5 (2021).

223. Letarov, A. & Kulikov, E. The bacteriophages in human- and animal body-associated microbial communities. *Journal of Applied Microbiology* vol. 107 1–13 Preprint at https://doi.org/10.1111/j.1365-2672.2009.04143.x (2009).

224. Clokie, M. R. J., Millard, A. D., Letarov, A. V. & Heaphy, S. Phages in nature. *Bacteriophage* **1**, 31–45 (2011).

225. Kęsik-Szeloch, A. *et al.* Characterising the biology of novel lytic bacteriophages infecting multidrug resistant Klebsiella pneumoniae. *Virol J* **10**, (2013).

226. Majkowska-Skrobek, G. *et al.* Capsule-targeting depolymerase, derived from klebsiella KP36 phage, as a tool for the development of anti-virulent strategy. *Viruses* **8**, (2016).

227. Lefkowitz, E. J. *et al.* Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res* **46**, D708–D717 (2018).

228. Latka, A., Maciejewska, B., Majkowska-Skrobek, G., Briers, Y. & Drulis-Kawa, Z. Bacteriophage-encoded virion-associated enzymes to overcome the carbohydrate barriers during the infection process. *Applied Microbiology and Biotechnology* vol. 101 3103–3119 Preprint at https://doi.org/10.1007/s00253-017-8224-6 (2017).

229. Majkowska-Skrobek, G. *et al.* Phage-borne depolymerases decrease Klebsiella pneumoniae resistance to innate defense mechanisms. *Front Microbiol* **9**, (2018).

230. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **19**, 455–477 (2012).

231. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* **39**, 578–585 (2021).

232. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

233. Terzian, P. *et al.* PHROG: Families of prokaryotic virus proteins clustered using remote homology. *NAR Genom Bioinform* **3**, (2021).

234. Nishimura, Y. *et al.* ViPTree: The viral proteomic tree server. *Bioinformatics* **33**, 2379–2380 (2017).

235. Moraru, C., Varsani, A. & Kropinski, A. M. VIRIDIC—A novel tool to calculate the intergenomic similarities of prokaryote-infecting viruses. *Viruses* **12**, (2020).

236. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* **37**, 632–639 (2019).

237. Pagès, H., Aboyoun, P., Gentleman, R. & DebRoy, S. Biostrings: Efficient manipulation of biological strings. *R Package* (2021).

238. Schliep, K. P. phangorn: Phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).

239. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**, 289–290 (2004).

240. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol Biol Evol* **25**, 1307–1320 (2008).

241. Rangel-Pineros, G. *et al.* From Trees to Clouds: PhageClouds for Fast Comparison of ∼640,000 Phage Genomic Sequences and Host-Centric Visualization Using Genomic Network Graphs. *PHAGE* **2**, 194–203 (2021).

242. Tisza, M. J. & Buck, C. B. A catalog of tens of thousands of viruses from human metagenomes reveals hidden associations with chronic diseases. *Proceedings of the National Academy of Sciences* **118**, (2021).

243. Gregory, A. C. *et al.* The Gut Virome Database Reveals Age-Dependent Patterns of Virome Diversity in the Human Gut. *Cell Host Microbe* **28**, 724-740.e8 (2020).

244. Camarillo-Guerrero, L. F., Almeida, A., Rangel-Pineros, G., Finn, R. D. & Lawley, T. D. Massive expansion of human gut bacteriophage diversity. *Cell* **184**, 1098-1109.e9 (2021).

245. Dion, M. B. *et al.* Streamlining CRISPR spacer-based bacterial host predictions to decipher the viral dark matter. *Nucleic Acids Res* **49**, 3127–3138 (2021).

246. Villarroel, J. *et al.* HostPhinder: A Phage Host Prediction Tool. *Viruses* **8**, 116 (2016).

247. South, A. *Rworldmap: A New R Package for Mapping Global Data*. http://www.un.org/millenniumgoals/bkgd. (2011).

248. Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C. & Hochreiter, S. msa: an R package for multiple sequence alignment. *Bioinformatics* **31**, 3997–3999 (2015).

249. Whelan, S. & Goldman, N. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Mol. Biol. Evol* **18**, 691–699 (2001).

250. Pertics, B. Z. *et al.* Isolation and Characterization of a Novel Lytic Bacteriophage against the K2 Capsule-Expressing Hypervirulent Klebsiella pneumoniae Strain 52145, and Identification of Its Functional Depolymerase. *Microorganisms* **9**, 650 (2021).

251. Bleriot, I. *et al.* «PemIK (PemK/PemI) type II TA system from Klebsiella pneumoniae clinical strains inhibits lytic phage». *Res Sq* (2021) doi:10.21203/rs.3.rs-679460/v1.

252. Cai, R. *et al.* Biological properties and genomics analysis of vB_KpnS_GH-K3, a Klebsiella phage with a putative depolymerase-like protein. *Virus Genes* **55**, 696–706 (2019).

253. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243–246 (2024).

254. ter Horst, A. M. *et al.* Minnesota peat viromes reveal terrestrial and aquatic niche partitioning for local and global viral populations. *Microbiome* **9**, 233 (2021).

255. Walker, P. J. *et al.* Changes to virus taxonomy and to the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). *Arch Virol* **166**, 2633–2648 (2021).

256. Turner, D., Kropinski, A. M. & Adriaenssens, E. M. A roadmap for genome-based phage taxonomy. *Viruses* **13**, (2021).

257. Shen, X. *et al.* Functional identification of the DNA packaging terminase from Pseudomonas aeruginosa phage PaP3. *Arch Virol* **157**, 2133–2141 (2012).

Chapter 8 Appendices

Appendix A – Results from lowering the identity threshold to 50 % for the BLAST search of *Klebsiella* genomes

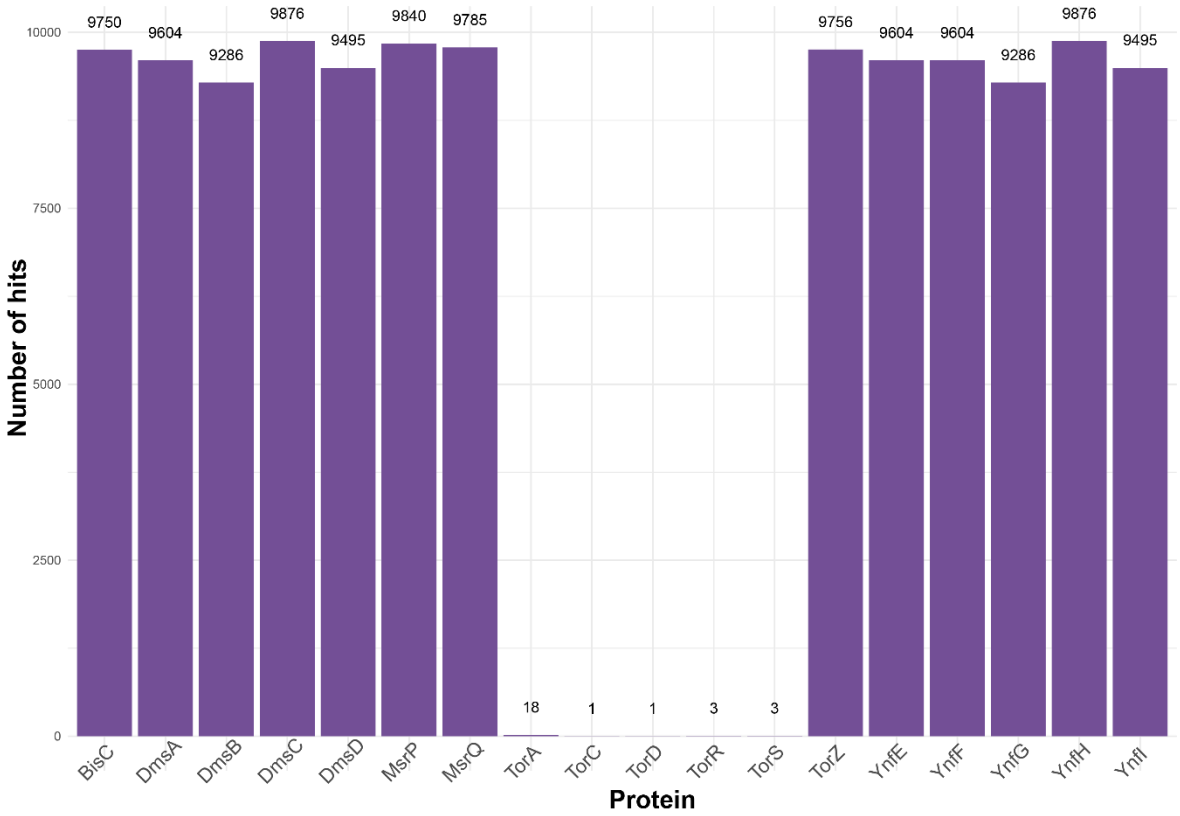


Figure A.1. Results from lowering the identity threshold to 50 % for the BLAST search of *Klebsiella* genomes.

A decrease in the identity threshold of the TMAO metabolism BLAST search did not yield a large increase in the number of TMAO metabolism proteins found in the genomes of *Klebsiella* spp.

Appendix B – Code for analyses of growth curve data

```
library(tidyverse) #v2.0.0
library(growthcurver) #v0.3.1
library(reshape2) #v1.4.4
library(matrixStats) #0.63.0
library(ggpubr) #0.6.0

setwd("~/Desktop")

plate1=read_tsv('TMAO_growth_curves.txt', col_names=T)
colnames(plate1)[1] = 'Time'


samples=c(rep('Control', 9), rep('TMAO', 9))
sample.info=data.frame(Treatment=samples, Cell=colnames(plate1)[2:19])
```

```
sample.info$Replicate = sample.info$Cell
sample.info$Replicate = sub(1, '', sample.info$Replicate)
sample.info$Replicate = sub(2, '', sample.info$Replicate)
sample.info$Replicate = sub(3, '', sample.info$Replicate)


gc_fit_full_plate1 = SummarizeGrowthByPlate(plate1)
gc_fit_full_plate1 = data.frame(gc_fit_full_plate1, sample.info)



data_with_controls = function(d=plate1) {
       num_analyses <- length(names(d)) - 1
       d_gc <- data.frame(sample = character(num_analyses),
                   k = numeric(num_analyses),
                   n0  = numeric(num_analyses),
                   r = numeric(num_analyses),
                   t_mid = numeric(num_analyses),
                   t_gen = numeric(num_analyses),
                   auc_l = numeric(num_analyses),
                   auc_e = numeric(num_analyses),
                   sigma = numeric(num_analyses),
                   stringsAsFactors = FALSE)

       trim_at_time <- 24 * 30


       par(mfcol = c(12,8))
       par(mar = c(0.25,0.25,0.25,0.25))
       y_lim_max <- max(d[,setdiff(names(d), "Time")]) - min(d[,setdiff(names(d),
"Time")])

       n <- 1    # keeps track of the current row in the output data frame
       for (col_name in names(d)) {

         # Don't process the column called "Time".
        # It contains time and not absorbance data.
         if (col_name != "Time") {

           # Create a temporary data frame that contains just the time and
current col
           d_loop <- d[, c("Time", col_name)]

         # Do the background correction.
          # Background correction option 1: subtract the minimum value in a
column
          #                                 from all measurements in that column
              min_value <- min(d_loop[, col_name])
           d_loop[, col_name] <- d_loop[, col_name] - min_value

           # Now, call Growthcurver to calculate the metrics using
SummarizeGrowth
           gc_fit <- SummarizeGrowth(data_t = d_loop[, "Time"],
                                     data_n = d_loop[, col_name],
                                   t_trim = trim_at_time,
                                    bg_correct = "none")

           # Now, add the metrics from this column to the next row (n) in the
         # output data frame, and increment the row counter (n)
          d_gc$sample[n] <- col_name
         d_gc[n, 2:9] <- c(gc_fit$vals$k,
                       gc_fit$vals$n0,
                       gc_fit$vals$r,
                       gc_fit$vals$t_mid,
                       gc_fit$vals$t_gen,
                       gc_fit$vals$auc_l,
                       gc_fit$vals$auc_e,
                       gc_fit$vals$sigma)
```

```
        n <- n + 1

        # Finally, plot the raw data and the fitted curve
       # Here, I'll just print some of the data points to keep the file size
smaller
        n_obs <- length(gc_fit$data$t)
       idx_to_plot <- 1:20 / 20 * n_obs
        plot(gc_fit$data$t[idx_to_plot], gc_fit$data$N[idx_to_plot],
            pch = 20,
             xlim = c(0, trim_at_time),
            ylim = c(0, y_lim_max),
             cex = 0.6, xaxt = "n", yaxt = "n")
         text(x = trim_at_time / 4, y = y_lim_max, labels = col_name, pos = 1)
       }
       }
}


#png('Plate1_with_controls.png')
#data_with_controls(plate1)
#dev.off()



#####GET AVERAGE VALUES FOR TECHNICAL REPLICATES
y = melt(plate1, id=colnames(plate1)[1])
all.plate1 = merge(y, sample.info, by.x='variable', by.='Cell')

df_mean.plate1 = all.plate1 %>% group_by(Replicate, Treatment, Time) %>%
summarise_at(vars(value), list(Average=mean))

mean.sd.plate1 =  df_mean.plate1  %>% group_by(Treatment, Time) %>%
summarise_at(vars(Average), list(mean=mean, sd=sd))

#####PLOT BIOLOGICAL REPEATS WITH AND WITHOUT ERROR BARS
a = ggplot(mean.sd.plate1, aes(x=Time, y=mean, color=Treatment)) +
geom_smooth(method='gam', se=F) + geom_errorbar(aes(ymin=mean-sd, ymax=mean+sd),
linewidth=0.5, width=1) + labs(y='Mean optical density at 600 nm', 'Time (min)')
+ scale_x_continuous(breaks = seq(0, 690, by = 30)) +
scale_color_manual(values=c('#bbbbbb', '#aa3377')) + theme_bw()


cfu=read_tsv('CFU.txt', col_names=T)
1
b = ggplot(cfu, aes(x=Time, y=Mean, fill=Treatment)) + geom_col(position='dodge',
width=80) +  geom_errorbar(aes(ymin=Mean-SD, ymax=Mean+SD), linewidth=0.5,
width=5, position=position_dodge(70)) + scale_x_continuous(name='Time (min)',
breaks = seq(0, 630, by = 90)) + scale_fill_manual(values=c('#bbbbbb',
'#aa3377')) + theme_bw() + ylab('Log10 of mean colony-forming units per
millilitre of L4-FAA5 culture') + coord_cartesian(ylim=c(4,8.5))

###AUC_E (empirical area under the curve data for biological replicates)
gc_fit_full_plate1 %>% filter(note != "")
#no errors identified

#Maximum possible population size in a particular environment, or the carrying
capacity, is given by K.
#Intrinsic growth rate of the population, r, is the growth rate that would occur
if there were no restrictions imposed on total population size.
#t_gen = doubling time (fastest possible doubling time)
#t_mid = point at which population density reaches 1/2K (which occurs at the
inflection point
#auc_e = empirical area under the curve


library(effectsize) #v0.8.5
library(broom) #v1.0.5
library(report) #v0.5.7
```

```
library(flextable) #v0.9.2
library(rempsyc) #v0.1.4


t_gen = gc_fit_full_plate1 %>% group_by(Replicate, Treatment) %>%
summarise_at(vars(t_gen), list(Average=mean))
t_gen.p = nice_t_test(data=t_gen, response='Average', group='Treatment')$p
A = ggboxplot(t_gen, x='Treatment', y='Average', add='jitter', fill='Treatment')
+ stat_compare_means(method='t.test')+ ylab('Mean doubling time (min)') +
scale_fill_manual(values=c('#bbbbbb', '#aa3377')) + theme_bw()


auc_e = gc_fit_full_plate1 %>% group_by(Replicate, Treatment) %>%
summarise_at(vars(auc_e), list(Average=mean))
auc_e.p = nice_t_test(data=auc_e, response='Average', group='Treatment')$p
B = ggboxplot(auc_e, x='Treatment', y='Average', add='jitter', fill='Treatment')
+ stat_compare_means(method='t.test')+ ylab('Mean empirical area under the curve
(relative units)') + scale_fill_manual(values=c('#bbbbbb', '#aa3377')) +
theme_bw()

ggarrange(a, b, ncol=1, nrow=2)
ggarrange(A, B, ncol=2, nrow=1)


t_gen.summary = t_gen %>% group_by(Treatment) %>% summarise_at(vars(Average),
list(Mean=mean, SD=sd))
t_gen.summary
#  Treatment  Mean    SD
#  <chr>     <dbl> <dbl>
#1 Control    61.6  3.78
#2 TMAO       45.4  1.93

auc_e.summary = auc_e %>% group_by(Treatment) %>% summarise_at(vars(Average),
list(Mean=mean, SD=sd))
#  Treatment  Mean    SD
#  <chr>     <dbl> <dbl>
#1 Control    87.1  1.99
#2 TMAO      121.   2.91

cfu.counts = read_tsv('cfu_counts.txt', col_names=T)
cfu.log10 = data.frame(Time=cfu.counts[,1], log10(cfu.counts[,2:7]))

df.cfu = data.frame(matrix(NA, ncol=2, nrow=8))
for (i in 1:nrow(cfu.log10)) {
      df.cfu[i,1]=cfu.log10[i,1]
      df.cfu[i,2] = round(t.test(cfu.log10[i,2:4], cfu.log10[i,5:7])$p.value, 3)
}
colnames(df.cfu) = c('Time (min)', 'p value')

#  Time (min) p value
#1          0   1.000
#2         90   0.015
#3        180   0.007
#4        270   0.037
#5        360   0.008
#6        450   0.081
#7        540   0.064
#8        630   0.015
```

Appendix C – CheckV information for phage genomes

contig_id	contig_length	gene_count	viral_genes	checkv_quality	miuvig_quality	completeness	completeness_method
KLPN5	49851	78	72	Complete	High-quality	100	DTR (high-confidence)
KLPN6	51570	83	70	Complete	High-quality	100	DTR (high-confidence)
KLPN7	49712	76	70	Complete	High-quality	100	DTR (high-confidence)
KLPN8	48785	78	69	Complete	High-quality	100	DTR (high-confidence)
NC_043469.1	51543	81	68	High-quality	High-quality	100	AAI-based (high-confidence)
MW042802.1	51633	82	70	Complete	High-quality	100	DTR (high-confidence)
NC_049844.1	43094	73	61	Medium-quality	Genome-fragment	86.71	AAI-based (high-confidence)
MZ221764.1	49159	75	67	High-quality	High-quality	98.89	AAI-based (high-confidence)
NC_049837.1	51780	80	67	High-quality	High-quality	100	AAI-based (high-confidence)
MW672037.1	50040	74	69	High-quality	High-quality	100	AAI-based (high-confidence)
NC_049836.1	51487	78	67	High-quality	High-quality	100	AAI-based (high-confidence)
NC_049835.1	54438	87	70	High-quality	High-quality	100	AAI-based (high-confidence)
MZ398242.1	49552	75	67	High-quality	High-quality	99.68	AAI-based (high-confidence)
NC_049834.1	52866	83	70	High-quality	High-quality	100	AAI-based (high-confidence)
NC_049843.1	37655	43	41	Medium-quality	Genome-fragment	75.11	AAI-based (high-confidence)
NC_049838.1	47844	77	70	High-quality	High-quality	96.24	AAI-based (high-confidence)
NC_049842.1	49316	77	69	High-quality	High-quality	99.2	AAI-based (high-confidence)
NC_049848.1	49835	77	71	Complete	High-quality	100	DTR (high-confidence)
MW417503.1	50107	78	65	Complete	High-quality	100	DTR (high-confidence)
MT894005.1	50346	80	70	High-quality	High-quality	100	AAI-based (high-confidence)
MZ598515.1	50675	80	70	High-quality	High-quality	100	AAI-based (high-confidence)
MW021764.1	51895	84	71	High-quality	High-quality	100	AAI-based (high-confidence)
MN395285.1	49472	77	70	Complete	High-quality	100	DTR (high-confidence)
NC_049845.1	49477	78	69	Complete	High-quality	100	DTR (high-confidence)
NC_055956.1	49276	76	68	High-quality	High-quality	99.14	AAI-based (high-confidence)
NC_049833.1	50713	79	67	High-quality	High-quality	100	AAI-based (high-confidence)
NC_049846.1	49045	76	68	High-quality	High-quality	98.67	AAI-based (high-confidence)
NC_049847.1	49916	78	71	High-quality	High-quality	100	AAI-based (high-confidence)
NC_049841.1	49935	76	67	High-quality	High-quality	100	AAI-based (high-confidence)
MT701592.1	51775	82	69	High-quality	High-quality	100	AAI-based (high-confidence)
NC_049839.1	50241	77	70	High-quality	High-quality	100	AAI-based (high-confidence)
LR757892.1	49131	76	70	Complete	High-quality	100	DTR (high-confidence)
MN013078.1	51562	80	68	High-quality	High-quality	100	AAI-based (high-confidence)
MN013087.1	51678	81	68	High-quality	High-quality	100	AAI-based (high-confidence)
MW722081.1	51632	81	71	High-quality	High-quality	100	AAI-based (high-confidence)

contig_id	contig_length	gene_count	viral_genes	checkv_quality	miuvig_quality	completeness	completeness_method
KLPN2	51264	81	68	Complete	High-quality	100	DTR (high-confidence)
KLPN3	47277	71	62	Complete	High-quality	100	DTR (high-confidence)
KLPN4	50522	79	68	Complete	High-quality	100	DTR (high-confidence)
MZ428227.1	48935	76	69	High-quality	High-quality	98.43	AAI-based (high-confidence)
MZ428228.1	48826	76	70	High-quality	High-quality	98.22	AAI-based (high-confidence)
MZ428222.1	45558	64	60	High-quality	High-quality	91.65	AAI-based (high-confidence)
MZ428223.1	49636	77	71	High-quality	High-quality	99.84	AAI-based (high-confidence)
MZ428224.1	51554	82	70	High-quality	High-quality	100	AAI-based (high-confidence)
MZ428225.1	49684	79	70	High-quality	High-quality	99.92	AAI-based (high-confidence)
MZ428226.1	48933	76	70	High-quality	High-quality	98.43	AAI-based (high-confidence)
MZ428221.1	51784	85	68	High-quality	High-quality	100	AAI-based (high-confidence)
GVDv1_Zuo_2017_SRR5677819_NODE_6_length_55276_cov_92.170424	55276	81	70	High-quality	High-quality	100	AAI-based (high-confidence)
uvig_574762	50405	79	71	High-quality	High-quality	100	AAI-based (high-confidence)
uvig_574399	47603	70	63	Complete	High-quality	100	DTR (high-confidence)
uvig_536741	49871	77	69	Complete	High-quality	100	DTR (high-confidence)
uvig_535962	49873	77	69	Complete	High-quality	100	DTR (high-confidence)
uvig_474523	16180	35	29	Low-quality	Genome-fragment	32.53	AAI-based (high-confidence)
uvig_467799	49631	75	64	Complete	High-quality	100	DTR (high-confidence)
uvig_464779	33319	48	39	Medium-quality	Genome-fragment	66.96	AAI-based (high-confidence)
uvig_437383	49366	72	66	High-quality	High-quality	99.28	AAI-based (high-confidence)
uvig_394929	51153	78	67	Complete	High-quality	100	DTR (high-confidence)
uvig_376089	49147	74	69	Complete	High-quality	100	DTR (high-confidence)
uvig_369684	49299	75	67	High-quality	High-quality	99.15	AAI-based (high-confidence)
uvig_354241	51621	80	68	High-quality	High-quality	100	AAI-based (high-confidence)
uvig_348444	51620	80	69	High-quality	High-quality	100	AAI-based (high-confidence)
uvig_347013	51606	80	69	High-quality	High-quality	100	AAI-based (high-confidence)
uvig_346479	14491	36	31	Low-quality	Genome-fragment	29.15	AAI-based (high-confidence)
uvig_340901	12377	13	13	Low-quality	Genome-fragment	24.89	AAI-based (high-confidence)
uvig_338855	14984	30	24	Low-quality	Genome-fragment	30.14	AAI-based (high-confidence)
uvig_334913	15190	16	15	Low-quality	Genome-fragment	30.56	AAI-based (high-confidence)
uvig_334911	34594	59	56	Medium-quality	Genome-fragment	69.6	AAI-based (high-confidence)
uvig_331247	51311	84	68	High-quality	High-quality	100	AAI-based (high-confidence)
uvig_330395	50715	81	69	Complete	High-quality	100	DTR (high-confidence)
uvig_329390	48979	74	68	High-quality	High-quality	98.45	AAI-based (high-confidence)
uvig_328591	47402	72	60	Complete	High-quality	100	DTR (high-confidence)
uvig_327471	50314	79	66	Complete	High-quality	100	DTR (high-confidence)
uvig_326277	50626	82	69	Complete	High-quality	100	DTR (high-confidence)
uvig_323103	28109	36	33	Medium-quality	Genome-fragment	56.54	AAI-based (high-confidence)
uvig_314355	49173	75	68	Complete	High-quality	100	DTR (high-confidence)

contig_id	contig_length	gene_count	viral_genes	checkv_quality	miuvig_quality	completeness	completeness_method
uvig_311634	17528	29	27	Low-quality	Genome-fragment	35.26	AAI-based (high-confidence)
uvig_293010	25170	40	35	Medium-quality	Genome-fragment	50.59	AAI-based (high-confidence)
uvig_288643	51565	77	67	Complete	High-quality	100	DTR (high-confidence)
uvig_288431	48102	73	63	Complete	High-quality	100	DTR (high-confidence)
uvig_287240	49273	78	72	Complete	High-quality	100	DTR (high-confidence)
uvig_285149	48646	76	69	High-quality	High-quality	97.84	AAI-based (high-confidence)
uvig_284377	50720	81	70	Complete	High-quality	100	DTR (high-confidence)
uvig_283917	51650	85	67	High-quality	High-quality	100	AAI-based (high-confidence)
uvig_279208	49250	76	67	Complete	High-quality	100	DTR (high-confidence)
uvig_278768	48961	74	68	Complete	High-quality	100	DTR (high-confidence)
uvig_255004	50328	79	67	Complete	High-quality	100	DTR (high-confidence)
uvig_243694	50577	81	69	Complete	High-quality	100	DTR (high-confidence)
uvig_239791	51119	77	70	Complete	High-quality	100	DTR (high-confidence)
uvig_234015	48858	74	66	Complete	High-quality	100	DTR (high-confidence)
uvig_227178	40153	65	56	Medium-quality	Genome-fragment	80.77	AAI-based (high-confidence)
uvig_224277	49371	76	66	Complete	High-quality	100	DTR (high-confidence)
uvig_223847	38875	62	52	Medium-quality	Genome-fragment	78.13	AAI-based (high-confidence)
uvig_223573	49633	77	67	Complete	High-quality	100	DTR (high-confidence)
uvig_219619	47173	67	63	Complete	High-quality	100	DTR (high-confidence)
uvig_215036	50916	79	70	High-quality	High-quality	100	AAI-based (high-confidence)
uvig_145376	49469	76	68	Complete	High-quality	100	DTR (high-confidence)
uvig_141073	27257	42	37	Medium-quality	Genome-fragment	54.82	AAI-based (high-confidence)
uvig_132550	49636	76	69	Complete	High-quality	100	DTR (high-confidence)
uvig_130754	48360	76	66	Complete	High-quality	100	DTR (high-confidence)
uvig_63387	10230	19	18	Low-quality	Genome-fragment	20.56	AAI-based (high-confidence)
uvig_63295	14106	32	26	Low-quality	Genome-fragment	28.36	AAI-based (high-confidence)
SAMN10080877_a1_ct19236_vs1	30632	62	53	Medium-quality	Genome-fragment	61.62	AAI-based (high-confidence)
SAMN05826713_a1_ct12717_vs1	19730	42	37	Low-quality	Genome-fragment	39.69	AAI-based (high-confidence)
SAMN05826713_a1_ct6131_vs1	12094	16	15	Low-quality	Genome-fragment	24.25	AAI-based (high-confidence)
SAMN00792055_a1_ct11403	50660	81	69	High-quality	High-quality	100	AAI-based (high-confidence)
SAMEA2737768_a1_ct34917	49581	75	69	High-quality	High-quality	99.71	AAI-based (high-confidence)
SAMEA2737751_a1_ct5309	50382	71	65	High-quality	High-quality	100	AAI-based (high-confidence)