High Throughput Genomic Analysis of *Helicobacter pylori* Within-host Diversity

Daniel J. Wilkinson

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

September 2019

I dedicate this thesis to my family – to them I owe everything. Although others may not have always believed in me, my family always have without any doubt. They gave me the confidence and freedom to pursue my dreams. I will never be able to repay them for all they have done, instead, I simply hope to make them proud.

Abstract

Helicobacter pylori is a globally significant human pathogen and is the causative agent of a wide range of diverse gastroduodenal diseases such as gastritis, peptic ulceration and gastric adenocarcinoma. Antimicrobial resistance is a growing problem, with *H. pylori* recently listed as one of the top ten antibiotic resistant pathogens of global concern by the World Health Organisation. This species has been shown to be a globally diverse pathogen expressing large genetic variation, even within geographically clustered sub populations. Furthermore, individuals infected with *H. pylori* are thought to harbour unique and diverse populations of quasispecies, but diversity between and within different niches of the human stomach and the process of bacterial adaptation to, and infection persistence within each niche are not yet well understood.

This study utilises whole genome deep population and single colony sequencing to quantify and characterise the within- and between-niche genetic diversity of *H. pylori* populations from paired antrum and corpus biopsies from the stomachs of individual patients. This revealed extensive genetic diversity both within and between different niches of the same stomach. Subsets of highly variable genes including outer membrane proteins, restriction modification systems, DNA repair, chemotaxis and virulence associated genes were observed.

In addition, this study investigated the between niche (antrum versus corpus) antimicrobial resistance profiles of individual patients. The within and between niche (antrum and corpus) diversity of two sequential datasets were also investigated and results from the same patient before and after failed eradication therapy were compared. For one sequential dataset there was a big increase in *H. pylori* allelic diversity both within and between niches of the patient's stomach approximately five months after failed eradication therapy.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Amx	Amoxicillin
BAM	Binary sequence Alignment Map
BSAC	British Society for Antimicrobial Chemotherapy
CDS	Coding Sequence
Clr	Clarithromycin
DNA	Deoxyribonucleic Acid
ELISA	Enzyme-Linked Immunosorbent Assay
EUCAST	European Committee on Antimicrobial Susceptibility Testing
gDNA	genomic Deoxyribonucleic Acid
GWAS	Genome Wide Association Study
IGR	Intergenic Region
Lvx	Levofloxacin
MAF	Minor Allele Frequency
MALT	Mucosa Associated Lymphoid Tissue
MHA	Mueller-Hinton Agar
MIC	Minimum Inhibitory Concentration
MLST	Multilocus Sequence Typing
Mtz	Metronidazole
OMP	Outer Membrane Protein
PAI	Pathogenicity Island
PCR	Polymerase Chain Reaction
PPI	Proton Pump Inhibitor
R-M	Restriction-Modification
Rif	Rifampicin
SAM	Sequence Alignment Map
SNP	Single Nucleotide Polymorphism
ssDNA	single stranded Deoxyribonucleic Acid
Tet	Tetracycline
VCF	Variant Call Format
WHO	World Health Organization

1. Chapter One: Introduction

1.1. Important highlights

Helicobacter pylori is a Gram negative, helical shaped, fastidious microorganism. The natural reservoir of *H. pylori* is the human stomach. Infection has occurred at least since anatomical modern humans migrated from Africa over 58,000 years ago (Linz et al., 2007).

This co-evolution has continued to the current day with approximately 50% of the global population infected by *H. pylori* with almost all infected individuals presenting some degree of gastritis (Kodaman et al., 2014). Other clinically important diseases are further attributed to *H. pylori* infection that include but are not limited to peptic ulceration and adenocarcinoma.

Helicobacter pylori was categorised as a class 1 carcinogen in 1994 by the International Agency for Research on Cancer, making this the first and presently the only known bacterial carcinogen (Møller, Heseltine and Vainio, 1995). In 2018, stomach cancer was second only to lung cancer in total cancer deaths worldwide (Ferlay et al., 2018).

The World Health Organization (WHO) listed clarithromycin-resistant *H. pylori* in their list of high priority organisms for which there is an urgent need for antibiotic research and development (World Health Organization, 2017).

Extreme genetic diversity as a result of a high mutation and recombination rate, immune evasion and development costs have hampered vaccine developments to date.

1.2. History of *Helicobacter pylori* discovery (and rediscovery)

The history of *Helicobacter* species and in particular the discovery of *H. pylori,* has been a complicated series of discovery and re-discovery. Among which, observations and hypotheses were made that were often written off by the scientific community that today are well recognised attributes of *H. pylori* infection. Furthermore, while *Helicobacter*-like organisms had been described prior to 1983, it was not until the work of Warren and Marshall (1983), that the stomach was disregarded as a sterile organ. This coupled with contradictory experiments over the past century, the incorrect notion that gastric

acidity was the cause of gastric ulcers, and the difficulties in culturing *H. pylori* largely contributed to the complex history of *H. pylori* discovery and its role in human disease.

Perhaps the first observation of *Helicobacter*- like organisms was by Bottcher and Letulle in 1875 who reported bacteria within gastric glands and gastric ulcers of animals (Kidd and Modlin, 1998; Bottcher, 1875). Astonishingly, Bottcher and Letulle are thought to have been the first to hypothesise that gastric ulcers were caused by these bacteria, however this was not widely accepted within the scientific community at the time (Kidd and Modlin, 1998).

The next notable discovery came in 1893 when 'spirochetes' were described to be colonising the gastric glands of canines (Bizzozero, 1893; Marshall, 2001). These *Helicobacter*- like organisms were most likely to be one or a combination of *Helicobacter bizzozeronii*, *H. felis*, *H. salomonis*, *H. heilmannii* and/or *H. canis* species (Van den Bulck et al., 2005; Hänninen et al., 1996; Wilcock, 2013; Canejo-Teixeira et al., 2014; Prachasilpchai et al., 2007). The work by Bizzozero was taken further by Salomon who ground up gastric epithelium from dogs infected with *Helicobacter*- like organisms and showed that these could infect mice (Salomon, 1896; Marshall, 2001). This represented the first time *Helicobacter*- like organisms were experimentally transferred into an uninfected host and to this day, mouse models are still used to study *H. pylori* pathogenesis. The work by Salomon was repeated a full 24 years later, in 1920, by Kasai and Kobayashi reproducing the results of the original experiment (Kasai and Kobayashi, 1919).

In the early 20th century, Krienitz identified three types of spirochetes by microscopic evaluation of the gastric contents belonging to a patient with gastric adenocarcinoma (Krienitz, 1906). However, no link was described between the presence of the bacteria and the gastric adenocarcinoma of the patient.

The first high prevalence of *Helicobacter*- like organisms were recorded in rhesus macaque monkeys from all gastric mucosa samples in a study by Doenges (1938). Additionally, Doenges investigated gastric samples taken from human autopsies in the USA and found that 43% of the humans sampled were infected by *Helicobacter*- like organisms (Kidd and Modlin, 1998). Freedberg and Baron followed up on these reports by investigating gastric specimens from patients who had undergone partial resection surgery and found that 40% of their patients were infected with spirochetes (Kidd and

Modlin, 1998; Marshall, 2001; Freedberg and Barron, 1940). However, it is worth noting that Freedberg and Barron did not associate an etiopathologic role of this infection due to histological inconsistencies and difficulty in identifying and culturing the organisms present (Kidd and Modlin, 1998). Despite these two confirmatory studies of the presence of spirochetes in human gastric specimens, an investigation by Palmer (1954), 14 years later of over 1,100 human gastric biopsies identified no *Helicobacter*-like organisms. Palmer concluded that the previous studies had isolated *Helicobacter*-like organisms as a result of post-mortem colonisation from cross contamination of the oral microbiota. With greater than 50% of the population thought to be infected at the time of the study, it is not known how such a result was attained (Marshall, 2001). However, a different staining technique was used in comparison to the earlier studies. It is argued that this study alone may have set back research into the human gastric microbiota by around 30 years (Kidd and Modlin, 1998).

In 1967, Ito published the first electron microscope picture of *Helicobacter*- like organisms colonising within a parietal cell gland (Ito, 1967). Further electron microscopy images of *Helicobacter*- like organisms were published in the years following this study (Lockard and Boler, 1970; Steer and Colin-Jones, 1975). However, it was the work by Steer and Colin-Jones (1975), that further noted adherence of *Helicobacter*- like organisms to the gastric epithelium and imaged the phagocytosis of these organisms. Furthermore, these authors suggested that white blood cells were actively recruited to these colonised areas of the gastric epithelium. However, despite these observations the authors were unable to culture the organisms present, as with all other studies previous.

It was not until the work of Warren and Marshall (1983), that the significance of 'unidentified curved bacilli' were realised in the role of human gastritis. By realising the similarities between these organisms and *Campylobacter* spp. they were the first to successfully culture *H. pylori* from human gastric samples. However, it must be noted that Warren and Marshall's first publication did not came without controversy with an initial submission emphatically rejected by the Australian Gastroenterology Association (Kidd and Modlin, 1998).

In order to complete Koch's postulates for the 'pyloric *Campylobacter*' and confirm that this organism was a disease causing agent, Marshall orally ingested a bacterial culture

isolated from a patient in 1985 (Marshall et al., 1985). Infection and gastritis were subsequently confirmed by gastric endoscopy approximately four weeks later.

Despite *Helicobacter*- like organisms being observed prior to the work of Warren and Marshall (1983), these were largely just that – observations with little to no etiology. However, their work not only associated infection with gastritis and peptic ulcers but went on to successfully culture and prove that *H. pylori* was responsible for disease. This was a paradigm-shifting discovery and, in 2005, Warren and Marshall were presented with the Nobel Prize in Physiology or Medicine for their work (Pincock, 2005).

To complete the complex history of *H. pylori*, it wasn't until the advent of 16S rRNA gene sequencing that the taxonomy of this bacteria was resolved to *Helicobacter pylori* and has previously been referred to as *Campylobacter pyloridis*, *Campylobacter pylori* and simply by the observed shape under microscopic examinations.

1.3. Transmission

The transmission routes of *H. pylori* have not been fully elucidated and are often debated. However, it is clear that infection with *H. pylori* is highly prevalent across the world with over 50% of the global population thought to be infected (Hooi et al., 2017). Therefore, it would not be unreasonable to suggest that there are multiple transmission routes leading to infection.

It has been reported that up to 90% of adults are infected in developing countries while less than 40% are infected in more developed countries (Leja, Axon and Brenner, 2016; Kayali et al., 2018). This deviation has been associated with differences in socioeconomic status, overcrowded living conditions, sanitation and hygiene (Bardhan, 1997; Cheng et al., 2009; Ahmed et al., 2007; Breckan et al., 2016).

Infection is thought to occur during early childhood with infection persisting life-long in the absence of eradication therapy.

Perhaps the most prevalent mode of transmission is person-to-person, particularly between close family members (Rothenbacher et al., 1999; Didelot et al., 2013; Krebes et al., 2014). Some studies have found that mother to child transmission was more

prevalent than father to child transmission (Osaki et al., 2015; Mamishi et al., 2016). Sibling to sibling transmission has also been identified between siblings of close age (≤4 years), with transmission from the older sibling most frequent (Goodman and Correa, 2000; Kivi et al., 2003). Spousal transmission has been reported, however, this is thought to be relatively rare (Linz et al., 2013; Gisbert et al., 2002; Kivi et al., 2003). Environmental sources of infection are also thought to play a role in transmission of *H. pylori* from a range of sources including food and water (Zamani et al., 2017; Goodman et al., 1996; Aziz, Khalifa and Sharaf, 2015).

Person-to-person transmission routes include gastro-oral (transmission via gastric juice such as vomiting in early childhood), oral-oral (through saliva) and faecal-oral (poor hygiene practices) (Kayali et al., 2018).

1.4. Colonisation

Helicobacter pylori has a specific tropism for gastric mucosa and epithelium of primates. The majority of the *H. pylori* load are contained within the gastric mucosa but also colonise the underlying epithelial cells. Colonisation of the antrum is thought to occur primarily and is often the most predominant colonisation niche, but colonisation can spread to or even become dominant in the corpus. Corpus predominant colonisation is associated with long term proton pump inhibitor (PPI) use, predisposed by the reduction of acid secretion by parietal cells (Mukaisho et al., 2014). Corpus predominant gastritis is more associated with the formation of gastric ulcers (Kusters, van Vliet and Kuipers, 2006). Pangastritis can also develop and encompass both the antrum and corpus stomach regions. Colonisation of *H. pylori* and the different local effects are depicted in figure 1.1.

A recent study by Ailloud *et al.*, (2019), revealed that intra stomach migration occurs within the stomach and is more frequent between niches of similar epithelia such as the oxyntic corpus and fundus. However, the ability to colonise as a lifelong infection is not fully understood, especially in the face of gastric mucosa flow and replacement, acidity and the host immune system.

Here, essential *H. pylori* colonisation attributes are discussed.

Figure 1.1 *Helicobacter pylori* colonisation and typical pathology

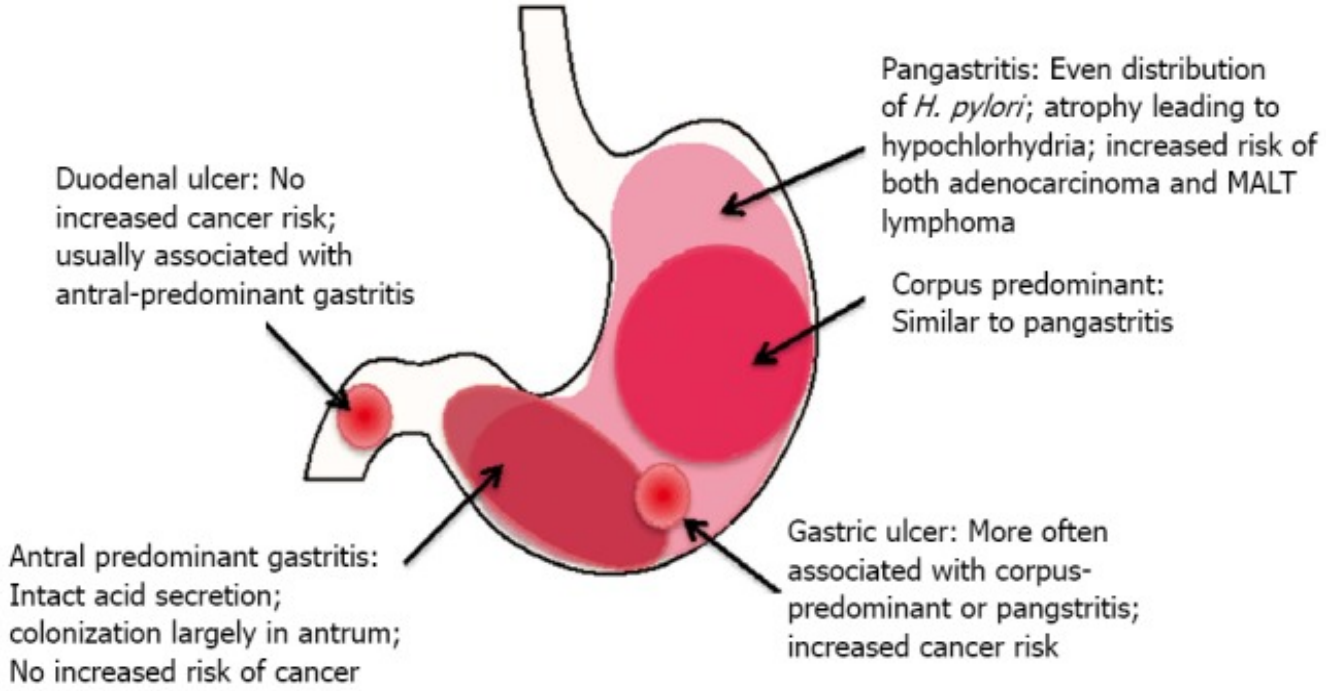


Figure adopted from Testerman and Morris (2014) depicting *H. pylori* colonisation and typical pathology with associated risk factors of gastric cancer.

1.4.1. Adhesion

Outer membrane proteins (OMPs) play a predominant role in bacterial adhesion to the gastric epithelium. *Helicobacter pylori* harbours a comparatively large number of OMPs comprising around 64 genes (Alm et al., 2000a). These are made up of five main OMP families, namely the Hop, Hof, Hom, iron-regulated OMPs and efflux pump OMPs.

While there are many characterised OMPs, perhaps the best characterised include; adherence-associated lipoprotein (AlpA and AlpB), blood group antigen binding adhesin (BabA), outer membrane inflammatory protein (OipA), sialic acid binding adhesin (SabA), HomB and HopZ (as reviewed by Oleastro and Ménard, 2013).

Outer membrane proteins also play a role in outer membrane vesicle formation and the OMP AlpB has been shown to impact biofilm formation and adherence to AGS cells (Yonezawa et al., 2017).

The lipopolysaccharide also contributes to the adhesion of *H. pylori* to the gastric epithelium, particularly through Lewis x mimicry in the O antigen side-chain (Sheu et al., 2007; Edwards et al., 2002).

The high number of OMP protein families and the extensive number OMP genes consisting within them and the LPS play an important role in gastric epithelial attachment and persistence of infection (preventing clearance by mucus flow).

1.4.2. Urease, motility and chemotaxis

Gastric juice within the stomach provides a harsh environment and barrier to bacterial colonisation, particularly due to hydrochloric acid and bile salts. In order to survive passage into the stomach, *H. pylori* must first survive the highly acidic conditions. Survival is facilitated by the secretion of the urease enzyme and pH-gated urea transport mechanisms. The pH-gated urea transport mechanism channels gastric urea to the cytoplasmic urease enzyme that hydrolyses urea into ammonia and carbon dioxide (McNulty et al., 2013; Weeks et al., 2000). This ultimately creates a more neutral micro-environment around *H. pylori* cells.

As a result of urease activity, the viscosity of the surrounding mucus is also reduced, facilitating the motility of the bacteria (Huang et al., 2016a).

The motility of *H. pylori* is mediated through 4-8 unipolar flagella (Gu, 2017). Colonisation and persistence of infection have been shown to be drastically reduced by non-motile mutants (Eaton, Morgan and Krakowka, 1992; Eaton et al., 1996; Foynes et al., 1999; Kim et al., 1999). Another study has linked increased *H. pylori* colonisation density with strains that are highly motile (Kao et al., 2012). These studies suggest that motility is essential to gastric colonisation and persistence of *H. pylori.*

Motility appears to be directional, with chemotaxis playing a major role. Four chemoreceptors have been identified within *H. pylori*, namely Tlps (TlpA, TlpB, TlpC, and TlpD), a CheA kinase, a CheY response regulator and additional paralogous coupling proteins (Abedrabbo et al., 2017).

Many studies have used chemotaxis gene mutagenesis to show the inability of *H. pylori* strains to colonise animal models (Howitt et al., 2011; Foynes et al., 2000; Terry et al., 2005; McGee et al., 2005), demonstrating that chemotaxis plays an essential role in gastric colonisation.

Chemotactic and mechanistic details of chemotaxis systems of *H. pylori* have not been fully elucidated. However, chemoattractants such as urea, arginine, bicarbonate and host-derived molecules as a result of host-cell injury have been identified to date (Cerda, Rivas and Toledo, 2003; Nakamura et al., 1998; Mizote, Yoshiyama and Nakazawa, 1997; Aihara et al., 2014). Chemorepellents have also been identified such as low pH and energy depleted associated environments relating to bacterial metabolism (Croxen et al., 2006; Schweinitzer et al., 2008).

Chemotaxis is also thought to play a role in gastric localisation within the stomach. A study by Rolig *et al.* (2012), found that *H. pylori* are attracted to specific niches within the stomach facilitated by unique chemotactic signals. They also found that localisation to the corpus was required by chemotaxis, but subsequent proliferation was not dependant on chemotaxis. Proliferation was dependent on chemotaxis in the antrum but localisation by chemotaxis was not. This suggests that nutrients are not limiting in the corpus in comparison to the antrum and that localisation might be determined via

chemotaxis to specific niches and not just towards the gastric mucosa, epithelium and glands.

Chemotaxis is thought to change during acute and chronic infection. Chemotaxis towards the antrum is thought to occur during acute infection, after which chemotaxis mutant strains are thought to shift towards corpus colonisation (Rolig et al., 2012; Johnson and Ottemann, 2018). Furthermore, chemotaxis mutant strains have been shown to colonise the gastric epithelium less abundantly and might play a role in inflammation (Williams et al., 2007).

1.4.3. Helical structure

The cork-screw helical shape of *H. pylori* is thought to assist movement into and through the thick gastric mucus layer. Mutants with loss of the helical cell shape have attenuated stomach colonisation and reduced motility in gel-like media mimicking the gastric mucus layer (Sycuro et al., 2012).

1.4.4. Antrum and corpus niche differences and *Helicobacter pylori* colonisation

As previously mentioned (section 1.4), *H. pylori* colonisation is thought to first initiate within the antrum. Antrum colonisation can persist life-long, however, in some cases colonisation can progress to other stomach niches such as the oxyntic corpus or fundus. In this case, an antrum or a corpus predominant colonisation often develops during the course of infection.

The antrum and corpus are specific niches within the stomach due to specific differences between these environments. The antrum possesses gastric glands that do not contain acid secreting cells but instead harbour somatostatin (D cells) and gastrin (G cells) and mucus producing cells (figure 1.2). Corpus glands contain parietal cells (acid secreting), chief cells (produce pepsinogen), mucus cells and enteroendocrine cells (D cells, G cells and enterochromaffin-like cells that release histamine, serotonin and atrial natriuretic peptide). Furthermore, the glands within the corpus niche are generally deeper than those in the antrum (Park and Kim, 2015; Fung et al., 2019).

The differences in gland cell makeup and pH between the antrum and corpus are reasonably well understood. However, the compositional matrix and concentration of

nutrients within specific niches of the stomach are not fully elucidated (Keilberg and Ottemann, 2016). Furthermore, it is not fully understood what promotes or limits *H. pylori* proliferation within the stomach. Studies have reported specific *H. pylori* strain differences between isolates taken from the antrum and corpus relating to virulence, antibiotic resistance and colonisation (Seo et al., 2019; Arévalo-Jaimes et al., 2019; Carroll et al., 2004; Rolig et al., 2012). While specific strain differences have been reported, the full extent of between niche *H. pylori* population diversity has yet to be studied in detail.

Figure 1.2 Diagram of the human stomach and niche specific glandular structure

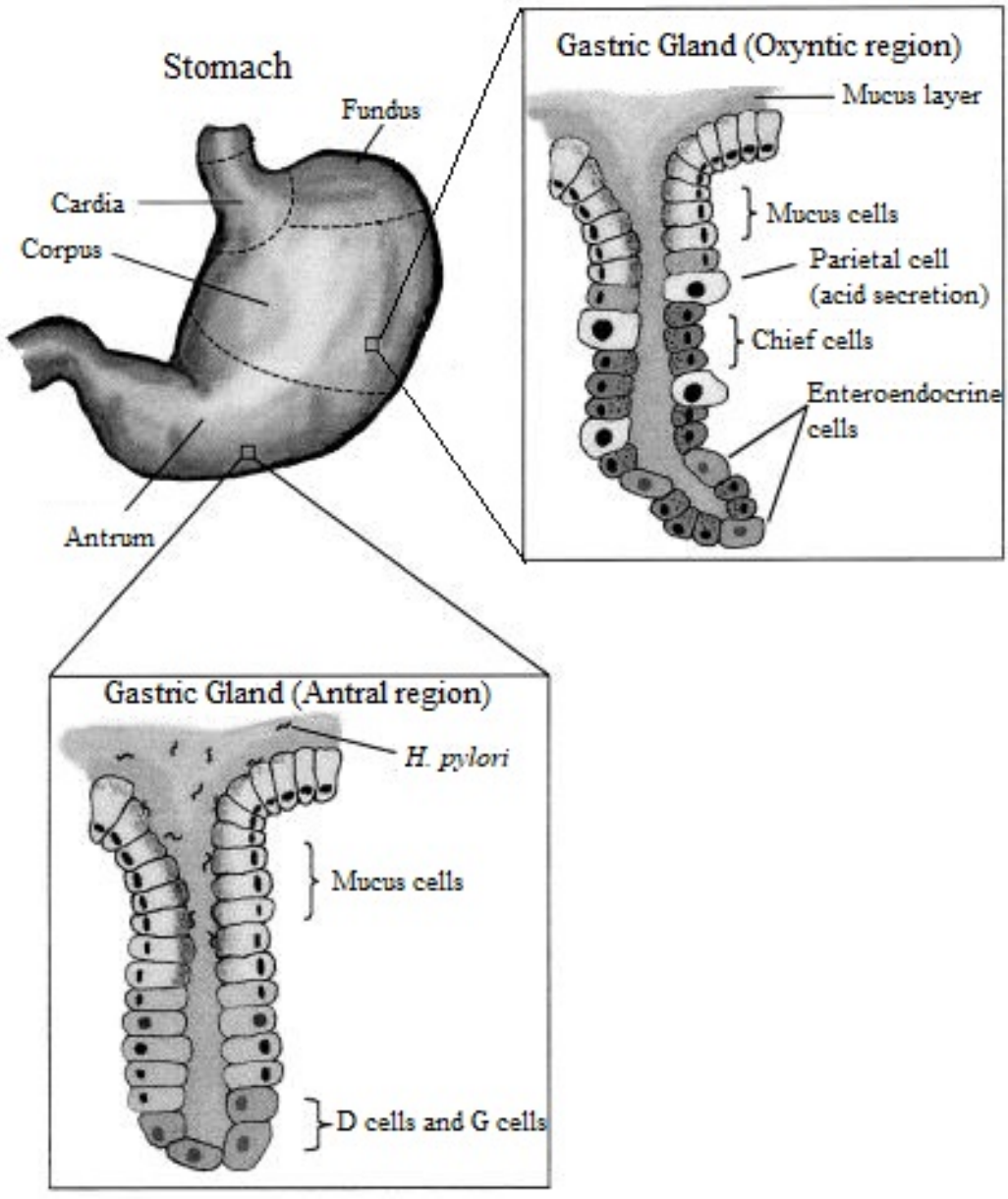


Figure adopted and adapted from Testerman, McGee and Mobley (2001). Schematic of the human stomach and niche specific glandular structure and cell composition.

1.4.5. Immune evasion

Despite the strong adaptive and innate immune responses, *H. pylori* is able to persist as a chronic, lifelong infection of the stomach. *Helicobacter pylori* is able to escape, disrupt and manipulate the host immune system facilitating survival and long-term infection. Virulence factors also play an important role in immune evasion. Immune evasion by *H. pylori* is expertly reviewed by Mejías-Luque and Gerhard (2017) and Karkhah et al., (2019).

Helicobacter pylori avoids or dampens down the innate immune system through a variety of ways. Detection by Toll-like receptors is reduced due to host cell antigen mimicry by OMPs, as previously mentioned. Furthermore, the LPS and flagella are less immunogenic than most other bacteria due to structural modifications (Stead et al., 2008; Mejías-Luque and Gerhard, 2017). Phase variation of antigens are also thought to aid in the evasion and modulation of the host immune response (Bergman et al., 2006).

In the face of high inflammation and phagocyte chemotaxis, *H. pylori* survive killing by reactive oxygen species by producing catalase and superoxide dismutase (Odenbreit, Wieland and Haas, 1996; Spiegelhalder et al., 1993). A recent study by Lekmeechai *et al.* (2018), have shown that outer membrane vesicles play an important role in reactive oxygen species neutralisation and thus immune evasion.

Immune clearance of *H. pylori* by phagocytosis has been shown to be independent of *vacA* and *cag*PAI negative mutants (Rittig et al., 2003). However, others suggest that *vacA* and *cag*PAI status play an important role in phagosome survival (Ramarao et al., 2000; Ramarao and Meyer, 2001; Zheng and Jones, 2003).

1.5. Toxins

1.5.1. The *cag* pathogenicity associated island and the role of cytotoxin-associated gene A (CagA)

The *cag* pathogenicity associated island (*cag*PAI) is either present or absent in *H. pylori* strains. When present, the island is either complete (consisting of ~30 genes) or

incomplete (<30 genes) (Nguyen et al., 2010). Some of these genes encode for a type IV secretion system that ultimately translocates the effector protein CagA into the gastric epithelial cells (Odenbreit et al., 2000).

Once translocated, CagA undergoes phosphorylation-dependent and phosphorylation-independent cell signalling modulation by interacting with over 25 host cellular binding partners (Backert and Tegtmeyer, 2017). These interactions effect the host cell in numerous ways including proliferation, changes in the cytoskeleton, formation of pedestals, and stimulation of IL-8 (Kao, Sheu and Wu, 2016). CagA has also been shown to downregulate B7-H2 expression which allows *H. pylori* to evade Th17-mediated clearance (Lina et al., 2013).

1.5.2. Vacuolating cytotoxin A (VacA)

Inflammation, induced in part by CagA, results in the recruitment of white blood cells to the site of *H. pylori* colonisation, as previously mentioned. VacA is an excreted pore forming toxin and is translocated into host cells by endocytosis (Karkhah et al., 2019b). This toxin has many host cell interactions such as cell damage by cellular vacuolation, cell proliferation, apoptosis and induction of IL-2 secretion by T cells (Amieva et al., 2003; Cover et al., 2003; Cover and Blaser, 1992).

VacA is also involved in immune evasion such as the prevention of phagosome maturation (Zheng and Jones, 2003). The interruption of the calcineurin signalling pathway by VacA has been shown to block the proliferation of T cells and down regulates IL-2 transcription (Gebert et al., 2003). Furthermore, VacA is able to bind to mitochondria resulting in cell apoptosis and the proliferation of T cells (Talebi Bezmin Abadi, 2017).

1.6. Prevalence

The global prevalence of *H. pylori* infection is around 50% but varies across the world (figure 1.3) (Hooi et al., 2017).

The prevalence of infection in the United kingdom was last reported by Vyse *et al.* (2002), revealing infection rates to be around 13%. However, it must be noted that this

was determined using serum serology from people aged between 1 and 84 years. Therefore, the prevalence of active infection might be different to that reported due to serological testing methodology that is not able to discriminate between active and current infection. Furthermore, the testing of children of such a young age might mask the true infection rate as infection in these patients might develop later on in childhood.

Figure 1.3 Global prevalence of *Helicobacter pylori*

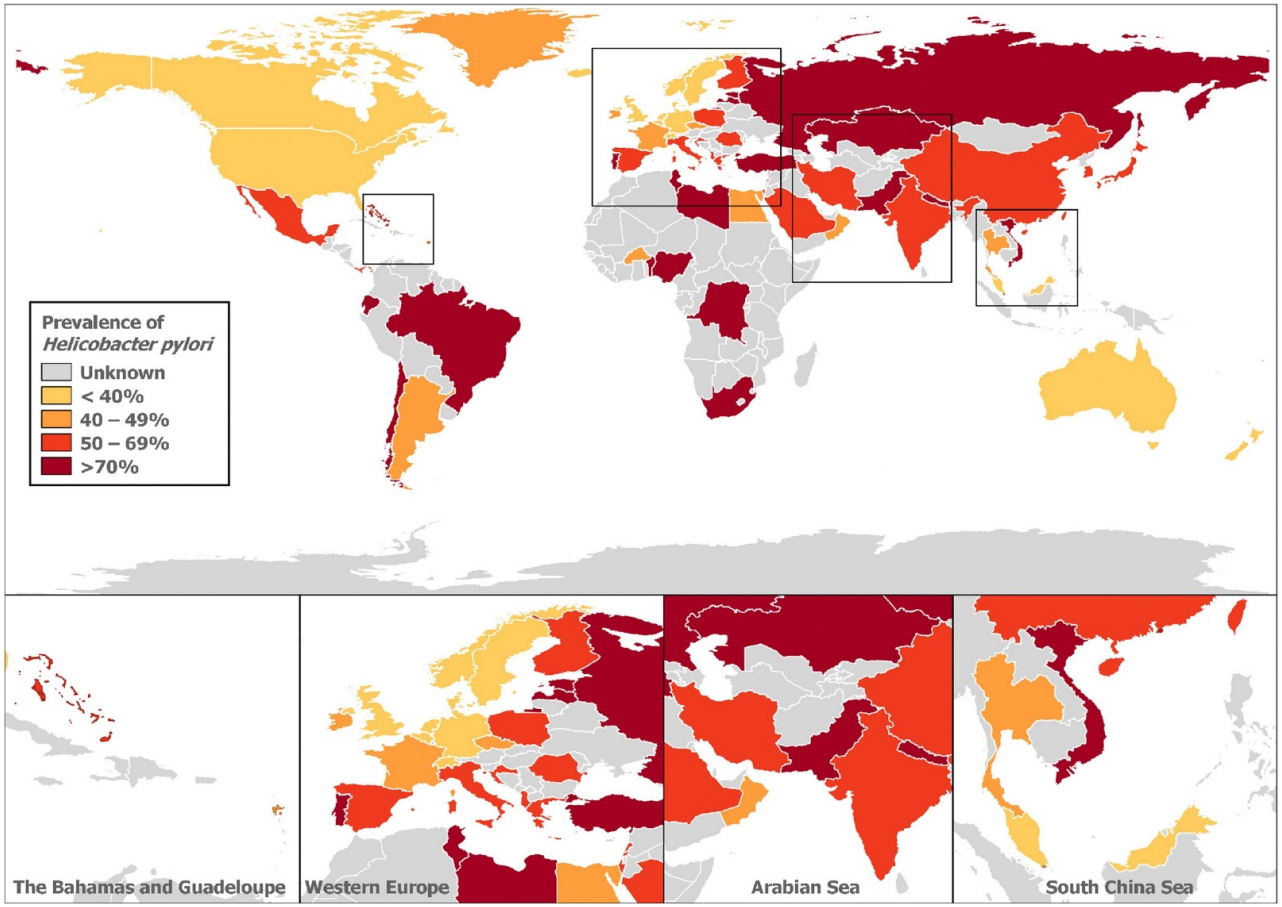


Figure adopted from Hooi *et al*., (2017). Global overview of *H. pylori* prevalence.

1.7. Clinical manifestations

Clinical symptoms of *H. pylori* infection include; dyspepsia and acid reflux that has not responded to antacids and lifestyle changes, abdomen pain, nausea, loss of appetite and weight loss.

1.7.1. Diseases caused

1.7.1.1. Gastritis

All *H. pylori* infected individuals develop some degree of gastritis, but most remain asymptomatic. Gastritis can be diagnosed by symptoms but often corelates poorly with histological or endoscopic gastritis (Sugano et al., 2015). Gastritis is therefore best diagnosed by gastric endoscopy (described below).

Gastritis is often a result of inflammation, intestinal metaplasia and atrophy with *H. pylori* being the leading cause of gastritis (Sugano et al., 2015).

Gastritis is often reported as three different types; pangastritis, antrum predominant, and corpus predominant (figure 1.1).

1.7.1.2. Ulcer disease

Helicobacter pylori is an etiological agent of both duodenal and peptic ulcer disease (Serin et al., 2015).

Different virulence factors and combinations thereof can increase the risk of ulcer disease. These virulence factors include; *cagA*, *vacA* s1/m1 genotype, *dupA*, *iceA1*, *oipA* and *babA* as recently reviewed by Chang, Yeh and Sheu (2018).

1.7.1.3. Gastric adenocarcinoma

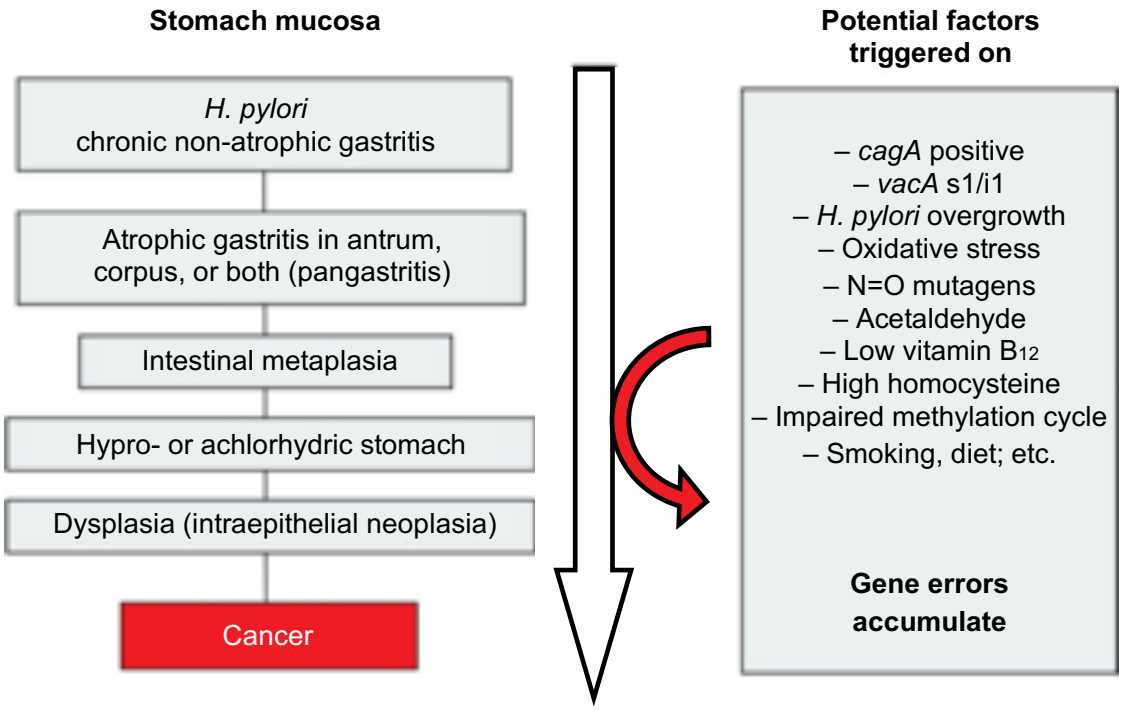
Infection with *H. pylori* is the strongest known risk factor for gastric adenocarcinoma (Correa and Piazuelo, 2011). Gastric cancer is multifactorial with *H. pylori* infection and the action of bacterial virulence factors, disease precursors and progressions, host

genetics and environmental factors thought to play a role (Correa and Piazuelo, 2011). Furthermore, gastric cancer is suspected to be a multistep process over the period of infection.

The human model of gastric carcinogenesis has been updated multiple times, but *H. pylori* is an etiological factor in each progression to adenocarcinoma (Correa, 1988b; Correa and Piazuelo, 2012). The first stage in the multistep cascade to gastric carcinogenesis is *H. pylori* infection and gastric inflammation (figure 1.4). The following step is to non-atrophic gastritis, potentially exasperated by *H. pylori* specific virulence factors (figure 1.4). Progression of atrophic gastritis may develop further as a result of *H. pylori* virulence factors, environmental factors and host genetic factors (figure 1.4). Atrophic gastritis is antecedent to intestinal metaplasia. Intestinal metaplasia can be slowed by the eradication of *H. pylori* infection, but healing can take many years or may not heal at all (Zullo et al., 2012; Walker, 2003). Gastric lesions associated with atrophic gastritis and intestinal metaplasia are the last step in the cascade to carcinogenesis (Correa and Piazuelo, 2012; Park and Kim, 2015).

Helicobacter pylori virulence factors such as *vacA* s1/i1 isoforms and *cagA* status are important risk factors of gastric cancer (Winter et al., 2014; Park and Kim, 2015). A recent genome wide association study (GWAS) by Berthenet *et al.* (2018) identified a number of individual gene presences, paired gene presences and specific gene polymorphisms that increased gastric cancer risk. This study identified both virulence and none virulence associated genes that increased risk of gastric cancer.

Figure 1.4 Gastric precancerous cascade



Gastric precancerous cascade adopted and adapted from Sipponen and Maaroos (2015).

1.7.1.4. Mucosa associated lymphoid tissue lymphoma

As with the development of gastric cancer, development of mucosa associated lymphoid tissue (MALT) lymphoma is thought to be a multistep progressive process starting from *H. pylori* infection and chronic gastritis (Farinha and Gascoyne, 2005).

Physiologically, the stomach does not contain lymphoid tissue. However, chronic infection with *H. pylori* and persistent gastric inflammation results in the formation of MALT in the gastric mucosa. Specifically, *H. pylori* infection results in the recruitment of B lymphocytes, T lymphocytes and neutrophils to the site of infection. B cells proliferate due to reactive T cells, cytokines and activation of the CD40 pathway by *H. pylori.* Continuous stimulation and prolonged growth of B cells coupled with the presence of reactive oxygen and nitrogen species results in the acquisition of genetic anomalies and precedes MALT lymphoma (Sagaert, 2016).

In approximately 75% of cases, MALT lymphoma regresses after *H. pylori* eradication suggesting that *H. pylori* infection is required to maintain malignancy (Hatakeyama, 2019).

1.7.1.5. Extragastric diseases

While *H. pylori* infection is localised to the gastric niche, there is a growing body of research into *H. pylori* associated diseases far from this primary site of infection. These have recently been reviewed by three recent publications (Ražuka-Ebela, Giupponi and Franceschi, 2018; Gravina et al., 2018; Franceschi, Covino and Roubaud Baudron, 2019).

In brief, *H. pylori* infection has been associated with extragastric diseases including; neurological diseases, cardiovascular diseases, pulmonary diseases, haematological diseases, autoimmune diseases, kidney diseases, metabolic syndromes, hepatobiliary diseases, pancreatic disease, inflammatory bowel disease, colorectal cancer, dermatologic diseases and obstetrical conditions (Franceschi, Covino and Roubaud Baudron, 2019).

1.8. Diagnosis

Patients presenting symptoms of a *H. pylori* infection such as dyspepsia and/or acid reflux that has not responded to antacids and lifestyle changes, usually follow a 'test and treat' strategy in the UK.

Diagnostic tests include both invasive and non-invasive tests. Patients are often diagnosed by non-invasive techniques due to the high sensitivity and specificity of some of these tests and the increased comfort these types of tests afford the patient undertaking them. However, non-invasive tests do not allow for the culture of the *H. pylori* infection and thus offer little insight into the active infection.

1.8.1. Non-invasive tests

1.8.1.1. Breath test by Isotope labelled urea

Patients are often diagnosed by non-invasive techniques such as the heavy isotope-labelled ^{13}C-urea or ^{14}C-urea breath test that takes advantage of the *H. pylori* induced enzymatic breakdown of the ^{13}C-urea/^{14}C-urea to ammonia and carbon dioxide by excreted urease. Enrichment of $^{13/14}CO_2$ expelled in the patient's breath is an indication of a current *H. pylori* infection. The sensitivity and specificity of this method are high, at 96% and 93% respectively (Ferwana et al., 2015). However, false negative *H. pylori* infection results are increased if the patient has not stopped PPI drug administration two weeks prior to the urea breath test, due to reduced urease activity as a consequence of higher gastric pH caused by these drugs (Graham et al., 2003).

1.8.1.2. Stool antigen test

The stool antigen test comprises of two versions, an enzyme immunoassay and an immunochromatography based test. Both have high sensitivity and specificity in identifying *H. pylori* infection, but the enzyme immunoassay has been shown to be the superior of the two methods with test sensitivity and specificity as high as 94% and 97% respectively (Gisbert and Pajares, 2004; Shimoyama, 2013). The stool antigen tests are thought to be more affordable than the urea breath, but patients have a much higher preference for the urea breath test than for the stool antigen test (Shimoyama, 2013; Cullen et al., 2002; McNulty and Whiting, 2007).

1.8.1.3. Serology

Diagnosis of *H. pylori* infection by the non-invasive blood and urine serology tests are falling out of favour across the world. This is due to the need for a secondary confirmatory test to diagnose a current *H. pylori* infection, such as the urea breath or stool antigen test. Since antibodies can persist long after a *H. pylori* infection has cleared, serological tests are not suitable for eradication confirmation after therapy. Serological based tests are usually performed by the enzyme-linked immunosorbent assay (ELISA) and can target a range of anti- *H. pylori* antibodies. However, the sensitivity and specificity of the blood serology test are low at 50% and 54% respectively (Kazemi et al., 2011). Therefore, it is no surprise that this diagnostic test is not preferred. However, a study by Stratton and Laczek, (2013) has shown that before the blood serology test was revoked as a diagnostic test within their institution, there was a much higher compliance to the *H. pylori* blood serology test as well as a higher total number

of *H. pylori* diagnostic tests being requested. This might indicate a patient preference towards the blood serology test that might improve patient compliance to diagnosis.

1.8.2. Invasive tests

1.8.2.1. Gastric endoscopy

Invasive diagnosis of *H. pylori* infection is carried out on biopsies taken from the stomach and confirmed through a biopsy urease test or through microscopy using the modified Giemsa stain (Rotimi et al., 2000; Chen, Chang and Lee, 1995). Invasive diagnosis by gastric endoscopy is recommended in patients over the age of 55 with recent, unexplained dyspepsia (Gisbert and Calvet, 2013). This is not to say that gastric endoscopy is never undertaken in patients under the age of 55. Patients below this age with persistent dyspepsia not alleviated by PPIs or lifestyle changes, sudden and unexplained weight-loss, persistent abdominal pain or patients with resistant *H. pylori* infection identified by a failed eradication therapy attempt are often referred for gastric endoscopy (Papastergiou, Georgopoulos and Karatapanis, 2014; Gisbert and Calvet, 2013). However, as the gastric cancer rate is low in people under the age of 55 in most Western European countries this age threshold for gastric endoscopy referral is thought to be adequate (Gisbert and Calvet, 2013). Furthermore, the 'test and treat' strategy has been shown to be as efficient as the more costly prompt endoscopy for dyspeptic patients further validating the use of this strategy in the clinical setting (Lassen, Hallas and Schaffalitzky de Muckadell, 2004; Lassen et al., 2000; Talley, 2005).

For these reasons, gastric endoscopy is infrequently carried out on *H. pylori* infected patients who generally follow the 'test and treat' strategy in the UK. However, gastric endoscopy is still the best method of detecting gastric cancers and gastric cancer risk as informed by histological examination of the pinch biopsy samples (Choi et al., 2018; Sakitani et al., 2018). The updated Sydney scoring system for histologically examined biopsies is used to score the environment for normal (0) and high (3) levels of inflammation, activity, intestinal metaplasia and atrophy (Dixon et al., 1996). High levels of intestinal metaplasia with gastric atrophy have been shown to be early markers of gastric cancer (Correa, 1988a; Correa, Piazuelo and Camargo, 2004). However, gastric cancer development is thought to be a multi-step process as a result of indirect inflammatory effects and/or the direct epigenetic effects of *H. pylori* infection as reviewed by Ishaq and Nunn, (2015).

1.9. Eradication therapy

1.9.1. Eradication therapy in England

1.9.1.1. First line treatment regime

For geographic regions with clarithromycin susceptibility > 80-85%, a seven day triple therapy course consisting of amoxicillin with either clarithromycin or metronidazole coupled with a PPI should be prescribed (Sugano et al., 2015; Public Health England, 2017). In case of penicillin allergy clarithromycin coupled with metronidazole and a PPI should be prescribed (Public Health England, 2017).

1.9.1.2. Second line treatment regime

If symptoms persist after first line eradication therapy a second line treatment regime should be prescribed consisting of a seven day course of PPI with amoxicillin and the antibiotic not prescribed from the first line therapy (Public Health England, 2017).

In the case of allergy to penicillin, seven days of PPI coupled with bismuth subsalicylate or tripotassium dicitratobismuthate with tetracycline hydrochloride and metronidazole are prescribed (Public Health England, 2017).

1.9.1.3. Third line treatment regime

Persistent symptoms after first line treatment regime and previous patient exposure to metronidazole and clarithromycin should be prescribed third line treatment.

Third line treatment consists of a PPI with amoxicillin and tetracycline hydrochloride or levofloxacin for a course of seven days (Public Health England, 2017).

In case of amoxicillin allergy, third line therapy consists of the second line therapy plus levofloxacin (if patient not previously exposed to levofloxacin) and metronidazole for seven days (Public Health England, 2017).

1.9.2. Other eradication regimes

1.9.2.1. Bismuth quadruple therapy

Bismuth coupled with a triple therapy consisting of a PPI and two antibiotics has been shown to be an effective eradication regime (Tursi et al., 2017; Fiorini et al., 2017).

However, bismuth quadruple therapy is recommended as a potential rescue regime or in areas with high antibiotic resistance to clarithromycin (Malfertheiner et al., 2017).

1.9.2.2. Sequential and concomitant therapy

Sequential therapy usually consists of a PPI coupled with an antibiotic for a seven day course immediately followed by a second seven day course of a different proton pump inhibitor and antibiotic (O'Morain et al., 2018).

Concomitant therapy consists of a PPI and three antibiotics (O'Morain et al., 2018).

Sequential therapy has been shown to be better at eradicating *H. pylori* infection and concomitant therapy has been shown to be superior to sequential therapy as reviewed by O'Morain *et al.* (2018). However, patient compliance to these regimes is low and no longer recommended (O'Morain et al., 2018).

1.9.2.3. Hybrid therapy

Hybrid therapy usually follows a course of 14 days with the first seven days consisting of a PPI and one – two antibiotic(s) followed by seven days with an additional two antibiotics (Miftahussurur et al., 2017; Ashokkumar et al., 2017).

However, eradication rates were reportedly reduced due to duel antibiotic resistance (O'Morain et al., 2018).

1.9.2.4. Culture-guided treatment

Culture-guided treatment has recently been recommended after a failed second line treatment. However, it is argued that as antibiotic resistance is ever increasing, culture-guided treatment should be used after a failed first line treatment (O'Morain et al., 2018).

Culture-guided treatment can be accomplished by antimicrobial susceptibility testing or resistance genotyping (O'Morain et al., 2018). However, an internationally recognised disk-diffusion methodology is not yet established (Ogata, Gales and Kawakami, 2014).

1.9.2.5. Probiotics

A recent meta-analysis revealed a potentially promising increase in eradication of *H. pylori* by inclusion of probiotics with existing eradication regimes (Si, Lan and Qiao, 2017).

Probiotics have been used in combination with quadruple eradication therapies with one study showing a 16% reduction in antimicrobial associated patient side effects such as dyspepsia, nausea/vomiting, diarrhoea and abdominal pain (Jung et al., 2018).

Probiotics, as a potential strategy against *H. pylori* infection was recently reviewed by Qureshi, Li and Gu, (2019), who report on the efficiency of probiotics as an alternative to *H. pylori* eradication treatment, as an adjunct to *H. pylori* eradication treatment and as a potential vaccine delivery vehicle.

1.10. Antibiotic resistance

Antibiotic resistance is a global One Health concern that threatens effective eradication and prevention of virtually all clinically relevant bacterial infections.

Antibiotic resistance in *H. pylori* is conferred by chromosomal mutations and not through mobile genetic elements. Therefore, spread of resistance phenotypes is thought to occur vertically from the resistant strain to its descendants (Mégraud, 2004). However, *H. pylori* is known to have an extremely high recombination rate suggesting that resistance genotypes could be passed horizontally between strains via homologous recombination (Falush et al., 2001a). Furthermore, *H. pylori* is capable of transformation suggesting that resistance genotypes could be integrated from the DNA of lysed cells (Bubendorfer et al., 2016a).

Resistant phenotypes might develop within a susceptible population during the course of infection due to the extremely high natural mutation rate of *H. pylori* (Falush et al., 2001a). As *H. pylori* can persist as a lifelong infection, antibiotics taken by a patient for an unrelated bacterial infection is unlikely to eradicate the resident *H. pylori* population (see eradication therapy, section 1.9). However, antibiotics taken by the patient throughout life, might contribute to the clearance of susceptible strains, leaving behind the more resistant phenotypes, driving *H. pylori* antimicrobial resistance as a result.

A recent study by Savoldi *et al.* (2018), systemically reviewed the prevalence of all WHO regions and found alarming levels of resistance worldwide. In particular, resistance to clarithromycin, metronidazole, and levofloxacin were ≥15% in almost all WHO regions.

Due to concerning levels of resistance to clarithromycin worldwide, a first line antibiotic in the eradication of *H. pylori* (see eradication therapy, section 1.9), the WHO designated clarithromycin-resistant *H. pylori* as a high priority bacterium for antibiotic research and development in 2017 (World Health Organization, 2017).

Heteroresistant *H. pylori* infections are becoming more common, but there is no standardised protocol for the detection and treatment in the clinical setting (Rizvanov et al., 2019). Furthermore, *H. pylori* strains taken between the antrum and corpus of same patients have presented with varying phenotypic resistances to antibiotics (Kim, Kim and Kwon, 2003; Selgrad et al., 2014). However, mixed *H. pylori* strain infections are difficult to investigate due to fastidious growth requirements hampering the selection of a wide range of single colony isolates to provide a representative sample. Furthermore, the presence of just one resistant strain has the potential to survive eradication therapy and then colonise the host after eradication of the sensitive population.

1.11. Genomics

The first complete genome sequence of *H. pylori* was published in 1997, using a random sequencing approach and dye-terminator sequencing (Tomb et al., 1997). This strain, denoted as strain 26695, was obtained from a patient in UK suffering from

gastritis. This strain was also found to elicit an immune and inflammatory response in animal models and was known to be toxigenic and transformable. The first complete sequence revealed *H. pylori* has a relatively small genome size of 1,667,867 bp in length, has a low GC content of 39% and contained approximately 1590 coding sequences (CDSs). This first complete *H. pylori* genome highlighted potential important genes that might be involved in host-pathogen interactions such as adhesins, lipoproteins and other outer membrane proteins due to their high abundance.

In 1999 a second research group published the second complete genome sequence of *H. pylori* denoted as strain J99 and compared it with 26695 using comparative genomics techniques (Alm et al., 1999). This study revealed a similar genomic organisation in terms of gene order and predicted proteomes, but with extensive genomic and allelic diversity. This study also reported that 7% of genes were unique between the two strained investigated suggesting a diverse accessory genome.

A follow up study a year later by Alm *et al.* (2000b), used comparative genetics to analyse OMP families using *H. pylori* strains 26695 and J99. They further investigated the large number of OMP present between the genomes and identified five paralogous gene families. This revealed a generally high orthologous protein pair identity, but some OMP genes had much lower nucleotide identity. Furthermore, it was found that gene duplication was not uncommon for some OMP related genes.

These early comparative genomics studies started to reveal the astonishing allelic diversity of *H. pylori* whilst also revealing comparatively similar genomic organisation and accessory genome flexibility.

1.11.1. Multilocus sequence typing and phylogeographic clustering

Multilocus sequence typing (MLST) is based on the nucleotide sequences of seven 'housekeeping' genes that groups isolates by homology (Maiden et al., 1998).

The largest MLST database is that of *H. pylori*, but this database is by no mean comprehensive with one study uploading ten genomes for which more than half of the alleles were not present within the database at the time of the study (Larsen et al., 2012). This further highlights the extraordinary genetic diversity of *H. pylori* strains and suggests so called 'housekeeping' genes are much more diverse than those generally

used in the MLST scheme. Such diversity will certainly limit the power of MLST analysis in the short term as more sequences will need to be uploaded and potentially in the long term as such diversity might begin to confuse analysis.

Despite this, MLST analysis was first applied to *H. pylori* strains from different geographical regions, revealing phylogeographic clustering (Achtman et al., 1999). Similar MLST studies have followed including the use of virulence genes in identifying local and global phylogeographical clustering of *H. pylori* strains and even identified past human migration events and an African origin around 58000 years ago (Falush et al., 2003; Linz et al., 2007; Moodley et al., 2009; Wirth et al., 2004; Yamaoka et al., 2002b)

1.11.2. Natural mutation, homologous recombination and natural transformation

Studies investigating the natural mutation rate of *H. pylori* agree on a rate between 1.38 $\times 10^{-5}$ and 0.7 $\times 10^{-6}$ per year per site (Falush et al., 2001a; Linz et al., 2014; Morelli et al., 2010; Didelot et al., 2013; Furuta et al., 2015a; Kennemann et al., 2011). This natural mutation rate is between 10 – 100 times higher than *Escherichia coli* (Dorer, Sessler and Salama, 2011).

The recombination rate of *H. pylori* has been shown to be around 6.9 $\times 10^{-5}$ per initiation site per year (Falush et al., 2001a). However, recombination has been shown to vary between *H. pylori* strains but are thought to introduce up to 100 times more substitutions than natural mutations (Didelot et al., 2013).

Natural transformation of *H. pylori* can result in the uptake of DNA resulting in up to 8% of sequence replacement through multiple transformation cycles (Bubendorfer et al., 2016a). High natural transformation in *H. pylori* introducing mosaic DNA imports is thought to be a driver of allelic diversity (Kulick et al., 2008).

1.11.3. Genetic diversity

Helicobacter pylori has colonised humans for at least 58,000 years and persists as a chronic lifelong infection (Linz et al., 2007). These properties have allowed *H. pylori* to co-evolve with humans with evidence of global, local and individual adaptations, conferring vast genetic diversity.

High natural mutation, homologous recombination and natural transformation rates of *H. pylori* all drive genetic diversity at a global, local and individual patient population level. Natural mutation drives clonal diversity over time while homologous recombination and natural transformation can introduce genetic diversity across sections of DNA by single or multiple events. Therefore, mixed *H. pylori* strain infections are thought to amplify within patient diversity more than a single strain infection.

Studies have highlighted OMP genes and other cell surface-related protein genes, restriction-modification genes, virulence associated genes and signal transduction genes to be especially genetically diverse (Oleastro et al., 2010; Furuta et al., 2015; Pride, Meinersmann and Blaser, 2001). It is thought that some genes may be highly diverse due to specific host adaptation (Dubois et al., 1999).

1.12. Study rationale

Patients are thought to be initially infected by single or multiple strains of *H. pylori* during early childhood. This can persist as a lifelong infection of the host. Considering the exceptionally high natural mutation and recombination rate of *H. pylori* the infecting strain is thought to diversify by natural mutation and homologous recombination that ultimately leads to sub-populations or quasispecies, with a mixed infection accelerating and amplifying this further. This potentially leads to low abundant strains within the larger or dominant population that are harder to isolate which might hold higher virulence and resistance phenotypes. Indeed, there is a large body of evidence that describe this extensive *H. pylori* genetic diversity both within and between patients.

Despite this knowledge, most studies follow a single colony isolation methodology, with all genomics-based studies following this protocol to date. This has potentially resulted in a low-resolution snapshot of the more *common* genetic diversity, which has had considerable impact on the progression of our understating of *H. pylori* infection. However, there are still questions outstanding and questions that have not been fully elucidated such as how *H. pylori* is able to persist as a lifelong chronic infection, why most patients remain asymptomatic while others develop more severe gastritis and disease progressions, what contributes to eradication therapy failure and if different niches within the human stomach select for different *H. pylori* strains.

In this thesis, single colony and population-based methodologies were designed and executed to better understand *H. pylori* genomic and phenotypic diversity within patients by investigating both within and between niche genomic diversity from paired antrum and corpus biopsies. This between and within niche line of study was designed to investigate potential niche specific adaptations due to the differences between the antrum and corpus niches (section 1.4.4).

To investigate the within and between niche (antrum and corpus) antibiotic resistance differences, an inexpensive and reliable disk diffusion based method was developed (Chapter Three).

Populations of clinically isolated *H. pylori* were screened for minor allelic variants using a deep population sequencing technique and read mapping based bioinformatics analysis (Chapter Four). This analysis aimed to elucidate the population level of genetic diversity at a single point in time.

Single colony isolates of *H. pylori* were also isolated from a subset of patient biopsies to allow for further analysis not permitted by deep population sequencing such as phylogenetic reconstructions and recombination detection. Additionally, the single colony isolate analysis was compared to the deep population sequencing dataset and used for data validation (Chapter Five).

Two patients with sequential biopsies taken from before and after failed eradication therapy were investigated by the deep population and single colony sequencing to investigate the effects eradication therapy had on these *H. pylori* populations (Chapter Six).

2. Chapter Two: Shared Materials and Methods

2.1. Sample acquisition, typing, culture, and storage information

Clinical biopsies were taken from patients attending for upper gastrointestinal endoscopy with suspected *H. pylori* infection at the Queens Medical Centre in Nottingham by Professor John Atherton and team at the University of Nottingham. Paired biopsies of the antrum and corpus of 15 patients, along with single biopsies from one region of three patients were taken. The updated Sydney scoring system (Dixon et al., 1996) was used to determine clinical inflammation, activity, atrophy and intestinal metaplasia at the time of endoscopy by Professor Atherton and team. Separate biopsies were taken for histology and isolation of *H. pylori* but were taken from the same region of the stomach.

Initial *H. pylori* culture from pinch biopsy samples was carried out by Professor Atherton and team at the University of Nottingham and not by the author. These pinch biopsies were plated onto blood base #2 agar plates (Oxoid, UK) containing 5% (v/v) horse blood (TCS Biosciences, UK) and incubated for 2-3 days at 37°C under microaerophilic conditions (10% CO_2, 5% O_2, 85% N_2). Single colony isolates and/or sweeping growth from each biopsy was picked and pooled, avoiding bacterial contaminants that might also be present. Picked growth was stored within iso-sensitest medium (Oxoid, UK) containing 15% (v/v) glycerol (Sigma Aldrich, UK) and stored at – 80°C.

The resulting *H. pylori* clinical sweep stocks were then provided to Nottingham Trent University by Professor Atherton and colleagues and sub-cultured by the author (as described later in section 2.2) for use in the studies described herein. Approval for use of clinical isolates was granted by the NHS National Research Ethics Service, Nottingham Research Ethics Committee (Ref: 08/H0408/195) 27th January 2009 (Appendix, figure 10.2.1).

It must be noted that virulence genotyping by PCR for *vacA*, *cagE*, *cagA* and CagA serology by ELISA was conducted and reported by Professor Atherton and colleagues at the University of Nottingham and not by the author. These were carried out as previously described by the research group (Atherton et al., 1995; Peek et al., 1995; Kidd et al., 2001; Reyes-Leon et al., 2007a).

It must be noted that the anonymised patient data presented within this thesis was collected by Professor Atherton and colleagues at the University of Nottingham and not by the author.

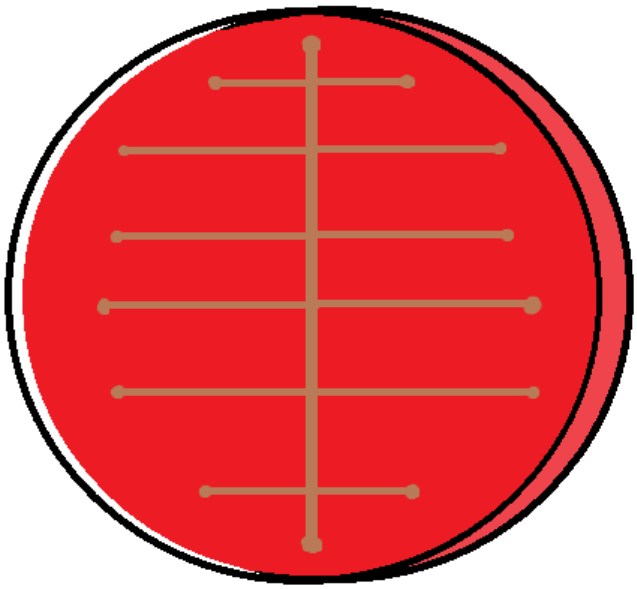
2.2. Culture from frozen stocks

Helicobacter pylori stocks were part thawed by gently warming the top section of the cryo-tube and a 75 μl quantity was immediately pipetted onto a pre-warmed (inoculum free culture plate placed in the 37°C incubator overnight) culture plate made up of blood base #2 agar containing 5% (v/v) horse blood. A pre-warmed culture plate was used to ensure there was no prior contaminant present on the culture plate before inoculation. A sterile inoculation loop/swab was then used to spread this droplet across the culture plate in cross sections (figure 2.1) to maximise the abundance of edge growth. This method also allowed for better determination of contaminants post incubation.

Special care was taken to minimise culture passage cycle numbers of *H. pylori* stocks to avoid laboratory-adaptation as large passage numbers can lead to loss of viability, loss of virulence and other genetic mutations (Leiser et al., 2015; Wirth et al., 1998; Gaillard et al., 2011).

Once inoculated, culture plates were incubated for 48-72 hours under microaerophilic conditions to allow for sufficient growth from frozen stocks and then subcultured onto a fresh culture plate using the culture pattern described below (figure 2.1) for 16-24 hours under microaerophilic conditions. Subsequent cultures were used for the experiments described herein, unless otherwise stated.

Figure 2.1 Culture pattern to maximise recovery of fresh *H. pylori* growth



Typical culture pattern of *H. pylori* from frozen stock. Brown cross sections show how stocks were spread across a blood agar base #2 culture plate supplemented with 7.5% defibrinated horse blood to allow for better contamination detection by manual observation. This method also allowed the researcher to extract more viable *H. pylori* growth by extracting from the edge of these lines.

2.3. Rapid urease test

A rapid urease test was carried out on all stocks obtained from the University of Nottingham in order to determine the presence of *H. pylori* and to help identify stocks that were contaminated.

A 10% weight/volume of urea powder (Sigma Aldrich, UK) was dissolved in sterile distilled water. A 1% phenol red solution (Sigma Aldrich, UK) was prepared in sterile distilled water and 2-3 drops were added to a 10 ml stock of the 10% urea solution to create a working stock for the rapid urease test.

Approximately 100 μl of the working stock of the rapid urease solution was added to the wells of a sterile 96 well microtiter plate (Appleton Woods, UK). Using a 1 μl inoculation loop, a small amount of bacterial growth was collected and used to inoculate a single well containing the rapid urease solution. A sudden colour change within 15

seconds of inoculation from orange to pink indicated the presence of the *H. pylori* urease and thus a *H. pylori* positive culture. Negative samples were discarded.

2.4. DNA extraction

Whole genomic DNA was extracted from clinical sweep populations from all patients described in Chapter Three. Samples selected for whole genome sequencing were cultured as described previously (Chapter Two section 2.2) and gDNA was extracted using the QIAGEN (Netherlands) QIAmp DNA Mini Kit following the manufacturer's instructions with minor alterations. These alterations were; reduction of incubation in lysis buffer AW to less than ten minutes, extension of incubation of proteinase K to 16-20 hours in buffer ATL, additional five minutes incubation at 70°C with 100 mg/ml RNase A (QIAGEN, Netherlands) in buffer AL, use of 4°C 100% (v/v) 200 proof molecular grade ethanol and additional final centrifugation step at 16,000 rpm on a benchtop centrifuge for ten minutes at room temperature. DNA was eluted in 90 μl non-DEPC treated nuclease free distilled water (Thermo Fisher Scientific, USA), after incubation of the extraction column with the nuclease free water at room temperature for a minimum of 30 minutes.

Reducing the incubation time of *H. pylori* culture in buffer AW ensured that proteinase K was promptly added to counter the effect of endonuclease action as *H. pylori* is known to harbour an abundance of restriction-modification (R-M) systems (Bubendorfer et al., 2016b; Xu et al., 2000). Extending proteinase K incubation improved bacterial lysis and protein degradation, resulting in better column flow through, more consistent homogeneity and reduced protein contamination in the final eluate. Pre-cooled ethanol allowed genomic DNA to precipitate out of solution more efficiently and an additional centrifugation step at maximum speed (16,000 rpm) ensured that spin columns were thoroughly dry and free of residual ethanol. Reducing the elution volume while increasing the incubation time increased the resulting DNA eluate concentrations.

During all incubation steps, with the exception of the incubation of nuclease free water prior to elution, samples were vortexed at frequent intervals and where possible, incubated on a shaking dry-heat block to prevent separation and autoagglutination of the bacterial component.

Transfer of samples to the spin columns was performed using broad 1,000 µl tips to avoid mechanical shearing of genomic DNA.

Taken together, these amendments to the standard genomic DNA extraction protocol improved DNA yield and quality.

2.5. DNA Qualification and Quantification

DNA quality was determined using the NanoDrop2000 spectrophometer (ThermoFisher Scientific, USA) using strict absorbance ratio cut-off values. Quality ratios considered acceptable were the A_{260}/A_{280} and A_{260}/A_{230} with a range of 1.8 - 1.9 and 1.9 - 2.2 respectively.

Samples with an irregular spectral curve or distortion were discarded and re-extracted due to suspected contamination.

A minimum sample concentration as determined by the NanoDrop2000 of 30 ng/µl was required with samples found to be below this discarded and DNA extraction repeated. However, this quantification was not relied upon for downstream processing such as sequencing library preparation and was used only as an inclusion criterion. This was due to the inaccuracy of ds-DNA quantification of the NanoDrop2000 platform as the concentration is determined by absorbance at 260 nm which include all nucleic acids present within the sample. Therefore, this reading also includes RNA and ssDNA which will not be sequenced, resulting in an overestimation of the ds-DNA sequencing target within the sample (Nakayama et al., 2016).

Samples with extracted genomic DNA passing the above criteria were frozen at -20°C for long term storage (up to 12 months).

Immediately prior to whole genome sequencing, stock DNA was diluted 1:5 in nuclease free water to create a sequencing stock. This sequencing stock was then accurately quantified for ds-DNA using the ds-DNA high sensitivity (HS) assay kit (ThermoFisher Scientific, USA) on the Qubit fluorometric quantification platform (ThermoFisher Scientific, USA). In contrast to the NanoDrop2000, the Qubit detects the fluorescence of PicoGreen which is a fluorochrome that specifically binds ds-DNA, detected by a

significant increase in fluorescence with very little background interference (Ahn, Costa and Rettig Emanuel, 1996). Sequencing stocks were diluted to the desired starting library preparation concentration of 0.3 ng/μl. To improve the detection accuracy of the Qubit, the input genomic DNA was loaded at 10 μl in 190 μl of loading dye.

2.6. Population whole genome sequencing considerations

Herein, deep sequencing is defined as an average nucleotide coverage depth of >100.

In order to obtain high sequencing coverage depth at each nucleotide base within the target genome, the number of samples loaded onto each run was significantly reduced so that the available sequencing output was dedicated to these samples. This allowed for greater resolution at each nucleotide base position within the genome permitting the detection of allelic variation observed within the sequenced bacterial population.

To this end, the MiSeq V3 reagent kit (Illumina, USA) was chosen over the V2 chemistry as the difference in sequencing output was 6.5 Gb with a difference in read output of 10 million reads. This permitted more samples to be loaded onto a single sequencing run while keeping the desired average sequencing coverage of >100 per sample. The expected sequencing depth per sample was calculated by utilising the Lander and Waterman (1988) equation as follows:

$$\text{Coverage per nucleotide sequenced} = \frac{\left(\frac{Read\ Length\ X\ Total\ read\ output\ of\ cartridge}{Target\ genome\ length}\right)}{number\ of\ samples}$$

Although this modified Lander and Waterman (1988) equation assumes a number of fixed parameters that are not achievable in actual sequencing operation, it acts as an estimation of sequencing coverage. These assumptions include; fixed fragment lengths, uniform fragment selection, stable target genome lengths and a fixed cartridge read output (Evans, Hower and Pachter, 2010). Therefore, to counter variations from actual sequencing operation and to more confidently obtain an average coverage of >100, we targeted a sequencing coverage of >200 using this equation. For these reasons, a maximum of 18 samples were used as input on any single deep sequencing run. Working can be seen here for the MiSeq V3 chemistry:

Coverage per nucleotide sequenced = $\frac{\frac{\left((250)\ \mathrm{X}\ (2.5\ \mathrm{X}\ 10^{7})\right)}{(1.68\ \mathrm{X}\ 10^{6})}}{\mathbf{18}}$ = 206.68

2.7. Single colony sequencing considerations

For single colony sequencing of *H. pylori* isolates, an average coverage of ≥30x at each nucleotide position within the genome was desired. A coverage of this size allows for more accurate downstream analysis and a better genome assembly. Using the equation described above (section 2.6) a sequencing run sample size of 48 single colony isolates was chosen as this provided a theoretical average coverage of 77.5 bases per nucleotide position on the genome. Again, this theoretical coverage was more than double the desired coverage to account for any variations in actual sequencing operation.

2.8. Whole genome sequencing

All whole genome sequencing runs were conducted inhouse at Nottingham Trent University on the Illumina MiSeq platform (Illumina, USA). Two deep sequencing and two single colony isolate runs were performed by the author while a third single colony sequencing run containing 23 single colony isolates was carried out by Pauline Ogrodzki (fellow postgraduate researcher) following the same methodology.

Following DNA qualification and quantification (section 2.5) and taking the library type into consideration (sections 2.6 – 2.7) the Nextera XT DNA Library Prep Kit Reference Guide (Document # 15031942 v02, Illumina, USA) was consulted and followed with minor adjustments. Briefly, genomic DNA (gDNA) underwent tagmentation using the Nextera XT transposase which fragments the ds-gDNA whilst tagging adapter sequences to the ends of the DNA fragments. Genomic libraries were then indexed using the Nextera XT Index Kit (Illumina, Cambridge, UK) and cleaned using Agencourt AMPure XP beads (Beckman Coulter, USA). Following this, indexed and cleaned libraries were manually normalised to 4 nM using the Qubit method described in section 2.5. In order to account for the molar conversion of ng/μl an average insert/fragment size was determined using the Agilent 2200 TapeStation (Agilent Technologies, USA) and input into the following equation:

$$\text{Molarity (nM)} = \frac{\text{concentration from Qubit (ng/µl)}}{660\ \text{g/mol} \times \text{average insert size}} \times 10^6$$

This normalisation method was preferred over the bead-based normalisation method described within the Nextera XT DNA Library Prep Kit Reference Guide as it allowed for more control over the library normalisation step.

Normalised libraries were pooled and denatured into single stranded DNA (ssDNA) using 0.2 N sodium hydroxide as described in the MiSeq System Denature and Dilute Libraries Guide (Document # 15039740 v01, Illumina, USA). Libraries were diluted to a loading concentration of 20 pM and spiked with 1% PhiX. PhiX was used as a control and to allow for better quality reporting and cluster density quantification by the MiSeq base calling software.

The MiSeq was set up to sequence in lengths of 250 bp in paired end mode. The post processing options were selected for adapter trimming, FastQ generation and demultiplexing into sample directories.

2.9. Quality control of sequencing reads

On completion of each sequencing run, the Illumina BaseSpace (Illumina, USA) online application produced run statistics were consulted to adjudge whether sequencing runs were successful. Statistics considered were; Q score distribution for all reads >Q30 above 80%, smooth intensity plots over all cycles that do not overlap, cluster density close to the recommended density of 1,300 k/mm^2 and near to equal read index distribution.

Although the Illumina software reporter automatically detects and removes sequencing adapters from the 5' of each paired end sequence, it does not always capture all sequencing adapter sequences and does not filter reads based on quality. Therefore, further curation processing steps were conducted on sequencing reads using additional software. Sequencing reads were trimmed for Illumina Nextera XT adapters and read through using Trimmomatic version 0.38 (Bolger, Lohse and Usadel, 2014) in paired-end mode employing both a palindrome and simple read trimming method. In paired-end mode, Trimmomatic maintains mate pair read correspondence, and uses

information obtained from both sequences to better identify and remove sequencing adapters. A 'simple' trimming event occurs when a sufficiently aligned match (by a user defined minimum alignment score) is found between the adapter sequence and the output sequencing reads. A palindrome trim detects adapter read through by ligating adapters to the beginning of mate pairs and their biological sequences are then aligned, if the biological sequence aligns with adapters present within the read of the reverse-complement mate pair a trimming event occurs, leaving only the biological sequence and removing the adapter read through. This is important as these sequences are not part of the target biological sequence. The following command was used:

```
trimmomatic PE -threads {user specified CPU/thread/core number}
-trimlog {user named log output} {path to forward/mate pair 1
reads} {path to reverse/mate pair 2 reads} {user named output
for paired-end trimmed forward/mate pair 1 reads} {user named
output for single-end trimmed forward/mate pair 1 reads} {user
named output for paired-end trimmed reverse/mate pair 2 reads}
{user named output for single-end trimmed reverse/mate pair 2
reads} ILLUMINACLIP:{path to fasta with adapters}:2:20:10
```

This allowed for a seed mismatch of 2 nucleotides so that trimming events were specific for the adapter sequence while still allowing for instances of incorrect base calling. The palindrome trimming threshold was lowered to an alignment score of 20 for paired-end reads to ensure the removal of adapter read through. For single ended reads, an alignment score of 10 was used for the palindrome trimming. A log file was created for troubleshooting purposes.

Further to the removal of Illumina sequencing adapters, resulting reads were trimmed for quality and length using Sickle version 1.33 (Joshi and Fass, 2011). Sickle uses a sliding window approach equal to 0.1 times the length of the read and trims the read when the resulting quality within the window drops below a sequence/base quality threshold (user specified) by identifying where in the window this drop occurs. In paired-end mode, Sickle takes two paired end inputs and outputs two trimmed paired-end sequencing read files. Where reads only pass in the forward or reverse mate pair, one of the mates is dropped and the passed mate is outputted as a 'single' read. The following command was used:

```
sickle pe -f {path to forward/mate pair 1 reads from Trimmomatic output} -r {path to reverse/mate pair 2 reads from Trimmomatic output} -t sanger -o {user named output of trimmed forward/mate pair 1 reads} -p {user named output of trimmed reverse/mate pair 2 reads} -s {user named output of single reads} -q 30 -l 50 -g
```

A quality cut off score of Phred 30 was chosen (*i.e.* base call accuracy >99.9%). This was chosen to improve data quality and enhance the reliability of downstream data processing steps such as variant discovery. In consideration of the possibility of trimming adapter read through, and to maximise the reads processed while removing very short reads, reads less than 50 bases in length were removed from the dataset.

Curated reads were passed to FastQC version 0.11.7 (Andrews, 2010) for confirmation of expected trimming behaviour. In particular, the removal of sequencing adapters, base quality >Phred 30 and length distribution between 50 and 250 bases. Individual reports were generated for each forward and reverse read set and manually inspected for each sample. The following command was used:

```
fastqc {path to input curated reads} --threads {user specified CPU/thread/core number} --outdir {path to output directory}
```

2.10. Contamination detection of sequenced libraries with non-*H. pylori*

Contamination within sequenced libraries was detected by Kraken version 1.0 (Wood and Salzberg, 2014) using the MiniKraken 8GB database which was constructed from complete bacterial, archaeal and viral genomes available within RefSeq as of 18th October 2017. Kraken uses a k-mer based classification method over sequence alignment that maintains high sensitivity while reducing run time and computing memory (Wood and Salzberg, 2014).

The following Kraken command was used:

```
(kraken --threads {user specified CPU/thread/core number} --preload --db {path to MiniKraken database} --fastq-input --gzip-compressed --paired {path to forward/mate pair 1 reads from
```

```
Sickle} {path to reverse/mate pair 2 reads from Sickle} > {Kraken
output}) 2> {user named Kraken log file directed from standard
output} ; kraken-translate --db {path to MiniKraken database}
{path to kraken output file} > {user specified output file for
sample taxonomic lables} ; kraken-report --db {path to MiniKraken
database} {input taxonomic lables file} > {user named output
report file}
```

For a sample library to be classed as contamination free (*i.e.* only contain reads belonging to the *H. pylori* species) > 92% of reads must be in support of *H. pylori* spp. classification with >95% of reads supporting the *Helicobacter* genus. This was to allow for a degree of leniency in relation to the vast genetic diversity observed within the *Helicobacter* genus.

Contaminated samples were retained for downstream analysis as it was important to observe how the methods described within this thesis behave with such datasets. Furthermore, due to the nature of *H. pylori* sampling by endoscopy, it is not always possible to obtain a contaminant free/pure culture of *H. pylori,* despite considerable efforts to do so. Therefore, it was decided that there was utility in retaining contaminated datasets in order to stress-test bioinformatics methodology and to determine if contaminated datasets could be analysed in the context of this project. However, all contaminated datasets were noted and considered at all analysis steps and where included are detailed as such in each section.

2.11. Whole genome assembly of sequenced populations and single colony isolates

Curated sequencing reads (section 8) were assembled using SPAdes version 3.11.1 (Bankevich et al., 2012). SPAdes applies the de Brujjn approach to assemble genomes by utilising varying k-mer strings from sequenced reads. Paired-end data is handled by reverse complementing one of the sequencing mates and assembling both together as a read pair, where resulting graphs show the reverse complement at their edge. SPAdes was invoked as follows:

```
spades.py --careful -t {user specified CPU/thread/core number} --
-pe-1 {path to forward/mate pair 1 reads from Sickle} --pe-2
```

```
{path to reverse/mate pair 2 reads from Sickle} -o {path to
output directory}
```

SPAdes was run in 'careful' mode which informs SPAdes to run the post processing tool 'MismatchCorrector' that better resolves short indels and assembly based mismatches, resulting in improved genome assemblies.

2.12. Curation of assembled genomes

All assembled contigs for each sample were then passed through the quality assessment tool (QUAST) version 5.0.2 (Gurevich et al., 2013) which provides assembly quality based statistics. Statistics used for assembly inclusion were: total number of contigs (<500), number of contigs with more than 1,000 base pairs (<200), N50 (>18,000), total length of all combined contigs (1.68 Mb ± 1.4 kb) and GC content (38.8 % ± 0.5 %). Although the expected genome size of *H. pylori* is ~ 1.68 Mb we allowed for ± 1.4 kb as only high-quality reads (reads with a Phred score ≥30) were used to create assembled genomes, potentially resulting in a reduced genome size in contrast to using all available sequencing reads. This was done to ensure high quality assembled contigs for downstream analysis steps. Furthermore, including lower quality reads from deep sequenced populations of *H. pylori* has the potential to introduce further assembly error as the likelihood of erroneous bases sequenced increases. This coupled with the naturally high mutation and recombination rate of *H. pylori* makes the curation of sequencing reads from population deep sequenced samples essential for further assembly based downstream analysis.

2.13. Whole genome annotation

Whole genome annotation was achieved using two different methods. Where genome annotations were used as part of an analysis pipeline, the associated tool used will also be described.

Whole genome annotation by the rapid prokaryotic genome annotation (PROKKA) tool version 1.13 (Seemann, 2014) was used by running the following command:

```
prokka --cpus {user specified CPU/thread/core number} --compliant
--centre C --locustag {user specified} --kingdom Bacteria --genus
Helicobacter --species pylori --outdir {user specified} --prefix
{user specified} --proteins {path to reference strain 26695
annotation file} --evalue 0.001 {path to assembled genome .fasta}
```

Prokka was run in 'compliant' mode to ensure a Genbank standard file format was produced as an output file. Further taxanomic information was written to the Prokka produced Genbank file in the form of kingdom, genus and species flags. The sequencing centre flag was used to shorten the centre ID that had proven to cause difficulties for some downstream analysis steps. The *H. pylori* reference strain 26695 (NC_000915.1) was used to guide genome annotations and to provide 'HP' gene names as accepted nomenclature in the research field. The 'evalue' of 0.001 was set to increase the similarity e-value cut-off threshold to allow for potential gene diversity.

A second annotation method was also employed for selected analysis pipelines using the rapid annotations using subsystems technology (RAST) online server (Aziz et al., 2008). Assembled genomes were submitted to the RAST server (http://rast.nmpdr.org) and the taxonomy identification number 210 was inputted, relating to the taxonomy string of the *Helicobacter* genus as assigned by the NCBI (Sayers et al., 2009; Benson et al., 2009). The species was manually input as *pylori* and the genetic code 11 (Archaea, most Bacteria, most Virii, and some Mitochondria) was selected. RAST was executed with the 'Classic RAST' annotation scheme, 'FIGfam' release 70, automatically fix errors, build metabolic model, back fill gaps and disable replication parameters. These settings were selected for all whole genome sequences to ensure all resulting annotations were comparable and standardised.

2.14. Single nucleotide polymorphism/variant annotation and location determination (CDS / IGR)

A pipeline was created to determine whether a single nucleotide polymorphism (SNP) was identified within a CDS or intergenic region (IGR) by intersecting SNP(s) on the corresponding fully annotated reference genome of the query dataset. All SNP/variant annotations were called from RAST annotated genomes (Chapter Two, section 2.13) to enable cross sample comparisons.

Briefly, all sequence sets were annotated through RAST (Chapter Two, section 2.13) and gene locations (chromosome number, start position, end position and gene product name) were extracted from GTF (general transfer format) files and input into individual files with four respective columns; CHR, FROM, TO, ANOTATION. These files were compressed and then indexed. Variant call format files containing SNPs from indexed same sample annotated genomes were then annotated through VCFtools version 0.1.16 (Danecek et al., 2011). The following command was used:

```
cat {path to VCF file} | vcf-annotate -a {path to compressed
annotation file for the corresponding query dataset} \
 -d key=INFO,ID=ANN,Number=1,Type=Integer,Description='My custom
annotation' \
 -c CHROM,FROM,TO,INFO/ANN > {path to output directory and user
specified output file name}
```

Resulting output VCF files contained an additional ‘ANN;’ field when the corresponding SNP was located within a CDS. However, if no ‘ANN;’ field was present for a particular SNP then this SNP was located within an IGR and was recorded as such.

3. Chapter Three: Sample selection and antimicrobial resistance

3.1. Introduction

Gastric endoscopy and histological scoring by the updated Sydney system of the antrum and corpus has revealed niche specific differences (Zhang et al., 2005). This could be an indicator of different virulence types of a mixed *H. pylori* infection acting within these environments. Indeed, a study by Carroll *et al.* (2004), identified virulence gene differences between *cagA* 3'-end repeats from paired antrum and corpus isolates taken from the same patients as well as *vacA* signal type differences (s1/s2) from phylogenetically related familial isolates. This intra-patient virulence diversity has also been observed by others, including virulence positive and negative genotypes for the *cag*PAI (Melo-Narváez et al., 2018; Ailloud et al., 2019; Armitano et al., 2013; Peña et al., 2017).

The *vacA* gene encodes a pore-forming toxin that has direct action towards gastric epithelial cells causing vacuoles to form (Cover and Blaser, 1992). However, the VacA toxin is known to be multi-functional and has other effects such as cell apoptosis and T-cell proliferation as well as other effects on different cell types (recently reviewed by Chauhan *et al.*, 2019). The *vacA* gene has been shown to be multi allelic, consisting of multiple genotypes of the signal (s1a-c, s2), intermediate (i1, i2, i3) and middle (m1, m2) region (Chauhan et al., 2019). A study by Winter *et al.* (2014), revealed that *H. pylori* with the more active s1/i1 genotypes of *vacA* induced more severe and extensive metaplasia as well as inflammation than the less active s2/i2 genotype. Therefore, differences in inflammation and intestinal metaplasia between different niches of the human stomach might indicate the presence of a mixed *H. pylori* infection. The different middle region genotypes have been shown to affect colonisation of *H. pylori* highlighting another potential marker of a mixed infection by virulence type (Letley et al., 2003; Ji et al., 2000)

Whilst the *vacA* gene is present in virtually all *H. pylori* strains, the *cag*PAI is either present or absent. Furthermore, strains can harbour an incomplete *cag*PAI which may affect the virulence potential of the *cag*PAI (Shimoyama, 2005; Ahmadzadeh et al., 2015). The *cag*PAI is made up of a compliment of ~30 genes (Olbermann et al., 2010b). The *cag*PAI encodes for a type IV secretion system that ultimately results in the translocation of the effector protein CagA into the gastric epithelial cells (Odenbreit et al., 2000). Translocated CagA is able to bind to ~25 cell signalling factors causing effects ranging from inflammation induction, cell proliferation and cell apoptosis (Knorr

et al., 2019). Due to this oncogenic property, patients harbouring a functional *cag*PAI have been shown to be at higher risk of gastric cancer (Shanks and El-Omar, 2009; Hatakeyama, 2014).

The *cagE* gene has been shown to be an essential core gene of the *cag*PAI, suggesting that the absence of this gene is important in the translocation of *cagA* (Pham et al., 2012; Backert, Tegtmeyer and Fischer, 2015; Olbermann et al., 2010b). Furthermore, it has been shown that strains lacking *cagE* are more likely to harbour an incomplete *cag*PAI with additional missing genes (Markovska et al., 2018). However, the *cagE* gene has been implicated as a potential virulence factor associated with duodenal ulceration, suggesting it might have more than one function (Day et al., 2000).

Antibiotic resistance profiles of *H. pylori* strains isolated between the antrum and corpus of the human stomach have also been identified (Kim, Kim and Kwon, 2003; Selgrad et al., 2014). Both studies concluded that patients harboured heteroresistant infections, most likely from a single infecting strain that has developed more resistant phenotypes as a result of natural mutation during chronic infection. One exception was noted by Selgrad *et al.* (2014), who identified one patient with a suspected multi strain infection. However, this was based on DNA fingerprinting by random amplified polymorphic DNA analysis, thus the true population diversity was not fully elucidated. Nonetheless, this methodology was able to detect one incidence of a mixed infection suggesting this technique could be employed to identify mixed strain *H. pylori* infections from patient biopsies. This is further supported by an early study by Kim, Kim and Kwon (2003) who investigated 220 strain pairs of *H. pylori* taken from antrum and corpus biopsies of 220 different patients and found that 50% of patients held antibiotic resistant strains with heteroresistance from paired biopsy strains found in 38% of patients harbouring antibiotic resistant strains.

Antibiotic resistance is usually associated with specific point mutations within the *H. pylori* chromosome rather than plasmid associated resistance acquisition. It is still possible for *H. pylori* strains to spread antibiotic resistance phenotypes to other susceptible *H. pylori* strains through homologous recombination, cell to cell conjugation-like transfer and natural transformation (Paul et al., 2001; Hua et al., 1998; Oyarzabal, Rad and Backert, 2007; Kao et al., 2014; Hoffman, 1999). However, for this to occur, the sensitive and resistant strains would need to interact within the same environment. Considering the extent of re-infection with a different *H. pylori* strain or a new infection

via adult-adult transmission is not fully understood and is a currently open debate, the opportunities for a naturally antibiotic resistant strain to interact with a sensitive strain of *H. pylori* are perhaps limited (Perry et al., 2006; Stone, 1999; Perez-Perez et al., 1991; Schutze et al., 1995). That being said, mixed *H. pylori* infections are not uncommon and are potentially a result of a multi-strain infection acquired during early childhood, most likely from an adult to child transmission from familial members (Malaty et al., 1998; Breckan et al., 2016; Ben Mansour et al., 2016). This presents an opportunity for different strains presenting varying antibiotic sensitivity to interact, potentially resulting in the spread of antibiotic resistance as adult *H. pylori* eradication failure rates increase.

This thesis focused on investigating *H. pylori* diversity across the stomachs of individual patients.

Due to financial and operational constraints, selection criteria were developed in order to target patients primarily with paired antrum and corpus samples with indicators of a mixed *H. pylori* infection. These indicators included *cagA* and *cagE* status and patient CagA serology status, antibiotic resistance differences between antrum and corpus *H. pylori* populations and histological differences between the two stomach niches.

3.2. Materials and methods

Clinical sweeps of *H. pylori* were obtained from patient pinch biopsies as described in Chapter Two section 2.1.

Helicobacter pylori clinical sweep stocks were provided to Nottingham Trent University by Professor Atherton and colleagues and sub-cultured by the author for use in this thesis as described in Chapter Two sections 2.2 – 2.3.

It must be noted that the anonymised patient data presented here was collected by Professor Atherton and colleagues at the University of Nottingham and not by the author.

It must be further noted that virulence genotyping by PCR for *vacA*, *cagE*, *cagA* and CagA serology by ELISA was conducted and reported by Professor Atherton and

colleagues at the University of Nottingham and not by the author. These were carried out as previously described by the research group (Atherton et al., 1995; Peek et al., 1995; Kidd et al., 2001; Reyes-Leon et al., 2007a).

3.2.1. Sample selection for use in this thesis

The University of Nottingham's *H. pylori* culture collection database was interrogated to identify patients that potentially harboured a mixed *H. pylori* infection. Patients with paired antrum and corpus samples were prioritised to enable between and within niche comparisons. Patients were further identified by one or more of the following criteria:

3.2.1.1. Virulence differences between different niches

Differences between paired antrum and corpus virulence factor genotyping for *cagA* and *cagE* status (presence or absence) were identified and highlighted across tables 3.2A and 3.2B in blue (opposite niche differences). This was an indication of a between niche mixed *H. pylori* infection.

Antrum and corpus *H. pylori* genotyping differences in *vacA* signal (s1/s2), intermediate (i1/i2) and mid (m1/m2) type were identified across tables 3.2A and 3.2B as indicators of between niche *H. pylori* strain differences.

3.2.1.2. Virulence differences within niches

The indicator of within niche mixed *H. pylori* infection was multiple *vacA* signal (s1/s2), intermediate (i1/i2) and mid (m1/m2) types within one niche. Such occurrences were highlighted in orange across tables 3.2A and 3.2B.

CagA serology was also used as an indicator. For example, if the patient was CagA serology positive but their *H. pylori* infection was *cagA* negative by PCR genotyping, this might be an indicator of a mixed infection where the genotyping missed the *H. pylori* strain(s) responsible for CagA delivery. A second example was where the genotyping result revealed a *cagE* negative and a *cagA* positive result for the *H. pylori* infection, but the patient was serology positive for CagA. A *cagE* negative result would indicate a potentially compromised type-4 secretion system, preventing the delivery of the CagA protein (Backert, Tegtmeyer and Fischer, 2015). Therefore, the CagA positive serology

result could indicate *H. pylori* strains with an intact type-4 secretion system held within the wider *H. pylori* infection. Other ambiguities such as a serology negative but a *cagA* or *cagE* and *cagA* positive genotype were also highlighted in orange across tables 3.2A and 3.2B.

3.2.1.3. Histologically determined Sydney score differences between antrum and corpus niches

Histologically determined Sydney scores were used to identify between niche environmental differences within the same patient. Sydney scores that were different between paired samples were highlighted in blue across tables 3.2A and 3.2B.

3.2.1.4. Sequentially isolated cultures

Although most patients in the Nottingham collection had been seen only once, there were two patients that had undergone gastric endoscopy on at least two different occasions. The first endoscopies were taken prior to eradication therapy while the second were post failed eradication therapy.

While only one patient had antrum and corpus cultures available both before and after eradication therapy, both patients were included in this thesis. This is described in more detail in Chapter Six, but these patients were included due to the unique and rare opportunity to investigate differences between *H. pylori* infection before and after failed eradication therapy.

3.2.2. Design of disk diffusion based antimicrobial resistance assay

The British Society for Antimicrobial Chemotherapy (BSAC) was once the recognised authority in the UK for antimicrobial susceptibility breakpoint guidance in the clinical setting. However, BSAC has now moved towards the European Committee on Antimicrobial Susceptibility Testing (EUCAST) antimicrobial breakpoint guidelines in an effort to standardise breakpoints across Europe (Brown, Wootton and Howe, 2016). Neither BSAC nor EUCAST have a disk diffusion based antimicrobial clinical breakpoint guidance for *H. pylori* as they do for many other clinically relevant bacterial species. Instead, they recommend determining the minimum inhibitory concentration (MIC) of *H.*

pylori, to which they advise MIC breakpoints for amoxicillin, clarithromycin, levofloxacin, metronidazole, rifampicin and tetracycline (EUCAST, 2019).

Therefore, a comprehensive search of the literature was undertaken to identify research studies that had employed a disk diffusion based antimicrobial susceptibility assay for use in this study. A combination of the EUCAST/BSAC and studies described within the literature were used to develop the disk diffusion based antimicrobial susceptibility assay described here with a focus on reproducibility and cost effectiveness.

3.2.2.1. Standardisation

Firstly, the reference *H. pylori* strain 60190 (ATCC 49503) was taken from long term storage (-80°C) and plated as described in Chapter Two section 2.2 onto three culture plates to ensure sufficient bacterial load.

Following the EUCAST/BSAC guidelines a McFarland 3 standard was determined using the reference *H. pylori* strain 60190 (ATCC 49503) (http://bsac.org.uk/wp-content/uploads/2014/06/Helicobacter-pylori.pdf). The *H. pylori* 60190 culture was used to inoculate 6 ml of sterilised (autoclave) 0.85% saline (Thermo Fisher Scientific, USA) within 18 mm, round bottomed, clear Pyrex® glass test tubes (wall thickness 1.2 mm ; Sigma Aldrich, UK) and made to a density of 0.582 ±0.008 (600 nm) using a spectrophotometer (model 6300; Jenway, UK). The spectrophotometer was blanked prior to reading using uninoculated 0.85% saline (NaCl). The *H. pylori* 60190 suspension within the test tube was then placed into a densitometer (DEN-1; Grant Instruments, UK). This was repeated in triplicate and a McFarland reading of 2.8 was reported each time. The standard deviation of this instrument was 0.1 for a McFarland 3 standard, as described in the manufacturer's instructions, so this was used as the acceptable deviation during subsequent sampling (Grant Instruments, UK).

A literature search of other research studies using a disk diffusion methodology for antimicrobial resistance typing inferred antibiotic concentrations and breakpoints as displayed in table 3.5. Antibiotic containing disks containing 10 μg amoxicillin, 15 μg clarithromycin, 1 μg levofloxacin, 5 μg metronidazole, 5 μg rifampicin and 30 μg tetracycline was sourced from Oxoid (Basingstoke, UK).

A total of four controls were selected to standardise the antimicrobial resistance assays. These strains were *H. pylori* 60190 (ATCC 49503), *Escherichia coli* DH5α (laboratory strain), *E. coli* 10418 (ATCC 10536) and *Staphylococcus aureus* (ATCC 9144). All control strains of bacteria were cultured as described in Chapter Two section 2.2, with the exception of the non- *H. pylori* cultures which were initially cultured from frozen stocks for 16-24 hours before subculture due to excessive growth over 48-72 hours of incubation compared to the fastidious *H. pylori* culture. Each culture was plated onto three culture plates during sub culturing (Chapter Two section 2.2) to ensure sufficient bacterial load.

The day prior to sub culturing the frozen stocks, Mueller-Hinton agar (MHA) (Oxoid, Basingstoke, UK) plates containing 7.5% defibrinated horse blood were made up and a precise 30 ml was added to each petri dish. Once set, these antimicrobial culture plates were stored at 4°C until use, for a maximum of five days. Prior to culture, these inoculation plates were incubated overnight at 37°C in order to dry and indicate any culture plates that were contaminated.

Each control bacterial culture was made up to a McFarland 2.8 ±0.1 using the densitometer (DEN-1) in 0.85% saline, with frequent light vortexing to keep the bacteria in suspension. Standard inoculums were stored at room temperature under aseptic conditions and used within 15 minutes of creation. Using a single cotton swab (wood-stick; Scientific Laboratory Supplies, UK) and taking care the swab did not touch the outer or inner rim of the test tube, each cotton swab was thoroughly inoculated with the standard inoculum. This swab was taken out (making sure not to touch the inner or outer rim of the test tube – to avoid potential contamination) and plated onto one half of a pre-poured MHA plate containing 7.5% defibrinated horse blood. The swabs were not inoculated more than once. An additional control strain was treated in the same way and used to inoculate the second half of the culture plate, ensuring an approximate 5 mm gap between the inoculums and avoiding cross contamination. This was repeated for five additional culture plate (six total). The six antibiotic containing disks were added to the centre of each plate (usually between the 5 mm gap – typically intersecting each inoculum by ~1 mm), one antibiotic disk per culture plate. This was also repeated for the remaining two bacterial controls (four total control strains).

Inoculated culture plates containing the antibiotic disks were immediately incubated at 37°C under microaerophilic conditions (10% CO_2, 5% O_2, 85% N_2) for five days (120

hours). Zones of inhibition were measured (mm) from the circumference edge of the antibiotic containing disks.

This methodology was repeated independently three times to gauge the reproducibility of this technique and to determine the expected zones of inhibition for these bacterial standards (table 3.1). The controls were used to validate subsequent antimicrobial resistance assays involving the *H. pylori* clinical sweeps used within this study (tables 3.2A-B).

3.2.3. Antimicrobial disk diffusion assay

Antimicrobial resistance assays were conducted on all clinical sweeps used within this thesis (tables 3.2A-B) following the same procedure as that described above. The control strains were cultured alongside each antimicrobial resistance assay performed to validate each particular batch.

Briefly, *H. pylori* clinical sweeps were cultured from frozen stocks and sub-cultured onto three culture plates to ensure sufficient bacterial load (Chapter Two section 2.2).

Post sub-cultured incubation, each clinical sweep was made up to a McFarland of 2.8±0.1 in 0.85% saline using the densitometer (DEN-1). Within 15 minutes, the standard inoculum was used to inoculate half of a 30 ml MHA plate supplemented with 7.5% defibrinated horse blood using a single cotton swab. This was repeated for a total of six MHA plates.

This was repeated for a different clinical sweep, inoculating the opposite half of the six MHA plates.

Six different antibiotic containing filter disks were placed in the centre of each of the six MHA plates inoculated with two different clinical sweeps, one antibiotic disk per plate.

This was repeated in triplicate for each clinical sweep.

Mueller-Hinton agar plates inoculated with the clinical sweeps and antibiotic disks were incubated for five days under microaerophilic conditions.

After incubation, zones of inhibition were recorded in mm from the edge of the antibiotic disks to the edge of the zone of inhibition. Other information was recorded such as the observation of resistant colonies within the zone of inhibition or where two zones of inhibition were observed. If two potential zones of inhibition were observed, the stronger inhibition boundary of the outer zone was recorded as the main population inhibition size. A full table of zones of inhibition and notations on clinical sweeps with resistant colonies and second zones of inhibition can be found in the Appendix (table 10.3.1).

3.3. Results and discussion

3.3.1. Sample selection

The initial target *H. pylori* clinical sweep list was much larger than those presented within this thesis (tables 3.2A-B). This was due to the failure of some cultures to grow or cultures that were contaminated, and so were excluded from the thesis.

While the selection of presumptively mixed patient *H. pylori* samples/cultures from the same patients were identified for use in this thesis, it is possible that other patient samples outside of this selection criteria held a mixed infection and were overlooked. This could highlight a potential selection bias. However, the implementation of this selection criteria helped to identify samples of interest for use in this thesis.

Furthermore, there is a potential selection bias relating to the way in which the *H. pylori* cultures were obtained from pinch biopsy samples. For instance, not all of the *H. pylori* strains within the sample might be culturable from the biopsy sample, resulting in a selection bias towards the culturable strains. One particular example of this could be related to *H. pylori* strains deep within gastric glands that might be difficult to extract whilst being passed over the agar plate (Fung et al., 2019). A second example could be *H. pylori* strains that are within a viable but non-culturable state (Boehnke et al., 2017; Buck and Oliver, 2010). Additionally, by avoiding bacterial contaminants on the culture medium, there is the potential to miss the selection of all strains cultured from the biopsy sample.

Information for the reader on patients and patient samples used within this thesis are presented in tables 3.1 – 3.4.

There were many gaps in the information provided surrounding patient metadata (tables 3.2A-B). However, such information was included to provide further information on the patients enrolled within this thesis.

There were more females (n=11) than males (n=7) enrolled in this thesis (table 3.3). The average age of the participants was 60.39 years (standard deviation = 11.16 years;

minimum age 40 to maximum age 79). The reasons for gastric endoscopy varied greatly but all were symptomatic of *H. pylori* infection (de Jong, Lantinga and Drenth, 2019).

All patients that were taking acid suppressant drugs had ceased use at least two weeks prior to gastric endoscopy (table 3.3). None of the patients had any other relevant drug history, such as antibiotic use before gastric endoscopy.

Table 3.1 Antibiotic zones of inhibition for control strains against six antibiotics

	Antibiotics and zones of inhibition (mm)					
Control strains:	**Amx (10 μg):**	**Clr (15 μg):**	**Lvx (1 μg):**	**Mtz (5 μg):**	**Rif (5 μg):**	**Tet (30 μg):**
***H. pylori* 60190 (ATCC 49503)**	25.33 (SD 2.56)	34.33 (SD 1.76)	15.00 (SD 1.00)	15.33 (SD 2.52)	26.33 (SD 1.53)	24.67 (SD 0.58)
***E. coli* DH5α**	9.00 (SD 0.00)	0.00	11.17 (SD 1.04)	0.00	0.00	12.67 (SD 1.16)
***E. coli* 10418 (ATCC 10536)**	9.00 (SD 1.00)	0.00	12.83 (SD 2.37)	0.00	0.00	12.33 (SD 1.53)
***S. aureus* (ATCC 9144)**	18.67 (SD 2.89)	12.50 (SD 1.32)	10.17 (SD 0.29)	0.00	14.67 (SD 0.29)	15.00 (SD 0.00)

Antibiotic zones of inhibition of control isolates for each antibiotic used within this study. The average zone of inhibition is denoted (from triplicate data) alongside the standard deviation (SD). All zones of inhibition were recorded in mm and measured from the edge of the antibiotic disk. These results were attained from the disk diffusion assay developed for this study as described in section 3.4.2. Antibiotic abbreviations; Amx – amoxicillin, Clr – clarithromycin, Lvx – levofloxacin, Mtz – metronidazole, Rif – rifampicin and Tet – tetracycline. A full table denoting triplicate data can be observed in the Appendix (table 10.3.1).

Table 3.2A Sample selection - antrum virulence typing and Sydney scores for each patient

Patient ID	*vacA* signal type	*vacA* intermediate type	*vacA* mid type	*cagE*	*cagA*	CagA serology	Inflammation	Activity	Atrophy	Intestinal metaplasia
45A	s1	i1	m1	Positive	Positive	N/A	2	2	1	1
77A	s1	i1	m1	Negative	Positive	Negative	2	2	1	0
93A	s1	i1	m2	Positive	Positive	Negative	2	1	1	0
120A	s1	i1	m1	Negative	Positive	Negative	1	0	0	0
194A	s1	i1	m2	Positive	Negative	Positive	2	2	1	2
265A	s1	i2	m2	Positive	Positive	N/A	3	2	0	0
295A*2	s1	i2	m2	Positive	Positive	Negative	2	2	0	0
308A	s1	i2	m2	Negative	Negative	Positive	2	2	0	0
322A	s1	i1	m1	Positive	Positive	Positive	1	0	1	3
326A*2	s1	i2	m2	Positive	Positive	Negative	2	2	0	0
439A	s1	i1	m2	Positive	Negative	N/A	1	2	0	3
444A	s1	i1	m1	Positive	Positive	N/A	2	2	1	2
495A	s1	i1	m1	Positive	Negative	Negative	2	0	0	0
537A*1	s1	i2	m2	Positive	Positive	Positive	2	2	0	0
565A	s2	i2	m2	Negative	Negative	Positive	2	2	1	0
621A	s1	i1	m1	Positive	Positive	Negative	2	2	0	0
732A	s1	i1	m2	Positive	Positive	Negative	2	2	0	1

Table 3.2B Sample selection - corpus virulence typing and Sydney scores for each patient

Patient ID	*vacA* signal type	*vacA* intermediate type	*vacA* mid type	*cagE*	*cagA*	CagA serology	Inflammation	Activity	Atrophy	Intestinal metaplasia
45C	s1	i1	m1	Positive	Positive	N/A	1	0	0	0
77C	s1	i1	m1	Negative	Positive	Negative	1	0	0	0
93C	s1	i1	m2	Positive	Positive	Negative	1	1	0	0
120C	s1	i1	m1	Negative	Positive	Negative	1	0	0	0
194C	s1	i1	m2	Positive	Negative	Positive	3	2	1	0
249C*[1]	s1	i1/i2	m2	Positive	Positive	Positive	1	0	0	0
265C	s1	i2	m2	Positive	Positive	N/A	1	0	0	0
295C*[2]	s1	i2	m2	Negative	Negative	Negative	1	0	0	0
308C	s1	i2	m2	Positive	Positive	Positive	2	1	0	0
322C	s1	i1	m1	Positive	Positive	Positive	3	2	0	1
326C*[2]	s1	i2	m2	Positive	Positive	Negative	1	0	0	0
439C	s1	i1	m2	Negative	Negative	N/A	1	0	0	0
444C	s1	i1	m1	Positive	Positive	N/A	1	0	0	0
495C	s1	i1	m1	Positive	Positive	Negative	1	0	0	0
565C	s2	i1/i2	m2	Positive	Positive	Positive	N/A	N/A	N/A	N/A
732C	s1	i1	m1	Positive	Positive	Negative	N/A	N/A	N/A	N/A

Tables 3.2A and 3.2B display the virulence typing and histologically scored Sydney results from the antrum and corpus respectively. Table fields are colour coded dependant on the criteria for sample inclusion of a presumptive mixed *H. pylori* infection (section 3.2.1). Blue = opposite sample pair (antrum/corpus) holds a different result, orange = sample specific indicator of a mixed infection (often inferred by adjacent cell results). Patients denoted with *[1] and *[2] represent returning patients providing sequentially sampled biopsies, referred to as sequential set 1 and 2 respectively. This is discussed in more detail within Chapter Six. The Sydney scores ranged from 0-3 with 0 = normal, 1 = low, 2 = moderate and 3 = high. All typing, and data reporting was provided by Professor Atherton and colleagues at the University of Nottingham.

Table 3.3 General patient information

Patient ID	Age	Sex	Ethnicity	Indication for endoscopy	Relevant patient medical history	Acid Suppressants	Antibiotics	Other Relevant Drug History
45	63	Female	Caucasian	Vomiting/ weight loss	N/A	N/A	N/A	N/A
77	57	Female	Caucasian	Bloating/ dysphagia	No	Gaviscon	No	No
93	77	Male	Caucasian	Heartburn	No	No	No	No
120	59	Female	Caucasian	Heartburn	No	No	No	No
194	63	Female	Asian	*H. pylori* culture	Perforated prepyloric ulcer 2 years ago	Ranitidine 150mg bi daily stopped 2 weeks ago	No	No
249*[1]	41	Male	Pakistan	Dyspepsia and *H. pylori* positive	No	Stopped 2 weeks ago	No	No
265	64	Female	N/A	Fe defi	No	No	No	No
295*[2]	68	Female	N/A	N/A	No	Stopped 2 weeks ago	No	No
308	50	Female	N/A	Eradication therapy failure	No	Unclear about PPIs	No	No
322	64	Female	N/A	Epi pain	No	No	No	No
326*2	68	Female	N/A	Eradication therapy failure	Previous eradication attempt	Stopped 2 weeks ago	No	No
439	51	Female	N/A	N/A	N/A	N/A	N/A	N/A
444	68	Female	N/A	N/A	N/A	N/A	N/A	N/A
495	40	Male	N/A	N/A	No	No	No	No

Patient ID	Age	Sex	Ethnicity	Indication for endoscopy	Relevant patient medical history	Acid Suppressants	Antibiotics	Other Relevant Drug History
537*[1]	45	Male	Pakistan	Eradication therapy failure	Previous eradication attempt	Stopped 2 weeks ago (omeprazole 40mg)	No	No
565	79	Male	Caucasian	Dyspepsia	No	No	No	No
621	66	Male	Caucasian	Dyspepsia, atypical chest pain	No	No	No	No
732	64	Male	Caucasian	Epi pain	No	No	No	No

General patient information for all patients used within this thesis. The age, sex, ethnicity, indication for gastric endoscopy, relevant patient medical history, recent acid suppressant history, recent antibiotic use and other drug history are displayed. Patients denoted with *[1] and *[2] represent returning patients providing sequentially sampled biopsies, referred to as sequential set 1 and 2 respectively. This is discussed in more detail within Chapter Six. Abbreviations used within this table include; Fe defi – iron deficiency anaemia and Epi pain – epigastric pain. Antibiotics were recorded if taken within 2 weeks of the endoscopy. All data recording was provided by the University of Nottingham from Professor Atherton and colleagues.

Table 3.4 Patient disease status

Patient ID	Oesophagus	Stomach	Duodenum	Other Notes	Disease Type	CLO Result
45	N/A	N/A	N/A	Deformed pylorus, duodenal ulcer disease (scars/craters)	Past DUD/ Normal	Positive
77	Minor erosion in HH	No	Duodenal ulcer and erosions	Minor HH erosion, DU and erosions	Acute DU	Positive
93	RO(2)	No	No	N/A	Normal	Positive
120	RO(2)	No	No	N/A	Normal	Positive
194	No	No	Scar in Di	Scar in Di (PPyU, DU scar)	Previous PPyU/ DU scar/ past DU	Positive
249*[1]	RO(2)	No	eDi and deformed pylorus	N/A	Acute DU	Positive
265	HH and erosions	No	No	N/A	Normal	Positive
295*[2]	RO(1)	Antral erosions eGi	2 tiny erosions possible eDi	N/A	N/A	Positive
308	No	No	No	N/A	Normal	Positive
322	RO(2)	GU	No	N/A	Acute GU	Positive
326*[2]	RO(2)	No, but previous eGi	No, but previous eDi	N/A	Normal previous eGi	N/A
439	N/A	N/A	N/A	N/A	N/A	Positive
444	N/A	N/A	N/A	N/A	N/A	Positive
495	RO(1)	No	eDi (4 erosions)	N/A	N/A	N/A

Patient ID	Oesophagus	Stomach	Duodenum	Other Notes	Disease Type	CLO Result
537*[1]	RO(2)	No	eDi, previous DUscar	N/A	N/A	N/A
565	N/A	No	N/A	N/A	N/A	N/A
621	RO(1)	Atrophic gastritis	No	N/A	N/A	N/A
732	RO(2)	2x GU on lesser curve	Erosive pyloric. No DU	N/A	N/A	N/A

Disease status of the oesophagus, stomach, duodenum is displayed for each patient as well as other general notes and disease type. Patients denoted with *[1] and *[2] represent returning patients providing sequentially sampled biopsies, referred to as sequential set 1 and 2 respectively. This is discussed in more detail within Chapter Six. The CLO test refers to the rapid urease test used to indicate *H. pylori* infection. Abbreviations used within this table include; HH – hiatus hernia, RO – reflux oesophagitis (1= mild, 2= moderate, 3= severe), eGi – erosive gastritis, GU – gastric ulcer, Di – inflamed duodenum, eDi – erosive duodenitis, DUscar – scar from a previous duodenal ulcer, DU – duodenal ulcer, PPyU – pre-pyloric ulcer and DUD – duodenal ulcer disease. All data was provided by Professor Atherton and colleagues at the University of Nottingham.

Table 3.5 Antibiotic breakpoints for disk diffusion assay

		Zone of inhibition (mm)		
Antibiotic class	**Antibiotic (μg)**	**Resistance**	**Susceptibility**	**Related studies**
β-Lactam antibiotic	Amoxicillin 10 μg	≤ 12.5 mm	> 12.5 mm	Lang and García, 2004; Ogata, Gales and Kawakami, 2014
Macrolide	Clarithromycin 15 μg	≤ 9 mm	> 9 mm	McNulty *et al.*, 2002; Ogata, Gales and Kawakami, 2014
Fluoroquinolone	Levofloxacin 1 μg	≤ 6 mm	> 12 mm	Yu *et al.*, 2011; Boyanova *et al.*, 2016
Nitroimidazole	Metronidazole 5 μg	≤ 10 mm	> 10 mm	McNulty *et al.*, 2002; Lang and García, 2004; Ogata, Gales and Kawakami, 2014
Rifampicin	Rifampicin 5 μg	≤ 17.5 mm	> 17.5 mm	Glocker, Bogdan and Kist, 2007; Chisholm and Owen, 2009
Broad spectrum polyketide	Tetracycline 30 μg	≤ 10.5 mm	> 10.5 mm	Lang and García, 2004; Ogata, Gales and Kawakami, 2014

Antibiotic breakpoints for the different antibiotics used within this study. Breakpoints were inferred by the literature from multiple studies and denoted in the 'related studies' column. These breakpoints refer to a disk diffusion methodology.

3.3.2. Antimicrobial resistance assays

The recording of zones of inhibition for the antimicrobial resistance assays proved difficult due to the dark discolouration during incubation of the culture plates containing defibrinated horse blood (section 3.4.2). This combined with the fastidious growth of the *H. pylori* cultures was difficult and has been recognised by other studies (Chisholm and Owen, 2009). However, placing the culture plates under a bright light source and moving the direction of the light helped with measurements.

Despite the difficulties in the recording of the zones of inhibition, the standard deviations for independent triplicate data for each antibiotic tested across all samples was small (figure 3.1). This suggests that the antimicrobial resistance assay designed for this study was reproducible.

The disk diffusion method employed within this study allowed for other useful data to be collected, such as clinical sweeps that held resistant colonies within the zone of inhibition and those with multiple zones of inhibition (Appendix, table 10.3.1). This technique has been recognised by McNulty *et al.* (2002), who also proclaim the cost effectiveness of this technique and ease of potential implementation into the clinical setting, standing alongside current techniques.

Antibiotic disk diffusion breakpoints for *H. pylori* are not defined by internationally recognised guidelines, such as EUCAST (EUCAST, 2019). However, many research studies have used a disk diffusion based antimicrobial resistance assay (reviewed by McNulty *et al.*, 2002). These studies and others (table 3.5) vary in the reported breakpoints for the disk diffusion method. Furthermore, most of these studies vary in the assay methods including length of incubation, concentration of antibiotics and culture media used. Therefore, there is a need to standardise an antibiotic disk diffusion assay for the study of *H. pylori* clinical breakpoints.

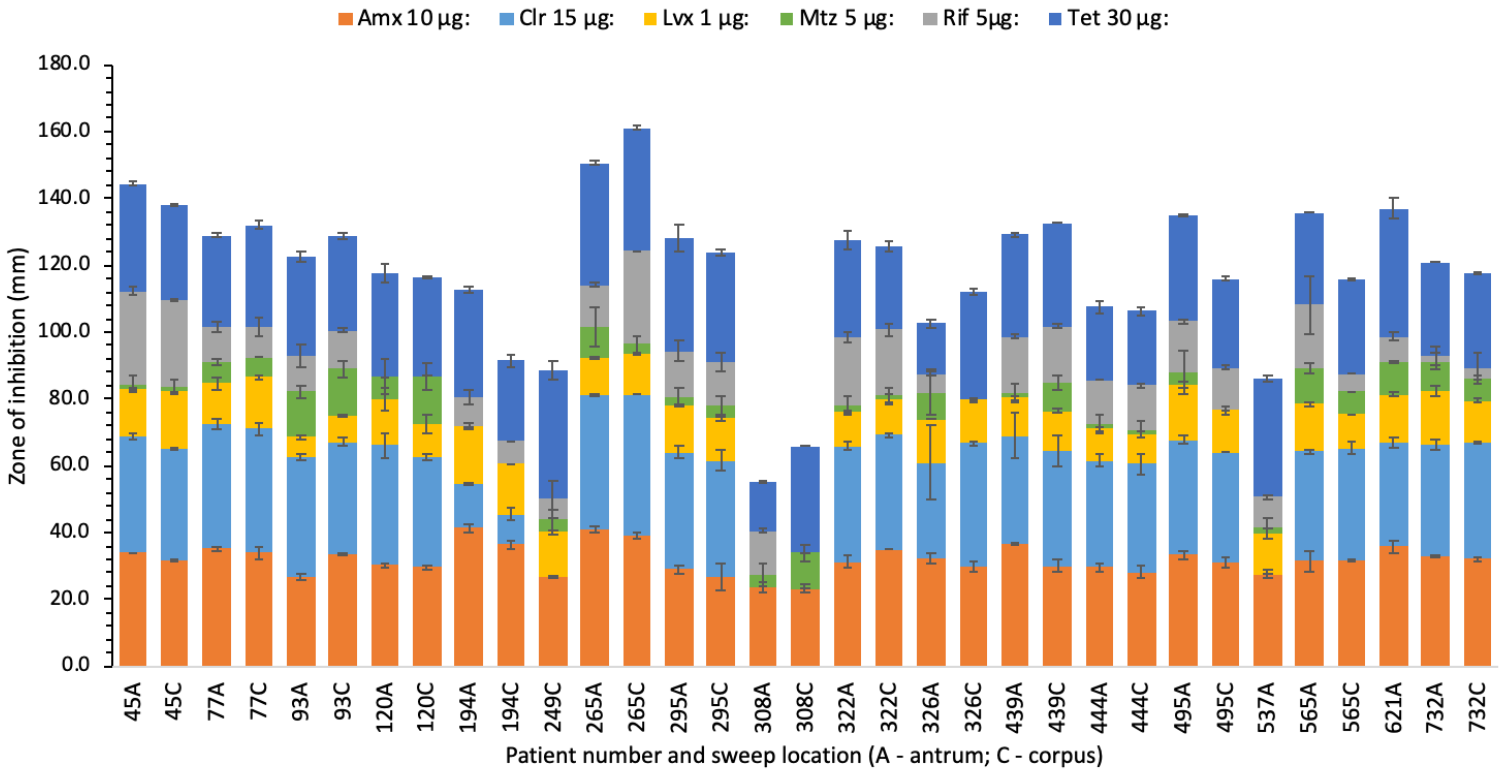
Nonetheless, the breakpoints used to define sensitive and resistant phenotypes within this study are denoted in table 3.5 and were converged upon by reviewing recent studies that employed the disk diffusion technique.

Some patients were found to hold *H. pylori* populations with differing sensitivity to the tested antibiotics by their zones of inhibition (figure 3.1). Of particular note were patients

265, 308 (later found to be contaminated – Chapter Four, table 4.1), 326 (later found to be contaminated – Chapter Four, table 4.1), 495 and 565. This correlates with other studies (Ailloud et al., 2019; Selgrad et al., 2014) and was used as a further indicator of a mixed strain *H. pylori* infection. These results were an indication that the sample selection criteria (section 3.2.1) were useful in identifying these patients with a presumptive mixed strain infection and justified the sequencing of these samples (Chapters Four – Six).

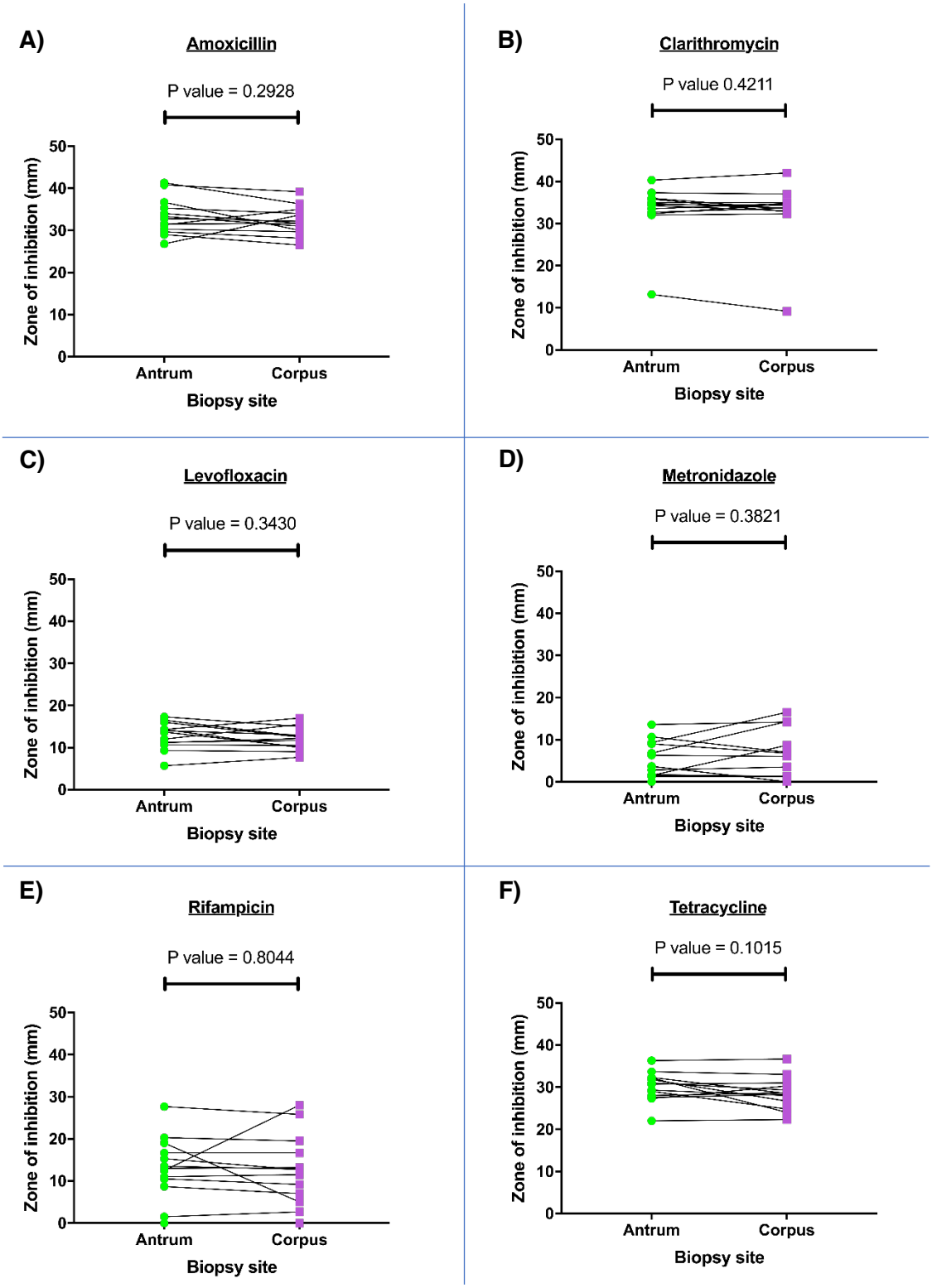
The antibiotic zones of inhibition of *H. pylori* populations taken from the antrum and corpus were compared using a paired t-test to identify if there was a general sensitivity difference between antrum and corpus populations (figure 3.2). No statistically significant differences were identified suggesting that while there are observable between niche differences for some antibiotic resistance profiles, the sample site might not be important as long as multiple biopsies were cultured for antimicrobial susceptibility testing. However, further investigation would be required to confirm or dispute this observation. In particular, multiple biopsies from each niche (antrum and corpus) would be beneficial in determining the best biopsy site for antimicrobial susceptibility testing.

Figure 3.1 Antibiograms of all clinical sweeps used within this study



Antibiograms of each patient clinical sweep used within this thesis. Antibiotics displayed; Amx (amoxicillin), Clr (clarithromycin), Lvx (levofloxacin), Mtz (metronidazole), Rif (rifampicin) and Tet (tetracycline). The six different antibiotics are overlaid as stacked bars for all clinical sweeps. Standard deviation bars are depicted for each antibiotic (triplicate data). Patients denoted with *1 and *2 represent returning patients providing sequentially sampled biopsies, referred to as sequential set 1 and 2 respectively. This is discussed in more detail within Chapter Six. This figure was created in Microsoft Excel (version 16.16.6). Raw data can be observed in Appendix table 10.3.1including populations that harboured resistant colonies within the zone of inhibition and potential second zones of inhibition.

Figure 3.2 Paired antrum and corpus antibiogram statistics



Paired t-test for antrum and corpus average zones of inhibition from each patient (where antrum and corpus pairs were available) for each antibiotic. Green circles represent antrum zones of inhibition where the opposite corpus pair per patient is depicted as purple squares. This analysis was conducted in GraphPad Prism (version 8.2.0).

3.4. Future work

Given ethical approval and agreement between both patients and the health practitioner, multiple (>3) pinch biopsies could be taken from the antrum and corpus of individual patients. Additionally, this could be extended to include the fundus region of the stomach. These multiple biopsies could be used to determine whether there are between niche as well as within niche antimicrobial resistance differences from the resident *H. pylori* populations. This would provide a more dynamic and robust study into *H. pylori* antimicrobial resistance differences within the stomach and might shed further light onto why some patients are not successfully eradicated of their infection. Furthermore, niche specific differences might be better studied with more samples. Such niche specific differences could potentially be a factor for increased antimicrobial resistance due to different selection pressures acting on the *H. pylori* populations.

Increasing the sample size would also help to further validate the reproducibility of the disk diffusion assay presented in this study and could aid in the creation of a more standardized disk diffusion methodology. Ideally, an internationally recognised disk diffusion assay and related clinical breakpoints is needed.

Isolating more resistant colonies found within zones of inhibition and carrying out whole genome sequencing on these strains might identify novel resistance mutations.

Next, population deep sequencing was carried out on presumptive mixed *H. pylori* strain populations to investigate the within and between niche genetic diversity (Chapter Four). Following this, single colony isolates were cultured from population sweeps to provide another layer of genetic analysis as well as deep sequencing data validation (Chapter Five). In Chapter Six, sequentially sampled patients from before and after eradication therapy were investigated using both deep population and single colony sequencing techniques.

4. Chapter Four: Whole genome deep population sequencing

4.1. Introduction

Bacterial chronic infection in humans is rare as bacterial infections are usually cleared by the immune system or are cleared by targeted antimicrobial therapies. *Helicobacter pylori* is unusual in the fact infection usually results in some level of gastritis but is most often asymptomatic in nature. Therefore, most infected individuals are unaware they are infected by this pathogen. *H. pylori* infects people in early childhood and can persist as a chronic, lifelong infection. Typically, eradication therapy is only sought if more symptomatic diseases are presenting. Due to the chronic nature of *H. pylori* infection and the high mutation and recombination rate, it is thought that a diverse population of quasispecies is present within infected individuals. It is also thought that some patients can be infected by more than one infecting strain, multiplying this within patient diversity further still.

Some studies have investigated the genetic diversity of *H. pylori* and have revealed a global population structure and diversity (Duncan et al., 2012; Latifi-Navid et al., 2010; Olbermann et al., 2010b; Montano et al., 2015; Vale et al., 2015a). This also holds true for some virulence associated genes (Duncan et al., 2012; Olbermann et al., 2010b). Such studies have employed a comparative genomics approach, usually of single colony isolates to investigate global population diversity. Other studies have looked at *H. pylori* strains isolated from specific geographical regions (Cortes et al., 2010; Kojima et al., 2016). This global and geographical genetic analysis has helped in the understanding of global and geographic diversity of *H. pylori.* Additionally, this has allowed for the identification of certain genotypes and their relation to disease types and severity. For example, a recent genome wide association study carried out by Berthenet *et al.* (2018), identified SNP and gene presence markers which were used to generate a risk score to assess gastric cancer risk. Another study identified six genes that were associated with peptic ulcer disease and adenocarcinoma by comparing multi-ethnic populations at high to low risk of these disease types (Gunaletchumy et al., 2015). Therefore, despite there being an astonishing level of genetic diversity, genomic analysis at a global and geographical level has provided important insights into virulence and disease progression.

Helicobacter pylori genomic diversity has also been investigated at a smaller scale. For example, some studies have investigated single colony isolates obtained from gastric biopsies taken from familial groups (Argent et al., 2008; Furuta et al., 2015; Didelot et

al., 2013; Krebes et al., 2014; Morelli et al., 2010; Kivi et al., 2003; Raymond et al., 2004). Although transmission routes of *H. pylori* are not yet fully elucidated, it is generally thought that infection is likely to be passed on by close family members during early childhood. In most cases the findings of these studies supported a familial transmission route, although this was not always the case. These studies provide insights into how *H. pylori* might adapt to a new host and in disease progression. Indeed, these studies have identified familial differences in virulence and colonisation associated genes. Several studies agree that *H. pylori* outer membrane proteins and other cell surface related genes are particularly variable. It has also been reported that recombination introduced much more variability in terms of SNPs than the general clock rate of mutation, but that a mixed infection likely facilitates this.

Few studies have investigated genetic diversity of *H. pylori* within individual patients (Didelot et al., 2013; Raymond et al., 2004; Kivi et al., 2003; López-Vidal et al., 2008; Matteo et al., 2007; Israel et al., 2001; Reyes-Leon et al., 2007b). Out of these studies, only two (Noto et al., 2017; Didelot et al., 2013) used a whole genome sequencing approach while the others used a PCR amplification and sequencing methodology to focus on specific genes/loci. López-Vidal *et al.* (2008), revealed differences in *cagA* genotype between niches as well as differences in gene sizes, suggesting that there can be virulence differences between different niches of the same stomach. This was further supported by earlier findings by Reyes-Leon *et al.* (2007), who identified a sequence truncation after the EPIYA-C motif between isolates taken from different niches of the same stomach. Didelot *et al.* (2013), showed that strains isolated from the antrum and corpus of some patients were likely originated from separate infecting strains, suggesting strain-specific niche adaptation. These findings were based on a relatively low sample size, with the whole genome studies taking just two representative single colonies from both the antrum and corpus of individual patients. Despite these low sample sizes, it is clear that insights into within host microevolution can be attained. However, the low sample size of these studies is likely to be unrepresentative of the whole population structure and not able to capture the *true* extent and implications of within niche genetic diversity. Further investigation into within niche diversity could have important implications in understanding how *H. pylori* is able to colonise a new host or new within host environment, how colonisation can persist life long, why some patients develop different disease types and why some niches of the stomach have varying disease severity.

Within patient genetic diversity of a chronically infecting population of *Burkholderia dolosa* of individuals with cystic fibrosis has been investigated by Lieberman *et al.* (2014), using a population deep sequencing approach. In this approach, the total genomic DNA of the whole bacterial population was extracted and sequenced to high sequencing coverage depth of greater than 100x. This is in contrast to single colony sequencing that is usually sequencing at an average depth of 30x which is considered reliable to call between single colony alignment-based SNPs. In the deep sequencing methodology described by Lieberman *et al.* (2014), if there are variant strains within the population, the genomic sequence of these strains are sequenced alongside all other alternative sequences. This generates a variable sequencing read set at genetically diverse regions within the genome. They then mapped these reads to an inferred reference genome that was isolated from a different outbreak patient. By mapping these reads to a reference genome, they were able to call polymorphisms where sequencing reads did not agree to that of the reference sequence. This methodology captured a snapshot of within patient genetic diversity of the whole population, something that would be extremely difficult to capture to this resolution using a single colony comparative genomics approach. This study found that diversifying quasispecies did not fix within the population but co-existed during chronic infection with polymorphic diversity providing a historical record of selection. Furthermore, this study revealed genes that were more prone to within patient polymorphic diversity, namely outer membrane associated genes, antibiotic resistance associated genes and iron scavenging genes.

Helicobacter pylori, like some bacterial infections of the cystic fibrosis lung, are chronic infections. Therefore, a population deep sequencing methodology might be useful in investigating within patient genetic diversity of the human stomach. Employing this method might better elucidate the within patient diversity and build on the findings of previous studies of this nature, as previously discussed.

In this Chapter, population deep sequencing was carried out on antrum and corpus samples from individual patients to investigate the within and between niche genetic diversity of *H. pylori* and compare findings between patients to identify genes that are more prone to polymorphic variation. This is the first time population deep sequencing has been used to investigate within patient *H. pylori* genetic diversity.

4.2. Materials and Methods

Sample selection, DNA extraction, whole genome deep sequencing, sequencing read curation, contamination detection, whole genome assembly, assembly curation and genome annotation was conducted as described in Chapter Two.

Definition of antrum and corpus differences are described in Chapter One section 1.4.4.

Samples used in this study are denoted in Chapter Three table 3.2A-B.

4.2.1. Within niche diversity

4.2.1.1. Read mapping based pipeline to detect within niche common and minor allelic variation

A read mapping and polymorphic detection pipeline was developed similar to that described by Lieberman et al. (2014), which employed numerous bioinformatic tools and processing steps.

Firstly, curated read sets from each sample were aligned to their corresponding *de novo* assembly (Chapter Two, sections 2.9 – 2.12) through Bowtie2 version 2.3.4.3 (Langmead and Salzberg, 2012). Bowtie2 was configured in 'very sensitive mode' to increase read mapping sensitivity and accuracy at the expense of speed and computing memory with the addition of the '`--no-mixed`' flag forcing bowtie2 to find alignments with paired-end reads sets only. A further modification was made by altering the maximum fragment length allowed for paired-end alignments to 2,000 bp, to account for *H. pylori* biology and in consideration of population deep sequenced data. A final adjustment was made using the '`--n-ceil`' flag by setting this to '`0,0.01`' in order to reduce the number of ambiguous characters allowed within an aligned read. This effectively reduced base ambiguity to less than 1%, further increasing read mapping sensitivity and accuracy. The following full command was used:

```
bowtie2-build --threads {user specified CPU/thread/core number}
-f {sample assembly file} {user named bowtie2 index file} ;
(bowtie2 -p {user specified CPU/thread/core number} -X 2000 --
no-mixed --very-sensitive --n-ceil 0,0.01 --un-conc-gz {path to
```

```
output file of paired-end reads that fail to align} -x {path to bowtie2 index file} -1 {path to forward/mate pair 1 reads from Sickle} -2 {path to reverse/mate pair 2 reads from Sickle} -S 120A_bt2_out.sam) 2>{path to output log file}
```

Secondly, the resulting sequence alignment map (SAM) file produced by Bowtie2 was passed through the SAMtools suite version 1.9 (Li et al., 2009) to sort and remove PCR duplicated reads. The following command was executed:

```
samtools view -@ {user specified CPU/thread/core number} -b -h -o {path to output BAM conversion} {path to input SAM from bowtie2} ; samtools sort -@ {user specified CPU/thread/core number} -n {path to input BAM file} -o {path to output sorted BAM file} ; samtools fixmate -@ {user specified CPU/thread/core number} -m {path to input sorted BAM file} {path to fixmate processed output BAM file} ; samtools sort -@ {user specified CPU/thread/core number} {path to input fixmate BAM file} -o {path to output sorted fixmate BAM file} ; samtools markdup -@ {user specified CPU/thread/core number} -S {path to input sorted fixmate BAM file} {path to output PCR duplicate removed BAM file}
```

Following SAM processing through the SAMtools suite, the resulting binary format SAM (BAM) files were passed to a haplotype caller, FreeBayes version 1.3.1 (Garrison and Marth, 2012). FreeBayes uses a Bayesian statistical inference model to call genetic variants from short-read alignments. FreeBayes was carefully configured using a number of arguments. The '`--pooled-continuous`' mode was selected in order to observe alternative nucleotide frequencies. Importantly, the '`-F`' flag was used to lower the minimum alternative fraction to support a particular alternative allele call to 3%. Additionally, a mapping quality of ≥ 34 and a base quality score of ≥ 30 was required to elevate an alternative allele call. This command is displayed below:

```
freebayes -f {path to sample de novo assembly) -F 0.03 --pooled-continuous -m 34 -q 30 {path to output BAM file from SAMtools} > {path to FreeBayes output VCF file}
```

FreeBayes produced VCF files were filtered using two different sets of parameters to call common and minor allelic variants using vcflib (Garrison,

https://github.com/vcflib/vcflib). To call common allelic variants the '`-f`' option was used to filter for SNPs, total number of alternative allele calls on the forward strand > 15 and total number of alternative allele calls on the reverse strand > 15. Minor allelic variation was determined by filtering for SNP sties only. The following example command was used to filter for common allelic variant sites:

```
vcffilter -f "TYPE = snp & SAF > 15 & SAR > 15" {path to input VCF file from FreeBayes} > {path to filtered output VCF file}
```

Filtered VCF files from the vcflib tools output were further filtered manually to ensure the called variants were of high quality and confidence. Each VCF file was inspected manually and variant sites were removed if they were identified in the first or last 500 bp of a contig. Next, the mapped reads in the BAM alignment files were loaded into Artemis (Carver et al., 2012) alongside the reference annotated genome. Variants located in a repeat region of 6 nucleotides or more were identified and inspected. These variant sites were removed if the reads supporting the alternative call were within the first or last 3 bases of the reads or if the supporting reads held other variants within a 20 bp range of the variant site. If the beginning of the reads aligned within the repeat region and the variant site was more than 3 bases from the beginning of the alignment, these were discarded due to the limitations of the Illumina sequencing by synthesis base calling in repeat regions. This limitation is a result of the increase in signal intensity when more than one base is added, which is harder to determine especially when taking into consideration the background *noise*. The support of just one read that passed these criteria was deemed enough evidence to classify the alternative call as valid, even if all other reads failed the measure.

Filtered variants were annotated as described in Chapter Two section 2.14 and manipulated manually to produce a list of annotated gene products that were found to harbour polymorphic sites. This approach allowed for the grouping of genes by associated gene product that harboured polymorphic sites. For example, two different polymorphic sites might be identified within the same hypothetical protein, but another SNP might also be found in a different hypothetical protein, in this case the hypothetical protein would be recorded three separate times as three SNPs were recorded associated with this gene product name, indicating a high level of diversity within this gene associated product. Using this data, a heatmap was created using ggplot2 and

the RColorBrewer through the R statistical software version 3.5.1 (R Core Team, 2018) where deeper colours represent gene products with higher counts of polymorphic sites.

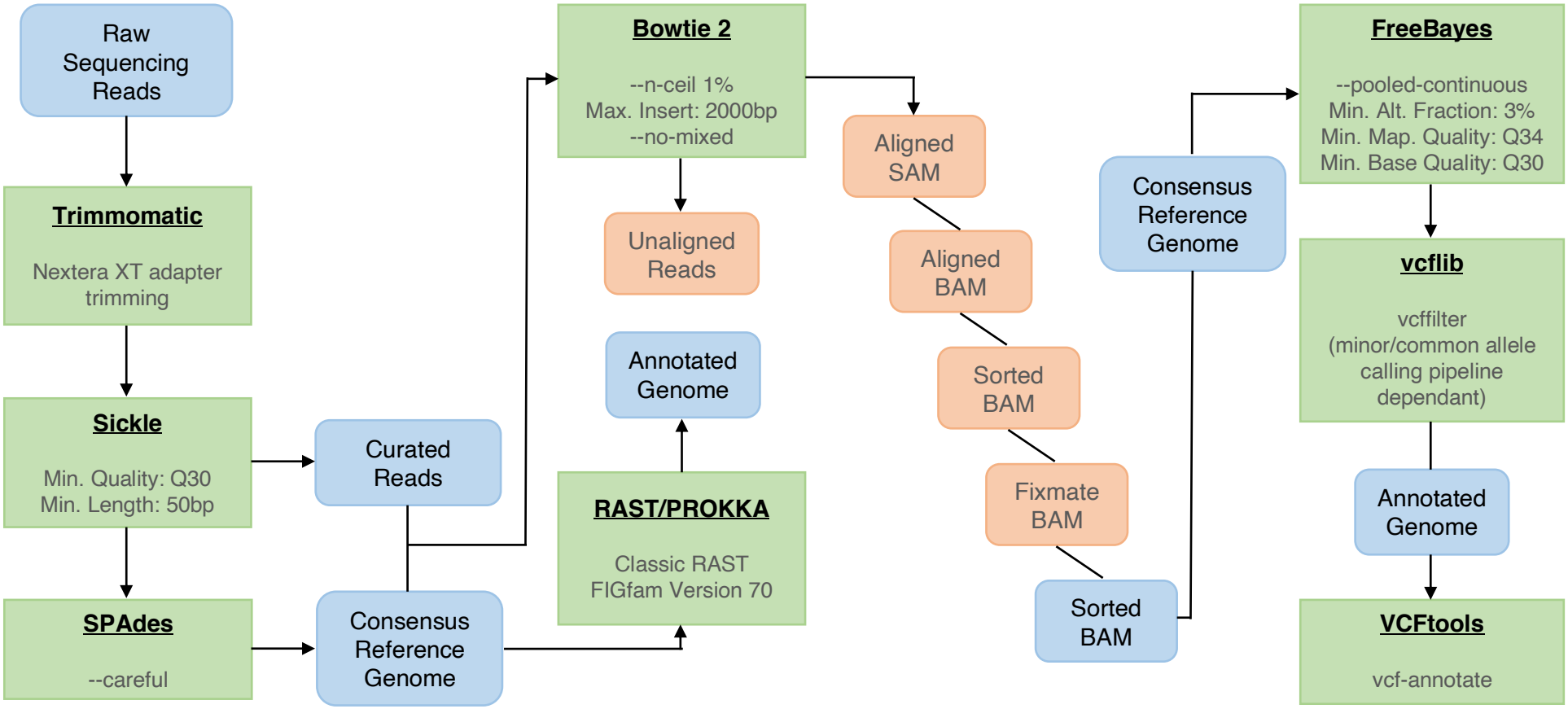
An overview of the bioinformatics deep sequencing pipeline is depicted in figure 4.1.

4.2.1.2. Per-base depth of coverage determination using sequence alignment mapping data

Sequence depth at a per nucleotide level was determined using mosdepth version 0.2.3 (Pedersen and Quinlan, 2018). This was used in the place of the SAMtools (Li et al., 2009) depth pipeline as the mosdepth algorithm prevents the double counting from ends of paired-end read sets, resulting in a more accurate determination of per-nucleotide sequencing coverage. Furthermore, mosdepth outputs a fraction based coverage statistic across each contig. This data was manually plot using ggplot2 and the RColorBrewer through the R statistical software version 3.5.1 (R Core Team, 2018) to present the coverage statistics per-contig of each dataset on a single graph (figure 4.3).

A second tool, weeSAM version 1.4 (https://github.com/centre-for-virus-research/weeSAM) was used to plot further depth statistics (figure 4.4).

Figure 4.1 Flow chart of the deep sequencing analysis pipeline



Green boxes represent software used to analyse the data with software names in bold and underlined. Key software optimisations are listed below software names. Blue boxes denote data input and output streams while orange boxes are intermediate or unused data streams.

4.2.2. Between niche diversity

4.2.2.1. Whole genome consensus alignment of paired antrum and corpus patient pairs

Deep sequenced populations resulting in a consensus assembled genome for each paired antrum and corpus sample were aligned in a pair-wise fashion using the multiple genome alignment (Mauve, version 2.4.0) tool in progressive mode with default settings (Darling et al., 2004). Mauve can align core genome sequence mosaics resulting from genomic rearrangement, recombination, deletions and insertions which makes this tool particularly suited to align *H. pylori* genomes.

Resulting alignments were used to extract variant positions/SNPs between aligned segments and input into a spreadsheet for further downstream analysis (sections 4.2.2.2 – 4.2.2.4).

4.2.2.2. Mapping patient niche specific deep sequencing read sets to the opposite niche consensus genome

To further investigate the between niche genetic diversity a second approach was employed, similar to the whole genome alignment of patient sample pairs as described above (section 4.2.2.1). The difference with this approach is the use of all sequencing read data available, rather than a reliance on two consensus assembled data sets. A consensus assembled genome of a deep sequenced population represents a snapshot of the most abundant sequence string taken from the population. Furthermore, it is possible that a variant position or run of positions within the consensus genome might cause a break in the assembly due to the assembler not being able to converge on a correct sequence pattern, resulting in a contig break. It is also difficult to determine how many reads cover a certain nucleotide position within the assembled genome where it is possible that a single read was used to determine a particular base, making this base position less likely to be part of the true biological sequence or a true variant position. For these reasons, it is not uncommon to observe an increased number of variant positions between aligned genomes, particularly around contig breaks and within contigs of short length. Despite this, these positions cannot be fully ignored as some of these variant positions are potentially true sites of variation between aligned genomes.

To better validate the variant SNP positions identified from the Mauve approach, a read mapping methodology was adopted to support the identification of true biological sequence variants between paired sample sets from the same patients.

The rapid haploid variant calling and core genome alignment (Snippy, version 4.4.0) tool (Seemann, https://github.com/tseemann/snippy) was selected for this purpose. Briefly, Snippy aligns sequencing reads to a reference sequence and then passes the alignment to a haplotype caller that identifies SNPs and indels which are subsequently reported. The following command was used where A = antrum and C= corpus:

```
snippy --cpus {user specified CPU/thread/core number} --report --minfrac 0.9 --mincov 30 --mapqual 30 --basequal 30 --ref {consensus deep sequenced genome of patient pair A} --outdir {user specified output directory} --R1 {path to forward/mate pair 1 of patient pair C} --R2 {path to reverse/mate pair 2 of patient pair C}
```

The reverse of this command was also executed for each paired sample set where the '`--ref`' genome was that of the deep sequenced corpus assembled genome and the sequencing read set for '`--R1`' and '`--R2`' was from the antrum deep sequencing dataset. This was done to map the SNP locations to the corresponding base positions within the genomes. The VCF files for each paired analysis were filtered to only include SNP variants. As with the Mauve method, these VCF files were annotated to determine whether variant SNPs were harboured within a CDS or IGR. This SNP annotation method is described in Chapter Two section 2.14.

For a SNP to be called by Snippy, 90% of reads had to be in support of the alternative allele with a minimum coverage of 30 reads overlapping the variant site. Additionally, only high-quality reads (Q30) and mapping qualities (Q30) were used to interrogate a variant site. These parameters were chosen to improve the confidence of variant calls between the datasets.

4.2.2.3. BLAST ring image generator: visually displaying between niche sites of variation by combining and comparing the Mauve and Snippy methodologies

The contigs from genome assemblies of paired datasets were filtered to remove contigs <500 bp in length (https://github.com/tinybio/filter_contigs) and individually concatenated to produce a single contiguous sequence. A BLAST (pairwise nucleotide basic local alignment search tool) approach was used to identify regions of nucleotide similarity between paired within patient antral and corpus genomes using the BLAST ring image generator (BRIG, version 0.95) (Alikhan et al., 2011; Altschul et al., 1990; Camacho et al., 2009). BRIG was used to align the sequences in a pair-wise fashion, requiring vice versa reference and query sequence overlays in order to fully investigate between niche genomic differences with an upper 98% and lower 95% nucleotide identity. Contig breaks were plotted to show contig break boundaries within the contiguous sequence. Approximate coverage depth was also plotted by extracting the per-base coverage from SAMtools depth from BAM files produced as output from section 4.2.2.2 by making use of the '`-aa`' flag and using a custom command to create a BRIG compatible graph file by executing the following:

```
len=$(wc -l < {path to input per base coverage statistics in tsv format}) ; paste <(seq 0 $((len-1))) <(seq 1 $len) {path to input per base coverage statistics in tsv format} > {path to BRIG compatible output .graph file}
```

This command adds an additional column to the beginning of the input tsv file containing 0 based coordinates required to plot the BRIG coverage graph.

The output SNPs from both Mauve and Snippy methodologies were imported into a spreadsheet and a manual mapping file was constructed to identify alignment SNPs that were detected by one or both methods. This was done in order to improve the high throughput capability of this analysis by reducing the reliance on manual filtering of the Snippy output SNPs with the aim of viewing high quality alignment SNP calls.

Briefly, the contig number(s) and SNP position(s) for both methodologies (Snippy and Mauve) were recorded into separate spreadsheets. Here, a unique mapping tag was appended to each SNP field by merging the contig identifier with the SNP position within the contig, separated by a colon (:). An additional column was added alongside the mapping locations with the number one (1) running consecutively. This was used as the presence absence identifier for the methodology detection step. Next, the data was then pooled into one spreadsheet, and duplicate mapping tags were removed.

Spreadsheet functions were used to determine whether each non-duplicated SNP was detected by one or both methods by comparisons to the mapping tags of the two methodologies, outputting the presence (1) or absence (0) for both Snippy and Mauve. This was done by using the '`=IFERROR(VLOOKUP())`' Microsoft Excel (version 16.16.6) function where errors/absences were reported as zero (0) and the matching mapping tags for each methodology was recorded as a one (1). The '`IF`' function was used to record the colour of the SNP to be displayed in the BRIG figure depending on whether the SNP was identified by both methodologies (red), Mauve only (black) or Snippy only (teal).

All unique variants (SNPs) identified by the Mauve and Snippy methodologies ('verified' red SNPs) were used to construct a custom VCF file for each antrum and corpus dataset. The VCF file was annotated as described in Chapter Two section 2.14 and annotations were imported into a spreadsheet.

Finally, a list of the between niche SNPs start and stop positions was constructed, in the order they appeared in the genome. The colour of each SNP depending on the methodology detection (as previously described) was added to an adjoining column. A run of numbers from one (1) to the end of the SNP list increasing by one each time was added to the next adjoining column to provide a numerical SNP identifier. These were used to create a tab-delimited file for custom BRIG annotations as described in the BRIG manual. The numbered SNPs allowed for identification of specific SNPs in the resulting figure that can be used to obtain the SNP annotations, providing the gene in which the SNP(s) occur or non-coding region if there was no gene associated.

This was repeated for each antrum and corpus dataset for each patient.

An example spreadsheet and associated functions can be interrogated by downloading the Appendix directory (table 10.4.1) from the OneDrive link (https://myntuac-my.sharepoint.com/:f:/r/personal/n0667645_my_ntu_ac_uk/Documents/OneDrive_link?csf=1&e=UuIp26).

4.2.2.4. Determination of synonymous and non-synonymous between niche variants

High-quality called variants identified by both Mauve and Snippy methodologies (red SNPs; Chapter Four section 4.2.2.3) for each paired antrum and corpus (same patient) data sets were pooled and recorded in a separate spreadsheet. Here, the Mauve alignment SNP file containing the locations of each SNP for both the antrum and corpus was imported into a spreadsheet and the unique mapping tags for the antrum and corpus alignment SNPs were added as previously described. Special care was taken to ensure the order of the antrum and corpus mapping locations were kept in the same order as the original alignment SNP output file (Mauve). The first alignment SNP pair from the antrum and corpus was given the value of one (1) in an adjoining empty cell and was ascended to the end of the SNP mapping list where the following cell was given the value of plus one (+ 1) from the cell before it. This provided a secondary mapping tag unique to each corresponding antrum and corpus alignment SNP location. This allowed for the cross comparison of the different contigs and SNP locations between the antrum and corpus alignments.

Next, the '`VLOOKUP`' function was used to report the secondary mapping numbers for both the antrum and corpus high-quality verified (red) SNPs. These numbers were pooled, and duplicates were removed. A '`VLOOKUP`' function was used to create an antrum reference list of high-quality (red) SNPs of the pooled antrum and corpus data.

A manually constructed custom VCF file containing the variant alignment positions (alignment SNPs) was created with the pooled antrum and corpus data. The VCF file was annotated as described in Chapter Two section 2.14.

An example spreadsheet and associated functions can be interrogated by downloading the Appendix directory (table 10.4.1) from the onedrive link (https://myntuac-my.sharepoint.com/:f:/r/personal/n0667645_my_ntu_ac_uk/Documents/OneDrive_link?csf=1&e=UuIp26).

This approach was taken as one dataset might call a hypothetical SNP by both methods (antrum), but this SNP might not be called on the reverse analysis by both methods in the other pair (corpus). One reason for this could be a fluctuation of sequencing coverage of sufficient quality on one dataset but not on the other, resulting in a variant detection by both methods on the first dataset, but not in the reverse pair for that variant position. In this example, the effect of coverage variation would have the highest impact on the Snippy method where an alternative nucleotide would not be called if the

coverage was below 30 or the alternative fraction was less than 90%. However, in this scenario, the called variant position by at least one antrum/corpus dataset is evidence that this position is a true site of variation between the niches. Furthermore, there are often more alignment SNPs identified around contig boundaries from aligned genomes (Mauve). This is mainly down to the *de novo* assembly of the contigs where a contig boundary is often a region that could not be joined/scaffolded to another contig. Therefore, these regions are often more variable and less reliable than further along the contig. Combining and comparing the Mauve and Snippy methodologies allows for a more accurate identification of higher quality alignment variants as the alignment variant by the Mauve method should be supported by the mapping of the opposite niche reads by the Snippy method and vice versa.

To determine whether the sites of variation were synonymous or nonsynonymous a tool named SnpEff version 4.3 (Cingolani et al., 2012) was used after resulting VCF files were annotated as described in Chapter Two section 2.14. SnpEff takes VCF files and predicts the effect of SNPs (such as SNP resulting in an amino acid change) within CDS.

The prebuilt database of *H. pylori* reference genomes could not be consulted by this method due to the choice of reference genomes used throughout this study. Furthermore, *H. pylori* genomes have an extremely high level of genetic diversity, aforementioned. Therefore, it was necessary to build custom databases within SnpEff for each reference genome (deep sequenced and consensus assembled genomes) in order to accurately determine the effect of SNPs within CDS.

Custom databases were created by modifying the SnpEff configuration file (snpEff.config) and manually adding each new reference entry to the list with the selection of the bacterial and plat plastid codon table. The genome assembly sequence files and annotated GTF files from RAST were compressed into sample/reference directories within the SnpEff data directory. Following this, the custom reference database was created by navigating to the SnpEff directory and executing the following command:

```
snpEff build -gtf22 -v {name of the database to build, as
described in the edited config entry}
```

After all reference databases were built, the VCF files containing the variant sites identified by both the Mauve and Snippy methodologies were passed through the SnpEff tool by the following command:

```
snpEff -v {name of the reference to use within the SnpEff database} {path to VCF file corresponding to the associated reference} > {path to user specified output SnpEff annotated VCF file}
```

On completion, the summary file was inspected to determine the overall effect of variants between the aligned genomes in terms of synonymous and nonsynonymous mutations.

4.2.2.5. Pan-genome analysis

A pan-genome analysis involving all deep sequenced patient paired samples was performed using the rapid large-scale prokaryote pan-genome analysis (Roary, version 3.8.2) pipeline (Page et al., 2015). Roary takes multiple annotated genomes from PROKKA (Chapter Two, section 2.13) and identifies the shared core and soft core (99% - 100% and 95% - 99% respectively) genes as well as the accessory shell and cloud (15% - 95% and 0% - 15% respectively) genes shared between the samples. From here, a gene presence absence list was generated and manually inspected to identify gene patterns between antrum and corpus sample groups. The following Roary command was executed:

```
cd {path to directory containing list of PROKKA output .gff files to pass to Roary} ; roary -p {user specified CPU/thread/core number} -e -n -v -f {path to output directory} *.gff
```

A core genome alignment of all consensus genomes was conducted and a phylogenetic tree by approximately-maximum-likelihood was constructed through Parsnp (version 1.2) which is part of the Harvest suite (Treangen et al., 2014). The clinical reference *H. pylori* genome J99 (NC_000921.1) was used as the phylogenetic root. The following command was executed:

```
parsnp -p {user specified CPU/thread/core number} -r {H. pylori
J99} -d {path to directory containing genomes to analyse} -a 13
-c -x
```

The resulting phylogeny was visualised in FigTree version 1.4.4 (https://github.com/rambaut/figtree) and re-rooted to the midpoint and ordered in decreasing order to improve visualisation. This was then uploaded into phandango version 1.3.0 (Hadfield et al., 2018) alongside the gene presence and absence matrix produced by Roary where the reference was manually appended to create a pan-genome gene presence and absence visualisation.

4.3. Results and discussion

All of the analysis presented in this Chapter was carried out on all patients and samples used within this study (table 4.1). Where appropriate and convenient, in some results figures and/or tables the representative patient 265 sample data were presented. All other patient results can be found in the Appendix (Chapter Ten) or the Appendix directory (https://myntuac-my.sharepoint.com/:f:/r/personal/n0667645_my_ntu_ac_uk/Documents/OneDrive_link?csf=1&e=UuIp26) as stated below. However, full patient and sample results are displayed where convenient.

4.3.1. Contamination detection of non-*Helicobacter pylori* biological sequences

Although the human stomach was once thought to be a sterile environment, this has proven to be untrue. The human stomach hosts a diverse abundance of bacterial taxa and microbiota studies have started to characterise these bacterial taxa and their relative abundance. Recent studies have indicated that there are no consistent microbiome signatures within the human stomach but *Helicobacter, Prevotella, Neisseria, Streptococcus, Bacteroidetes* and *Firmicutes* are notably abundant (Wurm et al., 2018; Chen et al., 2018; Kupcinskas and Hold, 2018). The human stomach has been shown to harbour a diverse abundance of bacterial species with one recent study identifying 110 different species within the stomach and 106 species within the duodenum (Mailhe et al., 2018). It has also been shown that *H. pylori* infection, different disease status of the stomach and use of proton pump inhibitors can change the microbiome of the human stomach (Wang et al., 2018; Parsons et al., 2017; Lopetuso

et al., 2018). As isolation of *H. pylori* population cultures are taken from biopsies within the human stomach, contamination is highly likely and takes the skill of the laboratory team to avoid and detect. This is further complicated by the endoscopy procedure which removes the biopsy sample from the patient through the oral cavity which hosts its own unique, diverse and highly abundant microbiome (Marsh, 2018; Wade, 2013), further increasing the risk of culture contamination when trying to isolate *H. pylori* from clinical biopsies.

For the reasons previously mentioned, it was imperative to detect potential contamination of population sweeps of *H. pylori* at the earliest opportunity. A k-mer raw sequence classification method was used for this purpose as described in Chapter Two section 2.10. The results of this analysis showed that all but two (308A and 326A) samples were contamination free (table 4.1). The two samples that were identified as contaminated did not match to any known bacterial, archaeal or viral genome within the RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/). This suggests that this unknown contaminant is an as yet unsequenced or unknown bacterial, archaeal or viral contaminant or is eukaryotic. Despite 308A and 326A being flagged as contaminated, these samples were passed forward to all analysis steps in order to stress test analysis pipelines and to determine the utility of using contaminated datasets as contamination is a real risk in this type of *H. pylori* population research.

Table 4.1 lists the bacteriophage detected within each sample, as determined by the sequence classifier. This analysis shows that some patient paired samples from different niches harbour the same bacteriophage composition (patients; 77, 322, 444, 565 and 732), but often with varying abundance. Conversely, the majority of paired patient samples (67%) showed differences in phage composition between antrum and corpus (patients; 45, 93, 120, 194, 265, 295, 308, 326, 439 and 495). This suggests that there were different circulating bacteriophage between different niches of the human stomach. This is a novel observation and could potentially be a source of genetic diversity between the populations. However, while bacteriophage can be beneficial to the host bacterium as seen amongst a range of bacterial species (Vale and Lehours, 2018; Harper et al., 2014; Torres-Barceló, 2018; Brussow, Canchaya and Hardt, 2004), no such benefit in terms of virulence or disease association has yet been reported in the literature for *H. pylori* phage (Vale et al., 2015b; Lehours et al., 2011). One study has correlated the *cagA* and *vacA* genotypes with orthologous phage genes (Kyrillos et al., 2016). Other studies have shown geographical clustering of *H. pylori* prophage

(Vale et al., 2017, 2015b). The varying phage composition observed in this study from between niche populations, could suggest that bacteria phage were circulating within but not between niches.

Table 4.1 Contamination detection of non-*Helicobacter pylori* biological sequences

Patient Sweep	Reads mapping to *H. pylori* (%)	Second highest match (%)	Phage detected (%)	Unclassified reads (%)
45A	98.55	*H. acinonychis* (0.3)	*Helicobacter* phage (<0.00)	0.84
45C	97.83	*H. acinonychis* (0.4)	*Helicobacter* phage KHP40 (<0.00)	1.38
77A	98.91	*H. acinonychis* (0.01)	*Helicobacter* phage (0.15)	0.52
77C	99.04	*H. acinonychis* (0.01)	*Helicobacter* phage (0.12)	0.48
93A	99.7	N/A	*Enterobacteria* phage phiX174 sensu lato (<0.00)	0.09
93C	99.71	N/A	*Helicobacter* phage (<0.00)	0.08
120A	98.41	*H. acinonychis* (0.3)	N/A	0.95
120C	99.55	*H. acinonychis* (0.01)	*Dickeya* phage + *Helicobacter* phage (<0.00)	0.19
194A	99.7	N/A	N/A	0.11
194C	99.68	N/A	Stx2-converting phage (<0.00)	0.11
249C	98.63	*H. acinonychis* (0.01)	N/A	0.76
265A	98.71	*H. acinonychis* (0.01)	N/A	0.76
265C	98.67	*H. acinonychis* (0.02)	*Dickeya* phage RC-2014 (<0.00) + *Enterobacteria* phage phiX174 sensu lato (<0.00)	0.78
295A	97.87	*H. acinonychis* (0.02)	*Helicobacter* phage 1961P (0.07) + KHP30 (0.03) + KH40 (<0.00)	1.3
295C	97.82	*H. acinonychis* (0.02)	*Helicobacter* phage 1961P (0.07) + KHP30 (0.03)	1.34
308A	56.17	*H. acinonychis* (0.01)	*Dickeya* phage RC-2014 (<0.00)	42.11
308C	98.55	*H. acinonychis* (0.01)	N/A	0.84
322A	96.02	*H. cetorum* (0.42) + *acinonychis* (0.19)	*Helicobacter* phage 1961P (0.03) + KHP30 (0.05) + KH40 (0.01)	3.76

Patient Sweep	Reads mapping to *H. pylori* (%)	Second highest match (%)	Phage detected (%)	Unclassified reads (%)
322C	95.33	*H. cetorum* (0.5) + *acinonychis* (0.25)	*Helicobacter* phage 1961P (0.03) + KHP30 (0.04) + KH40 (0.01)	4.47
326A	59.62	*Actinobacteria* (1.42)	*Helicobacter* phage (0.07) + *Dickeya* phage (<0.00) + *Enterobacteria* phage phiX174 *sensu lato* (<0.00)	38.33
326C	98.8	*H. acinonychis* (0.01)	*Helicobacter* phage (0.11)	0.65
439A	96.99	*H. cetorum* (0.17)	Unclassified *Siphoviridae* (<0.00) + *Myoviridae* (<0.00) + *Helicobacter* phage 1961P (<0.00)	1.95
439C	97.27	*H. cetorum* (0.15)	*Helicobacter* phage 1961P (<0.00)	1.76
444A	98.02	*H. cetorum* (0.28)	*Helicobacter* phage 1961P (<0.00) + KHP30 (<0.00)	0.86
444C	98.42	*H. cetorum* (0.16) + *acinonychis* (0.02)	*Helicobacter* phage 1961P + KHP30 (<0.00)	0.68
495A	99.37	*E. coli* (0.01)	N/A	0.36
495C	99.51	N/A	*Helicobacter* phage (<0.00)	0.23
537A	98.61	*H. acinonychis* (0.01)	N/A	0.77
565A	99.45	*E. coli* (0.01)	N/A	0.26
565C	99.61	N/A	N/A	0.14
621A	98.45	*H. acinonychis* (0.86)	*Enterobacteria* phage (<0.00)	0.22
732A	99.56	*H. acinonychis* (0.02)	*Helicobacter* phage (<0.00)	0.17
732C	99.57	*H. acinonychis* (0.02)	*Helicobacter* phage (<0.00)	0.16

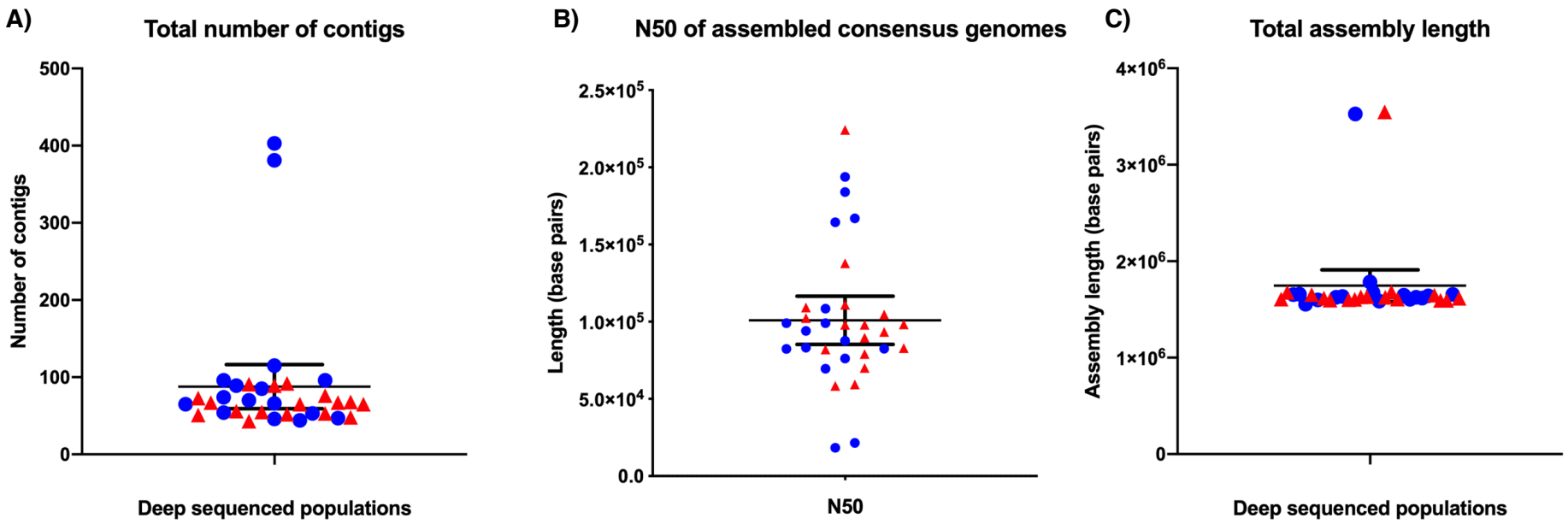
Contamination detection was detected as described in Chapter Two, section 2.10. Table denotes the percentage of reads mapping to complete bacterial, archaeal and viral genomes reference genomes in the RefSeq database as of 18th October 2017. Reads that did not map to any reference genomes were displayed in the last column. Samples with suspected contamination are highlighted in orange.

4.3.2. Quality statistics for deep sequenced *de novo* assembled consensus genomes

All deep sequenced populations were assembled into a consensus assembled genome meaning that all curated sequencing reads from the population were used as input (Chapter Two, section 2.9). Before any bioinformatics post-processing was conducted, assembly statistics were calculated (Chapter Two, section 2.12) and are presented in figure 4.2. Importantly, this figure shows that there are no assembly quality differences between both sequencing run sets (denoted by colour) and between antral and corpus samples (denoted by shape).

With the exclusion of the contaminated samples (n=31), the mean number of contigs was 89.48 (95% CL: 59.10 – 119.9) while the N50 mean was 94,245 (95% CL: 81,016–107,474) and total assembly mean length of 1,630,573 (95% CL: 1,615,608 – 1,645,539). The lower than expected mean genome size of approximately 1.63 Mbp compared to the reference genome (*H. pylori* J99) size of approximately 1.67 Mbp was considered acceptable due to the *de novo* assembly approach (*i.e.* a non-reference based assembly) and the use of high quality curated read sets (Chapter Two, section 2.9). Furthermore, other recent *H. pylori* whole genome studies report similar and often lower *de novo* assembled genome lengths of *H. pylori* (Kumar et al., 2012; Ali et al., 2015; Montano et al., 2015) than achieved within this study. Pilot testing on this dataset showed that in most cases, using raw sequencing reads resulted in only a small increase in total assembly length, but in rare cases also reduced the total genome length (data not shown). It is thought that the deep sequencing methodology utilised within this study allowed for removal of low-quality reads with very little effect on the final genome assemblies, probably due to the high depth of coverage. Therefore, high quality curated deep sequencing read sets were used to reconstruct consensus whole genome assemblies and this was thought to produce a more accurate consensus assembly, especially for downstream analysis steps such as paired whole genome sequence alignments.

Figure 4.2 Quality statistics for all deep sequenced de novo assembled consensus genomes



This figure depicts the quality metrics of deep sequenced consensus assembled genomes. Figure A – total number of contigs; figure B – N50 of assembled consensus genomes; figure C – total assembly length. This dataset includes the contaminated samples (n=33) 308A and 326A for reference purposes. Red = sequencing run 1, blue = sequencing run 2, circle = antrum, triangle = corpus. The mean and 95% CL are displayed as horizontal bars.

4.3.3. Whole genome depth of coverage

For the deep sequencing dataset, an average coverage depth of 231X was achieved. Each sample average coverage is denoted in table 4.2 along with details on the proportion of the genome covered by ≥100X coverage. The fraction of specific sample genomes covered by X coverage is depicted in figure 4.3 where a characteristic 'waterfall' shape is observed, suggesting little variation around the mean of each sample and small proportions of the genome at coverage extremes.

The representative sample for the deep sequencing dataset 265 (A and C) followed in this Chapter had a relatively even coverage distribution across the length of the genome, except for coverage spikes towards the close (figure 4.4). This was seen across all samples (Appendix, figures 11.4.1 – 11.4.18) and was due to short, high coverage contigs that were assumed to be assembly artefacts.

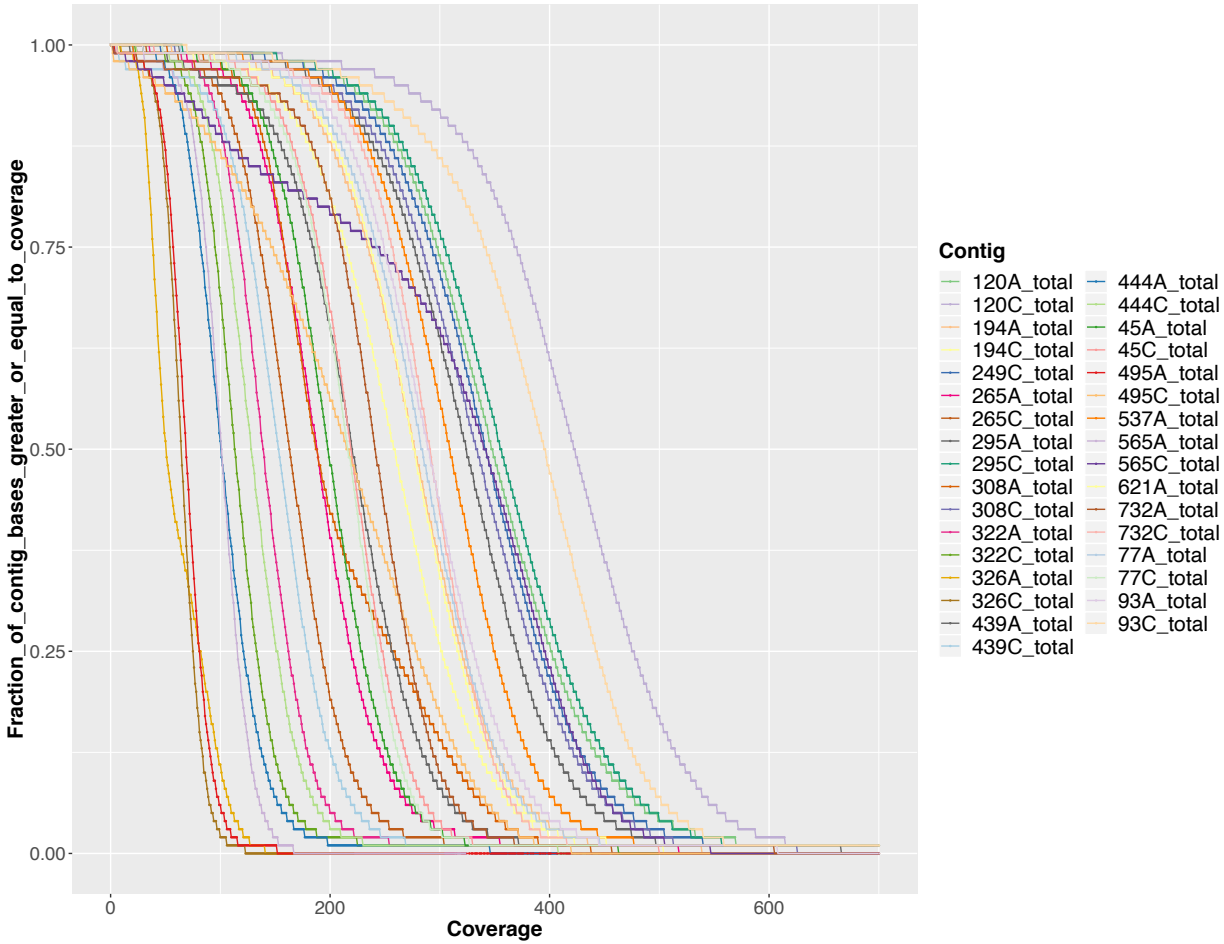
Table 4.2 Deep sequencing samples with average coverage and percentage of genome covered at greater than or equal to 100X

Sweep	Average coverage:	Bases with coverage ≥ 100
45A	198	98%
45C	219	98%
77A	284	99%
77C	216	98%
93A	290	99%
93C	396	99%
120A	348	99%
120C	423	99%
194A	277	99%
194C	259	98%
249C	342	99%
265A	188	97%
265C	163	94%
295A	324	99%
295C	356	99%
308A	188	98%

Sweep	Average coverage:	Bases with coverage ≥ 100
308C	334	99%
322A	139	90%
322C	113	69%
326A	50	11%
326C	64	2%
439A	220	95%
439C	153	91%
444A	100	51%
444C	128	82%
495A	68	6%
495C	217	87%
537A	309	99%
565A	101	52%
565C	342	89%
621A	277	98%
732A	241	96%
732C	291	98%

Coverage statistics were calculated as described in section 4.2.1.2. All samples are displayed with corresponding average and proportion of the genome covered by ≥ 100X coverage.

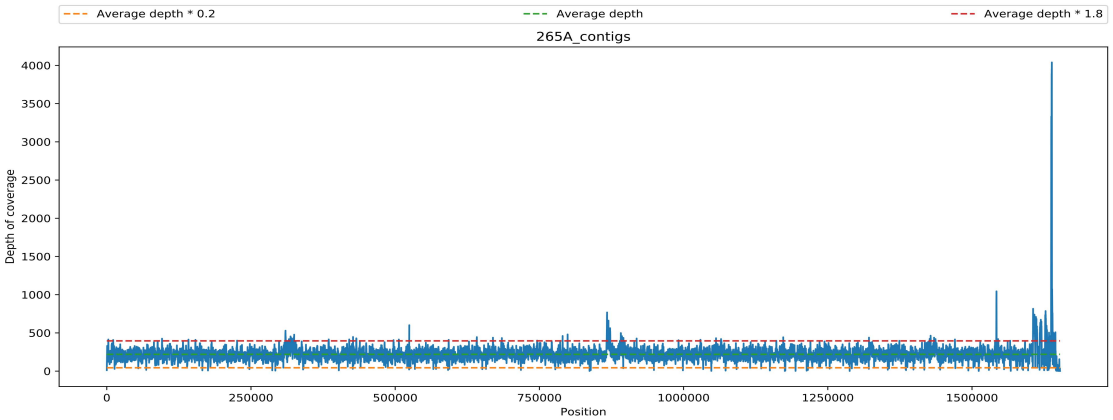
Figure 4.3 Coverage distribution across all deep sequenced samples



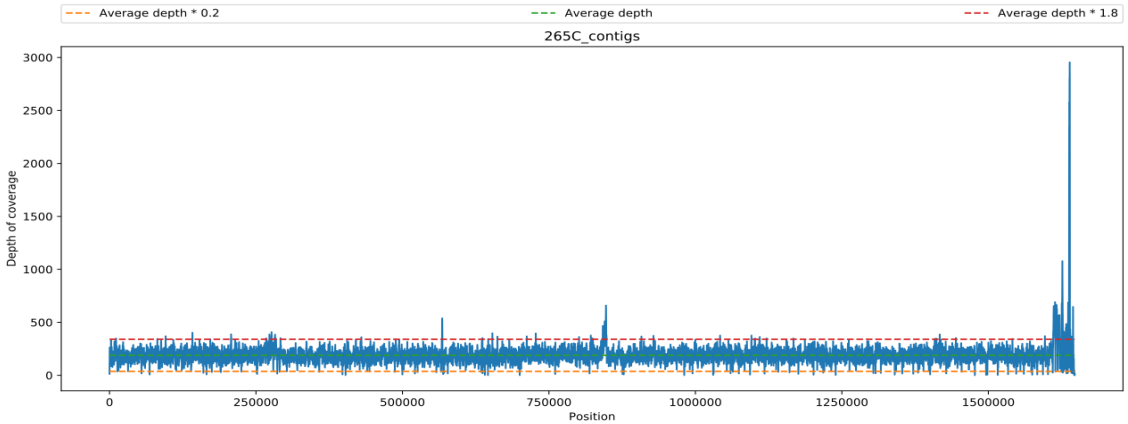
Coverage distribution displayed at each fraction of the genome for all samples. This figure was plotted in R from the statistics produced from mosdepth (Chapter Four, 4.2.1.2). A cut off at 700 bp was enforced because only a very small fraction of some genomes had greater than this coverage.

Figure 4.4 Coverage distribution across the genome length of sample 265A and 265C

A)



B)



This graph was drawn through WeeSam as previously described (Chapter Four, section 4.2.1.2) and depicts the coverage distribution across the length of the genome sequences of samples 265A (figure A) and 265C (figure B). Average coverage is plotted as a dashed horizontal green line and the multiple of 0.2 and 1.8 of the mean are plot as horizontal dashed red lines. Corresponding figures to all samples used within this study can be found in the appendix (figures 11.4.1 – 11.4.18).

4.3.4. Within niche polymorphic diversity

Within niche diversity was detected using two separate but similar pipelines to detect common and minor allelic diversity (Chapter Four, section 4.2.1.1). This approach was taken in order to highlight a strong evidence-based detection of common allelic variation while relaxing some parameters to detect more diversity within the datasets (minor allelic variation). Due to the higher thresholds of the common allelic detection pipeline, called positions by this method are assumed to be moving towards fixation within the sample due to the higher proportion of this variant within the population. Conversely, the allelic detection pipeline calls polymorphic positions at much lower frequencies, mainly attributed to the relaxation of the SAF (number of alternative reads mapping in the forward direction) and SAR (number of alternative reads mapping in the reverse direction) filters. This allowed for the detection of low frequency polymorphic sites that were potentially more recent polymorphisms or were variants that were persisting within the population but not moving towards fixation. One reason for this could be the lack of a selection pressure acting on the population resulting in a mixture of both fit and unfit populations persisting within the environment. Furthermore, the common and minor allelic pipelines were designed in order to detect against false positives at the expense of losing some true positives (common allelic variation) and the increase of true positives at the expense of potentially more false positives (minor allelic variation). However, in Chapter Five, it is further discussed how the minor allelic variants were matched with SNPs identified from single colony isolates and how well these different methods compared. As will be described later, a good crossover of detection was observed suggesting that the minor allelic detection was determined to be a reliable detection method. Furthermore, this methodology was adapted from work by Lieberman *et al.* (2014), where they tested both real and simulated datasets and show that identification of true positive polymorphic positions was high while the carefully tuned parameters reduced false positive calls.

All samples were found to harbour both common and minor allelic diversity (table 4.3). Not surprisingly, common allelic variants were detected more rarely than minor allelic variants, with a median increase of 12.4 $\times$ from the common to the minor allelic detection method. This was most likely down to parameters used for the common allelic variation calling pipeline where alternative base calls in relation to the reference had to observe $\geq$ 15 reads mapping in both the forward and reverse direction. Although this might appear very strict, it allowed for the calling and identification of highly polymorphic

sites that were proliferating within the populations, potentially moving to fixation within the populations or evidence of a mixed infection of two or more *H. pylori* strains where the total number of these variants are high.

Table 4.3 Total number of common and minor allelic positions within all samples

Patient/sample location:	Number of sites with common allelic variation	Number of sites with minor allelic variation
45A	2	24
45C	2	38
77A	7	39
77C	3	43
93A	201	262
93C	2	12
120A	4	31
120C	2	25
194A	3	18
194C	9	40
249C	7	51
265A	4	85
265C	6	124
295A	8	55
295C	7	51
308A	7	67
308C	2	46
322A	2	73
322C	50	384
326A	1	612
326C	1	483
439A	64	152
439C	2	98
444A	11	261
444C	14	147
495A	7	248
495C	80	225
537A	11	44

Patient/sample location:	Number of sites with common allelic variation	Number of sites with minor allelic variation
565A	6	498
565C	7765	12885
621A	8	123
732A	115	1034
732C	17	104

Colour code: light green: 1 –10 dark green: 11 – 20; light orange: 21 – 60; dark orange: 61 – 100; light red: 101 – 200; medium hue red: 201 – 400; dark red: > 401. Common allelic variation filtering of polymorphic positions where ≥ 15 alternative base calls have to be identified in both the forward and reverse direction and the alternative fraction must be above 3% of total reads mapping to support a call. Minor allelic detection filters on the minor allelic fraction of > 3% only. Both methods require a mapping quality of Q34 and a base quality of Q30 as described in section 4.2.1.1.

It is possible to comment on each gene expressing polymorphic diversity, however this is impractical and so only key genes and genes with the highest levels of diversity as well as genes with observable diversity across multiple samples will be discussed here. Nevertheless, this study offers a very rich dataset and further inspection of this dataset is welcomed by other researchers.

Excluding hypothetical proteins, the highest common allelic variation was observed within outer membrane associated genes (bab, hop, hof, hom, hef, hor, frp and lpt), with polymorphisms found among 73% (n=24 samples) (figure 4.5; Appendix figure 10.4.19). There are a wide range of outer membrane associated genes and protein families, where OMPs are thought to make up approximately 4% of the *H. pylori* coding genome (Alm et al., 2000a). Numerous studies have shown that OMPs are highly diverse and polymorphic with some strains harbouring different numbers of OMP genes as well as consisting of a mix of protein families (Kim et al., 2016; Solnick et al., 2004; Bauwens et al., 2018; Alm et al., 2000a; Oleastro et al., 2010; Pride, Meinersmann and Blaser, 2001). However, most studies have looked at polymorphic variation between single strains of *H. pylori* from between; patients, geographical regions, sequential isolates from animal models and familial isolated strains to reach these conclusions (Kim et al., 2016; Solnick et al., 2004; Yamaoka et al., 2002a; Akeel et al., 2019; Hansen et al., 2017; Furuta et al., 2015; Liu et al., 2015). The comparative genetics approaches

of such studies have contributed to the identification and understanding of OMP polymorphic diversity, but polymorphic diversity has yet to be shown within populations taken from the same time point, despite the identification of OMP phase variation and gene conversion (Solnick et al., 2004; Hansen et al., 2017; Liu et al., 2015; Yamaoka et al., 2006; Braga et al., 2019). This is perhaps hampered by the difficulty and increased workload in isolating single colonies from population sweeps, the availability of paired biopsy samples from the same patient and the increased sequencing costs such investigations incur.

The deep sequencing and analysis methodology adopted within this study reveals a snapshot of vast polymorphic diversity within OMP associated genes at a single time point within populations taken from single biopsies (figure 4.5). This study also demonstrates the resolution and power of this analysis method. Such high polymorphic diversity within OMP related genes has many potential biological implications, namely; adaptation to new hosts/environments, persistence of chronic infection and disease progression (Yamaoka et al., 2006; Oleastro and Ménard, 2013; Akeel et al., 2019; Braga et al., 2019; Furuta et al., 2015). Therefore, a continually diverse population with polymorphic diversity within OMPs could add to the picture of how chronic infection of *H. pylori* is established lifelong.

Some of the most frequently identified common allelic variant sites were within virulence related genes including *vacA* paralogue (HP0289, n=5 samples), *babA* (n=5 smaples), and *cagA* (n=4 samples). The *vacA* gene is present in almost all *H. pylori* strains but is known to have a number of polymorphic types defined within the signal (s1a-c/s2), intermediate (i1/i2/i3) and middle (m1/m2) regions (Van Doorn et al., 1998; Rhead et al., 2007; Chauhan et al., 2019). However, different allelic variants of *vacA* have different effects on colonisation, virulence and disease pathologies (Winter et al., 2014; Sheikh et al., 2018). Han *et al.* (1998), used restriction fragment length polymorphism and PCR techniques to investigate between and within patient diversity of the *vacA* genotype. These studies showed that each patient harboured a unique strain of *H. pylori* but revealed that there was no observable within patient *vacA* diversity between the antrum and corpus. The results presented in this study show that *vacA* presence can vary between antrum and corpus *H. pylori* populations from the same stomach (Chapter Three, tables 3.2A-B). Within niche genetic diversity of *vacA* was also observed but was not a common observation (4/33 deep sequenced populations; Appendix, figure 10.4.20).

This study revealed extensive within niche polymorphic diversity of *vacA* paralogues (figures 4.5 and 4.6). Paralogues of *vacA* have been described by other studies and have been shown to play a role in collagen degradation allowing access to essential amino acids and have also been shown to exacerbate the development of gastric ulceration (Castillo et al., 2008; Beswick, Suarez and Reyes, 2006; Kavermann et al., 2003). Again, such within niche diversity could help explain how *H. pylori* is able to colonise and persist as a lifelong infection and why some patients go onto develop disease while others do not. Furthermore, varying levels of variation, potentially attributed to various *vacA* allele paralogues could potentially aid in virulence by a balance of other interacting virulence genes. However, such intra population *vacA* paralogue diversity would need to be investigated further to better understand the biological significance of these extensive polymorphic profiles. It must also be noted that analysis of the within niche minor allelic diversity does reveal some patients with *vacA* genetic variability (Appendix, figure 10.4.20).

The *cagA* gene has been shown to be genetically diverse between individual patients and geographical regions (Peters et al., 2001; Olbermann et al., 2010b). Patients infected with *cag*PAI positive *H. pylori* strains are more likely to develop a range of diseases such as gastric ulcers and adenocarcinoma (Park et al., 2018; Backert and Tegtmeyer, 2017). However, what is not well documented prior to this study is evidence of within patient genetic diversity of the *cagA* gene. In addition to within host genetic diversity of the *cagA* gene, this study also reveals between niche *cagA* diversity (figure 4.5; patients 93 and 732). As this observation is not limited to one patient, this does not appear to be a special case or an outlier, especially considering this detection is based on stringent calling parameters of the common allelic variant detection methodology. The biological significance of this is yet to be determined. However, as *cagA* is a known virulence associated gene, within and between niche population diversity could result in variations in virulence across the stomach, potentially playing an important role in disease development.

Another notable common allelic gene product was the DNA-directed RNA polymerase beta subunit (HP1198; *rpoB*). Certain mutations within the *rpoB* gene of *H. pylori* have been shown to increase resistance to rifamycins (Heep et al., 2000b; a; Hays et al., 2018; Nishizawa et al., 2011). Therefore, polymorphic diversity within *rpoB* is a concerning observation and could indicate intra population variation in rifampicin

resistance within populations. In some cases, this observation might help to explain eradication therapy failure within patients whereby underlying resistant strains persist within the population. The samples with polymorphisms within the *rpoB* gene were samples 93A, 194C, 439A and 732A. In relation to the antibiogram results for these *H. pylori* populations, all except for sample 439A were shown to have two zones of inhibition, resistant colonies within the zones of inhibition, and or resistance to rifampicin (Appendix table 10.3.1). Therefore, polymorphisms within the *rpoB* gene from deep sequencing of these *H. pylori* populations is potentially related to phenotypic variation in rifampicin resistance. Despite the *H. pylori* population form sample 439A not exhibiting an obvious second zone of inhibition, resistant single colonies were observed outside the recorded inhibition zone (Appendix table 10.3.1). Other *H. pylori* populations from patient samples recorded as observing resistant colonies or second zones of inhibition for rifampicin were not displayed as polymorphic in the common allelic detection pipeline (Chapter Three, figure 3.1; figure 4.5; Appendix, table 10.3.1). However, additional samples were detected as showing polymorphic variation within the *rpoB* gene in the minor allelic pipeline for samples 77C, 265C, 322C, 326A, and 439C (figure 4.6). Sample 322C also showed an additional zone of inhibition for rifampicin where all other samples except for 439C presented with resistant single colonies within the recorded zone of inhibition. Therefore, there is a good match between *rpoB* polymorphic diversity and diversity in phenotypic rifampicin resistance. However, some patient samples showed the presence of resistant colonies within the zone of inhibition for rifampicin but were not shown as genetically variable for the *rpoB* gene. This suggests that other genes maybe further involved in rifampicin resistance. In relation to sample 326A which was shown to have no *rpoB* polymorphisms, the zone of inhibition was comparable to other samples harbouring polymorphic diversity suggesting that the resistance genotype was present within this sample and had fixed in the population resulting in no observable *rpoB* polymorphisms. Sample 439A showed polymorphisms within the *rpoB* gene but no resistant colonies or second zone of inhibition were seen. Furthermore, this sample had a comparably more sensitive zone of inhibition for rifampicin suggesting that not all polymorphic diversity or mutations within the *rpoB* gene results in higher rifampicin resistance.

What is abundantly clear between the common and minor allelic detection pipeline is that much more genetic diversity is detected from the latter pipeline (figures 4.5 – 4.6; Appendix figures 10.4.19 – 10.4.20). The minor allelic detection pipeline is able to detect minor sub-populations of genetic variation. From the minor allelic variation

pipeline (figure 4.6), the most polymorphic variation across samples was identified within the following genes/gene products; OMP associated (*hopQ*, *hopL*, *frpB*, *hopI*, *hefA*, *hopC* and *horD*) (23/33), Methyl-accepting chemotaxis protein TlpB (16/33), type I and III restriction modification associated (HP0464, HP0846, HP1371 and HP1521) (15/33), *vacA* paralogs *imaA* and *vlpC* (13/33), *cagY* of the *cag*PAI (12/33), Sialic acid-binding adhesin SabA (11/33), lipopolysaccharide biosynthesis protein (11/33) and glutathione-regulated potassium-efflux system protein (11/33). Again, high genetic diversity was observed within and between samples for OMP related and *vacA* paralogue genes (figure 4.6). However, additional genes were also identified with a genetically diverse signature across multiple samples as previously described.

Methyl-accepting chemotaxis associated genes were identified as the second most allelic gene showing within niche genetic diversity between multiple patients (figure 4.6). Chemotaxis is the process in which bacteria sense the external environment and are able to move towards or away from certain chemoattractants or chemorepellents respectively. To understand the significance of the genetic diversity observed within this study of methyl-accepting chemotaxis genes, further investigations into the nature of the chemotaxis attractant/repellent would need to be determined or further characterisation of the specific gene would need to be investigated. This would allow a more conclusive determination of the importance of this observed diversity. However, it could be speculated that genetic diversity of chemotaxis associated gees could aid in the survival of H. pylori strains. For example, it has been shown that the chemotaxis receptor tplB gene is essential for the pH taxis where acidic pH is a chemorepellent, allowing H. pylori strains to move towards more favourable, less acidic conditions (Croxen et al., 2006). Furthermore, the chemotaxis receptor tplA gene has been shown to sense arginine, bicarbonate, and acid (Cerda et al., 2011; Huang et al., 2017). Genetic diversity could result in more sensitive or insensitive chemotaxis, allowing for colonisation of different niches.

Another notable highly polymorphic group of genes identified by the minor allelic analysis pipeline were restriction-modification system associated. Furuta *et al.* (2015a), observed similar genetic diversity among restriction-modification associated genes by a comparative genetics approach of single colony isolates obtained from five different families. However, this project reveals the within patient population diversity rather than a between strain diversity from different members of a family. Furthermore, this study reveals a much higher extent of restriction-modification gene diversity, likely due to the

sampling technique of capturing diversity at a population level. Nonetheless, this observation is supported by that of Furuta *et al.* (2015a), suggesting an important role for restriction-modification system genetic diversity. Other studies have also observed restriction-modification system diversity between strains taken from different patients (Kojima et al., 2016; Nobusato, Uchiyama and Kobayashi, 2000; Aras et al., 2002; Yahara et al., 2016). Restriction-modification systems within bacteria have been described as a potential innate immune system that confer protection against invading foreign DNA such as that delivered from bacteriophages (Tock and Dryden, 2005; Vasu, Nagamalleswari and Nagaraja, 2012). Despite the ability to differentiate between self and non-self DNA, it has been shown that restriction-modification systems do not pose a barrier to homologous recombination (Bubendorfer et al., 2016a). Therefore, the observed within patient genetic diversity should not impact on the ability of subpopulations to recombine with one another suggesting another potential importance to this observed diversity. It has been shown that restriction-modification systems can play a role in global and specific gene expression (Vitoriano et al., 2013; Srikhanta et al., 2011; Furuta et al., 2014) while other studies have shown a role in adhesion and virulence (Lehours et al., 2007; Takeuchi et al., 2002; Ando et al., 2010; Gorrell and Kwok, 2017; Kumar et al., 2018; Gauntlett et al., 2014). Taken together, the observed within/between niche and between patient diversity of restriction-modification system associated genes could play a role in niche and host adaptation and may also play a role in chronic infection in addition to virulence.

Paralogs of the *vacA* gene have been shown to play a role in host colonisation and modulation of the host immune system (Ssuse, Castillo and Ottemann, 2012). Therefore, these highly variable genes could help in persistence of infection.

The *cagY* gene has been shown to be genetically diverse and is thought that ***cagY*** recombination can make this less or more immunogenic (Barrozo et al., 2016).

Lipopolysaccharide biosynthesis associated genes were also found to be a highly polymorphic within *H. pylori* populations across many different patients. The lipopolysaccharide is considered highly immunogenic and is often considered a target for vaccine development (Santos et al., 2010; Conde-Álvarez et al., 2013; Zhao et al., 2018). However, human sera antibodies have been shown to target specific *H. pylori* lipopolysaccharides but are generally less immunogenic due to host cell mimicry by blood group O-chains (Monteiro et al., 2011). Additionally, the lipopolysaccharide has

been associated with biofilm formation, adhesion, virulence, immune system evasion, increased antibiotic resistance and is essential to cell structure and integrity (Wong et al., 2016; Chang et al., 2011; Yu et al., 2016; Moran, 2008; Li et al., 2016; Stein et al., 2017; Khamri et al., 2005). Due to these wide-ranging functions, the extent of polymorphic diversity observed within this study is perhaps not surprising. However, the findings have many implications such as on the developing picture of how *H. pylori* is able to persist as a lifelong infection. Such genetic diversity of lipopolysaccharide biosynthesis associated genes might help enable this chronic infection. Another implication is in vaccine development, as lipopolysaccharide biosynthesis associated genes might not be an ideal vaccine target due to the intra- and inter-niche and interpatient genetic diversity observed here.

Glutathione-regulated potassium-efflux system protein KefB is involved in cellular homeostasis by regulating toxic electrophilic compounds and in doing so, modulating cytoplasmic pH (Roosild et al., 2010). This suggests a putative role in pH environmental adaptation which could be facilitated or modulated by genetic diversity. Furthermore, *kefB* mutations have been flagged as a gene candidate for *H. pylori* clarithromycin resistance, although this is not yet proven (Binh et al., 2014).

Figure 4.5 Common allelic variant gene products

Heatmap of most common allelic genes (polymorphic genes shared by two or more different samples). Heatmap was created using ggplot2 and the RColorBrewer through the R statistical software version 3.5.1 (R Core Team, 2018). Patient antral and corpus polymorphic diversity can be compared by looking between samples. Number of different polymorphic genes/associated gene products can be identified per sample. Colour intensity indicates a higher number of polymorphic positions within these gens/associated genes by product. This approach tries to keep together observed polymorphic diversity within genes by gene name and where no gene name is provided (by PROKKA) a unique gene number for each patient sample is provided. Sample 565C was excluded due to the extreme variation observed. An undocketed heatmap including sample 565C can be observed in the Appendix (figure 10.4.19) with full resolution images within the appendix directory (https://myntuac-my.sharepoint.com/:f:/r/personal/n0667645_my_ntu_ac_uk/Documents/OneDrive_link?csf=1&e=UuIp26).

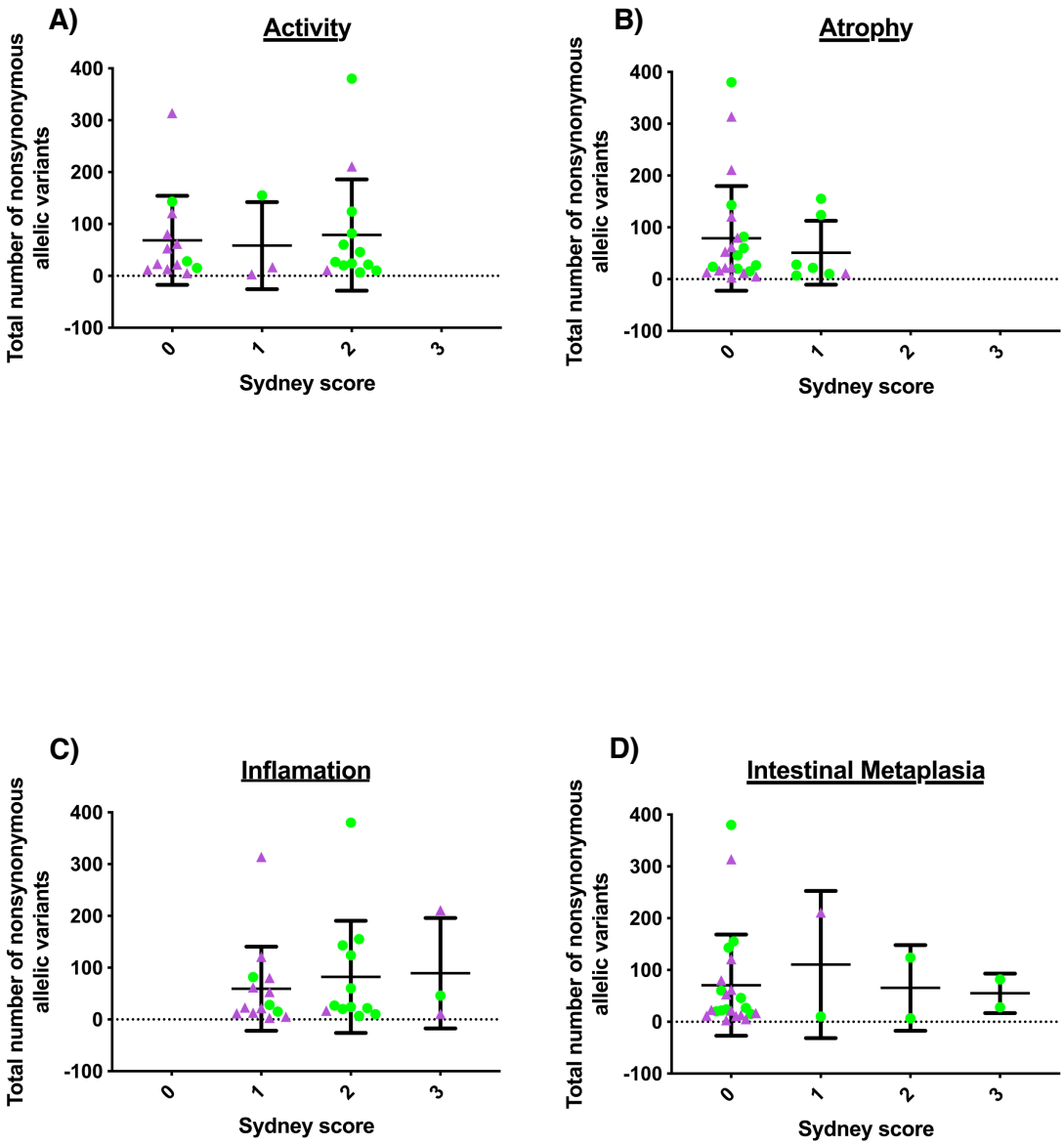
Figure 4.6 Minor allelic gene products identified between six or more samples

Heatmap of minor allelic variant genes/gene products shared between six or more different samples. Heatmap was created using ggplot2 and the RColorBrewer through the R statistical software version 3.5.1 (R Core Team, 2018). This figure can be interpreted as previously described in figure 4.5. Sample 565C was excluded due to the extreme variation observed. A heatmap displaying all minor allelic genes and the inclusion of all samples can be observed in the Appendix (figure 10.4.20). A full resolution image can be found in the Appendix directory (https://myntuac-my.sharepoint.com/:f:/r/personal/n0667645_my_ntu_ac_uk/Documents/OneDrive_link?csf=1&e=UuIp26).

After endoscopy the physician scores histological sections from stomach biopsies for indicators of disease severity, namely activity, atrophy, inflammation and intestinal metaplasia. This study hypothesised that adverse conditions (such as high levels of inflammation) would select for increased *H. pylori* population diversity or increase the selection of environment specific genes. In order to test these hypotheses, the total number of nonsynonymous mutations at polymorphic sites were recorded for each specific environment and corresponding disease severity (figure 4.7). No significant associations were found between the total number of nonsynonymous mutations at polymorphic positions and disease severities.

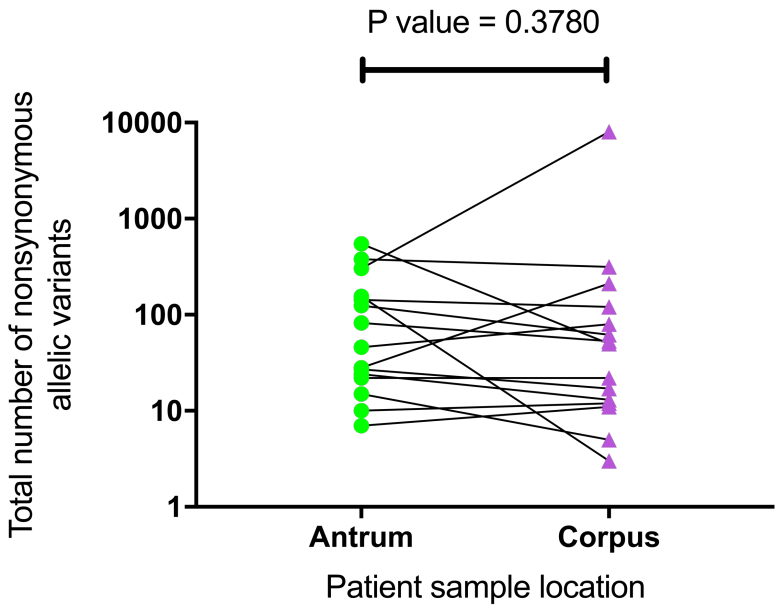
Additionally, total numbers of nonsynonymous mutations identified from within niche polymorphic diversity from the antrum and corpus of individual patients were compared via a paired t-test (figure 4.8). This analysis was performed to investigate whether there were differences in the number of nonsynonymous polymorphic mutations between the antrum and corpus. No significant differences in the number of nonsynonymous mutations were seen between the antrum and corpus populations of the same stomach at a single time point. This suggests a comparable level of selection and diversity within the separate niches taken from a single time point. However, patients 93, 322, 565 and 732 show big variation of total numbers of nonsynonymous mutations between the antrum and corpus populations which suggests that antrum and corpus differences might be patient specific.

Figure 4.7 Total number of nonsynonymous within niche mutations



Total counts of nonsynonymous mutations across different niche environments of activity (figure A), atrophy (figure B), inflammation (figure C) and intestinal metaplasia (figure D) with varying severity as described by the updated Sydney scoring method (Dixon et al., 1996; Stolte and Meining, 2001). This figure was created using a one-way ANOVA statistical test (GraphPad Prism version 8.1.2). No statistically significant associations (P value <0.05) were observed for any condition or severity of disease. Green circles represent antrum population samples while purple triangles represent corpus population samples.

Figure 4.8 Total number of nonsynonymous within niche polymorphic variants between the antrum and corpus of paired patient samples



Paired t-test of within niche nonsynonymous polymorphic variants found within the antrum (green circles) and corpus (purple triangles) populations taken from the same stomachs. Only patients with paired antrum and corpus populations were used.

4.3.5. Between niche genetic diversity

Between niche genetic diversity of individual patients was investigated using a multitude of techniques that taken together, resulted in an informative and reliable dataset. Firstly, a whole genome alignment of antral and corpus consensus genomes was conducted to identify alignment SNPs. A second approach was used to validate these alignment SNPs by mapping niche specific reads to the consensus genome of the opposite niche. The advantage of this duel methodology was to retain the consensus or majority base calls as determined by the consensus genome alignment method while removing alignment SNPs that were potentially artefacts of assembly reconstruction and thus not true SNPs or biological sequence. This phenomenon is usually confined to the start and end of assembled contigs and is due to the assembler being unable to resolve bases or find a continuous string of bases/reads resulting in a contig break, usually in regions of lower or higher sequence coverage. These methods were combined with a nucleotide basic local alignment search tool (BLASTN) identity score to further identify genetic differences between paired antral and corpus consensus genomes as this method takes into consideration sequence insertion and deletions as well as regions with multiple nucleotide polymorphisms (figure 4.9; Appendix, table 10.4.1; Appendix,

figures 11.4.21 – 11.4.35). This combined analysis approach shows that there are few regions of 100% nucleotide identity between paired antrum and corpus consensus genomes. Furthermore, there was good agreement between the alignment and alternative niche mapping methodologies (red SNPs) and where there was disagreement with the alignment only identification (black SNPs) these were located mainly around contig breaks, as predicted. Mapping only SNP locations (teal SNPs) were also identified, but often in lower fractions. These were thought to be identified due to minor allelic diversity at these positions that were not filtered out due to the mapping thresholds and thus were not the consensus base call at these positions. Therefore, this combined analysis methodology was successful and resulted in a subset of alignment SNPs that were of higher confidence of being *true* SNPs between the aligned genomes. All SNPs tagged by these methodologies were colour coded and mapped to their exact locations between the aligned genomes (figure 4.9) and numbered in ascending order from the start of the largest to the smallest contig sequence blocks. These SNPs were numbered so that individual SNPs or groups of SNPs could be identified and cross-referenced to determine whether each SNP was from a coding or non-coding section of the genome. If a SNP related to a coding region, the gene annotation was determined.

Noticeable gaps or regions with <95% BLASTN identity were identified between all aligned genomes except for samples 93A (population consensus reference) – 93C (Appendix, figure 10.4.23A), 194A - 194C (population consensus reference; Appendix figure 10.4.25B), 295A - 295C (population consensus reference; Appendix figure 10.4.27B), 308A - 308C (population consensus reference; Appendix figure 10.4.28B) and 439A - 439C (population consensus reference; Appendix figure 10.4.31B) where gaps were only confined to the smaller contigs towards the end of each genome. These gaps only found towards the end of the genome are potentially attributed to sequence artefacts as a result of sequence assembly as previously discussed. Sequence gaps between aligned genomes suggests variability of gene content and could be an additional determinant of between niche variability.

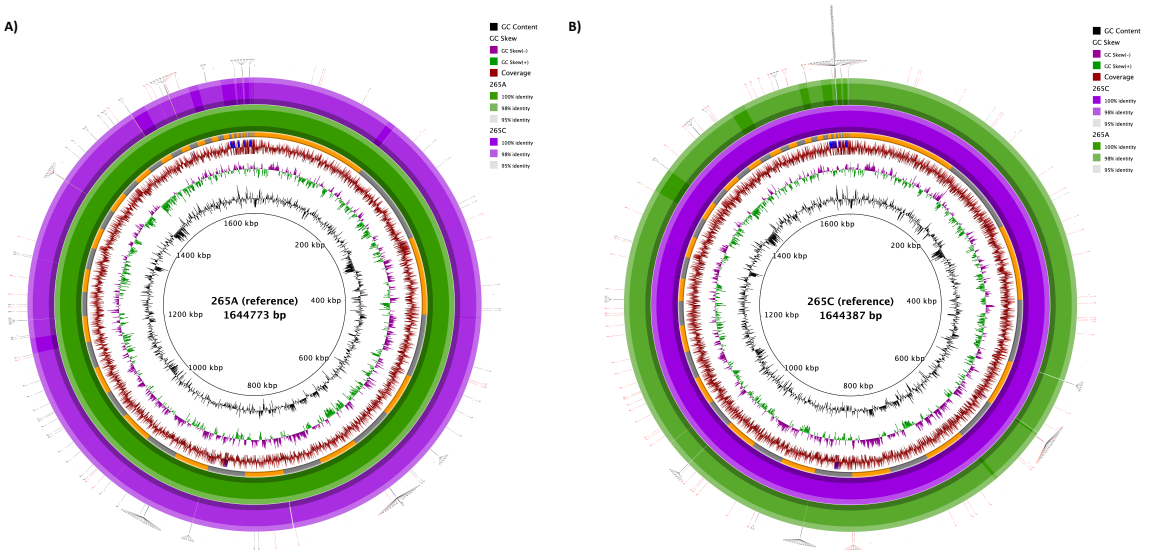
In addition to sequence presence and absence, hotspots of genetic diversity between aligned genomes can be identified by high density SNP regions such as SNPs 29 – 49 (figure 4.9A) which were located outside of a coding sequence. Outside of these hotspots of genetic diversity, alignment SNPs were uniformly distributed across the genome. In terms of bioinformatics analysis, this is an important observation as an

uneven distribution located on a single contig, could be an indicator of assembly bias, a SNP calling artefact or a region of homologous recombination. This is because spontaneous genetic mutations occur randomly across the genome thus distribution of such mutations should not be confined only to a single stretch of the genome (Rosche and Foster, 2000). However, selection pressures might act on a strain that drive the selection for, or loss of, certain genetic mutations within the population, resulting in genes with increased genetic diversity (Didelot et al., 2016). But, considering the high natural mutation rate of *H. pylori* strains from a population level, it would be alarming to observe diversity confined to a single, small stretch of DNA (Falush et al., 2001a).

Random mutation can have a significant effect on strain fitness and a mutant can rise to dominance if advantageous (positive selection) or drift out of the population if disadvantageous (negative selection). Therefore, in relation to within niche genetic variation, the most dominant polymorphic variant is likely to be the most successful at the sampling time point. For this reason, determination of these alignment SNPs can help aid in the identification of highly variable genes with potential niche specific adaptations. Furthermore, genes observing high levels of genetic diversity between populations could have wide ranging implications, such as niche specific adaptation, virulence and persistence of chronic infection.

Despite the known presence of sequencing read contaminants of samples 308A and 326A the same methodologies were applied to these samples to determine the utility of using contaminated *H. pylori* datasets and to stress test the analysis pipelines. Unexpectedly, the combined analysis pipeline revealed that the alternative niche aligned fully to select complete contigs, while other contigs remained alignment free or showed very small segments of genome alignment. It is hypothesised that the contaminant was assembled almost completely independent of the target *H. pylori* by the SPAdes assembler, presumably due to the contrasting GC percentage differences of the contaminant reads. This is evidenced by the GC content ring depicted in figures 11.4.28A and 11.4.30A (Appendix). This analysis shows that there is a potential application to remove these unaligned contigs from the dataset to better resolve the target *H. pylori* consensus sequence. However, this additional processing step fell outside the scope of this project and the presence of short length alignments to the unaligned contigs has unforeseeable downstream analysis consequences. Nevertheless, such a methodology could be applied in situations where re-sequencing is not possible.

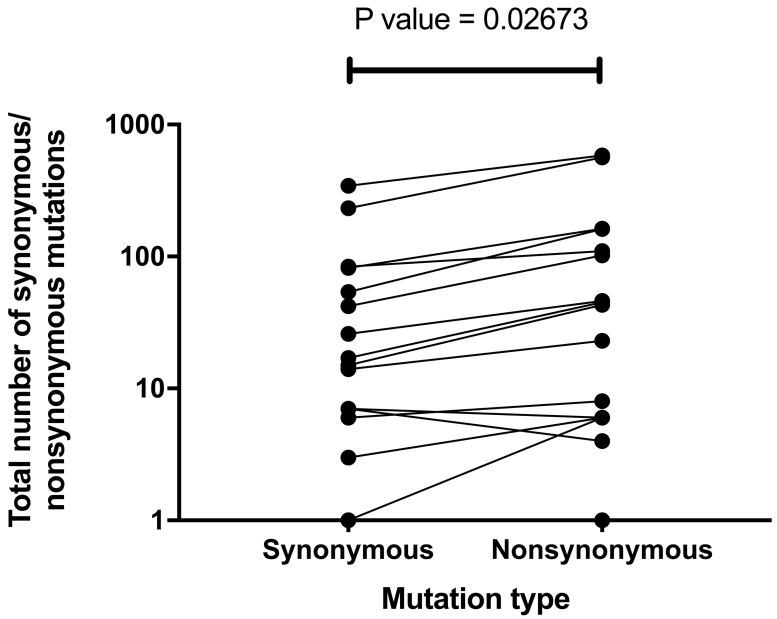
Figure 4.9 Between niche diversity using BLASTN and alignment-based SNPs – patient 265



BLAST Ring Image Generator (BRIG) plots of paired antrum and corpus consensus genome assemblies constructed as described in section 3.2.2.3. Figure A depicts the antrum consensus genome as the reference with the corpus consensus genome as the query where figure B depicts the reverse of this. The rings represent the following from the centre most ring outwards; GC percentage, GC skew of the mean, sequencing coverage, contig breaks, reference consensus genome, query consensus genome, SNP locations. The green concentric ring represents the antral consensus genome where the purple represents the corpus consensus genome. SNPs are a numbered for reference purposes so that they can be linked to a specific SNP within the VCF files. Additionally, SNPs were colour coded as follows; black = identified by whole genome alignment only, teal = identified only by the alternative niche read mapping approach, red = identified by both methodologies. An upper BLASTN identity threshold of 98% and lower identity of 95% were used. BRIG diagrams for all other patients with paired antrum and corpus data are presented in the Appendix (figures 11.4.21 – 11.4.35).

Validated (red) alignment SNPs were pooled from both antrum and corpus references and duplicate validated SNPs removed to create a final list of between niche validated alignment SNPs. These were then annotated to a single reference genome of the antrum or corpus from each patient and determined to be synonymous or nonsynonymous in nature to the alternative call on the opposite paired genome and statistically tested through a paired t-test to determine if there was a mutational significance to these between niche alignment SNPs (figure 4.10). These results show that there are statistically more nonsynonymous mutations between paired antrum and corpus populations. This suggests that there are selection pressures acting independently of these niches, resulting in more nonsynonymous fixed mutations between the populations. This finding is in contrast to the within niche genetic diversity when compared between the antrum and corpus (figure 4.8). An alternative interpretation of this result is one where there is no selection acting between the populations, proceeding to the accumulation of more nonsynonymous mutations that are not being positively or negatively selected against, resulting in intra population persistence of higher numbers of nonsynonymous mutations. Despite this alternative argument, the stark difference of nonsynonymous to nonsynonymous mutational differences of the within niche (figure 4.8) and between niche (figure 4.10) suggests differences in selection by the two analysis methodologies. Regardless of the favoured theory, these results further add to the picture of within and between niche genetic variability and differences between synonymous and nonsynonymous mutational profiles within the same patient.

Figure 4.10 Synonymous to nonsynonymous SNPs between aligned paired patient antrum and corpus consensus genomes



The validated (red) SNPs identified by the combined whole genome alignment and alternative niche mapping methodologies (figure 4.9; Appendix, figures 11.4.21 – 11.4.35) were annotated and the CDS SNPs were determined to be synonymous or nonsynonymous in nature between the populations. A paired t-test was performed using GraphPad (version 8.1.2) with the total numbers of synonymous and nonsynonymous mutations between the paired populations. The resulting p-value is displayed above the graph.

The annotated and validated (red) SNPs were further grouped by gene/gene product to identify the number of alignment SNPs found harboured between the aligned genomes. This was used to create a heatmap of genes showing SNPs between aligned genomes and their corresponding densities (figure 4.11). This figure shows that all patients have different levels of between niche alignment variability across genes with low (93, 194, 295, 308, 326, 349 and 495), moderate (77, 120, 256, 444 and 77) and high (45, 322, 565 and 732) levels of gene diversity between paired antrum-corpus isolates. In relation to the low, medium and high determination of within niche polymorphic variability (table 4.3), there appeared to be a disconnect between the total number of between niche alignment SNPs (figure 4.12) and the total number of within niche polymorphic variants (table 4.3) suggesting that high within niche polymorphic diversity does not correlate with high between niche alignment SNPs. Therefore, if only one of these analysis methodologies was followed, such as alignments of paired genomes from the same patient isolated from different niches (in the absence of deep

sequencing data) any underlying intrapopulation diversity might be missed. The opposite is also true, where a population is analysed following the allelic calling pipeline and there is an observed high polymorphic diversity, this does not necessarily mean there will be high levels of between niche alignment associated genetic diversity. This is an important distinction as future studies of within patient genetic diversity would require similar or improved sampling methodology (*i.e.* sampling more areas of the stomach such as the fundus or multiple samples from the same niche) to capture and better interoperate this type of analysis. In short, to better capture and understand within and between niche variability this duel methodology (within niche allelic variant calling and between niche consensus whole genome alignment) should be followed. However, there were some similarities to the within and between niche comparisons from the minor allele calling results (figure 4.6). For example, OMPs showed the most between niche diversity across all samples (figure 4.11). Other genes were equally common including *vacA* paralogue and restriction modification associated genes. This analysis confirms and characterises the highly diverse genetic variation of *H. pylori* populations observed within the stomachs of individual patients.

Bringing the within niche polymorphic diversity and between niche whole genome alignment results together (figure 4.12), the differences and similarities between the whole genome alignment and within niche allelic calling pipelines are clearer to observe. It is important to note that some genes identified with *between* niche diversity by alignment also had polymorphic diversity *within* one or both niches. One such example is the iron-regulated OMP within sample 265C that is polymorphic within this population but has also been identified as an alignment variable gene between the consensus genome sequences of the paired antrum and corpus derived patient samples (figure 4.12). This observation is not uncommon across the dataset and has more than one possible explanation. First, this gene is variable at multiple sites within this gene group and one of these positions is polymorphic within one population and not the other (within niche variation) where at a separate position there is a SNP between the two populations (consensus genome alignment SNP) that is no longer detected as polymorphic within the population due to fixation. Second, the polymorphic position within a specific niche (265C) has a consensus base call that is different to that of the opposite niche, resulting in a between niche consensus genome alignment SNP. However, in this scenario the base called as a between niche alignment SNP also observes within niche polymorphic diversity at this same base position resulting in one niche observing polymorphic diversity in comparison to the opposite niche population.

Such an observation may help in the understanding of which direction a SNP is occurring as the population observing polymorphic diversity at this site is potentially more likely driving the change than the population that shows no polymorphic diversity at this position. Further manual inspection of the VCF files from the within niche 265C and between niche whole genome alignment methodologies revealed that the latter was true where the between niche alignment SNP was identified as polymorphic at the same position within the minor allele calling pipeline.

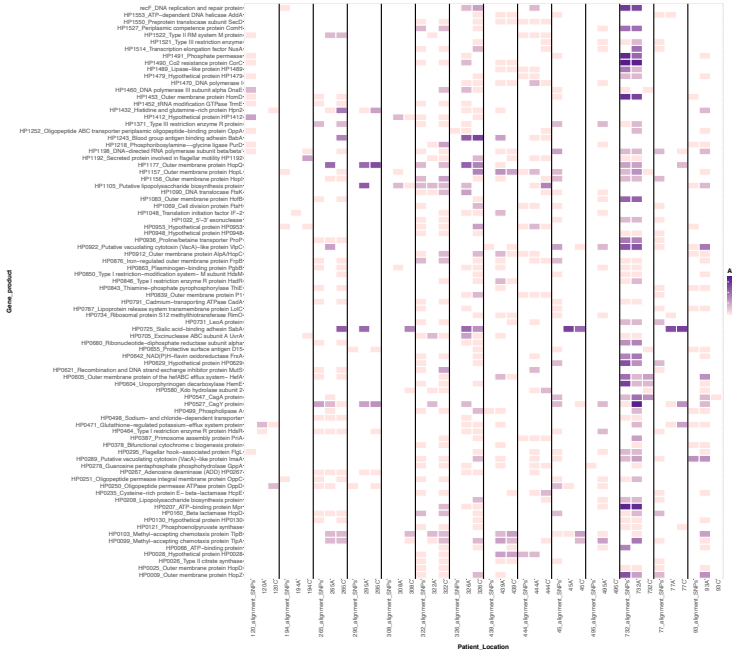
Some genes show both within niche polymorphic diversity from the antrum and the corpus, and between niche diversity from the alignment of consensus genomes (figure 4.12). One such example was the *cagY* gene for patient 265 (figure 4.12). The same explanations could still apply here as described for when only one niche is observing polymorphic diversity. However, if both niches observe polymorphic diversity at the same aligned position then this might indicate a position that is under similar selection across both niches and is starting to reach fixation within one population sooner than the other. Alternatively, this could indicate a position that is under no selection and is persisting within the population at different abundances due to the lack of clearance of unfit strains. Manual inspection of this particular gene revealed that the between niche alignment SNP position was polymorphic at the same alignment base position within both niches.

Finally, some genes were identified with genetic diversity by between niche consensus alignments but showed no within niche genetic diversity for the same gene. One such example was the *oppA* gene from patient 120 (figure 4.12). This is perhaps the strongest case of positive selection between the different niches as there is no evidence of the potentially less fit strain persisting within one niche. However, it is not possible to know in which direction this mutation occurred due to unknown sequence of the ancestral/originally infection isolate. Taken together, not only are these significant observations in terms of between genome alignment methodologies/comparative genomics of non-clonal samples due to the potential bias of the consensus call, but also these observations validate the reliability of the analysis pipelines developed within this study to capture the within patient genetic diversity dynamics. Additionally, studies that employ comparative genomics methodologies of non-clonal samples might not capture the full extent of sample diversity or allow for a consensus base call that would otherwise result in a between sample SNP which in fact is a position that is not variable among a subset of strains within the population and *vice versa*.

Figure 4.11 Heatmap of between niche genetic diversity by whole genome alignment verified SNPs

Heatmap of patient antrum and corpus consensus genome alignments using the validated (red) SNPs as described in the materials and methods section 4.2.2.3. Heatmap was created using ggplot2 and the RColorBrewer through the R statistical software version 3.5.1 (R Core Team, 2018). This figure represents variable genes found between three or more different patients. Only patients with paired antrum and corpus data were included. A full undocketed heatmap can be visualised in the Appendix (figure 10.4.36).

Figure 4.12 Heatmap of genetically diverse genes identified from whole genome alignment and within niche allelic variability



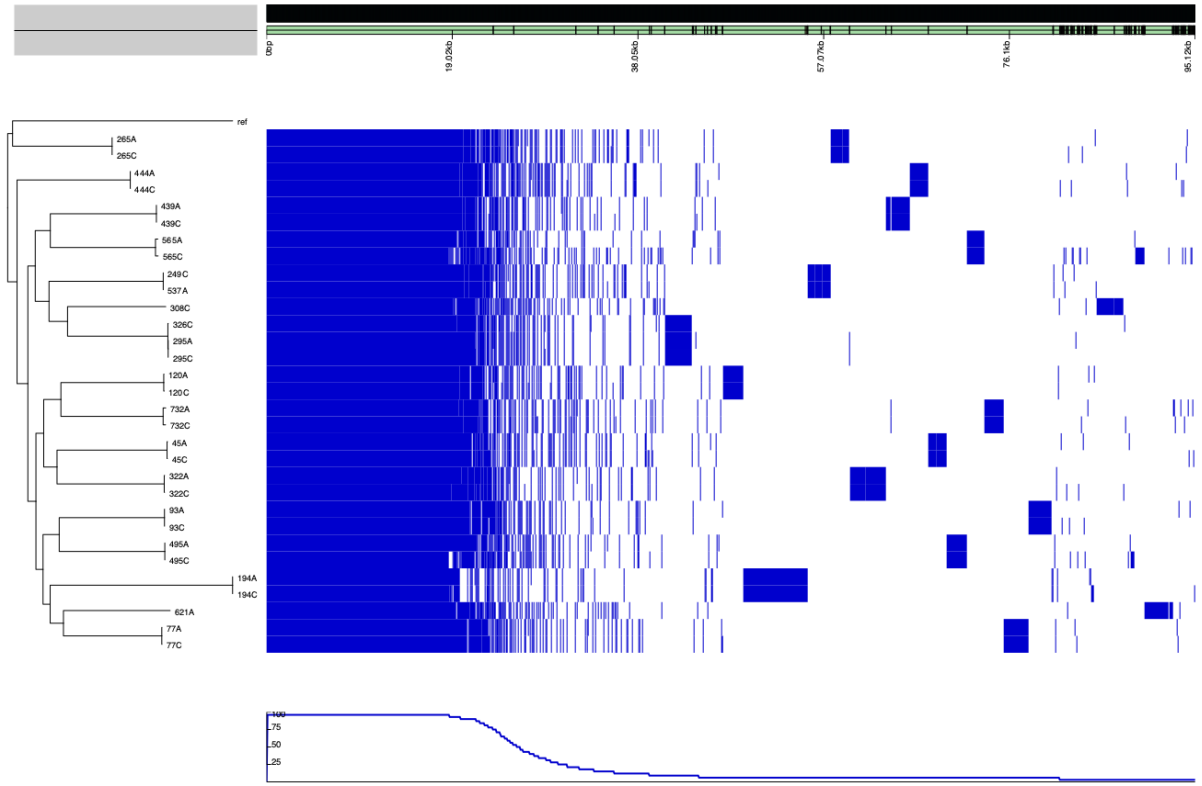
Heatmap created using ggplot2 and the RColorBrewer through the R statistical software version 3.5.1 (R Core Team, 2018). The data used to create figures 4.6 and 4.11 were pooled due to the use of the same allelic variant and alignment SNP annotation methodology (Chapter Two, section 2.14) and used to create this hybrid heatmap depicting genetic diversity both within each patient niche (antrum or corpus) by the within population minor allele calling pipeline (section 4.2.1.1) and between niche (antrum and corpus) by comparing the population minor allelic genes or through the verified niche consensus alignment (red) SNPs (section 4.2.2.3). Gene products/genes that were shown to be diverse six or more times by any methodology across the patient datasets were included in this figure. Hypothetical proteins and intergenic nucleotide diversity was removed from this dataset to improve visualisation. Only patients with paired antrum and corpus data were included. Patient 565 was excluded due to the extensive, outlier minor allelic diversity dataset to improve visualisation of the most genetically diverse genes/gene products. A undocketed figure including patient 565 can be found in the Appendix (figure 10.4.37).

To further investigate the between niche alignment gaps and regions of <95% BLASTN identified from the BRIG analysis (figure 4.9) a gene presence and absence analysis was conducted to distinguish differences in gene content. Roary (Page et al., 2015) was used to identify a total of 4,755 unique genes across 31 *H. pylori* populations (Appendix, figure 10.4.38). These were made up of; 935 core genes (99% - 100% strains), 55 soft-core genes (95% - 99% strains), 788 shell genes (15% - 95% strains) and 2,977 cloud genes (0% - 15% strains). This analysis revealed that there was a relatively small core genome shared between all populations and a genetically diverse accessory genome (appendix, figure 10.4.38).

To investigate how diverse these patient populations were to one another, a core genome phylogeny was constructed by approximately-maximum likelihood (Treangen et al., 2014) and the gene presence and absence was overlaid as a heatmap in order to better depict these differences (figure 4.13). As expected, each patient sample clustered independently on the phylogenetic tree. There were substantial genetic content differences between all patients outside the core genome. However, the paired antrum and corpus consensus genomes from each patient have remarkably similar accessory gene content, creating a unique gene barcode for each patient, with very little variability observed. The only exception to this was patient 565 which showed much more genetic content variability between the antrum and corpus. This result is similar to that observed in figures 11.4.34A-B (Appendix) and shows consistency across the different types of analysis. It is possible that this patient was infected with more than one strain of *H. pylori* that have co-existed and independently diversified over time, resulting in the analysis observations seen here. It also appears that the antral dominant population is present within the corpus but that the corpus harbours an additional genetically diverse *H pylori* population that is genetically distinct. This theory is backed up by the unusual gene content differences observed between the antrum and corpus of patient 565 whereby the antral gene presence absence is comparable to the corpus, but the corpus has many more additional genes. Furthermore, by *de novo* whole genome assembly, the consensus length of the corpus genome was significantly larger than that of the antrum (203,239 bp), suggesting that the corpus population was much more complex and held additional genetic content than that of the antrum. Considering the average whole genome length of *H. pylori* is approximately 1.67 Mbp in length, the antral population was comparable to this suggesting the presence of a single strain infection that has diversified over time. However, the genome size of the corpus consensus genome was 1.75 Mbp suggesting the presence of more than one infecting

isolate. In comparison to other paired patient samples, patient 565 showed one of the highest between niche alignment variability (788 verified CDS SNPs) and by far the highest number of within niche CDS polymorphic diversity (antrum = 463 polymorphic positions; corpus = 12,754). Again, the vast within niche polymorphic diversity of the corpus further highlights the possibility of this niche being populated by more than one distinct infecting strain, resulting in the inflation of observed polymorphic diversity.

Figure 4.13 Pan-genome gene presence/absence of all deep sequenced samples and core genome phylogeny



Pan-genome analysis of deep sequenced samples from this study, excluding the contaminated samples 326A and 308A. This figure was constructed as described in the materials and methods (section 4.2.2.5) and visualised by Phandango version 1.3.0 (Hadfield et al., 2018). The left-hand side of the figure displays the core genome phylogeny. The bottom graph represents the percentage of the strains harbouring each specific gene. The top graph represents the accumulating total length of all genes combined. At the centre is the gene presence/absence indicated by blue (presence) and white (absence). Other pangenome output statistics are presented in the Appendix (figure 10.4.38).

4.4. Future work

The findings described in this Chapter are novel both in terms of the application of *H. pylori* population deep sequencing and the pipelines used to analyse the datasets. Novel insights have been obtained by combining analysis pipelines in the understanding of intra- and inter-population genetic diversity. However, this Chapter focused on between niche diversity of the antrum versus corpus using single biopsies taken from these locations. It might be beneficial to take additional samples from the same stomach region to understand if there is further within niche diversity and/or to determine if the results observed here were down to sampling distance rather than niche specific diversity. Secondly, there are other regions within the stomach that were not studied, such as the fundus. Finally, the samples used within this study were taken from patients attending a single hospital within the UK and expanding the scope of sample collection to additional countries could prove informative, especially considering that *H. pylori* is known to be a globally diverse pathogen, with geographical genetic clustering. These additional samples might further elucidate the within patient genetic diversity of *H. pylori* and allow for further insights into niche specific adaptation. Obtaining further samples from a single patient could prove difficult due to the distress the patient endures during the endoscopy sampling procedure and may not be ethically viable.

The analysis presented in this study has presented a snapshot of genetic diversity within *H. pylori* sample populations. Chapter Five will go on to investigate single colony isolates extracted from the same populations described here, from a subset of patients. Thus, providing further data resolution in the form of additional analysis and allowing for the cross comparison of results that can be used to help validate the methodologies and analysis presented here.

5. Chapter Five: Single colony whole genome sequencing

5.1. Introduction

The advantage of using a duel deep population and single colony isolate sequencing approach from the same environments is that it allows for the analysis of the populations in a more comprehensive way. For example, there is no advantage to looking at the phylogeny of the two deep sequenced consensus antrum and corpus assembled genomes from an individual patient because population structure would be difficult to investigate. However, multiple single colony isolates taken from each of the antrum and corpus allow for the investigation of population structure between the different niches. Furthermore, paired deep sequenced populations do not allow for the investigation of recombination occurring within or between them as strain level comparison is impossible.

Previous studies on *H. pylori* population structure have revealed large scale phylogeographic clustering as well as more local geographical clustering. This is expertly reviewed by Suerbaum and Josenhans (2007). The advent of MLST and its use on *H. pylori* strains isolated globally revealed phylogeographic clustering of isolates (Achtman et al., 1999; Montano et al., 2015; Bullock et al., 2017). Phylogeographic clustering has also been observed by comparing coding sequences from single genes such as the antigen binding *babA* (Thorell et al., 2016) and virulence associated genes from the *cag*PAI in particular the *cagA* gene (Olbermann et al., 2010a). A study by Vale *et al.* (2015), identified *H. pylori* subpopulations within Europe that generally clustered according to wider geographical regions inferred from phylogenetic analysis of prophages. Such studies confer the idea of population specific adaptations, colonisation and virulence. Indeed, some studies have gone as far as associating phylogeographic clustering with differences in disease risk (Sheh et al., 2013; de Sablet et al., 2011; McClain et al., 2009).

Two important studies by Falush et al. (2003) and Linz *et al.* (2007), took advantage of the phylogeographical clustering approach and identified significant human migration events in human prehistory. This further highlighted the intimate relationship of *H. pylori* infection and co-evolution with humans since anatomically modern humans migrated out of Africa approximately 58,000 years ago (Linz et al., 2007). These studies prompted further research into the history of *H. pylori* and human infection with more recent studies identifying further insights into human migrations with one study identifying a second African migration in the last 52,000 years (Moodley et al., 2012).

Other studies have investigated evidence of human migrations in geographical areas that were otherwise hard to trace by human genetics or more traditional linguistic and archaeological approaches, through the analysis of *H. pylori* genomes taken from infected individuals from various countries (Breurec et al., 2011).

It is clear then, that human infection with *H. pylori* predates human migration out of Africa around 58,000 years ago and has resulted in the global scale phylogeographical clustering we observe today. However, few studies have looked at the *H. pylori* population structure within patients, especially from the different niches of the human stomach. A recent study by Fung *et al.* (2019), investigated the colonisation of isogenic *H. pylori* strains by fluorescent strain tagging in a mouse infection model. By increasing the transparency of the gastric tissue through CLARITY staining (Tomer et al., 2014), they were able to observe the colonisation of fluorescently tagged *H. pylori* using confocal microscopy. They proposed a model of infection whereby founder strains colonise deep within gastric glands where they spread to colonise adjacent glands, forming islands comprised of predominantly clonal strains. This study also suggests that the clonal colonisation islands persist preventing free-swimming strains in the mucosa from colonising pre-infected glands, potentially reducing their interactions. However, founder strains compete for space with adjacently colonised glands. Ultimately, the colonisation of the gastric niche is potentially dominated by sub-populations of *H. pylori* due to these founders. Although this study provides novel insights into how *H. pylori* colonise the gastric epithelium and persist in the face of constant turnover of epithelium cells as well as mucus, it is important to note that this study used genetically similar strains. Therefore, it is not fully representative of a potential multi strain infection or of a genetically diverse *H. pylori* infection that has evolved during a long-term chronic infection. A multi strain infection or a diverse *H. pylori* infection has the potential to hold fitness differences between strains, which might change the dynamics of the infection. Furthermore, very little colonisation of the corpus was observed, suggesting that niche specific differences between the antrum and corpus could be preventing colonisation. Therefore, *H. pylori* populations and isolates between the antrum and corpus could hold niche specialised strains and it is not clear how these different populations might differ or interact.

A recent study by Ailloud *et al.* (2019), isolated ten single colony isolates from the antrum, corpus and fundus of 16 patients. This study highlighted a patient specific degree of isolates clustering into niche specific clades. The isolates from some patients

generally clustered into niche specific clades, while other patients didn't show a defined niche specific phylogenetic clustering. The authors suggest the disparity between results could in part be explained by migration events between the different stomach niches of the antrum, corpus and fundus. Using a marginal state reconstruction migration model, it was shown that migration between the corpus and fundus was more common than migration between the corpus and the antrum. Migration events between the antrum and fundus was rare. These results suggested that the antrum population was generally segregated from the corpus and fundus, potentially attributed to the differences between the oxyntic epithelium of the corpus and fundus and the antrum that lacks parietal cells. While this study explores the local spatial phylogenetic clustering (or lack thereof) of isolates taken from the antrum and corpus it is important to note that this is a low resolution snapshot of genetic diversity taken from a relatively small sample size. Therefore, the populations of the antrum, corpus and fundus and the genetic differences observed between them could potentially be over simplified.

Homologous recombination plays an important role in *H. pylori* genetic diversity. Recombination can lead to faster diversification of the genome than spontaneous SNPs alone. This is especially true in a mixed strain *H. pylori* infection (Falush et al., 2001b; Kennemann et al., 2011; Krebes et al., 2014). As all *H. pylori* infections are different due to each individual harbouring their own unique *H. pylori* strains, the recombination rate of *H. pylori* is most likely to be variable. One main factor for this variation could potentially be the lack of opportunities for homologous recombination to occur, perhaps due to an absence of a mixed strain infection. This is noted by Didelot *et al.* (2013), who investigated the recombination of familial strains. However, in some patients an astonishing level of recombination was detected that introduced up to 100 times more substitutions then spontaneous mutations.

Other studies have investigated the number of recombination clusters within *H. pylori* genomes from familial and sequentially isolated samples, identifying between 16 and 441 import clusters (Kennemann et al., 2011; Krebes et al., 2014). The core genome, and more specifically the house keeping genes, are thought to be less prone to homologous recombination. However, this has been shown to be untrue in the case of *H. pylori* where recombination has been shown in housekeeping genes, core and accessory genes with little or no differences in recombination frequencies (Yahara, Lehours and Vale, 2019; Achtman et al., 1999). Such genetic diversity driven by

recombination is thought to help in colonisation and adaptation of *H. pylori* to new patients and niches as well as persistence of chronic infection.

It is difficult to investigate both the population structure and recombination of deep sequenced *H. pylori* populations. This is mainly because of the limited number of comparable genomes, in this case the consensus genomes, and the genetic diversity attributed to specific strains. While the deep sequenced populations allow a snapshot/ insight of the genetic diversity at a population level, this Chapter aims to investigate single colony isolates obtained from these populations. This allows for a comprehensive investigation of the patient samples from the antrum and corpus adding analysis that could not be attempted by the population deep sequencing methodology. Combining the data generated from both methodologies will provide a high-resolution analysis of the structure and diversity of each sampled population. Furthermore, this dual approach will allow for the cross comparison of results, providing a unique opportunity to scrutinise the results of both analysis pipelines.

5.2. Materials and Methods

Sample selection, DNA extraction, whole genome deep sequencing, sequencing read curation, contamination detection, whole genome assembly, assembly curation and genome annotation was conducted as described in Chapter Two.

5.2.1. Isolation of single colonies and whole genome sequencing

Population sweeps from frozen stocks were cultured in two different ways in an attempt to maximise the chance of isolating non-clonal single colony isolates and to better observe single colony diversity. Method 1 – frozen stocks were cultured as described in Chapter Two section 2.2. After incubation for 48-72 hours, edge growth was taken from multiple random areas of the agar plate using the same inoculation loop. A quadrant streak was performed on a fresh blood base #2 agar plate supplemented with 7.5% defibrinated horse blood and incubated for 48 – 120 hours at 37°C under microaerophilic conditions (10% CO_2, 5% O_2, 85% N_2). These were checked daily for the identification of single colony isolates. Once single colony isolates were visible approximately six well-spaced and isolated single colony isolates were carefully extracted using a 1 μl plastic inoculation loop and spread in a small quadrant onto a

fresh agar plate as described previously. Where appropriate, a range of different colony morphologies were extracted. For example, colonies of different size and gloss were taken. The significance of the small quadrant streak was to increase visualisation and density of bacterial growth from the initial single colony transfer after incubation. The remaining surface area of the agar plate was spread randomly whilst rotating the same inoculation loop to maximise the chance of bacterial transfer. After 72 – 120 hours of incubation under conditions previously described, the agar plates with patched up single colony growth were investigated for any signs of contamination and plates were discarded if contamination was suspected. To aid with this process a rapid urease test was carried out on each agar plate containing bacterial growth as described in Chapter Two section 2.3. Urease negative samples were discarded. A cotton swab was used to extract all bacterial growth that was urease positive and spread it onto two fresh agar plates which were incubated for 24 – 48 hours as previously described. Resulting growth over the two agar plates were harvested into Iso-Sensitest broth supplemented with 10% glycerol and stored at -80°C for long term storage.

Method two – the same process as method one was followed with the exception of the first step whereby 50 μl of the frozen stock was taken after gently thawing the top section of the cryogenic-tube and inoculated onto the outer section of the agar plate. From here a quadrant streak was performed in order to isolate single colony isolates. Therefore, this method required one fewer inoculation/patching of growth and incubation step than method one but, single colony isolates usually took longer to form by this method making the total incubation time comparable to that of method one.

Further to increasing the potential diversity of single colony isolates, method two was followed to mitigate against the potential effect of 'fitter' colony growth outcompeting the growth of slower growing *H. pylori* isolates from the initial culturing of the frozen stock in method one.

Once all single colonies were isolated they were numbered with the original patient number and biopsy location (A – antrum; C – corpus) followed by a unique single colony isolate number.

From the collection/database of single colony isolates, five – six single colony isolates were randomly selected from each patient biopsy location with equal number of strains

from each isolation methodology and brought forward for DNA extraction and single colony sequencing as described in Chapter Two sections 2.4 – 2.5 and 2.7 – 2.8.

A list of single colony isolates used within this study can be referred to in table 5.1.

Table 5.1 Single colony isolates used within this study

Patient number and biopsy location	List of single colony isolate numbers used within this study
194A	194A1, 194A2, 194A4, 194A6, 194A9, 194A12
194C	194C1, 194C2, 194C3. 194C4, 194C5, 194C6
249C	249C1, 249C3, 249C6, 249C8, 249C10, 249C16
295A	295A1, 295A2, 295A3, 295A4, 295A5, 295A6
295C	295C1, 295C2, 295C4, 295C6, 295C7, 295C8
322A	322A1, 322A2, 322A3, 322A4, 322A6, 322A7
322C	322C3, 322C4, 322C5, 322C6, 322C7, 322C8
326A	326A22, 326A23, 326A24, 326A25, 326A26, 326A27
326C	326C1, 326C2, 326C3, 326C4, 326C5, 326C6
439A	439A1, 439A4, 439A5, 439A6, 439A7, 439A8
439C	439C2, 439C3, 439C5, 439C6, 439C7, 439C8
444A	444A1, 444A2, 444A3, 444A4, 444A6, 444A8
444C	444C1, 444C2, 444C6, 444C8, 444C9, 444C10
495A	495A2, 495A3, 495A5, 495A6, 495A8
495C	495C1, 495C2, 495C3, 495C4, 495C5, 495C6
537A	537A1, 537A3, 537A4, 537A5, 537A7, 537A8
565A	565A1, 565A3, 565A4, 565A5, 565A6, 565A13
565C	565C1, 565C3, 565C6, 565C8, 565C12, 565C14
732A	732A2, 732A3, 732A4, 732A5, 732A7, 732A8
732C	732C1, 732C2, 732C3, 732C4, 732C5, 732C6

All *H. pylori* single colony isolates used within this study. A = antrum-derived isolate, C = corpus-derived isolate.

5.2.2. Sequencing read curation, contamination detection, whole genome assembly, assembly curation and genome annotation

Sequencing read curation, contamination detection, whole genome assembly, assembly curation and genome annotation was conducted as described in Chapter Two sections 2.9 – 2.13.

5.2.3. Creation of the patient reference consensus genome

A patient reference genome was created by merging all curated deep sequenced antrum and corpus sequencing reads from each patient and assembled as described in Chapter Two section 2.11. This was done to allow the comparison of all antrum and corpus single colony isolates to a single reference. This was in place of the selection of either the antrum or corpus deep sequencing reference. In essence, this was a hybrid assembly of the deep sequencing antrum and corpus reads resulting in a patient reference sequence.

It must be noted that there is still a potential assembly bias at play in this hybrid assembly. Of particular note are differences in sequencing coverage between the antrum and corpus read sets. For example, if the corpus reads were in support of a particular genetic sequence but there were more antrum reads in support of an alternative sequence then the patient reference consensus would be selected based upon the highest read depth, which in this case would be that of the antrum. This would cause an antrum biased reference. However, this hybrid approach was used because of the problems inherent in the alternative of selecting the reference as *either* the antrum or corpus deep sequenced consensus genome. Furthermore, due to the extremely high mutation and recombination rate as well as the differences between the accessory genomes of different *H. pylori* strains, a 'traditional' reference *H. pylori* genome such as *H. pylori* 26955 (NC_000915.1) would have been unsuitable for this research project.

The most suitable reference would have been the whole genome sequence(s) of the initial infecting isolate(s). However, obtaining these from naturally infected individuals presenting for gastric endoscopy many years after the suspected initial infection was not possible.

5.2.4. Phylogenetic analysis

A read mapping approach was used to create a core genome phylogeny for patients with single colony isolates from both the antrum and corpus (patients; 194, 295, 322, 326, 439, 444, 495, 565 and 732). Areas of recombination within the core genome were detected and removed from the phylogenies to create a more accurate phylogenetic tree construction.

The following pipeline was used:

```
snippy --cpus {user specified CPU/thread/core number} --report --minfrac 0.9 --mincov 6 --mapqual 30 --basequal 30 --ref {patient reference consensus genome GenBank annotation file} --outdir {user specified output directory} --R1 {path to forward/mate pair 1 of a single colony isolate} --R2 {path to reverse/mate pair 2 of a single colony isolate}

snippy-clean_full_aln {input file generated from previous command: core.full.aln} > {user specified output: clean.full.aln}

run_gubbins.py -p gubbins {input file generated from previous command: clean.full.aln}

snp-sites -c {input file generated from previous command: gubbins.filtered_polymorphic_sites.fasta > {user specified output: clean.core.aln}

FastTree -gtr -nt {input file generated from previous command: clean.core.aln > clean.core.tree
```

Snippy version 4.4.0 (Seemann, https://github.com/tseemann/snippy) was executed using the `snippy-multi` script which is part of the Snippy package. This was done in order to group all single colony isolates from each patient sample together but is essentially the same Snippy command (above) but run consecutively for each individual isolate. The `snippy-multi` script enacts the snippy-core script on completion which generates a core genome alignment of all single colony isolates used as input against the user defined reference.

The Snippy core genome alignment output file was cleaned using the `snippy-clean_full_aln` script from the Snippy package to remove ambiguous characters with 'N' bases. Recombination within the core genome alignment was detected and removed by Gubbins version 2.3.1 (Croucher et al., 2015). Polymorphic sites between the aligned genomes were extracted to create a SNP alignment file using the SNP-sites tool version 2.4.1 (Keane et al., 2016). Finally, SNP alignment-based phylogeny of the isolates core genome was constructed using FastTree version 2.1.1 (Price, Dehal and Arkin, 2010) which infers approximately-maximum-likelihood.

A SNP based core genome phylogeny was used as the single colony isolates from the same patient were potentially very similar and this methodology allows for fine scale resolution of the differences between the isolates.

5.2.5. Whole genome alignment of patient specific single colony isolates

The assemblies for all single colony isolates and patient reference consensus were filtered to remove contigs <500 bp in length (Payne, https://github.com/tinybio/filter_contigs).

Patient specific single colony isolates from both the antrum and corpus were aligned in a pair-wise fashion using the multiple genome alignment (Mauve, version 2.4.0) tool (Darling et al., 2004) in progressive mode with default settings as previously described. However, the patient consensus genome for each patient (Chapter Five, section 5.2.3) was used as the reference.

Mauve was used to reorder the contigs of each single colony isolate to the patient reference consensus genome.

5.2.6. BLAST ring image generator: single colony alignments and variant calling to the patent reference genome

A similar approach was taken to that described in Chapter Four section 4.2.2.3, but with some key differences.

Firstly, the patient reference consensus genome was used as the reference sequence. This allowed for clearer visualisation of between niche single colony variation by aligning all patient specific single colony isolates from the antrum and corpus.

Secondly, variant genes that were identified between single colony isolates were compared to the allelic genes identified by the population minor allele calling pipeline (Chapter Four, section 4.2.1.1). Unlike the deep population sequencing BRIG alignments, the exact variant positions could not be mapped due to the use of the patient (antrum plus corpus) reference genome. Instead, the single colony reads were mapped to the patient reference genome and the positions of the variants were annotated (Chapter Two, section 1.14). The annotated variant genes were harvested manually and compared to the allelic genes from the minor allele calling pipeline (Chapter Four, section 4.2.1.1). Concordant variant/allelic genes were then identified between the deep population sequencing and single colony analysis, as were the genes that were uniquely identified as variable for each analysis. These genes were identified in the annotation file (Chapter Two, section 2.13; PROKKA gbk/gff output files) of the patient consensus reference where each gene start and stop position was extracted for the gene names that matched the variable genes. A custom annotation file was constructed with these genes start and stop positions along with an abbreviated gene name (where appropriate) and colour coded. The variant genes were colour coded in relation to how they were detected. Red gene labels represented those that were detected as variable by the single colony sequencing and the minor allele population deep sequencing methodologies while the single colony only and minor allele only variable genes were coded black and teal respectively. This custom annotation file was loaded into BRIG (Alikhan et al., 2011) to better depict the variable genes along the genome and by which methodology they were detected by.

While this helps in the visualisation of variable genes across the genome there are three caveats to this display. One being that where there are more than one gene copy, both copies are displayed on the BRIG diagram even if only one of those genes was detected as variable. The next being that both methodologies might not identify the same base within a gene as being variable, showing a distorted picture of gene variability not site-specific concordance of gene variability like the deep population sequencing BRIG figures provide (Chapter Four, figures 4.9A-B). Leading on from this, there is no SNP density information within variable genes.

The single colony isolate reads were mapped through Snippy (Seemann, https://github.com/tseemann/snippy) using the following command:

```
snippy --cpus {user specified CPU/thread/core number} --report
--minfrac 0.9 --mincov 6 --mapqual 34 --basequal 30 --ref
{patient antrum/corpus consensus genome GenBank annotation file}
--outdir {user specified output directory} --R1 {path to
forward/mate pair 1 of a single colony isolate} --R2 {path to
reverse/mate pair 2 of a single colony isolate}
```

Finally, a higher BLASTN identity threshold was used for the BRIG plots as the single colony isolates are potentially less diverse at an individual strain level than the deep population dataset. Upper and lower identity thresholds of 99% and 96%, respectively, were used.

5.2.7. Pan-genome analysis

As described in Chapter Four section 4.2.2.5, a pan-genome assembly was performed using Roary (version 3.8.2) to investigate whole genome differences between the single colony isolates within and between the antrum and corpus of the same stomach.

To enable this, Roary (Page et al., 2015) was run in the same way as the deep sequencing dataset but instead included all single colony isolates used within this study. Once complete, the Roary script roary-query_pan_genome was executed for each patient with antrum and corpus isolate sets (patients; 194, 295, 322, 326, 439, 444, 495, 565 and 732) with the antrum isolates set as input 1 and the corpus isolates as input 2, using the difference flag. This output the differences between the two groups of isolates revealing antrum and corpus unique genes. The following pipeline was used:

```
cd {path to directory containing list of PROKKA output .gff files
to pass to Roary} ; roary -p {user specified CPU/thread/core
number} -e -n -v -f {path to output directory} *.gff
```

```
roary-query_pan_genome -a difference --input_set_one {Patient
specific antrum isolate list of PROKKA output .gff files) --
input_set_two {patient specific corpus isolate list of PROKKA
output .gff files)
```

The resulting unique set statistics (set_difference_unique_set_one_statistics.csv and set_difference_unique_set_two_statistics.csv files) for each patient isolate set was manually inspected and brought together into an antrum unique and corpus unique spreadsheet. Duplicate gene names were removed and a `COUNTIF` (Microsoft Excel version 16.16.6) function was used to count the number of instances each sample set harboured that unique gene. This resulted in a table containing information on how many single colony isolates harboured that specific antrum/corpus unique gene for each patient biopsy location and allowed comparisons across patients.

A core genome alignment of all single colony isolates was conducted and a phylogenetic tree by approximately-maximum-likelihood was constructed through Parsnp (version 1.2) which is part of the Harvest suite (Treangen *et al.*, 2014). The clinical reference *H. pylori* genome J99 (NC_000921.1) was used as the phylogenetic root. The following command was executed:

```
parsnp -p {user specified CPU/thread/core number} -r {H. pylori J99} -d {path to directory containing genomes to analyse} -a 13 -c -x
```

The resulting phylogeny was visualised in FigTree version 1.4.4 (https://github.com/rambaut/figtree) and re-rooted to the midpoint and ordered in decreasing order to improve visualisation. This was then uploaded into phandango (Hadfield *et al.*, 2018) alongside the gene presence and absence matrix produced by Roary where the reference was manually appended to create a pan-genome gene presence and absence visualisation.

5.2.8. Pan-genome wide association study

The gene presence and absence output from Roary (section 5.2.7) was interpreted by Scoary (Brynildsrud et al., 2016). Scoary calculates gene associations between all genes in the pan-genome to user specified traits of interest.

The single colony isolates from the antrum were compared against the isolates obtained from the corpus for each patient to investigate patient specific between niche differences. Additionally, all antrum isolates from the dataset were compared to all

corpus isolates from the dataset to investigate population specific antrum or corpus differences. This was extended to the isolates corresponding to histology Sydney scores (where available) for absent/low (Sydney scores 0-1) versus moderate/high (Sydney scores 2-3) for activity, atrophy, inflammation and intestinal metaplasia (Dixon et al., 1996).

The following command was executed:

```
scoary.py --genes {path to input gene presence/absence table output from Roary} --traits {path to table of traits to compare} --threads {user specified CPU/thread/core number} --outdir {path to output directory}
```

5.2.9. Recombination detection

Homologous recombination was inferred between all single colony isolates taken from each patient with isolates from both the antrum and corpus regions of the stomach (patients; 194, 295, 322, 326, 439, 444, 495, 565 and 732).

Sites of homologous recombination were inferred using two different methods employing different recombination detection algorithms. Genealogies unbiased by recombinations in nucleotide sequences (Gubbins; Croucher *et al.*, 2015) relies on a phylogenetic reconstruction that defines a base substitution rate on each branch of the phylogenetic tree and calls substitutions while measuring the distance between them. Each branch is then scanned to identify clusters of substitutions and recombination is inferred across branches or marked as an outside recombination site.

The second method fastGEAR (Mostowy et al., 2017), has four main steps which ultimately infer both *recent* and *ancestral* recombination events. Step one identifies sample lineages, step two identifies *recent* recombination events, step three identifies *ancestral* recombination and step four removes false positive recombinations by testing statistical significances. Steps one – three make use of the hidden Markov model (Husmeier, 2005) and is based on the recombination detection software STRUCTURE (Falush, Stephens and Pritchard, 2003).

The two different tools were used as they analyse the data and infer recombination by taking opposite approaches. Gubbins ultimately identifies substitution clusters within similar datasets that are flagged as outliers whereas fastGEAR locates similar genetic segments from diverse clusters and uses them to infer recombination. Combining both analyses allowed for a more comprehensive investigation of recombination between the input datasets.

Gubbins recombination detection pipeline:

```
mugsy -p {user defined output alignment prefix} --directory {path to directory containing assembled genomes from SPAdes to align} *.fasta

trimal -in {path to mugsy alignment} -out {user defined output file name} -gt .5

run_gubbins.py {path to trimmed alignment file from trimal output} --threads {user defined} --outgroup {name of reference to root the phylogeny} -i 20
```

Firstly, Mugsy (Angiuoli and Salzberg, 2011) was used to align the whole genome sequences of all single colony isolates isolated from each patient. Mugsy was chosen due to the balance between high throughput and alignment accuracy in terms of comparable alignment tools (Angiuoli and Salzberg, 2011). Furthermore, Mugsy does not depend on a reference sequence to inform an alignment of multiple genomes which was preferable in this study due to the potential increase in perceived intra-species diversity by the use of a diverse reference (Deloger, El Karoui and Petit, 2009). Mugsy also allows for genomic rearrangements, inclusion of gene loss and gain as well as duplications, all of which could be part of strain diversity.

A tool named trimAl (version 1.2) (Capella-Gutiérrez, Silla-Martínez and Gabaldón, 2009) was then used to lower the fraction of genomes with gaps allowed in their alignments to 50%. This was ultimately done to reduce the number of ‘gaps’ observed in the figures produced by Gubbins and fastGEAR and lowered the total genome sizes closer to the expected size of 1.67 Mbp. Furthermore, if only one sample was to hold an additional gene then it would be impossible to determine recombination within this

region as there would be no comparable genetic sequence to infer recombination. Depending on the sample size, this threshold can be lowered but 50% was deemed appropriate for this study.

Gubbins was enacted with default parameters with the exception of the number of iterations performed which was increased to 20, bringing it in line with the number performed by fastGEAR.

FastGEAR recombination detection pipeline:

```
mugsy -p {user defined output alignment prefix} --directory {path to directory containing assembled genomes from SPAdes to align} *.fasta

trimal -in {path to mugsy alignment} -out {user defined output file name} -gt .5

run_fastGEAR.sh {path to MATLAB version 901 runtime component} {path to input alignment file from Mugsy output} {output file and path} {path to input specifications file} ;

run_plotRecombinations.sh {path to MATLAB version 901 runtime component} {path to run_fastGEAR.sh output file} 1 1 ;

run_plotColors.sh {path to MATLAB version 901 runtime component} - {path to run_fastGEAR.sh output file} ;

run_plotMarginalsForStrain.sh {path to MATLAB version 901 runtime component} {path to run_fastGEAR.sh output file} 1 0
```

The pre-processing of recombination detection by fastGEAR was identical to the Gubbins pipeline (see above) for the whole genome alignments and alignment gap processing.

The fastGEAR script was then used to process the alignment file using a modified specification file as script arguments. The following specifications was used:

```
20 # Number of iterations
15 15 20 30 # Upper bound for the number of clusters (possibly
multiple values)
0 # Run clustering for all upper bounds (0=no / 1=yes)
- # File containing a partition for strains
0 # 1=produce reduced output, 0=produce complete output
```

The recombinations were then visually plotted, labelled and the marginal probabilities were plot using the `run_plotRecombinations`, `run_plotColors`, `run_plotMarginalsForStrain` scripts respectively.

Ancestral recombination was detected following the same pipeline but with a modified `run_plotRecombinations` command. This was:

```
run_plotRecombinations.sh {path to MATLAB version 901 runtime
component} {path to run_fastGEAR.sh output file} 2 1 ;
```

It is important to note that *ancestral* recombination does not necessarily mean a recombination event that occurred distantly in the past, rather it is a recombination event that has occurred on all samples in the data. Conversely, *recent* recombination detected by fastGEAR does not necessarily mean it was a very recent recombination event, rather it is a recombination event that has not occurred in all lineages.

5.2.10. Comparing and validating the deep sequencing minor allele detection with the single colony isolate variants

The curated reads obtained from the single colony isolates for each patient sample location (antrum and corpus) were mapped to the population deep sequenced consensus genomes of the corresponding patient sample locations. This was done to investigate and validate the results of the population deep sequenced minor allelic calling pipeline (Chapter Four, section 4.2.1.1).

The corresponding curated single colony sequencing reads (Chapter Two, section 2.9) were mapped to the antrum or corpus deep sequenced *de novo* consensus genomes (Chapter Two, section 2.11) using Snippy (Seemann, https://github.com/tseemann/snippy). The following parameters were used:

```
snippy --cpus {user specified CPU/thread/core number} --report --minfrac 0.9 --mincov 6 --mapqual 34 --basequal 30 --ref {patient antrum/corpus consensus genome GenBank annotation file} --outdir {user specified output directory} --R1 {path to forward/mate pair 1 of a single colony isolate} --R2 {path to reverse/mate pair 2 of a single colony isolate}
```

The minimum proportion of reads mapping to reference base calling an alternative base had to be greater than 90% to be called as a variant position with a coverage of 6 or more reads. A mapping quality of 34 was selected so that the parameters matched with the minor allele calling pipeline (Chapter Four, section 4.2.1.1) including a base quality of phred 30.

The resulting variant call format files were processed with vcflib (version 1.0.0) (Garrison, https://github.com/vcflib/vcflib) to select only fields containing SNP variant sites. The following command was executed:

```
vcffilter -f "TYPE = SNP” {path to VCF file from Snippy output} > {path to filtered VCF output}
```

Each filtered VCF file was further filtered by manual inspection as described in Chapter Four section 4.2.1.1. Briefly, variant sites were removed if they were identified in the first or last 500 bp of a contig. Next, the mapped reads in the BAM alignment files were loaded into Artemis (Carver et al., 2012) alongside the reference annotated genome. Variants located in a repeat region of 6 nucleotides or more were identified and inspected. These variant sites were removed if the reads supporting the alternative call were within the first or last 3 bases of the reads or if the supporting reads held other variants within a 20 bp range of the variant site. The support of just one read that passed these criteria was deemed sufficient evidence to classify the alternative call as valid, even if all other reads failed the measure.

This additional filtering was carried out to ensure the resulting variant files were standardised and comparable to those output from the minor allele pipeline (Chapter Four, section 4.2.1.1). The total number of variants were recorded for each single

colony isolate mapped to the corresponding patient sample deep sequenced consensus genome.

All single colony and antrum/corpus minor allele called VCF files were compressed with samtools (Li et al., 2009) `bgzip` and indexed with `tabix`. Each single colony variant file was used to intersect the minor allele called variant file of the associated patient sample location using bcftools which is part of the samtools package. The `-f` and `-c` flags were used to ignore column discrepancies plus variant call format version differences and to output unique records found in the second input file respectively. The following command was used:

```
vcf-isec -f -c {single colony filtered VCF file} {minor allele called VCF file for corresponding patient sample location} > {path to output file of single colony unique SNPs}
```

The resulting output files contained the single colony specific SNPs not identified by the minor allele called deep sequencing pipeline. The number of SNPs identified by each single colony and not by the deep sequencing pipeline were recorded.

The number of uniquely identified sites of variation by the single colony isolates was calculated to determine the number of sites potentially missed by the minor allele calling pipeline. This was done by concatenating all single colony output variant call format files from the previous command for each patient location and removing the duplicate positions. The resulting unique SNPs were recorded. The following command was used:

```
bcftools concat -a -D {path to all single colony isolate output files from the previous command} > {path to variant call format file containing unique sites of variation not identified in the deep sequenced minor allele detection pipeline}
```

The `-a` flag was used to allow for correct processing of the concatenation of the files so that overlaps could occur whereby the first coordinate of the next file to be processed can precede the last record of the current file. The `-D` flag was used to remove duplicate variants.

As some single colonies could hold duplicate SNPs between those isolated from the same niche, they were further investigated to calculate the total number of uniquely identified sites of variation so that they could be compared to the variant sites detected by the minor allele pipeline. This was achieved by combining all single colony variant files (VCF files) and removing all duplicate SNPs, leaving only unique variants. The total number of single colony detected sites of variation were recorded. The following command was executed through the samtools package, following the same format as the previous command:

```
bcftools concat -a -D {path to all single colony variant call format files from one patient sample location} > {path to output variant call format file containing unique list of variants detected by all single colony isolates}
```

The detection potential from the minor allele deep sequencing pipeline and the single colony isolates were compared by recording the total number of the deep sequencing unique, single colony unique and unique concordantly detected sites of variation. These were drawn as a venn diagram using the R statistical software version 3.5.1 (R Core Team, 2018) by the VennDiagram package version 1.6.20 (https://www.rdocumentation.org/packages/VennDiagram).

5.3. Results and discussion

All of the analysis presented in this Chapter was done on all patients and samples used within this study (table 5.1). Where appropriate and convenient, in some results figures and/or tables the representative patient 439 sample data were presented. All other patient results can be found in the appendix as stated below. However, full patient and sample results are displayed where convenient.

5.3.1. Detection of contaminating non-*Helicobacter pylori* sequences

Contamination detection is much easier using a single colony isolation methodology. This is because an experienced microbiologist can usually identify single colonies on a culture plate and distinguish between target and contaminant colonies based on a number of different criteria. From a population sweep this is much harder to do as the culture plate is usually heavy with bacterial growth, potentially disguising contaminant

bacteria. Visually, colony morphology that encompasses shape, depth, colour and gloss is often the first indicator of the target organism. Secondly, the biochemical characteristics of certain bacterial species can be exploited to indicate the target organism. Such is the basis of indicator and selection culture media.

Despite the care taken to minimise the contamination risk of isolating single colony isolates and the aid of the rapid urease test, contamination can still be present in the samples. This could be due to a number of reasons. For example, the presumptive bacterial colony could have the same morphology as another bacterial species. Furthermore, a rapid urease positive result could occur if there was a mixed population of the contaminant and the target *H. pylori* or if other urease positive bacteria are colonising the stomach and/or the oral cavity (due to oral microbiota contamination of the gastric biopsy) (Osaki et al., 2008; Mora and Arioli, 2014; Brandi et al., 2006).

Therefore, contamination detection was conducted *in silico* by mapping all curated single colony sequencing reads to all bacterial, archaeal or viral genomes within the RefSeq database (https://www.ncbi.nlm.nih.gov/refseq/). This method is described in Chapter 2 section 2.10. This output statistics from Kraken (Wood and Salzberg, 2014) which revealed that >95% of single colony isolate reads mapped to *H. pylori* reference genomes in all isolates with the exception of samples 322A1, 322A2, 322A3, 322A4, 322C5, 322A7 and 322C7, in which between 92.53% and 94.77% of reads mapped to *H. pylori* reference genomes. However, at least 95% of reads mapped to references from the *Helicobacter* genus in all cases. Therefore, these isolates were retained in this dataset as *H. pylori* has a very diverse genome and accessory genes which could all contribute to a drop in the proportion of reads mapping to the reference genomes. Furthermore, the percentage of unclassified reads were not close to comparable with the contaminated deep sequenced 308A and 326A samples where 42.11 and 38.33 percent of reads did not map to any bacterial, archaeal or viral reference genomes. In comparison, the datasets generated from isolates 322A1, 322A2, 322A3, 322A4, 322C5, 322A7 and 322C7 held between 3.57% to 4.88% of reads that failed to map to a reference genome. As isolates from both the antrum and corpus of patient 322 all showed a similar percentage of reads mapping to the reference *H. pylori* genomes it was shown as further evidence of inclusion of these samples as it would be unlikely for all 12 single colony isolates to be contaminants. Finally, in comparison to Kraken results from the deep sequenced antrum and corpus populations from patient 322 (Chapter

Four, table 4.1) a comparable percentage of mapped reads was shown. This is perhaps evidence at a low resolution of the diversity observed in the *H. pylori* species.

The assembly statistics were calculated for all single colony isolates, excluding 565C1 and 565C6 due to the unusually large genome sizes (n=117). The mean number of contigs was 39.37 (95% CL: 37.96 – 40.78) while the N50 mean was 93,899 (95% CL: 89,316 – 98,481) and total assembly mean length of 1,609,285 (95% CL: 1,602,833 – 1,615,738). These statistics were plot in figure 5.1.

The assembly statistics were generally comparable to the deep sequencing patient population consensus genome assemblies, but with some important differences (Chapter Four, figure 4.2). The mean number of contigs was 50.11 lower which might reflect the clonal nature of the single colony sequencing dataset. As each single colony isolate dataset derived from one single strain the SPAdes assembler was likely able to resolve more contig boundaries due to less read variation. The N50 was very similar (346 bp difference) indicating that the deep sequenced consensus assemblies were comparable to single colony assemblies. The N50 statistic refers to the minimum contig length that is able to cover 50% of the genome. Therefore, despite the inclusion of more read complexity by the addition of all population reads in the deep sequenced dataset, the genomes were comparable to those generated by single colony sequencing. These comparisons start to validate both methodologies.

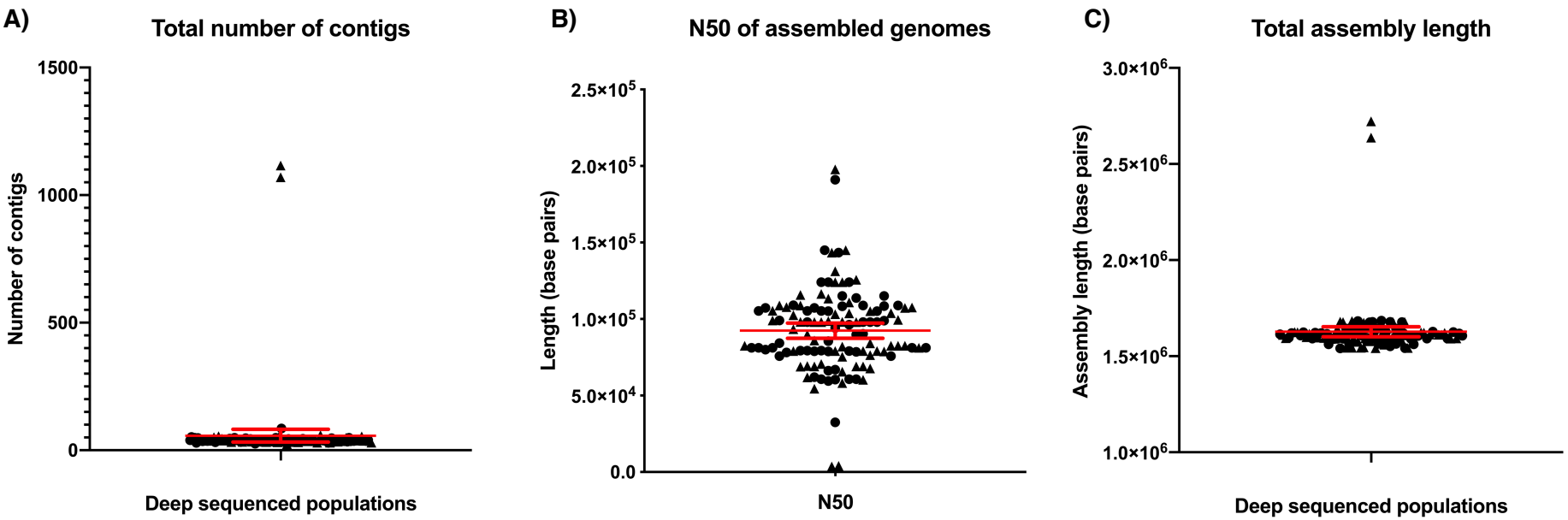
The difference between the genome length means of the deep sequenced populations and single colony isolates was 21.288 kb. This was unsurprising as the deep sequenced population dataset might be expected to have much higher genome lengths because accessory genes within the populations should be better captured with the population deep sequencing method. However, the reads covering these accessory genes might be in low abundance within the population read dataset, causing assembly issues. Nonetheless, the difference between the average genome lengths between the sequencing methods might be reflecting a low number of possible accessory genes within the deep sequenced populations.

Despite the outlier single colony isolates 565C1 and 565C6 with an unexpected genome length approximately 1 Mb higher than expected, there was no evidence of contaminated reads within the read sets. The Kraken contamination detection pipeline (Chapter Two, section 2.10) showed that 98.84% and 99.11% of reads mapped to the

H. pylori reference genomes respectively. These two isolates also had an unusually high number of contigs (565C1 – 1,501; 565C6 – 1,295) and low N50 (565C1 – 3,665; 565C6 – 3,977) which can be identified in figure 5.1A-B. One explanation to this could be the percentage of reads identified within these samples that might result in low coverage of the target genome. However, samples 565C1 and 565C6 had a share of 2.575% and 2.9246% respectively of the Illumina pass filtered sequencing reads generated by the MiSeq with the average share of sequencing reads per sample of 2.0239%. The most likely explanation of these outlier genomes is the possibility of inaccurate isolation of a single colony isolate, resulting in a non-clonal stock of two or more strains that are genetically distinct. This could result in the SPAdes assembler being unable to converge on a more contiguous assembly, causing an increase in the number of contigs observed in these sample.

Due to the helical shape of *H. pylori* it might be possible for individual bacterial cells to intertwine with one another. This might not be limited to two intertwined bacterial cells but numerous cells. If this was to occur, a single colony observed on a solid-state culture medium might look like any other *true* single colony isolate as the entwined cells will essentially have the same centre point of growth. To investigate this in more detail, the stocks for 565C1 and 565C6 would need to be streaked out to purity and new single colonies picked and sequenced. However, the deep sequenced corpus population of patient 565 also resulted in a higher than expected genome size, which supports the hypothesis that this patient was infected by more than one strain of *H. pylori*.

Figure 5.1 Quality statistics for all single colony isolate sequenced *de novo* assembled genomes



This figure depicts the quality metrics of the single colony assembled genomes. Figure A – total number of contigs; figure B – N50 of assembled genomes; figure C total assembly length. This dataset includes all single colony isolates (n=119). Circle = antrum, triangle = corpus. The mean and 95% CL are displayed as horizontal red bars. This figure was plot using GraphPad Prism (version 8.2.0).

5.3.2. Within patient phylogenetic analysis of single colony isolates

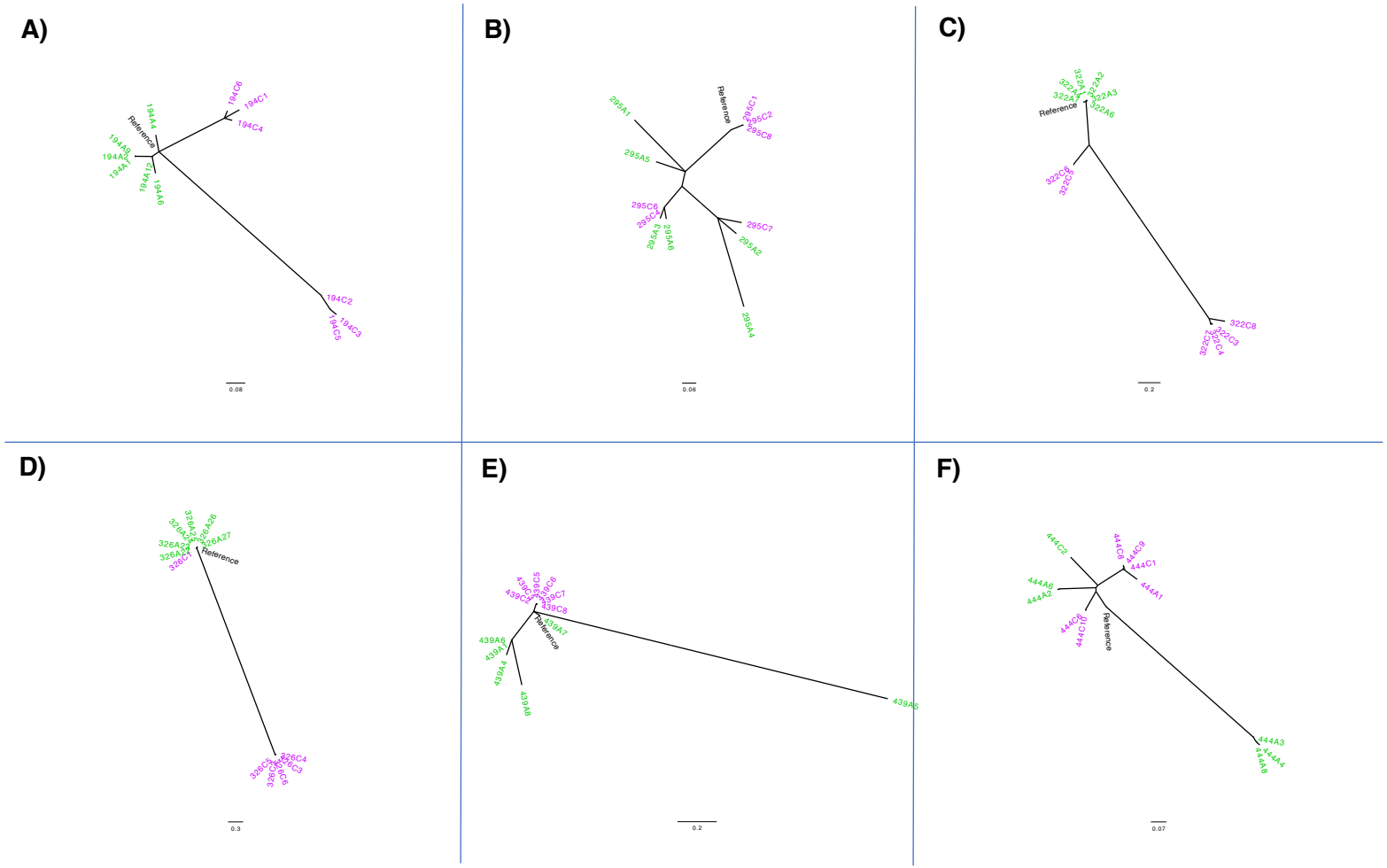
The single colony isolates from patients 326 and 732 show a good phylogenetic clustering of antrum and corpus clades. Patients 194, 322, 439 and 565 show a general clustering pattern between the antrum and corpus but usually with one additional cluster. For example, the antrum isolates from patient 322 have all clustered together, but the corpus isolates have separated into two separate clusters. In this case, it would suggest that the isolates isolated from the corpus are much more genetically diverse than the short-branched antrum cluster. However, the opposite is true for patient 439. The phylogeny of antral and corpus isolates isolated from patients 295 and 444 do not indicate an obvious clustering topology. This could suggest that the strains taken from these patients are highly diverse or that there is little genetic difference between the isolates taken from the antrum and the corpus. Study by Cao et al. (2015), found that strains taken from a single patient biopsy separated into two phylogenetic clades were of a mixed strain infection, which might suggest that the different clades are potentially an indicator of a mixed strain infection.

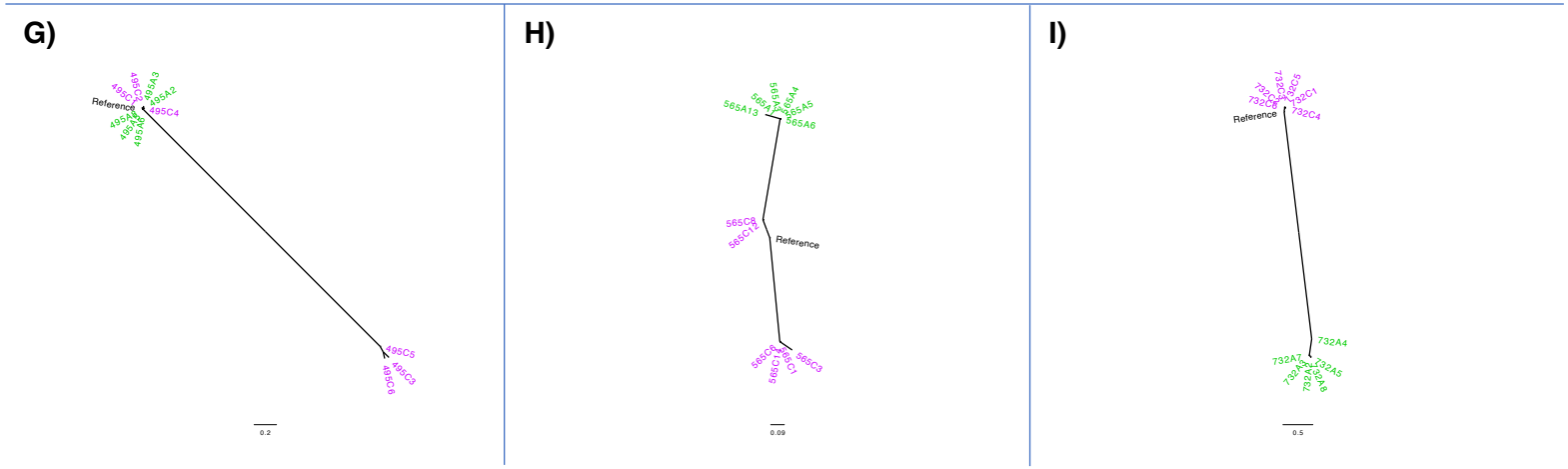
The phylogenetic structures of isolates obtained from the antrum and corpus closely match those observed by Ailloud *et al.* (2019), whereby some patients have a good antrum/corpus clustering while others are not as obviously split. Ailloud *et al.* (2019), also highlighted that migration of *H. pylori* strains from the antrum to the corpus are relatively infrequent whereas migration between the corpus and fundus is a more common event. They argued that the differences between the antrum and corpus niches hamper the migration of strains between the colonised environments. The phylogenetic tress displayed in figure 5.2A-H potentially highlight that where an antrum cluster is present, corpus strains are more frequently observed within antrum clades than the reverse to this. Particular examples of this can be seen in figures 5.2D (patient 326), figure 5.2G (patient 495) and more loosely associated in figure 5.2B (patent 295). The only exception to this is the 439A isolate clustering amongst the corpus isolates in patient 439 (figure 5.2E). This might suggest that while migrations between the antrum and corpus are less frequent, the corpus isolates are more likely to migrate to the antrum than the other way around. Again, this could be due to the differences between the antrum and corpus environments where antrum isolates are less fit or poorly adapted to colonise the harsher oxyntic epithelium whereas the corpus strains are able to colonise the more neutral antrum glands. Additionally, the antral niches of these patients held higher Sydney scores by histological examination compared to the corpus

niche, which might suggest higher histological scored niches allows for the colonisation of strains from other stomach niches. However, more patients and patient strains should be included to investigate this observation further and the biological significance investigated.

An alternative explanation to the non-clustering of antrum and corpus strains in patients 295 and 444 could be due to the biopsy sampling location and method. Fung *et al.* (2019), showed how founder strains initially colonise glands and then spread to adjacent glands in the immediate vicinity. This creates islands of closely related *H. pylori* strains, and where island boundaries occur the inhabitants may then compete for space. At these boundaries there are mixed strain glands or adjacent glands containing the different strains side by side. Furthermore, the transaction zone between the antrum and corpus is usually more of a mixture of *H. pylori* strains. Taken together, this may mean that if a biopsy is taken closer to a transition zone between the antrum or corpus or if a large biopsy is taken (that spans multiple strain islands) then the observed diversity might be greater. This could potentially disrupt the genetic clustering and phylogenetic topology of the antrum and corpus. Therefore, a standardised sampling location within the antrum and corpus might be beneficial to this type of analysis. However, this may not be implementable in practice as biopsies are usually taken from areas likely to harbour *H. pylori* infection, such as adjacent to visually diseased epithelium.

Figure 5.2 Phylogenetic analysis of all single colony isolates taken from paired antrum and corpus niches





Phylogenetic trees constructed as described in section 5.2.4 by approximately-maximum-likelihood. For improved visualisation, antrum-derived strains are shown in green and corpus-derived strains in pink/purple. The patient reference consensus genome is shown in black. Phylogenies were visualised in FigTree version 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/), rooted to the midpoint in decreasing order and drawn in a radial format.

5.3.3. Whole genome alignments of within patient single colonies taken from the antrum and corpus

The phylogenetic analysis of the *H. pylori* populations within individual patients highlighted a potential clustering of isolates within each niche (figures 5.3A, C – E, H – I). But this method was based on a core genome analysis and thus ignored all genes that were within the soft core and accessory genome. Therefore, this analysis alone could miss out important differences between the antrum versus corpus isolates. To address this, the single colony sequences from the antrum and corpus of patients with paired samples were aligned and visualised using a high BLASTN identity between 96% and 100%. This was done to identify small scales differences between the aligned genomes.

The representative patient example followed in this Chapter (patient 439) revealed that there were defined regions of 100% BLASTN identity in relation to the patient reference genome (figure 5.3). This generates a visual fingerprint that can differentiate the isolates taken from the antrum and corpus. However, there is also evidence of diversity *within* each niche. This analysis supports observations from the core genome phylogeny (figure 5.2E).

The deep population consensus genome alignments visualised using BRIG often highlighted regions of similarity and difference between the paired populations (Chapter Four, figure 4.9; Appendix, figures 11.4.21 – 11.4.35). This is supported by the single colony isolates where they are often grouped by their 100% BLASTN fingerprint. However, the perceived advantage of capturing the whole population diversity had a negative effect on the identification of potential pan-genome differences of the prevalent population and/or between strain differences. For example, if a single isolate within the antrum that was in low abundance harboured a gene that was not common within the population, this gene would still have been sequenced and potentially assembled into the consensus genome of the antrum population. This is further complicated when an accidental strain is taken from the corpus population due to human error such as during sample collection or biologically through migration or samples taken from strain island boundaries aforementioned (Fung et al., 2019). Furthermore, due to the more complex nature of whole genome assembly from a potentially mixed strain population, it is possible that not all genes within the population were assembled into the consensus

genome. In these examples, key strain differences within the population could be missed or overlooked.

The advantages of using a combination of deep sequencing and single colony methodology start to show in these analyses. In figure 5.3, there is a section of the genome that is either missing or has a BLASTN identity <96% approximately between positions 1.29 – 1.325 MBp in relation to the reference genome. This accounts for a missing genomic length of around 35 kb. This region appears to be missing or of low BLASTN identity in all corpus-derived strains and one antrum-derived strain from patient 439 but is present in the majority of the antrum-derived isolates (n=5). However, there is no evidence of a segment of this size missing from 439C in the comparison of the deep sequenced 439A and 439C consensus populations (Appendix, figures 11.4.31A-B). This shows that although sequencing of single colony isolates has utility in identifying this type of antrum-corpus difference, deep sequencing of the whole population completes the picture. Clearly, some isolates within the corpus must harbour this missing or higher identity genomic segment that is prevalent in the antrum-derived colonies. But, by sequencing a low number of single colony isolates it was possible that isolates harbouring this genetic segment were simply missed due to the unrepresentative sample size. This is better captured from a deep population sequencing approach as the entire genomic content of the population is captured.

On further examination of this region (1.29 – 1.325 Mbp to the reference genome), most of the genes were hypothetical (n=25). However, genes *xerH*, *virC1*, *traG*, *virB11*, *virB10*, *topA* and *ptlC* were annotated by PROKKA (Seemann, 2014). As these genes are not present within some strains they are likely to be mobile genetic elements. It is not possible from this analysis to determine whether these mobile genetic elements are associated with, a plasmid, bacteriophage or transposable element.

However, *xerH* is a recombinase gene shown to aid in chromosome segregation and colonisation. Knockout *xerH* mutants have been shown to affect colonisation and to have a higher sensitivity to ciprofloxacin, potentially attributed to the function of DNA joining and repair (Debowski et al., 2012). The strains harbouring this genomic segment harboured an additional *xerH* gene which could potentially cause differences in resistance and colonisation. Despite this, there were no obvious differences in levofloxacin sensitivity, at least at the population level, between the antrum- and corpus-derived *H. pylori* isolates from patient 439 (Chapter Three, figure 3.1). However,

another DNA damaging, and fragmentation associated antibiotic (Sisson et al., 2000; Jenks and Edwards, 2002), the antrum-derived *H. pylori* population was more resistant to metronidazole than the corpus-derived population (Chapter Three, figure 3.1). The additional *xerH* gene could potentially be contributing to the higher metronidazole resistance within this population, but experimental investigations would be needed to prove this.

The *virC1* gene is well defined in *Agrobacterium tumefaciens* and is associated with DNA transfer contributing in genome plasticity (Atmakuri et al., 2007). Mutation of *virC1* have been shown to reduce pathogenicity of *A. tumefaciens* in plants. The role of this gene in *H. pylori* is not fully understood but it is possible that it contributes to virulence, as part of the alternative type IV secretion system *tfs4* (Alandiyjany et al., 2017).

TraG like proteins in *H. pylori* have been shown to bind DNA non-specifically and are essential in DNA transfer by conjugation (Schröder et al., 2002). Therefore, the presence of this gene could suggest that these genes are part of a *H. pylori* acquired plasmid and would explain why some strains harbour this while other do not. Furthermore, the predominant absence of this presumptive plasmid in the corpus strains could be further evidence of antrum and corpus specific populations/divisions, albeit with acquired differences. The presence and absence of plasmids or even plasticity zones could potentially be used to track migration of *H. pylori* between niches of the human stomach and might explain why some, very closely related strains, can infect one niche but are not found in as high abundance in others, due to the acquisition of beneficial genes.

The *virB11* gene is usually associated with the *cag*PAI and codes for an ATPase that provides energy for T4SS apparatus assembly and/or CagA transport (da Costa, Pereira and Rabenhorst, 2015). The *virB11* and *cagE* genes are genetically linked and are thought to be associated with increased gastritis, particularly through increase of IL-1B and IL-8 (de Negreiros Bessa et al., 2014). The presence of the *virB11* gene could suggest that the antrum strains have a more complete *cag*PAI. This might have implications in virulence differences between the strains, and more widely between the antrum and corpus.

The *virB10* gene is essential for T4SS function, encoding a protein integral to the structure of the system and interacting with other T4SS encoded proteins both

functionally and structurally (Terradot et al., 2005). This is further evidence that the *cag*PAI is potentially more complete in the antrum isolates then in the corpus and again, could provide differences in virulence between them.

The virulence associated *topA* gene works by inducing negative DNA supercoiling in replication (Nitharwal et al., 2011; McNairn, Bhriain and Dorman, 1995). It has also been found to be highly conserved and positionally linked with the *flaB* gene and thus the expression of one gene effects the other (Suerbaum et al., 1998). Again, this might suggest that there are differences between the two antrum and corpus niches, in this case colonisation and motility differences. The *topA* gene was found to have only one promotor, low in comparison to the three promotors of the *topA* gene in *E. coli* (Suerbaum et al., 1998; Qi, Menzel and Tse-Dinh, 1997). Therefore, an additional copy of the *topA* gene within the strains isolated from the antrum could be conferring further gene expression relationships, potentially attributing to differences in virulence, colonisation or resistance.

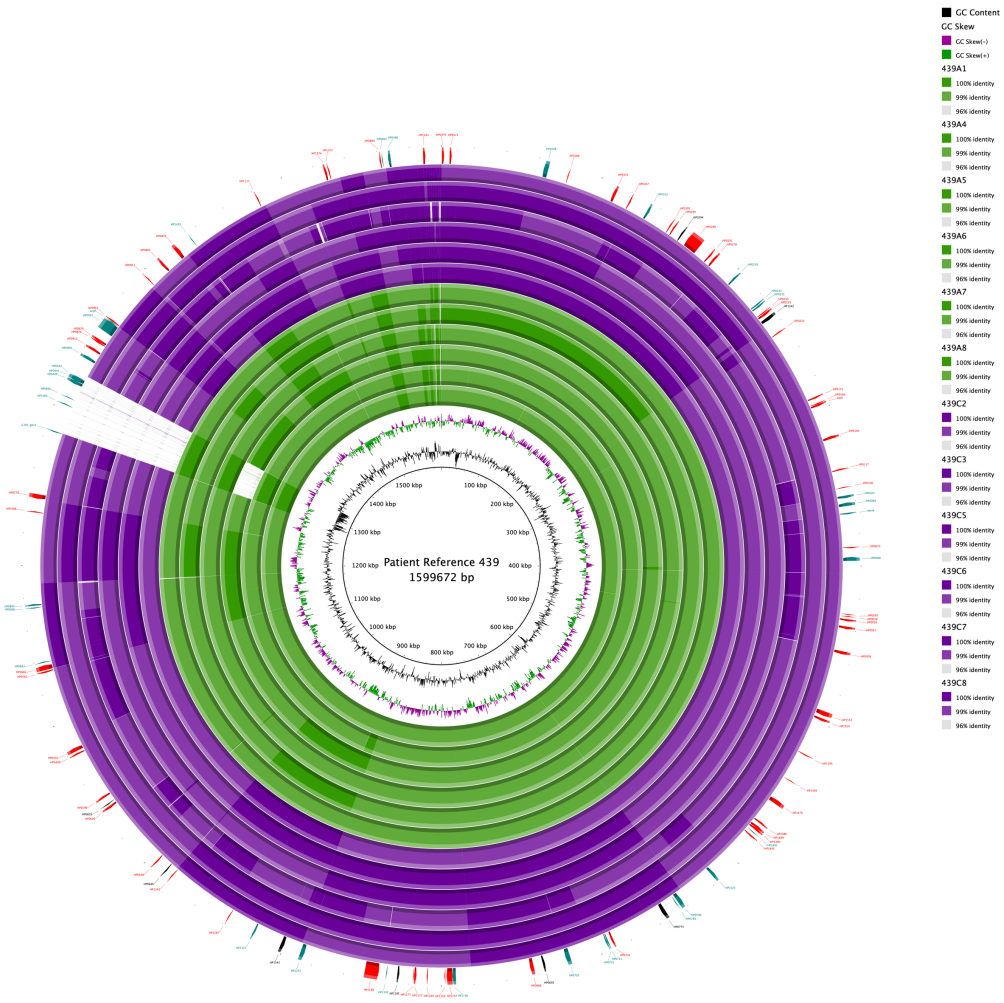
The *ptlC* gene has close homology to the *cagE* gene and this might explain why it was clustered with the *virB11* and *virB10* genes, perhaps providing further evidence of these genes being associated with a complete *cag*PAI (Censini et al., 1996). However, a *Helicobacter ptlC* gene has previously been identified in a *H. pylori* strain and has close homology to the *Bordetella pertussis* toxin. This toxin was shown to increase inflammation by induction of IL-8 (Tummuru, Sharma and Blaser, 1995).

Taken together, it is clear that there are gene content differences between strains and potentially between closely related strains taken from different environments within the same patient's stomach. These differences are potentially overlooked in the deep sequence population datasets because the presence of a single strain harbouring a gene could incorporate this gene into the consensus genome assembly of the population.

Although the discussion above has focused on strains isolated from patient 439 as a representative, this is not an uncommon observation across the wider dataset from all patients. Gaps and/or genes with <96% identity to the reference were observed in patients 194, 322, 326, 439, 495, 565 and 732. The highest density of and number of gaps were observed in patients 565 and 732. All of the analyses so far indicate that

these two patients potentially harboured multi strain infections that diversified over the course of their chronic infections.

Figure 5.3 Single colony isolate comparisons from paired antrum and corpus niches of patient 439



BLAST Ring Image Generator (BRIG) plots of paired antrum and corpus consensus genome assemblies constructed as described in section 5.2.6. The rings represent the following from the centre most ring outwards; GC percentage, GC skew, 439A1, 439A4, 439A5, 439A7, 439A8, 439C2, 439C3, 439C5, 439C6, 439C7 439C8, gene direction and gene names. The green concentric rings represent the strains isolated from the antrum and the purple rings represent the corpus-derived strains. Genes were colour coded as follows; black = identified as variable gene(s) by mapping single colony reads to the within niche deep sequenced consensus genome only, teal = identified allelic genes by the deep sequencing minor allele pipeline only, red = identified by both methodologies. An upper BLASTN identity threshold of 99% and lower identity of 96% were used. BRIG plots for all other patient single colony datasets presented in this study can be found in the Appendix figures 11.5.1 – 11.5.8.

5.3.4. Pan-genome analysis of within patient strains isolated from the antrum and corpus

Following on from the phylogenetic analysis (section 5.3.2) and the whole genome alignments of single colony isolates taken from each patient (section 5.3.3) questions arose around potential accessory gene differences between antrum and corpus populations. Therefore, an analysis of the pan-genome was carried out.

All single colony isolates were analysed together by Roary (Page et al., 2015) which clustered all genes and a core genome assembly was constructed by approximately-maximum likelihood (Treangen et al., 2014). This was done to see how the core genome phylogeny linked with the clustering of the genome wide gene clusters. This analysis is depicted in figure 5.4.

The pan genome analysis (figure 5.4) supported the analysis observed in the deep sequencing dataset (Chapter Four, figure 4.13), whereby a unique fingerprint of gene presence and absences was observed for strains derived from each patient. However, there were no obvious gene presence/absence differences between antrum and corpus isolates, at least on visual inspection. This may be due to the density of the pan-genome across all samples, making visual identification of small differences more difficult. Furthermore, the phylogeny, which orders the gene presence/absence heat map, did not always infer an obvious antrum-corpus phylogenetic split within individual patients. This is likely due to the smaller gene content that makes up the core genome of all the combined isolates across multiple patients, making small scale differences harder to observe. This is further compounded by the diversity of *H. pylori* isolates between the different patients. Additionally, the core genes shared across one patient's samples that ultimately contributed to antrum-corpus differences, could have been excluded from the analysis if these genes were not shared by all other isolates in the dataset.

Despite these limitations, this visual analysis (figure 5.4) does highlight some interesting observations. For example, the suspected mixed *H. pylori* samples 565C1 and 565C6 showed a close gene presence/absence pattern, but also harboured different genes. These strains still clustered on the patient 565 branch suggesting that they are more closely related to these patient samples. But, samples 565C3 and 565C14 were also branched off from the majority of strains along with 565C1 and 565C6. This might suggest that generally, the corpus isolates from this patient are much more diverse than

the antrum isolates, not just allelically (Chapter Four, table 4.3) but also in their gene content.

The strains taken from patients 295 and 526 clustered together as did strains from 249C and 537A. This was an expected result because these strains are part of sequential datasets and will be further examined in Chapter Six.

To investigate the gene presence and absences between the antrum and corpus strains taken within individual patients, a genome-wide association study was conducted as described in section 5.2.8. All strains belonging to separate niches within patients harboured niche specific genes, except for patents 295 and 444 (table 5.2; Appendix table 10.5.1 – 11.5.6). This was based on the sensitivity (niche specific gene presence to determine niche-positivity) and specificity (using gene absence to determine niche-negativity) greater 80% with a naïve p-value of < 0.05 (the presence/absence of a gene is unrelated to the niche status). Due to the relatively low sample size, p-values were not adjusted for multiple comparisons.

The genes identified as potentially missing between antrum and corpus isolates in figure 5.3 were confirmed to be missing by the GWAS (table 5.2). Genes *HP1001*, *HP1005*, *HP0894*, *HP1006*, *pfdA*, *HP1421* and *HP1002* were all found to have a sensitivity score of 83.33% and specificity of 100% for the antrum isolates of patient 439. This equates to each gene being found in 5/6 antrum strains and not present in any of the corpus strains. A further 26 genes were identified as potentially niche specific and should be investigated further to determine any potential role in niche adaptation and/or survival.

All of the hypothetical genes with high sensitivity to the antrum should be investigated further to determine their importance (tables 5.2 and 5.3).

There were only two corpus associated genes in patient 439 (table 5.2). One was a hypothetical protein (*HP1283*) while the other was a second copy of a *pldA* like gene (phospholipase A). The *pldA* gene is a phospholipase and has been shown to play an important part in colonisation of the gastric mucosa with a potential role in epithelium tissue damage (Dorrell et al., 1999). The *pldA* gene has also been shown to be phase variable, resulting in strains that are low or high in lysophospholipids content with further associations in urease and VacA secretions (Tannaes et al., 2001). This phase variation

was shown to be selected for under lower pH conditions and might explain why a second copy of this gene was identified in the corpus strains of patient 439, due to the oxyntic corpus environment (table 5.2). Alternatively, the second copy identified in these strains could be due to active phase variation of the strains during culture, that have then been picked up during the subsequent sequencing of these strains, resulting in the observation of two genes. Nevertheless, this was not observed in the antrum strains and thus highlights a potential difference between these strains isolated from the different stomach niches.

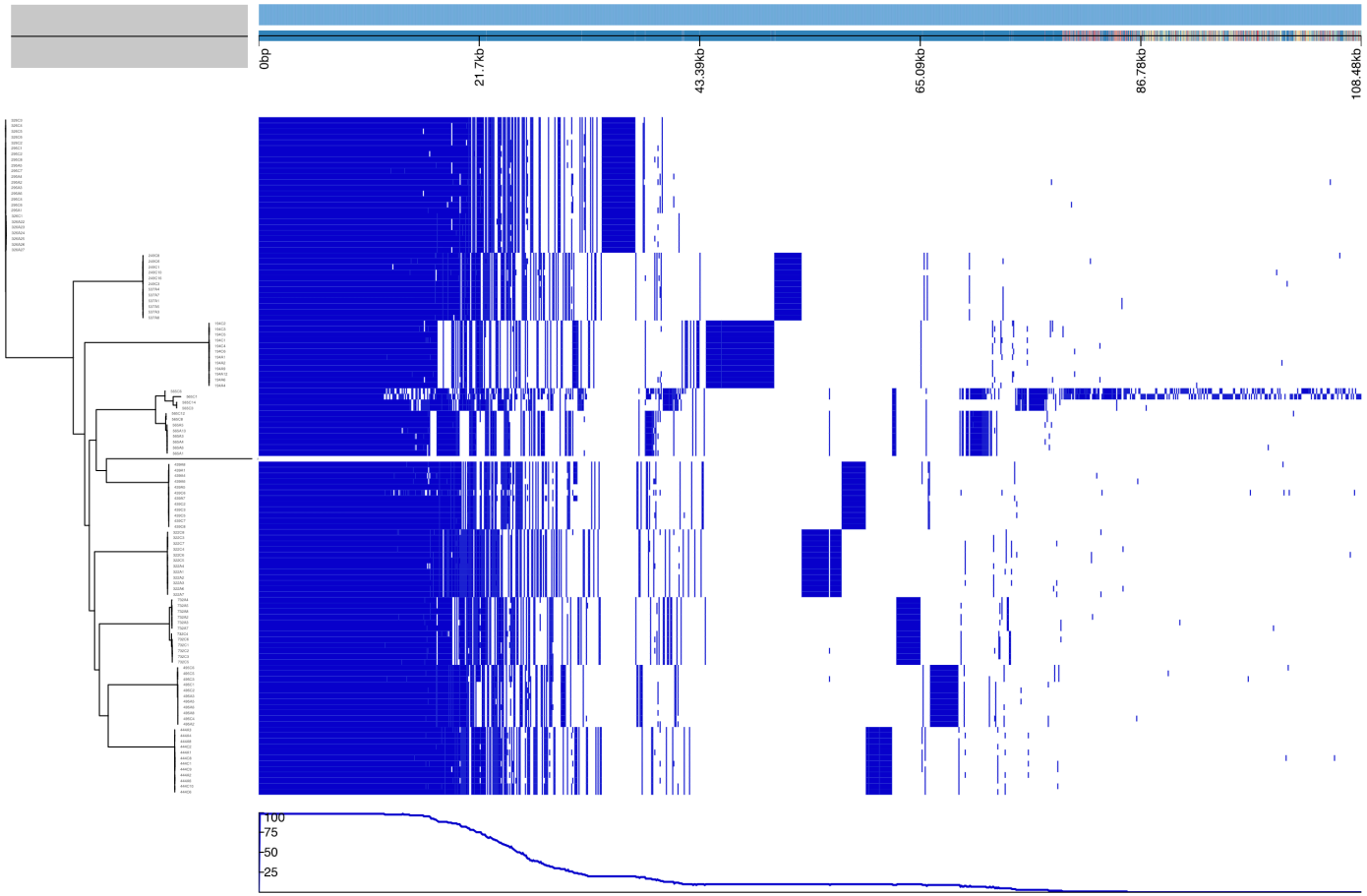
All of the antrum strains were compared against all of the corpus strains from each patient to determine if there were any universal antrum or corpus specific genes (table 5.3). No genes reached > 80% specificity and sensitivity in the antrum dataset suggesting that there are no clear antrum-associated genes. Furthermore, neither the Bonferroni correction nor the Benjamini-Hochberg correction for multiple comparisons which account for false positive and false negative adjusted p-values reached statistical significance, which likely reflects the low sample size used. However, genes *dxr* and a hypothetical protein (group 763) had relativity high sensitivity scores of 71.2% and 61% with low specificity scores of 6.7% and 13.3%, respectively. This suggests that while these genes are present in both niches, their presence or clustering is more variable in the corpus-derived isolates. While the hypothetical gene would need to be studied in more detail, the *dxr* gene is an essential gene in isoprenoid synthesis. Knockout mutants of *dxr* gene in *E. coli* have been shown to be lethal (Takahashi et al., 1998; Rodŕiguez-Concepción et al., 2000). Therefore, the potential higher variation of *dxr* in the majority of corpus strains is surprising. However, the group 136 gene in table 5.3 has a *dxr* non-unique gene name meaning that the same gene was identified by Roary but clustered separately (Page et al., 2015). Therefore, this gene is present in the corpus strains but clusters very differently to the antrum *dxr*. This might suggest that there are antrum and corpus low abundance strain specific *dxr* variants or that their position within the genome is different to that of presiding antrum associated strains. This is further supported by the high specificity of 95 and low sensitivity of the group 136/*dxr* variant gene cluster highlighting corpus isolate diversity of this gene. Further investigation into this finding is warranted. Especially as the dxr gene has been suggested as an antimicrobial target (Singh et al., 2007; Rodŕiguez-Concepción et al., 2000; Pérez-Gil et al., 2010).

The antrum strain dataset (table 5.3) also highlighted instances of 100% specificity but low sensitivity (8.4-10.2%). This suggests that there are low abundance strains with antrum specific genes. Most of these genes were hypothetical (n=11), but annotated genes included; *tlyA*, *selA*, *prmC*, *dxr*, *oppD* and *gsiA*. However, with only 5-6 strains harbouring these genes they could be associated with just one set of more clonal strains from one patient. This might suggest that these are most likely to be additional accessory genes within one patient population. However, as the gene names are non-unique, it would indicate that these genes have clustered independently and are potentially homologous but ultimately more genetically diverse, resulting in the observed sub-clustering (table 5.3).

Like the antrum strain GWAS, the corpus strain dataset had a number of genes with 100% specificity but low sensitivity (10%). This suggests that a low abundance of strains harbour corpus only genes. Again, the majority of these genes were hypothetical (n=8) but *HP1417m* (metal-dependent hydrolase), *HP1472* (type IIS restriction enzyme M protein Mod), *HP1513* (selenocysteine synthase SelA) and *HP0066* (ATP-binding protein) genes were annotated. Since the gene names are non-unique (*i.e.* gene names followed by '_'), it would suggest that these genes have clustered independently and are potentially not additional corpus associated genes but gene homologues.

While there were antrum and corpus specific genes identified from individual patients, no antrum or corpus specific genes could be identified from the dataset as a whole across all patients (table 5.3). Therefore, this suggests that where antrum/corpus gene differences are observed, these are patient-specific. In some patients, there was no evidence of antrum and corpus specific genes (samples 295 and 444)).

Figure 5.4 Pan-genome analysis of all single colony isolates used within this study



This figure was constructed as described in the materials and methods (section 5.2.7) and visualised by phandango (Hadfield et al., 2018). The left-hand side of the figure displays the core genome phylogeny. The bottom graph represents the percentage of the strains harbouring each specific gene. The top graph represents the accumulating total length of all genes combined. At the centre is the gene presence/absence indicated by blue (presence) and white (absence).

Table 5.2 Genome-wide association study of genes associated with the antrum-derived *H. pylori* strains from patient 439

		Antrum		Corpus	
Gene	**Annotation**	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
HP1283_2	Hypothetical protein	0.00	0.00	100.00	100.00
HP1001	hypothetical protein	83.33	100.00	0.00	16.67
virC1	PARA protein	83.33	100.00	0.00	16.67
group_1484	hypothetical protein	83.33	100.00	0.00	16.67
HP1005	PZ11b	83.33	100.00	0.00	16.67
HP0894	Addiction module toxin	83.33	100.00	0.00	16.67
HP1006	hypothetical protein	83.33	100.00	0.00	16.67
group_1766	hypothetical protein	83.33	100.00	0.00	16.67
group_2380	hypothetical protein	83.33	100.00	0.00	16.67
pfdA	hypothetical protein	83.33	100.00	0.00	16.67
group_2410	hypothetical protein	83.33	100.00	0.00	16.67
xerH	Tyrosine recombinase XerH	83.33	100.00	0.00	16.67
ptlC	VirB4-like protein	83.33	100.00	0.00	16.67
topA_2	DNA topoisomerase 1	83.33	100.00	0.00	16.67
group_2705	hypothetical protein	83.33	100.00	0.00	16.67
group_2745	hypothetical protein	83.33	100.00	0.00	16.67
group_2746	hypothetical protein	83.33	100.00	0.00	16.67
hisZ	hypothetical protein	83.33	100.00	0.00	16.67
group_2748	hypothetical protein	83.33	100.00	0.00	16.67
HP0446	hypothetical protein	83.33	100.00	0.00	16.67
group_3054	hypothetical protein	83.33	100.00	0.00	16.67

		Antrum		Corpus	
Gene	**Annotation**	**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
group_3105	hypothetical protein	83.33	100.00	0.00	16.67
HP1421_1	Type IV secretion system protein VirB11	83.33	100.00	0.00	16.67
group_3107	hypothetical protein	83.33	100.00	0.00	16.67
group_3108	hypothetical protein	83.33	100.00	0.00	16.67
HP0443	hypothetical protein	83.33	100.00	0.00	16.67
HP0442	hypothetical protein	83.33	100.00	0.00	16.67
group_1715	hypothetical protein	83.33	100.00	0.00	16.67
HP0499_2	Phospholipase A	0.00	16.67	83.33	100.00
group_3218	hypothetical protein	83.33	100.00	0.00	16.67
group_3828	hypothetical protein	83.33	100.00	0.00	16.67
group_4306	hypothetical protein	83.33	100.00	0.00	16.67
HP1002	hypothetical protein	83.33	100.00	0.00	16.67

Sensitivity and specificity of genes associated with either the antrum or corpus. Only genes with a naïve p-value (null hypothesis = the presence/absence of a specific gene is unrelated to the antrum/corpus status) < 0.05 are shown. Hypothetical genes were denoted by 'group' followed by a unique clustering number. Sensitivity; the sensitivity if using the presence of this gene as a diagnostic test to determine trait-positivity. Specificity; the specificity if using the non-presence of this gene as a diagnostic test to determine trait-negativity. Statistics and identification of antrum/corpus associated genes were calculated as described in the materials and methods (section 5.2.8). GWAS results for patient samples 194 (table 10.5.1), 322 (table 10.5.2), 326 (table 10.5.3), 495 (table 10.5.4), 565 (table 10.5.5) and 732 (table 10.5.6) can be found in the Appendix. Samples 295 and 444 did not identify any niche specific genes through the GWAS analysis.

Table 5.3 Identification of antrum- and corpus- associated genes from all patient isolates used in this study by genome-wide association study

Gene	Annotation	Antrum		Corpus	
		Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
dxr	1-deoxy-D-xylulose 5-phosphate reductoisomerase	71.19	6.67	93.33	28.81
group_763	hypothetical protein	61.02	13.33	86.67	38.98
HP0505	hypothetical protein	28.81	91.67	8.33	71.19
HP1471	Type IIS restriction enzyme R protein (BCGIB)	30.51	43.33	56.67	69.49
HP0980	RIP metalloprotease RseP	0.00	88.33	11.67	100.00
group_2834	hypothetical protein	10.17	100.00	0.00	89.83
HP1517_2	hypothetical protein	10.17	100.00	0.00	89.83
HP1471_1	hypothetical protein	10.17	100.00	0.00	89.83
selA	L-seryl-tRNA(Sec) selenium transferase	10.17	100.00	0.00	89.83
group_5051	hypothetical protein	10.17	100.00	0.00	89.83
HP0066_2	hypothetical protein	10.17	100.00	0.00	89.83
tlyA	16S/23S rRNA (cytidine-2'-O)-methyltransferase TlyA	10.17	100.00	0.00	89.83
group_5138	hypothetical protein	10.17	100.00	0.00	89.83
HP0629_2	hypothetical protein	10.17	100.00	0.00	89.83
prmC_3	Release factor glutamine methyltransferase	10.17	100.00	0.00	89.83
group_5171	hypothetical protein	10.17	100.00	0.00	89.83
HP0250_1	Oligopeptide permease ATPase protein OppD	1.69	85.00	15.00	98.31
dxr	1-deoxy-D-xylulose 5-phosphate reductoisomerase	18.64	95.00	5.00	81.36
HP1243	hypothetical protein	8.47	100.00	0.00	91.53
gsiA	Dipeptide ABC transporter- ATP-binding protein DppD	8.47	100.00	0.00	91.53

Gene	Annotation	Antrum		Corpus	
		Sensitivity (%)	Specificity (%)	Sensitivity (%)	Specificity (%)
group_2496	hypothetical protein	8.47	100.00	0.00	91.53
group_4306	hypothetical protein	8.47	100.00	0.00	91.53
HP1002	hypothetical protein	8.47	100.00	0.00	91.53
group_4775	hypothetical protein	8.47	100.00	0.00	91.53
group_209	hypothetical protein	0.00	90.00	10.00	100.00
HP1417m_2	metal-dependent hydrolase	0.00	90.00	10.00	100.00
group_2831	hypothetical protein	0.00	90.00	10.00	100.00
HP1283_2	Hypothetical protein	0.00	90.00	10.00	100.00
group_5181	hypothetical protein	0.00	90.00	10.00	100.00
HP1472	Type IIS restriction enzyme M protein Mod	0.00	90.00	10.00	100.00
HP1511	Hypothetical protein	0.00	90.00	10.00	100.00
HP1513	Selenocysteine synthase SelA	0.00	90.00	10.00	100.00
group_5186	hypothetical protein	0.00	90.00	10.00	100.00
HP0647	Hypothetical protein	0.00	90.00	10.00	100.00
HP0629_2	Hypothetical protein	0.00	90.00	10.00	100.00
HP0066_1	ATP-binding protein	0.00	90.00	10.00	100.00
acxA_2	Acetone carboxylase beta subunit	20.34	93.33	6.67	79.66

Sensitivity and specificity of genes associated with either the antrum or corpus. Only genes with a naïve p-value (null hypothesis = the presence/absence of a specific gene is unrelated to the antrum/corpus status) < 0.05 are shown. Hypothetical genes were denoted by 'group' followed by a unique clustering number. Sensitivity; the sensitivity if using the presence of this gene as a diagnostic test to determine trait-positivity. Specificity; the specificity if using the non-presence of this gene as a diagnostic test to determine trait-negativity. Statistics and identification of antrum/corpus associated genes were calculated as described in the materials and methods (section 5.2.8).

5.3.5. Within patient recombination

The isolation of multiple strains from the antrum and corpus of individual patients allowed for the investigation of homologous recombination and to determine if the populations recombine with one another.

A dual analysis approach was taken with differing algorithms as described in the materials and methods, section 5.2.9. This approach was taken to fully investigate the recombination of within patient isolates, because different methodologies can reveal differing results (Mostowy et al., 2017).

Recombination was observed between multiple strains in all patients by the Gubbins analysis (Croucher et al., 2015). Although recombination direction cannot be determined by this method, the continuous red (recombination shared between multiple isolates within the dataset) blocks observed between all strains from each patient infer that both within and between niche recombination might occur (figure 5.5A). This suggests that the antrum and corpus strains naturally interact, providing opportunities to recombine.

The Gubbins analysis for patient 439 (figure 5.5A) shows that while there appears to be shared recombination between strains from both the antrum and corpus (red blocks), this is mainly towards the end of the genomes. As Gubbins identifies regions of increased SNP densities between the genomes, the smaller contigs towards the end of the genome assembly are potentially more susceptible to variation due to the smaller contig sizes that usually have much higher average coverage from the *de novo* assembled genomes (an example of this can be observed in Chapter Four, figure 4.4). These contigs are likely of lower quality due to the artificial inflation of reads mapping to this region and potentially harbours more genetic variation. Therefore, the end of the genomes are likely to contain increased false positive detections of recombinant sites with this being a known limitation to this recombination detection methodology (Croucher et al., 2015). In consideration of this, there is a more defined antrum and corpus only strain recombination pattern, supporting the antrum and corpus strain division and inferring little between niche strain interaction (figure 5.5A). This is further supported by the fastGEAR analysis (figure 5.5B) where there is little to no inferred recombination between the lineages and between the antrum and corpus strains. The

only exception to this is one small blue antrum associated recombination block detected towards the end of the strain 439A7 (figure 5.5B).

Surprisingly, strains isolated from patient 732 that were shown to have a defined core genome-based population structure was inferred to recombine with strains from the opposite niche. The fastGEAR recombination analysis of patient 732 (Appendix, figure 10.5.16) indicates that there are shared recombination events between the antrum and corpus isolates identified by contaminating alternative lineage coloured blocks. These blocks are not defined towards the end of the genomes suggesting that they are not artefacts from genetic content contained in small contigs, providing evidence of between niche strain recombination (Appendix, figure 10.5.16). The flow of recombination seems to be mainly from corpus isolates (blue clade) to the antrum population (red clade). This observation perhaps shows corpus strain migration to the antrum is more common than antrum strain migration to the corpus. This might add to the findings of Ailloud *et al.* (2019), who found that 21% of migration events happen between the antrum and corpus.

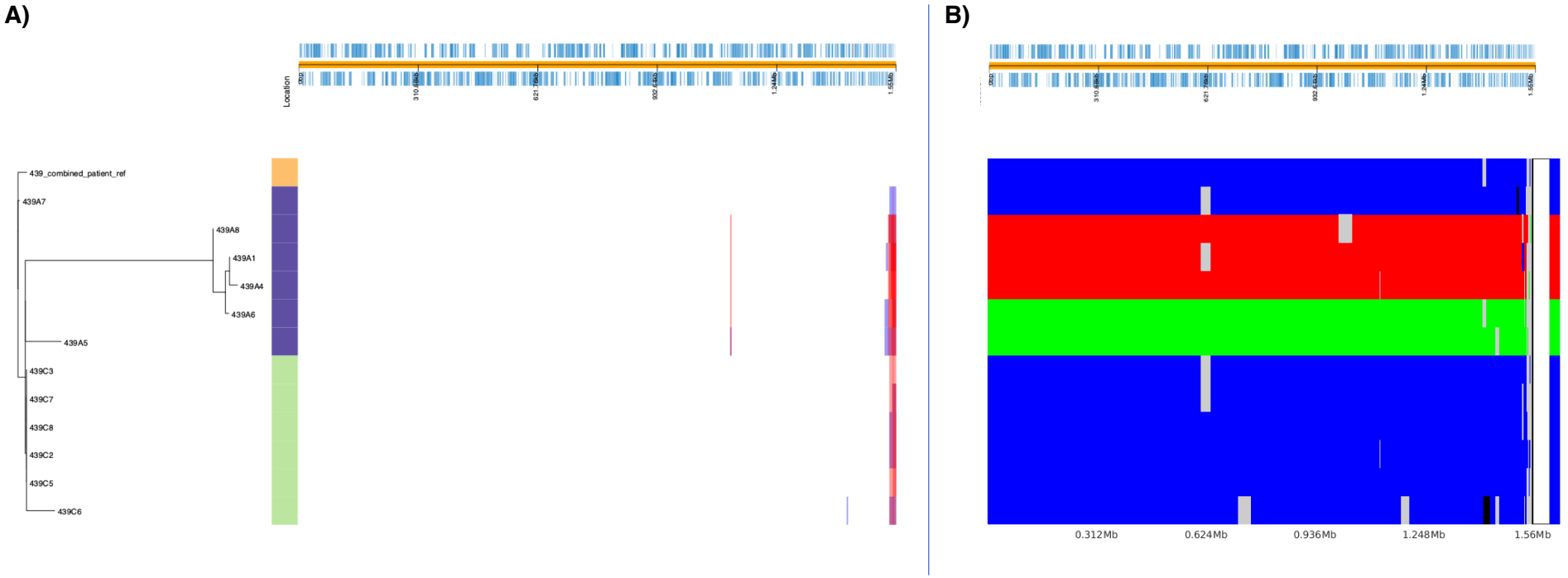
Strain 732A4 is also shown to harbour regions of recombination that have come from a strain outside the dataset (black blocks) by fastGEAR (Appendix, figure 10.5.16). These regions were not detected by the Gubbins analysis. This supports the dual recombination analysis approach and perhaps starts to explain why this strain is split on a different branch in the phylogeny (figure 5.2I; Appendix figure 10.5.16).

Single colony isolates from patients 194, 326, 439 and 444 were all shown to recombine between niches, but evidence was only observed towards the end of the genome sequences (figures 11.5.9, 11.5.12, 5.5 and 11.5.13 respectively).

Single colony isolates from the antrum and corpus were observed to recombine at genomic regions not defined towards the end of the genomes for patients 322, 565 and 732. This provided strong evidence of between niche interaction of strains. A study by Cao et al. (2015), investigated within niche microevolution of 18 strains taken from a single antral biopsy and found that recombination between clades representing a mixed strain infection was better identified than recombination between within-clade strains. This could suggest that patients 322, 565 and 732 hold a mixed strain infection.

No between niche recombination of isolates were observed for patients 295 and 495.

Figure 5.5 Recombination detection by Gubbins and fastGEAR – patient 439 single colony isolates



A) Recombination detected by Gubbins as described in the materials and methods (section 5.2.9). The genomic scale bar is depicted at the top of the figure with blue markings highlighting genes along the genome. The core genome phylogeny is depicted to the left followed by coloured blocks differentiating antrum (purple) and corpus (green) isolates. The reference strain is coloured orange. Figure centre depicts sites of recombination. White space (no recombination sites), blue blocks (single site of recombination) and red blocks (shared site of recombination). B) Recombination detected by fastGEAR as described in the materials and methods (section 5.2.9). Lineages detected by fastGEAR are colour coded and have been ordered to match the strain locations in the Gubbins analysis for ease of cross comparisons. Grey blocks represent recombination sites within lineages while black blocks represent recombination from an outside strain. Coloured blocks within lineages infer direction of recombination events between lineages. Results for all other samples can be found in the Appendix (figures 11.5.9 – 11.5.16).

5.3.6. Cross comparing the single colony identified SNPs with the deep sequenced population detected minor allele variants

The isolation of single colony isolates from the populations that were deep sequenced provided a unique opportunity to compare the minor allele population variants (Chapter Four, section 4.3.4) to the SNPs identified between the single colony isolates and the population consensus genome assembled from the deep sequencing data. This allows for the interrogation of results obtained from the minor allele calling pipeline to determine how well the data generated from each approach agree with each other.

If the minor allele calling pipeline was calling population variants as intended, then the majority of the single colony isolate variants should, in theory, be well captured. However, since the minor allele calling pipeline was designed with highly stringent criteria to reduce the false positive calling of variant sites, it is likely that the false negative rate was increased as a resulting trade-off. Therefore, it is important to evaluate the performance of this pipeline using real population data in the form of single colony isolates.

A total of 152 allelic variants were detected by the deep sequencing pipeline (table 5.4; Chapter Four, section 4.2.1.1) for the antrum population in patient 439 (Chapter Four, table 4.3). Mapping the single colony reads to the population consensus genome resulted in a range of isolate specific SNP counts (0 – 89 SNPs) (table 5.4). In total, 328 SNPs were detected by the minor allele calling pipeline, leaving only 39 SNPs across all 12 single colony isolates that were not picked up by this methodology (table 5.4). For this specific sample, this equated to an 89.37% detection of all single colony isolate SNPs by the deep sequencing minor allele calling pipeline. In terms of uniquely identified single colony variants across the six isolates, the minor allele calling pipeline missed 29 variants, equating to a detection rate of 83.98% (33% of unique variants detected by the minor allele calling pipeline only; 16.02% of variants missed by the minor allele calling pipeline; 50.28% of variants detected by both single colony and deep population analysis pipelines) (table 5.4).

While this was an impressive detection rate, it was not known to what extent the SNP counts were inflated due to appearing across multiple single colony isolates. For example, a total of 328 single colony SNPs were identified as concordant to the minor allele calling pipeline, but only 152 allelic variants were captured suggesting that this

detection rate might be different when duplicate SNPs were removed across the single colony isolates. This would allow for a more accurate detection as well as providing the proportion of low population allelic variants that were potentially filtered out of the allele calling pipeline.

Figure 5.6A represents the unique SNPs identified from the single colony isolates to the population consensus genome, compared with the minor allele variants detected by the population deep sequencing. There is a very good concordance between the two methodologies, validating the data generated using the different approaches. Approximately 50% of all detected SNPs were detected by both methodologies. However, only 16% of the SNPs were uniquely identified by the single colony isolate method while 34% were detected uniquely by the population deep sequencing methodology. This suggests that while there is good detection by both methods, the population deep sequencing identifies a much larger proportion of the population diversity than the single colony approach. Furthermore, as the number of single colony isolates increases the number of uniquely identified SNPs would likely decrease due to the rarer allelic variant positions within the population. However, the deep sequencing pipeline is more likely to capture the population as a whole, providing a more accurate snapshot of genetic diversity at population level.

The corpus population could not be interrogated for the 439C population due to no SNPs being detected by the single colony isolates (figure 5.6B). However, this is still an interesting observation, mainly because the single colony method would have suggested that there was no genetic diversity in this population. But, the deep sequencing of the population proves that this is incorrect. Furthermore, it highlights the limitations of sequencing single colony isolates from a clinical sweep. Successful isolation of single colony isolates, and representative sampling of the population is challenging and can only be judged after sequencing because most colonies look identical on the culture medium. Even then, it is still not possible to know how well the population was sampled. Whereas, a population deep sequencing approach, whilst also having its own limitations, relies less on the separation of single colony isolates and sequencing of the whole population allows for a potentially more accurate snapshot of population diversity. The random sampling of single colony isolates, and the return of no SNPs might also indicate that the pipelines used within this study are consistent and accurate. This result might indicate that there is little to no ‘background’ SNP variation

as a result of data handling (genome assembly and variant calling pipelines) validating the SNPs detected in the other samples as a result.

The better approach would be a culture independent sequencing approach. However, such a method would be difficult to apply to *H. pylori* populations colonising the human gastric mucosa. There would be ethical implications of sequencing directly from human tissue samples. Furthermore, as biopsy samples are usually pulled through the oral cavity, contamination by other bacteria is not uncommon, complicating the subsequent bioinformatics analysis.

The unique single colony SNP data for all single colony isolates were pooled and a combined detection percentage was calculated and depicted in figure 5.7. Samples from 565C were excluded due to the potential multiple *H. pylori* strain contamination of isolates 565C1 and 565C6. Of all uniquely identified SNPs by both methodologies, an average of 68.69% (95% CL: 59.13% – 78.25%) were detected only by the population deep sequencing, 8.13% (95% CL: 4.83% – 11.43%) by the single colony isolate analysis and 23.18% (95% CL: 14.93% – 31.43%) were concordant between the methodologies. Again, this suggests that the deep sequencing allelic calling pipeline worked well with only 8.13% of all identified SNPs not picked up in the population with a good overall concordance.

The deep sequencing methodology is superior in capturing the snapshot of population genetic diversity compared to a single colony approach and has many advantages. The main advantage is in better understanding of the genetic diversity within the population, which has implications in driving future research on key variable genes. Furthermore, in understanding this diversity, a suitable vaccine target might be identified as population variable genes would not be a good target. This analysis might help explain how *H. pylori* is able to colonise the human host as a life long chronic infection.

However, single colony analysis is still relevant and useful to provide insights where a population deep sequencing approach cannot. Examples include population structure and recombination.

Table 5.4 Concordance of single colony variants and variants detected by the deep population sequencing minor allele calling pipeline for patient 439A

Method	Sweep / isolate	Total variants	Single colony unique variants	Concordant variants
Population deep sequencing	**439A**	152	N/A	N/A
Single colony sequencing	**439A1**	89	6	83
Single colony sequencing	**439A4**	87	6	81
Single colony sequencing	**439A5**	30	16	14
Single colony sequencing	**439A6**	78	4	74
Single colony sequencing	**439A7**	0	0	0
Single colony sequencing	**439A8**	83	7	76
Pooled methodologies – unique variants	**All single colony isolates and deep sequenced antrum population**	**181**	**29**	**91**

Table displaying the total number of variants, single colony unique variants and concordant variants identified by the single colony (section 5.2.10) and deep population variant calling pipelines (Chapter Four, section 4.2.1.1). All variant sites detected within the single colony isolates were compared ('vcf-isec') to identify the single colony uniquely identified variants due to potential unique variant call duplication across the strains (as described in section 5.2.10). Further results of all samples used within this study can be found in the Appendix (tables 11.5.7 – 11.5.25).

Figure 5.6 Comparison of detection rates from the minor allele deep population and single colony variant calling pipelines for patient 439

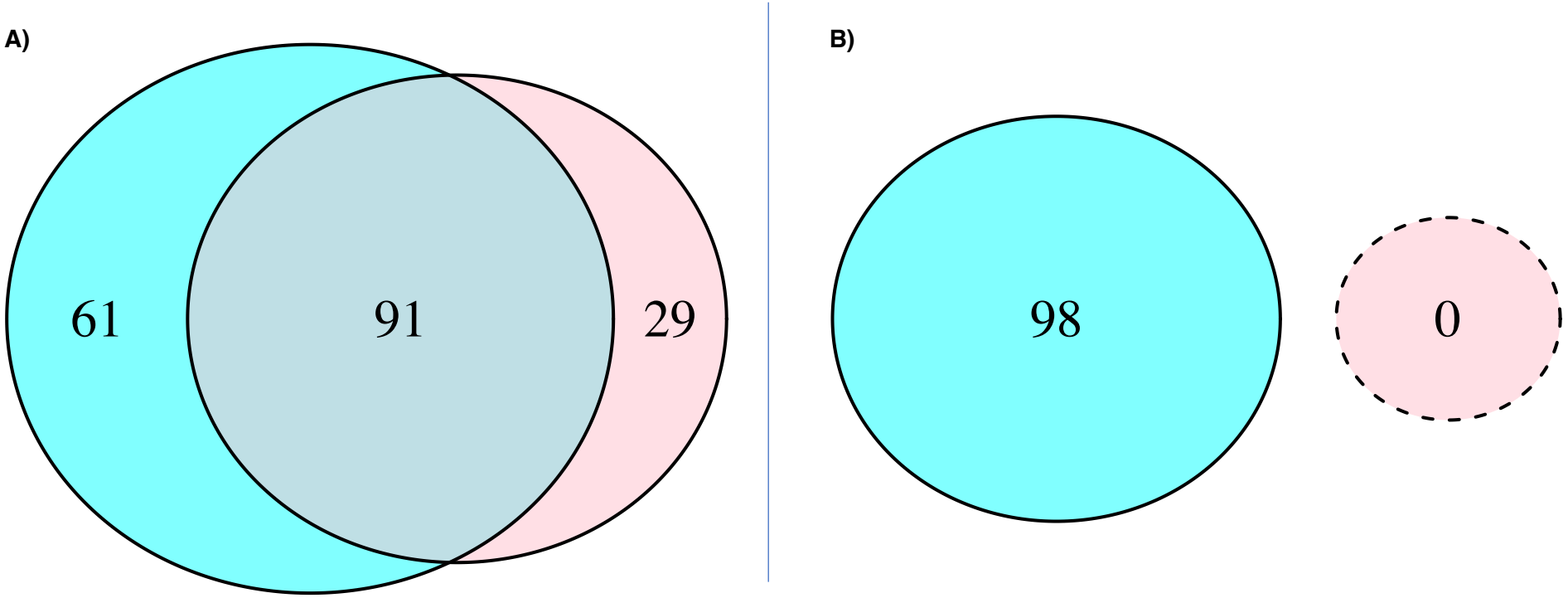


Figure depicting and denoting the proportion of uniquely identified deep population minor allele variants (blue), uniquely identified single colony variants (pink) and concordantly identified variants (pink/blue cross section). Figure A represents the antrum while figure B represents the corpus datasets. Venn diagrams for other patient samples used within this study are presented in the Appendix figures 11.5.17 – 11.5.26.

Figure 5.7 Total SNP detection rate between the minor allele deep population and single colony variant calling pipelines across all samples from all patients

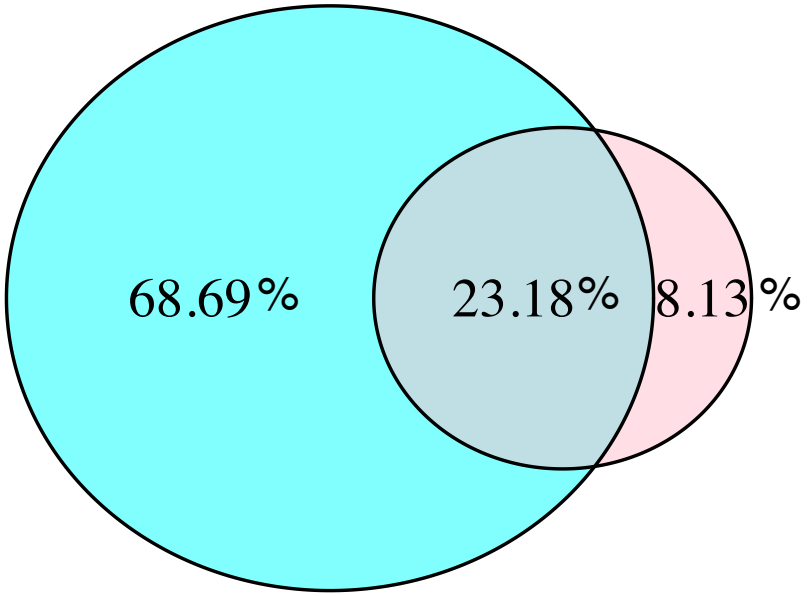


Figure depicting and denoting the proportion of uniquely identified deep population minor allele variants (blue), uniquely identified single colony variants (pink) and concordantly identified variants (pink/blue cross section). Sample 565C was excluded from this analysis (n=19) as both single colony and population variants were exceptionally high, with single colony variants being substantially higher than population variants and are thus outlier results. The average detection rate is depicted. Deep sequencing minor allele only detection (average) = 68.69% (95% CL: 59.13% – 78.25%), Single colony variant detection only (average) = 8.13% (95% CL: 4.83% – 11.43%), shared methodology detection = 23.18% (95% CL: 14.93% – 31.43%) (GraphPad Prism, version 8.2.0). This figure was constructed by the R statistical software version 3.5.1 (R Core Team, 2018).

5.4. Future work

While the deep population sequencing of *H. pylori* clinical sweeps was novel, the combination of deep population sequencing and single colony isolate sequencing was particularly powerful. This combination of approaches allowed a comprehensive investigation of the within patient and between patient genetic diversity and dynamics of *H. pylori* patient colonisation. However, only 5-6 single colony isolates were taken from each population providing insight but perhaps not complete depth of analysis. Therefore, an increased single colony sample size would have benefitted this project, particularly for the cross comparisons of the single colony and deep population methodologies and the population structure analysis between niches.

The GWAS lacked resolution/power due to the low number of samples, further suggesting that a bigger sample size would be highly beneficial to future studies. However, genes with high sensitivity and specificity from strains associated with the antrum or corpus should be investigated further to understand their importance in niche adaptation. For example, gene knockout mutants of *H. pylori* could be compared to wildtype strains to further investigate colonisation patterns across the stomach using mouse models and phenotypic assays.

This study investigated the gene differences between strains taken from individual patients and compared results to those presented in Chapter Four. However, gene expression levels were not investigated. Future studies might include expression of genes from different niches of the stomach. This could provide further information of between niche diversity of *H. pylori* strains and/or populations.

The two presumptive mixed strain contaminated samples 565C1 and 565C6 should be investigated further as these appeared to be extremely different to the other strains.

6. Chapter Six: Sequential datasets

6.1. Introduction

The collection of sequential *H. pylori* samples from the same patient over time is relatively rare due to the clinical practice of treating *H. pylori* infection with an eradication therapy regime in *H. pylori* positive patients. Furthermore, not all patients who present with *H. pylori* infection symptoms undergo gastric endoscopy and subsequent bacterial culture. The primary method of both initial diagnosis of *H. pylori* infection and confirmation of successful eradication are determined through the 13C urea breath test (European *Helicobacter pylori* Study Group, 1997). Therefore, relatively few patients undertake gastric endoscopy and fewer patients still return for gastric endoscopy after eradication therapy because a positive 13C urea breath test would usually be enough to prompt a second course of eradication therapy. Further still, antimicrobial susceptibility testing of *H. pylori* infection is rarely carried out (Dang and Graham, 2017). This hinders the effective tailoring of eradication therapies to clear *H. pylori* infections and limits the surveillance of antimicrobial resistance.

Despite this, there have been a number of studies involving sequential isolates. For example, the natural mutation rate of *H. pylori* has been determined from studies that used sequential isolates inferred by the number of observable mutations between the different sampling time points (Falush et al., 2001b; Morelli et al., 2010; Didelot et al., 2013). These same studies also determined the natural recombination rate between *H. pylori* strains. Other sequential *H. pylori* studies have investigated the size of recombination imports, that have been found to range from 261 – 3,853 bp (Kennemann et al., 2011; Kulick et al., 2008; Falush et al., 2001b).

Other studies have investigated *H. pylori* infection in families and populations where sequential samples were obtained (Nell et al., 2014; Kennemann et al., 2011; Morelli et al., 2010). The study by Morelli *et al.* (2010), investigated the within and between host evolution of *H. pylori* sequentially isolated strains by sequening 78 genes (total – 39,300 bp) from these strains. They calculated recombination rates of famillial isolated strains and sequential strains to determine the effect recombination had on these infections. Kennemann *et al.* (2011), used whole genome sequencing of sequential isolates from four patients to investigate the SNPs between the different time points and to investigate the recombinational imports and their relative lengths. They used this information to highlight specific genes of interest such as an increased recombinational imports and mutations within genes associated with adhesion. Another study

investigated within host evolution of the *H. pylori* Lewis b adhesion gene *babA* through sequentially isolated strains of *H. pylori* (median interval of 1.8 years; one strain sampled at two time points), and found that that in 12 out of 16 patients the *babA* gene was effected by recombination and/or mutation (Nell et al., 2014). These studies showed how sequential sampling can be used to track changes over time and can highlight genes of interest that could be involved in specific processes such as bacterial adhesion.

Some studies have followed *H. pylori* infected individuals over a period of time, taking sequential samples during the course of infection. For example, Kraft *et al.* (2006), investigated sequential isolates taken from an infected individual ranging from three to 48 months. They observed specific strain differences between time points across virulence associated genes, housekeeping genes, insertion elements and restriction-modification systems. These genetic changes associated with a range of genes over a relatively short period of time show how the *H. pylori* infection within a single host can change and develop during chronic infection. While this was an important study, it was based around whole-genome DNA microarrays and hybridisations modelled from the *H. pylori* reference genomes 26695 (Alm et al., 1999) and J99 (Tomb et al., 1997). Therefore, it is likely that only genes shared with strain 26695 were analysed, potentially missing out accessory genes in the strain isolated from the patient. Furthermore, DNA microarrays do not capture the whole target genome, only the coding genes. Therefore, intergenic regions and promoter regions are ignored in this type of analysis.

Krishna *et al.* (2016), investigated an archival J99 strain taken from the patient that provided the first sample six years prior to refusing eradication therapy (Alm et al., 1999). While these strains were not compared by whole genome sequencing, phenotypic acid tolerance differences were observed between strains taken from different time points. Deletion and reconstituted mutants confirmed that mutations in the *arsS* gene were responsible for the increased acid survival and this was further confirmed by the sequencing of these genes. The authors speculated that such genetic changes during chronic infection may result in niche migration from the less acidic antrum to the oxyntic corpus and might increase the risk of gastric cancer.

A longer term study of chronic infection by *H. pylori* was conducted by Liu *et al.* (2015), in a primate (rhesus macaque) model over ten years. Gastric biopsies were taken at 3-month intervals and cultured to collect ten single colony isolates from both the antrum

and corpus niches at each time point from 12 rhesus macaques. Whole genome sequencing was not carried out on any single colony isolates at any of the time points. Instead, DNA fingerprinting and comparative genomic microarray-based hybridisation was used to investigate the genetic differences between strains at the varying time points. The authors found that *H. pylori* single strain infection remained identical after one year of infection. However, after five years the single colony isolates had started to change genetically. This allowed the authors to identify the *babA* gene and conclude that it was under strong selection pressure potentially relating to colonisation of the host.

A second study using a primate (rhesus macaque) model took an opposite approach to that of Liu *et al.* (2015), and investigated the short term, acute infection phase of *H. pylori* infection (Linz et al., 2014). Linz *et al.* (2014), extracted antral gastric biopsies from seven days and one, two and three months after innoculating a single primate with *H. pylori* and used whole genome sequencing on the cultured isolates to compare them to the original infecting strain. The authours complimented the rhesus macaque model with two human volunteers who re-infected themselves with their previously infecting *H. pylori* cultures (cured through eradication therapy) and had gastric biopsies of the antrum taken at 20 days and 44 days after infection confirmation (urea breath test). The authors found that there was an increased mutation rate during acute infection that was ten times higer than during chronic infection. They termed this a 'mutational burst' during acute phase infection. This result was in direct conflict with the previous study by Liu *et al.* (2015), who did not observe genetic changes until after one year of infection. However, as previously mentioned, this result might have been limited by the analysis methodology employed (DNA fingerprinting and comparative genomic microarray-based hybridisation), whereas Linz *et al.* (2014), used whole genome sequencing that is more sensitive to SNP variation.

Taking these studies together, it is clear that sequential samples are rarely obtained from naturally infected patients with *H. pylori*. It is also clear that even fewer studies have utilised the information available within the datasets by undertaking whole genome sequencing and using whole genome comparative genetics techniques. Despite this, they have been used to generate some interesting and important findings in regard to how *H. pylori* infection develops during acute and chronic phases of infection.

To the best of this author's knowledge, sequentially isolated samples have not yet been tested for their sensitivity to antibiotics at different time points. Furthermore, characterisation of sequential *H. pylori* samples taken before and after failed eradication therapy within naturally infected humans has yet to be published within the literature. Therefore, this study provides a novel insight into how *H. pylori* infections might change once challenged by eradication therapy that subsequently fails to eradicate the infection. Both population deep sequencing and single colony isolate sequencing techniques were used in this study to comprehensively study two patients where sequential samples were taken both before and after failed eradication therapy.

6.2. Materials and methods

Sample selection, DNA extraction, whole genome deep sequencing, sequencing read curation, contamination detection, whole genome assembly, assembly curation and genome annotation were conducted as described in Chapter Two.

Antibiograms were carried out as described in Chapter Three section 3.4.2.

Single colony isolation and subsequent whole genome sequencing were carried out as described in Chapter Five section 5.2.1.

6.2.1. Samples used within this study

For two patients, sequential samples were taken before and after failed eradication therapy (table 6.1). However, due to patient confidentiality surrounding medical records, it was not possible to obtain information on the antibiotics prescribed in each case.

A *H. pylori* population sweep was taken from the corpus biopsy of patient 249 with a second population sweep taken from the antrum approximately 44.9 months (1,366 days) after the initial sampling and designated the patient ID 537. These samples are referred to as sequential set 1 (table 6.1). It is not known why paired antrum and corpus biopsies were not available. It is possible that the *H. pylori* sweep was not viable from the missing biopsy cultures or was simply a decision taken by the health practitioner at time of endoscopy to not take paired biopsies. Whilst Chapters Four and Five show there are antrum and corpus population differences, this data was still considered

important due to the rarity of sequential human samples and are therefore described here.

A second patient underwent *H. pylori* population culture from biopsies taken from both the antrum and corpus referred to as patient 295. This same patient had a second antrum and corpus biopsy taken approximately 4.9 months later with the patient identifier 326. This sample set is referred to as sequential set 2 (table 6.1).

The population sweeps were deep population sequenced as described in Chapter Two sections 2.6 and 2.8. Single colony isolates were also collected from each population (Chapter Five, section 5.2.1) and six of these were sequenced per population, as described in Chapter Two section 2.7 – 2.8.

Table 6.1 Patients and samples used within this study

Deep population sequenced samples	List of single colony isolate numbers used within this study	Sequential set	Notes
249C	249C1, 249C3, 249C6, 249C8, 249C10, 249C16	1	Initial biopsy taken at time point 0 from patient 249. Only the corpus sample was available.
537A	537A1, 537A3, 537A4, 537A5, 537A7, 537A8	1	Second/sequential biopsy taken 1,366 days after the initial biopsy. Only an antral sample was available.
295A	295A1, 295A2, 295A3, 295A4, 295A5, 295A6	2	Initial biopsy taken from the antrum at time point 0.
295C	295C1, 295C2, 295C4, 295C6, 295C7, 295C8	2	Initial biopsy taken from the corpus at time point 0.
326A	326A22, 326A23, 326A24, 326A25, 326A26, 326A27	2	Second/sequential biopsy taken 148 days from the antrum after the initial antral biopsy. The population sweep, 326A, was contaminated as shown in Chapter Four table 4.1. However, none of the single colony isolates were contaminated (Chapter Five section 5.3.1).
326C	326C1, 326C2, 326C3, 326C4, 326C5, 326C6	2	Second/sequential biopsy taken 148 days from the corpus after the initial corpus biopsy.

Samples used within this study including population sweeps and single colony isolates. Further patient, sampling and *H. pylori* samples details are denoted.

6.2.2. Patient reference genome for sequential samples

The reference genomes are described in each figure and analysis due to the use of multiple different references in this study. However, a consistent approach was taken to that described in Chapters Two, Four and Five, for each different type of analysis.

6.2.3. Antibiograms

Antibiotic resistance assays were carried out on the population sweeps for patient numbers 249, 295, 326 and 537 as described in Chapter Three section 3.4.2. Briefly, frozen cultures were thawed from -80 °C and plated onto a blood agar base #2 culture plate supplemented with 7.5% defibrinated horse blood which was incubated (37°C) under microaerophilic conditions for 48 – 72 hours. After incubation, a rapid urease test (Chapter Two, section 2.3) was used to confirm presence of *H. pylori* and bacterial edge growth was used to inoculate three additional culture plates and incubated for a further 16 – 24 hours under the same conditions (Chapter Two, section 2.2). Bacterial edge growth was taken from multiple culture plates (if required) and made up to a McFarland 2.8 ±0.1 in 0.85% saline using a densitometer (Chapter Three, section 3.4.2). Six MHA plates supplemented with 7.5% defibrinated horse blood were inoculated with the McFarland 2.8 cultures. Antibiotic disks containing 15 μg clarithromycin, 10 μg amoxicillin, 5 μg rifampicin, 1 μg levofloxacin, 30 μg tetracycline or 5 μg metronidazole were placed in the centre of the culture plates, one antibiotic disk per plate. These plates were incubated for 96 – 120 hours and zones of inhibition recorded.

Each antibiotic resistance assay was repeated in triplicate.

6.2.4. Deep population sequencing analysis of sequential samples

The antrum and the corpus populations were deep sequenced at time points 0 and 1 respectively for sequential set 1 (patient IDs 249 and 537). Both the antrum and corpus populations were sequenced at time points 0 and 1 for sequential set 2 (patient IDs 295 and 326) (table 6.1).

The minor allele detection pipeline was used on all datasets as described in Chapter Four section 4.2.1.1. Briefly, the sequencing reads from each population were assembled into a sample/niche consensus genome and the population reads were

mapped back to identify sites of allelic variation within each population. Duplicate reads were removed as they were likely to be PCR duplications. A read mapping quality score of Phred 34 and a base quality score of Phred 30 was used. To call minor allelic variant positions, variants with less than 3% of reads supporting a suspected allelic position were removed.

This data was then converted into a heatmap of variable/allelic genes.

6.2.4.1. Minor allele frequency determination

The minor allelic calling pipeline does not output minor allele frequency (MAF) statistics primarily due to the exclusion of this in FreeBayes (Garrison and Marth, 2012).

Minor allele frequencies were manually calculated by extracting minor allelic calls within the VCF output files output from the minor allele calling pipeline (Chapter Four, section 4.2.1.1). Next, the minor allelic positions within 'hypothetical protein' and 'outer membrane protein' were removed as these were excessively numerous and already described as genetically diverse within Chapters Four and Five. It was decided that removing these calls would potentially highlight other genes of interest.

The info field 'AD' (Number of observation for each allele – reference and alternative) within the VCF files were imported into Microsoft Excel (version 16.16.6) and added together for each separate minor allelic position to determine the total number of reads mapping to each position. Finally, the number of reads supporting the minor allele call were divided by the total number of reads mapping to that position, providing the MAF.

The VCF files were annotated as described in Chapter Two section 2.14.

6.2.5. Single colony isolate genetic variation before and after failed eradication therapy

To investigate the genetic variation of single colony isolates, isolate reads were mapped to the patient reference genome at time point 0. The following command was used:

```
snippy --cpus {user specified CPU/thread/core number} --report --minfrac 0.9 --mincov 6 --mapqual 34 --basequal 30 --ref {patient
```

```
reference at time point 0 consensus genome. GenBank annotation
file} --outdir {user specified output directory} --R1 {path to
forward/mate pair 1 of a single colony isolate} --R2 {path to
reverse/mate pair 2 of a single colony isolate}
```

Snippy (Seemann, https://github.com/tseemann/snippy) was enacted whereby 90% of reads had to support an alternative base call to be classed as a variant position between the isolate and the reference genome. A Phred 30 cut off score was used as this increases the base calling accuracy to the equivalent of 99.9%. Similarly, a Phred 34 mapping score was used to increase mapping accuracy equivalent to 99.96%. For a base to be considered polymorphic, it had to have at least 6 reads aligning to this position.

Output VCF files were processed as described in Chapter Two section 2.14 to annotate the genes in which the SNPs were found.

6.2.6. Whole genome alignments of the sequential samples

Taking a similar approach to that described in Chapter Four (section 4.2.2.3) and Chapter Five (section 5.2.6) all sequential single colony isolates were comparatively analysed using BLASTN to the patient consensus genome (Chapter Five, section 5.2.3) at time point 0.

Briefly, the assembled genomes were filtered to remove contigs of <500 bp in length. The single colony isolates isolated before and after failed eradication therapy for sequential set 1 were compared by BLASTN (upper 99% lower 96% nucleotide identity) to the population deep sequenced consensus genome assembly of patient 249C. A patient consensus genome was not created for sequential set 1 as there was no population data for the antral niche. The single colony isolates from sequential set 2 were compared against the patient (295) consensus genome where all population antrum and corpus reads were pooled and assembled by SPAdes (Chapter Two, section 2.11).

This method does not allow for the direct comparison of site-specific allelic variants or alignment SNPs to the comparative niche after failed eradication therapy. However, it does allow for a visualisation of all samples together facilitated by the patient reference

genome at time point 0. As previously mentioned (Chapter Five, section 5.2.6), the exact SNP locations could not be displayed because a different reference was used for the population minor allele detection and single colony read mapping methodologies. Therefore, to better depict the cross comparisons of the single colony variation and the minor allele population variants, genes that were detected as allelic within the population but not variable by the single colony read mapping to the respective reference genomes were highlighted as teal coloured gene labels. Genes observed to be variable by the single colony read mapping to the respective reference and not by the population minor allele detection pipeline were coloured as black gene labels. Finally, where both single colony read mapping and minor allele population gene variants matched by gene name, they were highlighted as red gene labels in the BRIG plot.

The single colony read mapping method is described in section 2.5.

6.2.7. Phylogeny

The single colony isolates from sequential set 1 and 2 were used to create a phylogenetic tree to infer genetic relationships. A read mapping approach was used where the sequencing reads from the single colony isolates were mapped against the patient reference genomes at time point 0 for each sequential dataset. This method was identical to that described in Chapter Five section 5.2.4.

Briefly, single colony isolate reads were mapped to the reference via Snippy (Seemann, https://github.com/tseemann/snippy) whereby 90% of reads had to align to an alternative base call to be called as variant. A minimum base coverage of 6 was also required, providing a higher alternative calling accuracy. Phred 30 and 36 scores were required for base quality and mapping quality, respectively.

Next, the regions of recombination were removed by Gubbins (Croucher et al., 2015) and sites of variation were output in relation to the reference sequence. This SNP alignment was run through FastTree to create a phylogenetic tree by approximately-maximum-likelihood (Price, Dehal and Arkin, 2010).

Phylogenies were drawn and manually labelled using the interactive tree of life (Letunic and Bork, 2007).

The alignment files were manually inspected to calculate the number of SNPs between strains in order to provide a SNP estimated scale bar.

6.2.8. Recombination

Regions of homologous recombination were investigated between all strains within sequential set 1 and 2 both before and after eradication therapy. This method is as previously described in Chapter Five section 5.2.7.

Briefly, all single colony isolate genomes from sequential set 1 and 2 were separately aligned to the time point 0 patient reference genome using Mugsy (Angiuoli and Salzberg, 2011) and alignments were filtered to remove alignment gaps that were observed in > 50% of the genomes. Two different recombination detection algorithms/programs were run, Gubbins (Croucher et al., 2015) and fastGEAR (Mostowy et al., 2017), to identify regions of recombination.

As previously described, Gubbins identifies substitution clusters within similar datasets that are flagged as outliers inferred through phylogeny whereas fastGEAR locates similar genetic segments from diverse clusters and uses these to infer recombination. FastGEAR does not depend on a phylogenetic reconstruction while Gubbins does. Analysis by two different methodologies allows for a better insight into regions of homologous recombination.

6.3. Results and discussion

6.3.1. Antibiograms before and after failed eradication therapy

Eradication therapy consists of two antibiotics, usually a combination of clarithromycin, amoxicillin and metronidazole coupled with a proton pump inhibitor (Cameron et al., 2004). It is common practice for a *H. pylori* infection to be prescribed eradication therapy without knowing the sensitivity profile of the infection. There are many reasons for this. For example, not all cases of *H. pylori* infection are identified through gastric endoscopy and subsequent histological and culture analysis. In the UK, the carbon-13 urea breath test is primarily used for *H. pylori* infection diagnosis (Shirin et al., 2001; Israeli et al.,

2003) and *H. pylori* are not routinely cultured for identification and sensitivity testing. The fastidious nature of *H. pylori* and the high risk of contamination due to the endoscopy passing through the oral cavity also make routine culturing difficult. Furthermore, there are currently no nationally recognised antibiotic typing profiles for sensitivity and resistance for *H. pylori* by disc diffusion. For example, the EUCAST only provide clinical MIC breakpoints and no guidance on breakpoints for disk diffusion based methodologies (EUCAST, 2019).

6.3.1.1. Sequential set 1

The antibiograms before and after failed eradication therapy were similar for sequential set 1 (figure 6.1). This is perhaps a surprising result considering treatment with antibiotics should, in theory, clear the infection or potentially have some notable impact on the infection. However, it is clear from the antibiograms that this patient was infected with a clarithromycin resistant *H. pylori* strain before eradication therapy was attempted. Clarithromycin resistance in *H. pylori* was recorded at around 5% from a study investigating resistance rates between the years 1991 and 2001 in the UK (Cameron et al., 2004). Although the resistance rates are potentially higher today, attributed to the increasing global trend of antimicrobial resistance, resistance to clarithromycin in the UK is thought to be low compared to other countries (Megraud et al., 2013). Despite the low level of clarithromycin resistance in the UK, a more recent European study into resistance found levels of resistance equivalent to 17.5% across European countries and 8% in the UK (Megraud et al., 2013). These results were attained through susceptibility testing via Etest strips. Due to the alarming increase of resistance to clarithromycin, a first line antibiotic in the triple therapy, the World Health Organisation (WHO) has recently placed *H. pylori* in the top ten pathogens of global concern where research and development are needed for new antibiotics (Savoldi et al., 2018).

Furthermore, the sequential set 1 time point 0 population was also resistant to metronidazole and rifampicin, making this a multi-drug resistant *H. pylori* infection (Chapter Three, table 3.5). Considering first line metronidazole and clarithromycin resistance, it is perhaps not surprising that this patient presented little change in resistances to the six antibiotics tested in this study. It is not known what combination of drugs were prescribed to this patient due to patient confidentiality restrictions. However, as the populations became slightly more sensitive to amoxicillin post antimicrobial therapy (figure 6.1) it might indicate that amoxicillin was not prescribed as

higher resistance to this antibiotic might be expected if resistant populations were present at time point 0. Due to suspected sensitivity to amoxicillin (Chapter Three, table 3.5) at time point 0, it would not be unreasonable to suggest that had amoxicillin been prescribed, it might have resulted in population clearance or reduction of *H. pylori* burden.

Resistance to rifampicin is another surprising observation for sequential set 1. This is because a 2009 study reported resistances as less than 1% in the UK and highlighted rifampicin as a potential antibiotic for use in first line drug resistant infections (Chisholm and Owen, 2009). Therefore, based on the recommendation of that study, treatment with rifampicin for this patient would likely be futile. Therefore, there is the desideratum to standardise the methodology for *H. pylori* antimicrobial resistance reporting.

Helicobacter pylori resistance to metronidazole in the UK was observed to be around 31.7% as identified in the aforementioned study by Cameron *et al.* (2004), suggesting that resistance is much higher for this antibiotic than for clarithromycin and rifampicin resistance. However, there is little reference to multi-drug resistant *H. pylori* infections in UK studies and within the EU generally, suggesting that the prevalence of such infections is not fully understood. (Megraud et al., 2013). This is supported by the observation of a three-drug resistant population in sequential set 1.

It must also be noted that only two notable studies on the antimicrobial resistance profiles of UK infections with *H. pylori* have been published (Cameron et al., 2004; Megraud et al., 2013). Megraud *et al.* (2013) commented on a surprisingly low number of clinical *H. pylori* representatives sampled from the UK in their study (including Germany, Italy and Poland). Therefore, there is perhaps the need for a current review into the antimicrobial resistance profiles of *H. pylori* infections in the UK today.

It is clear then, that the sequential set 1 patient holds a potentially difficult to treat (with first line triple therapy) multi-drug resistant *H. pylori* infection. This observation might have been missed without the antibiograms conducted in this study.

6.3.1.2. Sequential set 2

Contrary to sequential set 1, sequential set 2 exhibited changes in antibiotic resistance profiles after failed eradication therapy in both the antrum and corpus populations. This

striking difference between sequential set 1 and 2 is likely due to the differences in sensitivities of the *H. pylori* infections at time of eradication therapy. While sequential set 1 was resistant to two of the first line antibiotics before eradication therapy began, sequential set 2 was only resistant to one, metronidazole (figure 6.2A-B). Therefore, any combination of the three, first line antibiotics would most likely have an effect on the *H. pylori* infection. This provides a unique opportunity to investigate what might happen during and after a failed eradication regime.

It must also be noted that the patient of sequential set 2 held a multi-drug resistant *H. pylori* population due to metronidazole and rifampicin resistance (figure 6.2A-B; Chapter Three, table 3.5). Again, while metronidazole resistance is relatively high in the UK (~ 31.7%), rifampicin resistance was surprising considering previous studies finding that fewer than 1% of strains were rifampicin resistant (Cameron et al., 2004). Despite the multi-drug resistance of sequential set 2, this study is assuming that at least the first line drug therapy was attempted.

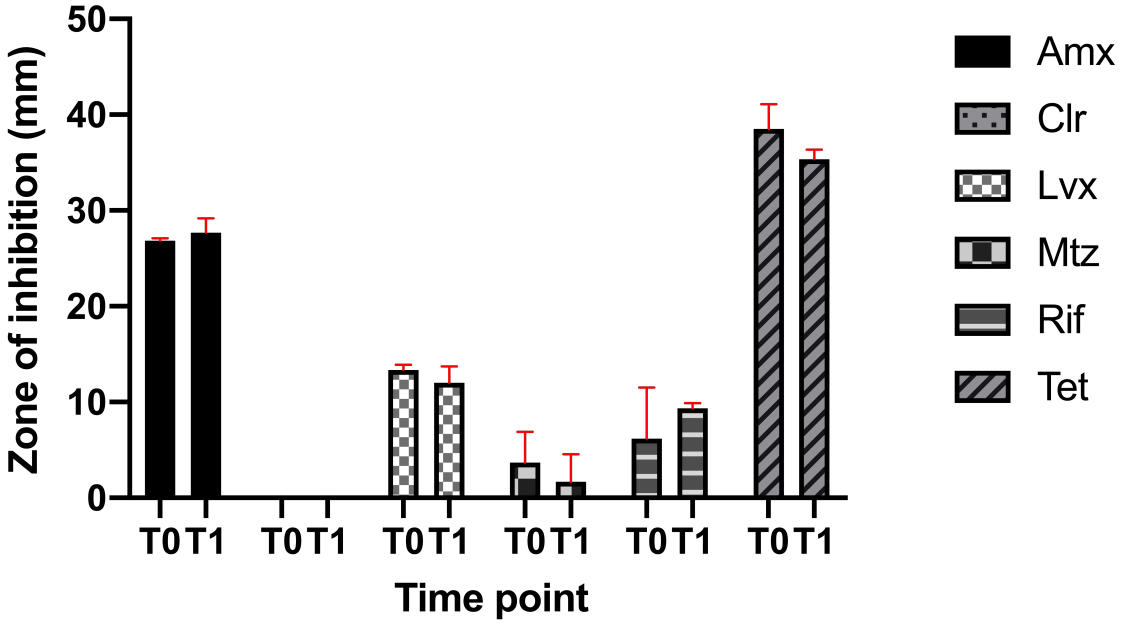
It should be further noted that the antrum population sweep at time point 0 for sequential set 2 (sample 326A) was most likely contaminated with an unknown contaminant (Chapter Four, table 4.1). Although the bacterial growth on the agar plates appeared uniform and consistent with the appearance of *H. pylori*, the antibiotic resistance assays for this population may be incorrect because the contaminant population that was detected by deep population sequencing could be contributing to the resistance profiles observed. However, the corpus population was not contaminated (Chapter Four, table 4.1) thus is a reliable data set and is the focus of the analysis presented here.

Excluding the antrum population (326A) at time point 0 for sequential set 2, it is clear that there are antibiotic resistance differences before and after failed eradication therapy. Of particular note is rifampicin that was significantly higher at time point 0 than time point 1 (p-value 0.015) (figure 6.2B). There were also slight increases in resistance of the *H. pylori* population to metronidazole and tetracycline following failed eradication therapy. Sensitivity to amoxicillin and clarithromycin was slightly higher while levofloxacin zones of inhibition were identical.

The antrum and corpus sensitivity profiles were also different post eradication failure, suggesting that the antibiotics had varying effects dependant on the niche. This could be down to differences in the delivery and/or effectiveness of the antibiotics between

these niches. Alternatively, *H. pylori* population/strain differences could be driving these differences in sensitivity to antibiotics. This observation highlights the need for standardisation of biopsy sampling location for antimicrobial resistance surveillance.

Figure 6.1 Antibiogram of sequential set 1 population sweeps



Time point 0 represents population sweep 249C while time point 1 refers to the population sweep 537A. Antibiograms were carried out as described in Chapter Three section 3.4.2. Triplicate results were averaged, and the standard deviation is plot with red error bars. This figure was plot using GraphPad Prism (version 8.2.0). Amx = amoxicillin, clr = clarithromycin, lvx = levofloxacin, mtz = metronidazole, rif = rifampicin, tet = tetracycline. A paired t-test was performed on each antibiotic triplicate data at each time point to investigate statistically different resistance between the timepoints by size of the inhibition zone. There were no statistically significant differences in antibiotic susceptibility between the pre- and post-treatment populations.

Figure 6.2 Antibiogram of sequential set 2 antrum and corpus population sweeps

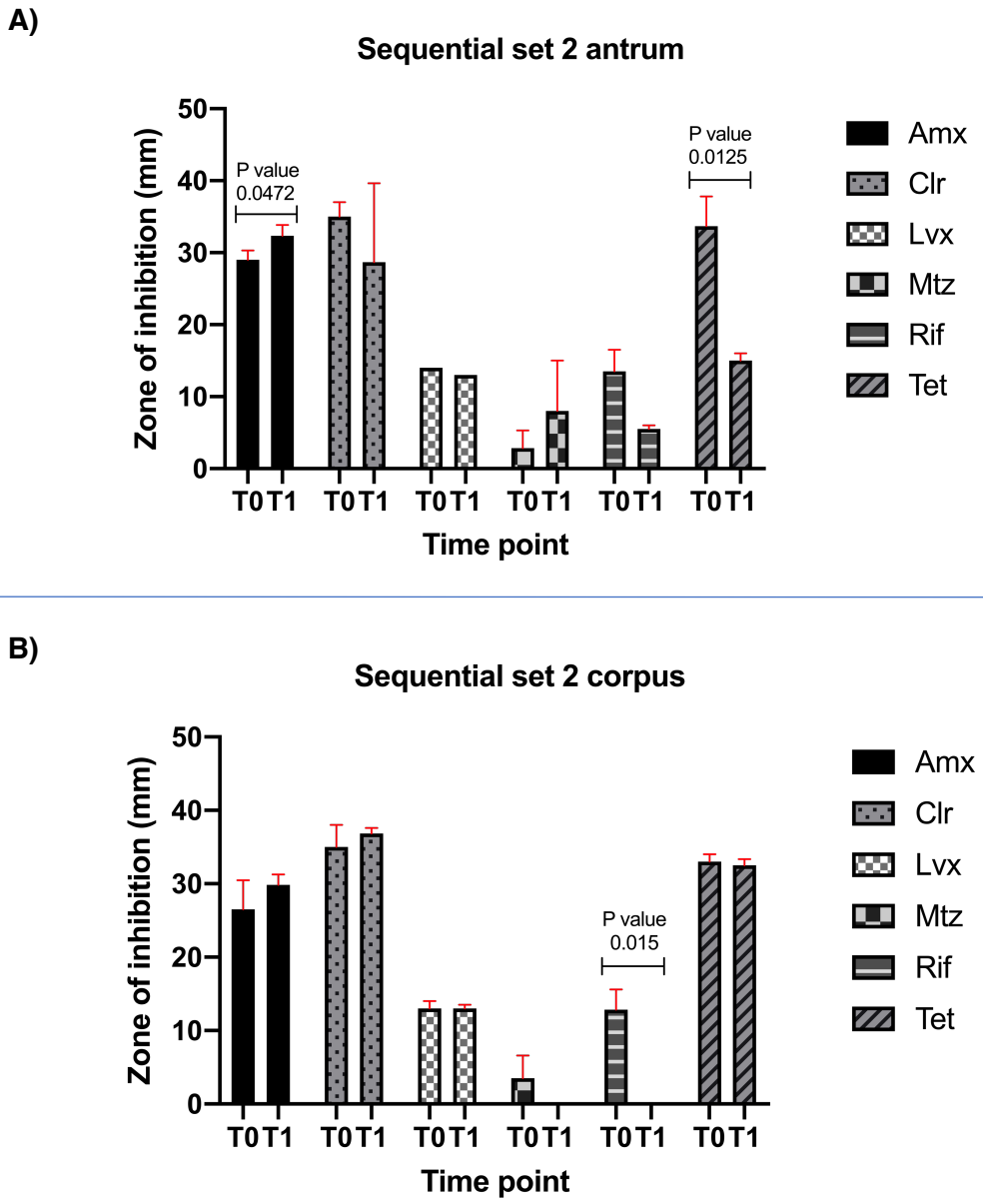


Figure A depicts the sequential set 2 antrum population antibiograms at sampling timepoints 0 (295A) and 1 (326A). Figure B depicts the sequential set 2 corpus population antibiograms at timepoints 0 (295C) and 1 (326C). Antibiograms were carried out as described in Chapter Three section 3.4.2. Triplicate antibiogram results were averaged where error bars represent the standard deviation. Amx = amoxicillin, clr = clarithromycin, lvx = levofloxacin, mtz = metronidazole, rif = rifampicin, tet = tetracycline. A paired t-test (unpaired with Welch's correction) was performed on each antibiotic triplicate data for each time point to investigate statistically different resistance between the timepoints by size of the inhibition zone. Antibiotics with statistically significant (P value < 0.05) differences between zones of inhibition at different time points were denoted above the data points. This figure was plot using GraphPad Prism (version 8.2.0).

6.3.2. Deep population minor allelic variation of sequential samples before and after failed eradication therapy

6.3.2.1. Sequential set 1

There were five more allelic genes after failed eradication therapy for sequential set 1 (figure 6.3). This was a surprisingly low increase in population genetic diversity, especially considering the sampling time gap of approximately 1,366 days. With the high mutation rate and recombination rate, a much higher level of population diversity might be expected after more than three years of chronic infection. Such low genetic diversity might suggest that the *H. pylori* population within this patient's stomach was very clonal and not of a mixed infection. This would limit opportunities for homologous recombination between genetically diverse strains. Furthermore, this infection appears to be successful with little change in genetic diversity. This is an interesting observation as high genetic diversity is thought to aid in the persistence of infection (Liu et al., 2015).

The most variable gene(s) in terms of the number of uniquely called allelic variants harboured within them, both before and after failed eradication therapy in sequential set 1, were the outer membrane protein genes (figure 6.3). Outer membrane protein variation is thought to aid in chronic infection due to differences in adhesion and recognition by the host immune system, preventing immune clearance (Liu et al., 2015; Huang et al., 2016b). The low overall genetic diversity between the time points combined with the high diversity observed within these genes, suggests an important role for outer membrane proteins in the persistence of chronic infection.

The most variable gene(s) in terms of called allelic variants with supporting alternative reads mapping to the alternative base call, encoded for outer membrane genes. These were shown to have a MAF > 0.125 pre-eradication therapy (table 6.2). These relatively high frequencies could suggest that these variant positions are starting to become elevated within the population, potentially moving towards fixation. Alternatively, these higher frequencies could be representative of a cluster of strains within the population.

In terms of the number of allelic variants with a MAF > 0.125 there were more observed post eradication (table 6.3). As observed in sequential set 1 (table 6.2), only three genes containing allelic positions > MAF 0.125 remained from before and after failed eradication therapy (HP1409, HP1177 and HP0725). This switch of genes with high

MAF variant positions potentially reflects genes with allelic variants being selected for at the different time points of infection. A study by Lieberman *et al.* (2014), found that when a strong selection pressure is acting on a gene, multiple gene mutations can occur but no specific adaptive mutation fixes within the population. This, they say, provides a genetic record of past selection events.

Diversity of OMP genes could be how this infection has been able to persist between the two sample points, perhapse preventing immumo recognition and clearance.

Despite there being little variation in the total number of allelic genes (figure 6.3) post eradication therapy in sequential set 1 compared to pre-therapy, there were changes in the number of minor alleles with frequencies > 0.125 within the populations (table 6.2). Post eradication therapy, ten genes were observed with very high (> 0.25 MAF) MAFs compared to only four genes with MAFs >0.25 pre-eradication therapy. This change might reflect minor allelic positions within genes that are moving towards fixation due to positive selection pressures. It is probably an over-simplification to suggest this could be due to antimicrobial treatment, because these samples were taken over three years apart.

6.3.2.2. Sequential set 2

In sequential set 2, there were comparable levels of genetic diversity between the antrum and corpus at time point 0. At time point 1 (after failed eradication therapy) there was a dramatic increase in overall diversity and there were more allelic genes within the antrum population compared to the corpus. The contaminated antrum population (Chapter Four, table 4.1), was not thought to be interfering with the minor allele detection pipeline results presented in figure 6.4 for four reasons. Firstly, the GC content of the contaminated assembled contigs were much higher to those that were aligned to the uncontaminated corpus dataset (Appendix, figures 11.4.45A-B). Therefore, the contaminated reads were likely to be distinct and less likely to map to the *H. pylori* genomic sequence. Secondly, any allelic positions that were identified within the non- *H. pylori* genome were manually filtered out. Thirdly, the high mapping quality score likely prevented contaminated reads from mapping to the *H. pylori* genome sequences. Finally, most of the genes that were found to be allelic were annotated as 'HP' reference genes.

The dramatic increase in the number of allelic genes after failed eradication therapy was surprising because we expected that antibiotic treatment would cause an evolutionary bottleneck. Such bottlenecks are thought to enhance the effects of genetic drift whereby strain extinction, fixation and sweeping events are increased (Didelot et al., 2016).

The increase in the number of genes harbouring allelic variation could be due to a number of factors. One theory could align with a study carried out by Linz *et al.* (2014), who observed a mutational burst during acute phase infection of humans and rhesus macaques after just 44 days of infection. Perhaps what happened after eradication therapy in this patient was similar to an acute phase infection, with antibiotic treatment clearing the chronic phase *H. pylori* infection, subsequently allowing a more acute phase type of infection with a mutational burst. This theory is supported by the MAFs of the genes observed within the populations after eradication therapy (tables 6.2 and 6.3). Out of 735 of the allelic positions between samples 326A and 326C at time point 1 (sequential set 2) there were 282 allelic variants with a MAF < 0.04 (i.e. < 4%). This MAF is very close to the minimum alternative fraction filter used to call minor allele variants, of ≥ 3% (Chapter Four, section 4.2.1.1). These low frequency allelic variants might be detected due to an increased mutation rate of the *H. pylori* strains within the populations inflating the observed genetic diversity within the populations.

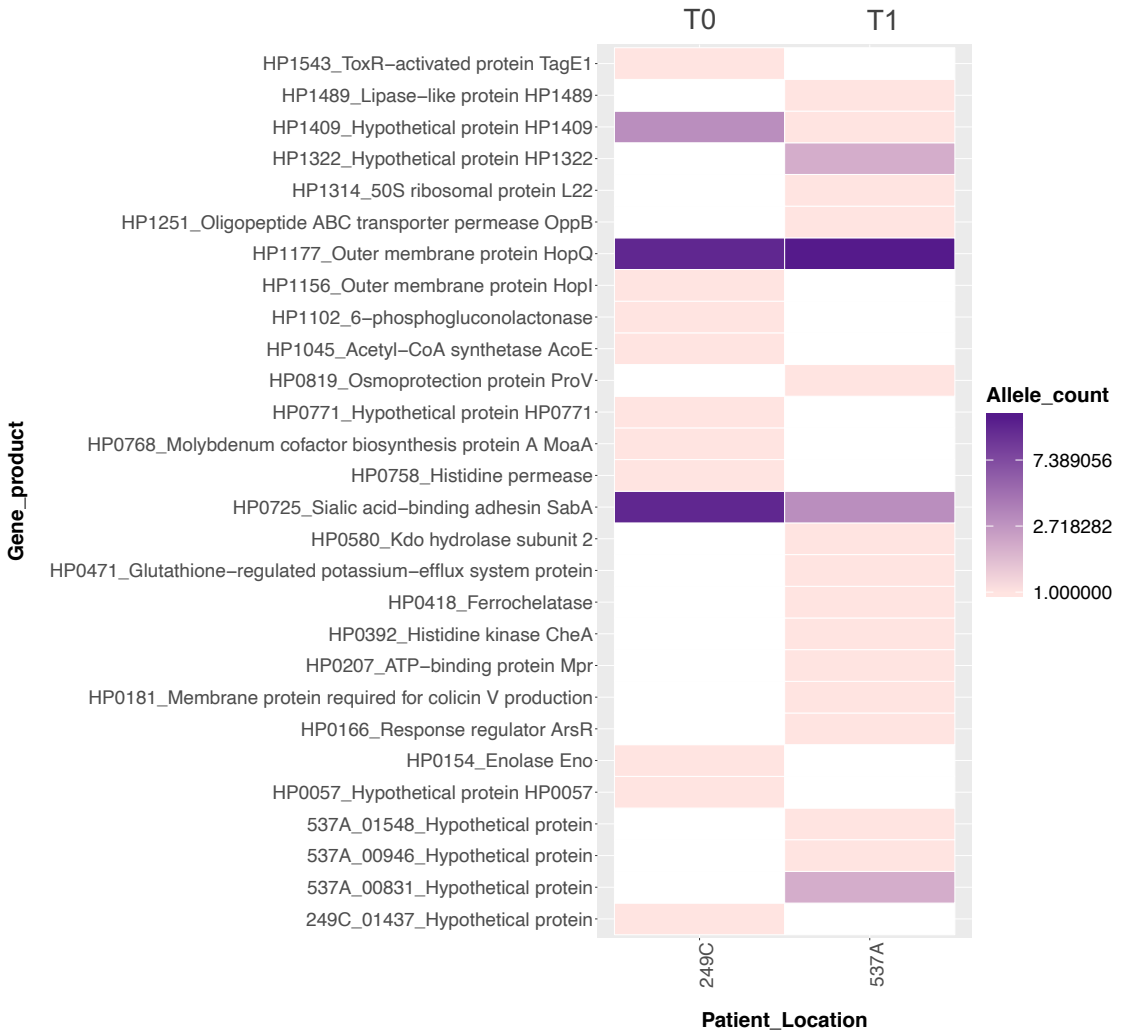
An alternative possible explanation for increased genetic diversity after failed eradication therapy relies on the concept of prior within population diversity and heteroresistance. Population diversity as a result of a mixed *H. pylori* strain infection or a population that has split into sub-populations/quasispecies due to mutation and homologous recombination during chronic infection, could present strains within the population that harbour higher antimicrobial resistance phenotypes. When challenged with the eradication therapy, it is possible that the previously dominant sensitive population is cleared, leaving behind heteroresistant *H. pylori* strains that are very closely related to the original population, but are ultimately more resistant to the antimicrobial challenge. If multiple sub-strains/quasispecies had acquired higher natural resistance, then it is possible that the observed increase in genetic diversity following failed eradication therapy is a result of outgrowth of these diverse resistant strains. How these resistant strains would develop back into a chronic infection is unknown. However, it could be speculated that over time, a genetically fitter strain is likely to then become the most dominant strain within the population, following genetic

drift and positive selection. It would also be possible for the cycle to repeat again, if challenged by another ineffective combination of prescribed antibiotics. Failed eradication therapy as a consequence could drive an increase in antimicrobial resistance in *H. pylori* infections. The caveat to this theory is that a sufficient level of population diversity would need to be present to increase the likelihood of multiple sub-strains harbouring naturally mutated resistance phenotypes. Such genetic diversity, resulting in an increased risk of eradication therapy failure, might therefore be limited to patients who have held a chronic *H. pylori* infection for many years, or patients with a potentially mixed *H. pylori* strain infection.

The last theory described here to potentially explain the increase of genetic diversity post eradication therapy failure refers to the triple therapy itself. First-line standard triple therapy consists of two antibiotics coupled with a PPI (Urgesi, Cianci and Riccioni, 2012). Proton pump inhibitors increase the gastric pH by reducing acid secretion within the stomach (Sachs, Shin and Howden, 2006). Furthermore, PPIs increase the efficacy of antibiotics against the *H. pylori* infection (Peterson, 1997; Yang, Lu and Lin, 2014). However, what is not well understood is the effect of PPIs on the *H. pylori* populations where eradication therapy fails. It is possible that by increasing the gastric pH, PPIs could reduce the environmental pressures acting on the *H. pylori* populations. The relief of acid stress on the bacterial population could drive further genetic variation within the populations due to less fit bacteria being able to survive the more neutral pH conditions. This could potentially drive the expansion in genetic diversity observed after eradication therapy failure.

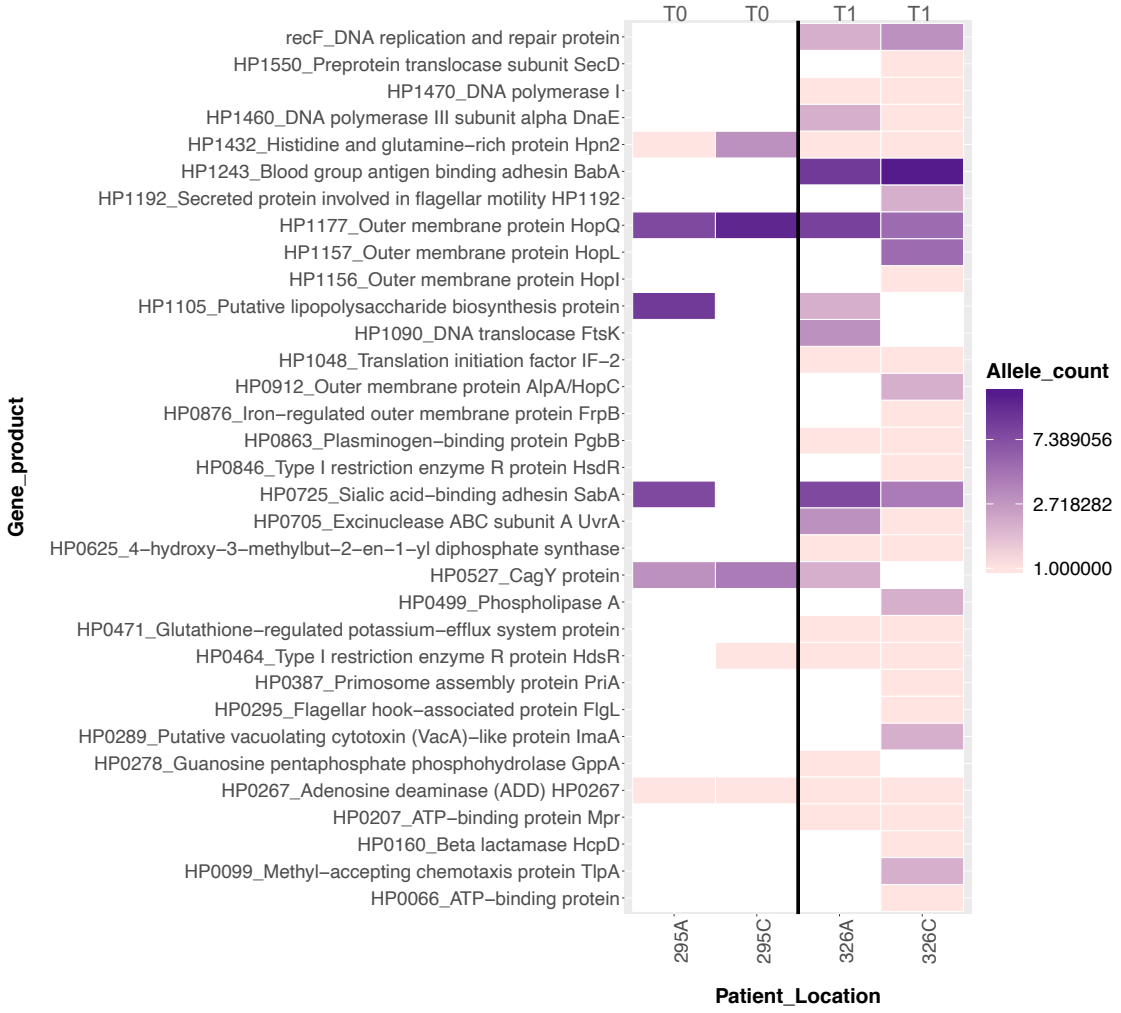
To determine which, if any, of these hypotheses is true would require further experimental investigation.

Figure 6.3 Heat map of minor allelic gene products before and after eradication therapy failure in patient 249/537 (sequential set 1)



Heat map of all minor allelic variant genes/gene products within sequential set 1 samples 249C (time point 0 – T0) and 537A (time point 1 – T1). Genetic diversity can be compared by looking between samples. Number of different polymorphic genes/associated gene products can be identified per sample. Colour intensity indicates a higher number of polymorphic positions within these genes/associated genes by product. This approach tries to keep together observed polymorphic diversity within genes by gene name and where no gene name is provided (by PROKKA) a unique gene number for each patient sample is provided.

Figure 6.4 Partial heat map of minor allelic gene products before and after eradication therapy failure in patient 295/326 (sequential set 2)



Heat map of all minor allelic variant genes/gene products within sequential set 2 samples 295A plus 295C (time point 0 = T0) and 326A plus 326C (time point 1 = T1). Genetic diversity can be compared by looking between samples. Number of different polymorphic genes/associated gene products can be identified per sample. Colour intensity indicates a higher number of polymorphic positions within these genes/associated genes by product. This approach tries to keep together observed polymorphic diversity within genes by gene name and where no gene name is provided (by PROKKA) a unique gene number for each patient sample is provided. Only the most diverse genes across all samples (Chapter Four, figure 4.6) are displayed in this figure due to the very high number of allelic genes within sequential set 2 time point 1 populations. Where there was no allelic variation within the genes depicted in Chapter Four figure 4.6, they were removed. A full version of this heat map of allelic genes can be observed in the Appendix (figure 10.6.1) and a full resolution image can be downloaded from the appendix directory (https://myntuac-my.sharepoint.com/:f:/r/personal/n0667645_my_ntu_ac_uk/Documents/OneDrive_link?csf=1&e=UuIp26).

Table 6.2 Sequential set 1 genes harbouring alleles with minor allele frequencies 12.5% or greater

Gene product (gene abbreviation)	Sample ID (time point)	
	249C MAF (T0)	537A MAF (T1)
537A_00946_hypothetical protein	N/A	0.445
HP0057_Hypothetical protein HP0057	0.174	N/A
HP0181_Membrane protein required for colicin V production	N/A	0.302
HP0392_Histidine kinase CheA	N/A	0.325
HP0471_Glutathione-regulated potassium-efflux system protein	N/A	0.271
HP0725_Sialic acid-binding adhesin SabA	0.279	N/A
HP0725_Sialic acid-binding adhesin SabA	0.422	N/A
HP0771_Hypothetical protein HP0771	0.138	N/A
HP1102_6-phosphogluconolactonase	0.176	N/A
HP1177_Outer membrane protein HopQ	0.230	0.290
HP1177_Outer membrane protein HopQ	0.388	0.391
HP1177_Outer membrane protein HopQ	0.379	0.385
HP1177_Outer membrane protein HopQ	N/A	0.173
HP1177_Outer membrane protein HopQ	N/A	0.143
HP1251_Oligopeptide ABC transporter permease OppB	N/A	0.477
HP1314_50S ribosomal protein L22	N/A	0.326
HP1322_Hypothetical protein HP1322	N/A	0.431

Table showing genes that harboured allelic positions with a minor MAF ≥ 0.125 (*i.e.* 12.5% of reads supporting a minor allele variant at that position). Genes with multiple instances refer to different allelic positions. Genes with allelic positions < 0.125 were filtered out to reduce table length and display of most significant frequency changes post eradication therapy. T0 = time point 0, T1 = time point 1. N/A refers to the specific allelic position not observing a MAF ≥ 0.125 at the denoted time point. Genes were annotated by PROKKA (Chapter Two, section 2.13).

Table 6.3 Sequential set 2 genes harbouring alleles with minor allele frequencies 12.5% or greater

	Sample ID (time point)			
Gene product (gene abbreviation)	295A (T1)	295C (T1)	326A (T2)	326C (T2)
295A_01271_hypothetical protein	0.455	N/A	N/A	N/A
326A_00310_hypothetical protein	N/A	N/A	0.151	N/A
326A_00311_hypothetical protein	N/A	N/A	0.171	N/A
326A_02270_hypothetical protein	N/A	N/A	0.269	N/A
326A_02270_hypothetical protein	N/A	N/A	0.375	N/A
326A_02270_hypothetical protein	N/A	N/A	0.125	N/A
326C_00844_hypothetical protein	N/A	N/A	N/A	0.125
ccpA_Catabolite control protein A	N/A	N/A	0.138	N/A
HP0162_Probable transcriptional regulatory protein HP0162	0.174	N/A	N/A	N/A
HP0207_ATP-binding protein Mpr	N/A	N/A	N/A	0.133
HP0349_CTP synthetase PyrG	N/A	N/A	N/A	0.143
HP0379_Alpha1,3-fucosyltransferase FutA	N/A	0.214	N/A	0.155
HP0379_Alpha1,3-fucosyltransferase FutA	N/A	N/A	N/A	0.210
HP0379_Alpha1,3-fucosyltransferase FutA	N/A	N/A	N/A	0.167
HP0394_UDP-2,3-diacylglucosamine hydrolase	N/A	N/A	0.273	N/A
HP0466_Hypothetical protein HP0466	0.242	N/A	N/A	N/A
HP0527_CagY protein	N/A	N/A	0.250	N/A
HP0543_CagF protein	N/A	N/A	N/A	0.137
HP0544_CagE protein	N/A	N/A	0.209	N/A
HP0559_Acyl carrier protein AcpP	N/A	N/A	0.232	N/A
HP0734_Ribosomal protein S12 methylthiotransferase RimO	N/A	N/A	N/A	0.155
HP0738_D-alanine--D-alanine ligase	N/A	N/A	N/A	0.170
HP0746_Hypothetical protein HP0746	N/A	N/A	0.180	N/A
HP0807_Iron(III) dicitrate transport protein FecA	0.153	N/A	N/A	N/A
HP0953_Hypothetical protein HP0953	N/A	N/A	N/A	0.127
HP1041_Flagellar biosynthesis protein FlhA	0.213	N/A	N/A	N/A
HP1057_Hypothetical protein HP1057	N/A	N/A	0.192	N/A
HP1110_Pyruvate flavodoxin oxidoreductase subunit alpha PorA	N/A	N/A	N/A	0.128
HP1177_Outer membrane protein HopQ	0.148	0.345	N/A	0.155
HP1177_Outer membrane protein HopQ	N/A	0.427	N/A	N/A
HP1177_Outer membrane protein HopQ	N/A	0.516	N/A	N/A
HP1177_Outer membrane protein HopQ	N/A	0.389	N/A	N/A
HP1192_Secreted protein involved in flagellar motility HP1192	N/A	N/A	N/A	0.125
HP1243_Blood group antigen binding adhesin BabA	N/A	N/A	0.366	0.463

HP1243_Blood group antigen binding adhesin BabA	N/A	N/A	0.357	0.407
HP1243_Blood group antigen binding adhesin BabA	N/A	N/A	0.279	0.264
HP1243_Blood group antigen binding adhesin BabA	N/A	N/A	0.272	0.272
HP1243_Blood group antigen binding adhesin BabA	N/A	N/A	0.648	0.531
HP1243_Blood group antigen binding adhesin BabA	N/A	N/A	0.607	0.529
HP1243_Blood group antigen binding adhesin BabA	N/A	N/A	0.605	0.522
HP1243_Blood group antigen binding adhesin BabA	N/A	N/A	0.388	0.326
HP1243_Blood group antigen binding adhesin BabA	N/A	N/A	N/A	0.153
HP1424_Hypothetical protein HP1424	N/A	N/A	N/A	0.180
HP1431_Ribosomal RNA small subunit methyltransferase A	N/A	N/A	N/A	0.161
smc_Chromosome partition protein Smc	N/A	N/A	N/A	0.176
smc_Chromosome partition protein Smc	N/A	N/A	N/A	0.149
smc_Chromosome partition protein Smc	N/A	N/A	N/A	0.169
smc_Chromosome partition protein Smc	N/A	N/A	N/A	0.137

Table showing genes that harboured allelic positions with a MAF ≥ 0.125 (*i.e.* 12.5% of reads supporting a minor allele variant at that position). Genes with allelic positions < 0.125 were filtered out to reduce table length and display only the most significant frequency changes post eradication therapy. T0 = time point 0, T1 = time point 1. N/A refers to the specific allelic position not observing a MAF ≥ 0.125 at the denoted time point. Genes were annotated by PROKKA (Chapter Two, section 2.13). Some gene products / genes are displayed more than once due to multiple allelic positions with a MAF ≥ 0.125.

6.3.3. Comparing the single colony isolates from sequential sets 1 and 2 to their respective patient reference genomes before and after failed eradication therapy using the nucleotide basic local alignment tool

6.3.3.1. Sequential set 1

The single colony isolates from sequential set 1 were compared by BLASTN and visualised by BRIG (figure 6.5A) (Alikhan *et al.*, 2011). The comparisons indicate that a similar pattern of 100% BLASTN identity and gaps can be observed for strains pre- and post- eradication therapy. This observation was also true for isolates taken from

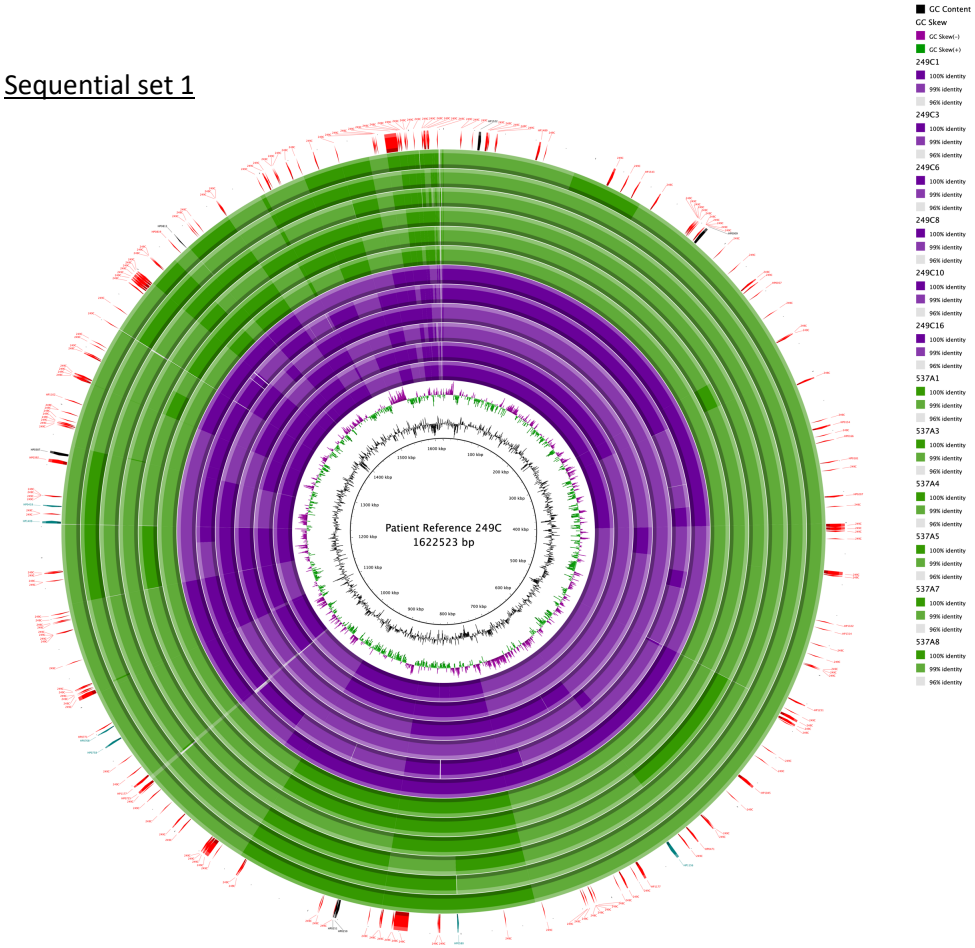
the antrum and corpus of other patients (Chapter Five, figure 5.3; Appendix figures 11.5.1 – 11.5.8). Despite these groups of isolates being taken from patient 249/537 1,366 days apart, they are remarkably similar. There is very little genetic variation between colonies isolated before and after eradication therapy and the variable genes appear to be spread evenly across the genome.

6.3.3.2. Sequential set 2

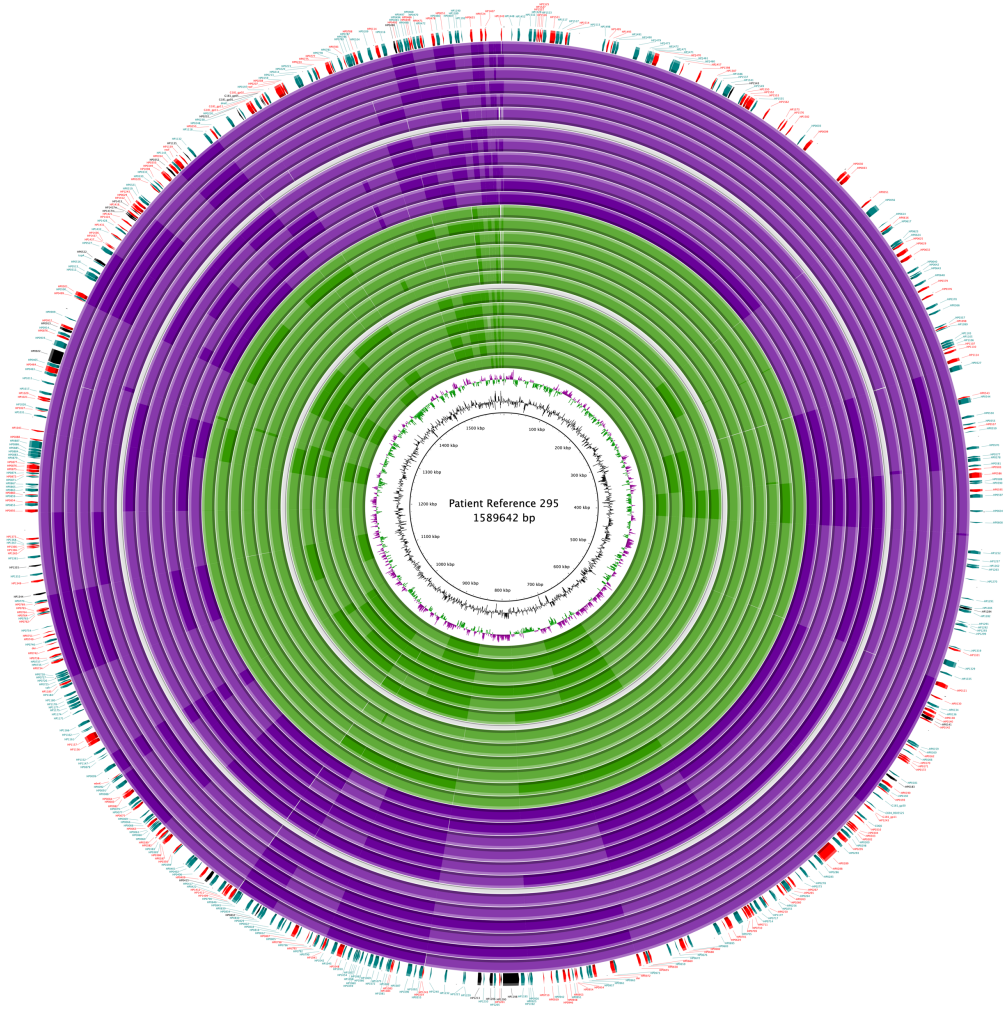
Sequential set 2 antrum and corpus isolates before and after failed eradication therapy also appear to be genetically related (>96 % BLASTN identity), with time point specific patterns of BLASTN identity (figure 6.5B). This suggests that all of the isolates originated from the same infecting strain. There was a dramatic increase in genetic diversity after failed eradication therapy and the variable genes are labelled on the outside concentric rings for reference (figure 6.5B).

Figure 6.5 Nucleotide identity comparison between sequential set 1 and set 2 isolates

A) Sequential set 1



B) Sequential set 2



BLAST Ring Image Generator (BRIG) plots of sequential set 1 (figure A) and sequential set 2 (figure B) single colony genome assemblies constructed as described in section 6.2.6. The rings are denoted in the legend to the right of each figure in the order they are described. The green concentric ring represents the strains isolated from the antrum where the purple represents the corpus strains. Genes were colour coded as follows; black = identified as variable gene(s) by mapping single colony reads to patient reference consensus genome only, teal = identified allelic genes by the deep sequencing minor allele pipeline only, red = identified by both methodologies (Chapter Five, section 5.2.6). An upper BLASTN identity threshold of 99% and lower identity of 96% were used (section 6.2.6). The legend was removed in figure B to improve visualisation. Figure B concentric rings from the inner most green ring to the outer purple ring depict the following strains: 295A1, 295A2, 295A3, 295A4, 295A5, 295A6, 326A22, 326A23, 326A24, 326A26, 326A25, 326A7, 295C1, 295C2, 295C4, 295C6, 295C7, 295C8, 326C1, 326C2, 326C3, 326C4, 326C5 and 326C6. A full resolution version can be found at https://myntuac-my.sharepoint.com/:f:/r/personal/n0667645_my_ntu_ac_uk/Documents/OneDrive_link?csf=1&e=UuIp26.

6.3.4. Phylogenetic analysis of sequential data sets

6.3.4.1. Sequential set 1

Sequential set 1 single colony isolates split into two clades (figure 6.6A). These clades definitively split into pre- and post- eradication therapy clusters. Despite this, there were very few genetic differences between the individual isolates with only 18 SNPs separating the longest branch tip to tip distance between isolates (249C and 537A4/7). This indicates that there was very little genetic divergence between the sampling points (1,366 days). The eradication therapy had very little effect on the *H. pylori* population. Considering the spontaneous mutation rate of *H. pylori* of 3×10^{-5} mutations per site per year (Björkholm et al., 2001), the patient reference genome (249C) size of 1.63 Mbp, would equate to an expected number of mutations over 3.74 years of ~20 SNPs. The expected number of SNPs by natural mutation is thus very similar to that observed between the longest phylogenetic branches. This further suggests that recombination was not driving diversification, but natural mutation was.

The limited diversity observed between the two time points is much lower than expected when considering the challenge of eradication therapy. The observed genetic diversity from single colony sequencing was consistent with the results obtained by the deep population sequencing analysis as previously discussed (section 6.3.2). Pulling together all of the genetic and phenotypic evidence, it is likely that this patient held a low diversity *H. pylori* infection before eradication therapy that was already completely resistant to the prescribed therapy. Under these circumstances, the drugs did not exert a selection pressure on the bacterial population, which remained essentially unchanged after treatment.

6.3.4.2. Sequential set 2

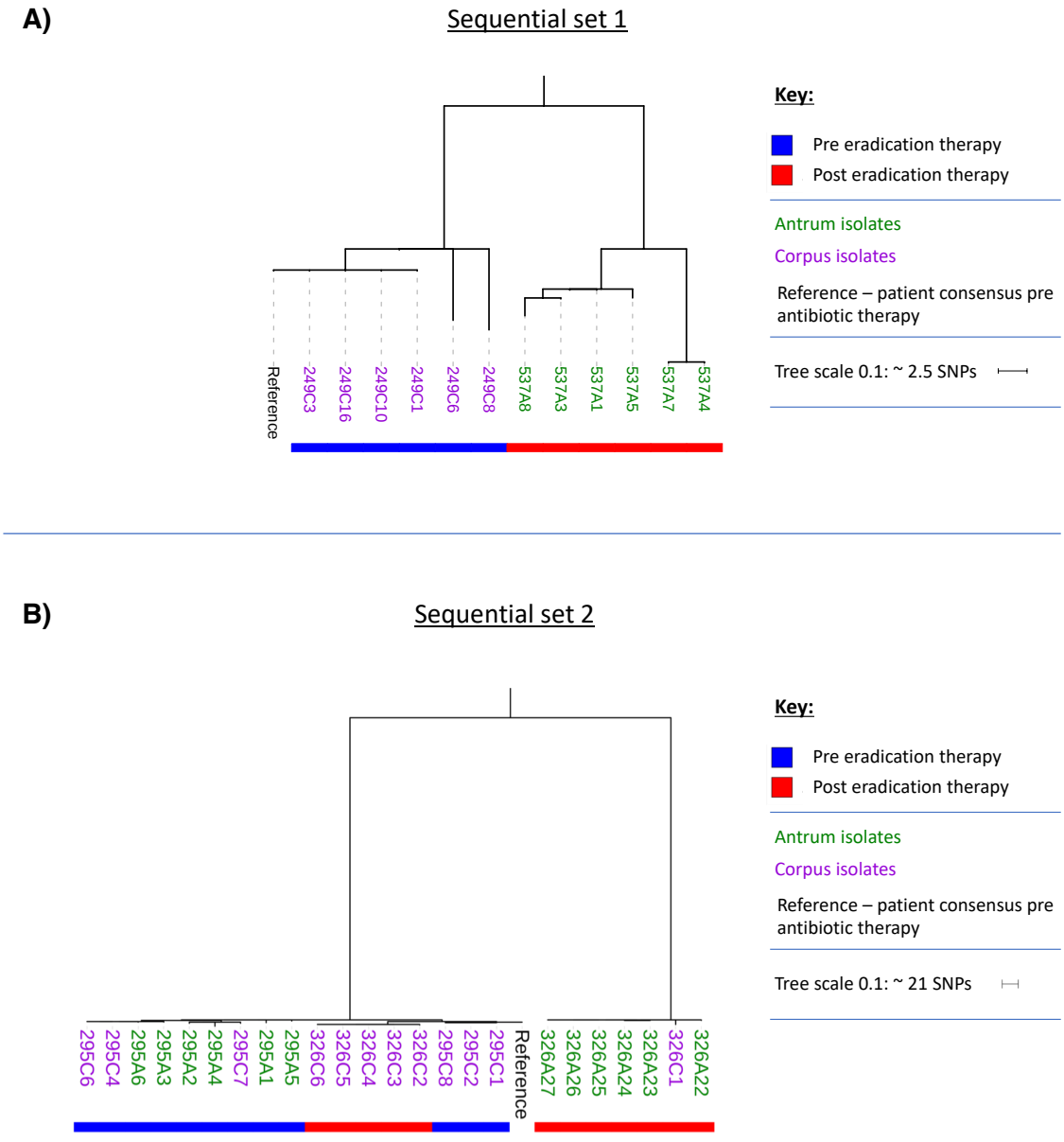
Sequential set 2 did not split into pre- and post- eradication therapy clusters (figure 6.6B). However, all antrum and one corpus strains post eradication therapy clustered away from all other strains. Five out of six corpus strains clustered with the majority of the pre-eradication therapy samples. This suggests that the corpus population strains underwent very little genetic change between the sampling points. This is somewhat contradictory to that observed by the population deep sequencing results (figure 6.4). The contradiction could be explained by inadequate single colony isolation, where a set

of very similar strains were taken that do not reflect the true population diversity. Alternatively, these results could suggest that while the corpus population is more genetically diverse after failed eradication therapy, there are strains from pre-therapy that are still abiding. If this is correct, it could suggest that the corpus strains are more naturally adapted to resist eradication therapy, despite being closely related to the antral strains. Furthermore, it is thought that antibiotics may degrade due to low pH, suggesting a possible reason why the *H. pylori* infection showed little genetic change, potentially attributed to antibiotic degradation in the oxyntic corpus niche (Vallve et al., 2002).

The antrum strains post eradication therapy were substantially different to the antrum strains isolated before therapy (figure 6.6B). Despite the big genetic differences between these strains, they were not thought to be from a re-infection event. The reasoning behind this focuses on the pan-genome of these isolates. In Chapter Five figure 5.4, the pan-genome clustering of all strains of sequential set 2 clustered almost indistinguishably with one another. Due to the big differences in pan-genomes between different strains and populations, a new or re-infection of *H. pylori* post eradication therapy would presumably change the genes clustering within the pan-genome (Chapter Four, figure 4.13; Chapter Five, figure 5.4). Therefore, this is strong evidence against a re-infection hypothesis. Furthermore, figure 6.5B shows that all the isolates from sequential set 2 share a close relationship to the patient reference genome with a BLASTN identity > 96% covering most of the genome with very few gaps.

The highly diverse antral strains post eradication therapy in sequential set 2 (figure 6.6B) compliment the analysis by the deep population dataset (figure 6.4). Not only is there a substantial increase in genetic diversity within the population post eradication therapy (figure 6.4), there is also a big increase in genetic diversity between the core genomes and population structure at a single colony resolution (figure 6.5B; figure 6.6B). Again, this suggests that while there were some persistent strains, the populations have substantially changed due to the challenge of eradication therapy.

Figure 6.6 Phylogenetic analysis of sequential datasets



Population structure of single colony isolates from before and after eradication therapy. Figure A shows the phylogeny of isolates taken from sequential set 2 while figure B shows isolates from sequential set 2. Phylogenetic trees were constructed as described in section 6.2.7. Phylogenetic trees were constructed and visualised by the interactive tree of life (iTOL) where they were rooted by the midpoint (Letunic and Bork, 2007). Sample labels were colour coded by their location (antral isolates = green, corpus isolates = purple). A colour strip was added to ease in the visualisation of isolates from the different time points (pre-eradication therapy / T0 = blue, post-eradication therapy / T1 = red).

6.3.5. Recombination detection between strains before and after failed eradication therapy

6.3.5.1. Sequential set 1

There was little observable recombination between isolates from sequential set 1 (figure 6.7A1-2). Where recombination was detected it was confined towards the end of the isolate genomes. As previously discussed in Chapter Five section 5.3.5, recombination observed towards the ends of the genomes might not be an accurate representation due to smaller contigs towards the end of the genomes that are usually of much higher sequencing depth, suggesting a region of the genome that was not well assembled. As a result, this could inflate the number of allelic variants in these segments. However, recombination is observable between isolates and from outside sources in both methodologies. The fastGEAR analysis highlighted one region of recombination between isolates 249C6 and 249C8 which was not detected by Gubbins, suggesting that the fastGEAR methodology is able to detect additional recombination events.

The low level of recombination for sequential set 1 is unsurprising, given the low level of genetic variation within the population (figure 6.3; figure 6.5A; figure 6.6A).

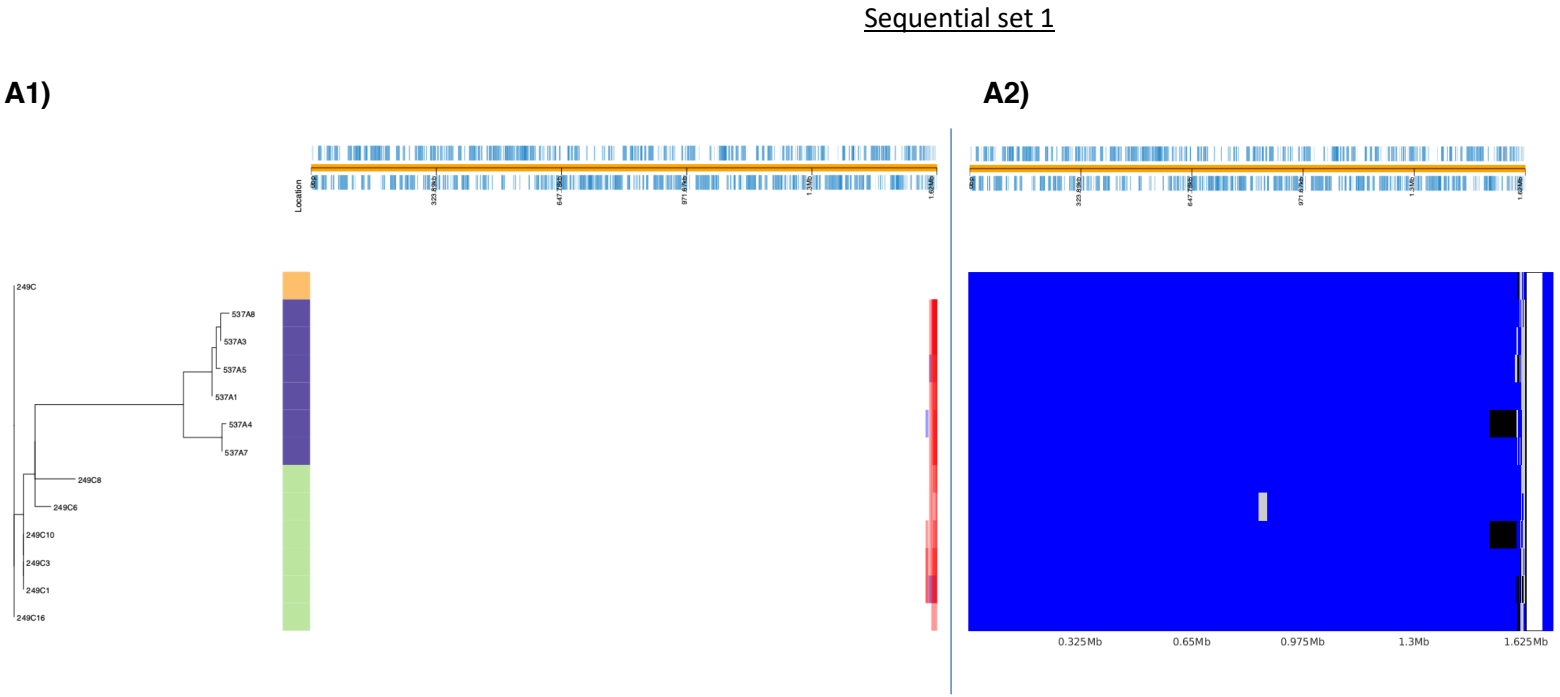
6.3.5.2. Sequential set 2

There were more recombination events detected between the sequential set 2 isolates (figure 6.7A-B). Recombination was not observed between the smaller clade and the larger clade of isolates. However, recombination was observed between isolates taken before and after eradication therapy by the fastGEAR analysis (figure 6.7B2). This suggests that single colony isolates were recombining during and/or after eradication therapy.

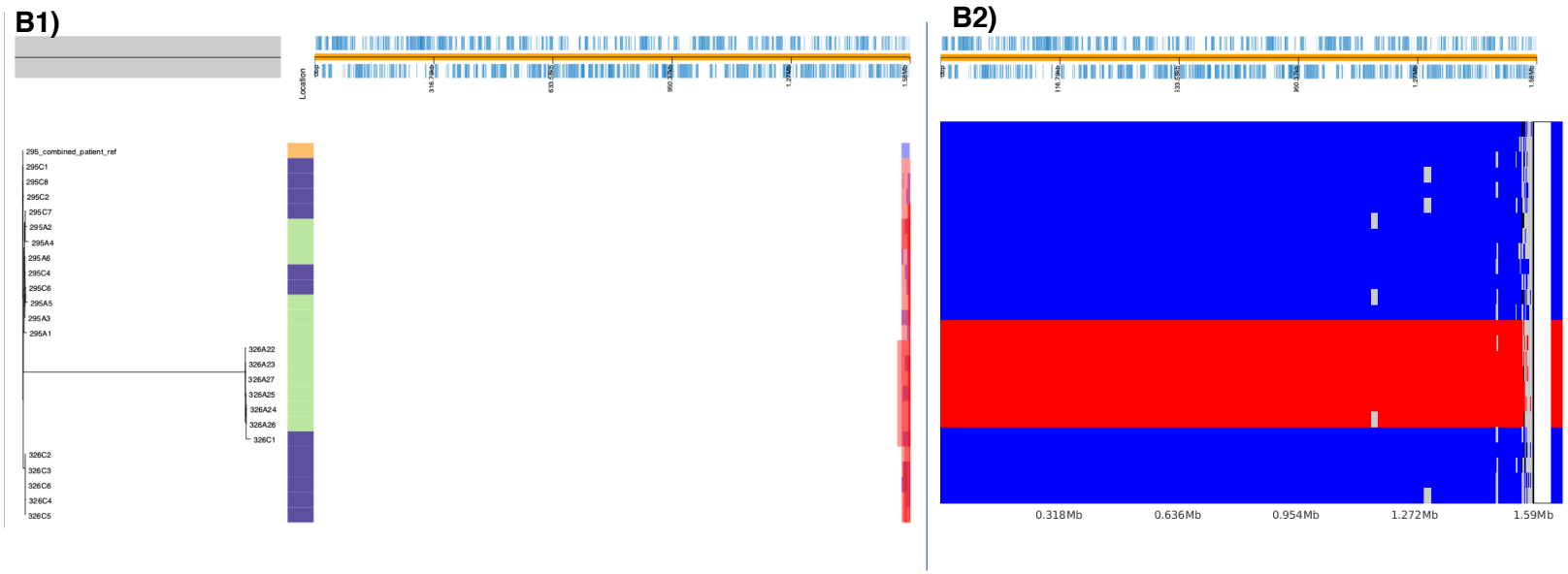
Due to the lack of evidence of between clade recombination (blue bars) and recombination from an outside source (black bars) in the divergent smaller clade (red lineage containing strains 326A22-27 and 326C1), it could be argued that the genetic diversity of this clade was not a result of extensive recombination (figure 6.7B2). Such an observation could suggest that this diversity is more a result of natural random mutation of sub-strains/quasispecies within the population. If true, this observation could support the aforementioned second theory designed to explain the increased

observation of genetic diversity post failed eradiation therapy (section 6.3.2). In short, this theory centres around the idea of within patient genetic diversity resulting in sub-populations/quasispecies prior to eradication. These sub-populations might be present in low abundance but may hold higher phenotypic resistance to antimicrobials due to random genetic mutation. These sub-population holding higher phenotypic resistance to antimicrobials might persist within the gastric niche but not fix due to competition from the dominant population. Lieberman *et al.* (2014), observed that *Burkholderia dolosa* intra-sample mutants coexist within the population and rarely sweep to fixation, supporting this theory, alongside the within patient genetic diversity observed in Chapter Four figure 4.6. Upon antibiotic challenge, the dominant *H. pylori* population that is sensitive would be cleared, leaving behind multiple small sub-populations of *H. pylori* with higher natural resistance to the eradication therapy drugs. Subsequent outgrowth of these populations could cause the observed increase in population genetic diversity after eradication therapy (figure 6.4). Since these low abundance sub-strains did not interact substantially with the dominant population prior to eradication therapy, little evidence of recombination between strains before and after eradication therapy was detected (figure 6.7B1-2). Recombination events were only observed between the lineages post eradication therapy.

Figure 6.7 Recombination between sequential set 1 and sequential set 2 isolates respectively



Sequential set 2



Recombination detected by Gubbins (figures A1 and B1) and fastGEAR (figures A2 and B2) as described in section 6.2.8. Figures A1 and A2 show sequential set 1 while figures B1 and B2 relate to sequential set 2. The genomic scale bar is depicted at the top of each figure with blue markings highlighting genes along the genome. The core genome phylogeny is depicted to the left followed by coloured blocks differentiating antrum isolates (purple), corpus isolates (green) and reference (orange). The centre of figures A1 and B1 depict sites of recombination. White space (no recombination sites), blue blocks (single site of recombination) and red blocks (shared site of recombination). Figures A2 and B2 represent recombination detected by fastGEAR as described in section 6.2.8. Lineages detected by fastGEAR are colour coded and have been ordered to match the strain locations in the Gubbins analysis for ease of cross comparisons. Grey blocks represent recombination sites within lineages while black blocks represent recombination from an outside strain. Coloured blocks within lineages infer direction of recombination events between lineages.

6.4. Future work

Determining the antibiotics provided to the patients would greatly help in the interpretation of both sequential set 1 and 2 results. Understanding of this would help to draw better conclusions of the effect of these drugs on the *H. pylori* population proceeding failed eradication therapy.

Genes observed to have higher MAF post eradication therapy would be good targets for mutagenesis experiments to observe their phenotypic effects on the antibiotic susceptibility and virulence of *H. pylori,* and to see if they contribute to eradication therapy failure. If so, these genes could be further targeted for improved eradication therapy or in diagnosis to highlight patients at risk of treatment failure.

Although ethically problematic, continued sampling after failed eradication at additional time points could further the understanding of how the *H. pylori* populations continue to change and adapt after failed eradication therapy, specifically in the context of sequential set 2.

This study only investigates sequential samples from two patients, one of which only had samples available from opposite niches of the stomach at the different time points. Ideally, additional patients should be identified where antrum and corpus samples are available for analysis before and after eradication therapy. Expanding the dataset would help provide a better picture of what happens to the *H. pylori* infection before and after eradication therapy.

Further investigations into the theories presented in this study regarding the possible reasons for increased genetic diversity after failed eradication (sequential set 2) is needed.

7. Chapter Seven: Conclusion

The prevailing aim of this thesis was to investigate the within niche and between niche genetic diversity of *H. pylori* populations taken from the same stomach. Secondary aims developed during the course of the thesis included between patient *H. pylori* population comparisons, antimicrobial resistance assay development with a focus on standardisation.

With increasing global antimicrobial resistance to virtually all clinically relevant bacteria, including *H. pylori* with a particular concern for clarithromycin resistance, it is perhaps past time to improve antimicrobial susceptibility testing for this bacterial infection. While antimicrobial susceptibility testing at initial diagnosis of *H. pylori* infection would be preferable, it is perhaps ethically more acceptable to implement susceptibility testing after the first unsuccessful eradication therapy attempt as a carcinogen and the invasive nature of culture collection. As it stands today, eradication failure triggers a second line eradication therapy attempt. It could be argued that in the era of increasing eradication therapy failure, screening for antimicrobial susceptibility is perhaps a way to slow the trend of increasing antibiotic resistance and provide better, more patient specific eradication therapy regimes. This might help preserve the efficacy and continued use of antibiotics in the eradication of *H. pylori* for generations to come.

However, limited understanding of within patient *H. pylori* antimicrobial susceptibility in terms of testing guidelines and how intra-strain diversity might affect phenotypic susceptibility has perhaps hampered culture-guided eradication therapy. These were addressed in Chapter Three, where a standardised antimicrobial resistance assay was presented, and a review of the literature identified potential breakpoints for an inexpensive but powerful (able to identify resistant colonies and second zones of inhibition) antimicrobial disk diffusion susceptibility assay. It was found that antrum and corpus antimicrobial susceptibility profiles can vary within individual patients. In terms of eradication therapy, these differences could be the difference in determining a *H. pylori* infection as sensitive or resistant suggesting that culture-guided eradication therapy is best informed from cultures taken from multiple gastric biopsy sites. However, future studies should investigate if there are intra-niche antimicrobial sensitivity differences, which might provide further guidance on which areas of the stomach should be selected for culture-guided eradication therapy.

Extensive genomic diversity is a hallmark of *H. pylori* infection both globally, locally and within infected patients. To date, all studies have investigated genomic diversity using

DNA fingerprinting techniques, gene specific analysis and more recently, using whole genome sequencing of multiple single colony isolates obtained from individuals, family members and animal models of infection. This has progressed our understanding of *H. pylori* genetic diversity and its potential role in host adaptation, disease development and persistence. However, these studies are limited by sample size, thus the true *H. pylori* diversity is yet to be elucidated at a population level.

In Chapter Four, a novel deep population sequencing method was applied to *H. pylori* clinical sweeps cultured from clinical biopsies. A novel read mapping approach was used to identify 'common' and 'minor' population allelic diversity. The former was used to identify alleles of high frequency while the latter was used to identify low frequency alleles. This novel approach presented a more accurate snapshot of population diversity, supporting observations from other studies of diverse genes and highlighted novel genes with allelic diversity.

A total of 585 and 4,872 polymorphic sites were detected across all samples (excluding sample 565C; n=32) by common and minor allelic calling pipelines respectively. Excluding hypothetical proteins, the most polymorphic genes across all patient samples were: OMP (26/32), Glutamyl-tRNA synthetase (17/32), OMP assembly factor YaeT precursor (17/32), Methyl-accepting chemotaxis gene *tlpA* (16/32), *vacA* paralog (14/32), Type III restriction-modification system methylation subunit (13/32), Methyl-accepting chemotaxis protein (12/32), Lipopolysaccharide biosynthesis protein (12/32), Type I restriction-modification system subunit R (12/32), Tetratricopeptide repeat family protein (11/32), Adenine specific DNA methyltransferase (10/32), Glutathione-regulated potassium-efflux system protein KefB (10/32) and Type I restriction-modification system subunit S (10/32). This study revealed that within niche polymorphic diversity was not preferably selected for or against within one niche to the other in relation to the antrum and the corpus. However, this study did reveal more nonsynonymous aligned *fixed* SNPs between the antrum and corpus than synonymous SNPs suggesting that selection pressures are potentially acting between these niches of the human stomach but presumably over a longer timescale. No associations between disease severity and total number of nonsynonymous mutations were found.

The results obtained from this study have important implications in vaccine rational design as genetically diverse antigens would not make good vaccine targets. This study further elucidates the within patient, between patient, within niche and between niche

genetic diversity of *H. pylori* populations by combining a multitude of genetic bioinformatics analysis methodologies. This novel approach provides further insights into *H. pylori* host and niche adaptation as well as how this pathogen is able to persist as a chronic infection. Furthermore, this author presents a rich resource of population gene diversity for other researchers to use and potentially follow up on. Finally, this study highlights important bioinformatics limitations. For example, so called 'fixed' SNPs identified by comparative whole genome alignments can often still be polymorphic at the SNP site, suggesting that these positions are not truly 'fixed' within the populations where polymorphic variation might hold insights into past selection pressures, as first suggested by Lieberman *et al.* (2014).

The methods and analysis presented in Chapter Four acted as a proof of concept for the capture of population diversity at single time point of infection. Future studies might benefit from this population method in sequential samples to observe how selection acts on specific genes over time.

By comparing within niche genetic diversity of single colony isolates using a read mapping approach to a reference genome, it was possible to validate the deep population minor allelic calling pipeline (Chapter Four). This revealed a good fit between the two sequencing methodologies suggesting that the population deep sequencing minor allele pipeline presented in Chapter Four was accurate.

The findings of Chapters Four and Five highlighted many genes that appear to be genetically diverse across numerous patient samples. While there was good agreement between the two different approaches in many instances, PCR-based validation could also be carried out. Furthermore, while genetic diversity is clear, gene expression levels are not. Therefore, DNA microarrays and/or reverse transcription PCR could be used to investigate the effects allelic diversity might have on gene expression. Additionally, while allelic diversity of specific genes might infer biological significance, the biological effect is less clear. In order to test this, more *conventional* studies would need to follow such as site directed mutagenesis of wild type strains and the comparison to wildtype strains through phenotypic assays.

Chapter Six consisted of a unique dataset of sequential *H. pylori* cultures taken before and after failed eradication therapy. Sequential set 1 showed little change in antimicrobial sensitivity before and after eradication therapy. Furthermore, the genetic

diversity between the different timepoints remained comparable. This suggests that the eradication therapy followed in this patient had little to no effect on the *H. pylori* population. This was potentially due to pre-existing resistance to the prescribed eradication regime. However, the sampling timepoints were ~3.7 years apart which might not be representative of changes during or immediately following eradication therapy.

Conversely, sequential set 2 showed antimicrobial sensitivity differences after failed eradication therapy ~4.9 months after the first sampling timepoint. High within niche population genetic diversity was observed post failed eradication therapy. This contradicted previous observations by other studies of restricted genetic diversity and suggested that failed eradication therapy might be a driver of genetic diversity. Increased diversity might drive additional resistance phenotypes to other antibiotics, resulting in further eradication therapy failures. However, further studies will be required to follow up these results to understand the implications of these findings as well as a more extensive sequential sample set from additional patients.

Deep population sequencing and analysis proved to be an effective method to investigate within population genomic diversity of *H. pylori*. However, this read mapping approach relied on *de novo* consensus assemblies of the population from short read sequencing data. This was useful in the analysis of population variation (due to extensive genetic diversity rendering popular reference genomes unsuitable). While advantageous, this method had its own limitations. Of particular note is that genomes are almost never complete, with contig breaks around areas of the genome that could not be resolved. This has implications in read mapping where reads might not align correctly in this region. Furthermore, Illumina short read sequencing (by synthesis) is prone to errors in repetitive sequences.

A comprehensive approach for the future might be to take a single colony isolate from the population to be studied, for sequencing on both the Illumina short read and Oxford Nanopore Technologies long read MinION platforms. This would provide an accurate representative 'reference' genome from the population to compare against. The population should then be deep sequenced as described in Chapter Four and the minor allele calling pipeline executed by mapping the population reads to this population representative 'reference' genome.

7.1. Summary

This thesis presents many novel *H. pylori* analysis techniques and findings. Entire *H. pylori* populations from single biopsies were studied, something that has not been presented before. This has contributed to further novel insights into *H. pylori* host adaptation and persistence. Deep sequencing of sequential samples taken before and after failed eradication was also novel and suggested that failed eradication therapy could be a driver of genetic diversity, with potential implications in eradication therapy.

It is hoped that this thesis will act as a useful platform to guide future research into allelically diverse genes detected within *H. pylori* populations (Chapter Four), within and between different niches of the human stomach (Chapter Four and Five) and from *H. pylori* populations after eradication therapy failure (Chapter Six). The development and standardisation of an internationally recognised *H. pylori* antimicrobial disk diffusion assay is needed, as is a better understanding of the *H. pylori* antibiotic resistance within the UK (Chapters Three and Six).

8. Acknowledgements

Dr. Jody Winter – I could write several pages about how much you have helped and guided me throughout my PhD study. But, put simply, I could not have hoped for a better Director of Studies. You have been a true inspiration to myself and your depth of knowledge is something I can only hope to attain one day. Your dedication, patience (which I am sure that I must have tested!) and kindness is something that has really made an impression on me. You have made me feel like part of the wider research group and have always encouraged me to do things that were initially out of my comfort zone. I have learnt so much from you over the years and for that I am truly grateful. It has been a pleasure working with you and I hope we can continue to do so in the future.

Dr. Ben Dickins – you have helped me considerably over the years and I appreciate that you have entertained, quite frankly, absurd requests and questions from myself. I have enjoyed our many discussions, many of which have gone off track from the meeting agendas. You have always made me think twice about my bioinformatics analysis and helped me more than you might realise. I often felt a little less stressed when seeing you run around campus at breakneck speed!

Dr. Jonathan Thomas – thank you for your advice and help throughout my study. You are one of the kindest and most patient people I have ever met. I could not have done the Oxford Nanopore Technologies MinION sequencing without you offering the spare capacity to me. Furthermore, I could not have done the sequencing on this platform or the subsequent analysis without your expert guidance. For that, I am very grateful. What has surprised me the most during my study is just how quickly alcohol could affect you!

Professor Alan McNally – originally my second supervisor, you have helped guide the microbial genomics aspect of my work and I greatly appreciate all the excellent advice you have given me throughout my study as an external advisor.

Dr. Steven Dunn – thank you for all of your advice during the start of my study, especially in regard to the use of the command line/BASH. I feel that I must have been a bit of a nuisance at times! I am also grateful that you showed me how to sequence on the Illumina MiSeq platform.

Professor John Atherton, Dr. Karen Robinson, Joanne Rhead and all of the team members at the University of Nottingham – thank you for all of the advice, guidance,

support, general requests and *Helicobacter pylori* strains you provided. Without your help, this project would not have been possible.

All members of the Microbial Resistance, Omics and Microbiome research group at Nottingham Trent University – regular group meetings have played an important role in my development as a researcher ranging from confidence building in informal presentations, to broadening my general knowledge in microbiology research. For that, I thank you all.

9. References

Abedrabbo, S., Castellon, J., Collins, K.D., Johnson, K.S. and Ottemann, K.M., 2017. Cooperation of two distinct coupling proteins creates chemosensory network connections. *Proceedings of the National Academy of Sciences of the United States of America*, 114(11), pp.2970–2975.

Achtman, M., Azuma, T., Berg, D.E., Ito, Y., Morelli, G., Pan, Z.J., Suerbaum, S., Thompson, S.A., van der Ende, A. and van Doorn, L.J., 1999. Recombination and clonal groupings within *Helicobacter pylori* from different geographical regions. *Molecular microbiology*, 32(3), pp.459–70.

Ahmadzadeh, A., Ghalehnoei, H., Farzi, N., Yadegar, A., Alebouyeh, M., Aghdaei, H.A., Molaei, M., Zali, M.R. and pour Hossein Gholi, M.A., 2015. Association of CagPAI integrity with severeness of *Helicobacter pylori* infection in patients with gastritis. *Pathologie Biologie*, 63(6), pp.252–257.

Ahmed, K.S., Khan, A.A., Ahmed, I., Tiwari, S.K., Habeeb, A., Ahi, J.D., Abid, Z., Ahmed, N. and Habibullah, C.M., 2007. Impact of household hygiene and water source on the prevalence and transmission of *Helicobacter pylori*: a South Indian perspective. *Singapore medical journal*, 48(6), pp.543–9.

Ahn, S., Costa, J. and Rettig Emanuel, J., 1996. PicoGreen quantitation of DNA: effective evaluation of samples pre- or post-PCR. *Nucleic Acids Research*, 24(13), pp.2623–2625.

Aihara, E., Closson, C., Matthis, A.L., Schumacher, M.A., Engevik, A.C., Zavros, Y., Ottemann, K.M. and Montrose, M.H., 2014. Motility and Chemotaxis Mediate the Preferential Colonization of Gastric Injury Sites by *Helicobacter pylori*. *PLoS Pathogens*, 10(7), p.e1004275.

Ailloud, F., Didelot, X., Woltemate, S., Pfaffinger, G., Overmann, J., Bader, R.C., Schulz, C., Malfertheiner, P. and Suerbaum, S., 2019. Within-host evolution of *Helicobacter pylori* shaped by niche-specific adaptation, intragastric migrations and selective sweeps. *Nature Communications*, 10(1), p.2273.

Akeel, M., Shehata, A., Elhafey, A., Elmakki, E., Aboshouk, T., Ageely, H. and Mahfouz, M., 2019. *Helicobacter pylori vacA, cagA* and *iceA* genotypes in dyspeptic patients from southwestern region, Saudi Arabia: distribution and association with clinical outcomes and histopathological changes. *BMC Gastroenterology*, 19(1), p.16.

Alandiyjany, M.N., Croxall, N.J., Grove, J.I. and Delahay, R.M., 2017. A role for the tfs3 ICE-encoded type IV secretion system in pro-inflammatory signalling by the *Helicobacter pylori* Ser/Thr kinase, CtkA. *PLOS ONE*, 12(7), p.e0182144.

Ali, A., Naz, A., Soares, S.C., Bakhtiar, M., Tiwari, S., Hassan, S.S., Hanan, F., Ramos, R., Pereira, U., Barh, D., Figueiredo, H.C.P., Ussery, D.W., Miyoshi, A., Silva, A. and Azevedo, V., 2015. Pan-genome analysis of human gastric pathogen *H. pylori*: Comparative genomics and pathogenomics approaches to identify regions associated with pathogenicity and prediction of potential core therapeutic targets. *BioMed Research International*, 2015.

Alikhan, N.-F., Petty, N.K., Ben Zakour, N.L. and Beatson, S.A., 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. *BMC Genomics*, 12(1), p.402.

Alm, R.A., Bina, J., Andrews, B.M., Doig, P., Hancock, R.E. and Trust, T.J., 2000a. Comparative genomics of *Helicobacter pylori*: analysis of the outer membrane protein families. *Infection and immunity*, 68(7), pp.4155–68.

Alm, R.A., Bina, J., Andrews, B.M., Doig, P., Hancock, R.E. and Trust, T.J., 2000b. Comparative genomics of *Helicobacter pylori*: analysis of the outer membrane protein families. *Infection and immunity*, 68(7), pp.4155–68.

Alm, R.A., Ling, L.-S.L., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., Smith, D.R., Noonan, B., Guild, B.C., deJonge, B.L., Carmel, G., Tummino, P.J., Caruso, A., Uria-Nickelsen, M., Mills, D.M., Ives, C., Gibson, R., Merberg, D., Mills, S.D., Jiang, Q., Taylor, D.E., Vovis, G.F. and Trust, T.J., 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature*, 397(6715), pp.176–180.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J., 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp.403–410.

Amieva, M.R., Vogelmann, R., Covacci, A., Tompkins, L.S., Nelson, W.J. and Falkow, S., 2003. Disruption of the Epithelial Apical-Junctional Complex by *Helicobacter pylori* CagA. *Science*, 300(5624), pp.1430–1434.

Ando, T., Ishiguro, K., Watanabe, O., Miyake, N., Kato, T., Hibi, S., Mimura, S., Nakamura, M., Miyahara, R., Ohmiya, N., Niwa, Y. and Goto, H., 2010. Restriction-modification systems may be associated with *Helicobacter pylori* virulence. *Journal of Gastroenterology and Hepatology*, 25, pp.S95–S98.

Andrews, S., 2010. FastQC - A Quality Control application for FastQ files. [online].

Angiuoli, S. V. and Salzberg, S.L., 2011. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, 27(3), pp.334–342.

Anon 1997. Current European concepts in the management of Helicobacter pylori infection. The Maastricht Consensus Report. European *Helicobacter pylori* Study Group. *Gut*, 41(1), pp.8–13.

Appelmelk, B.J., Simoons-Smit, I., Negrini, R., Moran, A.P., Aspinall, G.O., Forte, J.G., De Vries, T., Quan, H., Verboom, T., Maaskant, J.J., Ghiara, P., Kuipers, E.J., Bloemena, E., Tadema, T.M., Townsend, R.R., Tyagarajan, K., Crothers, J.M., Monteiro, M.A., Savio, A., De Graaff, J. and Graaff, J. De, 1996. Potential role of molecular mimicry between *Helicobacter pylori* lipopolysaccharide and host Lewis blood group antigens in autoimmunity. *Infection and immunity*, 64(6), pp.2031–40.

Aras, R.A., Small, A.J., Ando, T. and Blaser, M.J., 2002. *Helicobacter pylori* interstrain restriction-modification diversity prevents genome subversion by chromosomal DNA

from competing strains. *Nucleic acids research*, 30(24), pp.5391–7.

Arévalo-Jaimes, B.V., Rojas-Rengifo, D.F., Jaramillo, C.A., de Molano, B.M., Vera-Chamorro, J.F. and del Pilar Delgado, M., 2019. Genotypic determination of resistance and heteroresistance to clarithromycin in *Helicobacter pylori* isolates from antrum and corpus of Colombian symptomatic patients. *BMC Infectious Diseases*, 19(1), p.546.

Argent, R.H., Thomas, R.J., Aviles-Jimenez, F., Letley, D.P., Limb, M.C., El-Omar, E.M. and Atherton, J.C., 2008. Toxigenic *Helicobacter pylori* infection precedes gastric hypochlorhydria in cancer relatives, and H. pylori virulence evolves in these families. *Clinical Cancer Research*, 14(7), pp.2227–2235.

Armitano, R.I., Matteo, M.J., Goldman, C., Wonaga, A., Viola, L.A., De Palma, G.Z. and Catalano, M., 2013. *Helicobacter pylori* heterogeneity in patients with gastritis and peptic ulcer disease. *Infection, Genetics and Evolution*, 16, pp.377–385.

Ashokkumar, S., Agrawal, S., Mandal, J., Sureshkumar, S., Sreenath, G. and Kate, V., 2017. Hybrid therapy versus sequential therapy for eradication of *Helicobacter pylori*: A randomized controlled trial. *Journal of Pharmacology and Pharmacotherapeutics*, 8(2), p.62.

Atherton, J.C., Cao, P., Peek, R.M., Tummuru, M.K.R., Blaser, M.J. and Cover, T.L., 1995. Mosaicism in Vacuolating Cytotoxin Alleles of *Helicobacter pylori*. *Journal of Biological Chemistry*, 270(30), pp.17771–17777.

Atmakuri, K., Cascales, E., Burton, O.T., Banta, L.M. and Christie, P.J., 2007. *Agrobacterium* ParA/MinD-like VirC1 spatially coordinates early conjugative DNA transfer reactions. *The EMBO journal*, 26(10), pp.2540–51.

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

Aziz, R.K., Khalifa, M.M. and Sharaf, R.R., 2015. Contaminated water as a source of *Helicobacter pylori* infection: A review. *Journal of Advanced Research*, 6(4), pp.539–547.

Backert, S. and Tegtmeyer, N., 2017. Type IV Secretion and Signal Transduction of *Helicobacter pylori* CagA through Interactions with Host Cell Receptors. *Toxins*, 9(4).

Backert, S., Tegtmeyer, N. and Fischer, W., 2015. Composition, structure and function of the *Helicobacter pylori cag* pathogenicity island encoded type IV secretion system. *Future microbiology*, 10(6), pp.955–65.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A. V, Sirotkin, A. V, Vyahhi,

N., Tesler, G., Alekseyev, M.A. and Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, 19(5), pp.455–77.

Bardhan, P.K., 1997. Epidemiological Features of *Helicobacter pylori* Infection in Developing Countries. *Clinical Infectious Diseases*, 25(5), pp.973–978.

Barrozo, R.M., Hansen, L.M., Lam, A.M., Skoog, E.C., Martin, M.E., Cai, L.P., Lin, Y., Latoscha, A., Suerbaum, S., Canfield, D.R., Solnick, J.V.. *CagY* Is an Immune-Sensitive Regulator of the Helicobacter pylori Type IV Secretion System. Gastroenterology. 2016 Dec;151(6):1164-1175.e3. doi: 10.1053/j.gastro.2016.08.014. Epub 2016 Aug 26. PMID: 27569724; PMCID: PMC5124400.

Bauwens, E., Joosten, M., Taganna, J., Rossi, M., Debraekeleer, A., Tay, A., Peters, F., Backert, S., Fox, J., Ducatelle, R., Remaut, H., Haesebrouck, F. and Smet, A., 2018. In silico proteomic and phylogenetic analysis of the outer membrane protein repertoire of gastric *Helicobacter* species. *Scientific Reports*, 8(1), p.15453.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W., 2009. GenBank. *Nucleic Acids Research*, 37(Database), pp.D26–D31.

Bergman, M., Del Prete, G., van Kooyk, Y. and Appelmelk, B., 2006. *Helicobacter pylori* phase variation, immune modulation and gastric autoimmunity. *Nature Reviews Microbiology*, 4(2), pp.151–159.

Berthenet, E., Yahara, K., Thorell, K., Pascoe, B., Meric, G., Mikhail, J.M., Engstrand, L., Enroth, H., Burette, A., Megraud, F., Varon, C., Atherton, J.C., Smith, S., Wilkinson, T.S., Hitchings, M.D., Falush, D. and Sheppard, S.K., 2018. A GWAS on *Helicobacter pylori* strains points to genetic variants associated with gastric cancer risk. *BMC Biology*, 16(1), p.84.

Beswick, E.J., Suarez, G. and Reyes, V.E., 2006. *H. pylori* and host interactions that influence pathogenesis. *World journal of gastroenterology*, 12(35), pp.5599–605.

Binh, T.T., Shiota, S., Suzuki, R., Matsuda, M., Trang, T.T.H., Kwon, D.H., Iwatani, S. and Yamaoka, Y., 2014. Discovery of novel mutations for clarithromycin resistance in *Helicobacter pylori* by using next-generation sequencing. *Journal of Antimicrobial Chemotherapy*, 69(7), pp.1796–1803.

Bizzozero, G., 1893. Ueber die schlauchformigen drusen des magendarmkanals und die bezienhungen ihres epithels zu dem oberflachenepithel der schleimhaut. Arch. *Mikr Anat.*, 42(82).

Björkholm, B., Sjölund, M., Falk, P.G., Berg, O.G., Engstrand, L. and Andersson, D.I., 2001. Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori. Proceedings of the National Academy of Sciences of the United States of America*, 98(25), pp.14607–12.

Blatch, G.L. and Lässle, M., 1999. The tetratricopeptide repeat: a structural motif

mediating protein-protein interactions. *BioEssays*, 21(11), pp.932–939.

Boehnke, K.F., Eaton, K.A., Fontaine, C., Brewster, R., Wu, J., Eisenberg, J.N.S., Valdivieso, M., Baker, L.H. and Xi, C., 2017. Reduced infectivity of waterborne viable but nonculturable *Helicobacter pylori* strain SS1 in mice. *Helicobacter*, 22(4), p.e12391.

Bolger, A.M., Lohse, M. and Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15), pp.2114–2120.

Bottcher, G., 1875. No Title. Dorpater medicinische Zeitschrift, 184.

Boyanova, L., Ilieva, J., Gergova, G. and Mitov, I., 2016. Levofloxacin susceptibility testing against *Helicobacter pylori*: evaluation of a modified disk diffusion method compared to E test. *Diagnostic Microbiology and Infectious Disease*, 84(1), pp.55–56.

Braga, L.L.B.C., Batista, M.H.R., de Azevedo, O.G.R., da Silva Costa, K.C., Gomes, A.D., Rocha, G.A. and Queiroz, D.M.M., 2019. oipA "on" status of *Helicobacter pylori* is associated with gastric cancer in North-Eastern Brazil. *BMC Cancer*, 19(1), p.48.

Brandi, G., Biavati, B., Calabrese, C., Granata, M., Nannetti, A., Mattarelli, P., Di Febo, G., Saccoccio, G. and Biasco, G., 2006. Urease-Positive Bacteria Other than *Helicobacter pylori* in Human Gastric Juice and Mucosa. *The American Journal of Gastroenterology*, 101(8), pp.1756–1761.

Breckan, R.K., Paulssen, E.J., Asfeldt, A.M., Kvamme, J.-M., Straume, B. and Florholmen, J., 2016. The All-Age Prevalence of *Helicobacter pylori* Infection and Potential Transmission Routes. A Population-Based Study. *Helicobacter*, 21(6), pp.586–595.

Breurec, S., Guillard, B., Hem, S., Brisse, S., Dieye, F.B., Huerre, M., Oung, C., Raymond, J., Tan, T.S., Thiberge, J.-M., Vong, S., Monchy, D. and Linz, B., 2011. Evolutionary history of *Helicobacter pylori* sequences reflect past human migrations in Southeast Asia. *PloS one*, 6(7), p.e22058.

Brown, D.F.J., Wootton, M. and Howe, R.A., 2016. Antimicrobial susceptibility testing breakpoints and methods from BSAC to EUCAST. *Journal of Antimicrobial Chemotherapy,* 71(1), pp.3–5.

Brussow, H., Canchaya, C. and Hardt, W.-D., 2004. Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. *Microbiology and Molecular Biology Reviews*, 68(3), pp.560–602.

Brynildsrud, O., Bohlin, J., Scheffer, L. and Eldholm, V., 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology*, 17(1), p.238.

Bubendorfer, S., Krebes, J., Yang, I., Hage, E., Schulz, T.F., Bahlawane, C., Didelot, X. and Suerbaum, S., 2016a. Genome-wide analysis of chromosomal import patterns after natural transformation of *Helicobacter pylori*. *Nature Communications*, 7(May).

Bubendorfer, S., Krebes, J., Yang, I., Hage, E., Schulz, T.F., Bahlawane, C., Didelot, X. and Suerbaum, S., 2016b. Genome-wide analysis of chromosomal import patterns after natural transformation of *Helicobacter pylori*. *Nature Communications*, 7(1), p.11995.

Buck, A. and Oliver, J.D., 2010. Survival of spinach-associated *Helicobacter pylori* in the viable but nonculturable state. *Food Control*, 21(8), pp.1150–1154.

Van den Bulck, K., Decostere, A., Gruntar, I., Baele, M., Krt, B., Ducatelle, R. and Haesebrouck, F., 2005. In vitro antimicrobial susceptibility testing of *Helicobacter felis*, *H. bizzozeronii*, and *H. salomonis*. *Antimicrobial agents and chemotherapy*, 49(7), pp.2997–3000.

Bullock, K.K., Shaffer, C.L., Brooks, A.W., Secka, O., Forsyth, M.H., McClain, M.S. and Cover, T.L., 2017. Genetic signatures for *Helicobacter pylori* strains of West African origin. *PLoS ONE*, 12(11), pp.1–17.

De Bustos, A., Cuadrado, A. and Jouve, N., 2016. Sequencing of long stretches of repetitive DNA. *Scientific Reports*, 6(1), p.36665.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L., 2009. BLAST+: architecture and applications. *BMC bioinformatics*, 10, p.421.

Cameron, E.A.B., Powell, K.U., Baldwin, L., Jones, P., Bell, G.D. and Williams, S.G.J., 2004. *Helicobacter pylori*: antibiotic resistance and eradication rates in Suffolk, UK, 1991-2001. *Journal of Medical Microbiology*, 53(6), pp.535–538.

Canejo-Teixeira, R., Oliveira, M., Pissarra, H., Niza, M.M.R.E. and Vilela, C.L., 2014. A mixed population of *Helicobacter pylori*, *Helicobacter bizzozeronii* and "*Helicobacter heilmannii*" in the gastric mucosa of a domestic cat. *Irish Veterinary Journal*, 67(1), p.25.

Cao, Q., Didelot, X., Wu, Z., Li, Z., He, L., Li, Y., Ni, M., You, Y., Lin, X., Li, Z., Gong, Y., Zheng, M., Zhang, M., Liu, J., Wang, W., Bo, X., Falush, D., Wang, S., Zhang, J., 2015. Progressive genomic convergence of two *Helicobacter pylori* strains during mixed infection of a patient with chronic gastritis. *Gut*, 64(4),pp.554–561.

Capella-Gutiérrez, S., Silla-Martínez, J.M. and Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* (Oxford, England), 25(15), pp.1972–3.

Carroll, I.M., Ahmed, N., Beesley, S.M., Khan, A.A., Ghousunnissa, S., Moráin, C.A.., Habibullah, C.M. and Smyth, C.J., 2004. Microevolution between paired antral and paired antrum and corpus *Helicobacter pylori* isolates recovered from individual patients. *Journal of Medical Microbiology*, 53(7), pp.669–677.

Carver, T., Harris, S.R., Berriman, M., Parkhill, J. and McQuillan, J.A., 2012. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* (Oxford, England), 28(4), pp.464–9.

Castillo, A.R., Woodruff, A.J., Connolly, L.E., Sause, W.E. and Ottemann, K.M., 2008. Recombination-based in vivo expression technology identifies *Helicobacter pylori* genes important for host colonization. *Infection and immunity*, 76(12), pp.5632–44.

Censini, S., Lange, C., Xiang, Z., Crabtree, J.E., Ghiara, P., Borodovsky, M., Rappuoli, R. and Covacci, A., 1996. cag, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proceedings of the National Academy of Sciences*, 93(25), pp.14648–14653.

Cerda, O., Rivas, A. and Toledo, H., 2003. *Helicobacter pylori* strain ATCC700392 encodes a methyl-accepting chemotaxis receptor protein (MCP) for arginine and sodium bicarbonate. *FEMS Microbiology Letters*, 224(2), pp.175–181.

Cerda, O.A., Núñez-Villena, F., Soto, S.E., Ugalde, J.M., López-Solís, R. and Toledo, H., 2011. *tlpA* gene expression is required for arginine and bicarbonate chemotaxis in *Helicobacter pylori*. *Biological research*, 44(3), pp.277–82.

Cerveny, L., Straskova, A., Dankova, V., Hartlova, A., Ceckova, M., Staud, F. and Stulik, J., 2013. Tetratricopeptide Repeat Motifs in the World of Bacterial Pathogens: Role in Virulence Mechanisms. *Infection and Immunity*, 81(3), p.629.

Chang, P.-C., Wang, C.-J., You, C.-K. and Kao, M.-C., 2011. Effects of a HP0859 (*rfaD*) knockout mutation on lipopolysaccharide structure of *Helicobacter pylori* 26695 and the bacterial adhesion on AGS cells. *Biochemical and Biophysical Research Communications*, 405(3), pp.497–502.

Chang, W.-L., Yeh, Y.-C. and Sheu, B.-S., 2018. The impacts of *H. pylori* virulence factors on the development of gastroduodenal diseases. *Journal of biomedical science*, 25(1), p.68.

Chauhan, N., Tay, A.C.Y., Marshall, B.J. and Jain, U., 2019. *Helicobacter pylori* VacA, a distinct toxin exerts diverse functionalities in numerous cells: An overview. *Helicobacter*, 24(1), p.e12544.

Chen, T.S., Chang, F.Y. and Lee, S.D., 1995. A biopsy urease test in the detection of *Helicobacter pylori*: comparison of antral and body specimens. Zhonghua yi xue za zhi = *Chinese medical journal*; Free China ed, 55(5), pp.361–5.

Chen, X., Xia, C., Li, Q., Jin, L., Zheng, L. and Wu, Z., 2018. Comparisons Between Bacterial Communities in Mucosa in Patients With Gastric Antrum Ulcer and a Duodenal Ulcer. *Frontiers in Cellular and Infection Microbiology*, 8, p.126.

Cheng, H., Hu, F., Zhang, L., Yang, G., Ma, J., Hu, J., Wang, W., Gao, W. and Dong, X., 2009. Prevalence of *Helicobacter pylori* Infection and Identification of Risk Factors in Rural and Urban Beijing, China. *Helicobacter*, 14(2), pp.128–133.

Chisholm, S.A. and Owen, R.J., 2009. Frequency and molecular characteristics of ciprofloxacin- and rifampicin-resistant *Helicobacter pylori* from gastric infections in the UK. *Journal of Medical Microbiology*, 58(10), pp.1322–1328.

Choi, J.M., Kim, S.G., Choi, J., Park, J.Y., Oh, S., Yang, H.-J., Lim, J.H., Im, J.P., Kim, J.S. and Jung, H.C., 2018. Effects of *Helicobacter pylori* eradication for metachronous gastric cancer prevention: a randomized controlled trial. *Gastrointestinal Endoscopy*, 88(3), pp.475-485.e2.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X. and Ruden, D.M., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, 6(2), pp.80–92.

Collard, F., Collet, J.-F., Gerin, I., Veiga-da-Cunha, M. and Van Schaftingen, E., 1999. Identification of the cDNA encoding human 6-phosphogluconolactonase, the enzyme catalyzing the second step of the pentose phosphate pathway. *FEBS Letters*, 459(2), pp.223–226.

Conde-Álvarez, R., Arce-Gorvel, V., Gil-Ramírez, Y., Iriarte, M., Grilló, M.-J., Gorvel, J.P. and Moriyón, I., 2013. Lipopolysaccharide as a target for brucellosis vaccine design. *Microbial Pathogenesis*, 58, pp.29–34.

Correa, P., 1988a. A human model of gastric carcinogenesis. *Cancer research*, 48(13), pp.3554–60.

Correa, P., 1988b. Chronic gastritis: a clinico-pathological classification. *The American journal of gastroenterology*, 83(5), pp.504–9.

Correa, P. and Piazuelo, M.B., 2011. *Helicobacter pylori* Infection and Gastric Adenocarcinoma. *US gastroenterology & hepatology review*, 7(1), pp.59–64.

Correa, P. and Piazuelo, M.B., 2012. The gastric precancerous cascade. *Journal of Digestive Diseases*, 13(1), pp.2–9.

Correa, P., Piazuelo, M.B. and Camargo, M.C., 2004. The future of gastric cancer prevention. *Gastric Cancer*, 7(1), pp.9–16.

Cortes, M.C.C., Yamakawa, A., Casingal, C.R., Fajardo, L.S.N., Juan, M.L.G., De Guzman, B.B., Bondoc, E.M., Mahachai, V., Yamazaki, Y., Yoshida, M., Kutsumi, H., Natividad, F.F., Azuma, T. and Azuma, T., 2010. Diversity of the *cagA* gene of *Helicobacter pylori* strains from patients with gastroduodenal diseases in the Philippines. *FEMS Immunology & Medical Microbiology*, 60(1), pp.90–97.

da Costa, D.M., Pereira, E. dos S. and Rabenhorst, S.H.B., 2015. What exists beyond *cagA* and *vacA*? *Helicobacter pylor*i genes in gastric diseases. *World journal of gastroenterology*, 21(37), pp.10563–72.

Cover, T.L. and Blaser, M.J., 1992. Purification and characterization of the vacuolating toxin from *Helicobacter pylori*. *The Journal of biological chemistry*, 267(15), pp.10570–5.

Cover, T.L., Krishna, U.S., Israel, D.A. and Peek, R.M., 2003. Induction of gastric

epithelial cell apoptosis by *Helicobacter pylori* vacuolating cytotoxin. *Cancer research*, 63(5), pp.951–7.

Croucher, N.J., Page, A.J., Connor, T.R., Delaney, A.J., Keane, J.A., Bentley, S.D., Parkhill, J. and Harris, S.R., 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research*, 43(3), pp.e15–e15.

Croxen, M.A., Sisson, G., Melano, R. and Hoffman, P.S., 2006. The *Helicobacter pylori* chemotaxis receptor TlpB (HP0103) is required for pH taxis and for colonization of the gastric mucosa. *Journal of bacteriology*, 188(7), pp.2656–65.

Cullen, K.P., Broderick, B.M., Jayaram, J., Flynn, B. and O'Connor, H.J., 2002. Evaluation of the *Helicobacter pylori* stool antigen (HpSA) test in routine clinical practice--is it patient-friendly? *Irish medical journal*, 95(10), pp.305–6.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., McVean, G., Durbin, R. and 1000 Genomes Project Analysis Group, 1000 Genomes Project Analysis, 2011. The variant call format and VCFtools. *Bioinformatics* (Oxford, England), 27(15), pp.2156–8.

Dang, B.N. and Graham, D.Y., 2017. *Helicobacter pylori* infection and antibiotic resistance: a WHO high priority? *Nature Reviews Gastroenterology & Hepatology*, 14(7), pp.383–384.

Dankova, V., Balonova, L., Straskova, A., Spidlova, P., Putzova, D., Kijek, T., Bozue, J., Cote, C., Mou, S., Worsham, P., Szotakova, B., Cerveny, L. and Stulik, J., 2014. Characterization of tetratricopeptide repeat-like proteins in *Francisella tularensis* and identification of a novel locus required for virulence. *Infection and immunity*, 82(12), pp.5035–48.

Darling, A.C.E., Mau, B., Blattner, F.R. and Perna, N.T., 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome research*, 14(7), pp.1394–403.

Das, A.K., Cohen, P.W. and Barford, D., 1998. The structure of the tetratricopeptide repeats of protein phosphatase 5: implications for TPR-mediated protein-protein interactions. *The EMBO Journal*, 17(5), pp.1192–1199.

Day, A.S., Jones, N.L., Lynett, J.T., Jennings, H.A., Fallone, C.A., Beech, R. and Sherman, P.M., 2000. *cagE* Is a Virulence Factor Associated with *Helicobacter pylori* –Induced Duodenal Ulceration in Children. *The Journal of Infectious Diseases*, 181(4), pp.1370–1375.

Debowski, A.W., Carnoy, C., Verbrugghe, P., Nilsson, H.-O., Gauntlett, J.C., Fulurija, A., Camilleri, T., Berg, D.E., Marshall, B.J. and Benghezal, M., 2012. Xer recombinase and genome integrity in *Helicobacter pylori*, a pathogen without topoisomerase IV. P*loS one*, 7(4), p.e33310.

Deloger, M., El Karoui, M. and Petit, M.-A., 2009. A Genomic Distance Based on MUM Indicates Discontinuity between Most Bacterial Species and Genera. *Journal of Bacteriology*, 191(1), pp.91–99.

Didelot, X., Nell, S., Yang, I., Woltemate, S., van der Merwe, S. and Suerbaum, S., 2013. Genomic evolution and transmission of *Helicobacter pylori* in two South African families. *Proceedings of the National Academy of Sciences*, 110(34), pp.13880–13885.

Didelot, X., Walker, A.S., Peto, T.E., Crook, D.W. and Wilson, D.J., 2016. Within-host evolution of bacterial pathogens. *Nature Reviews Microbiology*, 14(3), pp.150–162.

Dixon, M.F., Genta, R.M., Yardley, J.H. and Correa, P., 1996. Classification and grading of gastritis. The updated Sydney System. International Workshop on the Histopathology of Gastritis, Houston 1994. *The American journal of surgical pathology*, 20(10), pp.1161–81.

Doenges, J., 1938. Spirochetes in Gastric Glands of Macacus rhesus and Humans without Definite History of Related Disease. *Experimental Biology and Medicine*, 38(4), pp.536–538.

Van Doorn, L.-J., Figueiredo, C.U., Sanna, R., Pena, S., Midolo, P., Ng, E.K.W., Atherton, J.C., Blaser, M.J. and Quint, W.G. V, 1998. Expanding Allelic Diversity of *Helicobacter pylori vacA*. *Journal of clinical microbiology,* 36, pp.2597-2603.

Dorer, M.S., Sessler, T.H. and Salama, N.R., 2011. Recombination and DNA repair in *Helicobacter pylori. Annual review of microbiology*, 65, pp.329–48.

Dorrell, N., Martino, M.C., Stabler, R.A., Ward, S.J., Zhang, Z.W., McColm, A.A., Farthing, M.J.G. and Wren, B.W., 1999. Characterization of *Helicobacter pylori* PldA, a phospholipase with a role in colonization of the gastric mucosa. *Gastroenterology*, 117(5), pp.1098–1104.

Dubois, A., Berg, D.E., Incecik, E.T., Fiala, N., Heman-Ackah, L.M., Del Valle, J., Yang, M., Wirth, H.P., Perez-Perez, G.I. and Blaser, M.J., 1999. Host specificity of *Helicobacter pylori* strains and host responses in experimentally challenged nonhuman primates. *Gastroenterology*, 116(1), pp.90–6.

Duncan, S.S., Valk, P.L., Shaffer, C.L., Bordenstein, S.R. and Cover, T.L., 2012. J-Western Forms of Helicobacter pylori *cagA* Constitute a Distinct Phylogenetic Group with a Widespread Geographic Distribution. *Journal of Bacteriology*, 194(6), pp.1593–1604.

Eaton, K.A., Morgan, D.R. and Krakowka, S., 1992. Motility as a factor in the colonisation of gnotobiotic piglets by *Helicobacter pylori*. *Journal of Medical Microbiology*, 37(2), pp.123–127.

Eaton, K.A., Suerbaum, S., Josenhans, C. and Krakowka, S., 1996. Colonization of gnotobiotic piglets by *Helicobacter pylori* deficient in two flagellin genes. *Infection and immunity*, 64(7), pp.2445–8.

Edwards, N.J., Monteiro, M.A., Faller, G., Walsh, E.J., Moran, A.P., Roberts, I.S. and High, N.J., 2002. Lewis X structures in the O antigen side-chain promote adhesion of *Helicobacter pylori* to the gastric epithelium. *Molecular Microbiology*, 35(6), pp.1530–1539.

Eriani, G., Delarue, M., Poch, O., Gangloff, J. and Moras, D., 1990. Partition of tRNA synthetases into two classes based on mutually exclusive sets of sequence motifs. *Nature*, 347(6289), pp.203–206.

EUCAST, 2019. The European Committee on Antimicrobial Susceptibility Testing. Breakpoint tables for interpretation of MICs and zone diameters. Version 9.0, 2019.

Evans, S.N., Hower, V. and Pachter, L., 2010. Coverage statistics for sequence census methods. *BMC Bioinformatics*, 11(1), p.430.

Falush, D., Kraft, C., Taylor, N.S., Correa, P., Fox, J.G., Achtman, M. and Suerbaum, S., 2001a. Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26), pp.15056–61.

Falush, D., Kraft, C., Taylor, N.S., Correa, P., Fox, J.G., Achtman, M. and Suerbaum, S., 2001b. Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock rates, recombination size, and minimal age. *Proceedings of the National Academy of Sciences of the United States of America*, 98(26), pp.15056–61.

Falush, D., Stephens, M. and Pritchard, J.K., 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4), pp.1567–87.

Falush, D., Wirth, T., Linz, B., Pritchard, J.K., Stephens, M., Kidd, M., Blaser, M.J., Graham, D.Y., Vacher, S., Perez-Perez, G.I., Yamaoka, Y., Mégraud, F., Otto, K., Reichard, U., Katzowitsch, E., Wang, X., Achtman, M. and Suerbaum, S., 2003. Traces of Human Migrations in *Helicobacter pylori* Populations. *Science*, 299(5612), pp.1582–1585.

Farinha, P. and Gascoyne, R.D., 2005. *Helicobacter pylori* and MALT Lymphoma. *Gastroenterology*, 128(6), pp.1579–1605.

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D.M., Piñeros, M., Znaor, A. and Bray, F., 2018. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *International Journal of Cancer*, 144(8), p.ijc.31937.

Ferwana, M., Abdulmajeed, I., Alhajiahmed, A., Madani, W., Firwana, B., Hasan, R., Altayar, O., Limburg, P.J., Murad, M.H. and Knawy, B., 2015. Accuracy of urea breath test in *Helicobacter pylori* infection: meta-analysis. *World journal of gastroenterology*, 21(4), pp.1305–14.

Fiorini, G., Saracino, I.M., Zullo, A., Gatta, L., Pavoni, M. and Vaira, D., 2017. Rescue therapy with bismuth quadruple regimen in patients with *Helicobacter pylori* -resistant strains. *Helicobacter*, 22(6), p.e12448.

Foynes, S., Dorrell, N., Ward, S.J., Stabler, R.A., McColm, A.A., Rycroft, A.N. and Wren, B.W., 2000. *Helicobacter pylori* possesses two CheY response regulators and a histidine kinase sensor, CheA, which are essential for chemotaxis and colonization of the gastric mucosa. *Infection and immunity*, 68(4), pp.2016–23.

Foynes, S., Dorrell, N., Ward, S.J., Zhang, Z.W., McColm, A.A., Farthing, M.J.. and Wren, B.W., 1999. Functional analysis of the roles of FliQ and FlhB in flagellar expression in *Helicobacter pylori. FEMS Microbiology Letters*, 174(1), pp.33–39.

Franceschi, F., Covino, M. and Roubaud Baudron, C., 2019. Review: *Helicobacter pylori* and extragastric diseases. *Helicobacter*, 24(S1).

Freedberg, A.S. and Barron, L.E., 1940. The presence of spirochetes in human gastric mucosa. *American Journal of Digestive Diseases*, 7(10), pp.443–445.

Fung, C., Tan, S., Nakajima, M., Skoog, E.C., Camarillo-Guerrero, L.F., Klein, J.A., Lawley, T.D., Solnick, J. V., Fukami, T. and Amieva, M.R., 2019. High-resolution mapping reveals that microniches in the gastric glands control *Helicobacter pylori* colonization of the stomach. *PLOS Biology*, 17(5), p.e3000231.

Furuta, Y., Konno, M., Osaki, T., Yonezawa, H., Ishige, T., Imai, M., Shiwa, Y., Shibata-Hatta, M., Kanesaki, Y., Yoshikawa, H., Kamiya, S. and Kobayashi, I., 2015a. Microevolution of virulence-related genes in *Helicobacter pylori* familial infection. *PLoS ONE*, 10(5), pp.1–17.

Furuta, Y., Namba-Fukuyo, H., Shibata, T.F., Nishiyama, T., Shigenobu, S., Suzuki, Y., Sugano, S., Hasebe, M. and Kobayashi, I., 2014. Methylome Diversification through Changes in DNA Methyltransferase Sequence Specificity. *PLoS Genetics*, 10(4), p.e1004272.

Gaillard, M.E., Bottero, D., Castuma, C.E., Basile, L.A. and Hozbor, D., 2011. Laboratory adaptation of *Bordetella pertussis* is associated with the loss of type three secretion system functionality. *Infection and immunity*, 79(9), pp.3677–82.

Garrison, E. and Marth, G., 2012. Haplotype-based variant detection from short-read sequencing. [arXiv:1207.3907v2].

Gauntlett, J.C., Nilsson, H.-O., Fulurija, A., Marshall, B.J. and Benghezal, M., 2014. Phase-variable restriction/modification systems are required for *Helicobacter pylori* colonization. *Gut pathogens*, 6, 35.

Gebert, B., Fischer, W., Weiss, E., Hoffmann, R. and Haas, R., 2003. *Helicobacter pylori* Vacuolating Cytotoxin Inhibits T Lymphocyte Activation. *Science*, 301(5636), pp.1099–1102.

Gisbert, J.P., Arata, I.G., Boixeda, D., Barba, M., Cantón, R., Plaza, A.G. and Pajares, J.M., 2002. Role of partner's infection in reinfection after *Helicobacter pylori* eradication. *European journal of gastroenterology & hepatology*, 14(8), pp.865–71.

Gisbert, J.P. and Calvet, X., 2013. *Helicobacter pylori* "Test-and-Treat" Strategy for Management of Dyspepsia: A Comprehensive Review. *Clinical and translational gastroenterology*, 4(3), p.e32.

Gisbert, J.P. and Pajares, J.M., 2004. Stool Antigen Test for the Diagnosis of *Helicobacter pylori* Infection: a Systematic Review. *Helicobacter*, 9(4), pp.347–368.

Glocker, E., Bogdan, C. and Kist, M., 2007. Characterization of rifampicin-resistant clinical Helicobacter pylori isolates from Germany. *Journal of Antimicrobial Chemotherapy*, 59(5), pp.874–879.

Goodman, K.J. and Correa, P., 2000. Transmission of *Helicobacter pylori* among siblings. *The Lancet*, 355(9201), pp.358–362.

Goodman, K.J., Correa, P., Aux, H.J.T., Ramirez, H., DeLany, J.P., Pepinosa, O.G., Quinones, M.L. and Parra, T.C., 1996. *Helicobacter pylori* Infection in the Colombian Andes: A Population-based Study of Transmission Pathways. A*merican Journal of Epidemiology*, 144(3), pp.290–299.

Gorrell, R. and Kwok, T., 2017. The *Helicobacter pylori* Methylome: Roles in Gene Regulation and Virulence. *Springer, Cham*, pp.105–127.

Graham, D.Y., Opekun, A.R., Hammoud, F., Yamaoka, Y., Reddy, R., Osato, M.S. and El-Zimaity, H.M.T., 2003. Studies regarding the mechanism of false negative urea breath tests with proton pump inhibitors. *The American Journal of Gastroenterology*, 98(5), pp.1005–1009.

Gravina, A.G., Zagari, R.M., De Musis, C., Romano, L., Loguercio, C. and Romano, M., 2018. *Helicobacter pylori* and extragastric diseases: A review. *World journal of gastroenterology*, 24(29), pp.3204–3221.

Gu, H., 2017. Role of Flagella in the Pathogenesis of *Helicobacter pylori*. *Current microbiology*, 74(7), pp.863–869.

Gunaletchumy, S.P., Seevasant, I., Tan, M.H., Croft, L.J., Mitchell, H.M., Goh, K.L., Loke, M.F. and Vadivelu, J., 2015. *Helicobacter pylori* Genetic Diversity and Gastro-duodenal Diseases in Malaysia. *Scientific Reports*, 4(1), p.7431.

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), pp.1072–1075.

Hadfield, J., Croucher, N.J., Goater, R.J., Abudahab, K., Aanensen, D.M. and Harris, S.R., 2018. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*, 34(2), pp.292–293.

Han, K.-D., Ahn, D.-H., Lee, S.-A., Min, Y.-H., Kwon, A.-R., Ahn, H.-C. and Lee, B.-J., 2013. Identification of chromosomal HP0892-HP0893 toxin-antitoxin proteins in *Helicobacter pylori* and structural elucidation of their protein-protein interaction. *The Journal of biological chemistry*, 288(8), pp.6004–13.

Han, K.-D., Matsuura, A., Ahn, H.-C., Kwon, A.-R., Min, Y.-H., Park, H.-J., Won, H.-S., Park, S.-J., Kim, D.-Y. and Lee, B.-J., 2011. Functional Identification of Toxin-Antitoxin Molecules from *Helicobacter pylori* 26695 and Structural Elucidation of the Molecular Interactions. *Journal of Biological Chemistry*, 286(6), pp.4842–4853.

Han, S.R., Schreiber, H.J., Bhakdi, S., Loos, M. and Maeurer, M.J., 1998. vacA genotypes and genetic diversity in clinical isolates of *Helicobacter pylori*. *Clinical and diagnostic laboratory immunology*, 5(2), pp.139–45.

Hänninen, M. -l., Happonen, I., Saari, S. and Jalava, K., 1996. Culture and Characteristics of *Helicobacter bizzozeronii*, a New Canine Gastric *Helicobacter sp*. *International Journal of Systematic Bacteriology*, 46(1), pp.160–166.

Hansen, L.M., Gideonsson, P., Canfield, D.R., Borén, T. and Solnick, J. V., 2017. Dynamic Expression of the BabA Adhesin and Its BabB Paralog during *Helicobacter pylori* Infection in Rhesus Macaques. *Infection and Immunity*, 85(6).

Harper, D., Parracho, H., Walker, J., Sharp, R., Hughes, G., Werthén, M., Lehman, S. and Morales, S., 2014. Bacteriophages and Biofilms. *Antibiotics*, 3(3), pp.270–284.

Hatakeyama, M., 2014. *Helicobacter pylori* CagA and gastric cancer: a paradigm for hit-and-run carcinogenesis. *Cell host & microbe*, 15(3), pp.306–16.

Hatakeyama, M., 2019. Malignant *Helicobacter pylori*-Associated Diseases: Gastric Cancer and MALT Lymphoma. *Springer, New York, NY*, pp.1–15.

Hays, C., Burucoa, C., Lehours, P., Tran, C.T., Leleu, A. and Raymond, J., 2018. Molecular characterization of *Helicobacter pylori* resistance to rifamycins. *Helicobacter*, 23(1), p.e12451.

Heep, M., Odenbreit, S., Beck, D., Decker, J., Prohaska, E., Rieger, U. and Lehn, N., 2000a. Mutations at four distinct regions of the rpoB gene can reduce the susceptibility of *Helicobacter pylori* to rifamycins. *Antimicrobial agents and chemotherapy*, 44(6), pp.1713–5.

Heep, M., Rieger, U., Beck, D. and Lehn, N., 2000b. Mutations in the beginning of the rpoB gene can induce resistance to rifamycins in both *Helicobacter pylori* and *Mycobacterium tuberculosis*. *Antimicrobial agents and chemotherapy*, 44(4), pp.1075–7.

Hoffman, P.S., 1999. Antibiotic Resistance Mechanisms of *Helicobacter pylori*. *Canadian Journal of Gastroenterology*, 13(3), pp.243–249.

Hooi, J.K.Y., Lai, W.Y., Ng, W.K., Suen, M.M.Y., Underwood, F.E., Tanyingoh, D.,

Malfertheiner, P., Graham, D.Y., Wong, V.W.S., Wu, J.C.Y., Chan, F.K.L., Sung, J.J.Y., Kaplan, G.G. and Ng, S.C., 2017. Global Prevalence of *Helicobacter pylori* Infection: Systematic Review and Meta-Analysis. *Gastroenterology*, 153(2), pp.420–429.

Howitt, M.R., Lee, J.Y., Lertsethtakarn, P., Vogelmann, R., Joubert, L.-M., Ottemann, K.M. and Amieva, M.R., 2011. ChePep Controls *Helicobacter pylori* Infection of the Gastric Glands and Chemotaxis in the Epsilonproteobacteria. *mBio*, 2(4).

Hu, Y., Guerrero, E., Keniry, M., Manrrique, J. and Bullard, J.M., 2015. Identification of Chemical Compounds That Inhibit the Function of Glutamyl-tRNA Synthetase from *Pseudomonas aeruginosa*. *Journal of Biomolecular Screening*, 20(9), pp.1160–1170.

Hua, J.-S., Zheng, P.-Y., Fong, T.K., Khin, M.M. and Bow, H., 1998. *Helicobacter pylori* acquistion of metronidazole resistance by natural transformation in vitro. World Journal of *Gastroenterology*, 4(5), p.385.

Huang, J.Y., Goers Sweeney, E., Guillemin, K. and Amieva, M.R., 2017. Multiple Acid Sensors Control *Helicobacter pylori* Colonization of the Stomach. *PLoS pathogens*, 13(1), p.e1006118.

Huang, Y., Wang, Q.-L., Cheng, D.-D., Xu, W.-T. and Lu, N.-H., 2016a. Adhesion and Invasion of Gastric Mucosa Epithelial Cells by *Helicobacter pylori*. *Frontiers in cellular and infection microbiology*, 6, p.159.

Huang, Y., Wang, Q., Cheng, D., Xu, W. and Lu, N., 2016b. Adhesion and Invasion of Gastric Mucosa Epithelial Cells by *Helicobacter pylori*. *Frontiers in Cellular and Infection Microbiology*, 6(November).

Hurdle, J.G., O'Neill, A.J. and Chopra, I., 2005. Prospects for aminoacyl-tRNA synthetase inhibitors as new antimicrobial agents. *Antimicrobial agents and chemotherapy*, 49(12), pp.4821–33.

Husmeier, D., 2005. Discriminating between rate heterogeneity and interspecific recombination in DNA sequence alignments with phylogenetic factorial hidden Markov models. *Bioinformatics*, 21(Suppl 2), pp.ii166–ii172.

Ishaq, S. and Nunn, L., 2015. *Helicobacter pylori* and gastric cancer: a state of the art review. *Gastroenterology and hepatology from bed to bench*, 8(Suppl 1), pp.S6–S14.

Israel, D.A., Salama, N., Krishna, U., Rieger, U.M., Atherton, J.C., Falkow, S., Peek, R.M. and Jr., 2001. *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proceedings of the National Academy of Sciences of the United States of America*, 98(25), pp.14625–30.

Israeli, E., Ilan, Y., Meir, S.B., Buenavida, C. and Goldin, E., 2003. A novel 13C-urea breath test device for the diagnosis of *Helicobacter pylori* infection: continuous online measurements allow for faster test results with high accuracy. *Journal of clinical gastroenterology*, 37(2), pp.139–41.

Ito, S., 1967. Section 6, Alimentary canal. In: Anatomic structure of the gastric mucosa. Handbook of physiology. *American Physiological Society*, pp.705–758.

Jenks, P.J. and Edwards, D.I., 2002. Metronidazole resistance in *Helicobacter pylori*. *International Journal of Antimicrobial Agents*, 19(1), pp.1–7.

Ji, X., Fernandez, T., Burroni, D., Pagliaccia, C., Atherton, J.C., Reyrat, J.M., Rappuoli, R. and Telford, J.L., 2000. Cell specificity of *Helicobacter pylori* cytotoxin is determined by a short region in the polymorphic midregion. *Infection and immunity*, 68(6), pp.3754–7.

Johnson, K.S. and Ottemann, K.M., 2018. Colonization, localization, and inflammation: the roles of *H. pylori* chemotaxis in vivo. *Current Opinion in Microbiology*, 41, pp.51–57.

de Jong, J.J., Lantinga, M.A. and Drenth, J.P., 2019. Prevention of overuse: A view on upper gastrointestinal endoscopy. *World journal of gastroenterology*, 25(2), pp.178–189.

Joshi, N. and Fass, J., 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. (Version 1 Available at https://github.com/najoshi/sickle).

Jung, J.H., Cho, I.K., Lee, C.H., Song, G.G. and Lim, J.H., 2018. Clinical Outcomes of Standard Triple Therapy Plus Probiotics or Concomitant Therapy for *Helicobacter pylori* Infection. *Gut and Liver*, 12(2), pp.165–172.

Kao, C.-Y., Lee, A.-Y., Huang, A.-H., Song, P.-Y., Yang, Y.-J., Sheu, S.-M., Chang, W.-L., Sheu, B.-S. and Wu, J.-J., 2014. Heteroresistance of *Helicobacter pylori* from the same patient prior to antibiotic treatment. *Infection, Genetics and Evolution*, 23, pp.196–202.

Kao, C.-Y., Sheu, B.-S., Sheu, S.-M., Yang, H.-B., Chang, W.-L., Cheng, H.-C. and Wu, J.-J., 2012. Higher Motility Enhances Bacterial Density and Inflammatory Response in Dyspeptic Patients Infected with *Helicobacter pylori. Helicobacter*, 17(6), pp.411–416.

Kao, C.-Y., Sheu, B.-S. and Wu, J.-J., 2016. *Helicobacter pylori* infection: An overview of bacterial virulence factors and pathogenesis. *Biomedical journal*, 39(1), pp.14–23.

Karkhah, A., Ebrahimpour, S., Rostamtabar, M., Koppolu, V., Darvish, S., Vasigala, V.K.R., Validi, M. and Nouri, H.R., 2019a. *Helicobacter pylori* evasion strategies of the host innate and adaptive immune responses to survive and develop gastrointestinal diseases. *Microbiological Research*, 218(October 2018), pp.49–57.

Karkhah, A., Ebrahimpour, S., Rostamtabar, M., Koppolu, V., Darvish, S., Vasigala, V.K.R., Validi, M. and Nouri, H.R., 2019b. *Helicobacter pylori* evasion strategies of the host innate and adaptive immune responses to survive and develop gastrointestinal diseases. *Microbiological Research*, 218, pp.49–57.

Kasai, K. and Kobayashi, R., 1919. The Stomach Spirochete Occurring in Mammals. *The Journal of Parasitology*, 6(1), p.1.

Kavermann, H., Burns, B.P., Angermuller, K., Odenbreit, S., Fischer, W., Melchers, K. and Haas, R., 2003. Identification and characterization of *Helicobacter pylori* genes essential for gastric colonization. *The Journal of experimental medicine*, 197(7), pp.813–22.

Kayali, S., Manfredi, M., Gaiani, F., Bianchi, L., Bizzarri, B., Leandro, G., Di Mario, F. and De' Angelis, G.L., 2018. *Helicobacter pylori*, transmission routes and recurrence of infection: state of the art. *Acta bio-medica : Atenei Parmensis*, 89(8-S), pp.72–76.

Kazemi, S., Tavakkoli, H., Habizadeh, M.R. and Emami, M.H., 2011. Diagnostic values of *Helicobacter pylori* diagnostic tests: stool antigen test, urea breath test, rapid urease test, serology and histology. *Journal of research in medical sciences : the official journal of Isfahan University of Medical Sciences*, 16(9), pp.1097–104.

Keane, J.A., Page, A.J., Delaney, A.J., Taylor, B., Seemann, T., Harris, S.R. and Soares, J., 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics*, 2(4).

Keilberg, D. and Ottemann, K.M., 2016. How *Helicobacter pylori* senses, targets and interacts with the gastric epithelium. *Environmental Microbiology*, 18(3), pp.791–806.

Kennemann, L., Didelot, X., Aebischer, T., Kuhn, S., Drescher, B., Droege, M., Reinhardt, R., Correa, P., Meyer, T.F., Josenhans, C., Falush, D. and Suerbaum, S., 2011. *Helicobacter pylori* genome evolution during human infection. *Proceedings of the National Academy of Sciences*, 108(12), pp.5033–5038.

Khamri, W., Moran, A.P., Worku, M.L., Karim, Q.N., Walker, M.M., Annuk, H., Ferris, J.A., Appelmelk, B.J., Eggleton, P., Reid, K.B.M. and Thursz, M.R., 2005. Variations in *Helicobacter pylori* lipopolysaccharide to evade the innate immune component surfactant protein D. *Infection and immunity*, 73(11), pp.7677–86.

Kidd, M., Lastovica, A.J., Atherton, J.C. and Louw, J.A., 2001. Conservation of the cag pathogenicity island is associated with *vacA* alleles and gastroduodenal disease in South African Helicobacter pylori isolates. *Gut*, 49(1), pp.11–7.

Kidd, M. and Modlin, I.M., 1998. A Century of *Helicobacter pylori*. *Digestion*, 59(1), pp.1–15.

Kim, A., Servetas, S.L., Kang, J., Kim, J., Jang, S., Choi, Y.H., Su, H., Jeon, Y.-E., Hong, Y.A., Yoo, Y.-J., Merrell, D.S. and Cha, J.-H., 2016. *Helicobacter pylori* outer membrane protein, HomC, shows geographic dependent polymorphism that is influenced by the Bab family. *Journal of Microbiology*, 54(12), pp.846–852.

Kim, J.J., Kim, J.G. and Kwon, D.H., 2003. Mixed-Infection of Antibiotic Susceptible and Resistant *Helicobacter pylori* Isolates in a Single Patient and Underestimation of Antimicrobial Susceptibility Testing. *Helicobacter*, 8(3), pp.202–206.

Kim, J.S., Chang, J.H., Chung, S.I. and Yum, J.S., 1999. Molecular cloning and characterization of the *Helicobacter pylori fliD* gene, an essential factor in flagellar

structure and motility. *Journal of bacteriology*, 181(22), pp.6969–76.

Kivi, M., Tindberg, Y., Sörberg, M., Casswall, T.H., Befrits, R., Hellström, P.M., Bengtsson, C., Engstrand, L. and Granström, M., 2003. Concordance of *Helicobacter pylori* strains within families. *Journal of clinical microbiology*, 41(12), pp.5604–8.

Knorr, J., Ricci, V., Hatakeyama, M. and Backert, S., 2019. Classification of *Helicobacter pylori* Virulence Factors: Is CagA a Toxin or Not? *Trends in Microbiology*, 27(9), pp.731–738.

Kodaman, N., Pazos, A., Schneider, B.G., Piazuelo, M.B., Mera, R., Sobota, R.S., Sicinschi, L.A., Shaffer, C.L., Romero-Gallo, J., de Sablet, T., Harder, R.H., Bravo, L.E., Peek, R.M., Wilson, K.T., Cover, T.L., Williams, S.M., Correa, P. and Correa, P., 2014. Human and *Helicobacter pylori* coevolution shapes the risk of gastric disease. *Proceedings of the National Academy of Sciences of the United States of America*, 111(4), pp.1455–60.

Kojima, K.K., Furuta, Y., Yahara, K., Fukuyo, M., Shiwa, Y., Nishiumi, S., Yoshida, M., Azuma, T., Yoshikawa, H. and Kobayashi, I., 2016. Population Evolution of *Helicobacter pylori* through Diversification in DNA Methylation and Interstrain Sequence Homogenization. *Molecular Biology and Evolution*, 33(11), pp.2848–2859.

Kraft, C., Stack, A., Josenhans, C., Niehus, E., Dietrich, G., Correa, P., Fox, J.G., Falush, D. and Suerbaum, S., 2006. Genomic changes during chronic *Helicobacter pylori* infection. *Journal of Bacteriology*, 188(1), pp.249–254.

Krebes, J., Didelot, X., Kennemann, L. and Suerbaum, S., 2014. Bidirectional genomic exchange between *Helicobacter pylori* strains from a family in Coventry, United Kingdom. *International Journal of Medical Microbiology*, 304(8), pp.1135–1146.

Krienitz, W., 1906. Ueber das Auftreten von Spirochäten verschiedener Form im Mageninhalt bei Carcinoma ventriculi. *Dtsch med Wochenschr*, 32(22), p.872.

Krishna, U., Romero-Gallo, J., Suarez, G., Azah, A., Krezel, A.M., Varga, M.G., Forsyth, M.H., Peek, R.M. and Jr, 2016. Genetic Evolution of a *Helicobacter pylori* Acid-Sensing Histidine Kinase and Gastric Disease. *The Journal of infectious diseases*, 214(4), pp.644–8.

Kulick, S., Moccia, C., Didelot, X., Falush, D., Kraft, C. and Suerbaum, S., 2008. Mosaic DNA imports with interspersions of recipient sequence after natural transformation of *Helicobacter pylori*. *PLoS ONE*, 3(11).

Kumar, N., Mukhopadhyay, A.K., Patra, R., De, R., Baddam, R., Shaik, S., Alam, J., Tiruvayipati, S. and Ahmed, N., 2012. Next-generation sequencing and *de novo* assembly, genome organization, and comparative genomic analyses of the genomes of two *Helicobacter pylori* isolates from duodenal ulcer patients in India. *Journal of bacteriology*, 194(21), pp.5963–4.

Kumar, S., Karmakar, B.C., Nagarajan, D., Mukhopadhyay, A.K., Morgan, R.D. and

Rao, D.N., 2018. N4-cytosine DNA methylation regulates transcription and pathogenesis in *Helicobacter pylori*. *Nucleic Acids Research*, 46(7), pp.3429–3445.

Kupcinskas, J. and Hold, G.L., 2018. Other *Helicobacters* and the gastric microbiome. *Helicobacter*, 23, p.e12521.

Kusters, J.G., van Vliet, A.H.M. and Kuipers, E.J., 2006. Pathogenesis of *Helicobacter pylori* Infection. *Clinical Microbiology Reviews*, 19(3), pp.449–490.

Kyrillos, A., Arora, G., Murray, B. and Rosenwald, A.G., 2016. The Presence of Phage Orthologous Genes in *Helicobacter pylori* Correlates with the Presence of the Virulence Factors CagA and VacA. *Helicobacter*, 21(3), pp.226–233.

Lander, E.S. and Waterman, M.S., 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3), pp.231–9.

Lang, L. and García, F., 2004. Comparison of E-test and disk diffusion assay to evaluate resistance of *Helicobacter pylori* isolates to amoxicillin, clarithromycin, metronidazole and tetracycline in Costa Rica. *International Journal of Antimicrobial Agents*, 24(6), pp.572–577.

Langmead, B. and Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4), pp.357–359.

Larsen, M. V, Cosentino, S., Rasmussen, S., Friis, C., Hasman, H., Marvig, R.L., Jelsbak, L., Sicheritz-Pontén, T., Ussery, D.W., Aarestrup, F.M. and Lund, O., 2012. Multilocus sequence typing of total-genome-sequenced bacteria. *Journal of clinical microbiology*, 50(4), pp.1355–61.

Lassen, A.T., Hallas, J. and Schaffalitzky de Muckadell, O.B., 2004. *Helicobacter pylori* test and eradicate versus prompt endoscopy for management of dyspeptic patients: 6.7 year follow up of a randomised trial. *Gut*, 53(12), pp.1758–63.

Lassen, A.T., Pedersen, F.M., Bytzer, P. and de Muckadell, O.B.S., 2000. *Helicobacter pylori* test-and-eradicate versus prompt endoscopy for management of dyspeptic patients: a randomised trial. *The Lancet*, 356(9228), pp.455–460.

Latifi-Navid, S., Ghorashi, S.A., Siavoshi, F., Linz, B., Massarrat, S., Khegay, T., Salmanian, A.-H., Shayesteh, A.A., Masoodi, M., Ghanadi, K., Ganji, A., Suerbaum, S., Achtman, M., Malekzadeh, R. and Falush, D., 2010. Ethnic and geographic differentiation of *Helicobacter pylori* within Iran. *PloS one*, 5(3), p.e9645.

Lehours, P., Dupouy, S., Chaineux, J., Ruskoné-Fourmestraux, A., Delchier, J.-C., Morgner, A., Mégraud, F. and Ménard, A., 2007. Genetic diversity of the HpyC1I restriction modification system in *Helicobacter pylori*. *Research in Microbiology*, 158(3), pp.265–271.

Lehours, P., Vale, F.F., Bjursell, M.K., Melefors, O., Advani, R., Glavas, S., Guegueniat, J., Gontier, E., Lacomme, S., Alves Matos, A., Menard, A., Mégraud, F., Engstrand, L.

and Andersson, A.F., 2011. Genome Sequencing Reveals a Phage in *Helicobacter pylori*. *mBio*, 2(6).

Leiser, O.P., Merkley, E.D., Clowers, B.H., Deatherage Kaiser, B.L., Lin, A., Hutchison, J.R., Melville, A.M., Wagner, D.M., Keim, P.S., Foster, J.T. and Kreuzer, H.W., 2015. Investigation of *Yersinia pestis* Laboratory Adaptation through a Combined Genomics and Proteomics Approach. *PloS one*, 10(11), p.e0142997.

Leja, M., Axon, A. and Brenner, H., 2016. Epidemiology of *Helicobacter pylori* infection. *Helicobacter*, 21, pp.3–7.

Lekmeechai, S., Su, Y.-C., Brant, M., Alvarado-Kristensson, M., Vallström, A., Obi, I., Arnqvist, A. and Riesbeck, K., 2018. *Helicobacter pylori* Outer Membrane Vesicles Protect the Pathogen From Reactive Oxygen Species of the Respiratory Burst. *Frontiers in Microbiology*, 9.

Lemoine, R., Pachlopnik-Schmid, J., Farin, H.F., Bigorgne, A., Debré, M., Sepulveda, F., Héritier, S., Lemale, J., Talbotec, C., Rieux-Laucat, F., Ruemmele, F., Morali, A., Cathebras, P., Nitschke, P., Bole-Feysot, C., Blanche, S., Brousse, N., Picard, C., Clevers, H., Fischer, A. and de Saint Basile, G., 2014. Immune deficiency-related enteropathy-lymphocytopenia-alopecia syndrome results from tetratricopeptide repeat domain 7A deficiency. *The Journal of allergy and clinical immunology*, 134(6), pp.1354-1364.e6.

Letley, D.P., Rhead, J.L., Twells, R.J., Dove, B. and Atherton, J.C., 2003. Determinants of Non-toxicity in the Gastric Pathogen *Helicobacter pylori*. *Journal of Biological Chemistry*, 278(29), pp.26734–26741.

Letunic, I. and Bork, P., 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*, 23(1), pp.127–128.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and 1000 Genome Project Data Processing Subgroup, 1000 Genome Project Data Processing, 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (Oxford, England), 25(16), pp.2078–9.

Li, H., Liao, T., Debowski, A.W., Tang, H., Nilsson, H.-O., Stubbs, K.A., Marshall, B.J. and Benghezal, M., 2016. Lipopolysaccharide Structure and Biosynthesis in Helicobacter pylori. Helicobacter, 21(6), pp.445–461.

Lieberman, T.D., Flett, K.B., Yelin, I., Martin, T.R., McAdam, A.J., Priebe, G.P. and Kishony, R., 2014. Genetic variation of a bacterial pathogen within individuals with cystic fibrosis provides a record of selective pressures. Nature genetics, 46(1), pp.82–7.

Lina, T.T., Pinchuk, I. V., House, J., Yamaoka, Y., Graham, D.Y., Beswick, E.J. and Reyes, V.E., 2013. CagA-Dependent Downregulation of B7-H2 Expression on Gastric Mucosa and Inhibition of Th17 Responses during *Helicobacter pylori* Infection. *The Journal of Immunology,* 191(7), pp.3838–3846.

Linz, B., Balloux Y\rManica, A, F., Liu, H., Roumagnac, P., Falush, D., Stamer, C., Prugnolle, F., van der Merwe, S.W., Yamaoka, Y., Graham, D.Y., Perez-Trallero T, E., Suerbaum, S. and Achtman, M., 2007. An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature*, 445(7130), pp.915–918.

Linz, B., Windsor, H.M., Gajewski, J.P., Hake, C.M., Drautz, D.I., Schuster, S.C. and Marshall, B.J., 2013. *Helicobacter pylori* genomic microevolution during naturally occurring transmission between adults. *PloS one*, 8(12), p.e82187.

Linz, B., Windsor, H.M., McGraw, J.J., Hansen, L.M., Gajewski, J.P., Tomsho, L.P., Hake, C.M., Solnick, J. V., Schuster, S.C. and Marshall, B.J., 2014. A mutation burst during the acute phase of *Helicobacter pylori* infection in humans and rhesus macaques. *Nature Communications*, 5(1), p.4165.

Liu, H., Fero, J.B., Mendez, M., Carpenter, B.M., Servetas, S.L., Rahman, A., Goldman, M.D., Boren, T., Salama, N.R., Merrell, D.S. and Dubois, A., 2015. Analysis of a single *Helicobacter pylori* strain over a 10-year period in a primate model. *International Journal of Medical Microbiology*, 305(3), pp.392–403.

Lockard, V.G. and Boler, R.K., 1970. Ultrastructure of a spiraled microorganism in the gastric mucosa of dogs. *American journal of veterinary research*, 31(8), pp.1453–62.

Lopetuso, L.R., Napoli, M., Rizzatti, G., Scaldaferri, F., Franceschi, F. and Gasbarrini, A., 2018. Considering gut microbiota disturbance in the management of *Helicobacter pylori* infection. *Expert Review of Gastroenterology & Hepatology*, 12(9), pp.899–906.

López-Vidal, Y., Ponce-de-León, S., Castillo-Rojas, G., Barreto-Zúñiga, R. and Torre-Delgadillo, A., 2008. High Diversity of *vacA* and *cagA Helicobacter pylori* Genotypes in Patients with and without Gastric Cancer. *PLoS ONE*, 3(12), p.e3849.

Luthy, L., Grutter, M.G. and Mittl, P.R.E., 2002. The crystal structure of *Helicobacter pylori* cysteine-rich protein B reveals a novel fold for a penicillin-binding protein. *The Journal of biological chemistry*, 277(12), pp.10187–93.

Maiden, M.C., Bygraves, J.A., Feil, E., Morelli, G., Russell, J.E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D.A., Feavers, I.M., Achtman, M. and Spratt, B.G., 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences of the United States of America*, 95(6), pp.3140–5.

Mailhe, M., Ricaboni, D., Vitton, V., Gonzalez, J.-M., Bachar, D., Dubourg, G., Cadoret, F., Robert, C., Delerce, J., Levasseur, A., Fournier, P.-E., Angelakis, E., Lagier, J.-C. and Raoult, D., 2018. Repertoire of the gut microbiota from stomach to colon using culturomics and next-generation sequencing. *BMC Microbiology*, 18(1), p.157.

Malaty, H.M., Graham, D.Y., Kumagai, T., Hosogaya, S., Misawa, K., Furihata, K., Ota, H., Sei, C., Tanaka, E., Akamatsu, T., Shimizu, T., Kiyosawa, K. and Katsuyama, T., 1998. Adult-to-child transmission of *Helicobacter pylori* infection: Results from an 8-year birth cohort study. *Gastroenterology*, 114, p.A211.

Malfertheiner, P., Megraud, F., O'Morain, C.A., Gisbert, J.P., Kuipers, E.J., Axon, A.T., Bazzoli, F., Gasbarrini, A., Atherton, J., Graham, D.Y., Hunt, R., Moayyedi, P., Rokkas, T., Rugge, M., Selgrad, M., Suerbaum, S., Sugano, K. and El-Omar, E.M., 2017. Management of *Helicobacter pylori* infection—the Maastricht V/Florence Consensus Report. *Gut*, 66(1), pp.6–30.

Mamishi, S., Eshaghi, H., Mahmoudi, S., Bahador, A., Hosseinpour Sadeghi, R., Najafi, M., Farahmand, F., Khodadad, A. and Pourakbari, B., 2016. Intrafamilial transmission of *Helicobacter pylori*: genotyping of faecal samples. *British Journal of Biomedical Science*, 73(1), pp.38–43.

Ben Mansour, K., Fendri, C., Battikh, H., Garnier, M., Zribi, M., Jlizi, A. and Burucoa, C., 2016. Multiple and mixed *Helicobacter pylori* infections: Comparison of two epidemiological situations in Tunisia and France. Infection, *Genetics and Evolution*, 37, pp.43–48.

Markovska, R., Boyanova, L., Yordanov, D., Stankova, P., Gergova, G. and Mitov, I., 2018. Status of *Helicobacter pylori* cag pathogenicity island (cagPAI) integrity and significance of its individual genes. *Infection, Genetics and Evolution*, 59(January), pp.167–171.

Marsh, P.D., 2018. In Sickness and in Health—What Does the Oral Microbiome Mean to Us? An Ecological Perspective. *Advances in Dental Research*, 29(1), pp.60–65.

Marshall, B.J., 2001. One Hundred Years of Discovery and Rediscovery of *Helicobacter pylori* and Its Association with Peptic Ulcer Disease. *Helicobacter pylori: Physiology and Genetics*. ASM Press.

Marshall, B.J., Armstrong, J.A., McGechie, D.B. and Glancy, R.J., 1985. Attempt to fulfil Koch's postulates for pyloric *Campylobacter*. *The Medical journal of Australia*, 142(8), pp.436–9.

Martin, S.L., Edbrooke, M.R., Hodgman, T.C., van den Eijnden, D.H. and Bird, M.I., 1997. Lewis X biosynthesis in *Helicobacter pylori*. Molecular cloning of an alpha(1,3)-fucosyltransferase gene. *The Journal of biological chemistry*, 272(34), pp.21349–56.

Matteo, M.J., Granados, G., Perez, C. V., Olmos, M., Sanchez, C. and Catalano, M., 2007. *Helicobacter pylori* cag pathogenicity island genotype diversity within the gastric niche of a single host. *Journal of Medical Microbiology*, 56(5), pp.664–669.

McClain, M.S., Shaffer, C.L., Israel, D.A., Peek, R.M. and Cover, T.L., 2009. Genome sequence analysis of *Helicobacter pylori* strains associated with gastric ulceration and gastric cancer. *BMC Genomics*, 10(1), p.3.

McGee, D.J., Langford, M.L., Watson, E.L., Carter, J.E., Chen, Y.-T. and Ottemann, K.M., 2005. Colonization and Inflammation Deficiencies in Mongolian Gerbils Infected by *Helicobacter pylori* Chemotaxis Mutants. *Infection and Immunity*, 73(3), pp.1820–1827.

McNairn, E., Bhriain, N.N. and Dorman, C.J., 1995. Overexpression of the *Shigella flexneri* genes coding for DNA topoisomerase IV compensates for loss of DNA topoisomerase I: effect on virulence gene expression. *Molecular Microbiology*, 15(3), pp.507–517.

McNulty, C., Owen, R., Tompkins, D., Hawtin, P., McColl, K., Price, A., Smith, G. and Teare, L., 2002. *Helicobacter pylori* susceptibility testing by disc diffusion. *Journal of Antimicrobial Chemotherapy*, 49(4), pp.601–609.

McNulty, C.A.M. and Whiting, J.W., 2007. Patients' attitudes to *Helicobacter pylori* breath and stool antigen tests compared to blood serology. *Journal of Infection*, 55(1), pp.19–22.

McNulty, R., Ulmschneider, J.P., Luecke, H. and Ulmschneider, M.B., 2013. Mechanisms of molecular transport through the urea channel of *Helicobacter pylori*. *Nature Communications*, 4(1), p.2900.

Mégraud, F., 2004. *H pylori* antibiotic resistance: prevalence, importance, and advances in testing. *Gut*, 53(9), pp.1374–84.

Megraud, F., Coenen, S., Versporten, A., Kist, M., Lopez-Brea, M., Hirschl, A.M., Andersen, L.P., Goossens, H., Glupczynski, Y. and Study Group participants, on behalf of the S.G., 2013. *Helicobacter pylori* resistance to antibiotics in Europe and its relationship to antibiotic consumption. *Gut*, 62(1), pp.34–42.

Mejías-Luque, R. and Gerhard, M., 2017. Immune Evasion Strategies and Persistence of *Helicobacter pylori*. *Springer, Cham*, pp.53–71.

Melo-Narváez, M.C., Rojas-Rengifo, D.F., Jimenéz-Soto, L.F., Delgado Perafán, M.D.P., Mendoza de Molano, B.E., Vera-Chamorro, J.F., Jaramillo, C.A., Melo-Narváez, M.C., Rojas-Rengifo, D.F., Jiménez-Soto, L.F., Delgado, M. del P., Mendoza de Molano, B., Vera-Chamorro, J.F. and Jaramillo, C., 2018. Genotipificación de *cagA* y de la región intermedia de vacA en cepas de *Helicobacter pylori* aisladas de pacientes adultos colombianos y asociación con enfermedades gástricas. *Revista Colombiana de Gastroenterología*, 33(2), p.103.

Mendz, G.L. and Hazell, S.L., 1991. Evidence for a pentose phosphate pathway in *Helicobacter pylori*. *FEMS Microbiology Letters*, 84(3), pp.331–336.

Miftahussurur, M., Cruz, M., Subsomwong, P., Jiménez Abreu, J.A., Hosking, C., Nagashima, H., Akada, J. and Yamaoka, Y., 2017. Clarithromycin-Based Triple Therapy Is Still Useful as an Initial Treatment for *Helicobacter pylori* Infection in the Dominican Republic. *The American Journal of Tropical Medicine and Hygiene*, 96(5), pp.16–0729.

Min, X., Zhang, X., Wang, H., Gong, Y., Li, M., Xu, W., Yin, Y. and Cao, J., 2012. Protection against pneumococcal infection elicited by immunization with glutamyl tRNA synthetase, polyamine transport protein D and sortase A. *Vaccine*, 30(24), pp.3624–3633.

Mini, R., Bernardini, G., Salzano, A.M., Renzone, G., Scaloni, A., Figura, N. and Santucci, A., 2006. Comparative proteomics and immunoproteomics of *Helicobacter pylori* related to different gastric pathologies. *Journal of Chromatography B*, 833(1), pp.63–79.

Mizote, T., Yoshiyama, H. and Nakazawa, T., 1997. Urease-independent chemotactic responses of *Helicobacter pylori* to urea, urease inhibitors, and sodium bicarbonate. *Infection and immunity*, 65(4), pp.1519–21.

Møller, H., Heseltine, E. and Vainio, H., 1995. Working group report on schistosomes, liver flukes and *Helicobacter pylori*. Meeting held at IARC, LYON, 7–14 june 1994. *International Journal of Cancer*, 60(5), pp.587–589.

Montano, V., Didelot, X., Foll, M., Linz, B., Reinhardt, R., Suerbaum, S., Moodley, Y. and Jensen, J.D., 2015. Worldwide Population Structure, Long-Term Demography, and Local Adaptation of *Helicobacter pylori. Genetics*, 200(3), pp.947–63.

Monteiro, M.A., Britton, S., Applebee, L.A. and Baqar, S., 2011. Synthesis and immunogenicity of a *Helicobacter pylori* lipopolysaccharide-based conjugate. *Vaccine*, 29(17), pp.3098–3102.

Moodley, Y., Linz, B., Bond, R.P., Nieuwoudt, M., Soodyall, H., Schlebusch, C.M., Bernhöft, S., Hale, J., Suerbaum, S., Mugisha, L., van der Merwe, S.W. and Achtman, M., 2012. Age of the association between *Helicobacter pylori* and man. *PLoS pathogens*, 8(5), p.e1002693.

Moodley, Y., Linz, B., Yamaoka, Y., Windsor, H.M., Breurec, S., Wu, J.-Y., Maady, A., Bernhoft, S., Thiberge, J.-M., Phuanukoonnon, S., Jobb, G., Siba, P., Graham, D.Y., Marshall, B.J. and Achtman, M., 2009. The Peopling of the Pacific from a Bacterial Perspective. *Science*, 323(5913), pp.527–530.

Mora, D. and Arioli, S., 2014. Microbial Urease in Health and Disease. *PLoS Pathogens*, 10(12), p.e1004472.

Moran, A.P., 2008. Relevance of fucosylation and Lewis antigen expression in the bacterial gastroduodenal pathogen *Helicobacter pylori*. *Carbohydrate Research*, 343(12), pp.1952–1965.

Morelli, G., Didelot, X., Kusecek, B., Schwarz, S., Bahlawane, C., Falush, D., Suerbaum, S. and Achtman, M., 2010. Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. *PLoS Genetics*, 6(7), pp.1–12.

Mostowy, R., Croucher, N.J., Andam, C.P., Corander, J., Hanage, W.P. and Marttinen, P., 2017. Efficient Inference of Recent and Ancestral Recombination within Bacterial Populations. *Molecular Biology and Evolution*, 34(5), pp.1167–1182.

Mukaisho, K., Hagiwara, T., Nakayama, T., Hattori, T. and Sugihara, H., 2014. Potential mechanism of corpus-predominant gastritis after PPI therapy in *Helicobacter pylori*-positive patients with GERD. *World journal of gastroenterology*, 20(34), pp.11962–5.

Nakamura, H., Yoshiyama, H., Takeuchi, H., Mizote, T., Okita, K. and Nakazawa, T., 1998. Urease Plays an Important Role in the Chemotactic Motility of *Helicobacter pylori* in a Viscous Environment. *Infection and immunity*, 66(10), pp. 4832-4837.

Nakayama, Y., Yamaguchi, H., Einaga, N. and Esumi, M., 2016. Pitfalls of DNA Quantification Using DNA-Binding Fluorescent Dyes and Suggested Solutions. *PLOS ONE*, 11(3), p.e0150528.

de Negreiros Bessa, P.P., Barbosa, F.C., do Carmo, A.P.S., Furtado, G.B., Barroso, F.C., Rabenhosrt, S.H.B., Bessa, P.P. de N., Barbosa, F.C., Carmo, A.P.S. do, Furtado, G.B., Barroso, F.C. and Rabenhosrt, S.H.B., 2014. Presence of the Genes *cagA*, *cagE*, *virB11* and Allelic Variation of *vacA* of Helicobacter pylori Are Associated with the Activity of Gastritis. Open Journal of Gastroenterology, 04(11), pp.347–355.

Nell, S., Estibariz, I., Krebes, J., Bunk, B., Graham, D.Y., Overmann, J., Song, Y., Spröer, C., Yang, I., Wex, T., Korlach, J., Malfertheiner, P. and Suerbaum, S., 2018. Genome and Methylome Variation in *Helicobacter pylori* With a cag Pathogenicity Island During Early Stages of Human Infection. *Gastroenterology*, 154(3), pp.612-623.

Nell, S., Kennemann, L., Schwarz, S., Josenhans, C. and Suerbaum, S., 2014. Dynamics of lewis b binding and sequence variation of the *babA* adhesin gene during chronic *Helicobacter pylori* infection in humans. *mBio*, 5(6), pp.1–10.

Nguyen, L.T., Uchida, T., Tsukamoto, Y., Trinh, T.D., Ta, L., Mai, H.B., Le, H.S., Ho, D.Q.D., Hoang, H.H., Matsuhisa, T., Okimoto, T., Kodama, M., Murakami, K., Fujioka, T., Yamaoka, Y. and Moriyama, M., 2010. Clinical relevance of cagPAI intactness in *Helicobacter pylori* isolates from Vietnam. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology*, 29(6), pp.651–60.

Nishizawa, T., Suzuki, H., Matsuzaki, J., Muraoka, H., Tsugawa, H., Hirata, K. and Hibi, T., 2011. *Helicobacter pylori* Resistance to Rifabutin in the Last 7 Years. *Antimicrobial Agents and Chemotherapy*, 55(11), pp.5374–5375.

Nitharwal, R.G., Verma, V., Dasgupta, S. and Dhar, S.K., 2011. *Helicobacter pylori* chromosomal DNA replication: Current status and future perspectives. *FEBS Letters*, 585(1), pp.7–17.

Nobusato, A., Uchiyama, I. and Kobayashi, I., 2000. Diversity of restriction-modification gene homologues in *Helicobacter pylori. Gene*, 259(1–2), pp.89–98.

Noto, J.M., Chopra, A., Loh, J.T., Romero-Gallo, J., Piazuelo, M.B., Watson, M., Leary, S., Beckett, A.C., Wilson, K.T., Cover, T.L., Mallal, S., Israel, D.A. and Peek, R.M., 2017. Pan-genomic analyses identify key *Helicobacter pylori* pathogenic loci modified by carcinogenic host microenvironments. *Gut*, pp.1–12.

O'Morain, N.R., Dore, M.P., O'Connor, A.J.P., Gisbert, J.P. and O'Morain, C.A., 2018. Treatment of *Helicobacter pylori* infection in 2018. *Helicobacter*, 23, p.e12519.

Odenbreit, S., Püls, J., Sedlmaier, B., Gerland, E., Fischer, W. and Haas, R., 2000. Translocation of Helicobacter pylori CagA into Gastric Epithelial Cells by Type IV Secretion. *Science*, 287(5457), pp.1497–1500.

Odenbreit, S., Wieland, B. and Haas, R., 1996. Cloning and genetic characterization of *Helicobacter pylori* catalase and construction of a catalase-deficient mutant strain. *Journal of Bacteriology*, 178(23), pp.6960–6967.

Ogata, S.K., Gales, A.C. and Kawakami, E., 2014. Antimicrobial susceptibility testing for *Helicobacter pylori* isolates from Brazilian children and adolescents: comparing agar dilution, E-test, and disk diffusion. *Brazilian journal of microbiology* : [publication of the Brazilian Society for Microbiology], 45(4), pp.1439–48.

Olbermann, P., Josenhans, C., Moodley, Y., Uhr, M., Stamer, C., Vauterin, M., Suerbaum, S., Achtman, M. and Linz, B., 2010a. A global overview of the genetic and functional diversity in the *Helicobacter pylori* cag pathogenicity island. *PLoS genetics*, 6(8), p.e1001069.

Olbermann, P., Josenhans, C., Moodley, Y., Uhr, M., Stamer, C., Vauterin, M., Suerbaum, S., Achtman, M. and Linz, B., 2010b. A global overview of the genetic and functional diversity in the *Helicobacter pylori* cag pathogenicity island. *PLoS Genetics*, 6(8).

Oleastro, M., Cordeiro, R., Menard, A. and Gomes, J.P., 2010. Allelic Diversity among *Helicobacter pylori* Outer Membrane Protein Genes *homB* and *homA* Generated by Recombination. *Journal of Bacteriology*, 192(15), pp.3961–3968.

Oleastro, M. and Ménard, A., 2013. The Role of *Helicobacter pylori* Outer Membrane Proteins in Adherence and Pathogenesis. *Biology*, 2(3), pp.1110–34.

Osaki, T., Konno, M., Yonezawa, H., Hojo, F., Zaman, C., Takahashi, M., Fujiwara, S. and Kamiya, S., 2015. Analysis of intra-familial transmission of *Helicobacter pylori* in Japanese families. *Journal of Medical Microbiology*, 64(Pt_1), pp.67–73.

Osaki, T., Mabe, K., Hanawa, T. and Kamiya, S., 2008. Urease-positive bacteria in the stomach induce a false-positive reaction in a urea breath test for diagnosis of *Helicobacter pylori* infection. *Journal of Medical Microbiology*, 57(7), pp.814–819.

Oyarzabal, O.A., Rad, R. and Backert, S., 2007. Conjugative transfer of chromosomally encoded antibiotic resistance from *Helicobacter pylori* to *Campylobacter jejuni*. *Journal of clinical microbiology*, 45(2), pp.402–8.

Page, A.J., Cummins, C.A., Hunt, M., Wong, V.K., Reuter, S., Holden, M.T.G., Fookes, M., Falush, D., Keane, J.A. and Parkhill, J., 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, 31(22), pp.3691–3693.

Palmer, E.D., 1954. Investigation of the Gastric Mucosa Spirochetes of the Human. *Gastroenterology*, 27(2), pp.218–220.

Papastergiou, V., Georgopoulos, S.D. and Karatapanis, S., 2014. Treatment of *Helicobacter pylori* infection: Meeting the challenge of antimicrobial resistance. *World Journal of Gastroenterology : WJG*, 20(29), p.9898.

Park, J.Y., Forman, D., Waskito, L.A., Yamaoka, Y. and Crabtree, J.E., 2018. Epidemiology of *Helicobacter pylori* and CagA-positive infections and global variations in gastric cancer. *Toxins*, 10(4).

Park, Y.H. and Kim, N., 2015. Review of atrophic gastritis and intestinal metaplasia as a premalignant lesion of gastric cancer. *Journal of cancer prevention*, 20(1), pp.25–40.

Parsons, B.N., Ijaz, U.Z., D'Amore, R., Burkitt, M.D., Eccles, R., Lenzi, L., Duckworth, C.A., Moore, A.R., Tiszlavicz, L., Varro, A., Hall, N. and Pritchard, D.M., 2017. Comparison of the human gastric microbiota in hypochlorhydric states arising as a result of *Helicobacter pylori*-induced atrophic gastritis, autoimmune atrophic gastritis and proton pump inhibitor use. *PLOS Pathogens*, 13(11), p.e1006653.

Paul, R., Postius, S., Melchers, K. and Schafer, K.P., 2001. Mutations of the *Helicobacter pylori* Genes *rdxA* and *pbp1* Cause Resistance against Metronidazole and Amoxicillin. *Antimicrobial Agents and Chemotherapy*, 45(3), pp.962–965.

Payne, J., Available at: https://github.com/tinybio/filter_contigs.

Pedersen, B.S. and Quinlan, A.R., 2018. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, 34(5), pp.867–868.

Peek, R.M., Miller, G.G., Tham, K.T., Perez-Perez, G.I., Zhao, X., Atherton, J.C. and Blaser, M.J., 1995. Heightened inflammatory response and cytokine expression in vivo to *cagA*+ *Helicobacter pylori* strains. *Laboratory investigation; a journal of technical methods and pathology*, 73(6), pp.760–70.

Peña, J., Rojas, H., Reyes, N., Fernández-Delgado, M., García-Amado, M.-A., Michelangeli, F. and Contreras, M., 2017. Multiple *cag* genotypes of *Helicobacter pylori* isolates colonize the oesophagus in individual hosts in a Venezuelan population. *Journal of Medical Microbiology*, 66(2), pp.226–235.

Pérez-Gil, J., Bergua, M., Boronat, A. and Imperial, S., 2010. Cloning and functional characterization of an enzyme from *Helicobacter pylori* that catalyzes two steps of the methylerythritol phosphate pathway for isoprenoid biosynthesis. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1800(9), pp.919–928.

Perez-Perez, G.I., Witkin, S.S., Decker, M.D. and Blaser, M.J., 1991. Seroprevalence of *Helicobacter pylori* infection in couples. *Journal of clinical microbiology*, 29(3), pp.642–4.

Perry, S., de la Luz Sanchez, M., Yang, S., Haggerty, T., Hurst, P., Perez-Perez, G. and Parsonnet, J., 2006. Gastroenteritis and Transmission of *Helicobacter pylori* Infection in Households. *Emerging Infectious Diseases*, 12(11), pp.1701–1708.

Peters, T.M., Owen, R.J., Slater, E., Varea, R., Teare, E.L. and Saverymuttu, S., 2001. Genetic diversity in the *Helicobacter pylori cag* pathogenicity island and effect on expression of anti-CagA serum antibody in UK patients with dyspepsia. *Journal of Clinical Pathology*, 54(3), pp.219–223.

Peterson, W.L., 1997. The role of antisecretory drugs in the treatment of *Helicobacter pylori* infection. *Aliment Pharmacol Ther*, 11, pp.21-25.

Pham, K.T., Weiss, E., Jiménez Soto, L.F., Breithaupt, U., Haas, R. and Fischer, W., 2012. CagI is an essential component of the Helicobacter pylori Cag type IV secretion system and forms a complex with CagL. *PloS one*, 7(4), p.e35341.

Pham, V.H., Maaroufi, H., Balg, C., Blais, S.P., Messier, N., Roy, P.H., Otis, F., Voyer, N., Lapointe, J. and Chênevert, R., 2016. Inhibition of *Helicobacter pylori* Glu-tRNA G ln amidotransferase by novel analogues of the putative transamidation intermediate. *FEBS Letters*, 590(19), pp.3335–3345.

Pincock, S., 2005. Nobel Prize winners Robin Warren and Barry Marshall. *Lancet*, 366(9495), p.1429.

Pohlmann, J. and Brötz-Oesterhelt, H., 2004. New aminoacyl-tRNA synthetase inhibitors as antibacterial agents. Current drug targets. *Infectious disorders*, 4(4), pp.261–72.

Prachasilpchai, W., Nuanualsuwan, S., Chatsuwan, T., Techangamsuwan, S., Wangnaitham, S. and Sailasuta, A., 2007. Diagnosis of *Helicobacter* spp. infection in canine stomach. *Journal of veterinary science*, 8(2), pp.139–45.

Price, M.N., Dehal, P.S. and Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, 5(3), p.e9490.

Pride, D.T., Meinersmann, R.J. and Blaser, M.J., 2001. Allelic Variation within *Helicobacter pylori babA* and *babB*. *Infection and Immunity*, 69(2), pp.1160–1171.

Public Health England, 2017. Test and treat for *Helicobacter pylori* (HP) in dyspepsia Quick reference guide for primary care: For consultation and local adaptation.

Qi, H., Menzel, R. and Tse-Dinh, Y.-C., 1997. Regulation of *Escherichia coli topA* gene transcription: involvement of a σs-dependent promoter. *Journal of Molecular Biology*, 267(3), pp.481–489.

Qureshi, N., Li, P. and Gu, Q., 2019. Probiotic therapy in *Helicobacter pylori* infection: a potential strategy against a serious pathogen? *Applied Microbiology and Biotechnology*, 103(4), pp.1573–1588.

Rajendran, V., Kalita, P., Shukla, H., Kumar, A. and Tripathi, T., 2018. Aminoacyl-tRNA synthetases: Structure, function, and drug discovery. *International Journal of Biological Macromolecules*, 111, pp.400–414.

Ramarao, N., Gray-Owen, S.D., Backert, S. and Meyer, T.F., 2000. *Helicobacter pylori* inhibits phagocytosis by professional phagocytes involving type IV secretion components. *Molecular Microbiology*, 37(6), pp.1389–1404.

Ramarao, N. and Meyer, T.F., 2001. *Helicobacter pylori* Resists Phagocytosis by Macrophages: Quantitative Assessment by Confocal Microscopy and Fluorescence-Activated Cell Sorting. *Infection and Immunity*, 69(4), pp.2604–2611.

Raymond, J., Thiberg, J.-M., Chevalier, C., Kalach, N., Bergeret, M., Labigne, A. and Dauga, C., 2004. Genetic and transmission analysis of *Helicobacter pylori* strains within a family. *Emerging infectious diseases*, 10(10), pp.1816–21.

Ražuka-Ebela, D., Giupponi, B. and Franceschi, F., 2018. *Helicobacter pylori* and extragastric diseases. *Helicobacter*, 23, p.e12520.

Reyes-Leon, A., Atherton, J.C., Argent, R.H., Puente, J.L. and Torres, J., 2007a. Heterogeneity in the activity of Mexican *Helicobacter pylori* strains in gastric epithelial cells and its association with diversity in the *cagA* gene. *Infection and immunity*, 75(7), pp.3445–54.

Reyes-Leon, A., Atherton, J.C., Argent, R.H., Puente, J.L. and Torres, J., 2007b. Heterogeneity in the Activity of Mexican *Helicobacter pylori* Strains in Gastric Epithelial Cells and Its Association with Diversity in the *cagA* Gene †. *Infection and Immunity*, 75(7), pp.3445–3454.

Rhead, J.L., Letley, D.P., Mohammadi, M., Hussein, N., Mohagheghi, M.A., Eshagh Hosseini, M. and Atherton, J.C., 2007. A New *Helicobacter pylori* Vacuolating Cytotoxin Determinant, the Intermediate Region, Is Associated With Gastric Cancer. *Gastroenterology*, 133(3), pp.926–936.

Rittig, M.G., Shaw, B., Letley, D.P., Thomas, R.J., Argent, R.H. and Atherton, J.C., 2003. *Helicobacter pylori*-induced homotypic phagosome fusion in human monocytes is independent of the bacterial *vacA* and *cag* status. *Cellular microbiology*, 5(12), pp.887–99.

Rizvanov, A.A., Haertlé, T., Bogomolnaya, L. and Talebi Bezmin Abadi, A., 2019. *Helicobacter pylori* and Its Antibiotic Heteroresistance: A Neglected Issue in Published Guidelines. *Frontiers in microbiology*, 10, p.1796.

Rodŕıguez-Concepción, M., Campos, N., Maŕıa Lois, L., Maldonado, C., Hoeffler, J.-F., Grosdemange-Billiard, C., Rohmer, M. and Boronat, A., 2000. Genetic evidence of branching in the isoprenoid pathway for the production of isopentenyl diphosphate and dimethylallyl diphosphate in *Escherichia coli. FEBS Letters*, 473(3), pp.328–332.

Rolig, A.S., Shanks, J., Carter, J.E. and Ottemann, K.M., 2012. *Helicobacter pylori* Requires TlpD-Driven Chemotaxis To Proliferate in the Antrum. *Infection and Immunity*, 80(10), pp.3713–3720.

Roosild, T.P., Castronovo, S., Healy, J., Miller, S., Pliotas, C., Rasmussen, T., Bartlett,

W., Conway, S.J. and Booth, I.R., 2010. Mechanism of ligand-gated potassium efflux in bacterial pathogens. *Proceedings of the National Academy of Sciences of the United States of America*, 107(46), pp.19784–9.

Rosche, W.A. and Foster, P.L., 2000. Determining mutation rates in bacterial populations. *Methods*, 20(1), pp.4–17.

Rothenbacher, D., Bode, G., Berg, G., Knayer, U., Gonser, T., Adler, G. and Brenner, H., 1999. *Helicobacter pylori* among Preschool Children and Their Parents: Evidence of Parent-Child Transmission. *The Journal of Infectious Diseases*, 179(2), pp.398–402.

Rotimi, O., Cairns, A., Gray, S., Moayyedi, P. and Dixon, M.F., 2000. Histological identification of *Helicobacter pylori*: comparison of staining methods. *Journal of clinical pathology*, 53(10), pp.756–9.

de Sablet, T., Piazuelo, M.B., Shaffer, C.L., Schneider, B.G., Asim, M., Chaturvedi, R., Bravo, L.E., Sicinschi, L.A., Delgado, A.G., Mera, R.M., Israel, D.A., Romero-Gallo, J., Peek, R.M., Cover, T.L., Correa, P. and Wilson, K.T., 2011. Phylogeographic origin of *Helicobacter pylori* is a determinant of gastric cancer risk. *Gut*, 60(9), pp.1189–1195.

SACHS, G., SHIN, J.M. and HOWDEN, C.W., 2006. Review article: the clinical pharmacology of proton pump inhibitors. *Alimentary Pharmacology and Therapeutics*, 23(s2), pp.2–8.

Sagaert, X., 2016. *Helicobacter pylori* Infection and MALT Lymphoma. In: *Helicobacter pylori Research*. *Tokyo: Springer Japan*, pp.423–441.

Sakitani, K., Nishizawa, T., Arita, M., Yoshida, S., Kataoka, Y., Ohki, D., Yamashita, H., Isomura, Y., Toyoshima, A., Watanabe, H., Iizuka, T., Saito, Y., Fujisaki, J., Yahagi, N., Koike, K. and Toyoshima, O., 2018. Early detection of gastric cancer after *Helicobacter pylori* eradication due to endoscopic surveillance. *Helicobacter*, 23(4), p.e12503.

Salazar, J.C., Ahel, I., Orellana, O., Tumbula-Hansen, D., Krieger, R., Daniels, L. and Soll, D., 2003. Coevolution of an aminoacyl-tRNA synthetase with its tRNA substrates. *Proceedings of the National Academy of Sciences*, 100(24), pp.13863–13868.

Salomon, H., 1896. Ueber das Spirillum des Säugetiermagens und sein Verhalten zu den Belegzellen. *Zentralbl Bakteriol*.

Santos, M.F., New, R.R.C., Andrade, G.R., Ozaki, C.Y., Sant'Anna, O.A., Mendonça-Previato, L., Trabulsi, L.R. and Domingos, M.O., 2010. Lipopolysaccharide as an antigen target for the formulation of a universal vaccine against *Escherichia coli* O111 strains. *Clinical and vaccine immunology*: CVI, 17(11), pp.1772–80.

Sause, W. E., Castillo, A. R., & Ottemann, K. M., 2012. The *Helicobacter pylori* autotransporter ImaA (HP0289) modulates the immune response and contributes to host colonization. Infection and immunity, 80(7), 2286–2296. doi:10.1128/IAI.00312-12

Savoldi, A., Carrara, E., Graham, D.Y., Conti, M. and Tacconelli, E., 2018. Prevalence

of Antibiotic Resistance in *Helicobacter pylori*: A Systematic Review and Meta-analysis in World Health Organization Regions. *Gastroenterology*, 155(5), pp.1372-1382.e17.

Sayers, E.W., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Mizrachi, I., Ostell, J., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Yaschenko, E. and Ye, J., 2009. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 37(Database), pp.D5–D15.

Schröder, G., Krause, S., Zechner, E.L., Traxler, B., Yeo, H.-J., Lurz, R., Waksman, G. and Lanka, E., 2002. TraG-like proteins of DNA transfer systems and of the *Helicobacter pylori* type IV secretion system: inner membrane gate for exported substrates? *Journal of bacteriology*, 184(10), pp.2767–79.

Schutze, K., Hentschel, E., Dragosics, B. and Hirschl, A.M., 1995. *Helicobacter pylori* reinfection with identical organisms: transmission by the patients' spouses. *Gut*, 36(6), pp.831–833.

Schweinitzer, T., Mizote, T., Ishikawa, N., Dudnik, A., Inatsu, S., Schreiber, S., Suerbaum, S., Aizawa, S.-I. and Josenhans, C., 2008. Functional Characterization and Mutagenesis of the Proposed Behavioral Sensor TlpD of *Helicobacter pylori*. *Journal of Bacteriology*, 190(9), pp.3244–3255.

Seemann, T., 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14), pp.2068–2069.

Seemann, T., n.d. SNIPPY. Available at: https://github.com/tseemann/snippy. Version 4.4.0.

Selgrad, M., Tammer, I., Langner, C., Bornschein, J., Meißle, J., Kandulski, A., Varbanova, M., Wex, T., Schlüter, D. and Malfertheiner, P., 2014. Different antibiotic susceptibility between antrum and corpus of the stomach, a possible reason for treatment failure of *Helicobacter pylori* infection. *World journal of gastroenterology*, 20(43), pp.16245–51.

Seo, J.W., Park, J.Y., Shin, T.-S. and Kim, J.G., 2019. The analysis of virulence factors and antibiotic resistance between *Helicobacter pylori* strains isolated from gastric antrum and body. *BMC Gastroenterology*, 19(1), p.140.

Serin, A., Tankurt, E., Şarkış, C. and Simsek, I., 2015. The prevalence of *Helicobacter pylori* infection in patients with gastric and duodenal ulcers - a 10-year, single-centre experience. *Przeglad gastroenterologiczny*, 10(3), pp.160–3.

Shanks, A.-M. and El-Omar, E.M., 2009. *Helicobacter pylori* infection, host genetics and gastric cancer. *Journal of Digestive Diseases*, 10(3), pp.157–164.

Sheh, A., Chaturvedi, R., Merrell, D.S., Correa, P., Wilson, K.T. and Fox, J.G., 2013.

Phylogeographic origin of *Helicobacter pylori* determines host-adaptive responses upon coculture with gastric epithelial cells. *Infection and immunity*, 81(7), pp.2468–77.

Sheikh, A.F., Yadyad, M.J., Goodarzi, H., Hashemi, S.J., Aslani, S., Assarzadegan, M.-A. and Ranjbar, R., 2018. *CagA* and *vacA* allelic combination of *Helicobacter pylori* in gastroduodenal disorders. *Microbial Pathogenesis*, 122, pp.144–150.

Sheu, S.-M., Sheu, B.-S., Yang, H.-B., Lei, H.-Y. and Wu, J.-J., 2007. Anti-Lewis X antibody promotes *Helicobacter pylori* adhesion to gastric epithelial cells. *Infection and immunity*, 75(6), pp.2661–7.

Shimoyama, T., 2005. Relation of CagA seropositivity to cag PAI phenotype and histological grade of gastritis in patients with *Helicobacter pylori* infection. *World Journal of Gastroenterology*, 11(24), p.3751.

Shimoyama, T., 2013. Stool antigen tests for the management of *Helicobacter pylori* infection. *World journal of gastroenterology*, 19(45), pp.8188–91.

Shirin, H., Kenet, G., Shevah, O., Wardi, J., Wardi, Y., Birkenfeld, S., Shahmurov, M., Bruck, R., Niv, Y., Moss, S.F. and Avni, Y., 2001. Evaluation of a novel continuous real time (13)C urea breath analyser for *Helicobacter pylori*. Ali*mentary pharmacology & therapeutics*, 15(3), pp.389–94.

Si, X.B., Lan, Y. and Qiao, L., 2017. A meta-analysis of randomized controlled trials of bismuth-containing quadruple therapy combined with probiotic supplement for eradication of *Helicobacter pylori*. *Zhonghua nei ke za zhi*, 56(10), pp.752–759.

Singh, N., Chevé, G., Avery, M.A. and McCurdy, C.R., 2007. Targeting the methyl erythritol phosphate (MEP) pathway for novel antimalarial, antibacterial and herbicidal drug discovery: inhibition of 1-deoxy-D-xylulose-5-phosphate reductoisomerase (DXR) enzyme. *Current pharmaceutical design*, 13(11), pp.1161–77.

Sipponen, P. and Maaroos, H.-I., 2015. Chronic gastritis. *Scandinavian journal of gastroenterology*, 50(6), pp.657–67.

Sisson, G., Jeong, J.Y., Goodwin, A., Bryden, L., Rossler, N., Lim-Morrison, S., Raudonikiene, A., Berg, D.E. and Hoffman, P.S., 2000. Metronidazole activation is mutagenic and causes DNA fragmentation in *Helicobacter pylori* and in *Escherichia coli* containing a cloned *H. pylori RdxA*(+) (Nitroreductase) gene. *Journal of bacteriology*, 182(18), pp.5091–6.

Solnick, J. V, Hansen, L.M., Salama, N.R., Boonjakuakul, J.K. and Syvanen, M., 2004. Modification of *Helicobacter pylori* outer membrane protein expression during experimental infection of rhesus macaques. *Proceedings of the National Academy of Sciences of the United States of America*, 101(7), pp.2106–11.

Spiegelhalder, C., Gerstenecker, B., Kersten, A., Schiltz, E. and Kist, M., 1993. Purification of *Helicobacter pylori* superoxide dismutase and cloning and sequencing of the gene. *Infection and immunity*, 61(12), pp.5315–25.

Srikhanta, Y.N., Gorrell, R.J., Steen, J.A., Gawthorne, J.A., Kwok, T., Grimmond, S.M., Robins-Browne, R.M. and Jennings, M.P., 2011. Phasevarion Mediated Epigenetic Gene Regulation in *Helicobacter pylori. PLoS ONE*, 6(12), p.e27569.

Stead, C.M., Beasley, A., Cotter, R.J. and Trent, M.S., 2008. Deciphering the unusual acylation pattern of *Helicobacter pylori* lipid A. *Journal of bacteriology*, 190(21), pp.7012–21.

Steer, H.W. and Colin-Jones, D.G., 1975. Mucosal changes in gastric ulceration and their response to carbenoxolone sodium. *Gut*, 16(8), pp.590–597.

Stein, S.C., Faber, E., Bats, S.H., Murillo, T., Speidel, Y., Coombs, N. and Josenhans, C., 2017. *Helicobacter pylori* modulates host cell responses by CagT4SS-dependent translocation of an intermediate metabolite of LPS inner core heptose biosynthesis. *PLOS Pathogens*, 13(7), p.e1006514.

Stolte, M. and Meining, A., 2001. The Updated Sydney System: Classification and Grading of Gastritis as the Basis of Diagnosis and Treatment. *Canadian Journal of Gastroenterology*, 15(9), pp.591–598.

Stone, M.A., 1999. Transmission of *Helicobacter pylori. Postgraduate medical journal*, 75(882), pp.198–200.

Stratton, A. and Laczek, J., 2013. Serology Improves Patient Adherence to *Helicobacter Pylori* Testing. *Hawai'i Journal of Medicine & Public Health*, 72(9 Suppl 4), p.71.

Suerbaum, S., Brauer-Steppkes, T., Labigne, A., Cameron, B. and Drlica, K., 1998. Topoisomerase I of *Helicobacter pylori*: juxtaposition with a flagellin gene (*flaB*) and functional requirement of a fourth zinc finger motif. *Gene*, 210(1), pp.151–161.

Suerbaum, S. and Josenhans, C., 2007. *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nature Reviews Microbiology*, 5(6), pp.441–452.

Sugano, K., Tack, J., Kuipers, E.J., Graham, D.Y., El-Omar, E.M., Miura, S., Haruma, K., Asaka, M., Uemura, N., Malfertheiner, P. and faculty members of Kyoto Global Consensus Conference, 2015. Kyoto global consensus report on *Helicobacter pylori* gastritis. *Gut*, 64(9), pp.1353–67.

Sycuro, L.K., Wyckoff, T.J., Biboy, J., Born, P., Pincus, Z., Vollmer, W. and Salama, N.R., 2012. Multiple Peptidoglycan Modification Networks Modulate *Helicobacter pylori*'s Cell Shape, Motility, and Colonization Potential. *PLoS Pathogens*, 8(3), p.e1002603.

Takahashi, S., Kuzuyama, T., Watanabe, H. and Seto, H., 1998. A 1-deoxy-D-xylulose 5-phosphate reductoisomerase catalyzing the formation of 2-C-methyl-D-erythritol 4-phosphate in an alternative nonmevalonate pathway for terpenoid biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 95(17), pp.9879–84.

Takeuchi, H., Israel, D.A., Miller, G.G., Donahue, J.P., Krishna, U., Gaus, K. and Peek, Jr., R.M., 2002. Characterization of Expression of a Functionally Conserved *Helicobacter pylori* Methyltransferase-Encoding Gene within Inflamed Mucosa and during In Vitro Growth. *The Journal of Infectious Diseases*, 186(8), pp.1186–1189.

Talebi Bezmin Abadi, A., 2017. Strategies used by *Helicobacter pylori* to establish persistent infection. *World Journal of Gastroenterology*, 23(16), p.2870.

Talley, N.J., 2005. Review: prompt endoscopy is not a cost effective strategy for initial management of dyspepsia. *Evidence-Based Medicine*, 10(6), pp.185–185.

Tannaes, T., Dekker, N., Bukholm, G., Bijlsma, J.J. and Appelmelk, B.J., 2001. Phase variation in the *Helicobacter pylori* phospholipase A gene and its role in acid adaptation. *Infection and immunity*, 69(12), pp.7334–40.

Terradot, L., Bayliss, R., Oomen, C., Leonard, G.A., Baron, C. and Waksman, G., 2005. Structures of two core subunits of the bacterial type IV secretion system, VirB8 from *Brucella suis* and ComB10 from *Helicobacter pylori*. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12), pp.4596–601.

Terry, K., Williams, S.M., Connolly, L. and Ottemann, K.M., 2005. Chemotaxis Plays Multiple Roles during *Helicobacter pylori* Animal Infection. *Infection and Immunity*, 73(2), pp.803–811.

Testerman, T.L., McGee, D.J. and Mobley, H.L.T., 2001. Adherence and Colonization. *Helicobacter pylori*: Physiology and Genetics. *ASM Press*.

Testerman, T.L. and Morris, J., 2014. Beyond the stomach: an updated view of *Helicobacter pylori* pathogenesis, diagnosis, and treatment. *World journal of gastroenterology*, 20(36), pp.12781–808.

Thorell, K., Hosseini, S., Palacios Gonzáles, R.V.P., Chaotham, C., Graham, D.Y., Paszat, L., Rabeneck, L., Lundin, S.B., Nookaew, I. and Sjöling, Å., 2016. Identification of a Latin American-specific *BabA* adhesin variant through whole genome sequencing of *Helicobacter pylori* patient isolates from Nicaragua. *BMC Evolutionary Biology*, 16(1), pp.1–16.

Tock, M.R. and Dryden, D.T., 2005. The biology of restriction and anti-restriction. *Current Opinion in Microbiology*, 8(4), pp.466–472.

Tomb, J.-F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E.F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H.G., Glodek, A., McKenney, K., Fitzegerald, L.M., Lee, N., Adams, M.D., Hickey, E.K., Berg, D.E., Gocayne, J.D., Utterback, T.R., Peterson, J.D., Kelley, J.M., Cotton, M.D., Weidman, J.M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W.S., Borodovsky, M., Karp, P.D., Smith, H.O., Fraser, C.M. and Venter, J.C., 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori. Nature*, 388(6642), pp.539–547.

Tomer, R., Ye, L., Hsueh, B. and Deisseroth, K., 2014. Advanced CLARITY for rapid and high-resolution imaging of intact tissues. *Nature protocols*, 9(7), pp.1682–97.

Torres-Barceló, C., 2018. The disparate effects of bacteriophages on antibiotic-resistant bacteria. *Emerging microbes & infections*, 7(1), p.168.

Treangen, T.J., Ondov, B.D., Koren, S. and Phillippy, A.M., 2014. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biology*, 15(11), p.524.

Treangen, T.J. and Salzberg, S.L., 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics*, 13(1), pp.36–46.

Tummuru, M.K., Sharma, S.A. and Blaser, M.J., 1995. *Helicobacter pylori picB*, a homologue of the *Bordetella pertussis* toxin secretion protein, is required for induction of IL-8 in gastric epithelial cells. *Molecular microbiology*, 18(5), pp.867–76.

Tursi, A., Di Mario, F., Franceschi, M., De Bastiani, R., Elisei, W., Baldassarre, G., Ferronato, A., Grillo, S., Landi, S., Zamparella, M., De Polo, M., Boscariolo, L. and Picchio, M., 2017. New bismuth-containing quadruple therapy in patients infected with *Helicobacter pylori*: A first Italian experience in clinical practice. *Helicobacter*, 22(3), p.e12371.

Urgesi, R., Cianci, R. and Riccioni, M.E., 2012. Update on triple therapy for eradication of *Helicobacter pylori*: current status of the art. Clinical and Experimental *Gastroenterology*, 5, p.151.

Vale, F.F. and Lehours, P., 2018. Relating Phage Genomes to *Helicobacter pylori* Population Structure: General Steps Using Whole-Genome Sequencing Data. *International journal of molecular sciences*, 19(7).

Vale, F.F., Nunes, A., Oleastro, M., Gomes, J.P., Sampaio, D.A., Rocha, R., Vítor, J.M.B., Engstrand, L., Pascoe, B., Berthenet, E., Sheppard, S.K., Hitchings, M.D., Mégraud, F., Vadivelu, J. and Lehours, P., 2017. Genomic structure and insertion sites of *Helicobacter pylori* prophages from various geographical origins. *Scientific Reports*, 7(1), p.42471.

Vale, F.F., Vadivelu, J., Oleastro, M., Breurec, S., Engstrand, L., Perets, T.T., Mégraud, F. and Lehours, P., 2015a. Dormant phages of *Helicobacter pylori* reveal distinct populations in Europe. *Scientific Reports*, 5(1), p.14333.

Vale, F.F., Vadivelu, J., Oleastro, M., Breurec, S., Engstrand, L., Perets, T.T., Mégraud, F. and Lehours, P., 2015b. Dormant phages of *Helicobacter pylori* reveal distinct populations in Europe. *Scientific Reports*, 5, pp.1–8.

Vallve, M., Vergara, M., Gisbert, J.P. and Calvet, X., 2002. Single vs. double dose of a proton pump inhibitor in triple therapy for *Helicobacter pylori* eradication: a meta-analysis. *Alimentary Pharmacology and Therapeutics*, 16(6), pp.1149–1156.

Vasu, K., Nagamalleswari, E. and Nagaraja, V., 2012. Promiscuous restriction is a cellular defense strategy that confers fitness advantage to bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 109(20), pp.E1287-93.

Vitoriano, I., Vítor, J.M.B., Oleastro, M., Roxo-Rosa, M. and Vale, F.F., 2013. Proteome variability among *Helicobacter pylori* isolates clustered according to genomic methylation. *Journal of Applied Microbiology*, 114(6), pp.1817–1832.

Vyse, A.J., Gay, N.J., Hesketh, L.M., Andrews, N.J., Marshall, B., Thomas, H.I.J., Morgan-Capner, P. and Miller, E., 2002. The burden of *Helicobacter pylori* infection in England and Wales. *Epidemiology and infection*, 128(3), pp.411–7.

Wade, W.G., 2013. The oral microbiome in health and disease. *Pharmacological Research*, 69(1), pp.137–143.

Walker, M.M., 2003. Is intestinal metaplasia of the stomach reversible? *Gut*, 52(1), pp.1–4.

Wang, L.-L., Liu, J.-X., Yu, X.-J., Si, J.-L., Zhai, Y.-X. and Dong, Q.-J., 2018. Microbial community reshaped in gastric cancer. *European review for medical and pharmacological sciences*, 22(21), pp.7257–7264.

Warren, J.R. and Marshall, B., 1983. Unidentified curved bacilli on gastric epithelium in active chronic gastritis. *Lancet*, 1(8336), pp.1273–5.

Weeks, D.L., Eskandari, S., Scott, D.R. and Sachs, G., 2000. A H+-Gated Urea Channel: The Link Between *Helicobacter pylori* Urease and Gastric Colonization. *Science*, 287(5452), pp.482–485.

Wilcock, B., 2013. Histopathology. *Canine and Feline Gastroenterology*, pp.333–385.

Williams, S.M., Chen, Y.-T., Andermann, T.M., Carter, J.E., McGee, D.J. and Ottemann, K.M., 2007. *Helicobacter pylori* Chemotaxis Modulates Inflammation and Bacterium-Gastric Epithelium Interactions in Infected Mice. *Infection and Immunity*, 75(8), pp.3747–3757.

Winter, J.A., Letley, D.P., Cook, K.W., Rhead, J.L., Zaitoun, A.A.M., Ingram, R.J.M., Amilon, K.R., Croxall, N.J., Kaye, P. V, Robinson, K. and Atherton, J.C., 2014. A role for the vacuolating cytotoxin, VacA, in colonization and *Helicobacter pylori*-induced metaplasia in the stomach. *The Journal of infectious diseases*, 210(6), pp.954–63.

Wirth, H.P., Beins, M.H., Yang, M., Tham, K.T. and Blaser, M.J., 1998. Experimental infection of Mongolian gerbils with wild-type and mutant *Helicobacter pylori* strains. *Infection and immunity*, 66(10), pp.4856–66.

Wirth, T., Wang, X., Linz, B., Novick, R.P., Lum, J.K., Blaser, M., Morelli, G., Falush, D. and Achtman, M., 2004. Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: Lessons from Ladakh. *Proceedings of the National Academy of Sciences*, 101(14), pp.4746–4751.

Wong, E.H.J., Ng, C.G., Chua, E.G., Tay, A.C.Y., Peters, F., Marshall, B.J., Ho, B., Goh, K.L., Vadivelu, J. and Loke, M.F., 2016. Comparative genomics revealed multiple *Helicobacter pylori* genes associated with biofilm formation in vitro. *PLoS ONE*, 11(11), pp.1–16.

Wood, D.E. and Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3), p.R46.

World Health Organization, 2017. Prioritization of pathogens to guide discovery, research and development of new antibiotics for drug-resistant bacterial infections, including tuberculosis. *Geneva: World Health Organization*.

Wurm, P., Dörner, E., Kremer, C., Spranger, J., Maddox, C., Halwachs, B., Harrison, U., Blanchard, T., Haas, R., Högenauer, C., Gorkiewicz, G. and Fricke, W.F., 2018. Qualitative and Quantitative DNA- and RNA-Based Analysis of the Bacterial Stomach Microbiota in Humans, Mice, and Gerbils. *mSystems*, 3(6).

Xu, Q., Morgan, R.D., Roberts, R.J. and Blaser, M.J., 2000. Identification of type II restriction and modification systems in *Helicobacter pylori* reveals their substantial diversity among strains. *Proceedings of the National Academy of Sciences*, 97(17), pp.9671–9676.

Yahara, K., Furuta, Y., Morimoto, S., Kikutake, C., Komukai, S., Matelska, D., Dunin-Horkawicz, S., Bujnicki, J.M., Uchiyama, I. and Kobayashi, I., 2016. Genome-wide survey of codons under diversifying selection in a highly recombining bacterial species, *Helicobacter pylori. DNA Research*, 23(2), pp.135–143.

Yahara, K., Lehours, P. and Vale, F.F., 2019. Analysis of genetic recombination and the pan-genome of a highly recombinogenic bacteriophage species. *Microbial Genomics*, 5(8), pp.e000282.

Yamaoka, Y., Kita, M., Kodama, T., Imamura, S., Ohno, T., Sawai, N., Ishimaru, A., Imanishi, J. and Graham, D.Y., 2002a. *Helicobacter pylori* infection in mice: Role of outer membrane proteins in colonization and inflammation. *Gastroenterology*, 123(6), pp.1992–2004.

Yamaoka, Y., Ojo, O., Fujimoto, S., Odenbreit, S., Haas, R., Gutierrez, O., El-Zimaity, H.M.T., Reddy, R., Arnqvist, A. and Graham, D.Y., 2006. *Helicobacter pylori* outer membrane proteins and gastroduodenal disease. *Gut*, 55(6), pp.775–81.

Yamaoka, Y., Orito, E., Mizokami, M., Gutierrez, O., Saitou, N., Kodama, T., Osato, M.S., Kim, J.G., Ramirez, F.C., Mahachai, V. and Graham, D.Y., 2002b. *Helicobacter pylori* in North and South America before Columbus. *FEBS letters*, 517(1–3), pp.180–4.

Yang, J.-C., Lu, C.-W. and Lin, C.-J., 2014. Treatment of *Helicobacter pylori* infection: Current status and future concepts. *World Journal of Gastroenterology*, 20(18), p.5283.

Yonezawa, H., Osaki, T., Fukutomi, T., Hanawa, T., Kurata, S., Zaman, C., Hojo, F. and

Kamiya, S., 2017. Diversification of the AlpB Outer Membrane Protein of *Helicobacter pylori* Affects Biofilm Formation and Cellular Adhesion. *Journal of bacteriology*, 199(6), pp.e00729-16.

Yu, C.-K., Wang, C.-J., Chew, Y., Wang, P.-C., Yin, H.-S. and Kao, M.-C., 2016. Functional characterization of *Helicobacter pylori* 26695 sedoheptulose 7-phosphate isomerase encoded by *hp0857* and its association with lipopolysaccharide biosynthesis and adhesion. *Biochemical and Biophysical Research Communications*, 477(4), pp.794–800.

Yu, C., Li, L., Chen, W., Jiao, Y., Yang, N., Yang, E., Zhang, J., Chen, L. and Li, Y., 2011a. Levofloxacin Susceptibility Testing for *Helicobacter pylori* in China: Comparison of E-Test and Disk Diffusion Method. *Helicobacter*, 16(2), pp.119–123.

Yu, Z., Lavèn, M., Klepsch, M., Gier, J.-W. de, Bitter, W., Ulsen, P. van and Luirink, J., 2011b. Role for *Escherichia coli* YidD in Membrane Protein Insertion. *Journal of Bacteriology*, 193(19), pp.5242–5251.

Zamani, M., Vahedi, A., Maghdouri, Z. and Shokri-Shirvani, J., 2017. Role of food in environmental transmission of *Helicobacter pylori*. *Caspian journal of internal medicine*, 8(3), pp.146–152.

Zhang, C., Yamada, N., Wu, Y.-L., Wen, M., Matsuhisa, T. and Matsukura, N., 2005. Comparison of *Helicobacter pylori* infection and gastric mucosal histological features of gastric ulcer patients with chronic gastritis patients. *World journal of gastroenterology*, 11(7), pp.976–81.

Zhao, Y., Arce-Gorvel, V., Conde-Álvarez, R., Moriyon, I. and Gorvel, J.-P., 2018. Vaccine development targeting lipopolysaccharide structure modification. *Microbes and Infection*, 20(9–10), pp.455–460.

Zheng, P.-Y. and Jones, N.L., 2003. *Helicobacter pylori* strains expressing the vacuolating cytotoxin interrupt phagosome maturation in macrophages by recruiting and retaining TACO (coronin 1) protein. *Cellular microbiology*, 5(1), pp.25–40.

Zullo, A., Hassan, C., Romiti, A., Giusto, M., Guerriero, C., Lorenzetti, R., Campo, S.M. and Tomao, S., 2012. Follow-up of intestinal metaplasia in the stomach: When, how and why. *World journal of gastrointestinal oncology*, 4(3), pp.30–6.

10. Appendix

Electronic version of this appendix can be found within the OneDrive appendix directory (https://myntuac-my.sharepoint.com/:f:/r/personal/n0667645_my_ntu_ac_uk/Documents/OneDrive_link?csf=1&e=Uulp26) or directly from the following OneDrive link https://myntuac-my.sharepoint.com/:f:/g/personal/n0667645_my_ntu_ac_uk/ErizNGJWWJ9lgB8DrfwO-NoB7hWGG_JoBHot01TfbYM4MQ?e=3lsa7d.