Genome evolution and virulence in H. pylori: Identifying the genes/alleles underlying phenotype variation

Berthenet, Elvire

GENOME EVOLUTION AND VIRULENCE IN *H. PYLORI*: IDENTIFYING THE GENES/ALLELES UNDERLYING PHENOTYPE VARIATION

Elvire Berthenet

Submitted to Swansea University in fulfilment of the requirements for the Degree of:
Doctor of Philosophy

Swansea University
2018

Summary

An estimated 50% of all people carry the stomach bacterium *Helicobacter pylori (H. pylori)*. This organism is responsible for gastric problems like gastritis and gastric ulcers, and is one of the major causes of gastric cancer worldwide. Large numbers of people carry this organism asymptomatically and many questions remain about why serious symptoms develop in a subset of infected humans.

These extremely recombinant bacteria may take different evolutionary trajectories in different people, and some genomic changes may be associated with gastric cancer. To test this, and learn more about the genetics of cancer-associated *H. pylori*, different approaches were used.

First, evolution of *H. pylori* populations was investigated looking at both core and accessory genomes and revealed traces of the long and complex history of the Americas in the bacterial genomes, as well as a similar evolution in core and accessory genome. This was the first time accessory genome of *H. pylori* was studied that way. Secondly, evolution occurring in the bacterial genome during colonisation of a single host was studied in mice model. This analysis revealed small changes during the passage from a human host to a mice host, and during the long-term colonisation of mice stomach. Then a Genome Wide Association Study (GWAS) approach was applied to a large isolate collection sampled across Europe comprising strains isolated from cancer patients and strains from asymptomatic or gastritis-suffering patients. This approach identified 11 polymorphisms in 9 genes (3 cagPAI genes, *babA*, *hpaA*, 1 outer membrane protein coding gene *HP1055* and 3 other core genes (*HP0747*, *HP0709* and *HP0468*) associated with cancer and a preliminary risk score was built to identify high risk strains. Finally, variations observed among clinical isolates of *H. pylori* from European patients with different pathologies in terms of motility and ability to trigger cytokine production in two types of cells were quantified. Motility variations were not associated with the disease type, but a link was observed for cytokine production. This was compared to genomic variations, confirming the role of known genomic factors such as cagPAI genes and sheding light to possible functions of a number of new genes.

Declarations and statements

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed .....Berthenet..... (candidate)

Date .....25/07/2018.....

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated.
Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .....Berthenet..... (candidate)

Date .....25/07/2018.....

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed .....Berthenet..... (candidate)

Date .....25/07/2018.....

Table of Contents

Acknowledgements

This section has to start with my supervision team. It changed a lot during these three years, but I want to thank every one of them, for taking me under their wings and helping me get through this thesis. I will start with those who were here when I started but are not in Swansea anymore, Dr Jane Mikhail and Prof Samuel Sheppard.

Even though our collaboration was not easy, Dr Mikhail was the one who introduced me to the world of *Helicobacter pylori*. She directed me in the laboratory and I would not have had the technical background I needed without her. She also started a lot of the collaborations with other groups from all over the world, and gathered up part of the precious collection of strains used in the laboratory.

Prof Samuel Sheppard left Swansea after my second year to go back to his home land, England. But he did not leave me behind, and kept an eye on me all the way until submission of this thesis. He introduced me to the international scientific stage, by taking me to conferences where I got to develop my own network in the bacterial genomics area, offering me opportunities of collaborations with other groups and sending me in another laboratory for a full month. He put trust in me when my imposture syndrome was trying to push me down. He valued my opinions and work and made me feel like I was exactly where I was supposed to be. He shared his knowledge and scientific ideas with me to bring my thesis as far and as high as it could go. For his incredible management, for the drinks we shared in pubs all over the world, and for the beginning of a long collaboration, I would like to thank him truly.

My current supervision team also deserves a big thank you. Prof Gareth Jenkins, who was my secondary supervisor, was there from day one and stayed until the end. I would like to thank him for keeping an eye on me and my work, even though my thesis subject was in a large part outside of his research area. We had a few interesting discussions which helped me greatly. He also read my posters and thesis when I needed him to. I thank him for all this. Prof Tom Wilkinson had the difficult task of becoming my first supervisor after the previous one left. His constant optimism helped me through the last year of this PhD, and his deep knowledge of immunology was a useful addition to my supervision panel. He was the one reading my thesis in detail and helping me with submission, and for this I would like to thank him.

This thesis being multi-disciplinary, there are a large number of people I need to acknowledge, for their inputs in this thesis work.

I thank Koji Yahara, for the execution of FineStructure (in Chapter 1 and 5), Chromopainting (in Chapter 1), Roary (in Chapter 5) and bugWAS (in Chapter 5). He was also of great help for me to understand those methods which allowed me to analyse the results by myself.

I thank Leonardos Mageiros, for creating the pan-genome script (used in all chapters) and teaching me how to use it, for introducing me to GWAS, and for creating and executing the ClonalFrame based GWAS used in Chapter 5.

I thank Guillaume Meric for creating the two GWAS results figures, based on the results I prepared (Chapter 5).

I thank Matthew Hitchings and Ben Pascoe for sequencing the DNA samples and assembling the sequences.

I thank my two summer students: Guislaine Zoller for her help with the motility assay, and Charlotte Lebrun for her help with AGS co-cultures and IL-8 ELISA (Chapter 6). It was a pleasure to supervise them and introduce them to the research world.

I thank Clément Viguier for his help with the accessory plots statistics in Chapter 1.

I thank Kaisa Thorell and Daniel Falush for all the discussions we shared, the fruitful collaborations we successfully completed, and their personal encouragements and advices.

I thank Prof Martin Blaser and Prof Guillermo Perez-Perez for welcoming me in New York for a full month and making me feel at home.

Of course, I thank all the scientists who shared clinical isolates with me. Without isolates, none of this work could have been done.

I thank Alexandra Elbakyan, without whom I would not have had access to publications I needed.

I thank all the Swansea University team, the ones who left during this PhD, those who arrived, and those who welcomed me and never left: for the pub lunches, the lunch time and tea time discussions, the after work drinks, the cute animal pictures when I needed a cheer up, the violin sessions, the pole fitness classes, but also their help with all my questions about science or welsh legislation, and their time spent reading my thesis.

A few words for my professional past: all the teachers and professors who gave me an interest in science or fed this interest, and my 6-months internship supervisor at Roslin institute, Jacqueline Smith, who was my first contact with research and bioinformatics. A few words for my professional future: Philippe Lehours and his group, who trust me to join them in Bordeaux from January.

But a PhD is not only a scientific adventure, and I cannot forget my friends and family, who were there for me during these three intensive years. First, I want to thank my fiancé, Xavier Martyn, who was "only" my boyfriend when I started in 2014, and is now my husband. He left France to follow me here, and it would have been much more difficult, not to say impossible without his support. I thank my parents and sisters who visited me several times, offering me some time off when I needed it and making sure I was enjoying French gastronomy at its maximum each time I was coming home. My parents also encouraged me towards the path of science when I was younger, and supported me emotionally and financially all the way. I also apologize for the lack of time I had for them. I thank all of my friends (met in real life or online), for visiting me in Swansea, travelling with me in countries of Europe, hosting me, chatting online, or even proof-reading my thesis (Marie, Charline, Jean, Alexis, Poncho and Medusa for reading the thesis). And I will finish by thanking all the artists I listened to during those long hours writing up these few hundred pages (in particular Postmodern Jukebox and Ibrahim Maalouf but there are so many others).

List of Figures

List of Tables

List of Abbreviations

ANOVA: analysis of variance

AP-1: activator protein 1

BB: brucella broth

BIGSdb: Bacterial Isolate Genome sequence database

BSA: bovine serum albumin

BSL-2: biosafety level 2

CagPAI: Cag Pathogenicity Island

CCL: Chemokine (C-C motif) ligand

CDS: coding sequence

CDH-1: cadherin-1

CFU: colony forming unit

CNRCH : Centre National de Référence des Campylobacters et Hélicobacters

DMSO: Dimethyl Sulfoxide

DNA: Deoxyribonucleic acid

DNP: double nucleotide polymorphism

EHMSG: European Helicobacter and Microbiota Study Group

ELISA: enzyme-linked immunosorbent assay

ExPASy: Expert Protein Analysis System

FBS: foetal bovine serum

FISH: fluorescence *in situ* hybridization

GC: gastric cancer

GIST : gastrointestinal stromal tumor

GWAS : genome-wide association study

H. pylori : *Helicobacter pylori*

HGT : Horizontal gene transfer

HPC Wales : High Performance Computing Wales

HpGP: *Helicobacter pylori* Genome Project

HRP : horseradish peroxidase

IL : interleukin

IM: intestinal metaplasia

iTOL: interactive tree of life

LPS: Lipopolysaccharides

MALT: mucosa-associated lymphoid tissue

MID: microbiology and infectious diseases

ML: MALT lymphoma

MLST: multi-locus sequence typing

NAG: non-atrophic gastritis

nBLAST: nucleotide Basic Local Alignment Search Tool

NCBI: National Center for Biotechnology Information

NF-κB: nuclear factor kappa-light-chain-enhancer of activated B cells

NS: non-synonymous

OD: optical density

OLGA: Operative Link for Gastritis Assessment

PATRIC: Pathosystem Resource Integration Center

PBS: Phosphate-Buffered Saline

PCR: polymerase chain reaction

PV: phase variation

PMA : Phorbol 12-myristate 13-acetate

PPI: proton pump inhibitor

Prog: progressive toward gastric cancer

RAST : Rapid Annotation Sequences Tool

RPMI medium: Roswell Park Memorial Institute medium

S: Synonymous

S media: Standard media

SAT: stool antigen test

SNP: single nucleotide polymorphism

T4SS: Type IV bacterial secretion system

TLR: Toll-like receptor

TNF-α: tumor necrosis factor α

TNM system: tumor, nodes and metastasis

TTC: 2,3,5,-triphenyltetrazolium chloride

UK: United Kingdom

USA: United States of America

VFdb: Virulence Factors database

WGS: whole-genome sequencing

1 Introduction

1.1 The human stomach

The gastrointestinal system is a complex interplay of organs allowing digestion and absorption of ingested food and liquids. The stomach is part of this system, and one of its functions is to regulate acid secretion to maintain sterilisation of ingested nutrients (Smolka and Schubert 2017).

1.1.1 Anatomy

The stomach is divided into 5 regions: The cardia which corresponds to the entrance of the stomach, the pylorus which is the exit of the stomach, and three areas inside named fundus, corpus (body) and antrum (Figure 1.1).

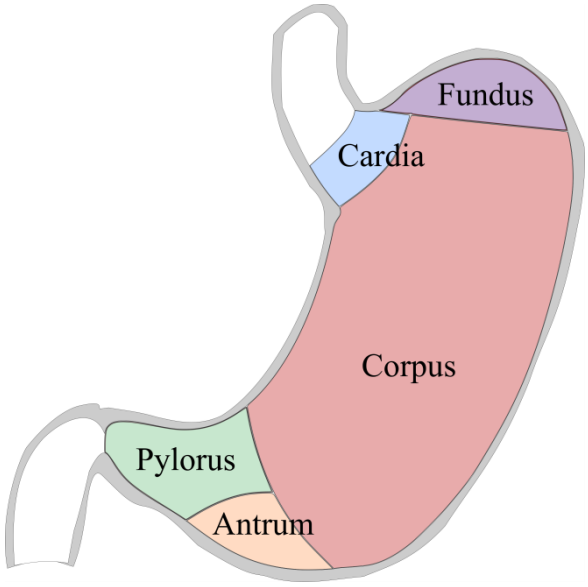


Figure 1.1 Normal anatomy of the human stomach

The fundus and corpus are acidic environments, due to the presence of acid-secreting glands. The epithelium of the antrum, on the other hand, produces alkaline secretions (Soybel 2005). The entire stomach is richly vascularized and comprises several layers of tissue. The inner lining is the mucosa (mucous membrane). Under this layer is the submucosa, composed of connective tissue. Below this there is a muscle layer called the muscularis propria (or muscularis externa). Finally, the serosa (or visceral peritoneum) is the fibrous membrane covering the outside of the stomach ("Canadian Cancer Society" 2017; Soybel 2005).

1.1.2 Physiology

Stomach physiology has three main functions. First function is to store food, for at least 2 hours. The specificities of the gastroesophageal junction and pylorus make this possible. Second function is to mix and mechanically breakdown food. This is achieved through contraction and relaxation of the muscle layers of the stomach. Finally, and maybe the most important function of the stomach is food digestion. This last function is largely dependent on the specificities of the gastric mucosa, which contain specialized cells and glands producing hydrochloric acid and digestive enzymes ("Canadian Cancer Society" 2017). The gastric mucosa can be divided into two main regions: acid-secreting and non-acid secreting. The corpus and fundus are acid- and pepsinogen-secreting, whereas the antrum is alkaline-secreting (Soybel 2005). The gastric lumen transports gastric juice (pH 1-2) during fasting periods. The surface epithelium, present in both the antrum and the corpus/fundus regions, secretes a mucus layer. The pH is around 5 to 6 inside this mucous layer, in contact with the epithelial cell surface (Quigley and Turnberg 1987; Talley et al. 1992). These less acidic conditions protect the gastric epithelium but also make it more suitable for bacterial colonisation.

1.1.3 Stomach microbiome

Numerous properties combine to limit bacterial growth in the stomach. Acidity and non-specific proteases are one of the main antibacterial properties of the stomach, but others exist. Bacterial growth is also inhibited by lactoferrin, antibacterial peptides such as LL-37, β-defensins 1 and 2 and components of gastric mucin and bile acids. Degradation of peptidoglycan, which is a cell-wall component of many bacteria, is achieved through the surfactant protein D (Algood and Cover 2006). Most ingested bacteria are killed in the stomach, but it is never entirely sterile. The mucus layer provides protection and an opportunity for survival. The microorganisms found in the stomach include those that colonise it, and those that pass through the gastric niche. The first organism discovered in the human stomach was *Helicobacter pylori* (*H. pylori*), then named *Pyloric campylobacter* (B. Marshall and Warren 1984). The three other genera dominating a normal acidic stomach free of *H. pylori* are *Veillonella*, *Lactobacillus* and *Clostridium*. There are, however, various organisms that are more

difficult to culture in the laboratory, and 16S rRNA sequencing has revealed more diverse genera (128 phylotypes) (Engstrand and Lindberg 2013; Bik et al. 2006).

1.1.4 Gastric Disorders and Gastric Cancer

Disturbance of stomach functions can be of many different types, and signs can range from slight discomfort to serious complications leading to death. Many gastric diseases in humans are associated with infection by *H. pylori* (Testerman and Morris 2014). Even though the signs observed in gastric disorders and diseases are usually similar, there is a wide diversity of symptoms that can be used to diagnose stomach diseases, ranging from asymptomatic gastritis (in more than 80% of infected population) to peptic ulcers (about 15%), gastric mucosa-associated lymphoid tissue (MALT) lymphoma (less than 0.5%) and gastric cancer (GC) (0.5-2%) (Figure 1.2) (John C. Atherton 2006; Sgouros and Bergele 2006).

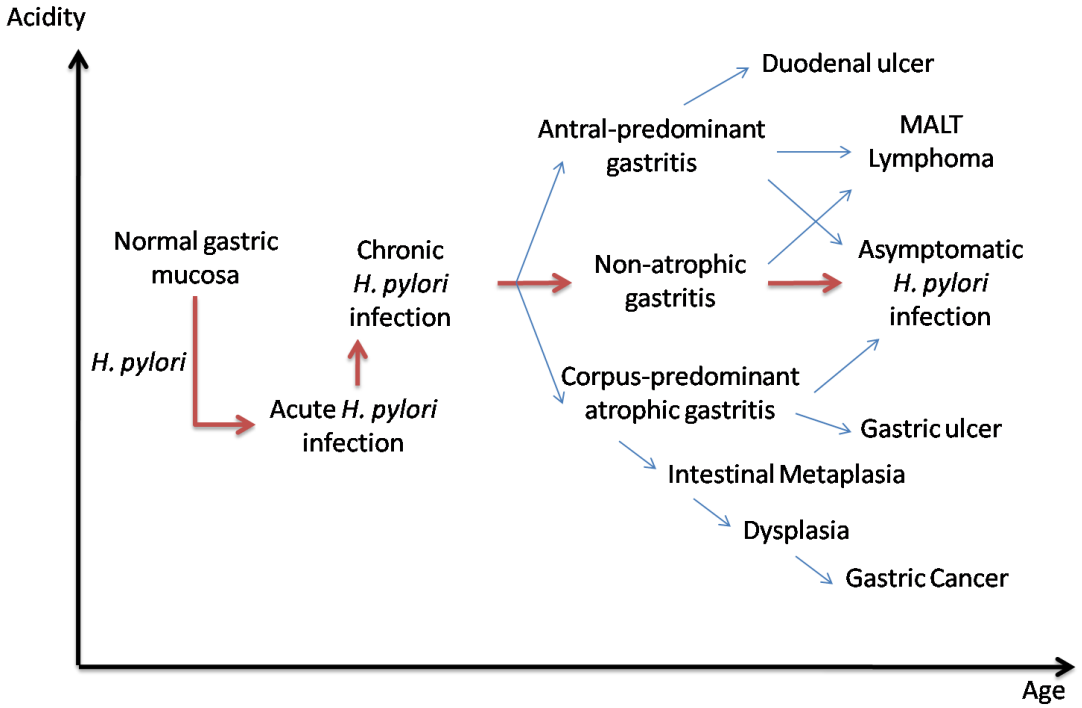


Figure 1.2 Flowchart of the different outcomes of *H. pylori* infection and their link to age and acid production.
Adapted from (Chung et al. 2005). Large red arrows represent the main outcome for each step.

The nature of signs and symptoms depends on numerous factors, including age and pH (Figure 1.2). For instance, a stomach with high acid production will favour the generation of duodenal ulcers or MALT lymphoma, whereas a less acidic environment will more likely lead to gastric ulcers or GC (Chung et al. 2005).

1.1.4.1 Gastritis

In most cases, gastritis, which is an inflammation of the stomach, is the first and only pathologic outcome resulting from an *H. pylori* infection. More than 80% of people infected with *H. pylori* will only get asymptomatic chronic gastritis, and about 80% of chronic gastritis cases are linked to the presence of *H. pylori* (Nordenstedt et al. 2013). Gastritis can evolve into different stages or move towards more serious issues. Types of gastritis encountered include superficial gastritis (close to asymptomatic), diffuse antral gastritis, postgastrectomy (reflux) gastritis, diffuse corporal atrophic gastritis and multifocal atrophic gastritis (Correa 1988). The last two stages of gastritis are considered pre-cancerous stages, as they are often linked to IM.

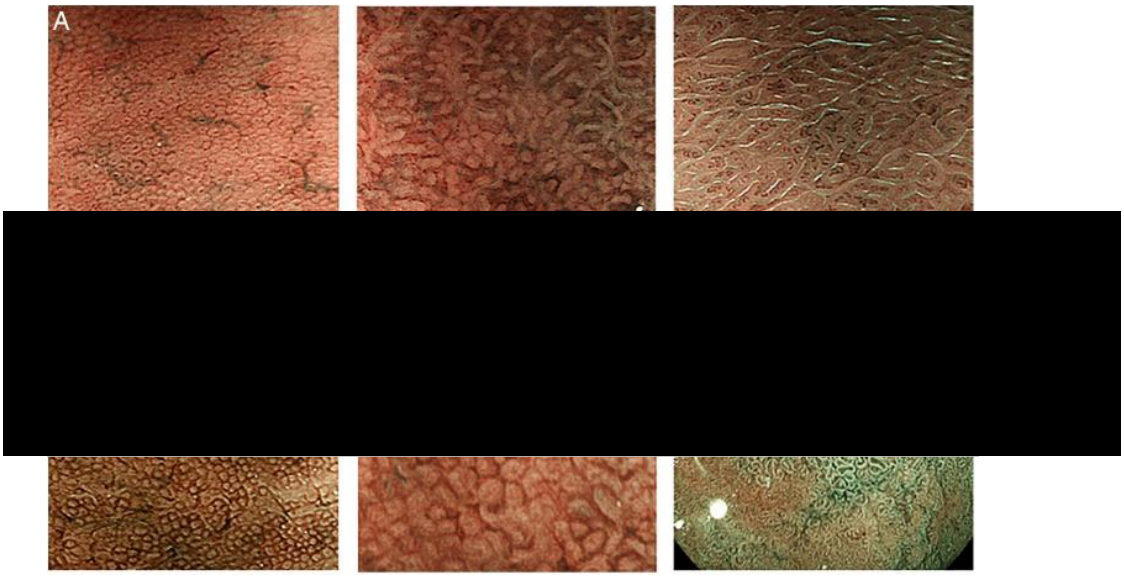


Figure 1.3: Image enhanced endoscopy of the gastric mucosa.
Adapted from (Sugano et al. 2015). **(A)** Narrow band imaging of the gastric mucosa. **(B)** Blue laser imaging of the gastric mucosa. *H. pylori* negative mucosa is shown on the left side images. *H. pylori*-infected mucosa with inflammation is shown on the central images. Intestinal metaplasia is shown on the right side images.

Classifications of gastritis (Capelle et al. 2010; Sipponen and Price 2011) are often inconsistent, as the information given by available classifications is not very useful to

clinicians (Sugano et al. 2015). The Operative Link for Gastritis Assessment (OLGA) is a classification of gastritis based on atrophy score and topography (Rugge et al. 2007). However, conventional endoscopy does not allow the diagnosis of atrophy and IM with certainty (Sugano et al. 2015). New techniques such as image-enhanced endoscopy are now available to achieve this classification (Figure 1.3). In general, a non-atrophic gastritis (NAG) is close to asymptomatic.

1.1.4.2 Peptic ulcers

Peptic ulcer is a non-fatal disease linked to gastritis. Peptic ulcer can be divided into two types of ulcer; duodenal or gastric (Figure 1.4), duodenal ulcers being more prevalent than gastric cancer in most countries (Calam 1998).

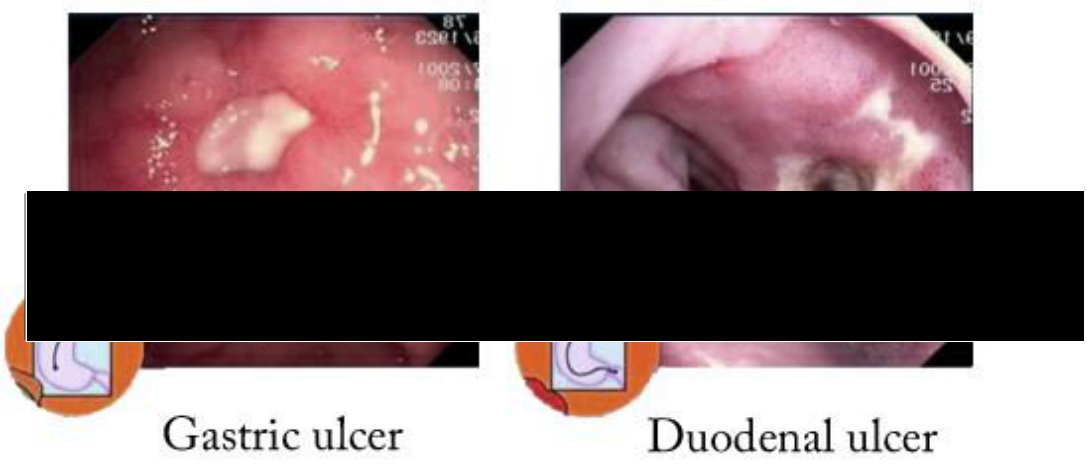


Figure 1.4: Endoscopic images of gastric ulcer and duodenal ulcer with positions from which the pictures were taken.
Adapted from ("TabletsManual.com" 2017).

Antral-predominant gastritis is associated with duodenal ulcer, whereas pangastritis is associated with gastric ulcer (Hwang et al. 2015; Lanas and Chan 2017). More than 90% of patients with dudodenal ulcer and 58-94% of patients with gastric ulcer are infected with *H. pylori* (Calam 1998),

1.1.4.3 Gastric cancer

Incidence rates are decreasing, but remain high. 9% (723,000 deaths) of all cancer deaths were due to gastric cancer (GC) in the world in 2012 (Stewart and Wild 2014), and 3% in the UK in 2014 (Cancer Research UK 2017b). Large geographical

disparities are observed both in terms of incidence and mortality (Stewart and Wild 2014). The survival rate is extremely poor, with 26% of 5-year survival rate in the USA between 1999 and 2005 (Jemal et al. 2010), and 20% in England and Wales ("Cancer Research UK" 2017c). GC is associated with chronic gastritis resulting in a very low acid secretion, and can be an evolution of a gastric ulcer (Hwang et al. 2015). GC can be split into two types: adenocarcinomas (90% of the GC cases) and non-adenocarcinomas (Kelley and Duggan 2003). Only 0.42 to 5.4% of all gastric cancer cases are *Helicobacter pylori* negative (Yamamoto et al. 2015).

Non-adenocarcinoma cancers include non-Hodgkin's lymphomas and leiomyosarcomas which make up almost 10% of GC cases, and more rare diseases and syndromes such as adenosquamous, squamous and undifferentiated carcinomas, choriocarcinomas, carcinoid tumors, rhabdomyosarcomas, hemangiopericytomas, and Kaposi's sarcoma (Kelley and Duggan 2003). Most non-adenocarcinoma cancers are rare and not associated with bacterial infection, and therefore will not be discussed further in this thesis. The only non-adenocarcinoma cancer discussed in this thesis is gastric MALT lymphoma (ML), which is associated with bacterial infection in 90% of the cases (Asano et al. 2015). The first description of ML was made in 1983 (Isaacson and Wright 1983; Son et al. 2010). ML can affect different organs, but our interest will focus on gastric ML. It is a tumor occurring in the stomach, a sub-type of non-Hodgkin's lymphoma (Cohen et al. 2006). ML is less prevalent than GC, representing only 7.6% of non-Hodgkin's lymphoma cases ("A Clinical Evaluation of the International Lymphoma Study Group Classification of Non-Hodgkin's Lymphoma. The Non-Hodgkin's Lymphoma Classification Project." 1997).

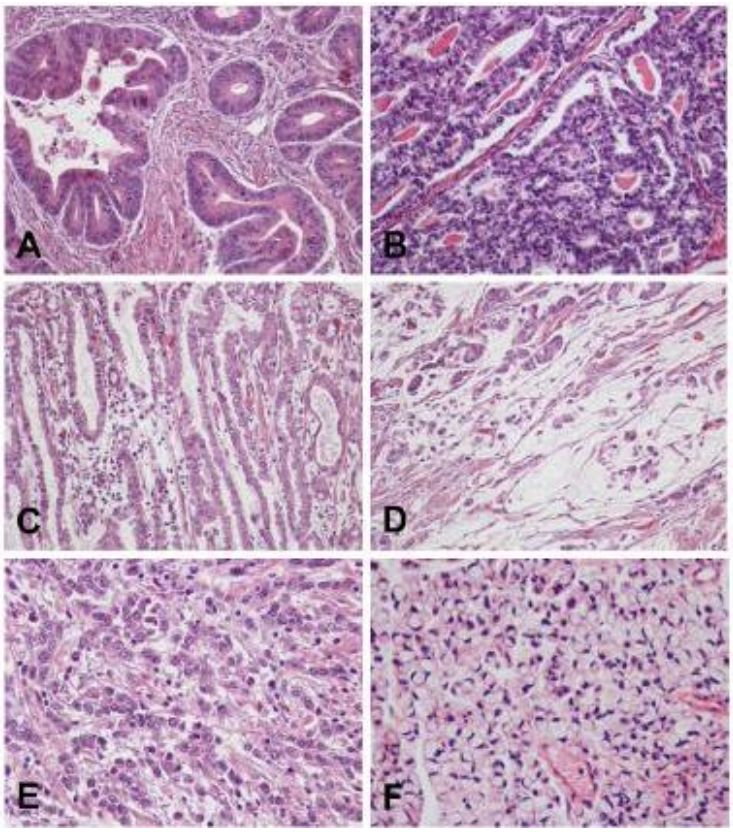


Figure 1.5: Gastric adenocarcinoma (HE stain).
Adapted from (Piazuelo and Correa 2013). Three different tumors of intestinal type are shown in panel **A**, **B** and **C**. Mucinous adenocarcinoma shown in panel **D**. Two types of tumors of diffuse type in panel **E** and **F**.

Histologically, adenocarcinoma can be divided into diffuse (Figure 1.5E-F) and intestinal sub-types (Figure 1.5A-C), and more rarely mucinous (Figure 1.5D), according to the Lauren classification (Hu et al. 2012). Clinically, two staging systems are currently used in the UK: a number system and the tumor, nodes and metastasis (TNM) system ("Cancer Research UK" 2017a). The intestinal subtype predominates in high-risk areas of GC, usually in individuals between 55 and 80 years old, and is more common in males. It is characterized by malignant epithelial cells that show cohesiveness and glandular differentiation and that are infiltrating the stroma (Figure 1.5A-C) preceded by a well-described sequence of histological lesions known as Correa's cascade (Figure 1.6).

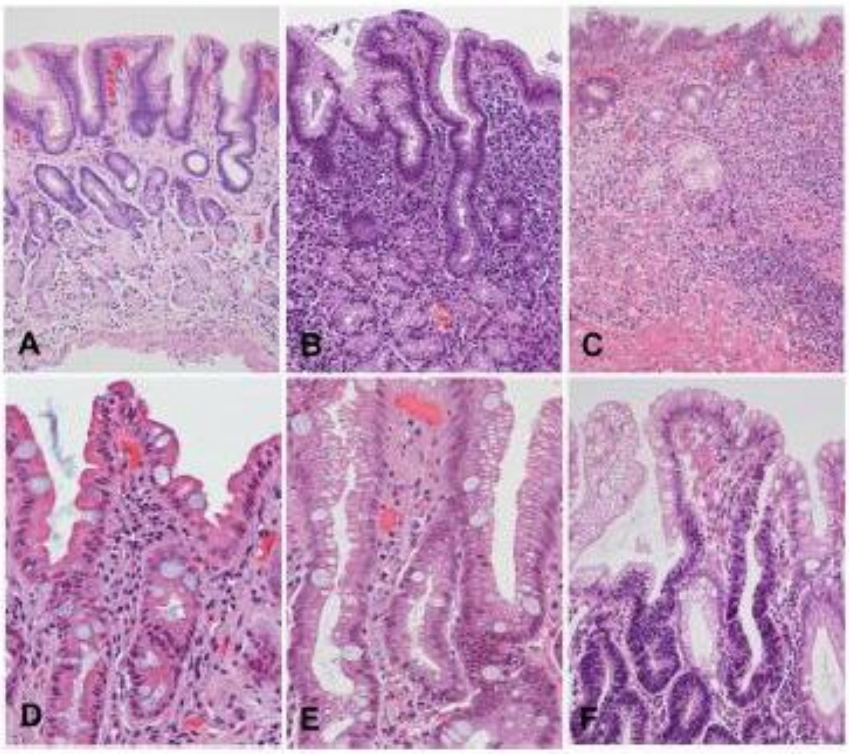


Figure 1.6: Correa's precancerous cascade (HE stain).
Adapted from (Piazuelo and Correa 2013). **A.** Normal gastric mucosa (magnification x100). **B.** Nonatrophic chronic gastritis (magnification x100). **C.** Multifocal atrophic gastritis without intestinal metaplasia (magnification x100). **D.** Intestinal metaplasia, complete type (magnification x200). **E.** Intestinal metaplasia, incomplete type (magnification x200). **F.** Dysplasia (magnification x200).

The diffuse type is often found in younger patients compared to the intestinal type and is not depending on sex. It is composed of discohesive cells that infiltrate the stroma (Figure 1.5E-F). Adenocarcinoma can also be described based on the localisation in the stomach, dividing GC into cardia and non-cardia types.

1.1.4.4 Predisposing factors

The development of a disease is usually due to interplay of genetic, environmental and other factors, such as sex and age. This is the case for gastric cancer.

1.1.4.4.1 Genetic factors

A family history of gastric cancer (GC) is a risk factor, as illustrated through the famous example of Napoleon Bonaparte (M.-G. Smith et al. 2006; Woolf and Isaacson 1961). It is difficult to be certain of the causes of his GC. *H. pylori* infection

may have played a role (Lugli et al. 2007), but there is also speculation of a genetic predisposition (Sokoloff 1938).

The role of mutations in the E-cadherin (encoded by *CDH-1*) is known in the potential development of GC (Becker et al. 1994; Guilford et al. 1998; Carneiro 2012), and is hereditary. Mutations in the *CDH-1* gene occur preferably in 50% of diffuse type adenocarinomas (M.-G. Smith et al. 2006). This is also the case for a polymorphism in the interleukin 1β (IL-1 β) gene (E M El-Omar 2001), with odds ratios of 1.6 or 2.9 for polymorphisms in this gene (M.-G. Smith et al. 2006) and in the tumor-necrosis factor α (TNF-α) gene, with a specific SNP significantly associated with an increased risk of gastric cancer (Yang et al. 2009). Specific combinations of genetic polymorphisms in both TNF-α and IL-10 increased the risk of non-cardia gastric adenocarcinoma with odd ratios up to 27.3 (Emad M El-Omar et al. 2003).

1.1.4.4.2 Environmental factors

Previous dogma suggested that salty and spicy food were responsible for gastric ulcers and other gastric disorders. The effect of salt has been confirmed, together with other preservatives such as nitrate (Joossens et al. 1996). Salt intake increases the risk for gastric cancer by 1.68 (high consumption) or 1.41 (moderately high consumption) compared to low consumption (D'Elia et al. 2012). However, spicy food is more likely to have an indirect effect through the acidity variations in the stomach provoked by such nutrients. Consuming alcohol is also increasing the risk for gastric cancer, with an odds ratio of 1.39 (Ma et al. 2017). Current smokers are more likely to develop gastric cancer, with an odds ratio of 1.69 compared to never smokers (La Torre et al. 2009)

1.1.4.4.3 Other predisposing factors

The main factor involved in development of gastric cancer (GC) is a bacterial infection with *Helicobacter pylori*. The risk of developing a GC is thought to be increased 2-6 fold when the patient is infected with *H. pylori* (Ford et al. 2014). According to a recent review, 89% of non-cardia GC cases are attributable to *H. pylori* infection (Plummer et al. 2015).

As in most cancers, age is also an important risk factor (J. Christie et al. 1997), with most cancer cases observed between 50 and 70 year-old, and only 15% of the cases

affecting adults of less than 41 year-old (T. Buffart et al. 2007). Risks are increasing faster in female populations than in male populations in some areas (W. Chen et al. 2016). However, the incidence remains globally higher in men (Ferlay et al. 2015; Olbermann et al. 2010). Epstein-Barr virus is also a cause in the tumorigenesis of some cases of GC (Camargo et al. 2016; Iizasa et al. 2012).

1.2 *Helicobacter pylori* (*H. pylori*)

H. pylori is a pathogenic Gram-negative microorganism infecting around half of the world's population (Peek and Blaser 2002). Its niche is the human stomach, and it was classified as a group I carcinogen by the International Agency for Research on Cancer in 1994 (IARC 1994). The origins of the association between *H. pylori* and the human species is thought to be at least 100,000 years (Moodley et al. 2012) before the human migration from the African cradle (Linz et al. 2007).

1.2.1 Epidemiology

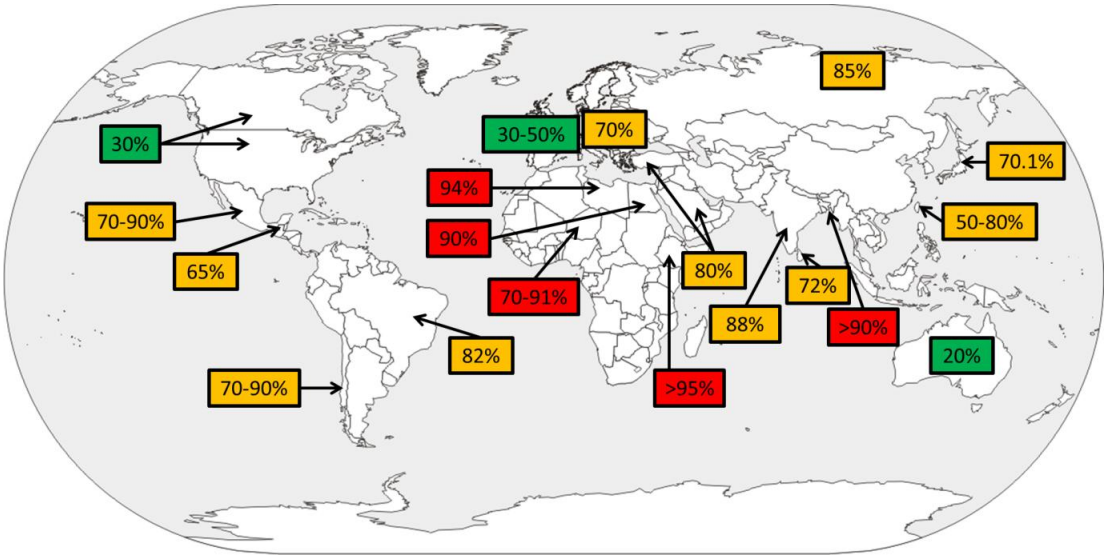


Figure 1.7: Prevalence of *H. pylori* infection in adult populations in the world.
This figure was made using data found in www.worldgastroenterology.org (consulted on 20/02/2017) in a global guideline published in 2010 on *Helicobacter pylori* in developing countries (World Gastroenterology Organisation Global Guidelines 2010). High prevalence of more than 90% (all of Africa and Bengladesh) is highlighted in red. Medium prevalence, between 50% and 90% is highlighted in yellow. Low prevalence of less than 50% (Western Europe, USA, Canada and Australia) is highlighted in green.

The prevalence of *H. pylori* varies geographically with a global average of about 50% (Parsonnet 1998; Taylor and Blaser 1991). Around 74% of people in developing countries and 58% in developed countries carry *H. pylori* (Figure 1.7) (World Gastroenterology Organisation Global Guidelines 2010).

It is usually acquired in childhood, and the microorganism colonises the stomach for years before provoking any symptoms. Moreover, even though chronic *H. pylori* infection is very frequent, up to 80% of carriers will never present any symptoms (M. J. Blaser and Atherton 2004; Dooley et al. 1989; Algood and Cover 2006).

1.2.2 Characteristics of *H. pylori*

H. pylori is a highly motile bacterium with a spiral or curved morphology and efficient flagella (Figure 1.8) (B. Marshall and Warren 1984).

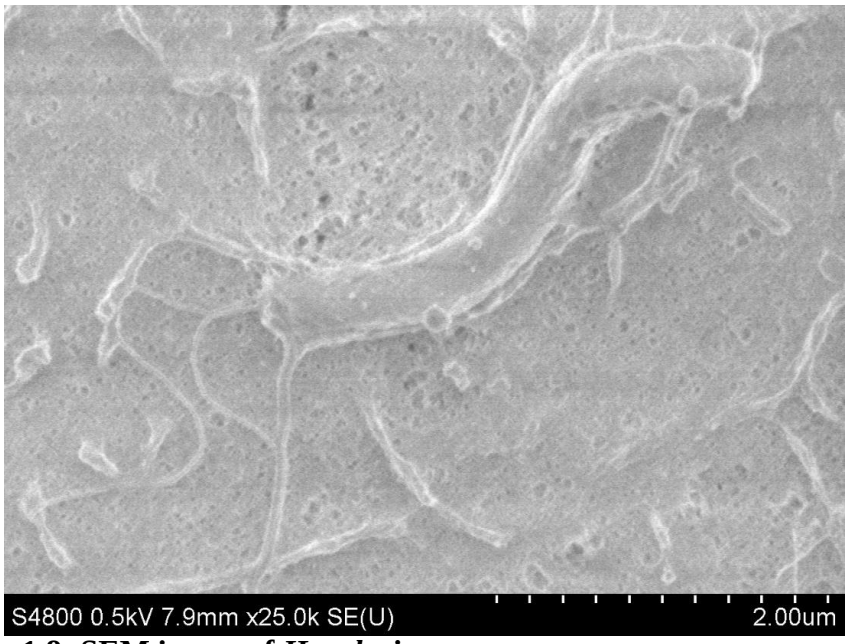


Figure 1.8: SEM image of *H. pylori*.
This image was taken on a Hitachi S4800 Scanning electron Microscope under conditions of 0.5k V for accelerating voltage with 10µA of current. Aperture was 2µm for a working distance of 8 µm, and the detector used was secondary emission upper. This image was shared by Dr Llinos Harris.

Colonies on agar plates are small, round and translucent. The laboratory conditions for growth of *H. pylori* are reduced oxygen and rich media. *H. pylori* is associated with human disease, with hazard of ingestion, and was therefore classified as a BSL-2 (biosafety level 2) pathogen. Because of this, important rules regarding hygiene and safety are applied while culturing this microorganism (Blanchard and Nedrud 2012).

1.2.3 *H. pylori* niche

H. pylori can grow at a pH ~5.1, which is the pH of the gastric content immediately after food ingestion (Rhee, Park, and Cho 2014). The bacterium can therefore resist these acidic conditions when ingested with food, allowing it to colonise the host's stomach. *H. pylori* is one of the few microorganisms that can thrive in the human stomach (Bik et al. 2006). Laboratory strains of *H. pylori* obtained from humans can be adapted to colonise other species, but it is very rare to find it naturally in non-human hosts. Other species of *Helicobacter* specialise in colonisation of other animals, but *H. pylori* is naturally found almost exclusively in humans. The composition of the gastric microbiota evolves alongside the *H. pylori* colonisation level and the acidic production of the stomach, with a reduced diversity in *H. pylori* positive stomachs due to the strong dominance of *H. pylori* in the population (Wroblewski and Peek 2016; Andersson et al. 2008), and disease develops following colonisation.

1.2.4 Diagnosis of *H. pylori* infection

Diagnostic methods for *H. pylori* infection can be divided into two categories: invasive and noninvasive. A large number of diagnostic tests have been developed but there is currently no gold standard based on a single test. A combination of culture and histopathological analysis, both obtained from biopsy samples, is an accepted standard in clinics (Cosgun et al. 2016).

Invasive methods include culture (sensitivity between 55% and 56% and specificity of 100%), rapid urease test (sensitivity over 75% and specificity over 84%), polymerase chain reaction (PCR) (sensitivity over 75% and specificity over 84%), fluorescence *in situ* hybridization (FISH) and histopathology (sensitivity over 66% and specificity over 94%) (S. K. Patel et al. 2014). All these methods require an endoscopy. Culture from human gastric biopsies is useful for testing antibiotic resistance and choose the correct treatment, but it is not usable for detection, as only 50 to 70% of infected biopsies will be detected as positive in culture, resulting in poor sensitivity (Loffeld et al. 1991). PCR-based methods and rapid urease test appear as the best alternatives for detection of *H. pylori* among invasive methods. The rapid urease test presents the advantages of a low cost and a fast diagnosis. However the influence of the bacterial density can affect the sensitivity (Nishikawa et al. 2000). Noninvasive methods

include urea-breath tests (sensitivity over 75% and specificity over 77%), whole blood serological tests through Immuno-globulin G enzyme-linked immunosorbent assay (ELISA) or finger-prick (World Gastroenterology Organisation Global Guidelines 2010) (sensitivity of 85% and specificity of 79%) (Rao et al. 2001), stool antigen test (SAT) (sensitivity over 67% and specificity over 65%) (S. K. Patel et al. 2014). Urea-breath test and serological tests are good noninvasive tests with high sensitivity but the specificity is below the specificity obtained with invasive tests (Cosgun et al. 2016; Xie et al. 2009). Moreover, urea-breath test can not be used if the patient received antibiotics or bismuth during the month preceding the test.

Verified infection with *H. pylori* leads to a choice of empirically prescribed treatments, which vary according to the geographic region, the prevalence of *H. pylori* infection, the presence of certain virulence factors, and resistance to antibiotics. Often a general approach of a proton pump inhibitor, a macrolide and a β-lactam for 7-10 days initiates therapy.

1.2.5 Treatment of *H. pylori* infection

The current guidelines state that an infection must be dealt with as soon as it is detected (Malfertheiner et al. 2017). The standard treatment has been, for the last two decades, triple therapies combining proton pump inhibitors (PPI), amoxicillin and clarithromycin or metronidazole. Eradication rates were originally high (>90% during the 90's), but they decreased during the following years, falling to lower than 70% (C.-C. Huang et al. 2017). This failure is caused by a worldwide increase of resistance to clarithromycin. Alternative strategies have therefore been proposed to overcome this resistance issue, such as quadruple therapy combining PPI, bismuth and two antibiotics (C.-C. Huang et al. 2017). A vaccine would be an efficient solution to prevent and reduce this global burden. However, there is none available yet, and an effective vaccine must protect despite the high variability of strains. Recently, a promising vaccine trial was completed in China (Zeng et al. 2015). The prophylactic vaccine tested in phase 3 was based on urease B subunit fused with heat-labile enterotoxin B subunit. Results showed that this vaccine was effective, safe and immunogenic, but the cohort was limited to a single geographic region and longer follow-up would be required to link the vaccine with *H. pylori* related disease incidence. Development of this vaccine has been discontinued. Increased global efforts are needed to build on these results (P. Sutton and Boag 2018).

1.2.6 *H. pylori* pathogenesis

Despite the majority of cases being asymptomatic, a wide diversity of disease symptoms are caused by *H. pylori* (Peek and Blaser 2002; Algood and Cover 2006; M. J. Blaser and Atherton 2004), amongst them gastric cancer (GC). The causal link between *H. pylori* and GC is strong, according to the Bradford Hill criteria (Bradford Hill 1965). *H. pylori*'s ability to provoke symptoms in its human host is a complex system involving interactions between the bacteria, its host and the environment. This thesis focuses mainly on the bacterial aspects. *H. pylori* virulence and pathogenicity factors have been a major research interest since the organism's discovery in 1984 (B. Marshall and Warren 1984) with the addition of *H. pylori* virulence factors to a virulence factor database (Table 1.1). However, those listed are only the virulence factors of *Helicobacter pylori* with supporting evidence for their role in virulence. Many others may exist, and to date may be genes with 'unknown function'.

The review of *H. pylori* pathogenesis is organised into the major processes involved during interaction with the host: colonisation, motility, adhesion, cell vacuolation, cytotoxicity, inflammation and evasion from the host immune system (Figure 1.9).

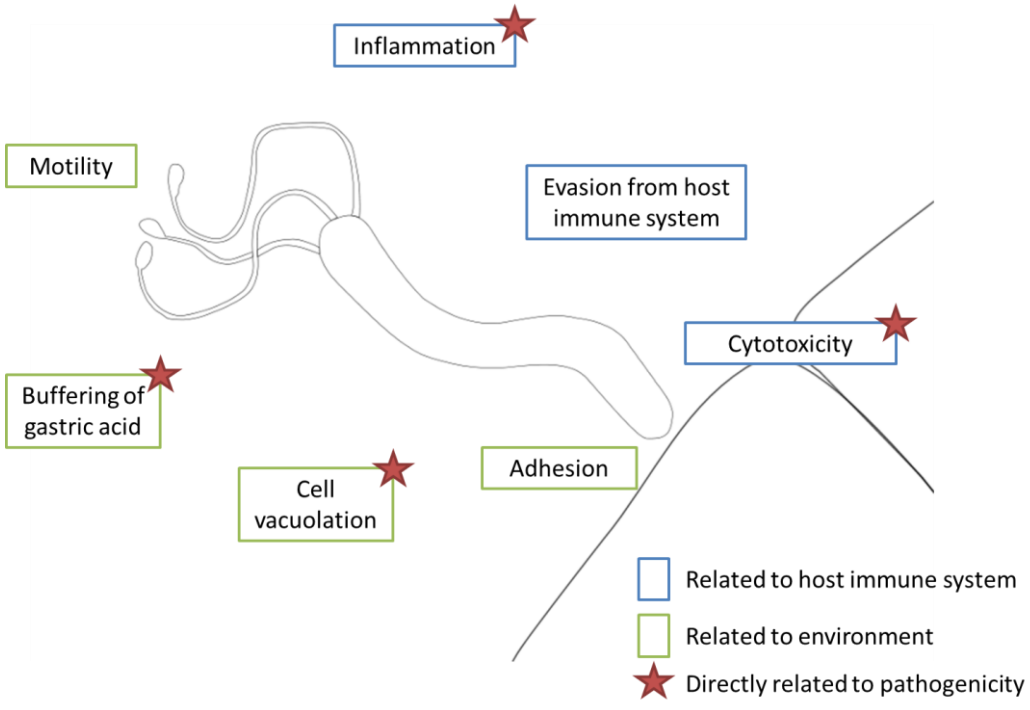


Figure 1.9: Summary of virulence factors linked to *H. pylori* pathogenesis.

Table 1.1: Main virulence factors in *H. pylori* and genes associated.
This table is based on the Virulence Factors Database (VFDB 2017) consulted on 31/07/2017.

Process	**Virulence factors**	**Full name / description**	**Related genes**
Adhesion	BabA	blood group antigen binding adhesin	*babA*/*hopS ; babB*/*hop*
	HopZ	*H.pylori* outer membrane protein	*hopZ*
	SabA	sialic acid-binding adhesin	*sabA*/*hopP*
Inflammatory activity Evasion from host immune system Adhesion	LPS	Lipopolysaccharide	*gluE*; *gluP*; *kdtB*; *lpxB*
Buffering of gastric acid	Urease	Urease	*ureA*; *ureB*; *ureE*; *ureI*
Evasion from host immune system	Lewis antigen	Lewis antigen	*futA*; *futB*; *neuA*/*flmD*;
Motility	Flagella	Flagella	*flaA*;*flaB*; *flgE_1*
Inflammatory activity	HP-NAP	neutrophil activating protein	*napA*
	OipA	outer inflammatory protein	*hopH*
Cytotoxicity	T4SS	Type IV secretion system	*cag1*; *cag2*; *cag3*; *cagI* *cagN*; *cagP*; *cagQ*; *cag* *virB11*; *virB2*/*cagC*; *vi* *virB7*/*cagT*; *virB8*/*cag*
Cell vacuolation	VacA	vacuolating cytotoxin A	*vacA*
Cytotoxicity	CagA	cytotoxin-associated antigen	*cagA*

1.2.6.1 Colonisation

The first stage of *H. pylori* infection involves colonisation of the stomach. This environment is too harsh for most bacteria to colonise, but the mucus can be a shield for *H. pylori*, protecting the bacteria from acidic juice and host defense factors. The process of colonisation of the stomach by *H. pylori* requires three functions: motility, adhesion and buffering of gastric acid (Figure 1.10). These three functions will be described individually in the following subsections.

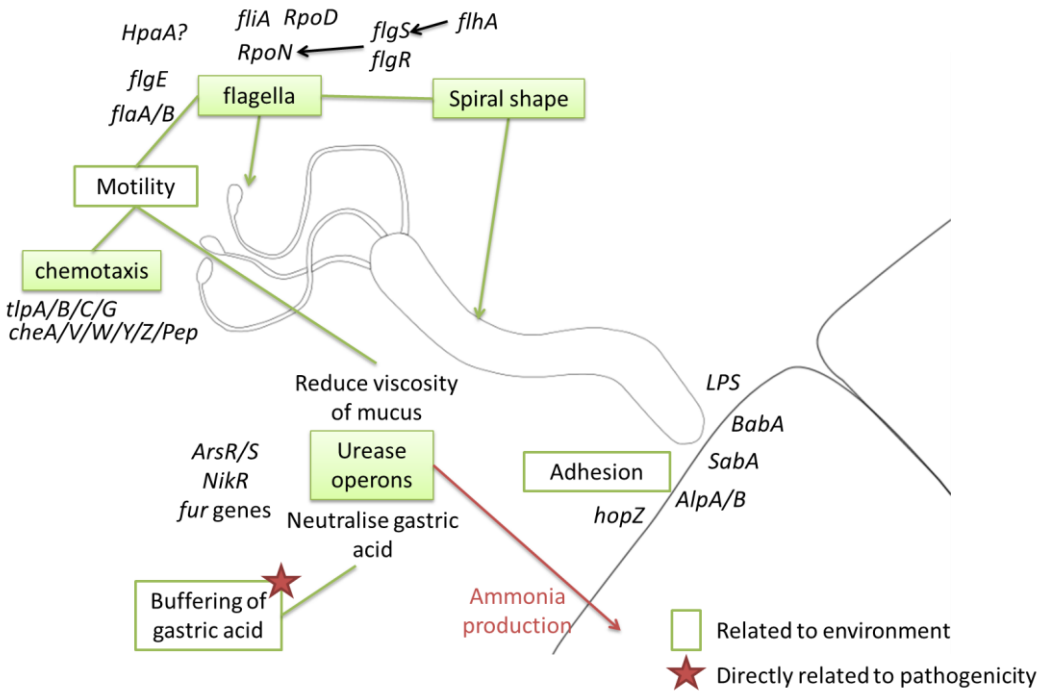


Figure 1.10: Overview of virulence factors involved in *H. pylori* colonisation of the stomach.

1.2.6.1.1 Motility

In *H. pylori*, motility is facilitated by three different functions: flagella, chemotaxis (*tlpB* and *HP1043*) and manipulation of the mucus layer viscosity surrounding the micro-organism (*nixA*, Urease operons). The helical shape of *H. pylori* is thought to enhance motility, and consequently aids the bacteria in penetrating the gastric mucus layer. In addition, *H. pylori* possess flagella (usually 4 to 6 per bacteria), which are essential for colonisation (Eaton et al. 1996). Only 3 genes linked to motility are described in *H. pylori* in the Virulence Factors database (VFdb) (VFDB 2017): *flaA*,

flaB and *flgE_1*. Specifically, *flaA* encodes for the major flagellin, composing the filaments in pair with the minor flagellin encoded by *flaB*. The *flgE_1* gene encodes for the hook proteins. However, motility is complex and further important genes are found in the literature (O′Toole, Lane, and Porwollik 2000). For instance, *rpoD*, *rpoN* and *rliA* are 3 sigma factor genes regulating expression of flagellar genes, while *flgS* (sensor kinase) and *flgR* (response regulator) regulate the transcription of genes alongside *rpoN*. FlhA is also known to bind to FlgS, adding to the complexity of motility function (Tsang et al. 2015). A flagellar sheath protein identical to HpaA was also identified in *H. pylori*, but its function is unclear (A. C. Jones et al. 1997). Other genes important in motility and flagellar biogenesis are *flbA* (Tsang et al. 2015), *fliI* (Jenks et al. 1997), *flaG*, *flmH*, *fliD*, *flgK*, *flgL*, *motA*, *motB*, *fliM*, *fliN* and *fliG* (O′Toole, Lane, and Porwollik 2000). Finally, motility may also be facilitated by the ability of *H. pylori* urease to lower the viscosity of mucus around the bacteria (Eaton and Krakowka 1994). Alongside motility, chemotaxis is also important in order to navigate through the pH gradient and efficiently penetrate the mucus layer. Genes involved in *H. pylori* chemotaxis include *tlpA/B/C/D, cheA/V/W/Y/Z/Pep* (O′Toole, Lane, and Porwollik 2000).

1.2.6.1.2 Adhesion

Once *H. pylori* has entered the mucus it needs to adhere to host cells. This is where many outer membrane proteins and adhesins have a role. Amongst them BabA (blood group antigen binding protein) (Aspholm-Hurtig et al. 2004), AlpA/B (adherence associated lipoproteins) (Senkovich et al. 2011), SabA (sialic acid binding and adhesion) (Unemo et al. 2005) and HopZ (outer membrane protein) (Peck et al. 1999). Lipopolysaccharide (LPS) and AlpA/B are also crucial, although independant (Odenbreit, Faller, and Haas 2002). All these surface components are potential targets for host immune defence. However, *H. pylori* is able to evade most of this defence through various mechanisms that we will review later in this introduction (section 1.2.6.5).

1.2.6.1.3. Buffering of gastric acid

The buffering of gastric acid participates in the motility function, but also in colonisation in a larger sense, by promoting survival of the bacterium. Modifications

of the gastric acid secretion, provoked by *H. pylori* through IL-1β, also has a more direct effect on pathogenicity through its cytotoxic effect on the host cells (Takashima et al. 2001).

Production and excretion of urease neutralise the low pH around the bacteria, through the *arsR/S, nikR, fur* genes (M. D. Jones et al. 2015; Mobley, Hu, and Foxal 1991). This neutralisation of pH produces ammonia, which has direct cytotoxic effects on the host epithelium. Cytotoxicity is not limited to one gene associated with the colonisation process and so further genes will be included in the following 3 sections.

1.2.6.2 Cell vacuolation (VacA)

Exotoxins, such as VacA (Vacuolating cytotoxin), can lead to gastric mucosal injury through cell vacuolation (Telford 1994). This cell vacuolation increases cell permeability and therefore facilitates the supply of essential nutrients (Cover and Blaser 1992; Iwamoto et al. 1999; Tombola et al. 2001). However, VacA also generates a host immune response through production of mast cell-derived proinflammatory cytokines resulting in chemotaxis and activation (Supajatura et al. 2002). VacA is unique to *H. pylori*, and the gene coding for it is present in all strains, but with significant polymorphisms (J C Atherton et al. 1995). Three highly variable polymorphic regions are identified in *vacA*: The signal sequence region found in two versions (s1, s2), the intermediate region found in three different versions (i1, i2, i3), and the mid region found in two different versions (m1, m2) (Junaid et al. 2016; M. J. Blaser and Atherton 2004). Mature VacA toxin is composed of two domains: a N-terminal p33 domain and a C-terminal p55 domain, linked by a protease-sensitive loop (Junaid et al. 2016).

1.2.6.3 Cytotoxicity (CagPAI)

The type IV bacterial secretion system (T4SS) is used by some *H. pylori* strains to inject effectors such as CagA (Cytotoxin associated gene A protein) or Tip-α (TNF-α inducing protein) into host cells, provoking multiple effects (Figure 1.11).

The cag pathogenicity island (CagPAI) is the most studied virulence factor in *H. pylori*. This 40-kb deoxyribonucleic acid (DNA) insertion element contains genes encoding the proteins forming this T4SS and the CagA protein that are secreted into host epithelial cells (S Odenbreit et al. 2000). The risk of developing gastric cancer

(GC) is higher for patients infected with CagA positive strains compared to strains lacking CagA (Parsonnet et al. 1997; J. Q. Huang et al. 2003). Adherence factors such as blood group antigen binding proteins encoded by the *babA* gene have also been shown to increase the delivery of CagA into the cell (Ishijima et al. 2011). Other important genes associated with the CagPAI include *tnpA* and *tnpB*, were also identified more frequently in strains isolated from patients with GC compared to other strains. This highlights its link to GC (Abadi et al. 2014).

CagPAI is also involved in evasion from host immune system function, by generating resistance to phagocyte killing and persistence within macrophages (Ramarao and Meyer 2001; Lina et al. 2014). CagPAI positive strains also tend to induce higher levels of IL-8 (Akopyants et al. 1998; Brandt et al. 2005; Fischer et al. 2001; Li et al. 1999; Segal et al. 1997), IL-10 and IL-12. They also increase the levels of nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) (Brandt et al. 2005; Li et al. 1999). Exposure to CagPAI positive strains show an activation of the expression of c-fos and c-jun, forming activator protein 1 (AP-1), a multipotential transcriptional factor associated with varied cytokines and chemokines, which can lead to GC (Mitsuno et al. 2001).

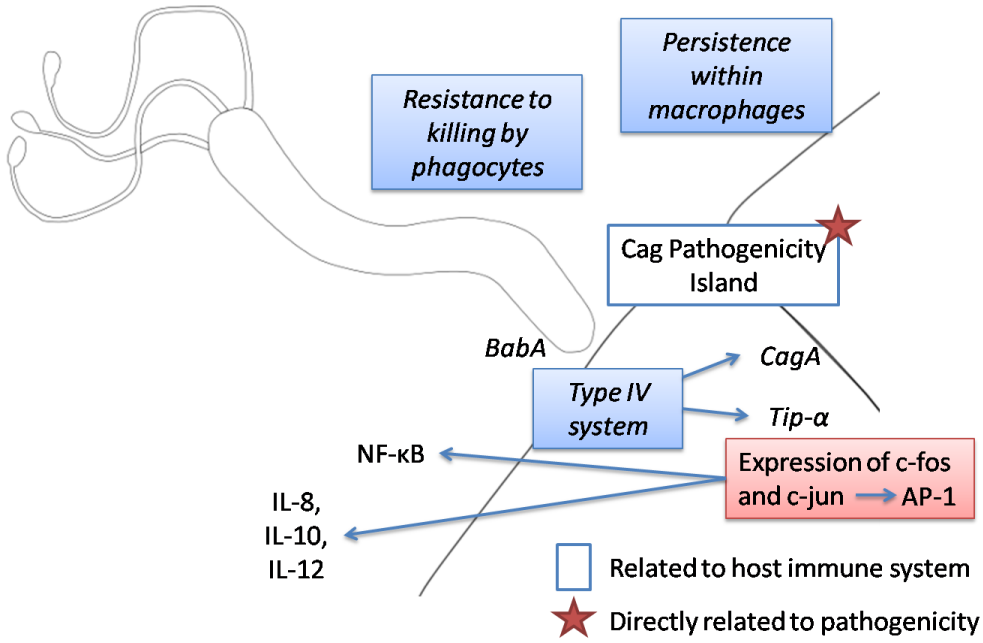


Figure 1.11: Overview of the role of CagPAI in *H. pylori* pathogenesis.

1.2.6.4 Inflammatory activity

Symptoms caused by *H. pylori* are variable, due to the variety of inter-related mechanisms. Inflammation (neutrophils, cellular exudate etc) is frequently observed in infected stomachs (Figure 1.12). LPS, OipA (outer inflammatory protein), CagA and neutrophil-activating protein (HP-NAP or NapA) (D. J. Evans et al. 1995; Satin et al. 2000) are all produced by *H. pylori* during infection of the host stomach. LPS from *H. pylori* is less proinflammatory than LPS from most other Gram-negative species (Moran and Aspinall 1998; Pérez-Pérez et al. 1995), due to a specific phosphorylation pattern and acylation in lipid A (Muotiala et al. 1992; Chmiela, Miszczyk, and Rudnicka 2014). SabA, an outer membrane protein, also participates in the recruitment of neutrophils (Unemo et al. 2005) which generate reactive oxygen species, leading to bacterial killing but also DNA damage in host cells, ultimately leading to gastric cancer.

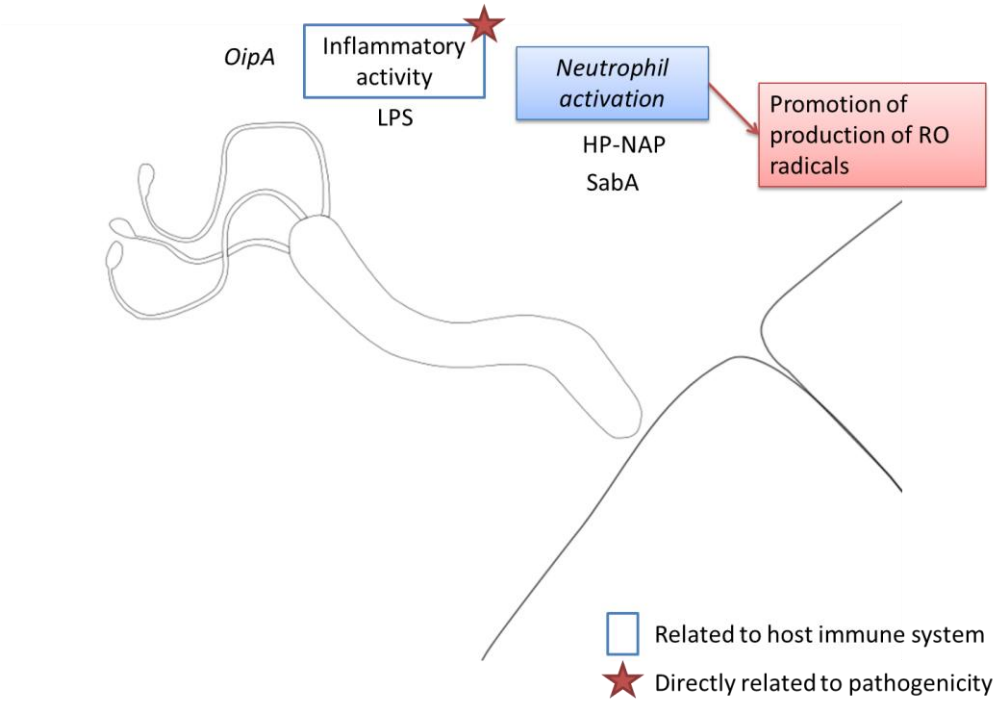


Figure 1.12: Overview of the inflammatory activity caused by *H. pylori* pathogenesis.

1.2.6.5 Evasion from host immune system

Inflammation is not only triggered by *H. pylori*, it is also regulated, resulting in a form of escape of the bacteria from the host immune system. For instance, the increase of regulatory T-cells provoked by *H. pylori* through interactins with dendritic cells,

results in a reduction of the TH17 immune response (J. Y. Kao et al. 2010). *H. pylori* can also evade the host immune system using a variety of virulence factors (Figure 1.13).

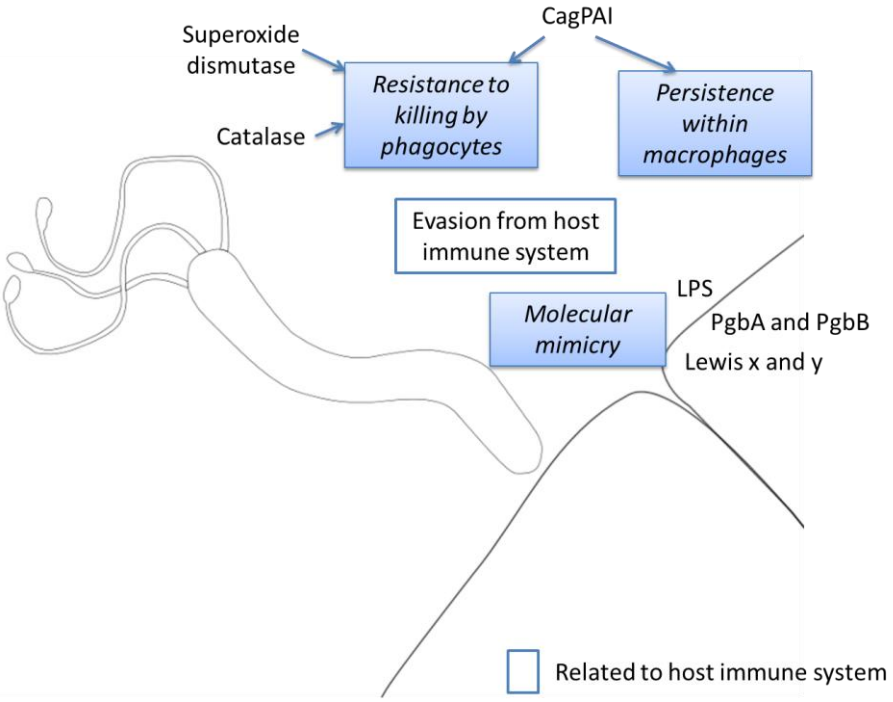


Figure 1.13: Overview of the mechanisms used by *H. pylori* to evade the host immune system.

As previously described, CagPAI is essential to immune evasion through persistence within macrophages, resistance to killing by phagocytes (Ramarao et al. 2000; Ramarao and Meyer 2001) and via downregulation of β-defensins (Bauer et al. 2012; S. R. Patel et al. 2013). Resistance to phagocytes killing is also achieved through two enzymes: superoxide dismutase (Spiegelhalder et al. 1993; Seyler, Olson, and Maier 2001) and catalase (S Odenbreit, Wieland, and Haas 1996; Ramarao, Gray-Owen, and Meyer 2000). Molecular mimicry of molecules Lewis x and y also helps the bacteria escaping the immune system, through expression of LPS O antigens (Aspinall and Monteiro 1996; Monteiro et al. 1998; Bergman et al. 2006). These Lewis antigens, through interaction with Macrophage inducible C-type lectin (Mincle), activate anti-inflammatory cytokine production (Devi, Rajakumara, and Ahmed 2015). The specifically low activity of *H. pylori* LPS is also a mechanism for the bacteria to keep a low profile by avoiding recognition by Toll-like receptors (TLR) (Cullen et al. 2011). *H. pylori* flagellin evades recognition by toll-like receptor 5 (TLR-5) (Lina et al. 2014; Gewirtz et al. 2004). VacA also plays a role in inhibition of T cells

activation (Boncristiano et al. 2003; J. M. Kim et al. 2011; M. Oertli et al. 2013). proteins like PgbA and PgbB bind plasminogen, coating the surface of the bacteria with host proteins (Jönsson et al. 2004). The capacity of *H. pylori* to control the balance between pro and anti-inflammatory responses, through all the mechanisms mentioned above, is the key to its persistence.

1.2.7 Beneficial effects of *H. pylori* infection

Despite the complications previously described (1.1.4), *H. pylori* colonisation can have beneficial effects on its host. For instance, the tuning down of immune system mentionned in the previous section is thought to be responsible for a protective effect of *H. pylori* against atopic diseases (Lionetti et al. 2014) such as allergic asthma (Mathias Oertli and Müller 2012; Arnold et al. 2011; D'Elios and Bernard 2010; Y. Chen and Blaser 2008), allergies (Hussain et al. 2016), eosinophilic oesophagitis (von Arnim et al. 2016) or conventional multiple sclerosis (LI et al. 2007). Therapeutic application of *H. pylori* extract has even been considered to reduce allergic airways disease (van Wijck et al. 2018). Other positive effects of *H. pylori* infection include a lesser risk for Barrett's oesophagus (Thrift et al. 2012) and obesity (O'Connor, O'Morain, and Ford 2017). Differentiating between 'beneficial' strains and those likely to cause cancer suggests better targeting of patients at risk.

1.3 Genomics of *H. pylori*

1.3.1 First *H. pylori* genome sequenced

The first complete genome sequence of *H. pylori* was published in 1997 (Tomb et al. 1997). The strain sequenced was named 26695, and was isolated from a patient in the United Kingdom suffering from gastritis. This first genome was 1,667,867 base pair (bp) with 1590 predicted coding sequences, and an average G+C content of 39%. These predicted coding sequences have since then been further described and the list has been amended. The list we will use was obtained from the National Center for Biotechnology Information (NCBI) in 2014, and contained 1573 genes (or loci). The method used for this sequencing was whole-genome shotgun sequencing with the Sanger method. Analysis of this first sequence of *H. pylori* was the basis for much of

the available knowledge on *H. pylori* pathogenesis, acid tolerance, antigenic variation and micro-aerophilic character.

The 26695 genome is often used as a reference strain (and we do the same in this thesis), as its genes were well described and referenced in the literature. Furthermore this strain is the reference strain used on the PATRIC (Pathosystem Resource Integration Center) database (PatricdB 2017b). All genes from the 26695 strain are annotated with the nomenclature HP followed by 4 digits.

1.3.2 Multi-Locus Sequence Typing (MLST)

Understanding the high variability of outcomes resulting from *H. pylori* infection depends upon an increased knowledge of the population genetic structure and the related phenotypic differences between isolates. Among the first DNA sequence based techniques that brought a deeper understanding of *H. pylori* population diversity was multi-locus sequence typing (MLST). MLST typing is based on the analysis of fragments from 7 house-keeping genes. Therefore, a simple PCR amplification is sufficient for MLST typing of isolates. The genes used in *H. pylori* MLST are *atpA, efp, mutY, ppa, trpC, ureI* and *yphC* (K. Jolley 2017; Achtman et al. 1999). Unlike *Campylobacter* species, *H. pylori* is not organized into clear cloud clonal complexes, but into clusters of related lineages, depending largely on the geographical origin of the samples (Achtman et al. 1999).

1.3.3 Whole-genome based methods

DNA sequencing methods determine the order of nucleotides in a DNA molecule. A large number of methods and technologies for sequencing were developed since the first Sanger sequencing method in 1977 (Pettersson, Lundeberg, and Ahmadian 2009; Sanger, Nicklen, and Coulson 1977). Using whole-genome sequencing (WGS) and analysis methods on *H. pylori* isolates, research has built and improved upon earlier MLST data.

1.3.3.1 Analysis of *H. pylori* genomes

Whole-genome sequencing has become faster, cheaper, and more efficient, and its application to microbiology has changed the face of research in this field. With the development of whole-genome sequencing methods, the costs and time associated

have decreased drastically (Loman et al. 2012), and the number of available *H. pylori* gene sequences have increased exponentially (Figure 1.14).

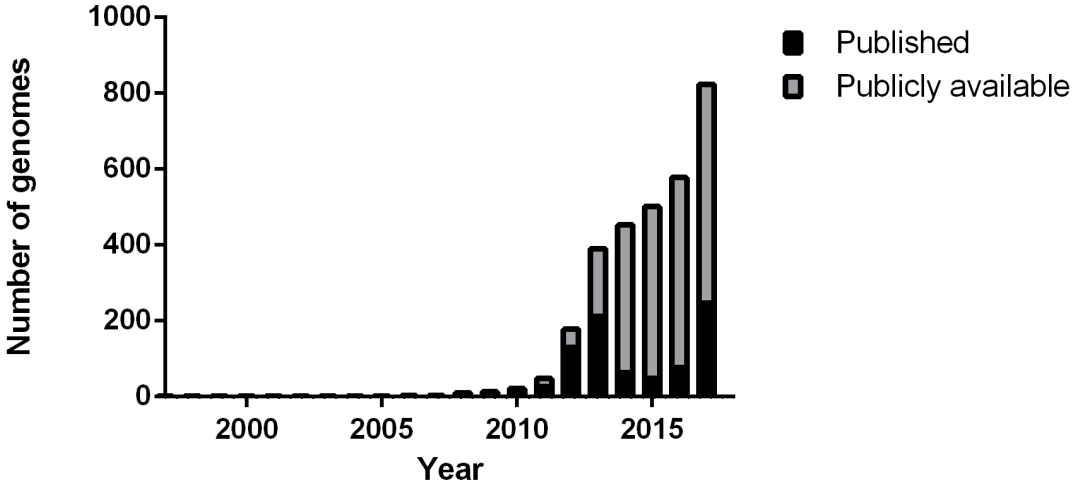


Figure 1.14: Cumulative number of *H. pylori* genomes available in NCBI from 1997 to October 2017.
Adapted from (Berthenet, Sheppard, and Vale 2016) presented in Appendix A.

At the time of writing (August 2017), 694 *H. pylori* genomes were available on the PATRIC db (PatricdB 2017a), with an average size of 1.63Mb. Sequences of *H. pylori* are now available from almost all areas of the world, and they often come with patient data, which allow researchers to perform detailed analysis of, not only the genes from MLST, but the whole genome.

1.3.3.2 Core and Accessory genome

Comparison of whole genomes of *H. pylori* reveal two types of genes, based on their presence in the dataset of interest: They are accessory and core genes (Uchiyama et al. 2016). It is important to remember that these definitions are dependent on the dataset studied. A gene can be a core gene in one dataset, and an accessory one in a different dataset including different strains.

The core genome comprises all the genes that are found in all the isolates from the dataset of interest. For a large dataset with genomes splitted into more than one contig, it is common to define core genome as all the genes found in at least 90% of the isolates. The size of the core genome will vary according to the size and variability of the dataset (Figure 1.15).

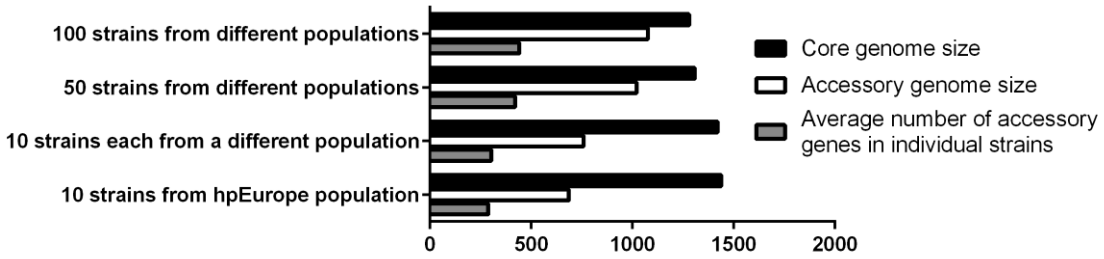


Figure 1.15: Variations in size of the core and accessory genome according to the dataset studied.
Strains used in this example are from the dataset used in Chapter 5 and Core genome is defined as all genes shared by 100% of the strains.

The accessory genome comprises all the genes that are found in at least one but not all the isolates from the dataset of interest. It often contributes to the acquisition of unique traits by *H. pylori* strains. The size of the accessory genome will vary according to the size and variability of the dataset (Figure 1.15).

1.3.3.3 A systematic approach to genome analysis

Among the most promising techniques for studying the bacterial genome are Genome wide association studies (GWAS), which were recently applied to *Campylobacter*, a related species within the epsilonproteobacteria (Sheppard et al. 2013). In this method, DNA sequence that is over-represented in one phenotype group compared to another is identified, in order to link accessory and core genome variations with the studied phenotype. This has the distinct advantage that sequence variation associated with phenotypes such as virulence can be identified without pre-selection bias. A GWAS will be carried out in Chapter 5 on a large dataset of *H. pylori*, aiming to identify genomic elements associated with gastric cancer. Two GWAS methods are used in this thesis: a ClonalFrame based method (Didelot and Falush 2007), and bugWAS (Earle et al. 2016). Both methods have been recently used in bacteria (Monteil et al. 2016; Méric et al. 2014; Sheppard et al. 2013; Suzuki et al. 2016), and take recombination and population structure into account which makes them suitable for bacterial genomes. However, the ClonalFrame based method relies on clonal complexes, and requires pairs of strains to be selected, resulting in a reduction in the number of strains included in the analysis. The nature of *Helicobacter pylori* population structure is also challenging for the ClonalFrame based method, as it does

not form clear clonal populations. The bugWAS method also has limitations, such as the risk of confounding true associations with the result of environment or sampling bias. Therefore both methods were used in parallel.

1.3.4 *H. pylori* genome variability

H. pylori has one of the highest known recombination rates, and presents the highest genetic variability among pathogenic bacteria (Figure 1.16), with an average number of alleles per locus of 11.2, resulting in a mean genetic diversity of 0.735 (Go et al. 1996).

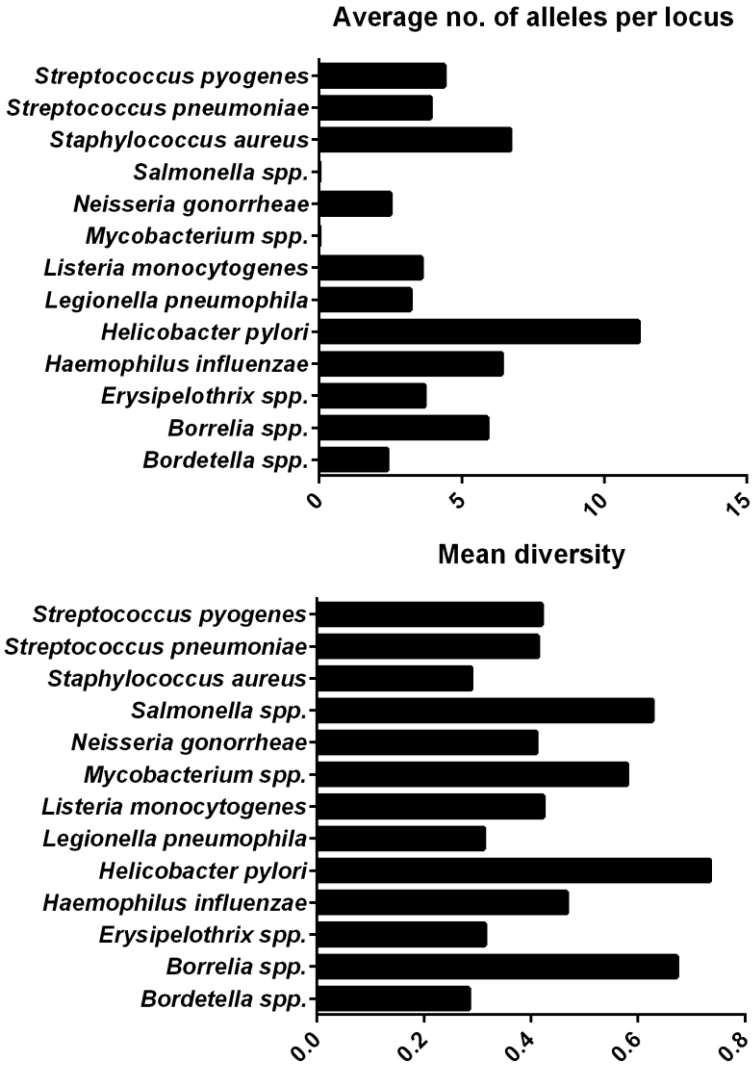


Figure 1.16: Genetic diversity among electrophoretic types in representative species of pathogenic bacteria.
Adapted from (Go et al. 1996).

This rapid evolution and the resulting genomic variability lead to resistance to treatment, challenging scientists toward new ways of eradicating infection, and make the development of an efficient vaccine difficult. A good example of diversity can be seen by the fact that two strains of *H. pylori* found in different patients (or even in patients from a same family) are often extremely different (Kivi et al. 2007). Different strains of *H. pylori* can also colonise a same host (Kibria et al. 2015; J. W. Kim et al. 2004; Ben Mansour et al. 2016). The *H. pylori* genome is incredibly diverse, with variation in genome size (Dong et al. 2014), gene presence, gene order, and allelic profile (Thorell et al. 2016). Despite most strains having an average number of genes of about 1637.5 (Dong et al. 2014), the number of genes shared by all strains (core genes) from a large size sample of strains (n=100) is closer to 1200.

The high level of structuring in bacterial populations, the clonal nature of cell division, together with most processes needing many genes, make it difficult to identify genomic elements that are directly related to specific cellular function from a background of genes that are simply inherited by clonal decent.

1.3.4.1 Variability linked to geography

Helicobacter pylori is present in all inhabited continents, but isolates differ in different parts of the world (Ierardi et al. 2013). Geographic variations in the prevalence of *H. pylori* are known, but variations are also observed in the genome (Falush et al. 2003; Linz et al. 2007). Although *H. pylori* infection is considered a pre-requisite for development of most gastric cancers, there is no strict correlation between *H. pylori* prevalence and gastric cancer rates. For instance, *H. pylori* prevalence is extremely high in Africa, but gastric cancer is rare. This is known as the African enigma (Holcombe 1992). The reason for this difference is unknown so far, but some research showed a genetic instability in the host differing between European and African populations that could be part of the answer (T. E. Buffart et al. 2011). Resistance to antibiotics is different in each part of the world, which leads to variations in the recommended first line treatment (Ierardi et al. 2013). This is due to selection pressure caused by antibiotic treatment, or intake of antibiotics from other sources (Ling et al. 1996). However, there are other differences which are not linked to antibiotic resistance. This is the case for the presence or absence of the CagPAI island (Kumar, Kumar, and Dixit 2010; Olbermann et al. 2010; Yakoob et al. 2009). This island of genes is more prevalent in strains coming from East Asia than from

Europe (Maeda et al. 1998; Yakoob et al. 2009). Variations in the *cagA* gene, among CagPAI positive strains, are also observed. Specifically, the *cagA* gene is composed of repetitions of EPIYA motifs. Different EPIYA motifs are observed, and the motifs present and their number of repetitions vary following geographic patterns. For instance, East Asian strains show in large majority the motifs: A-B-D where other strains present the motifs A-B-C, A-B-C-C. or A-B-C-C-C (Y Yamaoka et al. 2000; Xia et al. 2009; HATAKEYAMA 2017). Variations in the *VacA* gene also match a geographic distribution (Diaz et al. 2005; Van Doorn et al. 1999; Maeda et al. 1998).

1.3.4.2 Variability linked to pathogenicity

Variations in the CagPAI island and *VacA* gene are not only linked to geography, but have been linked to pathogenicity. For instance, the risk of developing complications due to infection by *H.pylori* is higher for patients infected with CagA positive strains compared to strains lacking CagA (Ahmadzadeh et al. 2015; Rizzato et al. 2012; Yamazaki et al. 2005; Khatoon et al. 2017; Parsonnet et al. 1997; J. Q. Huang et al. 2003). Allelic variations in *vacA* are also associated with pathogenicity (Yamazaki et al. 2005). Combination of *vacA* types s1 and m1 are associated with GC more strongly than other types (Kidd et al. 1999; Miehlke et al. 2000). Other genes show a link with gastric cancer or with other gastric diseases, for instance for *babA*, *oipA*, *dupA* and *iceA* (Miftahussurur and Yamaoka 2015; S. Y. Kim et al. 2001).

1.3.4.3 Mechanisms behind variability

Two forces are driving the variability in bacterial genomes. Replication errors or damage which generate point mutations, rearrangements or deletions of various sizes on one hand, and horizontal transfer which incorporate genetic material from an external source by recombination. The relative rates of those two forces are shaping the bacterial population genetic structure (Guttman and Dykhuizen 1994).

Bacterial mutation rates are generally low. However under strong selective pressure, such as antibiotic use or a drastic change in the environment, hyper-mutator phenotypes can emerge in the population, with elevated mutation rates. Most of the mutations observed are deleterious (Kimura 1967), but some can induce variability in the genome that will help the bacteria interact with its environment. For instance, phase variation is a famous mechanism consisting in a simple change in the number of

repeat of a single nucleotide or a pair of nucleotides that will result in a modification in the frame of reading (Bergman et al. 2006; Appelmelk et al. 1999; G. Wang et al. 2000). This mechanism allows for a quick adaptation of the gene between on and off versions.

Horizontal gene transfer (HGT) or genetic exchange can be divided into two mechanisms: Homologous recombination, which consist in the replacement of homologous DNA with a sequence from another organism (often from the same species) on one hand, and non-homologous genetic exchange, which consist in the introduction of DNA segments or whole genes into the bacterial genome. *H. pylori* genomes show evidence of high levels of homologous recombination compared to most other bacterial species (Dorer, Sessler, and Salama 2011; Vos and Didelot 2009).. Non-homologous recombination is also an important source of genomic diversity for *H. pylori*. Large fragments of DNA called mobile genetic elements can be transferred into the genome, such as bacteriophages, pathogenicity islands, transposons, insertion sequenes, or plasmids. They can be inserted into specific sites in the chromosome or be part of the extra-chromosomal DNA, as autonomously replicating elements. These mechanisms are central in the evolution of accessory genome. As a proof of the importance of non-homologous recombination in the evolution of *H. pylori*, one can mention the CagPAI, a pathogenicity island unique to *Helicobacter pylori* conferring virulence to the strains (Hacker and Kaper 2000; Fernandez-Gonzalez and Backert 2014).

HGT can confer novel function (e.g. antibiotic resistance), but is also of more general interest in understanding the driving forces of bacterial evolution. Rapid recombination between geographically isolated populations can lead to local genomic signatures. For example, several studies have used local signals of recent admixture between strains to describe the migration of human hosts (Nell et al. 2013; Linz et al. 2007). By co-existing for such a long time, the genomes of these two species have evolved together leaving traces of human migrations in *H. pylori* population genomes (Falush et al. 2003). The recent admixture occurring in the Americas will be investigated in Chapter 3. Geographic signatures are also observed in *H. pylori* prophages (Vale et al. 2017, Appendix B).

Admixture within *H. pylori* genomes is also a signal for microevolution. While traditional typing technologies may not be sensitive enough to detect variation between closely related isolates, such as those in a single-family transmission

network, whole-genome sequencing provides opportunities for enhanced resolution. Transmission pathways among individuals of the same family have been characterized revealing genomic adaptation to child hosts as a probable part of the infection pathway (Furuta et al. 2015). Another WGS study focused on multiple colonies isolated from a single patient, demonstrating the co-existence of different lineages and HGT between isolates from these lineages resulting in a progressive genomic convergence (Cao et al. 2015). The variations occurring during long-term infection of a mouse model with clinical *H. pylori* strains will be investigated in Chapter 4.

Finally, signatures of selection have been investigated in *H. pylori* genomes by estimating the ratio of non-synonymous to synonymous substitutions (dN/dS) in genes present in more than 90% of a 29 genome collection (Koji Yahara et al. 2016). Codons with evidence of diversifying selection (dN/dS>1) were widely distributed, accounting for ∼0.2% of the genome, and were commonly associated with gene functions of host interaction, cell surface expression and genome maintenance. Different methods are available to account for this ratio, depending on the purpose of the comparison. This association of specific functions with enrichment of non-synonymous Single Nucleotide Polymorphisms (SNP) will also be investigated in this thesis.

1.3.4.4 Remarkable strains of *H. pylori*

Since the first genome 26695 was sequenced, (Tomb et al. 1997) further strains have been isolated and sequenced. Some of them are remarkable, due to either the conditions of their isolation or the features of their genomes. The first remarkable strain highlighted is an ancient strain of *H. pylori* sampled from a 5300 year old European mummy. DNA from this strain showed no admixture between Asian and African *H. pylori*, which are commonly seen today, therefore suggesting admixture between these two populations occurred after this time in Central Europe (Maixner et al. 2016). Strains such as this one are especially important in attempts to date evolutionary events, but it is very rare to obtain isolates from ancestral human populations. For this reason, much of the work to understand past acquisition of genes is based upon inference of phylogenies and identifying lineages sharing genes that can

be traced to an acquisition event on a phylogenetic tree. This will be the basis for chapter 1.

Two strains, with their genomes published in 2015, are also of interest (Kersulyte et al. 2015). They were isolated from a Canadian arctic aboriginal community in Aklavik, Northwest Territories, Canada. The population in this region suffers from a high prevalence of *H. pylori* (Carraher et al. 2013). One of these two strains presented a new lineage of *H. pylori*, close to the hspAmerind population. A project involving the local population is on-going, which should result in a large collection of samples from this isolated population, alongside clinically-linked data. This collection will be important for studying mechanisms of evolution in *H. pylori*, and its link to gastric cancer without interference from co-infections of hosts with *H. pylori* from diverse origins.

1.4 Aims

Current treatments against *H. pylori* infection use multiple antibiotic and drug regimens. Considering the high risks of developing a cancer, this approach makes sense. However, the rise of antibiotic resistance suggests that better antimicrobial stewardship is required for the control of *H. pylori* in the clinic. The versatile nature of the *H. pylori* genome is a sign that part of the risk associated with individual strains might be predictable through sequencing of the infecting strain. The final aim of this project is to open the way towards new guidelines for treatment, based on sequencing of the infected strains, which could reduce the rise of antibiotic resistance.

Chapter 3 will investigate the genome variability of a global collection of *H. pylori* strains. Two hypotheses will be tested:

- The genomic variability of *H. pylori* strains from the Americas reflects the history of recent and ancient migrations which built the identity of these regions,
- Core and accessory genomes are evolving in a similar way.

Chapter 4 will investigate the genome variability occurring in hosts during a long-term infection, and we will verify the following hypotheses:

- A *H. pylori* strain evolves when changing from one host to another,
- A *H. pylori* strain infecting a stomach for a long time evolves alongside the development of symptoms.

Chapter 5 will then focus on a European population of strains to address a range of hypotheses:

- The GWAS method can be applied to *H. pylori* genome despite its high variability,
- Specific genomic traits in specific genes can be linked with the progression of gastric cancer (GC),
- A risk score can be built in order to target strains with a higher risk for triggering GC.

Finally, Chapter 6 will investigate phenotypic characteristics of strains in relation with their genomic variability. It will address the following hypotheses:

- Motility varies according to the pathology of the patient from which the strain was isolated,
- Immune response is triggered differently according to the pathology of the patient from which the strains was isolated,
- Some genes covary with phenotypic differences observed among strains.

2 Material and Methods

2.1 *H. pylori* strains

A large number of strains coming from different collections were used in order to perform my analyses (Figure 2.1). I will introduce them briefly and highlight their origin and characteristics. Details are available in Appendix C. Collections will be presented according to the number of strains.

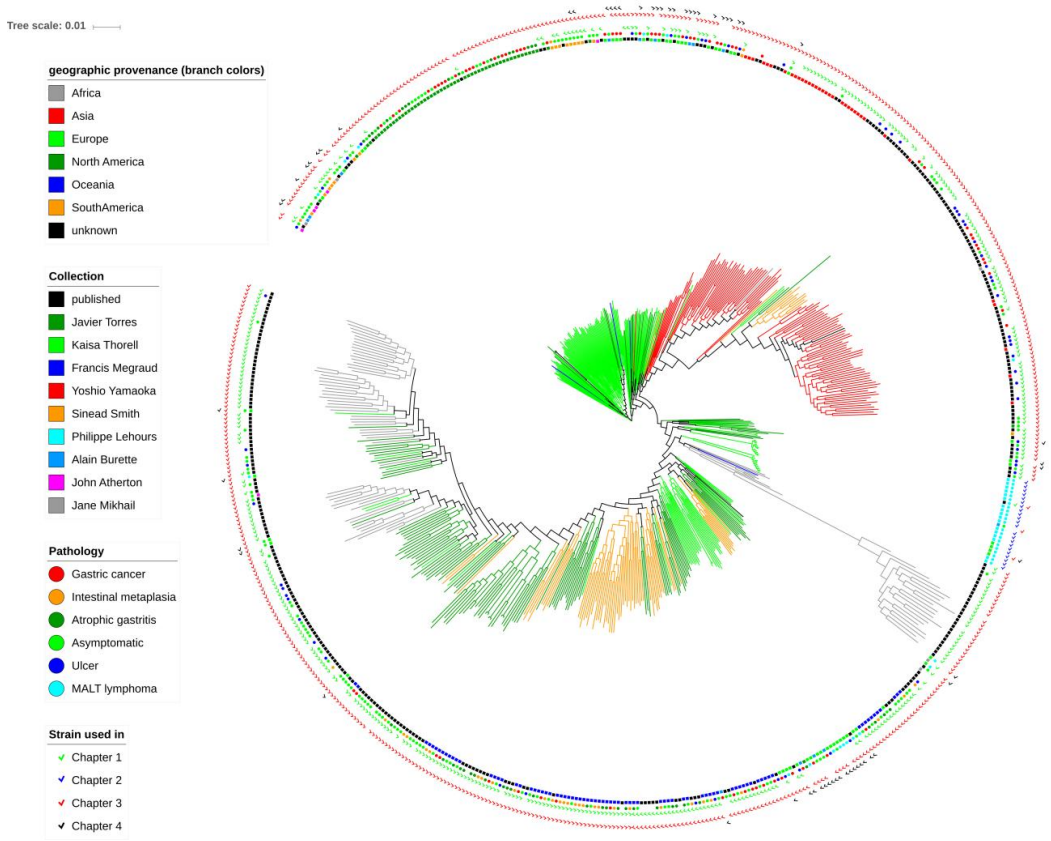


Figure 2.1: Circular View of a genomic neighbour-joining tree built with FastTree from an alignment of the 604 strains used in this thesis based on the reference strain 26695 genome.
The tree was annotated with iTOL (Letunic and Bork 2016) according to the geographic provenance, collection from which the strain was obtained, pathology associated and use in this thesis. An online version is available for this figure at this address: http://itol.embl.de/tree/13744115311634150183430 2#.

The largest collection of strains used is the publically available strains (321 strains). Those sequences were obtained from the NCBI Genbank database. They were from varied sources, which can be traced back from the publications in which they were mentioned. Geographic origin of the samples was global:

- 86 strains were isolated in Africa,
- 74 were isolated in Asia,
- 17 were isolated in Europe,
- 106 were isolated in North America,
- 4 were isolated in Oceania,
- 33 were isolated in South America,
- 1 was from unknown geographic origin.

Pathologies associated with those strains comprised:

- asymptomatic or non atrophic gastritis (NAG) (76 strains),
- atrophic gastritis (15 strains),
- gastric cancer (GC) (15 strains),
- intestinal metaplasia (IM) (9 strains),
- MALT lymphoma (ML) (4 strains),
- ulcer (45 strains),
- various complex, rare or undefined pathologies (27 strains).

Pathology related information was not available for the 130 remaining strains.

The second largest collection (79 strains) was shared by Javier Torres. 35 strains were from North America (Mexico) and 44 were from South America (Colombia). Isolates were already sequenced and assemblies were directly shared with us. Pathologies associated comprised:

- asymptomatic (28 strains),
- atrophic gastritis (14 strains),
- GC (14 strains),
- IM (21 strains),
- ulcer (2 strains).

This collection was used in both Chapter 3 and Chapter 5, and sequences were published in the publications linked to these two chapters.

Another large collection (56 strains), used exclusively in Chapter 5, was shared with us by Kaisa Thorell. All the strains were originating from Swedish hospitals, and were part of a large case-control study (Enroth et al. 2000). Extracted DNA sent to us for sequencing comprised strains associated with gastric cancer (20 strains), non-atrophic gastritis (20 strains) and atrophic gastritis (16 strains).

53 strains were generously shared with us by Francis Megraud. Cultures samples were sent to us, from which we were able to extract and sequence DNA. Genomes of these strains were used in all chapters, and some of the culture samples were used in chapter 6. Host pathologies associated with those strains included:

- asymptomatic (11 strains),
- GC (14 strains),
- ML (18 strains),
- ulcer (10 strains).

Yoshio Yamaoka shared with us 35 sequences of strains isolated in Asia. Host pathology was not known for these strains, but the sequences were very useful as part of our global datasets used in Chapter 3 and Chapter 5.

A small collection of 18 strains cultures was shared with us by Sinead Smith, from Dublin in Ireland. DNA was extracted and sequenced in Swansea. Genomes of these strains were used in Chapter 3 and Chapter 5. 4 culture samples were also used in Chapter 6, alongside the corresponding sequences.

Another collection was shared from Bordeaux later during my thesis, this time by Philippe Lehours. This collection of 17 strains was isolated from two patients suffering from ML and used in infection studies in mice models (Chrisment et al. 2014). The strains were re-isolated after passage in mice and were sent to us for sequencing. Genomes from these strains were the basis for Chapter 4.

Fourteen cultures of strains were sent to us by Alain Burette, from Brussels in Belgium, through John Atherton. Pathologies associated with these strains were carefully selected, and were equally distributed between asymptomatic (7 strains) and GC (7 strains). Those strains were used in Chapter 5 and Chapter 6.

Six extra cultures of strains were shared with us by John Atherton, from Nottingham, in the UK. Pathologies associated with these strains were asymptomatic (4 strains) and ulcer (2 strains). They were used in Chapter 5.

Five strains were from the collection of Jane Mikhail. 3 strains were isolated in Singleton hospital, in Swansea, and 2 were shared with her by collaborators (unknown geographic origin). All these strains were sequenced in Swansea and used in Chapter 3 and Chapter 5.

2.2 Laboratory

All laboratory work involving mammalian cells and bacteria was performed inside a type 2 biological safety cabinet. All bacterial waste was autoclaved, and cell waste bleached or autoclaved. Incubations were achieved in an Air-Jacketed Automatic CO_2 Incubator. Pure cell cultures were kept isolated from cultures involving bacteria in two independent incubators in order to prevent cross-contamination. Suppliers for equipment are presented in Table 2.1. Suppliers for Consumables are presented in Table 2.2.

Table 2.1: List of Equipment

Equipment	**Product name**	**Supplier**
Type 2 Biological Safety Cabinet	ScanLaf MARS	LaboGene™
Incubator	Air-Jacketed Automatic CO_2 Incubator	NuAire, Inc.
Hermetic jar	Hermetic jar	Oxoid™
Spectrophotometer	Model 3710	Jenway
Centrifuge	Heraeus™ Megafuge™ 16R	ThermoFisher Scientific
Nanodrop spectrophotometer	ND1000	NanoDrop Technologies, inc.
Sequencer	HiSeqSystem	Illumina, San Diego, CA
Centrifuge	Centrifuge 5415 R	Eppendorf
Plate reader	FLUOStar® OMEGA	BMG LabTech
Chemidoc	Chemidoc MP system	BioRad

Table 2.2: List of Consumables

Consumables	**Supplier**
Columbia Blood Agar (CBA) plates	Oxoid™
Sterile disposable 10μL loop	Microspec©
CampyGen™ sachet	Thermo Scientific™ Oxoid™
Brucella Broth powder	BD BBL™
Foetal Bovine Serum (FBS)	Gibco®
L-shaped spreader	Microspec©
T25 flask	Greiner
Sterile disposable 1μL loop	Microspec©
Cryogenic vial	Starlab International©
Phosphate-Buffered Saline (PBS)	Gibco®
QIAmp DNA Mini Kit	Qiagen, Crawley, UK
RPMI 1640 media	Corning®
TrypLe Express	Gibco®
Dimethyl Sulfoxide (DMSO)	Merck
Trypan Blue stain (0.4%)	Gibco®
Human IL-8/CXCL8 DuoSet kit	R&D Systems
Human CCL4/MIP-1 beta DuoSet kit	R&D Systems
RayBio® C-Series kit	RayBiotech, Inc.

2.2.1 Culture of *H. pylori* on solid medium

Helicobacter pylori strains were recovered from glycerol stocks stored at -80°C. The stocks were maintained on ice to avoid thawing, and were spread onto fresh CBA plates at room temperature using a sterile disposable 10µL loop. The plates were inverted and incubated at 37°C in a hermetic jar with a CampyGen sachet which generated and maintained microaerophilic conditions for 4 to 6 days (Blanchard and Nedrud 2012). The CampyGen sachet had to be replaced every 2 to 3 days. The advancement of growth was checked on this occasion by estimation of the diameter of the colonies and the purity of the culture was verified by observation of plates for small, round, translucent colonies representative of *H. pylori* (Blanchard and Nedrud 2012). Homogeneity was also verified in order to identify potential mixed stocks. A verification of the species using the catalase, urease and oxidase tests (Blanchard and Nedrud 2012) was not necessary as these were performed prior to original storage.

2.2.2 Culture of *H. pylori* in liquid medium

The medium used for liquid culture of *H. pylori* was Brucella Broth (BB) (Table 2.3).

Table 2.3: Composition of Brucella Broth liquid medium

Brucella Broth powder	28 g/L
MilliQ water	200 mL
FBS	10 %

In brief, once enough growth was observed on solid medium, 1mL of BB was added to the surface of the plate and pushed to one side using a L-shaped spreader to resuspend the colonies. Then, 200 to 400 µL (according to the amount of growth observed on plates) of the suspended colonies were aliquoted and added to 10mL of BB in a T25 flask. A sterility flask with only BB was cultured in parallel to identify potential contamination from the medium. After manual agitation to homogenise the bacterial solution, the flasks were incubated at 37°C for 20-24 hours and then checked under the microscope for density of the culture and motility of the strains to confirm the absence of contaminants and viability of the strains for further experiments (Blanchard and Nedrud 2012).

2.2.3 Enumeration of *H. pylori*

Enumeration of *H. pylori* was performed to ensure quantification of the bacteria/cell ratio in co-culture experiments. *H. pylori* strains were first cultured on plates as

described in 2.2.1. Then, strains were cultured in liquid for 20 to 24 hours as described in 2.2.2. Optical density at 600nm was then assessed using a spectrophotometer, and bacterial cultures were diluted in BB Table 2.3 to obtain an optical density of 0.1 (OD 0.1). These normalised cultures of *H. pylori* were then successively diluted to 1/2,000, 1/10,000 and 1/20,000. 100 µL of these dilutions were plated in triplicate on CBA plates and incubated at 37°C under microaerophilic conditions until the size of the colonies was sufficient for enumeration (5 to 10 days). The number of colony forming units (CFU) in 1 mL of OD 0.1 normalised culture was then calculated using the appropriate dilution plates with colonies ranging between 25 and 1000. The lower limit of detection was 25 as recommended by (Tomasiewicz 1980; S. Sutton 2006). The upper limit was estimated empirically based on the small size of the colonies compared to bacteria used as standards and their separation on plates (Breed and Dotterrer 1916; Tomasiewicz 1980; S. Sutton 2006). The average CFU/mL was calculated for each strain, based on the calculation protocol presented below.

- Using the number of colony forming units ($nCFU_i$) for each plate (number of plates = N), an average was calculated ($aCFU = \frac{\sum_{i=1}^{N} nCFU_i}{N}$)
- All plates for which $nCFU_i \geq 0.2 * aCFU$ were discarded. (number of remaining plates = N*)
- If $N^* \geq 3$, a corrected average was calculated ($aCFU^* = \frac{\sum_{i=1}^{N^*} nCFU_i}{N^*}$). If less than three plates remained, a new enumeration was achieved.

2.2.4 Maintenance of stocks of *H. pylori*

New stocks were made from the *H. pylori* cultures on solid medium (2.2.1) prior to experiments, in order to keep a sufficient amount of the original stocks.

These stocks were made by collecting the colonies on the surface of the plates with a 1 µL loop and shaking the content of the loop into 700 µL of BB in a cryogenic vial. Then, 300 µL of a 50% solution of glycerol was added to the vial prior to vortexing. The new stock was then logged into Swansea Microbiology and Infectious Diseases (MID) group sample records and stored at -80°C for future use. All *H. pylori* used throughout the thesis period were subjected to minimal passages to avoid laboratory linked mutations while obtaining sufficient growth for experiments.

2.2.5 Motility of *H. pylori*

After culture of *H. pylori* on solid medium (2.2.1) followed by culture in liquid medium (2.2.2), cultures were centrifuged at 3000g for 5 minutes and the supernatant discarded. The bacterial pellet was resuspended into 2 mL of PBS and optical density (OD) was assessed using a spectrophotometer. Bacterial suspensions were then diluted to a normalised OD ranging from 0.5 to 1, according to the minimum OD of the batch following first measurement. Then, 0.5 µL of diluted bacteria solution was injected in triplicate into the centre of a 6-well agar plate prepared as described in Table 2.4. These motility plates were then incubated horizontally at 37°C under microaerophilic conditions until measurement, 2 - 6 days after inoculation.

Table 2.4: Composition of the motility assay plates

Brucella Broth	28 g/L
Agar	0.37%
FBS	10%
MilliQ Water	200 mL
2,3,5,-triphenyltetrazolium chloride (TTC)	1%

The diameter of growth was measured with a decimetre for each well, and an average for each strain i was calculated (D_i) (C.-Y. Kao, Sheu, and Wu 2014; C.-Y. Kao et al. 2012). These average measurements were then corrected to the positive control strain (B24) measured on the same day. Formula given by: $Motility_{index} = {D_i}/{D_{B24}}$. Every strain was studied twice to obtain an average index based on two independent experiments, and a third replicate was made if the standard deviation for the two values of normalised measure was more than 0.1.

2.2.6 DNA extraction and sequencing from *H. pylori* strains

Total DNA was extracted using the QIAmp DNA Mini Kit from solid cultures showing sufficient growth. Quantification of DNA was assessed with a Nanodrop spectrophotometer prior to sequencing. High-throughput genome sequencing was performed using a HiSeqSystem sequencer, and *de novo* assembling was performed using Velvet (version 1.2.08) by Matthew Hitchings and Ben Pascoe. All the contigs obtained from our samples were imported into the SheppardLab Bacterial Isolate Genome sequence database (BIGSdb http://zoo-

dalmore.zoo.ox.ac.uk/perl/bigsdb/bigsdb.pl?db=sheppard_hpylori_isolates) for genomic analysis.

2.2.7 Culture of AGS cells

AGS cells were cultured in Standard media (S media) composed of RPMI 1640 media supplemented with 2 mM of L-glutamine and 10% FBS. They were passaged 1 in 8 every 2 to 3 days for a maximum of 30 passages before use of a new stock. For cell detachment, media was removed, washed with PBS (8 mL) and then incubated with 4 mL of TrypLe Express, for 8 minutes. Detached cells were removed and neutralised in 8 mL of S media prior to centrifugation at 300g for 5 minutes. The media was removed and the cells resuspended in fresh S media. Cells were supplemented with Penicillin (100 units/mL) and Streptomycin (100 µg/mL) to avoid contamination during maintenance. Stocks were regularly made from the earlier passages to ensure sufficient stocks of cells for experiments, and were stored in 1 mL aliquots of FBS supplemented with 10% DMSO in Liquid Nitrogen.

2.2.8 Culture of THP-1 cells

THP-1 cells were cultured in S media. They were passaged 1 in 3 every 2 to 3 days for a maximum of 30 passages before use of a new stock. A centrifugation step was performed at 300g for 5 minutes to eliminate all remaining old media. The media was removed and the cells resuspended in fresh S media. Cells were supplemented with Penicillin (100 unit/mL) and Streptomycin (100 µg/mL) to avoid contamination during maintenance. Stocks were regularly made from the earlier passages to ensure a sufficient resource of cells for the experiments, and were stored in 1 mL aliquots of $5x10^6$ cells/mL in a solution of FBS supplemented with 10% DMSO in Liquid Nitrogen.

2.2.9 Viability testing

For both AGS and THP-1 cells, viable cells were enumerated prior to experiments. This assay consisted in 1 in 2 dilution of the resuspended cells in Trypan Blue stain (0.4%), then enumeration of the viable cells (white cells) using a hemocytometer. The average number of cells contained in 5 of the hemocytometer squares was calculated. Multiplication by $2x10^4$ was giving the concentration of cells in cells/mL.

2.2.10 Infection of AGS / THP-1 cells with *H. pylori*

AGS and THP-1 cells were resuspended from a viable culture obtained as described in 2.2.7 and 2.2.8 in S media without Penicillin/Streptomycin and diluted to obtain a concentration of respectively $50x10^3$ and $100x10^3$ cells/mL. Phorbol 12-myristate 13-acetate (PMA) was added (10 ng/mL) to the THP-1 cells to differentiate them into macrophages (Park et al. 2007), and both types of cells were seeded separately into 24 well plates (1 mL per well) and incubated for 24 hours at 37°C.

In parallel, the *H. pylori* strains were cultured in BB (2.2.2) for 22 hours at 37°C. After 22 hours of bacterial growth, the density of the cultures was assessed using a spectrophotometer and cultures were diluted down to an OD of 0.1. Bacterial cultures were then centrifuged at 3000g for 5 minutes and re-suspended in the same volume of S media.

AGS and THP-1 cells were washed twice with RPMI media supplemented with L-glutamine without Penicillin/Streptomycin, and 500 µL of the bacterial solution were added to the wells, along with 500 µL of S media without Penicillin/Streptomycin. PMA at 10 ng/mL (Park et al. 2007) was used as positive control and DMSO as negative control. Each sample was studied in triplicate wells. The plates were then centrifuged at 300g for 5 minutes at 28°C and incubated for 24 hours at 37°C. Supernatant was collected after 24 hours, centrifuged at 4°C at maximum speed (13.2 rpm) for 10 minutes, the supernatants decanted and then stored at -20°C until analysed.

2.2.11 Concentration of interleukin-8 in supernatants

Measure of the concentration of interleukin-8 (IL-8) in supernatants from infection experiments was obtained by enzyme-linked immunosorbent assay (ELISA) using a Human IL-8/CXCL8 DuoSet kit. First, a half-area plate with a flat bottom was incubated overnight with 50 µL of capture antibody. Wash buffer (Table 2.5) and Blocking buffer (Table 2.6) were freshly made.

Table 2.5: Composition of Wash buffer for ELISA

PBS tablets	1 per 200 mL
MilliQ Water	400 mL
Tween20	200 µL

Table 2.6: Composition of Blocking buffer for ELISA

Bovine serum albumin (BSA)	1%
PBS	50 mL

Plate was washed three times with wash buffer, and incubated with 150 µL of block buffer for one hour minimum. During this incubation step, standards (31.2-2000 pg/mL) were diluted into PBS and samples were prepared. Samples from AGS cells did not need dilution to be in the range of the analysis, but samples from THP-1 cells were diluted 1:50 in wash buffer. Plate was washed three times, and 50 µL of standards or samples were applied according to plate map recorded. After 1 hour 30 of incubation, the plate was washed three times, and 50 µL of detection andibody were added to the wells. After 1 hour 30 of incubation, the plate was washed three times, and 50 µL of streptavidin were added to the wells. After 20 minutes of incubation in the dark, plate was washed three times and peroxidase SureBlue was added to the wells. After 15 to 20 minutes of incubation, the optical density in each well was measured at 450 nm and 570 nm using a plate reader. The online software elisaanalysis.com was used to analyse the results, based on the corrected value of OD (450-570). An r^2 of minimum 0.995 was used to ensure reliable results, and for each sample the coefficient of variation between technical replicates was of a maximum of 20%. Each infection of AGS or THP-1 cells by an *H. pylori* strain was repeated in three independent experiments. To reduce experimental variations between experiments, the average negative control from each experiment was used as a unit value.

2.2.12 Concentration of CCL4 in supernatants

Measure of the concentration of Chemokine (C-C motif) ligand 4 (CCL4) in supernatants from infection experiments was obtained by ELISA using a Human CCL4/MIP-1 beta DuoSet kit. First, a half-area plate with a flat bottom was incubated overnight with 50 µL of capture antibody. Wash buffer (Table 2.5) and Blocking buffer (Table 2.6) were freshly made. Plate was washed three times with wash buffer, and incubated with 150 µL of block buffer for one hour minimum. During this incubation step, standards (15.6-1000 pg/mL) were diluted into PBS and samples from THP-1 cells were diluted 1:50 in PBS. Plate was washed three times, and 50 µL of standards or samples were applied according to plate map recorded. After 1 hour 30 of incubation, the plate was washed three times, and 50 µL of detection andibody

were added to the wells. After 1 hour 30 of incubation, the plate was washed three times, and 50 µL of streptavidin were added to the wells. After 20 minutes of incubation in the dark, plate was washed three times and peroxidase SureBlue was added to the wells. After 15 to 20 minutes of incubation, the optical density in each well was measured at 450 nm and 570 nm using a plate reader. The online software elisaanalysis.com was used to analyse the results, based on the corrected value of OD (450-570). An r^2 of minimum 0.995 was used to ensure reliable results, and for each sample the coefficient of variation between technical replicates was of a maximum of 20%. Each infection of THP-1 cells by an *H. pylori* strain was repeated in three independent experiments. To reduce experimental variations between experiments, the average negative control for each experiment was used to adjust the results of each experiment.

2.2.13 Human Inflammation Antibody Array

Detection of 40 human proteins was achieved for 4 samples from infection experiments. Two samples were obtained from infection of THP-1 cells with each of the two strains and two from infection of AGS with each of the same two strains. The two strains used were 30950 (gastric cancer and CagPAI positive) and 31235 (non-cancer strain and CagPAI negative). The assay was performed using the RayBio® C-Series kit according to the manufacturer instructions and intensity of the spots was analysed using ImageJ. Briefly, the membranes were blocked with 2 mL of blocking buffer for 30 minutes at room temperature. Blocking buffer was then removed, and 1 mL of sample was incubated for 3 hours at room temperature. Samples were removed by aspiration, and two washes were performed using two wash buffers with volumes of 2 mL. After removal of the second wash buffer, 1 mL of biotinylated antibody cocktail was added and incubated for 2 hours at room temperature. Membranes were washed again with the same two wash buffers, and 2 mL of HRP-streptavidin were added to the wells and incubated for 2 hours at room temperature. Membranes were washed again with the same wash buffers, and placed on a provided plastic sheet after removal of excess buffer. 500 µL of detection buffer mixture were added onto the membranes. After two minutes of incubation, the membranes were sandwiched between two plastic sheets and chemiluminescence was measured with a Chemidoc MP System. For each membrane, the maximum intensity of each spot (grey scale) was adjusted to the intensity of the 4 positive control spots from the top right corner.

This intensity of each pair of spots was compared between the cancer/non-cancer strains on AGS cells or on THP-1 cells. A difference of more than 0.1 in maximum intensity was investigated further by quantitative ELISA.

2.3 Genomics

2.3.1 BIGSdb

All the publicly available sequences for *H. pylori* genomes were uploaded onto the Sheppard lab Bacterial Isolates Genomic Sequences database (BIGSdb http://zoo-dalmore.zoo.ox.ac.uk/perl/bigsdb/bigsdb.pl?db=sheppard_hpylori_isolates) (K. A. Jolley and Maiden 2010). The strains sequenced on site and the ones shared by collaborators were also added to allow analysis to be run with both publicly available and new sequences. Strains with genome size below 1.3 Mbp were removed from analyses for bad quality. Strains with genome size above 1.9 Mbp were checked for sample contamination using nBLAST of suspected contigs against public databases and removed if contamination was verified. Contamination by the *phiX* gene, an artefact from Illumina sequencing (Mukherjee et al. 2015), was also cleared by nBLAST of the *phiX* gene. All available information about the strains, such as geographic or ethnic origin of the patient and symptoms associated, were also added to the database (Appendix C).

2.3.2 Genome Comparator

The first step of the genomic analysis was to run a genome comparator on the selected strains. Genome Comparator is a tool available on BIGSdb (K. A. Jolley and Maiden 2010). To run a Genome Comparator analysis, strains of interest were selected in the database. Then a list of genes was created. This list is usually either the list of genes from a reference strain (in our studies, we used the strain 26695), or a pan-genome (a list of all the genes present in at least one of the strains from the specific dataset). The Genome Comparator performs a gene-by-gene nBLAST (Altschul et al. 1990) alignment of all the genes from the given list in all the given strains, and gives back different output files, including a results table, an alignment, and an xmfa file. The results table shows the presence or absence of all the list genes in each strain, and also gives information of the allele version of the gene in each of these strains. The

alignments were used to make phylogenetic trees. The xmfa file shows the gene-by-gene alignment, allowing investigation of specific gene variations in more detail.

2.3.3 Genomic trees

Neighbour-joining trees were built using FastTree v2.0 (Price, Dehal, and Arkin 2010) performed using the High Performance Computing Wales (HPC Wales) system (HPC Wales 2017), and annotated with Evolview or iTOL v3 (Letunic and Bork 2016), based on alignments obtained from the output files from genome comparator analysis.

2.3.4 Pan-genome approach

The pan-genome approach consists of creating a list of all the genes present in at least one strain from our dataset. First the list of strains was selected. The mean size of a *H.pylori* genome was 1.635 Mb and the mean number of genes was 1616, based on the 695 sequences publicly available on NCBI Genbank (NCBI 2017). Therefore, any strain sequence deviating from the average in terms of either genome size or gene numbers was discarded suggesting a sequence of bad quality (low size or low number of genes with a lot of truncated genes) or contaminated sequence (high size with a lot of genes only found in this specific strain). Once the dataset was established, the fasta files of each strain sequence were downloaded from the BIGSdb, and a pan-genome was constructed using our group script (described in 2.3.4.1) or Roary (described in 2.3.4.2).

2.3.4.1 Pan-genome script

This method was developed by Leonardos Mageiros for part of his PhD at Swansea University (Mageiros, L 2013-2017) (Méric et al. 2014). Execution of the script was achieved by me. Genome-Wide Association Study based on ClonalFrame, and all minor genomic analysis requiring a pan-genome were performed using a pan-genome built with this script (Figure 2.2). Briefly, the fasta files exported from BIGSdb were submitted to the Rapid Annotation Sequences Tool (RAST) (Overbeek et al. 2014). The list of annotated genes for each of the strains was downloaded from the RAST server. Similarity between each pair of open reading frames was checked through a BLASTn (Altschul et al. 1990) search and a list of all the genes present in at least one

strain of our dataset was created (Sheppard et al. 2013). Genome Comparator and another upload into RAST were used to reduce gene duplicates to alleles of a same gene.

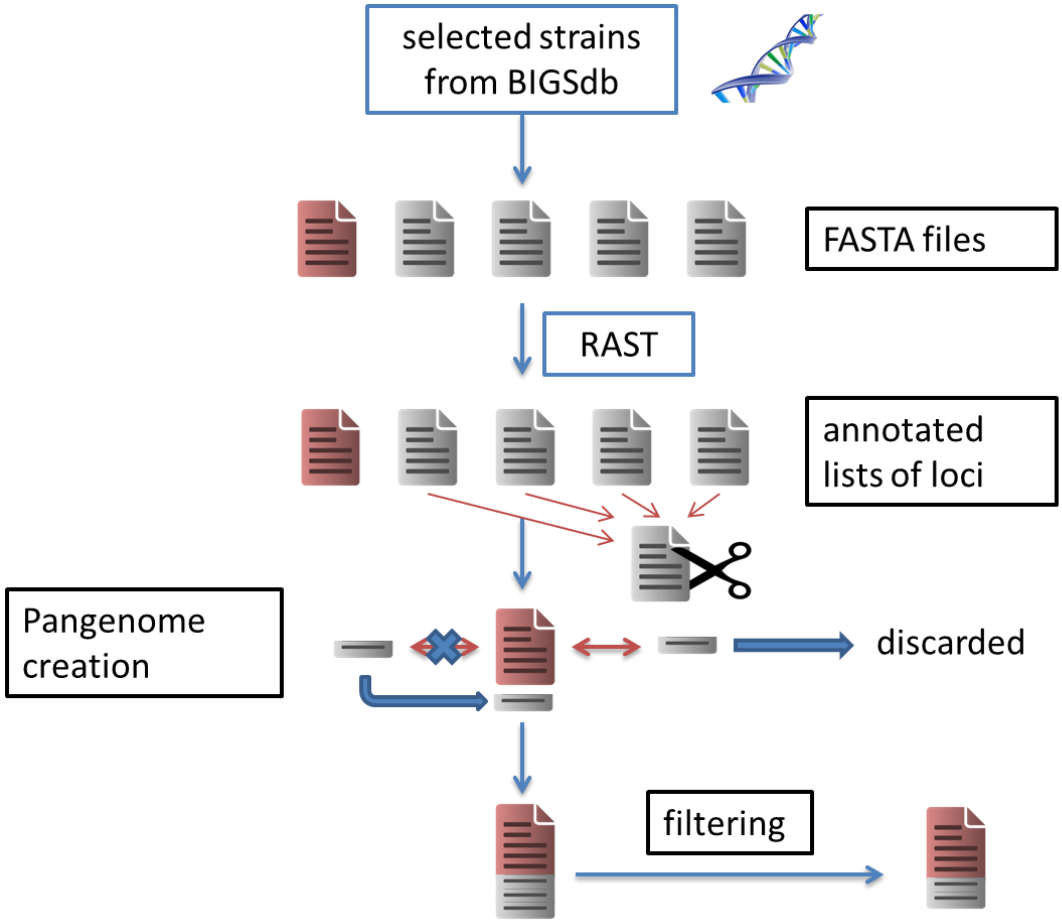


Figure 2.2: Pan-genome creation process using an in-house method developed by Leonardos Mageiros.
The script used RAST annotation files as input and produced a list of one allele for each gene present in at least one of the strains (Méric et al. 2014).

2.3.4.2 Roary

This method was used to run the publicly available bugwas package (Earle et al. 2016) used in Chapter 5. Lowering the threshold for Roary (Page et al. 2015) was not shown to strongly affect the size of the pan-genome, so default parameters were used.

2.3.5 FineStructure and ChromoPainter

FineStructure (Lawson et al. 2012) and Chromopainter (K. Yahara et al. 2013) analyses were performed by Koji Yahara in order to identify population structure in datasets based on paired similarity of core genome alignments. Preparation of the datasets and analysis of the results were achieved by me. Both the heatmap and tree from the FineStructure results were used to separate populations among isolates.

2.3.6 Genome-Wide Association Study

2.3.6.1 GWAS based on ClonalFrame

The first GWAS method was based on one previously published on other bacteria (Sheppard et al. 2013). It relied on the construction of a pan-genome from the dataset chosen and ClonalFrame (Didelot and Falush 2007). Briefly, a tree was built using a core genome alignment with FastTree v2.0 (Price, Dehal, and Arkin 2010). Pairs of strains were selected around that tree to create two replicate datasets. A pan-genome was created using the pan-genome script (Méric et al. 2014) method (2.3.4.1) on the joint replicate datasets. This pan-genome was used for each of our two replicate datasets as the reference gene list for the GWAS. The genes were split into 30bp words, or k-mers. The prevalence of those words in each group of strains (gastric cancer or non-cancer) was assessed using a ClonalFrame (Didelot and Falush 2007) based model, and the hits were identified. The script execution was performed by Leonardos Mageiros. Preparation of the datasets and analysis of the results were achieved by me.

2.3.6.2 GWAS based on bugwas

The second GWAS method was conducted using a pipeline recently applied in another study by Koji Yahara (Suzuki et al. 2016). This method was based on the bugwas package (Earle et al. 2016), and was executed in two approaches.

The first version was a k-mer-based approach (Sheppard et al. 2013) in which the genome sequence of each isolate was fragmented into unique overlapping 31-bp DNA words, or k-mers, that were used to identify genetic variations. The 31-bp words significantly associated with gastric cancer were explored after accounting for the inter-dependence of the strains and population structure. The script calculated an n x n

relatedness matrix that summarized all genetic covariance among the strains, and employed statistical tests for a given k-mer using a linear mixed regression model. Unlike the ClonalFrame method, this method does not depend on a single clonal tree that is difficult to construct reliably due to the high rate of recombination in *H. pylori*. The second version was a SNP-based approach in which the nucleotides present in all positions of each of the genes was compared between pairs of isolates.

Bugwas, like the ClonalFrame based method, was also based on a pan-genome analysis. Annotation was done using prokka, and pan-genome creation was made using Roary (2.3.4.2). Selection of hits was based on the odds ratio and p-value. The hits were then analysed individually to investigate their function and the effect of the SNPs identified. Execution of bugWAS was performed by Koji Yahara. Preparation of the datasets and analysis of the results were achieved by me.

2.3.7 Accessory genome analysis

In one of our analyses (see chapter 1), some populations were hybrids derived from a few derived populations, and an accessory genome analysis was developed to highlight markers of this evolution in the populations. The accessory genome from our dataset was studied based on the output of a genome comparator executed on the dataset of 401 strains (Appendix C) using the pan-genome script method (2.3.4.1). A binary presence/absence matrix was built for all the accessory genes in all the 401 strains.

2.3.7.1 Accessory tree

For each pair of strains, the ratio of genes present or absent out of the total number of accessory genes was calculated to create a relatedness matrix. This matrix was then used to build a tree using Matlab, and was visualised with Evolview (He et al. 2016).

2.3.7.2 Accessory plots

The binary presence/absence matrix was also used to calculate the frequency of presence of each individual gene in each FineStructure population of our dataset, called prevalence. These data were used to build 3 dimensional plots, using the prevalence in each of the three groups selected as coordinates. On the X and Y axis

were two suspected ancestor populations, and on the Z axis was the suspected hybrid population.

All the genes were split into 7 categories of profiles:

- "no difference between prevalence in the 3 populations",
- "prevalence in hybrid = prevalence in ancestor 1",
- "prevalence in hybrid = prevalence in ancestor 2",
- "prevalence in hybrid = average between ancestor 1 and ancestor 2",
- "prevalence in ancestor 1 = prevalence in ancestor 2",
- "prevalence in ancestor 1 = average between ancestor 2 and hybrid",
- "prevalence in ancestor 2 = average between ancestor 1 and hybrid".

The repartition of the genes in these plots gave indications of the likeliness of the hypothesised hybrid population being a hybrid between the two ancestral populations, as well as the closeness of the hybrid population to one or another ancestor.

Statistics were performed based on the distance of each gene from the equi-prevalence straight line (X=Y=Z), by comparing each category of genes (apart from the first one) in each plot via an ANOVA. In each case, the equality of variances was not verified (p-value < 0.05 for Levene's statistic test), and sample size (the number of genes) was small in some of the categories of genes, so a Dunnett's T3 test was chosen to perform the ANOVA.

2.3.8 Analysis of individual gene variations

Strain differences between individual genes were investigated using BioEdit. Effects on the amino-acid sequence were checked using an amino-acid alignment obtained from BIGSdb, to differentiate synonymous and non-synonymous hits. Non-synonymous hits were further studied by creation of figures showing proportion of amino-acids in each position according to characteristics of the strains, using WebLogo (Crooks et al. 2004).

2.4 Statistical Analysis

All statistical analyses were performed using GraphPad Prism. When two groups were compared, unpaired t-tests were used. ANOVA analyses were used when more than two groups were compared. For each statistical test (except in GWAS analyses), the level of significance used (unless otherwise stated) was 0.05.

3 Long-term genomic evolution of *H. pylori* in Americas

Helicobacter pylori has had the ability to colonise human stomachs for thousands of years (Moodley et al. 2012). It can also live within its host for years (Rhee, Park, and Cho 2014), and different strains can cohabit within the same host (Cao et al. 2015). This long-term colonisation has resulted in co-evolution of human and *H. pylori* genomes in populations around the world. Human migrations have carried the bacteria throughout history, and traces of ancient human migrations can be found by studying genomic admixture in *H. pylori* genomes (Falush et al. 2003). The Americas are an excellent place to study population admixture, because of the recent history of human migration. A rapid colonisation of the New World, principally by European migrants and African slaves, massacred indigenous populations. The new-comers brought with them pathogens, including *H. pylori* that were different from native populations. The *H. pylori* species seems to have benefited from this new gene-pool created by this human migration, facilitating rapid recombination and mutation (Suerbaum and Josenhans 2007).

The three human populations investigated here include Europeans, Africans and Native Americans and each carry genetically distinct populations of *H. pylori*, named hpEurope, hpAfrica1, hpAfrica2 and hspAmerind (Montano et al. 2015; Falush et al. 2003; Linz et al. 2007). The prefix hp indicates a population and hsp a subpopulation. Subpopulations are genetically distinct from each other but less differentiated than populations. The relationships between bacterial populations reflect differentiation that occurred during the complex migration history of humans (Falush et al. 2003). hspAmerind strains are presumed to be descendants of the strains present in the Americas prior to 1492, and are a subpopulation of hpEAsia, which is found in East Asian countries such as China and Japan. This heritage is linked to the ancient colonisation of the New World by Asian populations (Marangoni, Caramelli, and Manzi 2014). hspAmerind subpopulation is rarely found, even within groups with strong Native American ancestry and may be dying out in competition with other strains, possibly due to low diversity within the population (Domínguez-Bello et al. 2008). hpEurope bacteria are themselves ancient hybrids between two populations, whose close relatives are currently found in un-admixed populations in North East Africa (hpNEAfrica) and central Asia (hpAsia2). A study of a 5300-year old mummy found in central Europe showed that he was infected with *H. pylori* which was an

hpAsia2 type, with little or no African ancestry, suggesting that the admixture probably took place within the last few thousand years (Maixner et al. 2016).

H. pylori is associated with gastric cancer (GC), which is one of the most lethal cancers (Parsonnet et al. 1997; Plummer et al. 2015). GC is a global health issue, but it is even more so in Latin America, as some countries in this region have among the highest mortality rates worldwide (Ferlay et al. 2015). Mortality due to GC is ranked 3rd in South America and 5th in Central America (Globocan 2012). Mortality rates vary between neighbouring countries (for instance 5.6 in Argentina against 15.0 in Chile in terms of Estimated age-standardized mortality rates for males per 100,000), and within nations (for instance in Colombia with higher rates in mountains populations than coastal populations) (Ferlay et al. 2015; Torres et al. 2013). In addition, phylogeographic origin of the bacteria, as well as discordant origin of bacteria and hosts have been linked to increased risk of gastric cancer development (de Sablet et al. 2011; Kodaman et al. 2014). However, these studies were based on MLST analysis which compares only seven housekeeping genes, therefore whole-genome based analysis would increase the resolution for such studies.

There is a need for better understanding of the dynamics among *H. pylori* populations. MLST is a good tool and was precious to start understanding the link between *H. pylori* populations and their hosts, however whole-genome based methods will bring a better resolution.

The development of whole-genome sequencing makes it possible to study in depth the variations in *H. pylori* genomes and the ancestry of these populations. This will lead to a better understanding of the mechanisms leading to such a wide variability in *H. pylori*, and the relationship between this variability and human activities. Recent research (Thorell et al. 2017) formed part of this project and is outlined in Appendix D. For this study, a global collection of *H. pylori* genomes, combining both publicly available and newly-sequenced genomes, was studied. This study aimed to confirm two things. First, that the genomic variability of *H. pylori* strains from the Americas reflects the history of migrations which shaped these regions. Second, that both the core and accessory genomes reflect this in a similar way.

3.1 Materials and Methods

3.1.1 Genomic data set

The data set used in this study was composed of both publicly available strains and newly sequenced strains. The original dataset available on the Sheppard lab BIGSdb (BIGSdb http://zoo-dalmore.zoo.ox.ac.uk/perl/bigsdb/bigsdb.pl?db=sheppard_hpylori_isolates) consists of 825 strains. This dataset was processed to remove all sequences from non-human sources or those that were of poor quality. A genome comparison was performed (2.3.2) and a neighbour-joining tree was built using FastTree v2.0 (Price, Dehal, and Arkin 2010) to identify clones. Clones are defined as strains that are almost identical, often isolated from a same patient, either from different areas of the stomach or at different time points, or from patients belonging to the same family. The final dataset comprised 401 strains. Average number of contigs was 85.9, average length was 1644038.5 bp and average GC content was 38.95% (Figure 3.1).

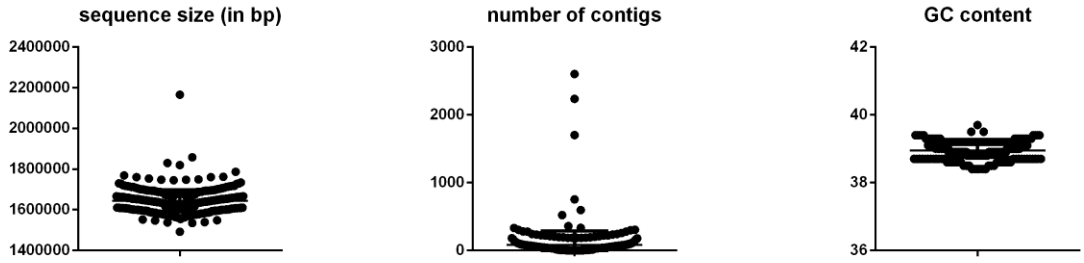


Figure 3.1: Distribution of the genomic characteristics of the sequences used in a 401 strains dataset.

3.1.2 Chromopainter and FineStructure

Genes from the reference strain 26695 were searched by nBLAST in the 401 strains by genome comparator, in order to prepare core genome alignments used as input for ChromoPainter (K. Yahara et al. 2013) (2.3.5). ChromoPainter infers chunks of DNA donated from a donor to a recipient. Results are summarized into a co-ancestry matrix, highlighting the relatedness of some populations by comparing the profile of the strains from this population, and the likeliness of their possible hybridisation. FineStructure (2.3.5) (Lawson et al. 2012) uses the co-ancestry matrix to cluster

individuals based on paired-similarity. The results from FineStructure are summarised in a heatmap.

3.1.3 Pan-genome approach

A Pan-genome was built using an in house developed script (2.3.4.1). Briefly, fasta files from all 401 strains were downloaded from the BIGSdb and submitted to the Rapid Annotation Sequences Tool (RAST) (Overbeek et al. 2014). A list of one allele of each gene found in at least one of the strains was then compiled and filtered. The final pan-genome for this analysis was composed of 2457 genes. A genome comparator executed on the 401 strains from the dataset against this pan genome list revealed that 990 of these genes were core genes (present in all the strains from this dataset). The 1467 remaining genes were accessory genes (present in at least one, but not all strains from this dataset). Truncated genes were considered as present, as truncation was in most cases caused by contig limits and therefore did not reflect a biological truncation of the genes.

3.1.4 Accessory genome analysis

The accessory genome from our dataset was studied based on the table output of a genome comparator executed on the dataset of 401 strains using the pan-genome list of 2457 genes. A binary presence/absence matrix was built for all the 1467 accessory genes from this 401 strains dataset.

Using this binary presence/absence matrix, an accessory tree was built and annotated (2.3.7.1). The binary presence/absence matrix was also used to calculate the frequency of each individual gene in each FineStructure population of our dataset (called prevalence of the gene). In addition, it was used to build 3-dimensional plots, using the prevalence in each of the three groups selected as coordinates (2.3.7.2). ANOVA-based statistics were performed to describe the hybridisation using Graphpad Prism v6.

3.2 Results

3.2.1 Phylogeny of the data set

The final data set comprised 401 strains of *Helicobacter pylori* that were sampled globally. The topology of the genomic tree (Figure 3.2: Neighbour-joining tree based on the whole-genome alignment of 401 *H. pylori* strains.) was typical of *H. pylori*, with strains distributed on a cline going from hspEAsia to hpAfrica1 and with hpAfrica2 splitting away from the general cline.

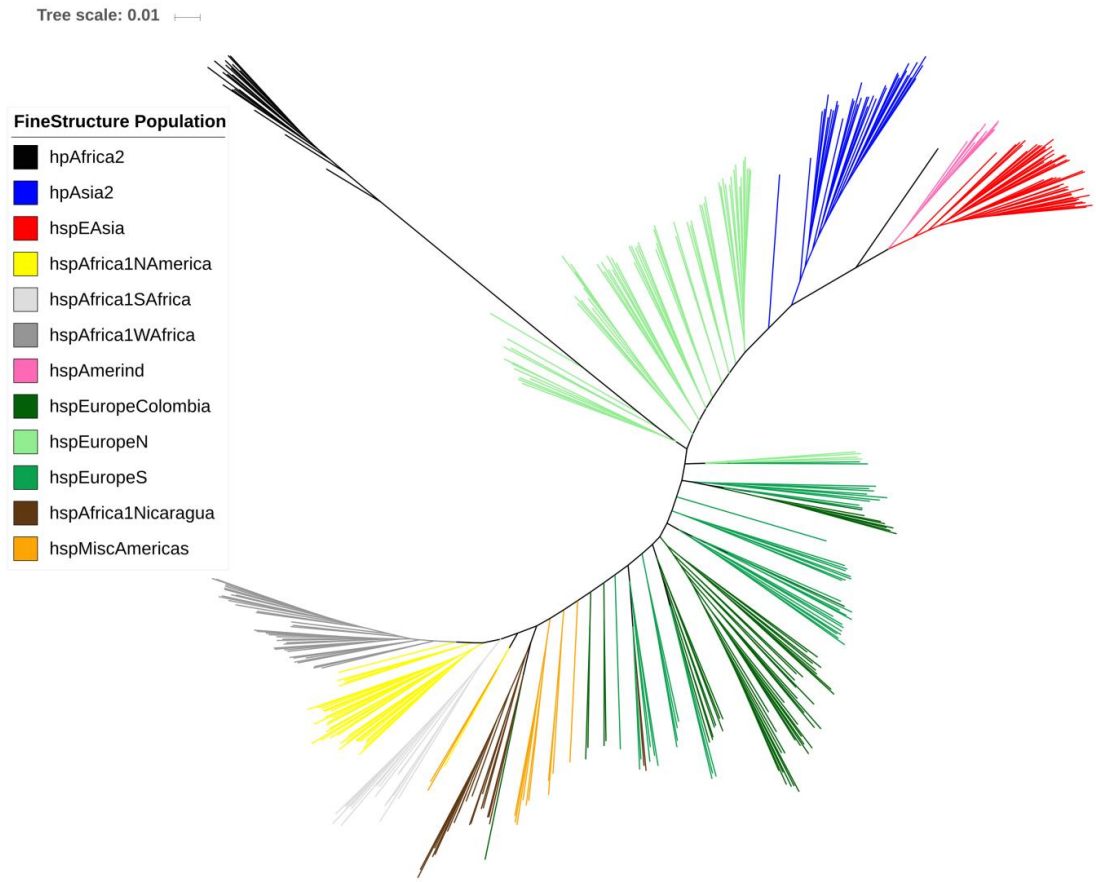


Figure 3.2: Neighbour-joining tree based on the whole-genome alignment of 401 *H. pylori* strains.
The population of the strains determined by FineStructure is represented through color of the branches.

The geographic origin of these 401 clinical strains of *H. pylori* isolated in human clinical cases can be broken down as described in Table 3.1.

Table 3.1: Geographic origin of the 401 *H. pylori* strains used in FineStructure analysis.

Colonisation group	Geographic origin	Number of strains
New World	North America	57
	Mexico	35
	Central America	33
	South America	73
Old World	Europe	41
	West Africa	38
	South Africa	43
	Asia	57
	East Asia	22
	Oceania	1
	Experimental	1

3.2.2 Core Genome Analysis

3.2.2.1 FineStructure

A FineStructure analysis was performed to identify the populations found among the dataset (Figure 3.3). The coloured matrix demonstrates how much of the genomes are shared between strains. The darker the colour, the more shared ancestry. Twelve populations were identified based on this analysis. Five of them are found mainly in the New World. Some populations were more distinct from other populations. For instance, hpAfrica2, hspAmerind and hpAsia2 show a large amount of intra-population importation and a low amount of importation from or to other populations. Conversely, hspEuropeColombia and hspEuropeS show almost the same amount of importation from or to other populations than intra-population (Figure 3.3).

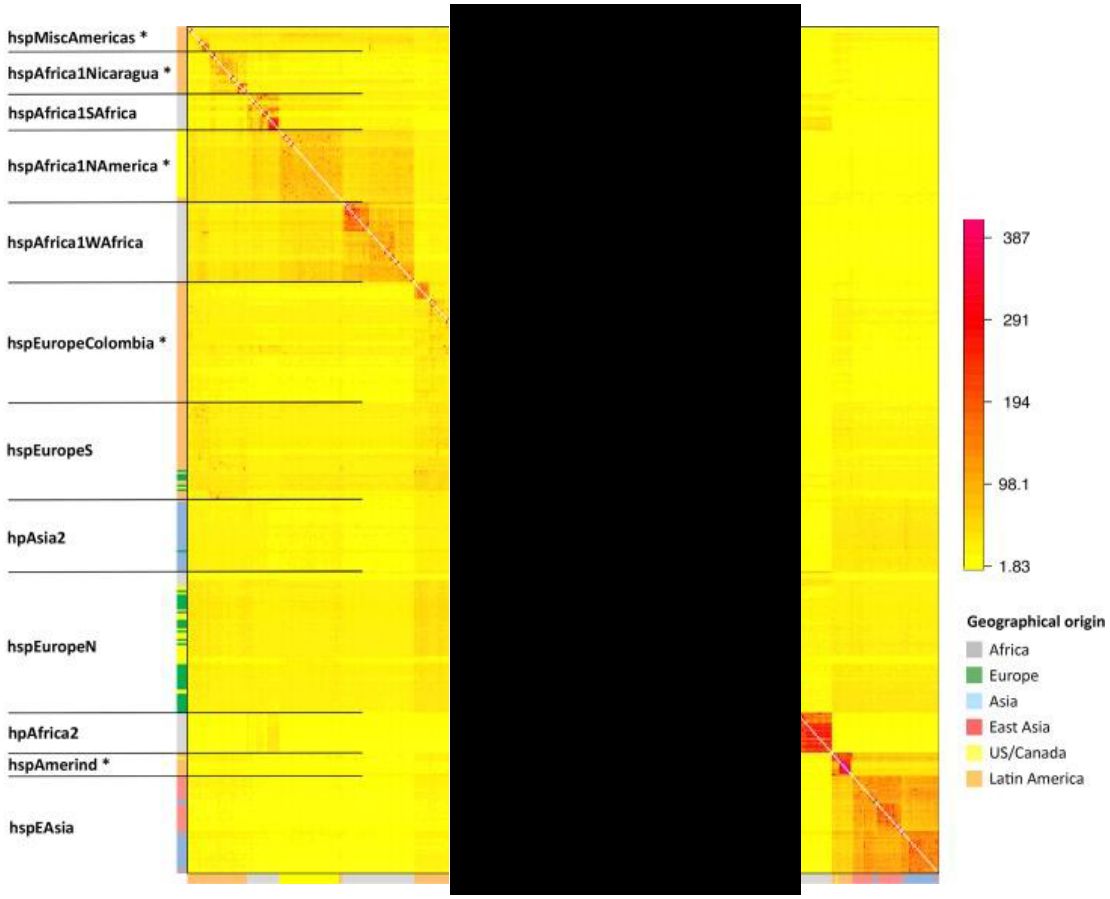


Figure 3.3: Identification of 12 populations of *H.pylori* in a dataset of 401 global strains by FineStructure.
Adapted from (Thorell et al. 2017). Asterisks are for populations mainly found in the New World. The colour of each cell in the matrix represents the expected number of DNA chunks imported from a donor genome (column / x-axis) to a recipient genome (row / y axis). The colour bars indicate the geographic origin of the recipient and donor strains.

3.2.2.2 Chromosome painting

Two chromosome painting analyses were performed and results were arranged to show the link between geographic origin of the strains and their population, as well as the relationship between populations (Figure 3.4).

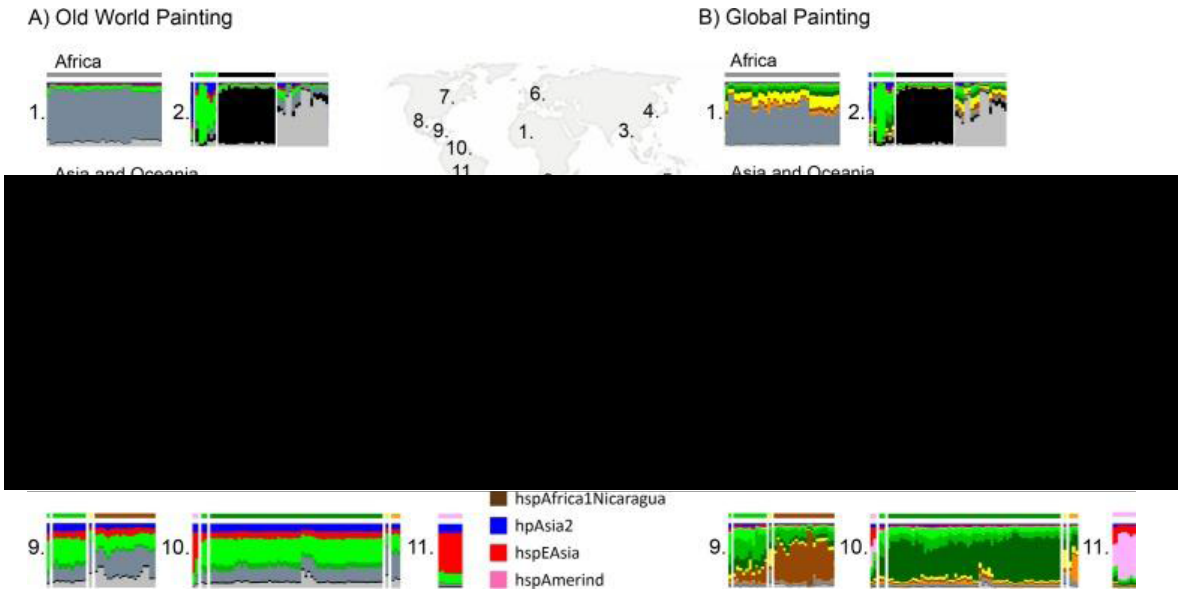


Figure 3.4: Chromosome painting results showing repartition of the global and ancestral population in the world.
Adapted from (Thorell et al. 2017). Each vertical bar represents one isolate. The colour composition of each bar represents the amount of DNA donated from each subpopulation in the core genome of the isolate. **Panel A** is made with only old world populations (hspAfrica1SAfrica, hspAfrica1WAfrica, hspEuropeS, hpAsia2, hspEuropeN, hpAfrica2 and hspEAsia) from old world geographic origin (1 to 6 on map) used as donors. **Panel B** is made using all strains as potential donors.

Each colour represents one of the twelve populations of *H. pylori* identified with FineStructure. On Figure 3.4A, the donor panel used was made with only isolates from old world geographical origin (Europe, Africa and Asia) belonging to Old World populations (hspAfrica1SAfrica, hspAfrica1WAfrica, hspEuropeS, hpAsia2, hspEuropeN, hpAfrica2 and hspEAsia). hspEuropeS seems to have a larger fraction of chunks coming from African populations, while hspEuropeN has a larger portion of DNA coming from hpAsia2. These two populations also have a larger portion of their palette coming from other old world population, compared to hpAfrica1, hpAfrica2, hpEAsia and hpAsia2. This is due to the ancient hybridisation between hpAfrica1 and hpAsia2 resulting in the modern hpEurope population (Falush et al. 2003). The 5 other old world populations are highly distinct, showing more than half of their palette coming from their own population.

Our hypothesis is that most of the recombination occurred from Old World to New World strains. This OldWorld Painting allows us to investigate the origins of these New World *H. pylori* populations without the more recent admixture that happened since the colonisation of New World by Old World human populations.

Data shown in Figure 3.4B, on the other hand, uses all strains as potential donors, allowing a better view of the recent admixture events, and the relationship between New World populations of *H. pylori* strains.

Regarding the isolates found in the Americas, some were strains belonging to Old World populations, others revealed 5 new subpopulations that were not found in Old World territory and included: hspEuropeColombia, hspAfrica1NAmerica, hspMiscAmericas, hspAfrica1Nicaragua and hspAmerind. Most of these populations were derived from European and African Old World populations (Figure 3.4A). The exception was hspAmerind, described in previous research as closely related to hspEAsia (Falush et al. 2003), originating from a more ancient colonisation of the Americas by Asian people, who formed the Native American population. hspEuropeColombia was found mainly in Colombian isolates and demonstrated a high level of European ancestry. As shown in Figure 3.4B, this population shows high intra-population recombination. hspAfrica1NAmerica and hspAfrica1Nicaragua were found in isolates from North America and Nicaragua respectively. Both populations showed a high level of African ancestry, but did not show exactly the same pattern (Figure 3.4B). hspAfrica1NAmerica showed a dominance of African ancestry, whereas hspAfrica1Nicaragua also had imported European and Asian ancestry. Global painting (Figure 3.4B) also revealed more intra-population recombination in hspAfrica1Nicaragua than there was in hspAfrica1NAmerica.

As both Finestructure and Chromopainting analyses are based on core genome analysis, methods were developed to study the accessory genome in that specific aspect of hybridisation of populations.

3.2.3 Accessory genome analysis

3.2.3.1 Accessory genome phylogeny

For each strain, the pattern of presence-absence of each gene was determined by genome comparator to calculate frequency (or prevalence) in the different populations. A neighbour-joining tree was built based on the presence/absence relatedness matrix (Figure 3.5).

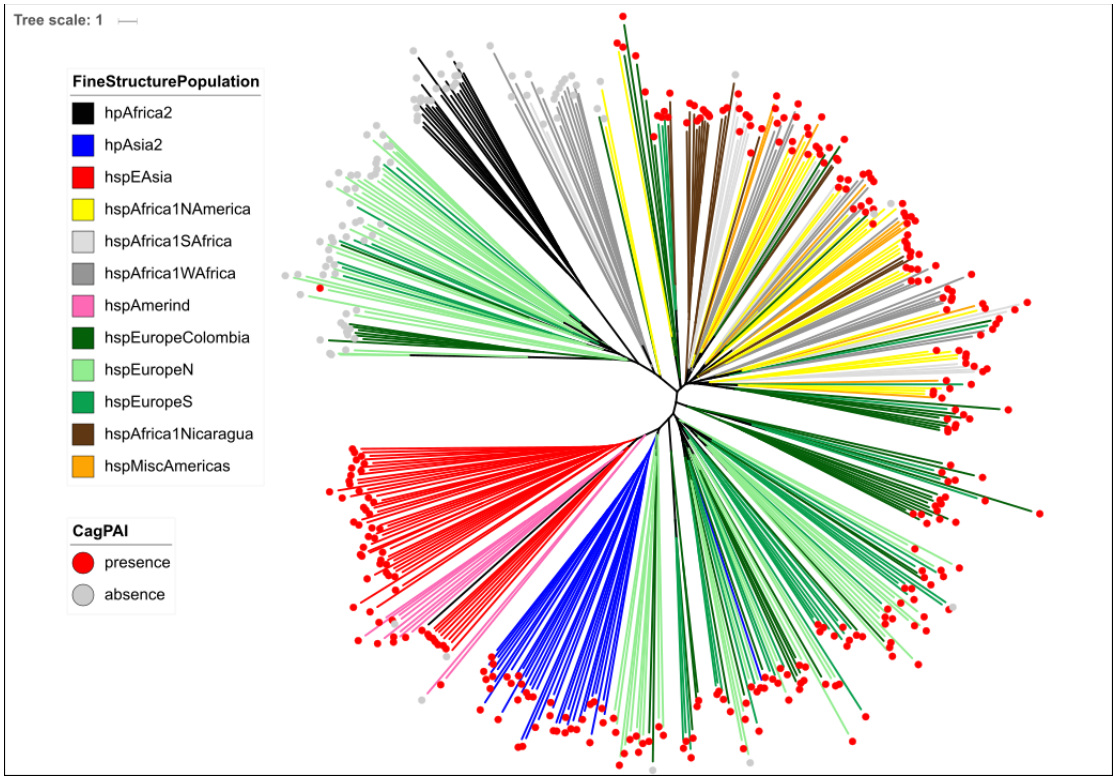


Figure 3.5: Neighbour-joining accessory genome tree based on gene sharing distance (absence and presence of genes).
This tree was annotated with iTOL (Letunic and Bork 2016). Colour of the branches represents the FineStructure population. Red circles indicate strains with the Cag Pathogenicity Island (CagPAI), and grey circles indicate strains without the CagPAI. An online version is available for this figure at this address: https://itol.embl.de/tree/13744115340155150929017 9#

Figure 3.5 highlights that the accessory genome shows similar segregation of the strains as the one found with FineStructure, which was based only on core genome (Figure 3.2). However the relatedness of the strains is not only made according to hp populations, as some of the populations are now splitted into different areas of the tree. Presence of the Cag pathogenicity island (CagPAI) was investigated, as this group of genes was often identified as a 'whole' in many of the strains, and we hypothesised that this would affect the clustering of strains in this accessory genome tree. That was the case, as shown in Figure 3.5, most of the CagPAI positive strains clustered in the same area of the tree. Interestingly, most of the CagPAI negative European strains (hspEuropeN and hspEuropeS) and hspEuropeColombia cluster with the CagPAI negative African strains, whereas CagPAI positive European strains cluster with the Asian strains (in large majority CagPAI positive). American populations such as hspMiscAmericas and hspAfrica1Nicaragua and hspAfrica1NAmerica cluster with the CagPAI positive African strains.

3.2.3.2 Accessory genome plots

Additional analysis of the accessory genome was performed on hpEurope and 3 of the New World populations (hspEuropeColombia, hspAfrica1NAmerica and hspAfrica1Nicaragua), to verify if their accessory genome profiles were consistent with the Chromosome painting and FineStructure results based on the core genome.

The method led to the creation of 3 dimensional plots using the prevalence in the potential hybrid population as vertical axis, and the prevalence in the two ancestral populations suggested by chromosome painting on the horizontal axis. A randomisation of the group to which the strains belonged was performed as a negative control. All strains were allocated to a random group without any population structure taken into consideration, and the prevalence in these 3 randomised groups was used to build the negative control to check that the model showed divergence in accessory genome prevalence which was linked to the population structure (Figure 3.6A). The negative control showed genes to be clustered in the centre of the plot (labelled in black), with no outliers being closer to one of the axis (labelled in colours according to Figure 3.6B). A positive control was also built using data generated for a hypothetical "perfect hybrid" population with genes sharing either the exact same prevalence than one of the ancestor or the average prevalence between the two ancestors (Figure 3.6C).

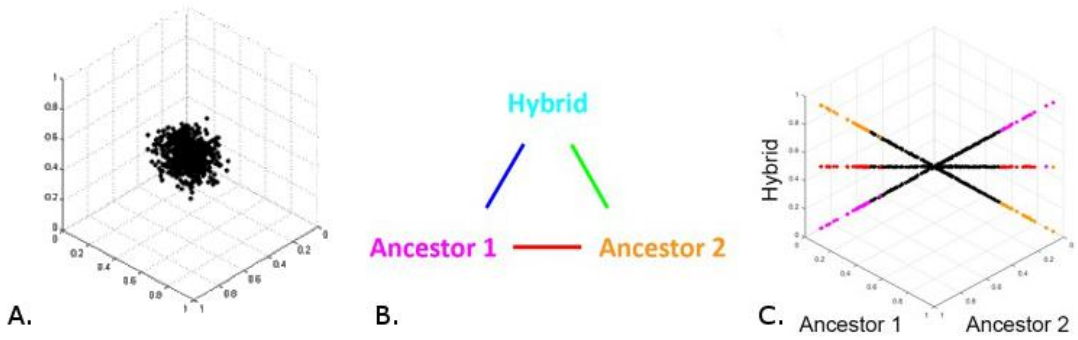


Figure 3.6: Controls used to develop a method for hybridisation analysis based on 3-dimensional plots of accessory genomes.
Adapted from (Thorell et al. 2017) **A.** Negative control plot, built using random population assignment. **B.** Colour Legend **C.** Positive control plot, built using artificial hybrid population with frequency of genes either identical to Ancestor 1 (in magenta), to Ancestor 2 (in orange) or a 50-50 hybrid (in red). In panels A and C, each dot is a gene, and the coordinates correspond to the frequencies of this gene in the three randomised populations. The graphs are orientated in order to have genes with identical frequencies in all three populations in the centre of the plot. These genes and those that had small variations of frequencies are represented in black. Genes with greater frequency differences appear in colours, according to the triangular legend presented in panel B. Colours used for the text are for genes that differ substantially between the population named and the other two. The criteria used to say a frequency X is larger than another one Y was [X-Y>= 0.5], [X>=0.5 and Y<0.1], or [X>0.9 and Y<=0.5]. Colours on the lines are used for genes that show high differences in frequency between the two populations on each side of the line, with the last one having an intermediate frequency.

The accessory genome method was applied to different sets of populations (Figure 3.7). The first set used was hpEurope as a hybrid population from hpAsia2 and hpAfrica1. The hpEurope population is understood to be the result of an old recombination between hpAsia2 and hpAfrica1 (Falush et al. 2003), based on MLST analysis. This was also confirmed by the core genome analysis (Figure 3.3 and Figure 3.4). The three other hybrid populations studied were hspAfrica1NAmerica, hspEuropeColombia and hspAfricaNicaragua, which were identified as hybrids from African and European origins in the core genome analysis (Figure 3.4).

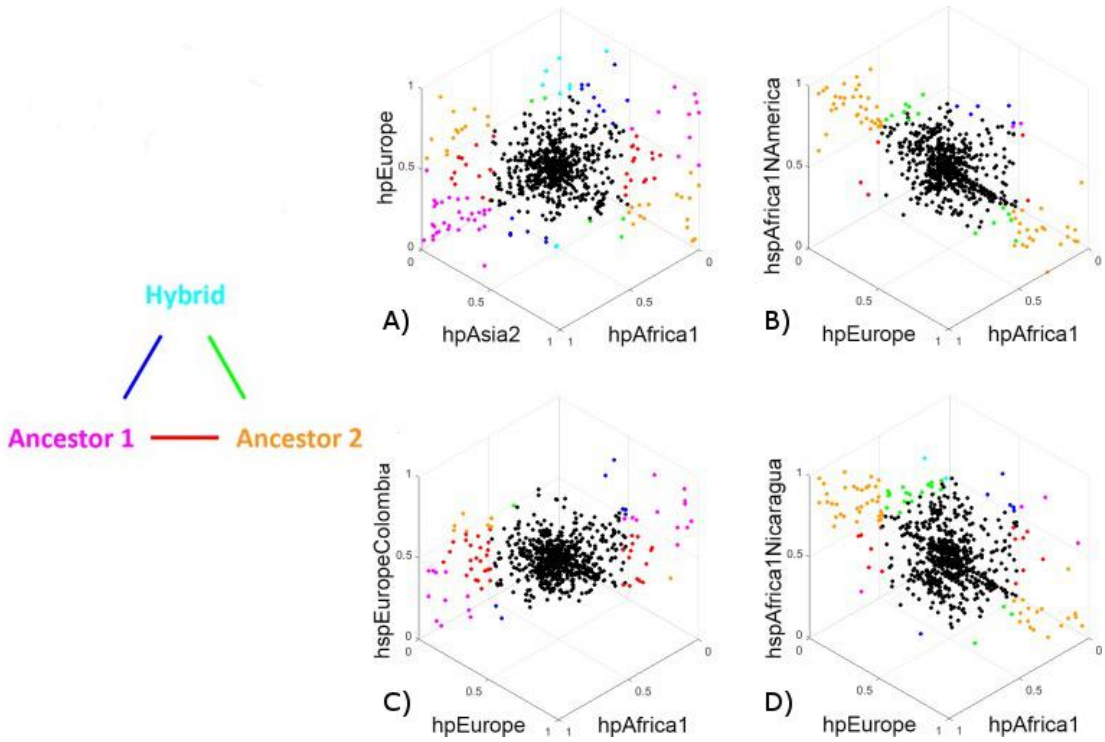


Figure 3.7: 3-dimensional plots of accessory genomes in hybrid populations.

Adapted from (Thorell et al. 2017). Each dot is a gene, and the coordinates correspond to the frequencies of this gene in three populations. The graphs are orientated in order to have genes with identical frequencies in all three populations in the centre of the plot. These genes and those that had too smaller variations of frequencies to be studied are represented in black. Genes with greaterfrequency differences are labelled in colours, according to the triangular legend. Colours used for the text are for genes that differ substantially between the population named and the other two. The criteria used to say a frequency X is larger than another one Y was [X-Y>= 0.5], [X>=0.5 and Y<0.1], or [X>0.9 and Y<=0.5]. Colours on the lines are used for genes that show a high difference in frequency between the two populations on each side of the line, with the last one having an intermediate frequency. **A.** Comparison of hpEurope to hpAsia2 and hpAfrica1, from strains isolated in old world.**B.** Comparison of hspEuropeColombia to hpEurope and hpAfrica1. **C.** Comparison of hspAfrica1Nicaragua to hpEurope and hpAfrica1. **D.** Comparison of hspAfrica1NAmerica to hpEurope and hpAfrica1.

Plots in Figure 3.7 show different types of hybrids based on the repartition of gene prevalence compared to hypothetical ancestors. hpEurope (Figure 3.7A) appear to have genes equally distributed between those showing an African profile, those showing an Asian profile, and those showing an average repartition between African and Asian prevalence. The three other hybrid populations seem to be much more driven by one of the ancestors over the other one. hspAfrica1NAmerica (Figure 3.7B)

and hspAfrica1Nicaragua (Figure 3.7D) show a larger number of genes with a distribution similar to that of hpAfrica1. However, there are more genes of different ancestry in hspAfrica1Nicaragua than in hspAfrica1NAmerica which has a more linear repartition along the hpAfrica1 line. hspEuropeColombia (Figure 3.7C) has a profile close to hpEurope, but with a small drift to the side, showing a significant difference in the pattern of prevalence.

Statistics were performed on the plots from Figure 3.7, based on the distance of each gene from the equi-prevalence line. The ANOVA analysis for the plot from Figure 3.7A (hpEurope) showed a significantly higher distance for genes belonging to the profile "hybrid = hpAfrica1" and "hybrid = hpAsia2" (p-value < 0.05), which confirms that the population hpEurope is a balanced hybrid between hpAfrica1 and hpAsia2. Regarding New World populations, our accessory genome study identifies two types of hybrids. In Figure 3.7B and Figure 3.7D, populations are closer to hpAfrica1 than to hpEurope. The ANOVA for these two plots show that only "hybrid = ancestor 1" is significantly higher (p-value < 0.05) in distance than the other groups. In Figure 3.7C the population is closer to hpEurope than to hpAfrica1, with statistics showing that only "hybrid = ancestor 2" is significantly higher (p-value < 0.05) in distance than the other group.

3.3 Discussion

A large number of strains collected in Northern and Latin America were from hpEurope populations, with chromosome painting profiles undistinguishable between Old World and New World isolates. These hpEurope strains were clustering into two close but distinct subpopulations, which tend to segregate between North and South. The southern population showed a higher hpAfrica1 ancestry compared to the northern one. This difference in Old World could be explained by a Mediterranean melting pot, and there is a cline in the European strain-based tree that is consistent with this observation. The New World part of these hpEurope subpopulations can be explained by the colonisation history of North and South America. These observations draw a direct relationship between human and bacterial ancestries.

However, human and bacterial ancestries are not always perfectly concordant. Colonisation of the Americas by European populations resulted in a new physical and dietary environment for the bacteria, as well as a new ethnic mix of hosts. Colonisation of a host from a certain ethnic background by a bacteria from a distinct

population background can lead to modifications of the host-bacterial interactions, as observed in other studies (de Sablet et al. 2011; Kodaman et al. 2014). This can also have an effect on long-term evolution. The hspAmerind population is a good example of these effects. Strains from this population are extremely rare, even in populations with substantial Native American ancestry. Even though this might be biased by a more limited access to clinical samples from such populations, there is strong evidence that these strains are dying out due to the benefit of hpEurope strains when these populations are in a competitive environment (Domínguez-Bello et al. 2008). This is reflected in the present study due to the fact that none of the populations only found in the New World show a considerable amount of hspEAsia apart from hspAmerind. This subpopulation did not mix with the others to produce a new subpopulation-it appears to be going extinct.

Variations observed amongst American populations can be partially explained in different ways. First of all, the sample collections from the United States all come from Cleveland, which is a large cosmopolitan city, whereas samples from South and Central Americas come from hospitals where the patients come from various parts of their respective countries. Most of the samples coming from Old World populations found in the New World are from the USA, but that could be a reflection of this difference in the type of collection. However, there is another plausible explanation for this. The prevalence in the USA has been decreasing (Stewart and Wild 2014) for the last 2 decades, whereas it is still extremely high in Latin American countries. High prevalence increases the possibilities for horizontal transmission and mixed infections, which can result in emergence of local populations, whereas a low prevalence tends to conserve the original imported populations.

The long-term evolution of *H. pylori* strains is in part driven by human history. Population migration events can make highly divergent bacterial populations meet and those populations can recombine to result in new populations. This evolutionary path can be of great importance in understanding host/pathogen interactions, as the creation of new populations or the eradication of old populations of *H. pylori* can be linked to prevalence of diseases such as gastric cancer.

It is interesting to note that the Old World populations identified by FineStructure and ChromoPainter matched the ones previously described (Falush et al. 2003). However, the increased number of strains and use of whole-genome sequencing instead of MLST allowed for the differentiation between two subpopulations which were

considered as one in previous studies (Falush et al. 2003): hspEuropeS and hspEuropeN. The two populations are closely related, but distinct in terms of ancestry, as hspEuropeS has a stronger input in African populations, and hspEuropeN has a stronger input in Asian populations. Study of the accessory genome in hpEurope strains also confirmed the hybrid origin of this population between Asian and African ancestry, which was demonstrated using MLST (Falush et al. 2003) and core genome analysis, with FineStructure. This is the first study based on complete accessory genome, as most studies are based on core genome analysis. This enabled us to investigate the relationship between core and accessory genomes and have a more complete picture of the evolution taking place in *H. pylori*. Accessory analysis of hybridisation ancestry is a good way to represent and analyse the prevalence of genes between 3 populations. However, it requires more than 20 strains from each population, which could be considered a limitation as samples are not always available in abundance.

In conclusion, we succeeded in showing that the genomic variability of *H. pylori* strains from the Americas reflects the migration history. Our study highlighted two distinct ancestral hpEurope, that we named hspEuropeN and hspEuropeS, spreading to the New World in different ways. These Old World populations mixed in the Americas with Asian and African strains to give birth to new populations, unique to the Americas. Our study studied both core and accessory genome, and we showed that both genomes were evolving in similar ways. Our observations on the link between *H. pylori* and human populations were concordant, showing the same tendancy in the different New World populations.

4 Rapid genomic evolution in *Helicobacter pylori* strains infecting mice

Helicobacter pylori can colonise the stomach for years without causing any symptoms, but often causes asymptomatic or symptomatic inflammatory responses (Correa 1988; Supajatura et al. 2002; D. J. Evans et al. 1995; Satin et al. 2000; Moran and Aspinall 1998; Pérez-Pérez et al. 1995; Unemo et al. 2005). This prolonged inflammation can also be responsible for the development of various gastric disorders, such as gastric cancer but also the less prevalent gastric MALT Lymphoma (ML). The exact mechanisms leading to gastric complications remain unclear (Bessède et al. 2015), but it is thought that the extreme diversity of the *H. pylori* genome could be one answer. A few studies have investigated the effects of long-term host colonisation on the *H.pylori* genome (Israel et al. 2001; Kersulyte, Chalkauskas, and Berg 1999; Kuipers et al. 2000). However, there is a need to link this long-term colonisation with not only the changes occurring in the bacterial genome, but also with the development of disease (e.g ML) and associated symptoms. Murine models of *H. pylori* colonisation have been established (D. H. Kim et al. 2003), with comparable symptoms to those observed in humans (She et al. 2003). This model is a good option both economically and biologically (S. Zhang et al. 2014), and so an infection study was carried out at the Centre National de Référence des Campylobacters et Hélicobacters (CNRCH) in Bordeaux (Chrisment et al. 2014).

The reasons why only part of the population infected with *H. pylori* develops symptoms and the mechanisms underlying the wide range of symptoms that can occur are still unknown. In the murine study previously mentionned (Chrisment et al. 2014), observation of lesions and markers of low-grade ML similar to those seen in humans were made regardless of the infecting strain, which would suggest the importance of host (mouse) rather than bacterial factors. However, to add to this complexity, symptoms were significantly more acute in one of the ML associated strains (B47). This suggests that the diversity observed in *H. pylori* population may contribute to the disease process. The genome evolution of a *H. pylori* strain during a long-term infection was previously studied in a primate model (Liu et al. 2015) and attempts were also made in humans (Avasthi et al. 2011), but these studies require very long incubation times to be relevant and focused on other pathologies. There is, so far, no

study of the evolution of *H. pylori* strains during the development of MALT lymphoma symptoms. A recent publication studied the differences between a specific experimental strain (SS1) before and after passage in mice exist but this study was only based on one strain, and highlighted variability (Draper et al. 2017). Studies on other strains could confirm or complete their observations.

This chapter uses the isolates from a previous study where neonatal thymectomised BALB/c mice were infected with *H. pylori* strains in order to assess the capacity of the strains to promote ML (Chrisment et al. 2014). Two of the strains used in this study, B38 and B47, were isolated from European patients suffering from ML. At each stage of this study the bacterial strains were isolated from mice and stored, offering the opportunity to sequence them and have a deeper look at the variations that occur in the bacterial genomic during the long-term colonisation of mice stomach.

This study aims to investigate two statements: i) A *H. pylori* strain evolves when changing its host niche from human to mice; and ii) A *H. pylori* strain infecting a stomach for a long time evolves alongside the development of ML symptoms.

4.1 Material and Methods

4.1.1 Isolation of *H. pylori* from mice

in vivo experiments took place at Laboratoire de Bacteriologie in Bordeaux (Chrisment et al. 2014), prior to this PhD. The two strains selected for this study (B38 and B47) were capable of colonising the stomachs of BALB/c mice. One of them, B38, went through a re-isolation step prior to colonisation experiment in order to obtain a sufficient amount of infecting bacteria. They were both CagPAI negative and were tested against two control strains also known for their ability to colonise mice but not associated with ML. Bacterial strains were cultured under microaerophilic conditions at 37°C. The strains were grown on selective agar plates and collected in Brucella broth medium. Six-week-old mice were fasted in order to facilitate bacterial colonisation, then gavaged for 3 consecutive days with the required strain at a dose of 10^8 *H. pylori* per mouse. When the mice were sacrificed (after 6 weeks, 6 months or 12 months post-infection), biopsies were used for culture and frozen at -80°C. *H. pylori* culture was performed from the collected stomach tissue, homogenised in PBS

and grown on a selective medium under microaerophilic conditions at 37°C for 3 to 10 days. These bacterial cultures were kept at -80°C until used for DNA extraction and sequencing (Figure 4.1).

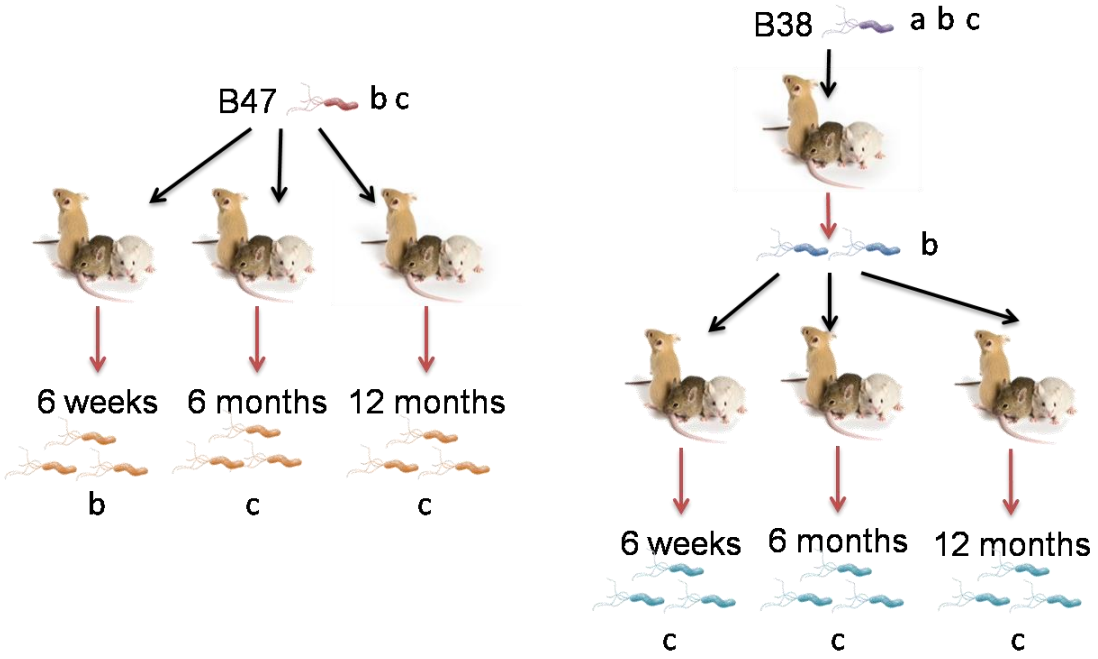


Figure 4.1: *In vivo* microevolution of *H. pylori*.
Original strains B38 and B47 were both isolated from human patients suffering from ML. B38 was passaged through a mouse prior to infection. Then both B38 and B47 were administered to mice and re-isolated at defined times (6 weeks, 6 months and 12 months) following infection. Infection is shown with black arrows. Isolation is shown with red arrows. The strain labelled a was already sequenced and published. All other strains were sequenced specifically in this genomic study. Strains labelled b were used in the change of host analysis. Strains labelled c were used in the long-term colonisation analysis.

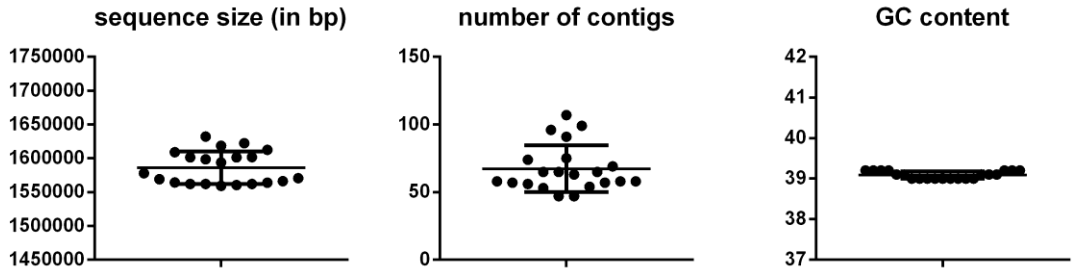
4.1.2 DNA extraction and sequencing

The B38 strain sequence was already publicly available. The remaining 21 other isolates were DNA extracted and their genomes sequenced (Figure 4.1). Total DNA was extracted using the QIAmp DNA Mini Kit (Qiagen, Crawley, UK) according to the manufacturer's instructions at Laboratoire de Bacteriologie, Bordeaux. The samples were whole-genome sequenced at Swansea University. Quantification of DNA was assessed with a Nanodrop spectrophotometer prior to sequencing. High-throughput genome sequencing was performed using a MiSeqSystem (Illumina, San Diego, CA), and *de novo* assembly was performed using Velvet (version 1.2.08). All contigs from the 22 strains were imported into the SheppardLab Hp Bacterial Isolate

Genome sequence database (BIGSdb http://zoo-dalmore.zoo.ox.ac.uk/perl/bigsdb/bigsdb.pl?db=sheppard_hpylori_isolates) for genomic analysis (K. A. Jolley and Maiden 2010).

4.1.3 Genomic analysis

The average number of contigs for the 21 newly sequenced strains was 67.3, average length was 1586210 bp and average GC content was 39.09 % (Figure 4.2). Size of the B38 genome was 1576758 bp, on a single contig, with a GC content of 39.2%.



A. B. C.

Figure 4.2: Distribution of the genomic characteristics of the 21 newly sequenced strains.
These characteristics were analysed using BIGSdb (K. A. Jolley and Maiden 2010). Characteristics shown are sequence size (**A**), number of contigs (**B**) and GC content (**C**).

Coding sequences (CDS) were identified for each strain using RAST (Overbeek et al. 2014) (Figure 4.3). The average number of CDS in those 22 strains was 1620.6. B38 derived strains had a smaller number of CDS (1611 on average) compared to B47 derived strains (1632 on average). This difference was significant (p-value < 0.0001).
A pan-genome was constructed with all loci present in at least one of the 22 strains used in this study, with the genes contained in the reference strain 26695 used as a reference gene list to facilitate identification of the genes (see details of the method in 2.3.4.1). Gene-by-gene alignment was performed by genome comparator (2.3.2) using the 1703 CDS sequences of the pan-genome as reference, and the alignments were exported from the database to investigate gene-by-gene variations. A neighbour-joining tree was built from the gene-by-gene alignment of the strains using FastTree v2 (Price, Dehal, and Arkin 2010) and annotated with iTOL v3 (Letunic and Bork 2016).

The genome comparator tool produced a results matrix which included gene prevalence, and different alleles. Key differences measured focused on i) adaptation to a new host; ii) changes during the long-term colonisation of this host (Figure 4.1).

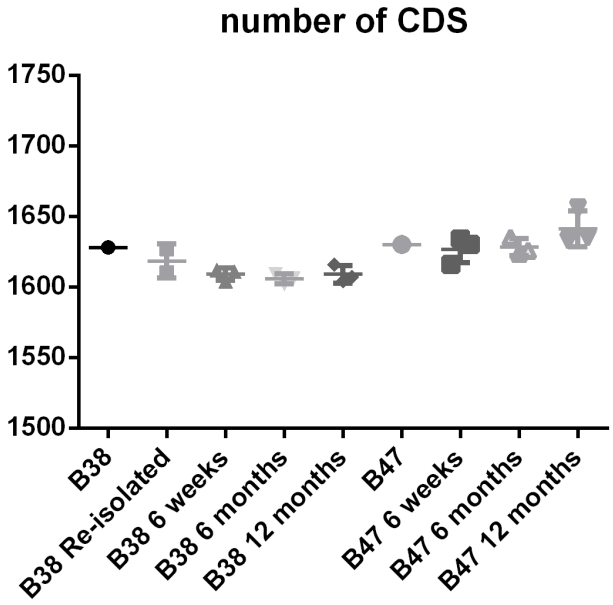


Figure 4.3: Number of CDS annotated with RAST in the 22 strains isolated from ML patients or after passage in mice.

Gene-by-gene variations were studied using BioEdit. Five types of changes were searched for: (i) single nucleotide polymorphism (SNP), corresponding to a change in only one position from one nucleotide to another; (ii) double nucleotide polymorphism (DNP), corresponding to a change in two neighboring nucleotides; (iii) Phase Variation (PV), corresponding to a change in the number of repeats of one or two nucleotide; (iv) Deletion or Insertion, inducing a gap in the alignment; (v) Change in the number of repetitions of a large sequence. PV and Deletion/Insertion are non-synonymous, as they are changing the reading frame. Change in the number of repetitions of a large sequence is non-synonymous, but usually results in no change in the reading frame. SNP and DNP can be either synonymous or non-synonymous. Analysis of corresponding amino-acid sequences was achieved based on translations from the gene sequences using the online tool ExPaSy.

4.2 Results

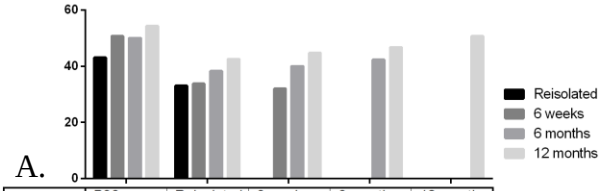
4.2.1 Population biology of the *H. pylori* dataset

Genomic variability was first verified on a neighbour-joining tree (Figure 4.4).

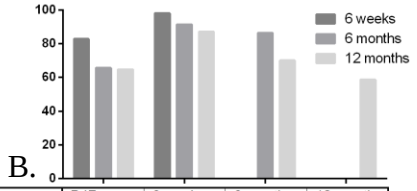
Figure 4.4: Neighbour-joining tree of the 22 strains isolated from ML patients or after passage in mice.
The tree was built using the alignment produced by the genome comparator tool from BIGSdb and annotated with iTOL v.3.There was a clear segregation between B38 isolates on the left side and B47 isolates on the right side, confirming the isolates came from two distinct strains.

The two strains used, B38 and B47, formed two distinct clusters, confirming firstly that they were different strains, as expected, and second that the genomic variations occurring during long-term infection in mice are smaller than the genomic difference between the two different strains. The results of a genome comparator performed with the pan-genome of these two strains and their variants confirmed these differences, with 1584 genes showing allelic variations between the two original strains or present in only one of them.

The gene-by-gene comparison of the passaged isolates compared to the isolates originally administered showed a higher variability within the B47 isolates than within the B38 isolates (64.7 to 82.7 genes presenting variations in B47 against 43 to 54.3 B38) (Figure 4.5).



	B38	Reisolated	6 weeks	6 months	12 months
Reisolated	43	33			
6 weeks	50.7	33.8	32.0		
6 months	50.0	38.3	40.0	42.3	
12 months	54.3	42.5	44.7	46.7	50.7



	B47	6 weeks	6 months	12 months
6 weeks	82.7	98.0		
6 months	65.7	91.3	86.3	
12 months	64.7	87.1	70.0	58.7

Figure 4.5: Average number of genes showing allelic variations between the different sets of strains derived from B38 (Panel A) and B47 (Panel B).

For B38 isolates, the number of genes presenting allele variations between isolates from a single time-point was lower than the number of genes presenting allele variations when compared to the original isolate. It was not the case for B47 isolates, in which the number of genes with allele variations between isolates from a single time-point was higher than the number of genes presenting allele variations when compared to the original isolate (Figure 4.5).

4.2.2 Evolution during change of host

Effects of the change of host on the strains is studied based on the genes changing between the original clinical strain and the batch of strains obtained after a first passage of 6 weeks in mice. A larger number of genes changed during the change of host in B47 derived strains (53 genes) compared to the number of genes changing during the change of host in B38 derived strains (38 genes). After selection of only genes with at least 2 identical alleles in the re-isolated strains that are different from the original strain allele and suppression of artefacts due to contig limits creating truncated genes, only 2 genes have changed in B47 and 10 genes have changed in B38. Artefacts due to contig limits were identified as genes for which the truncated gene was positioned at an extremity of one of the contigs. The sequence variations observed in B38 during change of host were of three types: SNP, PV and deletions/insertions (Table 4.1).

Table 4.1: Changes observed in genes during change of host in strain B38 or B47.

Gene	Change in	Type of change	Effect on Amino-Acid sequence
HP0217/*cgtA*	B38	PV	Change in the length of the protein
HP0251/*oppC*	B38	PV	Change in the length of the protein
HP0464	B38	PV	Change in the length of the protein
		SNP	Alanine to threonine
HP0499	B38	PV	Change in the length of the protein
HP0685	B47	PV	Change in the length of the protein
HP1054	B38	SNP	Tyrosine to Histidine
HP1088	B38	SNP	Arginine to Glutamine
HP1237	B38	1SNP	Synonymous
HP1243/*babA*	B38	insertion	Change in the length of the protein
		SNP	Glycine to Arginine
		SNP	Histidine to Tyrosine
		SNP	Threonine to Alanine
HP1251/*oppB*	B47	PV	Change in the length of the protein
HP1252/*oppA*	B38	PV	Change in the length of the protein
HP1365	B38	SNP	Histidine to Tyrosine

For strain B47, only PV was observed during change of host. In most of the genes, only one sequence variation was identified. However both a PV and a SNP were found in *HP0464*, and 3 SNP and a deletion/addition were found in *HP1243*.

Interestingly, one of the genes identified in B38 (*HP1252*, or *oppA*) was in close proximity to one of the two genes identified in B47 (*HP1251*, or *oppB*). Another gene, identified only in B38, coded for another protein from the Opp family: *HP0251*, or *oppC*. All three were identified as going through PV during host transfer. These genes code for an oligopeptide ABC transporter substrate-binding protein and two oligopeptide ABC transporter permeases. All the phase variations occurring in these three genes have had consequences in the length of the protein, causing a shorter version of the protein in either the original strain (*OppB* and *OppC*) or the reisolated strains (*OppA*).

HP0217, also known as *cgtA*, is a gene coding for a beta-1,4,N-acetylgalactosamyltransferase and has a PV in B38. Both the original strain version and the re-isolated strain version were shorter than the version found in the reference strain 26695. However the version present in the re-isolated strains is slightly shorter than in the original strain.

HP0464 codes for hsdR, a type I restriction-modification system endonuclease and has both a SNP and a PV in B38. Both the SNP and PV are non-synonymous, causing an amino-acid change from an alanine in the original strain to a threonine in the re-

isolated strains, and a shorter version of the protein in the original B38 strain than in the re-isolated strains.

HP0499 codes for a precursor for an outer membrane phospholipase A1 (pldA, or DR-phospholipaseA) and has a PV in B38. This PV causes a shorter version of the protein in B38 than in the re-isolated strains.

HP1054 is annotated as a hypothetical protein, but is likely to code for the murein hydrolase activator NlpD and shows a SNP variation in B38. This SNP was non-synonymous, causing an amino-acid change from a tyrosine in B38 to a histidine in the re-isolated strains.

HP1088 codes for a transketolase (tktA) and has a SNP variation in B38. This SNP was also non-synonymous, causing an amino-acid change from an arginine in B38 to a glutamine in the re-isolated strains.

HP1237 codes for a carbamoyl phosphate synthase small subunit (pyrAa) and has a SNP variation in B38. This SNP was the only synonymous SNP found in this B38 change of host study.

HP1243, also known as *babA*, is the gene that has the most variations in this study with 3 SNPs and 1 insertion. This gene codes for the outer membrane protein omp28. All 4 nucleotide variations found in this gene were non-synonymous. Effects of the insertion were important, creating a shorter version of the protein in the re-isolated strains compared to both B38 and 26695 versions. All 3 SNPs were similar in B38 and in 26695, differentiating the re-isolated strains from the human ones.

HP1365 codes for a DNA binding response regulator from the OmpR family and has a non-synonymous SNP in B38, causing an amino-acid change from a histidine to a threonine.

Finally, *HP0685* codes for a flagellar biosynthesis protein FliP and has a PV in B47. The Phase Variation in this gene resulted in a shorter version of the protein in the re-isolated strains than in the original B47 strain.

4.2.3 Evolution during long-term colonisation

Changes during long-term colonisation of mice were studied. A larger number of genes changed during long-term colonisation in B47 derived strains (79 genes) compared to the number of genes changing during long-term colonisation in B38 derived strains (63 genes). From these, genes for which the change was caused only by the change of host, artefacts caused by contig limits, and genes that were not

showing at least 2 identical alleles among the strains from a same time-point were removed. Artefacts due to contig limits were identified as genes for which the truncated gene was positioned at an extremity of one of the contigs. This selection left 3 genes in B47 derived strains and 4 genes in B38 derived strains. Changes included SNP, DNP, PV and long sequence repetition variations (Table 4.2). The two long sequence repetition variations observed concerned the same pattern of 21bp, inducing a repetition of the amino-acid sequence DDLRVNY.

Table 4.2: Changes observed in genes during long-term colonisation of mice in strain B38 or B47.

Gene	Change in	Type of change	Effect on Amino-Acid sequence
HP0379	B38	SNP	Aspartic acid to asparagine
		Number of repetitions	DDLRVNY 5 to 14 times
HP0629	B47	SNP	Alanine to Aspartic acid
HP0651	B38	SNP	Aspartic acid to asparagine
		Number of repetitions	DDLRVNY 5 to 14 times
HP0855	B38 & B47	PV	Change in the length of the protein
HP1041	B47	SNP	Glutamic acid to Alanine
0010_8940_0104	B38	PV	Change in the length of the protein

One of the genes, *HP0855*, presented a variation linked to long-term colonisation in both B47 and B38 derived strains. Again, like in the change of host, the variation observed in this gene was a PV. *HP0855* codes for an alginateO-acetylation protein (algI), with functions linked to cell wall, membrane and envelope biogenesis.

Two of the genes that were identified as having evolved during long-term colonisation of mice with B38 are especially interesting: *HP0379* and *HP0651*. Both code for a fucosyltransferase, which has a crucial role in LPS biosynthesis. Each of these two genes had a SNP placed inside the same short sequence, replacing an aspartic acid with an asparagine in the amino-acid sequence. In addition to this, they also had a unique type of variation: there was a repetition of a small sequence (21bp), resulting in the amino-acid repeat sequence DDLRVNY from 5 to 14 times. This has been described in another study (Rasko et al. 2000). More investigation was necessary on these sequences, as they were often found on the extremities of contigs. Completion of the missing parts from the genome comparator results for all the versions of *HP0379* in B38 derived strains was possible, but not for all the versions of *HP0651*.

Two different endings were found in the complete sequences for *HP0651*, which did not allow us to infer the end of the sequence based on the sequences available. However, the contig limit was placed after the repetitions in all these incomplete genes, thus the count of the number of repetitions is robust.

The final gene presenting variations in B38 has not been described yet. This gene was named *0010_8940_0104* by the RAST annotation, but it was not an nBLAST match for any of the 26695 genes during the pan-genome construction. The variation observed was a PV, leading to variations in the length of the protein. A pBLAST of the longest version found in our study against the proteome of 26695 matched for a type IIG Restriction Modification System with 98% of coverage for 91% identity.

Genes with variations in B47 included *HP1041*, which code for the flagellar biosynthesis protein FlhA. This gene contained a non-synonymous SNP resulting in the change from a glutamic acid in the B47 original and 6 weeks re-isolated strains to an alanine in the 6 months and 12 months re-isolated strains.

The final gene, *HP0629*, also contained a non-synonymous SNP, inducing a change from an alanine in B47 to an aspartic acid in the 6 months re-isolated strains. Its function was unknown.

During this long-term colonisation, lesions of the stomach were assessed (see (Chrisment et al. 2014) for details). Mucosal inflammation was present in all infected mice 12 months post-infection, and absent from the non-infected mice. However the level of lymphoid infiltrates was only statistically different from the non-infected mice for mice infected with B47. These infiltrates were associated with lymphoepithelial lesions, which is a signature of a lymphoma stage. Lesions were more extensive and more severe in B47 infected mice compared to B38 infected mice.

4.3 Discussion

The variability between the original and re-isolated strains after the first passage in mice was quite low compared to what would have been expected from highly recombining bacteria. This suggests that the change of host from a human with ML symptoms to healthy mice does not affect the strains greatly. This observation confirms that studying ML symptoms provoked by *H. pylori* strains isolated from ML-suffering patients using a mouse model is suitable and does not infer changes too important in the bacterial strains. This makes this model not only good to study the development of ML in the stomach tissues, which was the aims of the Chrisment et al

(2014) study, but also to study the bacterial genome and its evolution during colonisation and development of symptoms.

However, despite our results showing genomic variations in *H. pylori* strains colonising mouse models, variability was extremely high between samples from the same time-point, mainly due to the fact that the strains are evolving in different hosts, and mice are sacrificed at each time point making it difficult to be certain that the intermediate time point is truly an intermediate evolution between the original strain and the strain found in the 12 months post-infection stomach. Despite the fact that a single strain was used to infect the mice, it is also possible that at the re-isolation time, a variety of strains were present, undergoing different evolutionary pathways. Only one single colony was isolated for each mouse, and this hidden variability could have an effect on the results. A wider range of single colonies from different biopsies (n=10) of a same animal would help assessing this variability.

The design of this study did not include an evolutionary control, in the form of the strains cultured outside a new host. Such an experiment would have been needed in order to ensure that the variability observed was indeed the result of the evolutionary response of the bacteria to its new host, and not spontaneous variability (Jee et al. 2016; Draper et al. 2017).

Another limitation of this study was the high number of artefacts identified by the genome comparator. This could be due to the quality of the sequencing and assemblies. Transport of the samples was not optimal, due to a transporter issue, which could have affected the quality of the samples, causing a drop in the quality of the sequences. However, the sequences were of expected size and content, therefore the results once the artefacts were removed should be trusted. Contig sizes were a limitation in gene-by-gene alignment methods used in this study. Increased variability observed in B47 derived strains compared to B38 derived strains, and the increased number of artefacts due to contig limits in B47 could be explained by the fact that the B38 genome was a complete genome sequence assembled in a single contig, whereas B47 was splitted in different contigs. This could be solved by using another method for sequencing, such as Pacific Bio Systems. However, this technology was not available in Swansea but is an opportunity for a future collaboration.

This high variability issue could be improved by increasing considerably the number of samples, to allow statistics to be performed on the genomes. Indeed a maximum of 3 isolates for each time point were available to us for this study, which is too small to

perform any sort of statistical measure and can hide some of the variability by the chances of not picking them up in our sample population.

Another way to handle the variability issue would be to use a larger animal model which would allow the use of endoscopies instead of sacrifices. Indeed, mouse model present limitations. Their small size is one of them, as endoscopies are not impossible to perform but involve a lot of critical steps and require specific setups (Brückner et al. 2014). The life span is also shorter than the average duration of a *H. pylori* chronic infection in humans, which limits the study of true long-term colonisation and its effect on the bacterial genome. Mice also are poor models to study inflammation which is a crucial part of the *H. pylori* infection outcomes. Primates can be considered, as shown recently (Liu et al. 2015), but it is a costly model requiring more complicated ethical project licences. In addition, the adaptation of human strains to primates hosts has been less studied than for mice (Guo et al. 2014; S. Zhang et al. 2014; Ameri Shah Reza et al. 2012). However, it does present the advantage of a longer life-time, closer to the human one, and endoscopies are achievable more easily on an animal of such scale. On an intermediate level between mice and primates, one could also consider gerbils. Cases of infections with *Helicobacter* species have been described in gerbils, which could also constitute a suitable model (Asim et al. 2015; Kodama et al. 2005). In order to study the evolution of *H. pylori* strains according to the time spent in a specific host, it is also possible to study.families of patients infected with *H. pylori*. Indeed, families are often infected with a common strain, or are transmitting their infecting strains to other members of the family. Therefore sequencing strains isolated from patients belonging to a same family is a way to approximate longitudinal study of the bacterial genome (Raymond et al. 2004; Osaki et al. 2015; Raymond et al. 2008; Kivi et al. 2007).. At rare occasions, it is also possible to obtain pairs of isolates from a same patient at two different timepoints, which is a perfect opportunity to study genome evolution during human infection (Kennemann et al. 2011).

Despite those limitations of our model, a number of genes changed in this study, some of which presenting interesting patterns of evolution, or linked to functions of importance in virulence.

Outer membrane proteins are present on the cell membrane and are therefore likely to be in contact with the host cells and adapting to a new host in the context of the

immune response. A number of outer membrane proteins were highlighted in our change of host study, such as BabA or members of the Opp family. Interestingly, *babA* was also highlighted in a study of long-term colonisation performed on a primate model (Liu et al. 2015), and members of the hop family (comprising *babA*) were identified as showing an increased frequency of imports after 3 years in a human host (Kennemann et al. 2011). In their study, a PV was identified in this gene, which was not the case during the current study, where an insertion and 3 SNP were identified. However, the fact that this gene was highlighted by both these studies highlights the importance of *babA* as a virulence factor for colonisation of a host. *BabA* is one of the main virulence factors in *H. pylori* (VFDB 2017). The PV highlighted in members of the Opp family highlight a possible hot spot for short-term evolution of *H. pylori* strains during change from human to mouse hosts. These phase variation mechanisms have been described before in *H. pylori* and are well-known mechanisms for rapid adaptation (Bergman et al. 2006). It is not surprising to find them in our experimental conditions and it leads us to hypothesise that this set of genes coding for outer membrane proteins is a potential hot spot for evolution in the context of a change of host. Changes in *oppA* and *oppB* were also described in a study comparing the famous mouse-adapted strain SS1 with the pre-change of host equivalent PMSS1 (Draper et al. 2017). A number of changes were also identified in the long-term colonisation study, such as a gene linked to motility, *HP1041*, which was affected by a non-synonymous SNP. Another interesting mechanism of evolution highlighted in the long-term colonisation study was the number of repetition of a 21bp sequence in two fucosyltransferase-coding genes. These change highlighted by our study in *futB*, a gene involved in LPS modification, were also highlighted in the SS1/PMSS1 study (Draper et al. 2017), confirming that these genes are hot spots. of evolution, suggesting that the bacteria adapt to its host's stomach while provoking changes in it.It is interesting to note that the B47 strain showed a smaller number of genes having evolved during change of host and long-term colonisation, compared to B38. This B47 strain was also the one which provoked the most advanced and acute ML-like symptoms in a significant number of mice (Chrisment et al. 2014). This could suggest that B47 was already well-adapted to the ML environment, therefore having a faster effect on its new environment to turn it into a ML-like mucosa. Although such an adaptation cannot be proven with only these strains, an additional study could be carried out with a larger number of strains.

In conclusion, changes were observed in the bacterial genome during the passage from human to mice host. Some of these changes confirmed the variability observed in the study on SS1/PMSS1 (Draper et al. 2017). Strains also evolve during long-term colonisation, and evolution seemed to be more important in the strain causing the highest damages to the mice stomach.

5 A Genome Wide Association Study of *Helicobacter pylori* in cancer-causing European strains

Helicobacter pylori is a Gram negative bacteria that colonises the human stomach. Despite the fact that this bacteria has colonised human stomachs for at least 100,000 years (Moodley et al. 2012), it was only discovered in 1984 (B. Marshall and Warren 1984). The link between the presence of *H. pylori* in the stomach and the development of peptic ulcer disease is well known (B. J. Marshall n.d.; A. C. Smith 1989; B. Marshall and Warren 1984), and this organism can go on to cause gastric cancer in some patients (Parsonnet et al. 1997). *H. pylori* can colonise the stomach for years without causing any symptoms (Peek and Blaser 2002). However, the prolonged low grade inflammation associated with this colonisation could be responsible for the development of various gastric disorders, such as gastric cancer (Wroblewski and Peek 2016). Gastric cancer (GC) is the third most common cause of cancer deaths worldwide, causing approximately 723,000 deaths every year (Ferlay et al. 2015), and treatments are not yet optimal (Shi and Gao 2016). Therefore, recent guidelines have encouraged complete clearance of the bacterial infection, in an attempt to prevent future complications (Ierardi et al. 2013). However, the procedure for eradication of *H. pylori* is not easy, and increased antibiotic resistance has been reported throughout the World (Binh et al. 2015).

Despite the above, some studies have shown a possible positive impact of *H. pylori* infections in some health-related issues (Whalen and Massidda 2015). The definition of health given by the World Health Organization in 1948 (WHO 1948) states that "Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity". Considering the large number of cases where colonisation by *H. pylori* does not provoke serious symptoms, systematic eradication of *H. pylori* might not always be beneficial to the patient's health. The long-term co-evolution between humans and *H. pylori* might be beneficial for both the bacterium and the human host, and the effects of complete eradication of the bacteria are unknown, which could result in the rise of other health issues, that could be more complex to treat (Vincent et al. 2013). What if it was possible to predict whether a strain was going to cause cancer? Would it then be necessary to eradicate the bacteria if sufficient relevant proof showed that complications were unlikely? Improved

antimicrobial stewardship could reduce the resistance issue, while improving patient comfort with low risk of cancer.

The exact mechanisms leading to complications remain unclear (Bessède et al. 2015), but it is thought that the extreme variability of the *H. pylori* genome could be part of the answer. Gastritis is always the first symptom after an *H. pylori* infection (Sugano et al. 2015). After long-term colonisation, symptoms can evolve towards more serious issues, or remain stable. Different pathways towards different complications can occur. Antral-predominant gastritis is often associated with a higher level of acid production, and is more likely to evolve into duodenal ulcers or MALT lymphomas, whereas corpus-predominant atrophic gastritis is associated with a lower level of acid production, and can lead to gastric ulcers or GC. Pathways towards GC from atrophic gastritis include the development of IM and dysplasia. Atrophic gastritis is reversible to asymptomatic gastritis, but can also progress towards gastric cancer (Chung et al. 2005). Even though some genomic factors are known to be associated with cancer, such as CagPAI genes and VacA, these genes are not sufficient to explain the development of cancerous lesions in some of the patients, while others only get superficial symptoms (da Costa, Pereira, and Rabenhorst 2015). There is a need for deeper analysis of the differences between cancer-causing strains and non-cancer causing strains. Identification of risk genotypes in the bacterial strains could lead to new ways to face treatment of the infection.

Development of GC is not only due to bacterial genomic factors, but could also be driven by host genetics (Cristescu et al. 2015) or the environment, and most certainly by a combination of those three factors. One study identified an increase in the risk of cancer when there is divergence between the host and bacterial population (de Sablet et al. 2011). This chapter focuses on bacterial genomics factors, but host and environmental factors will be used to explain the variations found in the bacteria genome. Another element to keep in mind is the fact that in the event of a *H. pylori* infection, the stomach is not colonised by a unique strain, but a mix of strains, sometimes closely related, but sometimes resulting from a mixed infection (Kibria et al. 2015; J. W. Kim et al. 2004; Ben Mansour et al. 2016). Therefore linking bacterial genotypes with gastric cancer is not straightforward. Studies can be biased by random selection of certain strains, and screening applications that could result of such studies in clinics could also suffer from this fact and produce false results.

The GWAS method has previously been used successfully to study human genetics. It has recently been applied to bacterial genomics; through adaptation of the original method to take into account specificities of the bacterial genomes (Sheppard et al. 2013). The challenge faced by using GWAS in *H. pylori* is its extremely high genomic variability. Indeed, *H. pylori* strains are not organised in clusters like the related species *Campylobacter*. Two strains isolated from two different patients are often extremely different, with allelic differences between common genes but also genes which are absent in one and present in the other (Alm and Trust 1999). There is a clear population structure in *H. pylori* genomes, which is strongly associated with the geographic or ethnic origin of the patient from whom the sample was isolated. This creates difficulty when applying methods currently used, increasing the risk of obtaining false positive hits. To address this population issue, this study will only consider European strains, so that the number of strains is sufficient while reducing the effect of population structure. Methods of GWAS based on ClonalFrame, which were proven efficient with other bacteria, are not optimal for *H. pylori*. Therefore, this ClonalFrame method will be compared with another method named 'bugwas', which does not rely on a pair-wise selection of strains on a tree. Results from bugwas are included in a first-author submitted publication which has not been accepted or published at the time of writing.

In this chapter, the feasibility of two different GWAS methods on *H. pylori* datasets will be verified. Specific genomic traits linked to the progression of GC will be identified using GWAS. Finally a risk score will be used to lead the way towards a better targeting of strains with a higher risk for triggering GC.

5.1 Materials and Methods

5.1.1 Dataset

All 578 publicly available genomes from *H. pylori* strains and 198 strains sequenced in Swansea or by collaborators were used to search for the complete list of genes present in the *H. pylori* reference strain 26695 using a nBLAST (ref Section Genome Comparator). Strains that were not from a human clinical source, had an unusual sized genome or number of genes were removed (2.3.1), and a tree was built from an alignment of the remaining strains using FastTree v2.0 (Price, Dehal, and Arkin 2010)

to identify clones, leaving a total of 565 strains. Metadata was retrieved when available from existing publication or patient information collected with the strains. Those 565 strains were used in a FineStructure (Lawson et al. 2012) analysis in order to identify large populations (2.3.5).

5.1.2 Genome-wide association study based on ClonalFrame

This GWAS method was based on one previously used in other bacteria (Sheppard et al. 2013). This relied on the construction of a pan-genome from the dataset chosen and ClonalFrame (Didelot and Falush 2007). A dataset comprising only strains from hpEurope populations (identified by FineStructure analysis) was used. Strains isolated from patients with gastric cancer (GC), Intestinal Metaplasia (IM) and Atrophic Gastritis were part of the Cancer group. Strains isolated from patients with gastritis or labelled as normal were part of the Non Cancer group. Strains with unclear or unknown pathology, or with pathology which were not part of the cancer pathway, were discarded, leaving only 196 strains. A tree of this data set was built using a core genome alignment with FastTree v2.0 (Figure 5.1).

Pairs of strains were selected uniformly spread across the tree to create two replicate datasets of 30 and 31 pairs of strains respectively. The pan-genome was created with these 122 strains (Appendix E) using the method described in 2.3.4.1. Briefly, the complete genomes of the strains were downloaded from BIGSdb (Overbeek et al. 2014) and annotated using RAST (Overbeek et al. 2014). Similarity between each pair of open reading frames was checked through a BLASTn (Altschul et al. 1990) search and a list of all the genes present in at least one strain of our dataset was created (Sheppard et al. 2013). This pan-genome was then used on the two replicate datasets as the reference gene list for the GWAS (2.3.6.1). Then, genes were split into 30bp words, the prevalence of the words in each group of strains (cancer or non-cancer) was assessed using a ClonalFrame based model (Méric et al. 2014; Sheppard et al. 2013). A Fisher's exact test was used to determine the association score for each word, with the following null hypothesis: this word is present in all isolates equally. This association score was used to select hits which were then mapped back to the reference pan-genome to identify the corresponding genes. Those genes were then analysed individually to investigate their function.

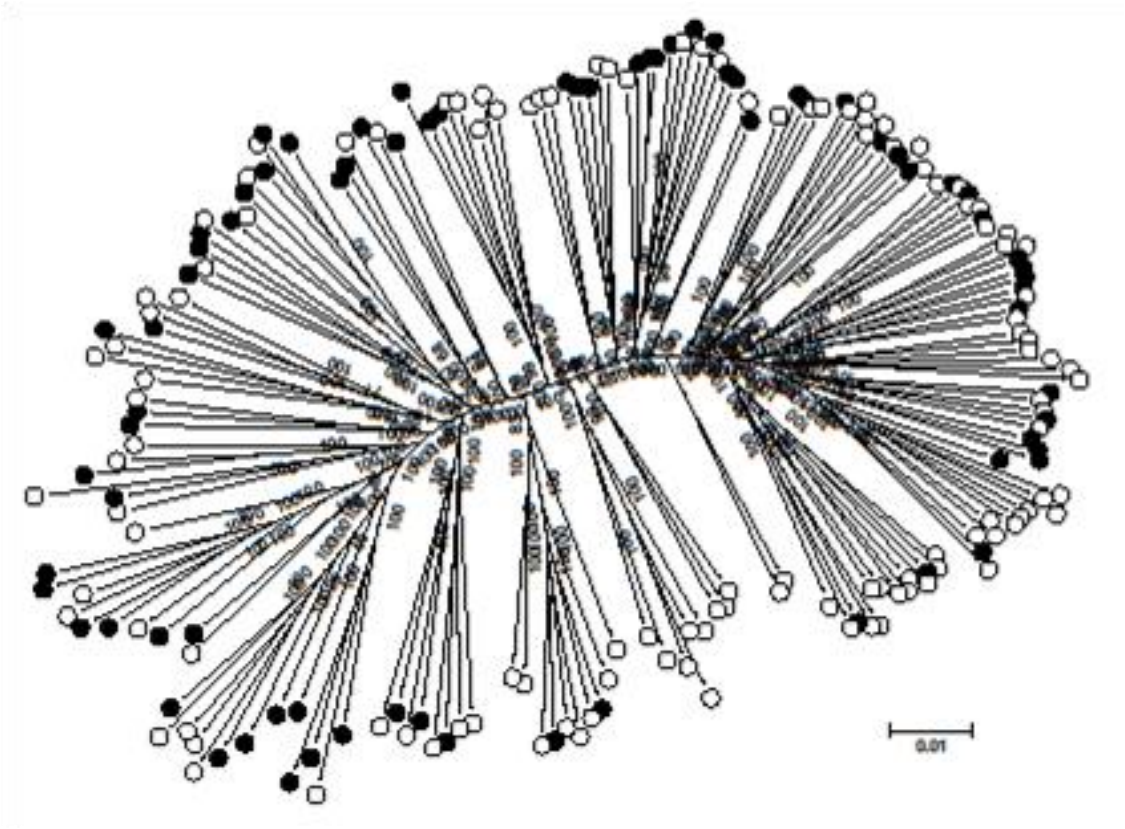


Figure 5.1: Neighbour-joining tree based on whole genome sequence alignment of 196 strains from Europe used in the ClonalFrame GWAS method.
Leaves are labelled according to cancer status (Cancer group = black circle, non-cancer group = white circle).

5.1.3 Genome-wide association study based on bugwas

An alternative GWAS method was used for the larger dataset available (including patient data) from a unique FineStructure large population (hpEurope). It was composed of 173 strains (Appendix E). Patient information associated with those strains and composition of the 3 pathology groups are summarised in Table 5.1.

Table 5.1: Composition of the GWAS groups used in the bugwas method.

Pathology Group	***Precise Phenotype***	***Number of strains***
Gastric cancer	Cancer	11
	Gastric Cancer	36
	GIST	1
	Stomach fundus tumor	1
	Total	**49**
Progressive towards cancer	Intestinal Metaplasia and Atrophy	1
	Intestinal Metaplasia	2
	Metaplasia	19
	Atrophic Gastritis	14
	Control (with atrophic gastritis)	16
	Total	**52**
Non atrophic gastritis	Asymptomatic	6
	Control (without atrophic gastritis)	23
	Gastritis	40
	Normal	3
	Total	**72**

The three groups were (i) Gastric cancer (GC), Progressive towards cancer (Prog) and Non atrophic gastritis (NAG). Only diagnosis information was used. For strains with more information available, full ethics were obtained by collaborators who collected the strains.

A tree based on the alignment of these 173 strains was constructed with genes from the reference strain 26695 using FastTree v2.0 and annotated using iTOL v3 (Letunic and Bork 2016) (Figure 5.2).

This GWAS method was conducted using a pipeline recently applied in another study by Koji Yahara (Suzuki et al. 2016). Two different methods based on the bugwas package (Earle et al. 2016) were used. The first was a k-mer-based method, in that aspect similar to the ClonalFrame approach (Sheppard et al. 2013), in which the genome sequence of each isolate was fragmented into unique overlapping 31-bp DNA 'words' that were used to identify genetic variations (2.3.6.2). This method was based on a pan-genome, built using Roary (with default parameters) after annotation by prokka (2.3.4.2).

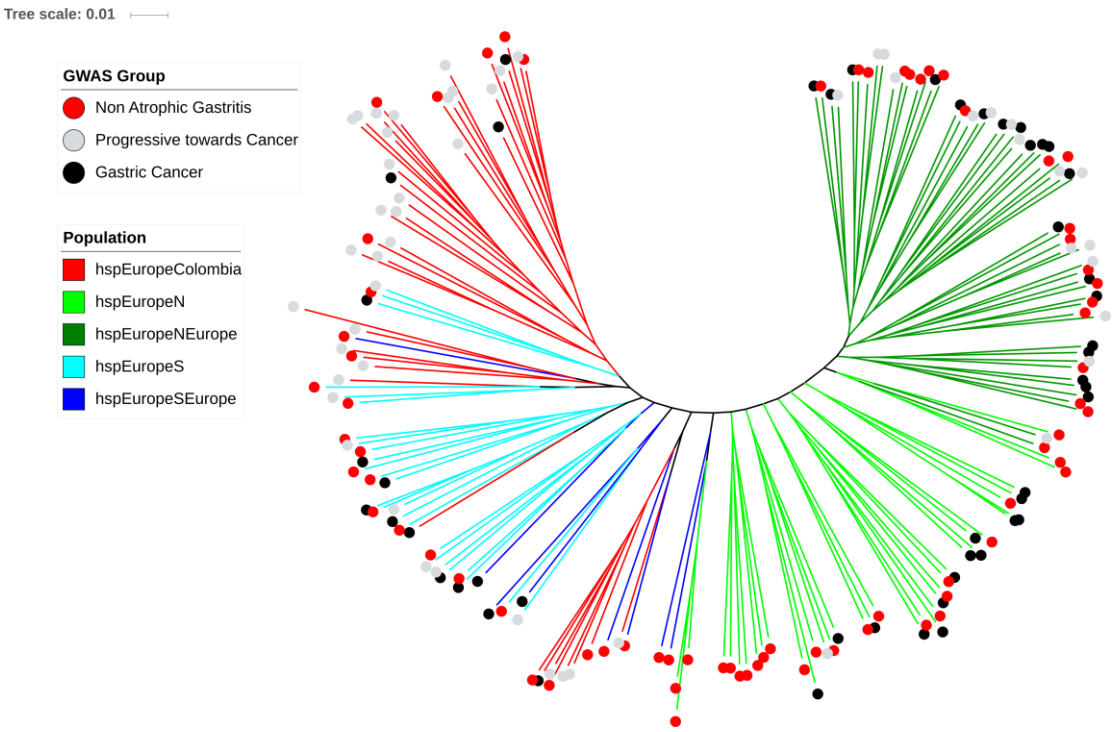


Figure 5.2: Neighbour-joining tree based on whole-genome sequence alignment of all 173 strains from hpEurope derived populations used in the bugwas method.
Colour of the branches represents the population identified by FineStructure, and leaf labels represent the GWAS group in which the strain was classified based on the pathology of the patient. An online version is available at this adress : http://itol.embl.de/tree/8110356195775515108283 66#

The second method was based on SNPs instead of k-mers (2.3.6.2). This method relied on a global alignment performed using ELS37 strain as a reference, a CagPAI positive GC strain. Only the SNPs belonging to a CDS were considered in the analysis, for consistency with the k-mer based method. It also had the advantages of being less sensitive to the high variable nature of the *H. pylori* genome, but does not pick up genes that are absent from the reference strain ELS37. Therefore, both methods were used in combination.

As GWAS requires a binary dataset, and the cancer isolates were spread into 3 groups (NAG, Prog and GC), two different datasets were used: (i) NAG vs rest and (ii) GC vs rest (Figure 5.3).

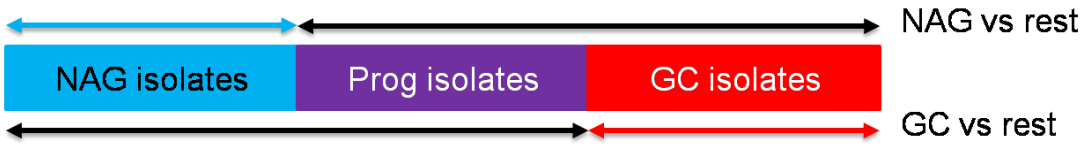


Figure 5.3: Composition of the two GWAS binary datasets.

Selection of hits for the bugwas method was made based on the odds ratio (20% difference was used as a cut-off) and p-value. The hits were then analysed individually to investigate their function and the effect of the genomic variations identified.

5.1.4 Analysis of gene hits

All the genes containing associated elements identified by the bugwas method were investigated individually using BioEdit. Effects on the amino-acid sequence were also checked using two distinct alignments: a global alignment based on a cancer strain (ELS37), obtained from the SNP GWAS, and a gene-by-gene alignment obtained from BIGSdb, using the sequences from ELS37 as reference. These alignments were used to differentiate synonymous and non-synonymous hits, using the ExPASy translate webtool (http://web.expasy.org/translate/). Non-synonymous hits were further studied by creation of figures showing the proportion of amino-acids in each position according to the GWAS group of each strain, using WebLogo (Crooks et al. 2004).

5.1.5 Non-synonymous enrichment

The ratio of non-synonymous SNPs to the total number of SNPs was calculated in the GWAS hits from the bugwas method. The same ratio was calculated in the 7 MLST genes: *atpA*, *efp*, *mutY*, *ppa*, *trpC*, *ureI*, *yphC*. This calculation was made using the "highlight variable sites" function in MEGA7 and was limited to SNPs present in at least 15% of the strains.

5.1.6 Risk score

The most significant GWAS hits from the bugwas method (level of significance of 10^{-6}) were used for calculation of a rudimentary risk score. First, the correlation between the presence of a risk or safe genotype A and the presence of another risk or safe genotype B was verified for each isolate pair, using a Pearson's correlation test. A Pearson's correlation of more than 0.9 with a p-value < 0.05 was used to define correlated genes. For each genotype, a genotype score (g_s) was determined as such: (i) For accessory variation: 1 if the gene is present, -1 if the gene is absent; (ii) For allelic variation: 1 if the risk genotype is present, -1 if the safe genotype is present, 0 if another genotype is present. The weight of correlated genotypes was then brought to the weight of one genotype for calculation of the risk score (using average genotype score). The risk score was calculated using this formula:

$$Risk\ score = \sum_{for\ each\ genotype} g_s \times -log^{10}(pvalue)$$

5.2 Results

5.2.1 GWAS based on Clonal Frame

GWAS based on ClonalFrame showed an abnormally high number of hits. 1724 genes contained hits with a p-value below 0.05, out of the 2240 genes of the pan-genome (Figure 5.4).

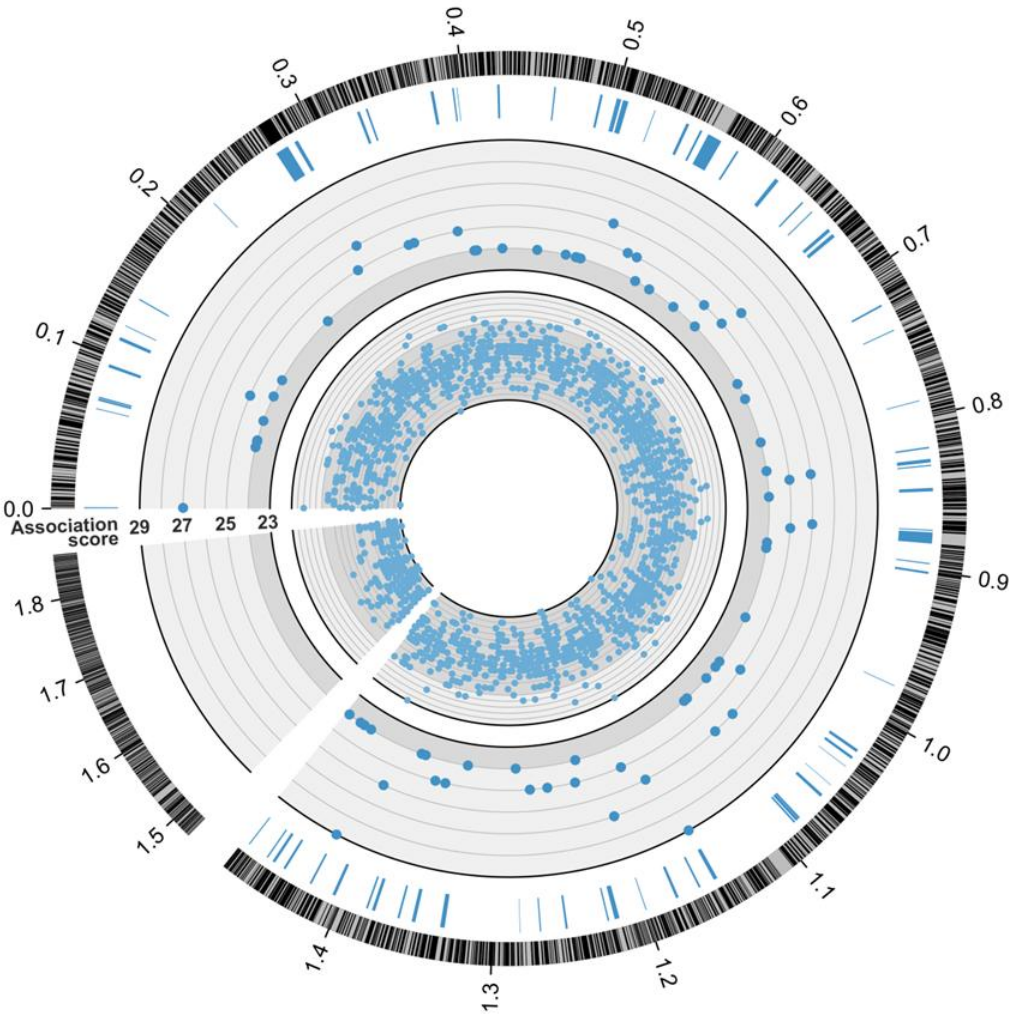


Figure 5.4: Results of the ClonalFrame based GWAS on two datasets of 30 and 31 pairs of strains highlighting differences between cancer-related and non-cancer-related strains.
Portion of the pan-genome between positions 0 to 1.485 correspond to the reference genome 26695. The remaining section corresponds to genes from the pan-genome which did not match genes from 26695. Blue bands indicate positions of genes with an association score over 24. Inner association score circle represent all the hits from the ClonalFrame based GWAS. Outer association score circle represents association scores restricted to over 24. Figure was built using R.

Those hits were filtered by association score to reveal interesting genes. CagPAI genes were included in the 71 gene hits with association scores of more than 24 (Appendix F). Of the 71 genes, 13 were hits in the Virulence Factors Database (Table 5.2), among them two of the most important CagPAI genes: *cagA* and *cagE*. *flgL* was the highest association score recorded amongst genes referenced in the Virulence factors database.

Table 5.2: Hits with an association score over 24 matching virulence factors from the Virulence Factors database (VFDB, 2017).

Gene Tag	Average Association Score	Hits in VFdB	Functional Group
HP0068	24	*ureG*	Buffering of gastric acid
HP0099	24	*tlpA*	Motility and chemotaxis
HP0295	26	*flgL*	Motility and chemotaxis
HP0529	24	*virB6/cagW*	CagPAI and typeIV SS
HP0544	24	*virB4/cagE*	CagPAI and typeIV SS
HP0547	24	*cagA*	CagPAI and typeIV SS
HP0685	24	*fliP*	Motility and chemotaxis
HP0867	24	*lpxB*	LPS
HP1031	24	*fliM*	Motility and chemotaxis
HP1119	24	*flgK*	Motility and chemotaxis
HP1177	24	*sabB/hopO + sabA/hopP + hopZ + babA/hopS + babB/hopT*	Outer membrane protein
HP1243	24	*babA/hopS + babB/hopT*	Outer membrane protein
HP1582	24	*pdxJ*	Motility and chemotaxis

In addition to these 13 genes, 7 genes were also identified as being linked to functions related to virulence, based on their RAST annotations (Table 5.3).

Table 5.3: Hits with an association score over 24 with gene functions linked to virulence

Gene Tag	Gene Name	Average Association Score	Reason of interest
HP0289		25	Outer membrane protein
HP0605	*hefA*	25	Outer membrane protein
HP0610		24	Cytotoxin (VacA paralog)
HP0920		26	Membrane protein
HP0922		25	Cytotoxin (VacA paralog)
HP1027	*fur*	24	Buffering of gastric acid
HP1156	*hopI*	26	Outer membrane protein

HP1156 (*hopI*), *HP0289* and *HP0605* (*hefA*) code for outer membrane proteins and *HP0920* codes for a membrane protein. Two genes, *HP0922* and *HP0610* code for cytotoxins and were identified as VacA paralogs. *HP1027* (*fur*) is known to be associated with the buffering of gastric acid. Assignment of functions to the genes identified with this method showed important links to motility and adhesion (outer membrane protein) (Figure 5.5).

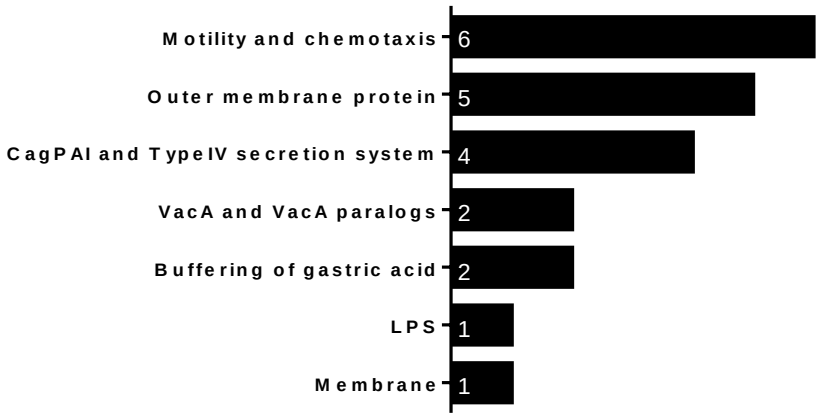


Figure 5.5: Assignment of functions to genes identified by ClonalFrame based GWAS performed on 122 strains belonging to hpEurope derived sub-populations.

Prevalence of the 71 genes with an association score over 24 in the complete dataset (combining the 2 replicates) highlighted two different types of genomic variations: i) accessory variations, with prevalence of the genes showing a difference between cancer and non-cancer groups of more than 0.1, and ii) allelic variations, with identical prevalence between cancer and non-cancer groups. Of our 71 genes, 5 show accessory variations (Figure 5.6).

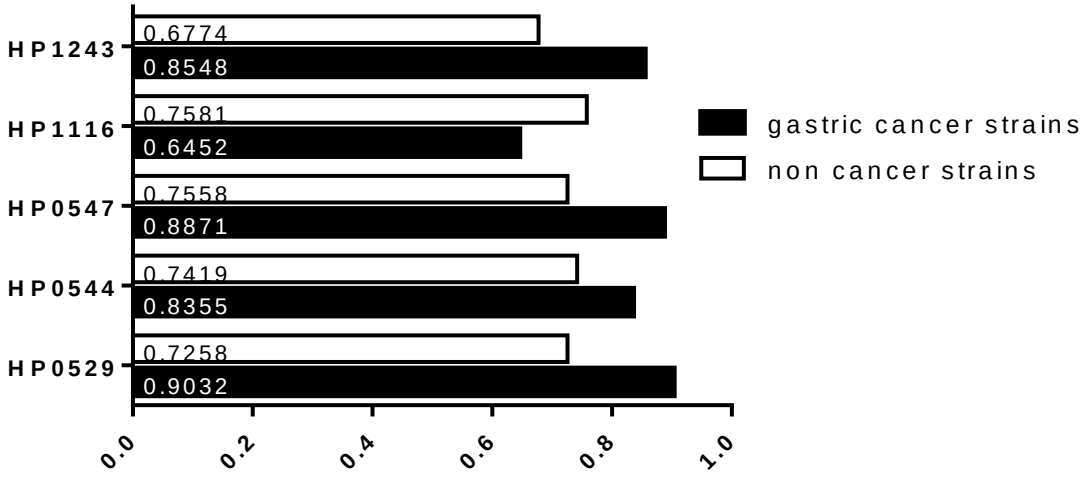


Figure 5.6: Prevalence of top hit genes from ClonalFrame based GWAS presenting an accessory variation.

Of these genes, three (*HP1243*, *HP0544* and *HP0547*) were previously identified as hits in the Virulence Factor db and code for proteins with functions of high importance in the CagPAI pathway: *babA*, *cagA* and *cagE*. Another gene from the CagPAI island also shows accessory variations: *HP0529*. The presence of these 4 genes was higher in the cancer strains than in the non-cancer strains in this dataset. Interestingly and in contrast, the last gene, *HP1116*, showed higher presence in non-cancer strains than in cancer strains. This gene codes for a hypothetical protein.

5.2.2 GWAS based on bugwas

The bugwas GWAS method used 4 different analyses involving two methods (k-mer or SNPs) on two datasets (GC vs rest or NAG vs rest) (2.3.6.2). In total, 642 hits (432 SNPs and 210 k-mers) in 32 genes with a p-value $\leq 10^{-5}$ (Table 5.4), and 118 hits (64 SNPs and 46 k-mers) in 12 genes with a p-value $\leq 10^{-6}$ (Figure 5.7, Table 5.4).

Table 5.4: Summary of the hits obtained in 4 bugwas based GWAS performed on 173 strains from hpEurope derived sub-populations according to patient pathology.

GWAS experiment	Number of hits with p-value		Number of hits with p-value	
	$\leq 10^{-5}$	$\leq 10^{-6}$	$\leq 10^{-5}$	$\leq 10^{-6}$
GC vs others (k-mer)	166	39	20	6
NAG vs others (k-mer)	44	15	10	2
GC vs others (SNP)	237	33	4	3
GC vs others (SNP)	195	31	4	2

Amongst the 32 genes identified in our study, 20 were annotated with functions known to be associated with virulence of *H. pylori*, and 12 had unknown functions or functions not known to be linked to virulence (Figure 5.8).

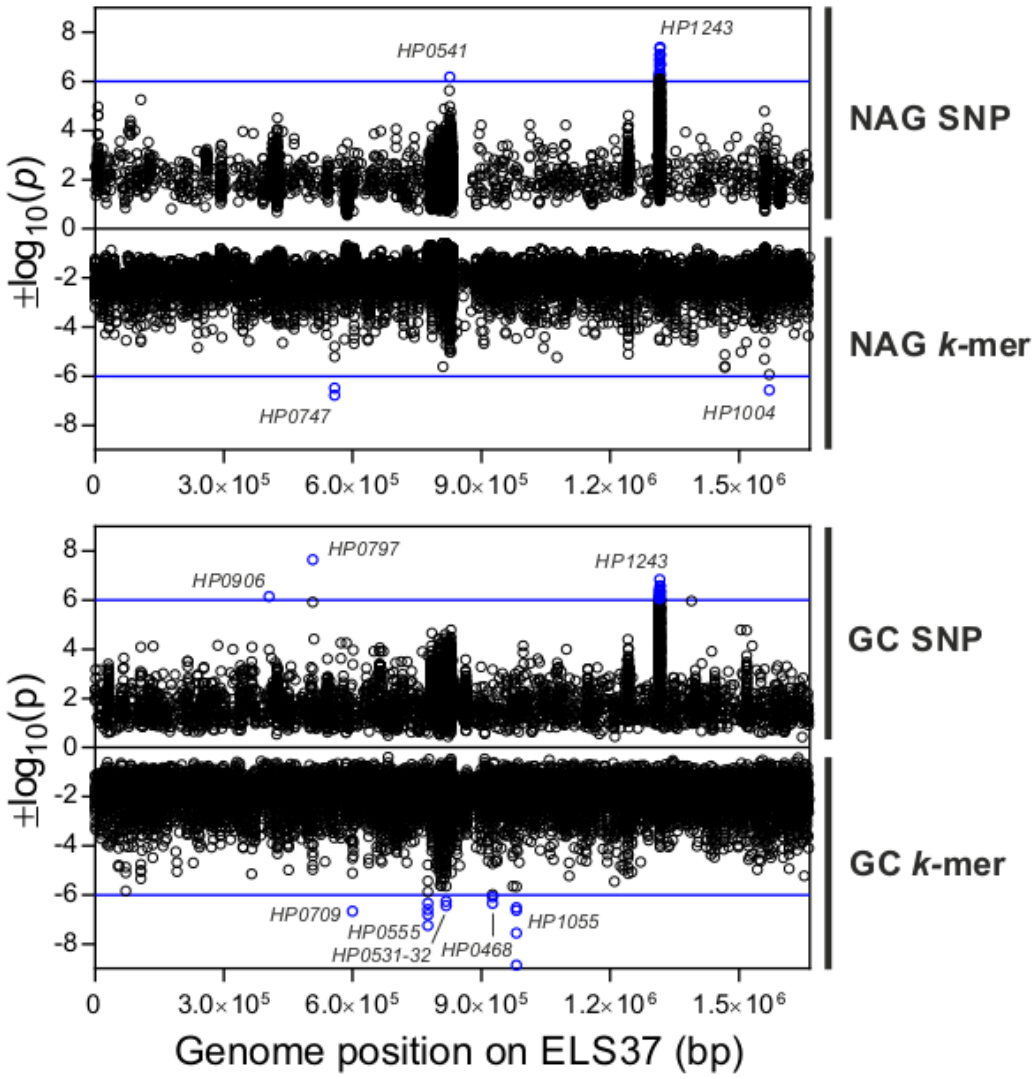


Figure 5.7: Location of genetic elements associated with gastric cancer on ELS37 genome highlighted in 4 bugwas based GWAS performed on 173 strains from hpEurope derived sub-populations according to patient pathology.

GWAS comparing isolates from patients with (**A**) non-atrophic gastritis to those with gastric cancer and precancerous progression and (**B**) gastric cancer to those with non-atrophic gastritis and precancerous progression. Two GWAS tests were performed with bugwas software for each panel, one based on SNPs (upper panels) and the other based on k-mers (lower panels). Positions of the genomic elements are represented on the horizontal axis. Log 10 of the p-value for each hit is recorded on the vertical axis. The blue line indicates a p-value $\leq 10^{-6}$.

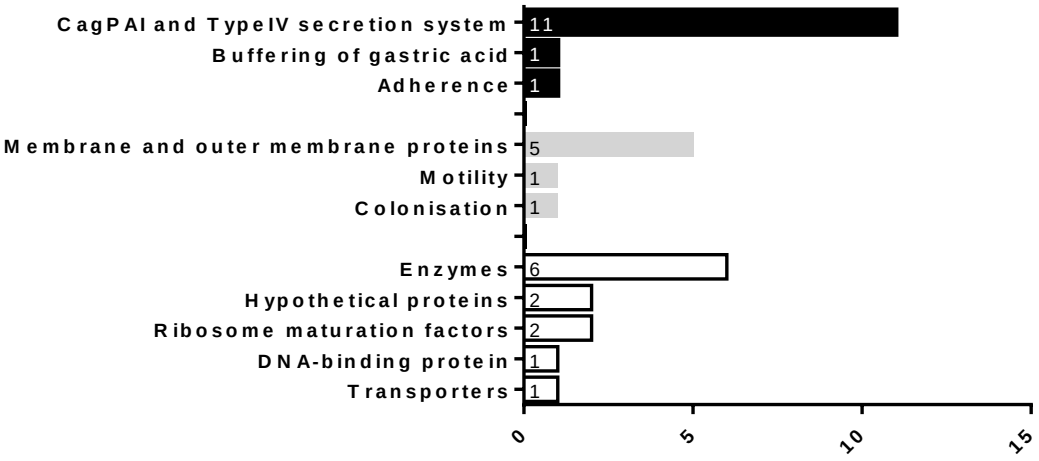


Figure 5.8: Assignment of functions to genes recording hits with a p-value ≤ 10^{-5} in at least one of the 4 GWAS performed on 173 strains from hpEurope derived sub-populations according to patient pathology.
Bars shaded in black represent functions with a direct effect on virulence. Bars shaded in grey represent functions with an indirect effect on virulence. White bars represent genes with functions not known to be associated with virulence in *H. pylori*.

Amongst the 32 genes presenting hits with a p-value ≤ 10^{-5}, 6 genes recorded hits in two of the four GWAS tests: *HP0102*, *HP0468*, *HP0531* (*cag11*), *HP0532* (*cag12*), *HP0541* (*cag20*), *HP0544* (*cagE*/*cag23*), *HP1177* (*hopQ*), and *HP1243* (*babA*). Distribution of gene frequency for these 32 genes showed that 22 genes belonged to the core genome (>90%) and 10 were accessory genes (≤ 90%) (Figure 5.9). Two major types of genomic variations were identified in the hits highlighted by GWAS: accessory variations and allelic variations.

Figure 5.9: Prevalence of the genes recording hits with a p-value ≤ 10^{-5} in at least one of the 4 GWAS performed on 173 strains from hpEurope derived sub-populations according to patient pathology.

5.2.2.1 Accessory variations in top gene hits

An accessory variation is a difference in the presence of a particular gene between the GWAS groups, with prevalence of the gene increasing with the probability of the isolate being obtained from a patient with gastric cancer. Prevalence of all the accessory genes except one (*HP0555*) was not equally distributed amongst our GWAS groups, showing that presence of all remaining 9 genes was higher in the GC group than in the NAG group, with an intermediate prevalence in the Prog group (Figure 5.10).

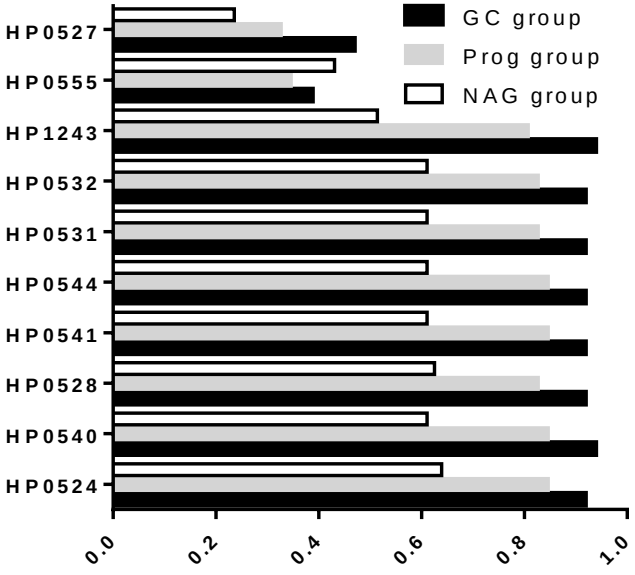


Figure 5.10: Distribution of the gene prevalence for the genes recording hits in at least one of the 4 bugwas based GWAS performed on 173 strains from hpEurope derived sub-populations according to patient pathology.

The 9 genes with accessory variations were all part of the cagPAI island or coded for a gene associated with Cag function (*HP1243*/*babA*). Four of these genes (*HP0531*, *HP0532*, *HP0541* and *HP1243*) had a p-value $\leq 10^{-6}$. *HP1243* was also the only gene with a p-value $\leq 10^{-6}$ in both of our SNP-based GWAS tests (GC vs rest and NAG vs rest).

5.2.2.2 Allelic variations in top gene hits

There were 22 gene hits within the core genome with a p-value $\leq 10^{-5}$. In addition, *HP0555*, which was part of the accessory genome, also had allelic variations: presence of the gene did not vary, but instead two or more versions of the gene were

present in different proportions in each of the GWAS groups, having an effect on the population biology.

The 12 genes with a p-value $\leq 10^{-6}$ were further investigated, individually, using the genomic alignments obtained by GWAS. Two of those genes (*HP0906* and *HP1004*) did not show clear alignments either on data extracted from the GWAS results and on gene-by-gene alignments produced using BIGSdb. These alignments issues might be responsible for the hits by provoking false positive results. Therefore, they were excluded from all further analyses. Analysis of the remaining 10 genes revealed a total of 12 genomic variations (Table 5.5), with 4 accessory variations (described above in section 5.2.2.1) and 8 allelic variations described below. When more than one SNP was identified in a unique gene, they were paired. Out of the 8 sequence variations, 6 were non-synonymous, or had non-synonymous effects when associated with another SNP found in the same codon. The 2 synonymous variations found were in genes where a non-synonymous variation was also identified.

The ratio of non-synonymous SNPs to the total SNPs in the GWAS hits was significantly (p-value of 0.03 with t-student test) higher than the ratio in MLST genes (Figure 5.11).

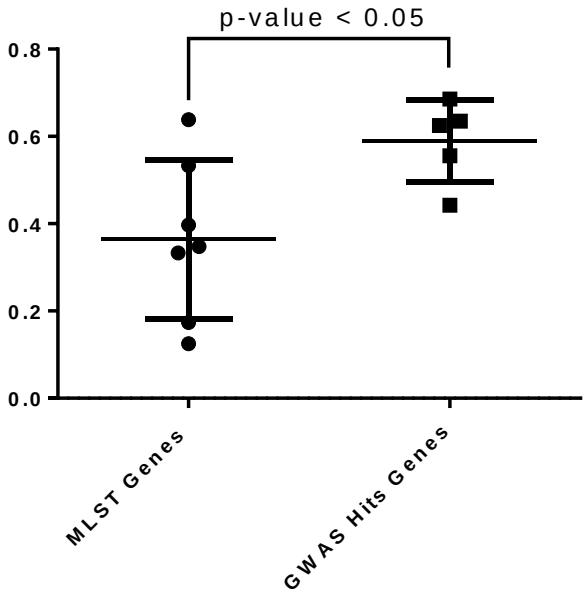


Figure 5.11: Ratio of non-synonymous SNPs to the total SNPs in genes recording hits with a p-value ≤ 10-6 in at least one of the 4 bugwas based GWAS and in MLST genes.
Calculation was based on SNPs occurring in at least 15% of the strains.

Table 5.5: Cancer risk genotypes identified in 4 bugwas based GWAS performed on 173 strains populations according to patient pathology.

Risk and safe genotypes are over-represented among isolates from patients presenting gastric ca respectively, with p-value corresponding to the minimum in each gene (p-value $\leq 10^{-6}$).

Gene name[1]	**p-value (min)**	**Risk genotype**	**Position**[2]	**Safe genotype**	**Frequency**[3]	**Effect on Amino-sequence**[4]
HP1055 [981621-982565] (-)	$1.4.10^{-9}$	A	798	C	0.469/0.125	S, associated with G t substitution at position NS with T in safe, A in ris
HP0797 [506543-507325] (+)	$2.24.10^{-8}$	C + T	325 and 334	T + G	0.592/0.181	NS: L/S in safe , F/A in ris
HP1243,babA1 [1314192-1316405] (-)	$3.99.10^{-8}$	presence	all gene	absence	0.94/0.51	
HP0747 [317158-317757] (+)	$1.69.10^{-7}$	GGAA	934 to 937	AAAA/ GGAG	0.531/0.264	NS: KA in safe, GT in risk
HP0709 [598549-599451] (-)	$2.13.10^{-7}$	A	145	G	0.327/0.153	NS: D in safe, N in risk
		A	159	G	0.959/0.792	S
* *HP0532,cag12* [817677-818519] (+)	$3.62.10^{-7}$	presence	all gene	absence	0.92/0.61	
HP0468 [925539 - 927026] (+)	$4.59.10^{-7}$	CGCC	705 to 708	CACG/ TGCG	0.694/0.514	NS: T in safe, A in risk
		A	729	G	0.796/0.5	S
* *HP0531,cag11* [816985-817641] (+)	$5.4.10^{-7}$	presence	all gene	absence	0.92/0.61	
* *HP0541,cag20* [825334-826446] (-)	$6.6.10^{-7}$	presence	all gene	absence	0.92/0.61	

[1]Position in ELS37 genome [ ], + and – strand is denoted ( ).
[2]Position in gene.
[3]Frequency GC strains/ NAG strains.
[4]The effect on the amino acid sequence is indicated as synonymous (S) and non-synonymous (NS).
*Correlated genes.

Three hits were found in the k-mer GWAS GC vs rest with p-value ≤ 10^{-6} in *HP0468*. Two variations in the genome were identified in this sequence, one of which caused an amino-acid change (Figure 5.12).

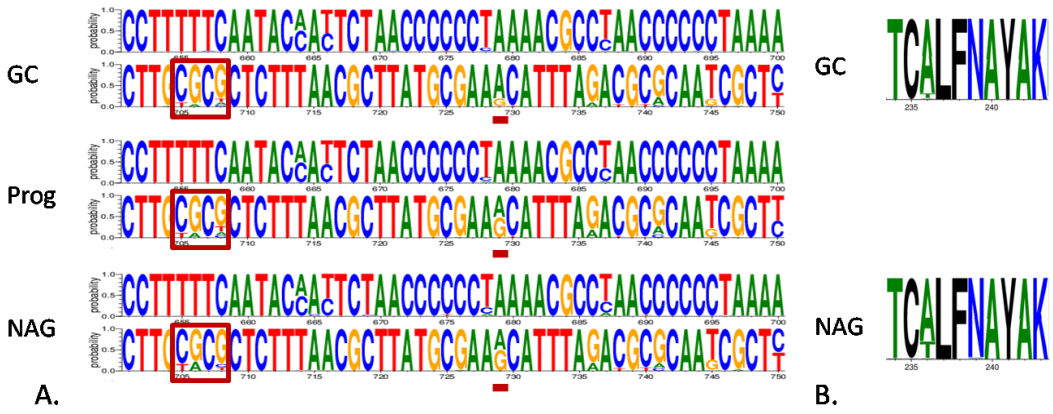


Figure 5.12: Allelic variations observed for the two hits found in *HP0468* and effects on amino-acid sequence.
A. allelic variations identified based on the k-mer GWAS GC vs rest results. Non-synonymous changes are boxed in red, and synonymous changes are boxed underlined in red. **B**. variations in the amino-acid sequence correspondent to non-synonymous change in the nucleic acid sequence.

Based on this analysis, the presence of the sequence CGCC in positions 705 to 708 (three last nucleotides coding for an alanine) was considered a marker for strains at risk of causing gastric cancer (GC), whereas sequences CACG or TGCG (three last nucleotides coding for a threonine) was considered a marker for strains which are not likely to cause GC. Similarly, the presence of an A in position 729 was associated with GC-causing strains and a G at the same position was a marker for strains less likely to cause GC.

HP0555 was not part of the core genome and did not show a clear presence/absence pattern amongst the GWAS groups like the other accessory genes, despite its functional association with the CagPAI. Ten hits were found in the k-mer GWAS GC vs rest with p-value ≤ 10^{-6} in *HP0555*, and all the associated allele variations were non-synonymous (Figure 5.13). Like the prevalence pattern for this gene, there was no clear selection for one of the alleles in associated cancer groups. Instead, associations were observed in the strains belonging to the Prog group, and the two other groups showed a similar distribution of alleles. Therefore, a specific allele marker for the risk or the safety of the strains could not be identified.

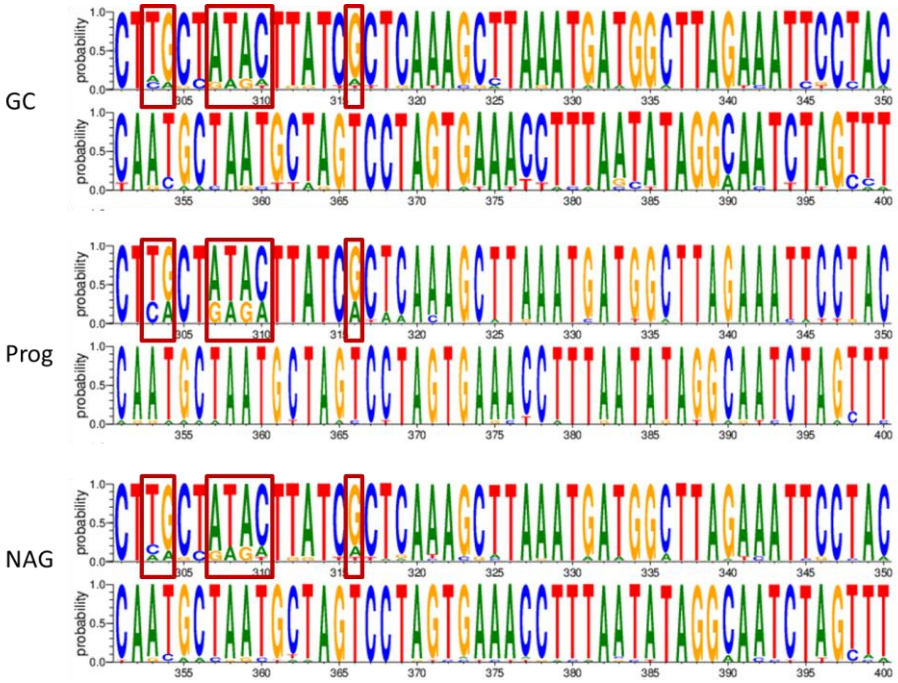


Figure 5.13: Allelic variations observed for the hits found in *HP0555* and effects on amino-acid sequence.
Allelic variations were identified based on the k-mer GWAS GC vs rest results. Non-synonymous changes are boxed in red.

Four hits were found in the k-mer GWAS GC vs rest with p-value $\leq 10^{-6}$ in *HP0709*. Two variations in the genome were identified in this sequence, one of which caused an amino-acid change (Figure 5.14).

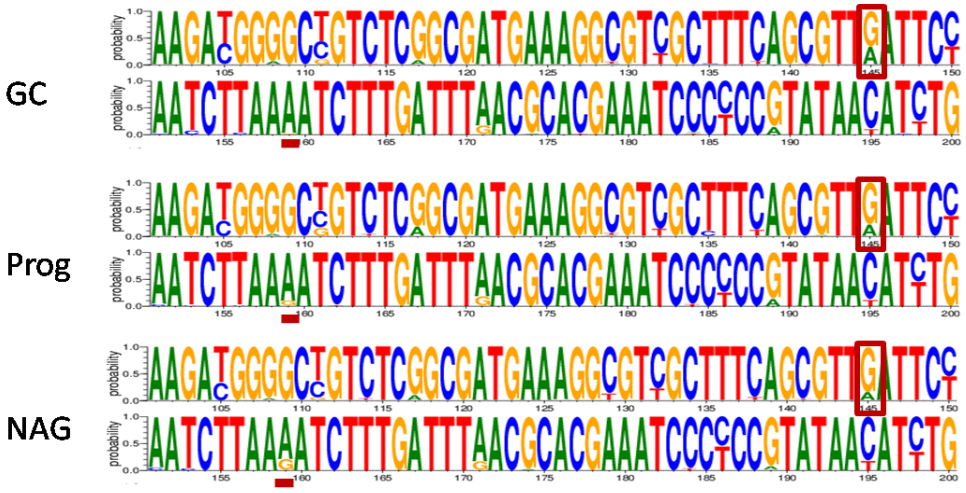


Figure 5.14: Allelic variations observed for the two hits found in *HP0709* and effects on amino-acid sequence.
Allelic variations are identified based on the k-mer GWAS GC vs rest results. Non-synonymous changes are boxed in red, and synonymous changes are underlined in red.

Based on this analysis, the presence of the base pair A in positions 145 (coding for asparagine) was considered a risk marker for strains causing GC, whereas a G (coding for aspartic acid) was considered a marker for strains which were not at risk of causing GC. Similarly, the presence of an A in position 159 was associated with GC-causing strains and a G at the same position was a marker for strains not at risk of causing GC.

Twelve hits were found in the k-mer GWAS NAG vs rest with p-value $\leq 10^{-6}$ in *HP0747*. Only 1 variation in the genome was identified in this sequence and it caused an amino-acid change (Figure 5.15).

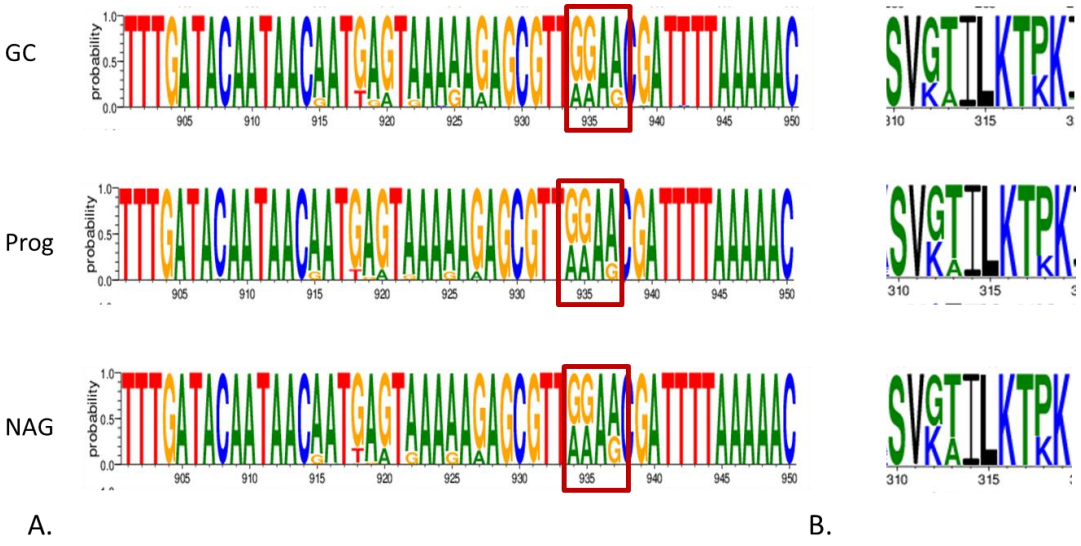


Figure 5.15: Allelic variations observed for the hits found in *HP0747* and effects on amino-acid sequence.
A. Allelic variations identified based on the k-mer GWAS NAG vs rest results. Non-synonymous changes are boxed in red. **B.** variations in the amino-acid sequence correspondent to non-synonymous change in the nucleic acid sequence.

Based on this analysis, the presence of the sequence GGAA in positions 934 to 937 (coding for a glycine followed by a threonine) was considered a marker for strains at risk of causing GC, whereas sequences AAAA or GGAG (coding for a lysine followed by a threonine or a glycine followed by an alanine respectively) were considered a marker for strains which are unlikely to cause GC.

Only one hit was found in the SNP GWAS GC vs rest with p-value ≤ 10^{-6} in *HP0797*, which identified a variation in the genome provoking an amino-acid change (Figure 5.16).

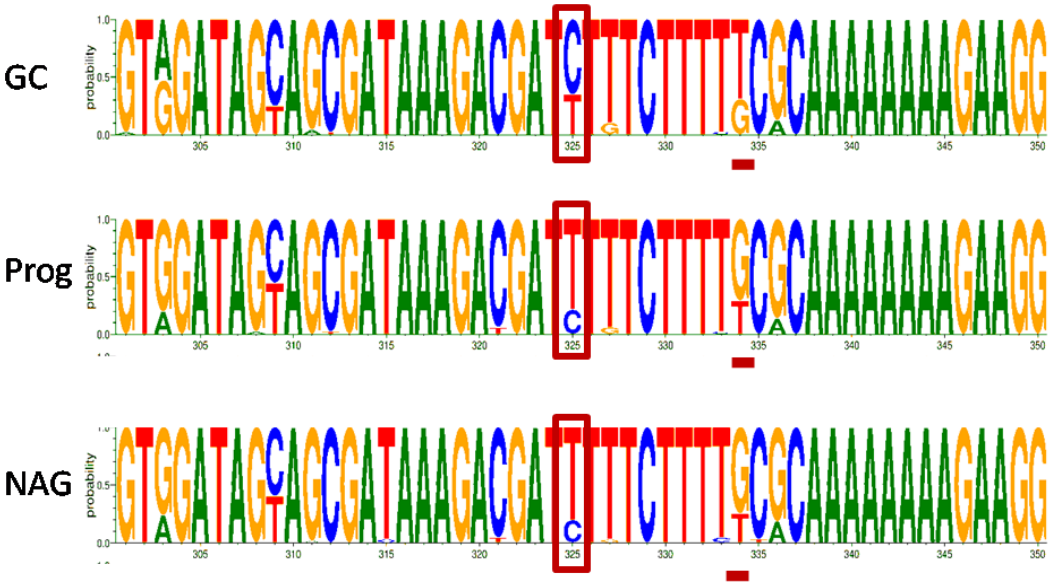


Figure 5.16: Allelic variations observed for the two hits found in *HP0797* and effects on amino-acid sequence.
Allelic variations are identified based on the SNP GWAS GC vs rest results. Non-synonymous changes with a p-value < 10^{-6} are boxed in red. Non-synonymous changes with a p-value >10^{-6} are underlined in red.

Another SNP was visible on the same sequence and was found to pair with one found in position 325. Although it was not originally selected as its p-value was over 10^{-6}, it remains interesting as it is also a non-synonymous SNP. This second SNP was observed in position 334. Based on this analysis, the presence of a C in position 325 and a T in position 334 (coding for a phenylalanine and an alanine) were considered markers for strains at risk of causing GC, whereas a T and a G (coding for a leucine and a serine) were considered a marker for strains which are unlikely to cause GC.

Seventeen hits were found in the k-mer GWAS GC vs rest with p-value ≤ 10^{-6} in *HP1055*, which identified a variation in the genome. This hit was the one with the strongest p-value ($1.4x10^{-9}$), but it was synonymous (Figure 5.17). However, when the SNP was identified in gastric cancer alongside another, more rarely found SNP, positioned 2bp before the first one, it had non-synonymous consequences. This was only observed on a small number of strains.

Based on this analysis, the presence of an A in position 798 (coding for an alanine) was considered a marker for strains at risk of causing GC, whereas a C (coding for a threonine when associated with a G in position 796) was considered a marker for strains which were unlikely to cause GC.

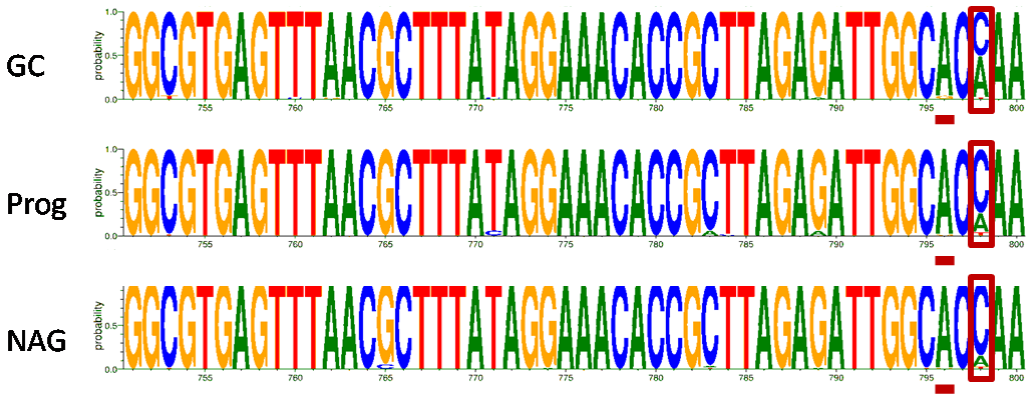


Figure 5.17: Allelic variations observed for the two hits found in *HP1055* and effects on amino-acid sequence.
Allelic variations were identified based on the SNP GWAS GC vs rest results. Synonymous changes with a p-value < 10^{-6} are boxed in red. SNPs causing the synonymous change to become non-synonymous are underlined.

5.2.2.3 Risk score

Based on analysis of accessory and allelic variations in the hits with p-value ≤ 10^{-6}, a list of 11 risk genotypes was assembled (Table 5.5). The 3 cagPAI genes (*HP0531*, *HP0532* and *HP0541*) were correlated, based on Pearson's correlation analysis. No other genes showed correlation. These 11 risk genotypes were used to build a risk score (5.1.6). Distribution of risk scores in our dataset was significantly (ANOVA p-value <0.0001 between the gastric cancer group and each of the other groups) associated with the pathology (Figure 5.18).

Strains isolated from patients with gastric cancer all showed risk scores over -25, whereas 20% of patients showing only non-atrophic gastritis symptoms had a risk score below this limit.

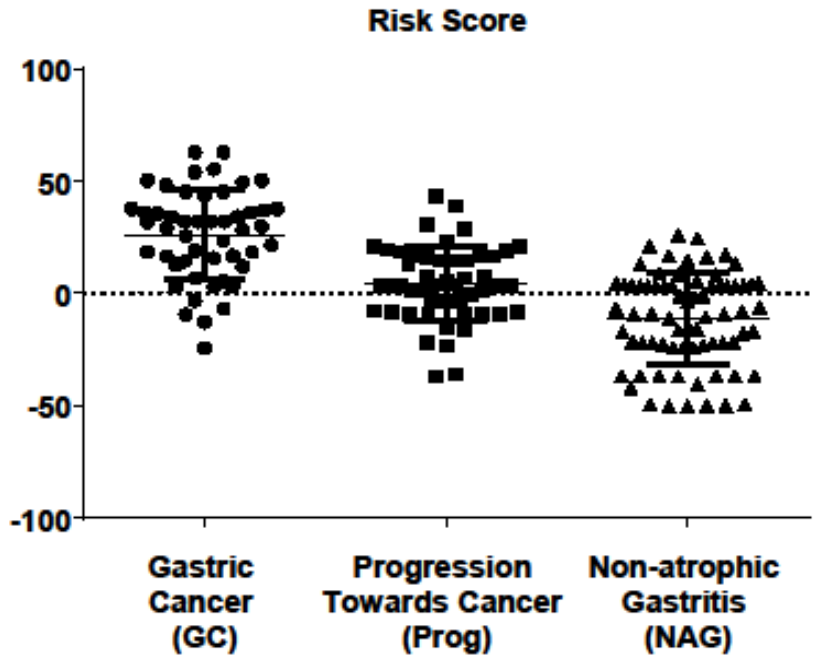


Figure 5.18: Assignment of risk scores to 173 strains from hpEurope derived sub-populations according to patient pathology.
Each dot corresponds to the risk score associated with a single strain. This risk score was calculated based on the presence of risk or safe genotypes for each of the 9 genes considered (Table 5.5).

5.3 Discussion

The GWAS method, applied to *H. pylori*, can identify genes associated with complex phenotypes such as gastric cancer. The choice of the GWAS model must be considered with care. Two methods were used in this Chapter, based on two different softwares, ClonalFrame and bugwas. Gene functions identified in this chapter were in large part linked to host-bacteria interactions. The major function identified, already highly associated to gastric cancer, was the injection of CagA through the type IV secretion system. Identification of cagPAI genes with both methods supports the validity of the GWAS. Functions linked to colonisation, such as motility, buffering of gastric acid and adherence were also identified, alongside *vacA* and *vacA* paralogs, were also highlighted by our results. *HP0068*, also known as *ureG*, was a hit in both methods, highlighting the important role of urease in pathogenicity. Two genes coding for outer membrane proteins were identified in both GWAS: *HP1177* (*hopQ*) and *HP1243* (*babA*). Membrane proteins are associated with functions linked to

host/bacteria or environment/bacteria interactions, and therefore are likely to be linked with pathogenicity. A hypothetical protein, *HP0468*, was also found in both methods. This is the only gene highlighted by both GWAS methods used with unknown function. Despite the lack of knowledge about the function of this gene, it is likely to have a role in the development of gastric cancer, and therefore should be investigated further.

ClonalFrame based methods, although well-adapted for other bacteria species such as *Campylobacter*, was not optimal for highly recombinogenic bacteria such as *H. pylori*. The ClonalFrame based method used in our chapter showed a high number of hits, forming a confusing background signal. This is likely to be due to the versatile nature of the *H. pylori* genome and its ability to recombine. The structure in *H. pylori* populations was probably interfering with the results and showing false positive hits. More precise data on pathology were also made available to us after completion of this method, which would have changed the assignment of a couple of strains from this dataset. Despite these limitations, 71 genes with high association scores were highlighted, showing both accessory and allelic variations. Most accessory variations were CagPAI related, with genes more present in gastric cancer strains than in non cancer strains. This group of genes have been previously described as being present in a large majority of cancer strains (Parsonnet et al. 1997; López-Vidal et al. 2008). Considering the high variability of *H. pylori* genome, a pattern of presence or absence of a gene is reasonably easier to identify, even without the high-speed genomics techniques that are now at our disposal. Therefore it is not surprising that the research on this group of genes is more advanced than for genes presenting allelic variations. However, one of the genes showing accessory variations (*HP1116*) was more present in non cancer strains compared to cancer strains. This suggests that this gene of unknown function could have a protective effect against development of cancerous pathologies due to *H. pylori* infection.

Bugwas-based GWAS produced more encouraging results. The environmental and host factors did not mask the signal in the GWAS, and genes related to virulence, as well as genes which were not previously associated with virulence, were identified with associations to healthy or diseased phenotypes. Again, both accessory and allelic variations were identified. When two (or more) allelic variations were identified in a

unique gene, changes were almost systematically found in the same strains, indicating that either the strains derived from two distinct ancestors, one with the changes and one without, or that selection pressure selected for those specific traits. Once again, genes identified comprised genes belonging to the CagPAI pathway. On top of these, genes coding for membrane proteins were identified (*HP0555*, *HP1055*). *HP1055* was shown to be essential in transposon mutagenesis experiments (N. R. Salama, Shepherd, and Falkow 2004), but its exact function remains unknown. *HP0797*, also known as *hpaA*, was originally described as a sialic acid binding protein involved in adhesion (D. G. Evans et al. 1993). However it is now thought to be a lipoprotein (A. C. Jones et al. 1997; O'Toole et al. 1995). A recent study also highlighted this gene as being essential for *in vivo* colonisation of the mouse stomach (Carlsohn et al. 2006). The product of this gene, HpaA, shows strong immunogenic properties (P. Sutton et al. 2007). It has been considered as a target for vaccines development (Tobias et al. 2017; R. Zhang et al. 2016). Two enzyme-coding genes were also highlighted by the bugwas GWAS. *HP0709* codes for an enzyme, S-adenosyl-I-methionine hydroxide adenosyltransferase. Conflicting annotations were found regarding this gene and its product. This enzyme could be involved in either methylation of DNA and proteins, or in the synthesis of the branched amino acids valine, leucine and isoleucine (Deng and O'Hagan 2008). *HP0747*, also known as *trmB*, codes for a predicted S-adenosylmethionine-dependent methyltransferase regulated by *HP1021* (Pflock et al. 2007). It is possibly involved in the regulation of acetone metabolism. It was also identified in a previous study as a gene with a large number of radical substitutions in fast-evolving regions (Zheng, Roberts, and Kasif 2004). Finally, *HP0468* codes for a hypothetical protein, poorly conserved outside the *Helicobacter* genus. A study on chemolithoautotrophically enhanced growth of *H. pylori* identified this gene as being upregulated by molecular hydrogen (Kuhns et al. 2016), but no exact function was described.

It is important to note that the GWAS presented in this chapter have limitations. Some are due to the nature of the bacteria, and can be addressed by a careful choice of model. A new phylogenetic approach to GWAS named treeWAS (Collins and Didelot 2018) has been developed recently, which presents good specificity and power, and could limit the potential false-positive hits observed in the two methods used in this thesis.

Another limitation of this study was the collection of strains used. Due to the need for a high number of strains and the use of collections from different origins, it was not possible to control for possible confounding effects linked to other gastric cancer risk factors (alcohol consumption, smoking, age of the patient, genetic polymorphisms in patient genes, ratio male/female…), because information for each factor was not available for all strains used in our study. Our results can be impacted by this, and future studies should take these effects into account. Moreover, both cancer and non-cancer groups had some internal variability across pathologies included (Appendix E). Strains came from diverse collections across the world, carried out by different clinic or research groups using different staging systems to record lesions. Another potential bias is the fact that only one strain was included in our dataset for each patient. Stomachs can be colonised by multiple strains (Kibria et al. 2015; J. W. Kim et al. 2004; Ben Mansour et al. 2016). Some of the strains could have been linked to gastric cancer, when they were actually not the strains driving the carcinogenesis. This would also be a limitation for clinical application. Sequencing one strain from a single colony would not be sufficient to make sure that the strain is safe and that the patient should not be treated. More than one strain would have to be sequenced.

Moreover, as gastric cancer occurs after a long-term infection, some of the non-cancer strains could have subsequently evolved into cancer strains if left longer in their host, and causative strains could have disappeared at the time of isolation, or evolved into the strains isolated. This variability can limit the power of the GWAS method. Moreover, the genes identified as associated with gastric cancer can, indeed drive the disease, but can also be the result of the changes occurring in the environment, such as disruption of the epithelium and reduction of acidity. The genome of *H. pylori* strains could have evolved alongside those changes in order to adapt and survive in this changed environment. GWAS studies on complex phenomenon such as gastric cancer, based on strains from healthy or diseased patients, cannot discriminate between genomic traits that caused the cancer and genomic traits that were provoked by the environmental changes in cancerous regions.

To address this issue, a prospective study would have been preferable, by collecting strains from healthy patients and following the evolution of their symptoms. This was not possible here but with suitable ethical and clinical support would be a study for

the future. Such a study on other mammalian models might be a solution. For instance, long-term infection of mice or primates with strains isolated from healthy patients and analysis of the consequences on the stomach to assess the *in vivo* virulence of the strain and perform a GWAS. However, not all human strains are easy to use for assays in different hosts. Furthermore, another source of variability due to interactions of the bacteria with this new host would be introduced (see Chapter 4).

Despite these limitations, GWAS study in *H. pylori* represents an important advance in the way *H. pylori* is studied in relation to gastric disease. A large diversity is a necessity in nature, and the definition of health is not the absence of disease, but "a state of complete physical, mental and social well-being". By systematically clearing *H. pylori* from patients, even in the cases where it does not have negative impact on its host, this diversity is reduced, and it frees the niche to other pathogens or disorders to develop. Moreover, research uncovered positive impacts of *H. pylori* colonisation. A robust GWAS, and a risk score such as the one built in this chapter, could be a foundation for a sequencing-based detection method used in clinics to target high-cancer risk strains and limit the pressure towards global increase of antibiotic resistance.

In conclusion, we successfully applied the GWAS method to *H. pylori* cancer and non-cancer strains, by comparing two different methods and using a large dataset with a control on the population structure. This study identified genomic traits in 9 genes that correlate with gastric cancer. These traits were used to build a risk score that could be the first stone towards a new treatment strategy targeting only the strains at risk while leaving the others to keep the microbiota balance in place and reduce the rise of antibiotic resistance.

6 Phenotypic characteristics of *Helicobacter pylori* European strains

Helicobacter pylori is a Gram-negative bacteria that inhabits a unique and harsh environment, namely the human stomach (B. Marshall and Warren 1984). This bacterium is well adapted to such a highly acidic environment, through expression of factors such as enhanced motility and secretion of urease. The presence of *H. pylori* induces inflammation in the host stomach, which may progress to more serious health complications (e.g. ulcer, mucosa-associated lymphoid tissue (MALT) lymphoma or gastric cancer). Indeed, *H. pylori* is associated with ulcer formation (Chamberlain and Peura 1990; Oderda et al. 1990), and increases the risk of gastric cancer from 2 to 6 times compared to those without *H. pylori* infection (Parsonnet et al. 1997; Wroblewski and Peek 2016; Ferlay et al. 2015). The exact mechanisms linking presence of *H. pylori* in the stomach with gastric cancer are still partially unknown, and a more precise view of what links *H. pylori* colonisation and carcinogenesis is essential to be able to treat and prevent cancer development. The genomic variability in *H. pylori* is one of the highest amongst bacteria, due to the ability of this micro-organism to recombine (Go et al. 1996). Phenotypic variability has also been observed between some specific strains of *H. pylori*.

Motility is essential for *H. pylori* pathogenesis in the human host (Eaton et al. 1996; O′Toole, Lane, and Porwollik 2000). Most of the genes involved in motility are core genes, present in all strains of *H. pylori*. However, allele variations in these genes are common. In addition, a number of genes associated with motility were highlighted in the genomic analyses carried out in previous chapters:

- Highly statistically significant gene hits in Genome Wide Association Studies (GWAS) comparing gastric cancer strains with non-cancer strains (Chapter 5) (*HP0295*, *HP0685*, *HP1031* and *HP1119*, respectively known as *flgL*, *fliP*, *fliM* and *flgK*)
- Identification of a gene showing Phase Variation (PV) in a MALT Lymphoma strain re-isolated from mice (Chapter 4) (*HP0685*, also known as *fliP*)
- Identification of a gene showing a SNP during long-term colonisation in mice (Chapter 4) (*HP1041*, also known as *flhA*)

When infecting a host, *H. pylori* causes inflammation through the mobilisation of cells of innate and adaptive immunity (D. J. Evans et al. 1995; Satin et al. 2000;

Unemo et al. 2005). Important differences exist between those inflammatory pathways. Some research suggests that the type of strain is one of the key variables. There is already some evidence of a difference in IL-8 production in response to strains possessing the Cag Pathogenicity Island (CagPAI positive) and strains without this island (CagPAI negative) (Fischer et al. 2001). NF-κB activation is one of the pathways linking CagPAI with IL8 production (Brandt et al. 2005). The role of TLR-2 and TLR-5 has also been studied and linked to both cagPAI dependant and independent signalling (Kumar Pachathundikandi et al. 2011). A complete CagPAI island seems to be needed to induce IL-8 production in AGS cells (Nilsson et al. 2003). The sequence variability observed in cagPAI genes, segregating between East Asian-type and Western-type, is also linked to the production of IL-8, with the East Asian-type being more virulent (Yuan et al. 2017). In brief, there is a wide variety of pathways and actors involved in the CagPAI dependant induction of IL-8. Identifying the bacterial genes linked to this IL-8 induction in different types of cells might help bringing light on some of these pathways. A large number of genes from the CagPAI island, alongside genes with functions associated with CagA secretion were identified in GWAS comparing gastric cancer strains with non-cancer strain in Chapter 5. This chapter will focus on two functions of *H. pylori* which were highlighted in the previous chapters: i) motility and ii) induction of an immune response in epithelial and inflammatory cells. Although motility of *H. pylori* and inflammation following infection has been studied, as described in the previous paragraphs, questions remain about the genomic basis explaining the variability of phenotypes between different strains.

Motility will be studied in 56 strains of *H. pylori*, all isolated in Europe from defined patients groups (including gastritis, gastric cancer, MALT lymphoma and ulcer). Differences in motility linked to the pathology will be investigated. The genomes of these strains will also be studied using whole-genome-based methods to identify a genetic basis that could explain phenotypic differences.

Differences in inflammation triggered by *H. pylori* strains will also be studied in a subset of 15 strains. Two types of cells (epithelial and macrophages) will be infected with strains of *H. pylori* to confirm the importance of CagPAI in generating early inflammatory responses, and compare CagPAI positive strains isolated from patients with different diseases. Again, genomic analysis will be performed to highlight genes that could be involved in the possibility of differential immune responses.

Three specific aims will be addressed in this chapter:

- Motility varies according to the pathology of the patient from which the strain was isolated,
- Immune response is triggered differently according to the pathology of the patient from which the strains was isolated,
- Some genes covary with phenotypic differences observed among strains.

6.1 Materials and Methods

6.1.1 Dataset

Thanks to a global effort in collecting samples during endoscopies and isolating clinical strains of *H. pylori*, there is a large collection of strains available to be studied in the laboratory. Fifty-seven strains of *Helicobacter pylori* isolated from European patients were used in this study (kind gifts from Dr Alain Burette, Prof Francis Megraud and Dr Sinead Smith). These samples come from a wide range of patient clinical outcomes (Appendix C):

- 18 strains were isolated from patients with normal mucosa or asymptomatic gastritis with no intestinal metaplasia,
- 18 strains were isolated from patients suffering from gastric cancer or gastrointestinal stromal tumor,
- 13 strains were isolated from patients suffering from MALT Lymphoma,
- 8 strains were isolated from patients suffering from ulcer.

6.1.2 DNA extraction and sequencing

Whole genome sequences for 41 of these strains were already available. The average genome size for these strains was 1614846.4 bp and GC content was 39.03%, inside the normal range for this species (1.3.3.1). Three sequences were available but failed quality control, so they were re-sequenced along with the 13 newly sequenced strains. DNA was extracted and sequenced as described in 2.1.6. The average genome size for these newly sequenced strains was 1623319.1 bp and GC content was 38.95%. All available and newly sequenced strains were entered into BIGSdb to allow genomic analyses.

6.1.3 Enumeration of *H. pylori*

A subset of 10 *H. pylori* strains (from the 57) were chosen based on the patient groups above and their ability to grow reproducibly. These were used in an enumeration assay (Table 6.1). Five isolates were cancer strains and 5 were gastritis or control strains. Those strains were grown on CBA plates in micro-aerophilic conditions as described in 2.2.1. Once sufficient growth was observed on plates, the colonies were resuspended into BB media and cultured for 20 to 24 hours at 37°C in an incubator as described in 2.2.2. Enumeration was then performed following the protocol described in 2.2.3, using 3 different dilutions spread in triplicate.

Table 6.1: Summary of the characteristics of the 10 *H. pylori* strains used for enumaration.
CagPAI status is defined as positive when more than 90% of the CagPAI genes were present

Strain Name	Pathology	Cag PAI status	Geographic provenance
3697	gastritis	Negative	France
3802	gastritis	Negative	France
29373	control	Negative	Belgium
3699	gastritis	Positive	France
3824	gastritis	Positive	France
GC23	Gastric cancer	Positive	France
GC54	Gastric cancer	Positive	France
GC65	Gastric cancer	Positive	France
30950	Gastric cancer	Positive	Belgium
38185	Gastric cancer	Positive	Belgium

6.1.4 Motility of *H. pylori*

Fifty-six *H. pylori* strains (Table 6.2) were grown on CBA plates as described in 2.2.1.

Once sufficient growth was observed on plates, the colonies were resuspended into BB media and cultured in liquid media for 20 to 24 hours at 37°C in an incubator (2.2.2). A motility assay was then performed as described in section 2.2.5 with normalised cultures aliquoted onto motility agar plates. The motility was measured in triplicate wells at least twice for each strain and the diameter of growth was compared to the positive control (strain B24) in order to minimize experimental bias due to differences in growth (C.-Y. Kao, Sheu, and Wu 2014; C.-Y. Kao et al. 2012).

Table 6.2: List of the 56 *H. pylori* strains used for a motility assay and pathology associated.

Strain Name	Pathology	Strain Name	Pathology
29009	control	**31181**	control
33375	control	**3735**	gastritis
SSR2	mild chronic gastritis	**SSR5**	moderate chronic gastritis
SSR13	moderate chronic gastritis	**3755**	gastritis
3770	gastritis	**3802**	gastritis
3824	gastritis	**3754**	gastritis
3745	gastritis	**3697**	gastritis
3699	gastritis	**30908**	normal
31235	normal	**27935**	gastric cancer
28861	gastric cancer	**34320**	gastric cancer
GC11	gastric cancer	**GC23**	gastric cancer
GC26	gastric cancer	**GC27**	gastric cancer
GC31	gastrointestinal stromal tumor	**GC34**	gastric cancer
GC54	gastric cancer	**GC65**	gastric cancer
GC30	gastric cancer	**GC69**	gastric cancer
21580	gastric cancer	**30950**	gastric cancer
38185	gastric cancer	**19027**	gastric cancer
GC62	gastric cancer	**B23**	MALT lymphoma
B24	MALT lymphoma	**B25**	MALT lymphoma
B26	MALT lymphoma	**B27**	MALT lymphoma
B29	MALT lymphoma	**B30**	MALT lymphoma
B31	MALT lymphoma	**B37**	MALT lymphoma
B40	MALT lymphoma	**B41**	MALT lymphoma
B44	MALT lymphoma	**B47**	MALT lymphoma
3843	Ulcer	**ANT170**	Ulcer
BON254	Ulcer	**CHA185**	Ulcer
GRA185	Ulcer	**PHI092**	Ulcer
3738	Ulcer	**3774**	Ulcer

6.1.5 Infection of AGS and THP-1 cells with *H. pylori*

Fifteen strains of *H. pylori* were used in two independent infection experiments using AGS and THP-1 cells (Figure 6.1).

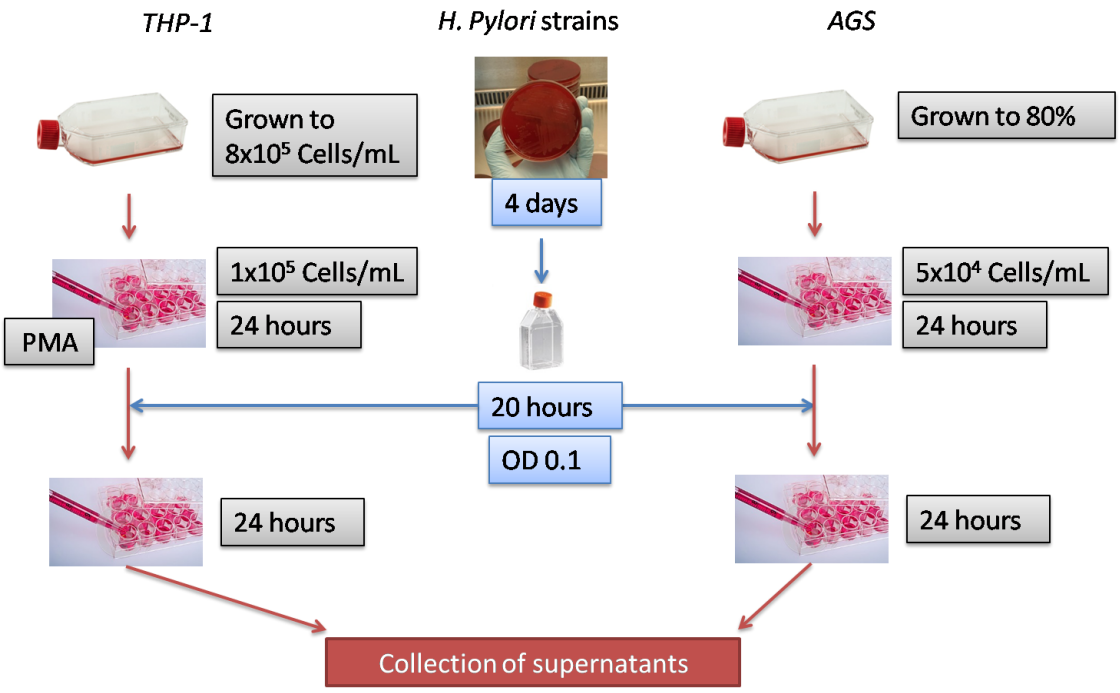


Figure 6.1: Protocol used for infection of AGS or THP-1 cells with *H. pylori* strains.
AGS and THP-1 cells were grown in RPMI media supplemented with L-glutamine and 10% Foetal bovine serum (FBS) as described in 2.1.7 and 2.1.8. In parallel, *H. pylori* strains were grown on blood agar plates until sufficient growth was observed (2.2.1). AGS and THP-1 cells were seeded in 24 well plates at $50x10^3$ and $100x10^3$ cell/mL respectively (2.2.9). THP-1 cells were differentiated into macrophages, and both AGS and differentiated THP-1 cells were incubated for 24 hours at 37°C (2.2.10). In parallel, the *H. pylori* strains were cultured in liquid and diluted to an OD of 0.1 (2.2.2). Cultures were induced with PMA as positive control and DMSO as negative control (2.2.10). Each sample was done in triplicate wells. Supernatants were collected after 24 hours, centrifuged at 4°C at maximum speed for 10 minutes then stored at -20°C until analysed.

Five of the strains used in this experiment were from patients with gastric cancer, and 10 were from patients with gastritis. CagPAI status is defined as positive when more than 90% of the CagPAI genes were present. Amongst the gastritis strains, 5 were cagPAI positive and 5 were cagPAI negative. All the gastric cancer strains were cagPAI positive (Table 6.3). For each strain, infections were repeated in triplicate, during at least two independent experiments.

Table 6.3: Summary of the characteristics of the 15 *H. pylori* strains used for infection of AGS and THP-1 cells.
CagPAI status is defined as positive when more than 90% of the CagPAI genes were present

Strain Name	Pathology	Cag PAI status	Geographic provenance
3697	gastritis	Negative	France
3745	gastritis	Negative	France
3802	gastritis	Negative	France
29373	control	Negative	Belgium
31235	control	Negative	Belgium
3699	gastritis	Positive	France
3824	gastritis	Positive	France
29009	control	Positive	Belgium
30908	control	Positive	Belgium
31181	control	Positive	Belgium
GC23	Gastric cancer	Positive	France
GC54	Gastric cancer	Positive	France
GC65	Gastric cancer	Positive	France
30950	Gastric cancer	Positive	Belgium
38185	Gastric cancer	Positive	Belgium

6.1.5.1 Human Inflammation Antibody Array

Detection of 40 human proteins was carried out on 4 samples obtained from infection experiments. Two samples were obtained from infection of THP-1 cells with two strains and two from infection of AGS cells with the same two strains. The two strains used were 30950 (CagPAI positive gastric cancer strain) and 31235 (CagPAI negative non-cancer strain). The assay was performed using the RayBio® C-Series kit according to the manufacturer instructions and intensity of the spots was analysed using ImageJ as described in 2.2.13. Only differences of more than 0.1 were considered for further confirmatory investigation using ELISA.

6.1.5.2 Interleukin-8 and CCL4 ELISA

Concentrations of interleukin-8 (IL-8) were measured using an ELISA kit (DuoSet) in half-area 96-well plates in all samples from infection of AGS or THP-1 cells following the protocol described in 2.2.11. Concentrations of CCL4 were measured using an ELISA kit (DuoSet) in half-area 96-well plates only in samples from infections of THP-1 cells following protocol described in 2.2.12.

Each infection experiment with a strain of *H. pylori* on AGS or THP-1 was repeated 3 times. To reduce experimental variations between experiments, the average negative control from each experiment was used as a unit value.

Supernatants from THP-1 cells were highly concentrated in cytokines, therefore dilutions were needed for both IL-8 and CCL4 ELISA to obtain concentrations in the kit range. For IL-8 concentration, dilutions in PBS did not show linear variation of concentration with dilution, therefore dilutions were achieved in wash buffer (containing Tween) to disaggregate proteins. Dilutions of 1 in 50 in wash buffer were chosen for IL-8 (Figure 6.2A). Dilutions of 1 in 50 in PBS were chosen for CCL4 (Figure 6.2B). Positive and negative controls were not diluted. Results presented in sections below are concentrations in non-diluted supernatants.

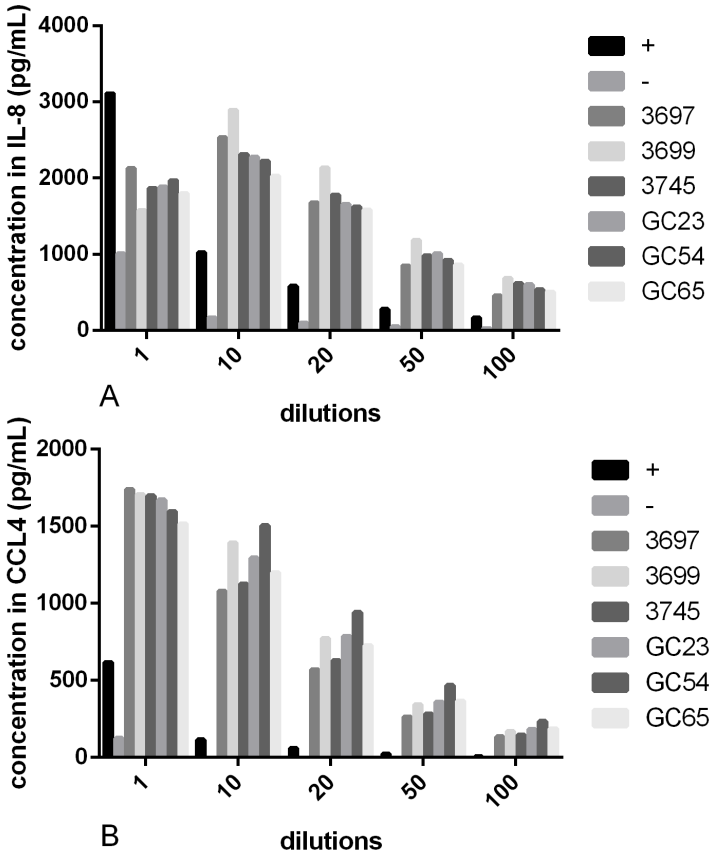


Figure 6.2: Concentration of IL-8 (A) and CCL4 (B) in dilutions from supernatants obtained after 24h infection of differentiated THP-1 cells with *H. pylori* strains

A. Concentration of IL-8 in diluted supernatants. Diluent used was wash buffer. Range of detection was between 31.2 and 2000 pg/mL. **B.** Concentration of CCL4 in diluted supernatants. Diluent used was PBS. Range of detection was between 15.6 and 1000 pg/mL.

6.1.6 Genomic analyses

Two types of genomic analyses were performed based on the phenotype results. For both methods, the dataset was divided into groups according to the results of the studied phenotype. A binary dataset was created separating strains in the upper and lower 33rd percentiles for each phenotype. Both genomic analyses were based on a genome comparator (2.3.2).

6.1.6.1 Identification of genes associated with phenotypes

This first analysis was based on a genome comparator performed with the genes from a pan-genome built using the reference strain 26695 genes and all 57 strains used in phenotypic assays following the method described in 2.3.2. This genome comparator aimed to show whether some of the genes had higher incidence over the range of phenotypes. Non-accessory variations were also investigated briefly by observation of the allele numbers obtained by genome comparator. Functions of the genes were searched for on NCBI and PATRICdb, and interactions of the gene products with other proteins were queried on PATRICdb.

6.1.6.2 Attribution of functions to genes targets

The second analysis was a genome comparator (2.3.2) performed using all genes identified in chapters 3, 4 and 5. This analysis aimed to link the genes highlighted *in silico* to phenotypic characteristics of the strains studied *in vitro*.

6.2 Results

6.2.1 Enumeration of *H. pylori*

An enumeration experiment was performed on 10 strains of *H. pylori*, from different collections and associated with different symptoms (Table 6.1). The number of colony forming units (CFU) per mL of a bacterial solution with OD=0.1 was variable (Figure 6.3A). There was no significant difference in calibrated counts of *H. pylori* from gastritis or gastric cancer patients (Figure 6.3B). The average for these 10 strains was $1.58.10^7$ CFU/mL of OD 0.1 bacterial solution, which is consistent with values found in the literature (Blanchard and Nedrud 2012). This average was used to estimate the concentration of bacteria used in the remaining experiments.

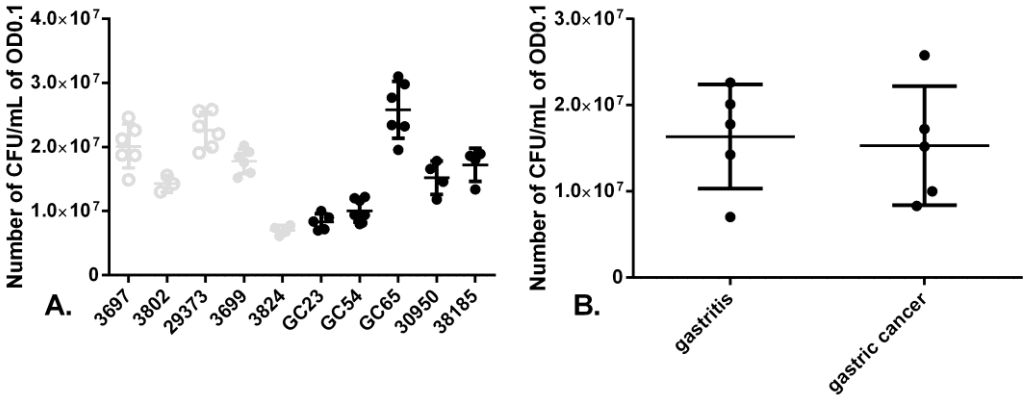


Figure 6.3: Enumeration carried out on 10 clinical *H. pylori* strains.
A. Each dot represents the number of colony forming units per mL of OD 0.1 bacterial solution calculated from one plate. Strains in black are gastric cancer strains. Strains in grey are gastritis strains. Plain circles represent CagPAI positive strains and empty circles CagPAI negative strains. **B.** Each dot represents the average number of colony forming units per mL of OD 0.1 calculated from at least 3 plates for one strain. No significant difference was observed between gastritis and gastric cancer strains (t-student).

6.2.2 Variability of motility in *H. pylori*

The relationship between the motility of 56 strains (Table 6.2) and disease background was investigated (Figure 6.4).

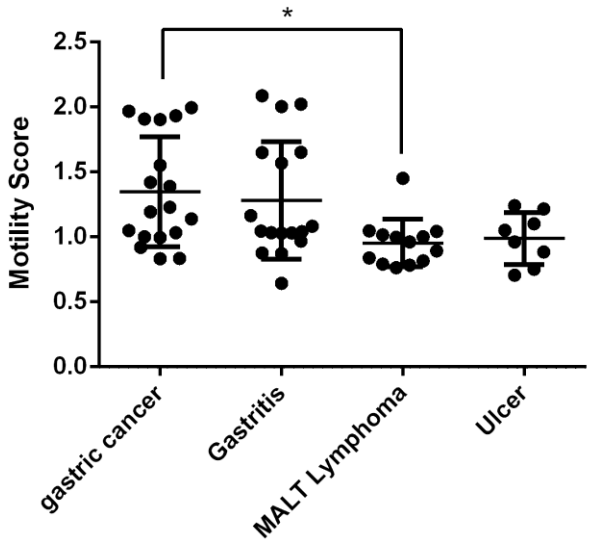


Figure 6.4: Motility measured in *H. pylori* strains from different patient pathology.
Each dot represents the average motility score for one strain, measured at least twice and corrected using the measure for control strain B24. Star represents a p-value < 0.05 between groups (one-way ANOVA).

Differences in motility between disease types were only statistically significant between gastric cancer strains and MALT lymphoma strains (p-value < 0.05 with one-way ANOVA). However, the standard deviations for MALT lymphoma and ulcer strains were much smaller compared to gastric cancer and normal/gastritis strains.

6.2.3 The host-immune response triggered by *H. pylori*

Fifteen strains of *H. pylori* chosen for their disease background and CagPAI status (Table 6.3) were used to infect AGS or THP-1 cells to investigate inflammatory cytokine response to the live bacteria. Supernatants from these co-cultures were used in three experiments focusing on a cytokines screen using a human inflammation antibody array and further confirmatory ELISA assays.

6.2.3.1 Human Inflammation Antibody Array

Four samples of supernatants from infection experiments were used to perform a human inflammation antibody array assay. This assay aimed to identify human cytokines produced by AGS and THP-1 cells when challenged by different strains of *H. pylori*. The strains used were 30950 (a cag-PAI positive cancer strain) and 31235 (a cag-PAI negative non-cancer strain). Optimised exposures of the assay membranes are presented in Figure 6.5.

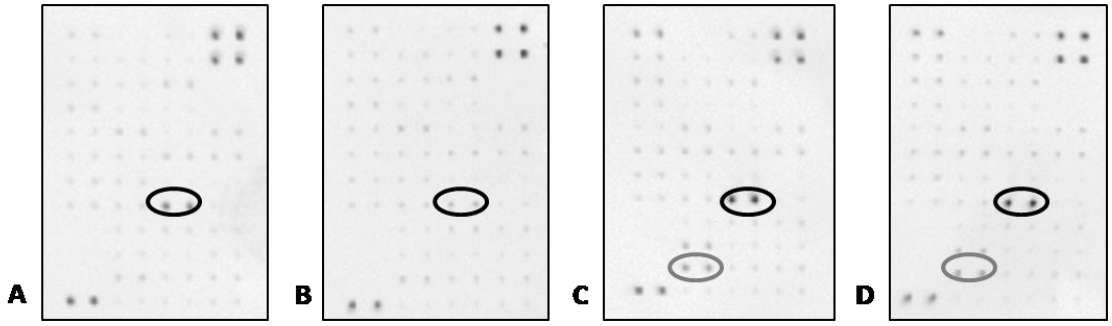


Figure 6.5: Human inflammation antibody array comparing a clinical *H. pylori* isolate associated with cancer (30950) against one associated with gastritis (31235).
Each antibody is present in duplicate. In the top right corner are the two positive controls, followed by the two negative controls. IL-8 is boxed in black. CCL4 is boxed in grey. Complete map of the array is available in Appendix G. **A.** AGS challenged with strain 30950 of *H. pylori*. **B.** AGS challenged with strain 31235 of *H. pylori*. **C.** THP-1 challenged with strain 30950 of *H. pylori*. **D.** THP-1 challenged with strain 31235 of *H. pylori*.

In AGS cells the only strong difference between the cancer and non-cancer strains was for IL-8, with intensities of 0.82 and 0.32 respectively. This difference in IL-8 was also detected in THP-1, but was not as strong as for AGS cells, with intensities of 1.46 and 1.13 in the cancer and non-cancer strains respectively.

Other important differences revealed by this screening procedure were observed in THP-1 cells, between cancer and non-cancer strains, in the MIP-1 family (CCL3, 4 and 15), and for IL-10, IL-12 p70 and I-309 (TCA-3/CCL1). These differences were weaker than the ones observed for IL-8, but were over 0.1, which was our defined threshold.

This screening assay successfully identified two cytokines of interest, IL-8 and CCL4, which were further investigated by ELISA assays on a larger number of strains in order to confirm the observed difference in the two test strains. IL-8 was investigated in both AGS and THP-1 supernatants. CCL4 was investigated in THP-1 supernatants.

6.2.3.2 IL-8 response to *H. pylori* infection in AGS cells

IL-8 was measured in supernatants from AGS cells infected with 15 *H. pylori* strains in triplicate. Most of the strains showed little variation between replicate experiments, with the exception of 4 of the CagPAI-positive non-cancer strains (Figure 6.6A).

CagPAI negative strains induced no IL-8 response. A statistically significant difference was observed between CagPAI-positive and CagPAI-negative strains (Figure 6.6B) in non-cancer patients.

Despite a trend towards non-cancer strains inducing a higher IL-8 production in AGS cells, there was no significant difference compared to the CagPAI positive strains causing cancer (Figure 6.6B).

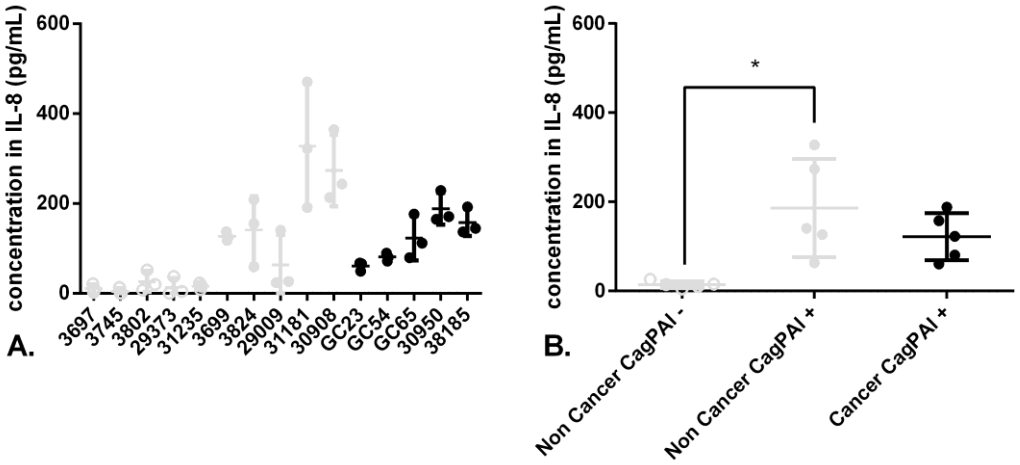


Figure 6.6: Concentration of IL-8 in supernatants obtained after 24h infection of AGS cells with *H. pylori* strains.
IL-8 was measured by ELISA assay and adjusted to the negative control for each experiment. Grey dots represent gastritis strains and black represent gastric cancer strains. Plain circles are for CagPAI positive strains and empty circles for CagPAI negative strains. **A.** Each dot represents the average concentration of IL-8 from triplicate wells of the same experiment. **B.** Each dot represents the average from 3 independent experiments for one strain. Star represents a significant difference between the groups, with a p-value < 0.05 (One-way ANOVA).

6.2.3.3 IL-8 response to *H. pylori* infection in THP-1 cells

The production of IL-8 in relation to CagPAI status has been largely studied in epithelial cells. Little is known about the influence of CagPAI to trigger the production of IL-8 in other types of cells associated with gastric disease, such as macrophages. To address this, an infection experiment using differentiated THP-1 cells as model macrophages was carried out, based on the cytokine array data.

Despite the difference in IL-8 production between the two sample strains used in the human inflammation antibody array, measurement of IL-8 in supernatants from the infection of THP-1 cells with 15 strains of *H. pylori* did not show differences, either with respect to cancer diagnosis or CagPAI status of the strains (Figure 6.7B). A closer look at the results from individual strains also shows a very wide variability between replicates from the same sample (Figure 6.7A).

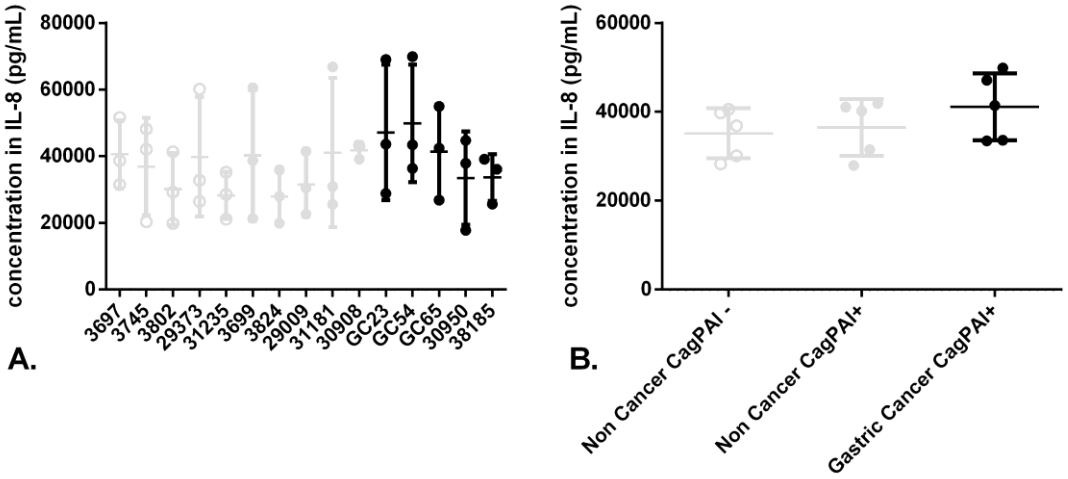


Figure 6.7: Concentration of IL-8 in supernatants obtained after 24h infection of differentiated THP-1 cells with *H. pylori* strains.
IL-8 was measured by ELISA assay and adjusted to the negative control for each experiment. Grey dots represent gastritis strains and black represent gastric cancer strains. Plain circles are for CagPAI positive strains and empty circles for CagPAI negative strains. **A.** Each dot represents the average concentration of IL-8 in triplicate wells from the same experiment. **B.** Each dot represents the average from 3 independent experiments for one strain.

6.2.3.4 CCL4 response to *H. pylori* infection in THP-1 cells

Based on the results from the membrane assay showing a difference in CCL4 response from THP-1 cells between one cancer and non-cancer isolate, further samples were investigated by ELISA.

All 15 samples previously described for study of IL-8 concentration in co-culture with THP-1 were used for this experiment. Variability between replicates with the same strain was significantly lower than for IL-8 in the same samples (Figure 6.7A and Figure 6.8A). The results showed production of CCL4 in all samples, without a clear difference between strains causing cancer and strains causing gastritis. No statistical difference was observed between groups of strains based on disease or CagPAI status (Figure 6.8B).

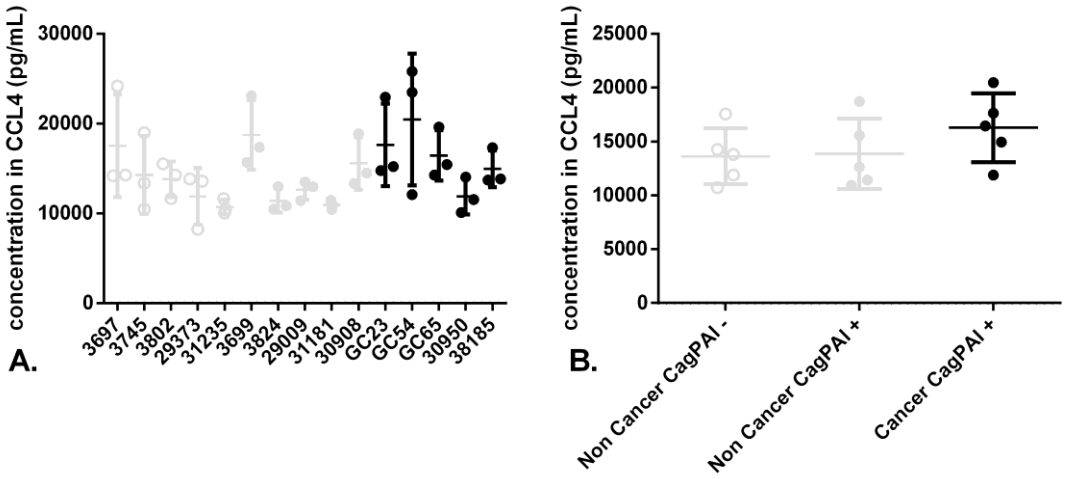


Figure 6.8: Concentration of CCL4 in supernatants obtained after 24h infection of differentiated THP-1 cells with *H. pylori* strains.
CCL4 was measured by ELISA assay and adjusted to the negative control for each experiment. Grey dots represent gastritis strains and black represent gastric cancer strains. Plain circles are for CagPAI positive strains and empty circles for CagPAI negative strains. **A.** Each dot represents the average concentration in IL-8 from triplicate wells in the same experiment. **B.** Each dot represents the average from 3 independent experiments for one strain.

6.2.4 Genomic origin for phenotypic variability

6.2.4.1 Pan-genome of strains used in phenotypic analyses

The pan-genome built from the set of 57 strains used for phenotypic assays included a total of 1913 genes. A genome comparator using this pan-genome as a reference and the 56 strains dataset identified 1319 core genes (present in more than 90% of these strains), and 594 accessory genes (Figure 6.9).

6.2.4.2 Identification of genes associated with motility

The 56 strains used for the measurement of motility were ranked according to motility score. These scores were used to separate strains into three 33rd percentiles:

- 19 strains with low motility,
- 18 strains with average motility,
- 19 strains with high motility.

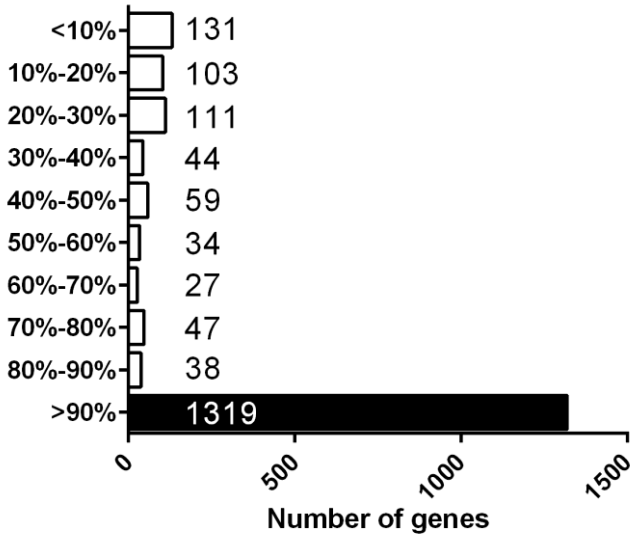


Figure 6.9: Composition of the *H. pylori* pan-genome based on the 56 strains used in phenotypic assays.
Core genome (black bar) was composed of all genes present in at least 90% of the strains. Accessory genome (white bars) was composed of all genes present in at least 1 strain, but less than 90% of the strains.

High and low motility strains were then compared using a genome comparator to identify genes with function linked to an increase in motility. This analysis highlighted 139 genes showing a difference of more than 0.2 in prevalence between high and low motility (Figure 6.10). Of these genes, 62 were present in the reference strain 26695.

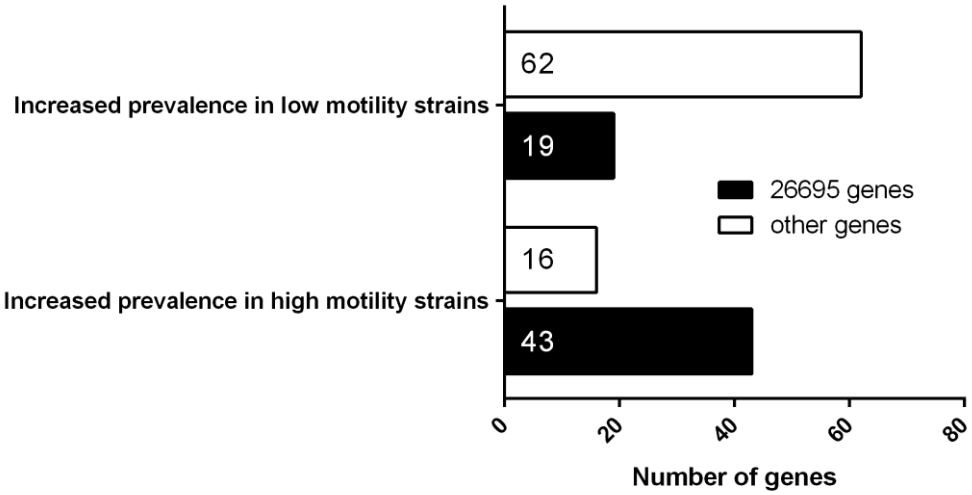


Figure 6.10: Pan-genome approach showing the number of genes with a difference in prevalence of more than 20% between high motility strains and low motility strains.

Only genes with a greater prevalence in high motility strains compared to low motility strains was studied further. The 43 genes from 26695 genome were located across the genome (Figure 6.11). Details for 26695 genes (outside the CagPAI island genes and *HP1243* described below) are presented in Table 6.4. Fifteen genes were not found in the 26695 genome.

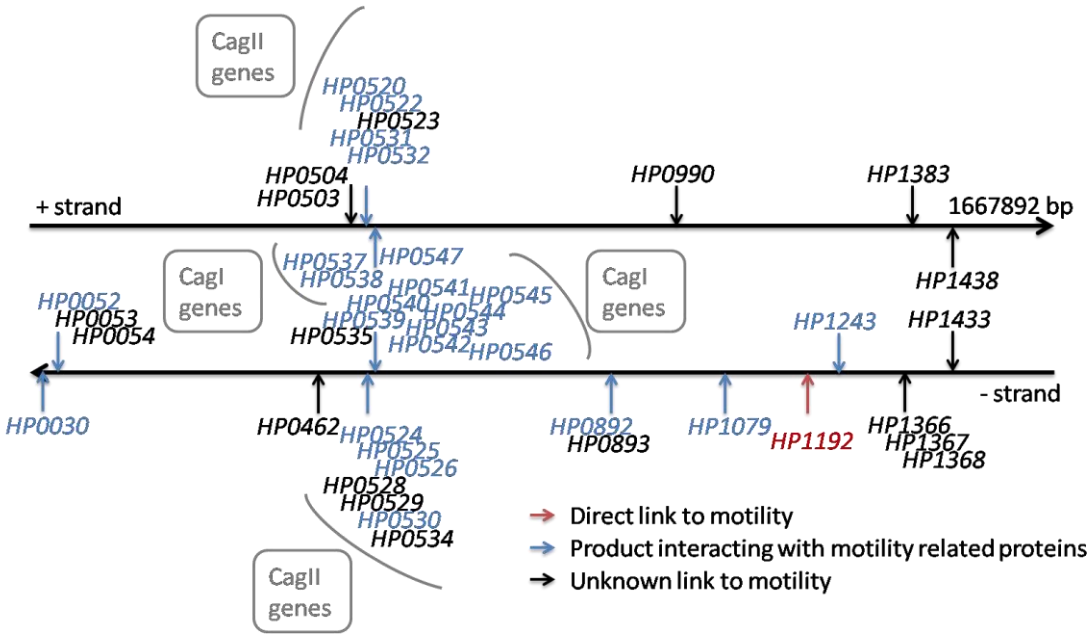


Figure 6.11: Position (in 26695 genome) of the 43 genes showing an increased prevalence in high motility strains.
Direct link to motility identifies genes for which the function is linked to motility (flagella, urease, chemotaxis). Product interactions are based on queries on the interactions of the gene products in PATRICdb revealing links to motility.

HP1243 and cagPAI genes

HP1243 codes for babA, an outer membrane protein known to be highly associated with the CagPAI pathway. This pathogenicity island is associated with gastric cancer. *HP1243* gene was highlighted as having higher presence in high motility strains in our motility assay.

The CagPAI island has been divided in the literature into two distinct zones: cagII (composed of 14 genes) and cagI (composed of 13 genes). Among the 14 genes composing CagII, only 12 were showing increased prevalence in high motility strains.

Table 6.4: Genes with increased prevalence in high motility strains, product functions and predicte
Interactions referenced in this table were obtained on PATRICdb. Only relevant interactions are listed.

Locus tag	Product	Predicted interactions			
		Flagella	**Urease**	**Chemotaxis**	**Others**
HP0033	ClpA (ATP-dependent Clp protease ATP-binding subunit)	*HP0753* (*fliS*)			
HP0052	Putative type II DNA modification enzyme	*HP0325* (*flgH*)			
HP0053	Hypothetical protein				
HP0054	Adenine/cytosine DNA methyltransferase				
HP0462	Type I restriction-modification system, specificity subunit S				*HP0523*
HP0503	Hypothetical protein				
HP0504	Hypothetical protein				
HP0892	YafQ (mRNA interferase)		*HP0073*		*HP0893*, *H*
HP0893	Hypothetical protein				*HP0892*
HP0990	Hypothetical protein				*HP0527*, *H*
HP1079	ATP/GTAP phosphatase		*HP0068*, *HP0072*		*HP1243* (b
HP1192	secreted protein involved in flagellar motility	*HP0115*, *HP0601*, *HP1585*	*HP0068*, *HP0069*, *HP0070*, *HP0072*, *HP0073*	*HP0019*, *HP0103*, *HP0393*, *HP0599*, *HP1067*	*HP0520*, *H*
HP1366	type IIS restriction enzyme R				*HP1578* (L associated)
HP1367	type IIS restriction enzyme M1				*HP1578* (L associated),
HP1368	type IIS restriction enzyme M2				*HP1578* (L associated),
HP1383	modification system S subunit				*HP1208* (u
HP1433	Hypothetical protein				
HP1438	lipoprotein				*HP0523*

Genes not found in 26695

Fifteen coding sequences with increased prevalence in high motility strains were not part of the annotated 26695 genome used as a reference for the construction of the pan-genome. Annotations from RAST did not aid in elucidating the functions for these genes. Twelve genes were annotated as hypothetical proteins. One gene was annotated as an RloF (R-linked ORF F) (005_4_1517). The remaining two genes were closely positioned in the genome of the strain in which they were found. One was annotated as coding for a transposase (004_3_0002), and one for a mobile element protein (004_3_0003).

6.2.4.3 Identification of genes associated with cytokine production in AGS or THP-1 cells

The 15 strains used for infection work were ranked by the concentration of either IL-8 in AGS cells, IL-8 in THP-1 cells or CCL4 in THP-1 cells. A genome comparator was performed using these 15 strains and the pan-genome constructed on the set of 57 strains in order to identify differences in prevalence between strains associated with high production of these cytokines in AGS or THP-1 cells and strains associated with low production of these cytokines in AGS or THP-1 cells. Only genes with differences of more than 0.5 in prevalence between groups were considered relevant (Figure 6.12).

This analysis highlighted very clearly (prevalence of 1 in high producing strains against prevalence of 0 in low producing strains) a link between most cagPAI genes and IL-8 production in AGS cells, and with *babA*. *babA* was also identified in our GWAS analysis as a risk factor for gastric cancer, and has been associated with CagPAI. Confirmation of the relevance of these genes (cagPAI genes and *babA*) with the cytokine production is already present in the literature (Ishijima et al. 2011; Fischer et al. 2001). A few other genes were covarying with IL-8 production in AGS cells, with a lower difference (0.6) between high and low producers. Amongst them were the replicates of *tnpA* and *tnpB*, which were also identified in the study of IL-8 in THP-1 cells. *tnpA* has been linked to more severe disease (Abadi et al. 2014; Mattar et al. 2010), but its exact role has so far not been identified. Our finding could be a lead to further investigation on the link of *tnpA* and *tnpB* with IL-8 production.

HP1499 was identified as more frequent in non atrophic gastritis patients than in gastric cancer or duodenal ulcer, and *HP0962* was specific to gastric cancer strains (Romo-González et al. 2009), but no role was identified regarding IL-8 production. One of the cagPAI genes, *HP0524*, encodes for a protein belonging to the TraG-like protein (Schröder et al. 2002), therefore *traG* and the two genes positioned around *traG*, *HP1003* and *HP1005*, could be linked to immune response, but there is so far no proof of it. *HP0079*, *HP0356*, *HP0462*, *HP0593*, *HP0682*, *HP1078*, *HP1276*, *HP1366/1368*, *HP1471*, *HP1517/1518/1519* were all uniquely identified by our study.

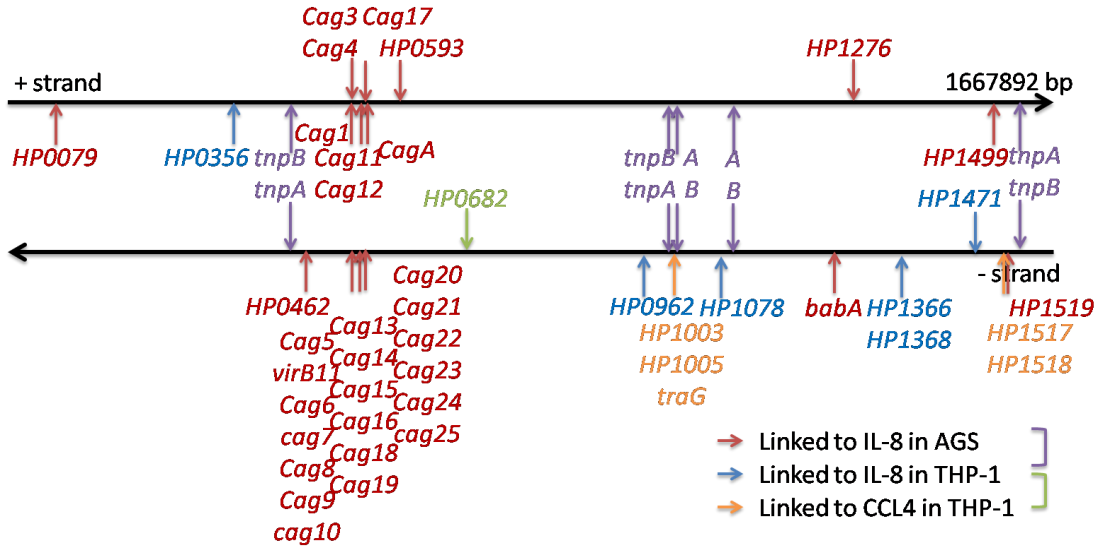


Figure 6.12: Position (in 26695 genome) of the main genes with increased prevalence in strains triggering a high production of cytokines.
Only genes with a difference in prevalence of at least 0.5 between low and high cytokine are represented.

6.2.5 Attribution of functions to genes highlighted by genomic analyses

A genome comparator was performed on the 57 strains dataset using all 110 genes highlighted in previous chapters of this thesis. It showed that 93 genes were core (present in more than 90% of the strains from this dataset) and 17 were accessory.

6.2.5.1 Motility

Prevalence of the genes highlighted in previous chapters was compared in high and low motility strains. A large majority (14 out of 17) of the accessory genes showed a

variation of more than 0.1 between prevalence in high and low motility. Two genes which were part of the core genome in the global dataset were also identified to be core in one of the motility groups but accessory in the other one, with a difference in prevalence of more than 0.1 (Figure 6.13).

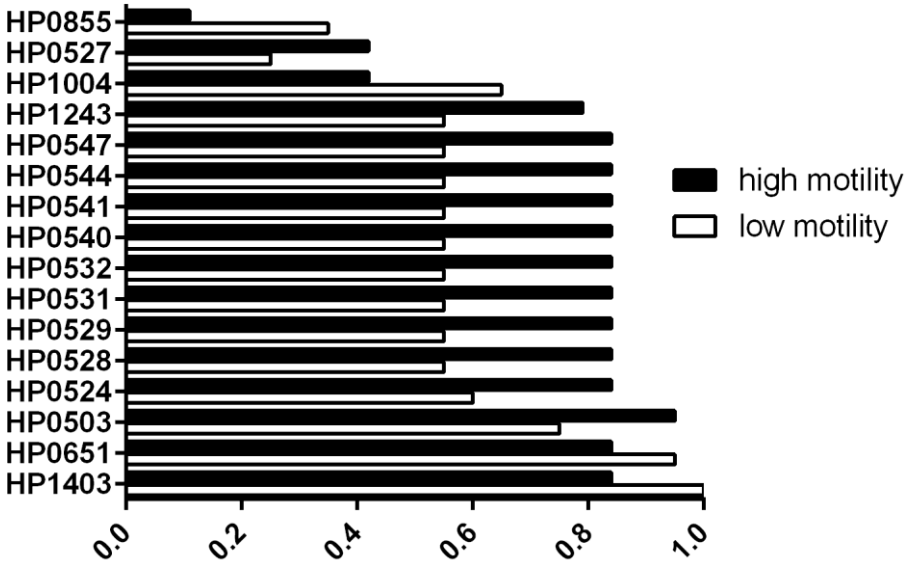


Figure 6.13: Genes highlighted in previous chapters showing a difference in prevalence between strains with high and low motility.

A threshold of 0.2 difference in prevalence was used to select genes covarying with motility, leaving 12 genes.

Two genes showed a higher prevalence in low motility strains than in high motility strains, namely *HP0855* and *HP1004*. *HP0855* showed significant genomic variation based on PV in both strains studied in long-term colonisation in mice. *HP1004* was a hit in the bugwas GWAS with a strong p-value of $2.73.10^{-7}$. Function of this gene is unknown as it codes for a hypothetical protein. The difference in prevalence observed suggests that the function of this gene implies a reduction of strain motility.

All other genes showed a higher prevalence in high motility strains compared to low motility strains. These included cagPAI genes (*HP0524*, *HP0528*, *HP0529*, *HP0531*, *HP0532*, *HP0540*, *HP0541*, *HP0544*, *HP0547*) and genes known to be associated with CagPAI (*HP1243* also known as *babA*).

An investigation of gene-by-gene alignment of the 7 non-accessory risk genotypes identified in Chapter 5 was performed using Bioedit. This analysis recorded an increased proportion of the risk genotype in low motility strains for *HP0797* (Figure

6.14A) and *HP0747* (Figure 6.14B), and an increased proportion of the safe genotype in low motility strains for the synonymous change of *HP0468* (Figure 6.14C). Other genotypes did not show increase or decrease related to motility.

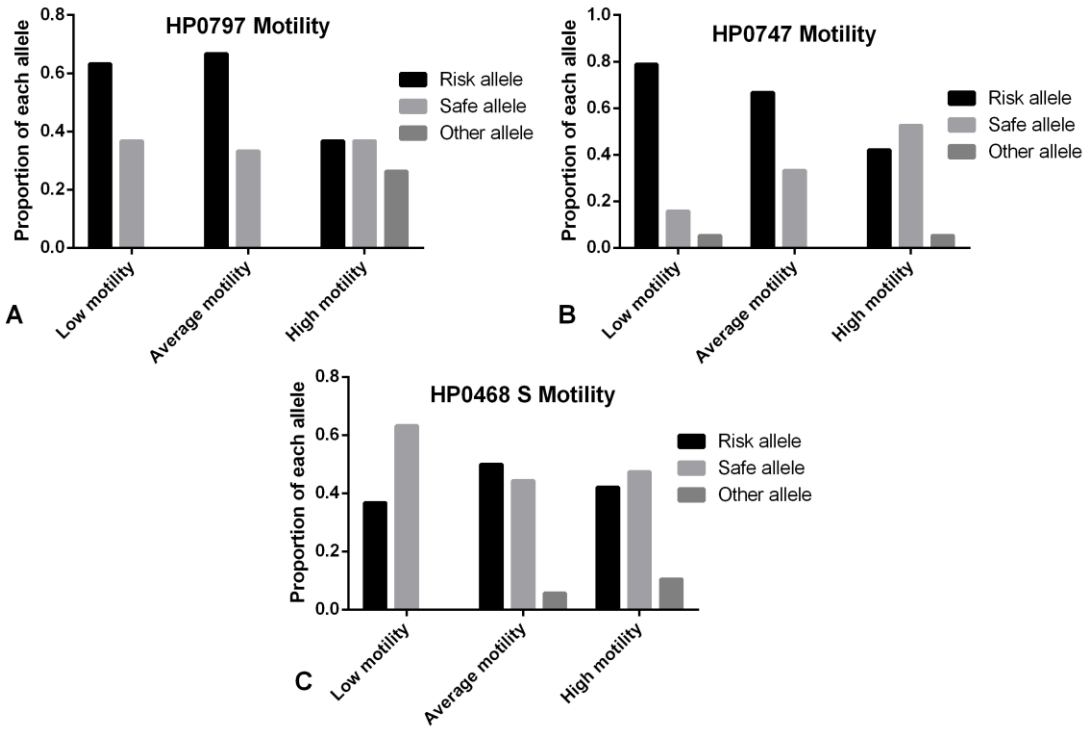


Figure 6.14: Proportion of risk and safe genotypes from Chapter 5 showing an increased or decreased allele presence according to motility. **A.** Non-synonymous allelic change identified in *HP0797*. **B.** Non-synonymous allelic change identified in *HP0747*. **C.** Synonymous allelic change identified in *HP0468*.

6.2.5.2 Triggering of cytokine production

Prevalence of highlighted genes and IL-8 production in AGS cells

Prevalence of the genes highlighted in previous chapters was also compared in strains triggering high and low production of IL-8 in AGS cells. All but one (*HP1045*), accessory genes showed a variation of more than 0.1 between prevalence in high and low motility. However, as the number of strains used in infection experiments was much lower than in the motility assay, the threshold was increased to 0.5. This limit identified 10 genes previously identified in this thesis, all of them part of the CagPAI island (*HP0524*, *HP0528*, *HP0529*, *HP0531*, *HP0532*, *HP0540*, *HP0541*, *HP0544*,

HP0547) or associated with CagPAI (*HP1243* also known as *babA*) that were absent from all 5 strains triggering low production of IL-8 in AGS cells and present in all 5 strains triggering high production of IL-8 in AGS cells. *HP0527*, also part of the CagPAI island, was also absent from all 5 strains triggering low production of IL-8 in AGS cells, but was only present in 3 out of the 5 strains triggering high production of IL-8 in AGS cells (difference in prevalence of 0.6).

Prevalence of highlighted genes and IL-8 production in THP-1 cells

Prevalence of the genes highlighted in previous chapters was also compared in strains triggering high and low production of IL-8 in THP-1 cells. Difference in prevalence was lower than in AGS cells. No gene presented a difference in prevalence of more than 0.5 between groups of strains in this comparison.

Prevalence of highlighted genes and CCL4 production in THP-1 cells

Prevalence of the genes highlighted in previous chapters was finally compared in strains triggering high and low production of CCL-4 in THP-1 cells. Difference in prevalence was again lower than in IL-8 production in AGS cells. However, one gene, *HP1004*, was present in all 5 strains triggering high production of CCL4 and only present in 2 of the 5 strains triggering low production of CCL4, therefore presenting a difference in prevalence of 0.6. *HP1004* was a hit in the bugwas GWAS with a strong p-value of $2.73.10^{-7}$. The function of this gene is unknown as it codes for a hypothetical protein. It was identified as being more present in low motility strains.

Non-accessory risk genotypes and cytokine production

An investigation on gene-by-gene alignment of the 7 non-accessory risk genotypes identified in Chapter 5 was performed using Bioedit for each cytokine. Analysis of the production of IL-8 in AGS cells recorded an increased proportion of the risk genotype in low producers strains for *HP0747* (Figure 6.15A) and both synonymous and non-synonymous changes of *HP0468* (Figure 6.15B-C). An increased proportion of the risk genotype in low producers strains for the non-synonymous change of *HP0709* (Figure 6.15D) was also identified. Other genotypes did not show increase or decrease related to production of IL-8 in AGS cells.

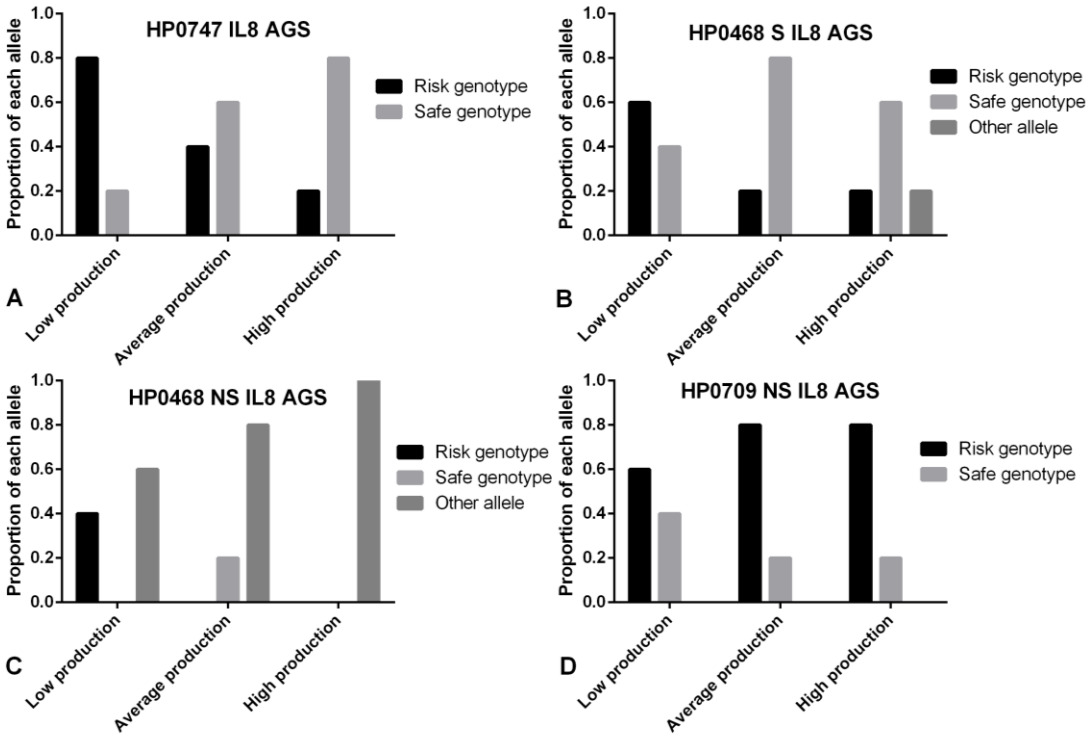


Figure 6.15: Proportion of risk and safe genotypes from Chapter 5 showing an increased or decreased allele presence according to ability to trigger IL-8 production in AGS cells.
A. Non-synonymous allelic change identified in *HP0747*. **B.** Synonymous allelic change identified in *HP0468*. **C.** Non-synonymous allelic change identified in *HP0468*. **D.** Non-synonymous allelic change identified in *HP0709*.

Analysis of the production of IL-8 in THP1 cells recorded an increased proportion of the risk genotype in low producers strains for *HP1055* (Figure 6.16A) and both synonymous and non-synonymous change of *HP0468* (Figure 6.16B-C). An increased proportion of the risk genotype in high producers strains for the non-synonymous change of *HP0709* (Figure 6.16C) was also identified. Other genotypes did not show increase or decrease related to production of IL-8 in THP1 cells.

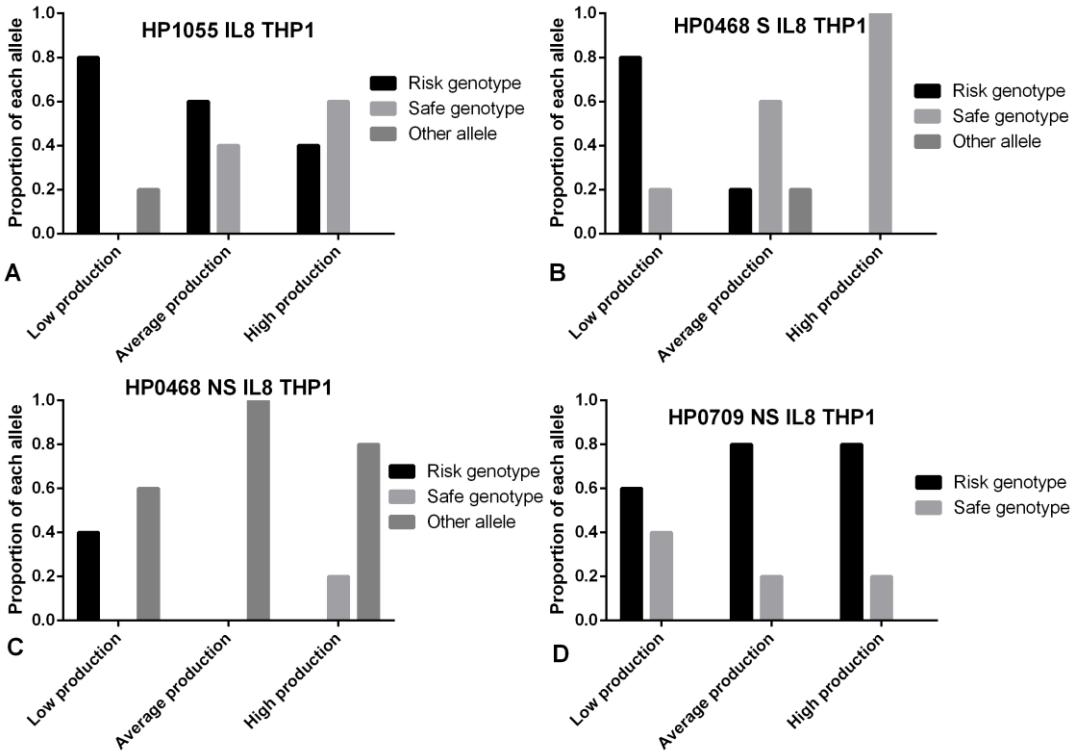


Figure 6.16: Proportion of risk and safe genotypes from Chapter 5 showing an increased or decreased allele presence according to ability to trigger IL-8 production in THP1 cells.
A. Non-synonymous allelic change identified in *HP1055*. **B.** Synonymous allelic change identified in *HP0468*. **C.** Non-synonymous allelic change identified in *HP0468*. **D.** Non-synonymous allelic change identified in *HP0709*.

Analysis of the production of CCL4 in THP1 cells recorded an increased proportion of the risk genotype in low producers strains for the synonymous change of *HP0709* (Figure 6.17C) and both synonymous and non-synonymous changes of *HP0468* (Figure 6.17A-B), and an increased proportion of the safe genotype in low producers strains for the non-synonymous change of *HP0709* (Figure 6.17C). Other genotypes did not show increase or decrease related to production of CCL4 in THP1 cells.

Noticeably, the same increase was observed for all 3 cytokine production experiments regarding the three risk genotypes *HP0468* (synonymous and non-synonymous) and *HP0709* (non-synonymous).

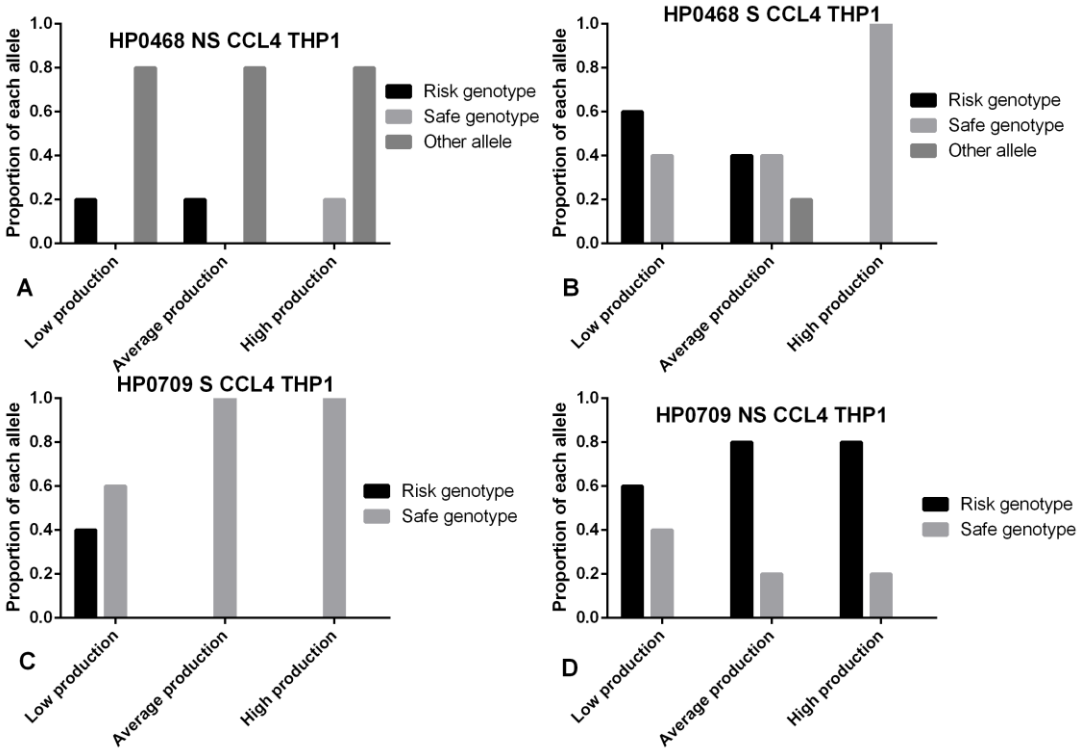


Figure 6.17: Proportion of risk and safe genotypes from Chapter 5 showing an increased or decreased allele presence according to ability to trigger CCL4 production in THP1 cells.
A. Non-synonymous allelic change identified in *HP0468*. **B.** Synonymous allelic change identified in *HP0468*. **C.** Synonymous allelic change identified in *HP0709*. **D.** Non-synonymous allelic change identified in *HP0709*.

6.3 Discussion

The study of phenotypic characteristics of clinical strains of *H. pylori* isolated from patients for whom pathologies were described sheds light on the phenotypic diversity observed in sets of strains.

Motility is essential for colonisation, as a strain with no motility will not be able to colonise the stomach long term, but not all strains with efficient motility will be responsible for strong inflammation and progress towards gastric cancer. Motility is only a pre-requisite for the pathogenicity of the strains to take its part, and it is only indirectly linked to the development of complications in the host (1.2.6). Despite what was previously described (C.-Y. Kao et al. 2012) motility was not associated with the pathology associated with the strains in our experiment. However, a wide variability was observed amongst our collection of isolates. Differences in terms of standard deviations were also observed between pathologies. This could be explained by the

smaller number of isolates available for MALT lymphoma and ulcer, but could also be representative of a difference in the type of disease. MALT lymphoma and duodenal ulcer are associated with a high level of acid production in stomach, whereas gastric cancer is associated with a low level of acid production. *H. pylori* strains isolated from MALT lymphomas and ulcers might be better adapted to a high level of acidity, therefore reducing their motility capacity in the more neutral conditions used in this in vitro experiment.

Pathologies left aside, a number of genes were highlighted as important in relation to motility. One of them, *HP1192* was already associated with motility in a motility assay involving deletion and over-expression mutants for this specific gene (Tsao et al. 2009). Products of other genes with increased prevalence in high motility strains were interacting with colonisation-related genes. This was the case for the famous cagPAI genes (Vannini, Roncarati, and Danielli 2016) and *babA* gene (Rad et al. 2002). Products of *HP0033*, and of the group of 3 genes *HP0052*, *HP0053* and *HP0054* were respectively interacting with flagellar proteins fliS and flgH (J. S. Kim et al. 1999). Products of the couple of genes *HP0892* and *HP0893* as well as product of *HP1079* were interacting with urease gene products *HP0073*, *HP0068* and *HP0072* (Mobley, Hu, and Foxal 1991). Finally, some genes had never been associated to motility and/or were coding for unknown functions, which were found with higher prevalence in higher motility isolates: *HP0462*, *HP0503*/*HP0504*, *HP0990*, *HP1366*/*HP1367*/*HP1368*, *HP1383*, *HP1433* and *HP1438*. These genes could be associated with high motility and would require further research, for instance by motility assays or colonisation experiments on mice with mutant strains.

IL-8 and CCL4 were identified in a semi-quantitative human inflammation antibody assay as being differently produced in infection experiments with strains isolated from cancer patient compared to a non-cancer patient. The smaller difference in IL-8 concentration for THP-1 cells compared to AGS cells can be explained by higher concentrations in the THP-1 supernatants. In the membrane assay, samples were diluted 1 in 2 for both types of cells, and therefore the dots corresponding to IL-8 in THP-1 cells showed close to saturation intensities. The compactness of high concentrations could reduce the observed difference. Measurement of the cytokines in different types of mammalian cells significantly confirmed the role of CagPAI genes in the epithelial cells host response (Ahmadzadeh et al. 2015; Hammond et al. 2015; Censini et al. 1996). IL-8 production in AGS cells was lower in gastric cancer cagPAI

positive strains compared to non cancer cagPAI positive strains, which could suggest that the gastric cancer strains could more easily elude the inflammation developed by the epithelial cells and allow the bacteria to trigger other routes of inflammation in order to develop cancer. However this difference was non-significant, and the number of strains used very small. Pathology set aside, a number of genes were highlighted as covarying with the IL-8 productions in AGS cells, on top of them are the cagPAI genes and genes related to the CagPAI, such as *babA*, *tnpA* and *tnpB* (Abadi et al. 2014). Other genes covarying with IL-8 production in AGS include *HP0079* (membrane protein), *HP0462* (type I restriction-modification system specificity protein), *HP0593* (DNA methyltransferase (Banerjee and Rao 2011)), *HP1276*, *HP1499* (restriction endonuclease) and *HP1519*.

No difference was highlighted regarding both IL-8 and CCL4 production in THP-1, neither according to CagPAI status nor pathology. Considering the human inflammation antibody assay was using samples diluted only 1 in 2, whereas the ELISA measures were using dilutions of 1 in 50, the disaggregation of the proteins obtained by dilution in wash buffer could have altered the results. The higher concentrations observed in THP-1 cells could be due to the nature of the cells which are prompt to adapt and react in a stronger way than AGS cells. Pathology set aside, a number of genes were highlighted as covarying with the IL-8 productions in THP-1 cells, among them are *tnpA* and *tnpB*, which were also covarying with the IL-8 production in AGS cells. *HP0356*, *HP0962* (acyl carrier protein), *HP1078*, *HP1366* (type IIS restriction-modification system endonuclease), *HP1367* (type IIS restriction-modification system methyltransferase) and *HP1471* were also co-varying with IL-8 production in THP-1 cells. *HP0682* was covarying with both cytokines tested in THP-1 cells, IL-8 and CCL4, but its function is unknown as it is described as a hypothetical protein. Also covarying with CCL4 in THP-1 cells were *HP1003*, *HP1005*, *traG* (involved with T4SS (Schröder et al. 2002)), *HP1517* and *HP1518*.

Genes identified in the previous results chapter (specifically chapter 4 and 5) also showed co-variations with motility or cytokine production, which shed light on the potential mechanisms involving those genes with change of host, long-term colonisation or gastric cancer.

Experiments such as knocking out some of the genes identified could help confirm the questions remaining on the exact functions linked to motility and triggering of immune response in the host (O'Toole, Kostrzynska, and Trust 1994; Schmitz,

Josenhans, and Suerbaum 1997; C.-Y. Kao, Sheu, and Wu 2014). This approach would be especially interesting on genes for which function is unknown but are highlighted in more than one aspect of our study. Other phenotypic experiments could also have been carried out on this set of clinical strains in order to link disease development in the host with phenotypic characteristics of the strains, and to identify potential genes linked to these variations. Reactive Oxygen Species produced by host cells when undergoing stress, could for instance be an interesting approach to the study of *H. pylori* strains (Satin et al. 2000). Infection of THP-1 and AGS cells co-cultures by *H. pylori* strains could also be performed, to model interactions between these three actors during chronic inflammation caused by the bacteria (Fox et al. 2015).

In conclusion, variations in motility are not linked to the pathology of the patient infected. However, some genes are co-varying with motility and could be investigated further. Immune response measured in co-culture experiments with clinical strains varies with the pathology of the patient from who the strain. Genes already known to be associated with triggering of IL-8 production in epithelial cells were identified in our study but new ones were also co-varying with cytokine triggering phenotype. These could be targets for experiments in order to confirm or dispute the association.

7 General Discussion

This chapter will focus on answering the questions defined in introduction using the whole results exposed in previous chapters. More precisely, the two first results chapters, Chapter 3 and Chapter 4, focused on defining the genome evolution occurring in *H. pylori* species, addressed four hypotheses:

- The genomic variability of *H. pylori* strains from the Americas reflects the history of recent and ancient migrations which built the identity of these regions,
- Core and accessory genomes are evolving in a similar way,
- *H. pylori* strains evolve when changing from one host to another,
- *H. pylori* strains infecting a stomach for a long time evolve alongside the development of symptoms.

Mimicking genomic variability of *H. pylori* strains, a wide phenotypic variability is observed in clinical strains. This was addressed in Chapter 6, with three hypotheses:

- Motility varies according to the pathology of the patient from which the strain was isolated,
- Immune response is triggered differently according to the pathology of the patient from which the strains was isolated,
- Some genes covary with phenotypic differences observed among strains.

Finally, GWAS methods were used on *H. pylori* clinical strains in an attempt to identify genomic traits linked to gastric cancer (GC). Three hypotheses were tested in Chapter 5:

- The GWAS method can be applied to *H. pylori* genome despite its high variability,
- Specific genomic traits in specific genes can be linked with the progression of GC,
- A risk score can be built in order to target strains with a higher risk for triggering GC.

Limitations of the methods used in this study, as well as leads for future research, will also be discussed.

7.1 Genome evolution in *H. pylori*

The *H. pylori* species is particularly diverse (Dorer, Sessler, and Salama 2011). Genomic variation can be observed at different levels. Firstly, traces of the history of human migrations are conserved in its genome, due to their long co-existence (Falush et al. 2003; Moodley et al. 2012). This was studied by MLST methods for old patterns of migrations such as the ones originating in the current European population (Falush et al. 2003). Indeed, European *H. pylori* populations are known to be a mix of Asian and African populations. This was confirmed in a published study (Thorell et al. 2017) based on complete genomes (see Chapter3), both in the core genome, through the use of FineStructure and Chromopainter (3.2.2), and in the accessory genome through the use of a new method of analysis of accessory 3-dimentional plots (3.2.3). These higher resolution methods, combined with the large global dataset which included 198 strains from the Americas, allowed the identification of more recent events of migration which are part of the history of the Americas, highlighting differences between regions of this continent (Marangoni, Caramelli, and Manzi 2014). Different subpopulations were newly identified in the Americas, with different levels of inputs from hpAfrica1 and hpEurope. Both core and accessory methods were confirming this evolution in a similar way.

However, the genome of *H. pylori* not only comprises traces of ancient hosts, but also evolves over a shorter scale (Cao et al. 2015; Avasthi et al. 2011). Colonisation of a human host by *H. pylori* usually occurs in childhood, but can persist for years without any symptoms (O'Ryan et al. 2015; Dooley et al. 1989). During this time, a single strain of *H. pylori* can evolve, adapting to this host and sometimes causing gastritis or more severe outcomes such as gastric ulcer (Lanas and Chan 2017), MALT lymphoma (ML) (H.-C. Wang et al. 2015) or gastric cancer (GC) (Wroblewski and Peek 2016; Figura et al. 2016). Different strains of *H. pylori* can also co-habit, increasing the potentiality of genomic evolution through recombination (Cao et al. 2015; Kersulyte, Chalkauskas, and Berg 1999). An attempt to assess the usability of murine model to study long-term colonisation evolution of *H. pylori* genome was made in Chapter 4. The isolation of strains of *H. pylori* after colonisation of a mouse host for different durations was an opportunity to study the evolution of a single strain in its host during long-term colonisation, and to assess the changes occurring in *H. pylori* strains when changing host. The two strains used in this study were originally

isolated from human patients suffering from ML (Chrisment et al. 2014). Gastric lesions similar to those observed in the human original host developed in the infected mice after long-term colonisation, suggesting a bacterial origin to the symptoms over a host pre-disposition. The genomic study made on these strains showed that specific changes were observed during the change of host (4.2.2), and during long-term colonisation (4.2.3). The genomic evolutions that were observed were of different types, backed-up by literature: PV (Appelmelk et al. 1999; Bergman et al. 2006), SNP (Furuta et al. 2015), deletion/insertion (Tsang et al. 2013) and repetition of a 21bp sequence (Rasko et al. 2000). Functions of the genes evolving in this study comprised outer-membrane proteins (Furuta et al. 2015), LPS biosynthesis (Stefan Odenbreit, Faller, and Haas 2002; Chmiela, Miszczyk, and Rudnicka 2014; Appelmelk et al. 1999) and flagella (C.-Y. Kao et al. 2012; O′Toole, Lane, and Porwollik 2000), which are all crucial for colonisation.

7.2 Phenotypic variations in *H. pylori* strains

Strains of *H. pylori* are not only variable in terms of genomes. Indeed, in most cases, genomic variations affect the phenotypes of the strains. A collection of European clinical isolates was gathered in order to study the phenotypic variations amongst them (Chapter 6). The focus of our experiments was to target phenotypes that were both (i) achievable for large number of isolates (ii) representative of the virulence of the strains. No significant variation was found in motility between isolates from different pathology. However, the ability to trigger an IL-8 response in AGS cells was higher in gastric cancer isolates than in non cancer isolates reflecting the presence or absence of the CagPAI island genes (Li et al. 1999; Sheh et al. 2013; Khatoon et al. 2017). No difference in the ability to trigger either an IL-8 or CCL4 response in THP-1 cells was associated with the pathology of the isolates. However, 113 genes were co-varying with the phenotypic variations measured in our set of isolates. These covariations highlighted genes already known to be associated with such phenotypes (CagPAI genes and *babA* with IL-8 in AGS cells), but also genes that were not expected to covary with those phenotypes. Studying those genes, to confirm or refute their association with phenotypes could be of interest to better understand the mechanisms behind pathogenicity of *H. pylori* strains.

7.3 Prediction of virulence

GWAS methods, initially developed for human genomics, have been increasingly used in bacterial genomics (Power, Parkhill, and de Oliveira 2016; P. E. Chen and Shapiro 2015; Alam et al. 2014; Sheppard et al. 2013). These methods, relying on a statistical analysis of the prevalence of genomic traits in strains for which a binary phenotype can be clearly identified, are powerful tools to investigate the virulence of *H. pylori* strains. Application of GWAS to clinical strains of *H. pylori* from a single population (hpEurope) highlighted 12 genotypes associated with gastric cancer (Chapter 5). These genotypes were used to build a risk score that could allow prediction of the virulence of a strain.

In the different chapters of this thesis, a large number of genes were highlighted, linked to different evolutionary mechanisms. Markers of evolution in an animal model, linked to both change of host or long-term colonisation, were identified in Chapter 4. Co-variations of certain genotypes with phenotypes such as motility or immune system triggering were highlighted in Chapter 6. Finally association of some genomic traits with gastric cancer was verified using different GWAS methods in Chapter 5. All these genes are referenced in Appendix H. Some of these genes were highlighted by independent aspects of this thesis work, as well as other studies from the literature. Independent studies converging towards individual genes are also a way to understand and predict virulence of a strain. In total, 35 genes were highlighted for more than one reason in this work. cagPAI genes and *babA* have already been highly studied, and can be considered positive controls. Other genes highlighted at least twice in this thesis are presented in Table 7.2. This list of genes could be investigated more deeply to understand the functional reasons behind their identification in this study.

7.4 Clinical applications of genomic-based prediction of virulence

Genomic-based prediction of virulence using GWAS presents the advantage of not being based on assumptions of the function of the genes. In our method, all the genes from all the strains from the dataset are equally considered, allowing the identification of potential new virulence factors. In this work, a first attempt of risk score was made. Such a tool could be of use in clinics, to decide whether or not the infection should be treated. Indeed, most of the infected population remains asymptomatic (Dooley et al.

1989). On the other hand, the consequences, when the infection is not asymptomatic, can be lethal (Ferlay et al. 2015). For this reason, the current guidelines are to treat an infection as soon as it is detected (Malfertheiner et al. 2017). All lines of treatment for *H. pylori* infections being based on a combination of antibiotics, the resistance issue is on the rise (Ierardi et al. 2013). A more careful use of antimicrobial therapy, reserved to those with the most virulent strains, could be a solution to the rise in antibiotic resistance. Risk score, in combination with evaluation of symptoms, could determine the choice of treatment.

Table 7.1: Summary of places in this thesis where cagPAI genes and babA were highlighted

Gene Tag	Gene name	Chapter 2	Chapter 3		Chapter 4	
		Change of Host	ClonalFrame GWAS	bugWAS	Motility	IL-8 in AGS
HP1243	*babA*	X	X	X	+	+
HP0544	*cag23*		X	X	+	+
HP0524	*cag5*			X	+	+
HP0528	*cag8*			X	+	+
HP0529	*cag9*		X		+	+
HP0531	*cag11*			X	+	+
HP0532	*cag12*			X	+	+
HP0540	*cag19*			X	+	+
HP0541	*cag20*			X	+	+
HP0547	*cag26*		X		+	+
HP0520	*cag1*				+	+
HP0522	*cag3*				+	+
HP0523	*cag4*				+	+
HP0525					+	+
HP0526	*cagZ*				+	+
HP0527	*cag7*			X		+
HP0530					+	+
HP0534	*cag13*				+	+
HP0537	*cag16*				+	+
HP0538	*cag17*				+	+
HP0539	*cag18*				+	+
HP0542	*cag21*				+	+
HP0543	*cag22*				+	+
HP0545	*cag24*				+	+
HP0546	*cag25*				+	+

Table 7.2: Genes highlighted in one or more study in this thesis

Gene tag	Gene name	description	Chapter 2		Chapter 3	
			Change of Host	**Colonisation**	**ClonalFrame GWAS**	**BugWAS**
HP0068	*ureG*	urease accessory protein UreG			X	X
HP0462	*hsdS*	type I restriction-modification system specificity protein				
HP0468		hypothetical protein			X	X
HP0503		hypothetical protein			X	
HP0685	*fliP*	flagellar biosynthesis protein FliP	X		X	
HP0855		peptidoglycan O-acetyltransferase		X		
HP1004		hypothetical protein				X
HP1177	*hopQ*	membrane protein			X	X
HP1252	*oppA*	oligopeptide ABC transporter substrate-binding protein	X		X	
0010_9_0525		hypothetical protein				

Leaving less virulent populations of *H. pylori* in place could also be beneficial. Indeed, a small number of studies have highlighted beneficial effects of *H. pylori* infection for the host (O'Connor, O'Morain, and Ford 2017). These positive effects do not outweigh the negative effects, but if one could predict the virulence of the strain, this could change the treatment regimens in the future. Sequencers are becoming cheaper and easier to use, and it becomes more and more realistic to be able to use such techniques for standard diagnosis in a public health context (Pallen, Loman, and Penn 2010). New sequencers are developed and some of them are small and portable, such as the Oxford Nanopore MinION (Lu, Giordano, and Ning 2016; Walter et al. 2017). It could soon be a reality to sequence the strain from a biopsy, run an application and obtain a risk score, used to help the clinician in the decision to treat the infection.

7.5 Limitations of the thesis

Although prediction of virulence through GWAS methods shows promises, there are a number of limitations to this method, some of which could be avoided by future research on the subject. The first limitation is the fact that GWAS requires a binary dataset. *H. pylori* linked pathologies are not binary, and therefore choices have to be made to turn the dataset into a usable binary dataset. The robustness of this binary dataset relies on the quality of information given with the sequences. Moreover, virulence of *H. pylori* expresses itself in many different ways, and there are many outcomes with symptoms often highly divergent. In our study, genotypes linked to gastric cancer were identified, leading to the construction of a risk score.

This risk score was built as a short proof of concept study. Indeed, there was no dataset at the time of this work to allow validation of the risk score in an independent dataset. A validation in a dataset outside hpEurope, the population used in the creation process, could be considered, but the genomic differences between two *H. pylori* populations are too important and there is a high risk that the risk score could not be used in another population. For these reasons, the risk score built in this thesis should not be used as it is for clinical decisions. A GWAS and risk score should be performed for each *H. pylori* population, validated on independent datasets, and the relevant risk score should be used in clinics, based on the population to which the clinical isolate belongs. Population can be determined quickly by the position of the

isolate on a reference phylogenetic tree containing 5 isolates per main population (Vale et al. 2017).

Finally, all genomic analyses rely on clinical isolates. As it is considered unethical to not treat a patient when the infection is detected, it is impossible to have isolates from symptomatic patients before the development of the symptoms. This is a limitation, as symptoms can develop after a long-term infection during which the infecting strain can evolve. Considering the slow and large alterations of the stomach in some of the outcomes, such as gastric cancer, it is highly possible that the strain triggering the cancer is not present in the same version anymore, if at all, at the time of sampling. This is therefore a bias, which could lead to GWAS identifying genomic traits common to strains able to survive in the gastric cancer conditions, instead of genomic traits common to strains triggering this gastric cancer.

Limitations in terms of isolates used also impact Chapter 4. Indeed, this analysis relied on only 2 clinical strains from MALT lymphoma patients. Each of these two strains were used to infect mice, but only 3 isolates were re-isolated for each time-point, which is a very low number for genomic analysis. A larger number of isolates would be necessary to increase the reliability of the results and allow statistics to be made.

7.6 Future directions

If this thesis work could be taken further, different leads could be followed. First, there is a need for strong GWAS based not only on gastric cancer strains, but also ulcer or MALT lymphoma. More rigorous patient data associated with the strains used would also increase the quality of such analyses. To achieve this, a large number of strains would be needed, with information about the patients, and controlled population structure. Indeed, GWAS requires a large number of strains to be robust (a minimum of 100 strains per binary group), due to its fundamental principle based on statistics. A collaborative project named Helicobacter pylori Genome Project (HpGP), presented at the European Helicobacter and Microbiota Study Group (EHMSG) conference in 2017 (http://ehmsg.org/2017/programme.htm), is leading the way toward an increase in quality and volume of *H. pylori* strains collections. Once suitable collections are available, it will be possible to achieve GWAS, on different datasets corresponding to the different outcomes of an *H. pylori* infection, which will give us a more global view of the virulence of *H. pylori*. These GWAS will have to be

achieved on different Hp populations, to take into account the differences between those populations (Thorell et al. 2017). Validation of risk scores using independent datasets from the same population are also a necessity for risk scores to be used in clinics.

A new GWAS method, based on phylogeny (Collins and Didelot 2018), could also be considered in order to account for population structure and recombination while achieving a high power and specificity. Indeed, despite the recombination being taken into account in the two methods used in this thesis, these methods presented limitations in terms of specificity.

Even with strong datasets made available for GWAS studies, and optimal GWAS methods, validation of the genotypes identified in genomics studies remains essential. This can be done with conventional laboratory methods such as creation of mutants, and studies of these mutant strains *in vitro* and *in vivo* (Schmitz, Josenhans, and Suerbaum 1997; C.-Y. Kao, Sheu, and Wu 2014). Even though construction of mutant strains was not an option during this thesis, phenotypic studies were achieved. They could have been taken further through the co-culture of *H. pylori* clinical strains AGS in combination with THP-1 (Fox et al. 2015). This would model more closely the reality of the stomach environment. pH could also be controlled, as we identified a few genes related to the buffering of pH in our genomic studies. Relationship between *H. pylori* and other components of the gastric microbiota could also be investigated.

Long-term colonisation *in vivo* experiments would also be highly valuable to *H. pylori* research. Animal models such as the one presented in Chapter 4 could be taken further, with larger number of replicates, but also with controls for the evolution of the experimental strains in a laboratory or in different hosts. Colonisation of healthy human hosts and genomic analysis of original and re-isolated strains from such experiments would also shed light on the genomic aspects of long-term colonisation by *H. pylori*. Such a study would require strong ethics, and a research group from Germany presented an on-going project that could address this question at the CHRO conference in September 2017 (http://www.chro2017.com/content.php?PAGE=11).

In conclusion, multi-disciplinary collaborations are the key to a complete picture of *H. pylori* role in the diverse clinical outcomes following infection. And this picture, when complete, will have to be included in a bigger one, by taking into account the role of host genetics and environmental factors in the development of disease, as well as the complex interactions between these three interdependent factors.

Appendices

Appendix A: Published article: Recent “omics” advances in *Helicobacter pylori*

This review article was published in the Helicobacter Journal in 2016 on invitation by Prof Francis Megraud.

DOI: 10.1111/hel.12334

REVIEW ARTICLE

WILEY Helicobacter

Recent “omics” advances in *Helicobacter pylori*

Elvire Berthenet[1] | Sam Sheppard[2] | Filipa F. Vale[3]

[1]College of Medicine, Institute of Life Science, Swansea University, Swansea, UK

[2]Departments of Biology and Biochemistry, University of Bath, Claverton Down, Bath, UK

[3]Host-Pathogen Interactions Unit, Research Institute for Medicines (iMed-ULisboa), Instituto de Medicina Molecular, Faculdade de Farmácia da Universidade de Lisboa, Lisboa, Portugal

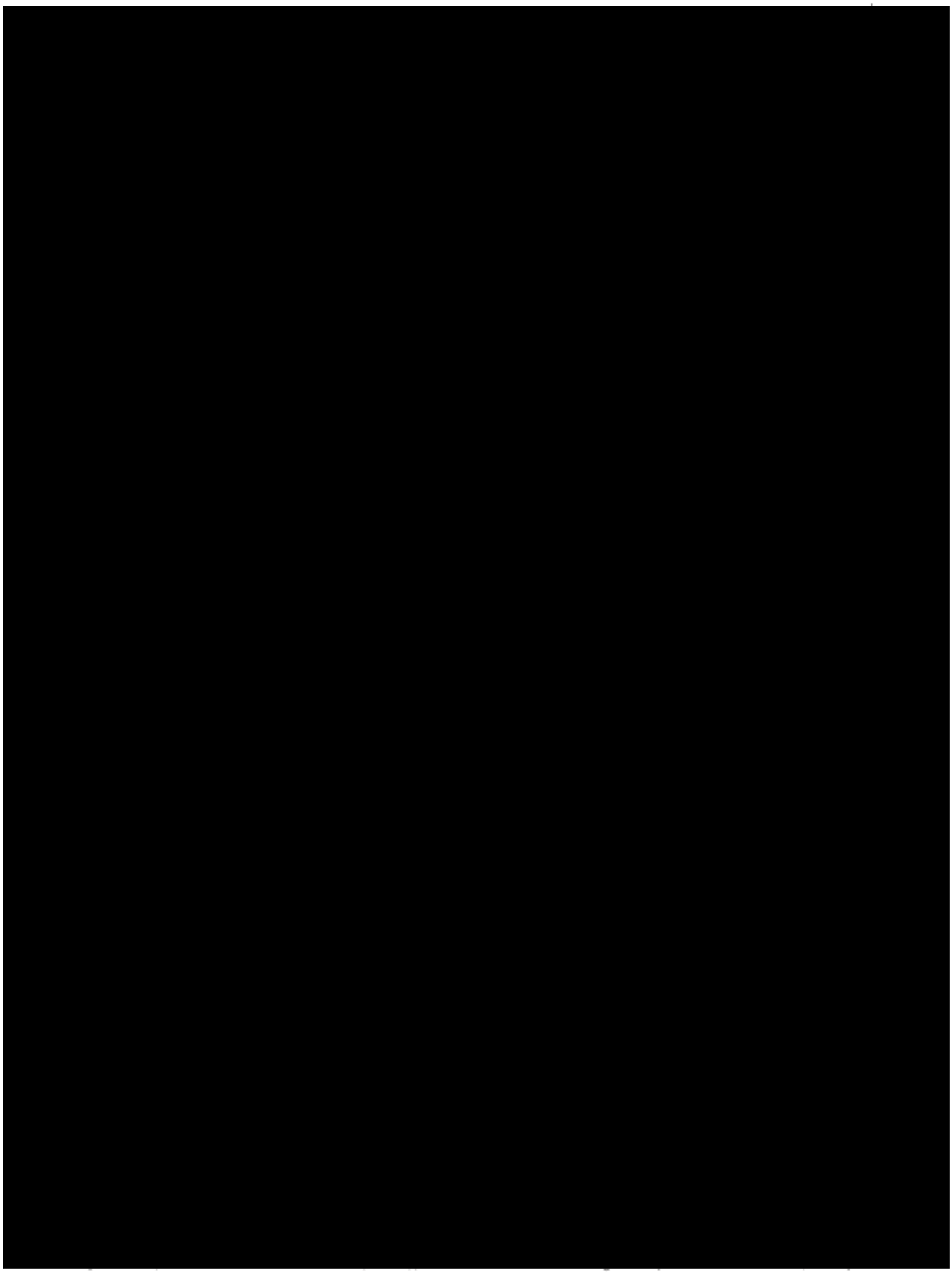
Correspondence
Sam Sheppard, Departments of Biology and Biochemistry, University of Bath, Claverton Down, Bath, UK.
Email: s.k.s.sheppard@bath.ac.uk
and
Filipa F. Vale, Host-Pathogen Interactions Unit, Research Institute for Medicines (iMed-ULisboa), Instituto de Medicina Molecular, Faculdade de Farmácia da Universidade de Lisboa, Lisboa, Portugal.
Email: vale.filipa@gmail.com

Abstract

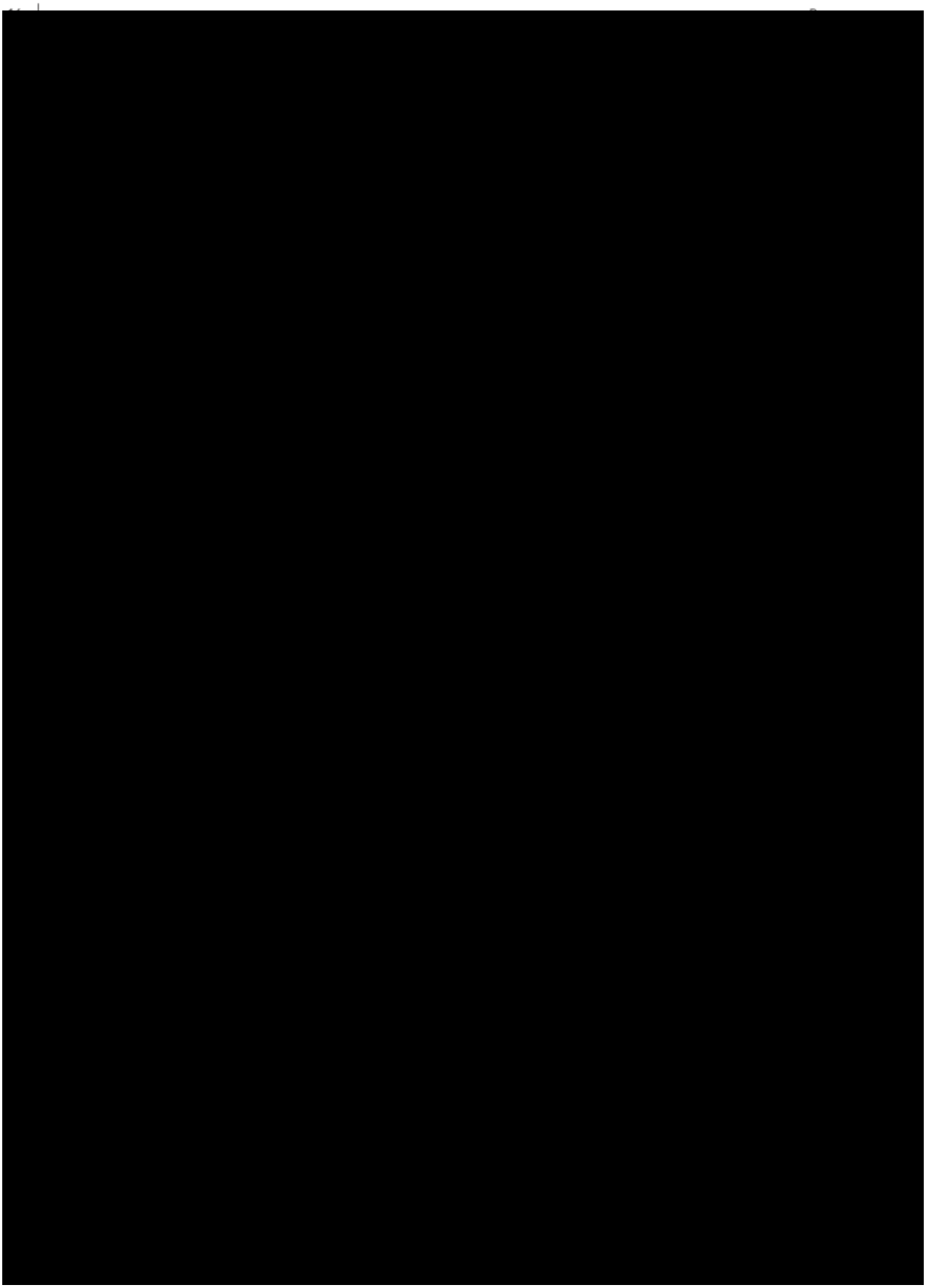
The development of high-throughput whole genome sequencing (WGS) technologies is changing the face of microbiology, facilitating the comparison of large numbers of genomes from different lineages of a same organism. Our aim was to review the main advances on *Helicobacter pylori* "omics" and to understand how this is improving our knowledge of the biology, diversity and pathogenesis of *H. pylori*. Since the first *H. pylori* isolate was sequenced in 1997, 510 genomes have been deposited in the NCBI archive, providing a basis for improved understanding of the epidemiology and evolution of this important pathogen. This review focuses on works published between April 2015 and March 2016. Helicobacter "omics" is already making an impact and is a growing research field. Ultimately these advances will be translated into a routine clinical laboratory setting in order to improve public health.

KEYWORDS

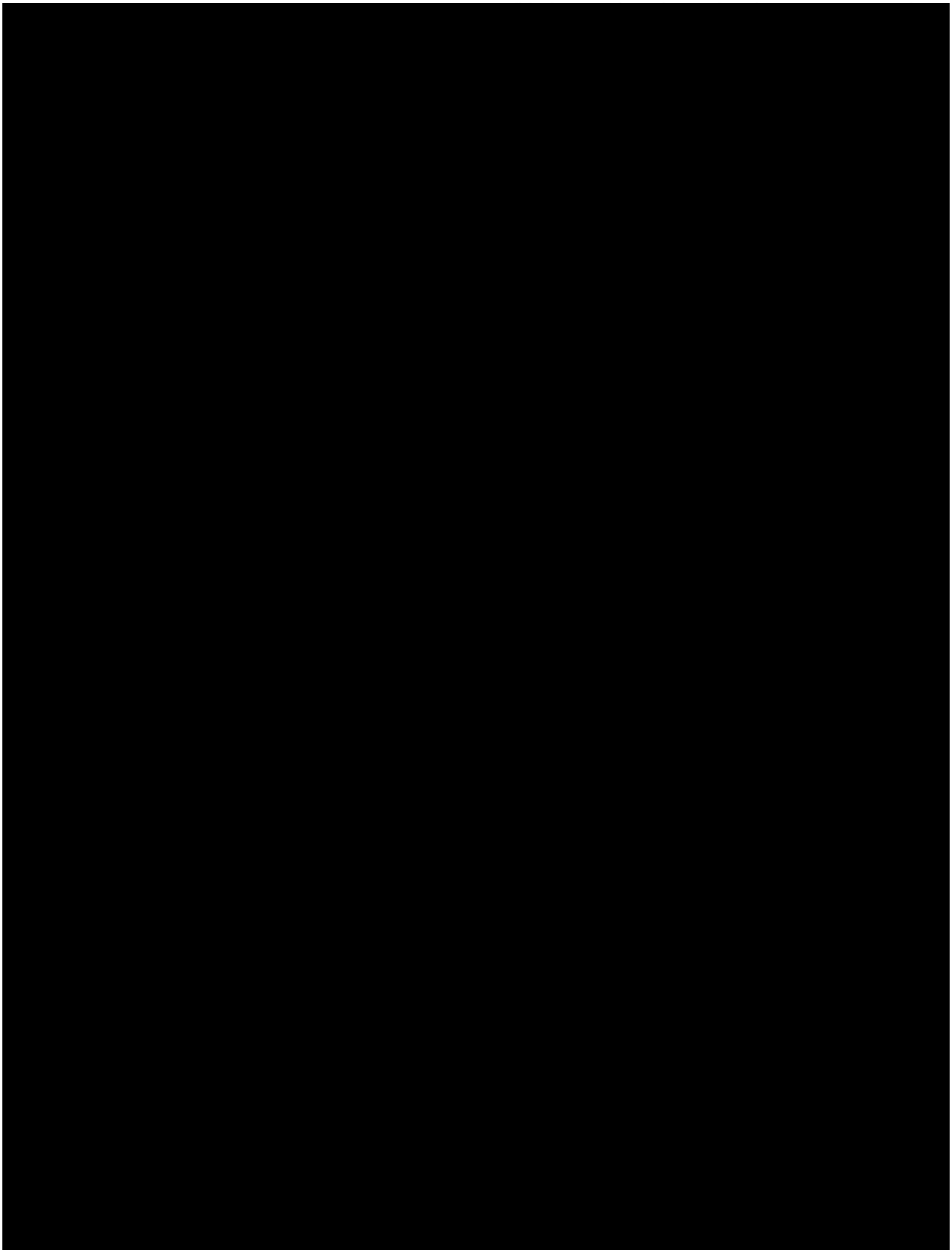
genetic admixture, genomics and transcriptomics, mobile elements, pathogenicity islands, target new therapy



Appendix A: Published article: Recent “omics” advances in *Helicobacter pylori*



Appendix A: Published article: Recent “omics” advances in *Helicobacter pylori*



Appendix A: Published article: Recent “omics” advances in *Helicobacter pylori*

Appendix B: Published article: Genomic structure and insertion sites of *Helicobacter pylori* prophages from various geographical origins

This research article was published in the Scientific Reports Journal in 2017 by Filipa Vale. My participation in this article was in identification of populations of the *H. pylori* genomes used.

SCIENTIFIC REPORTS

OPEN

Received: 05 July 2016
Accepted: 10 January 2017
Published: 16 February 2017

Genomic structure and insertion sites of *Helicobacter pylori* prophages from various geographical origins

Filipa F. Vale[1,2], Alexandra Nunes[3], Mónica Oleastro[4], João P. Gomes[3], Daniel A. Sampaio[5], Raquel Rocha[4], Jorge M. B. Vítor[6], Lars Engstrand[7], Ben Pascoe[8], Elvire Berthenet[9], Samuel K. Sheppard[8], Matthew D. Hitchings[9], Francis Mégraud[2], Jamuna Vadivelu[10] & Philippe Lehours[2,11]

Helicobacter pylori **genetic diversity is known to be influenced by mobile genomic elements. Here we focused on prophages, the least characterized mobile elements of *H. pylori*. We present the full genomic sequences, insertion sites and phylogenetic analysis of 28 prophages found in *H. pylori* isolates from patients of distinct disease types, ranging from gastritis to gastric cancer, and geographic origins, covering most continents. The genome sizes of these prophages range from 22.6–33.0 Kbp, consisting of 27–39 open reading frames. A 36.6% GC was found in prophages in contrast to 39% in *H. pylori* genome. Remarkably a conserved integration site was found in over 50% of the cases. Nearly 40% of the prophages harbored insertion sequences (IS) previously described in *H. pylori*. Tandem repeats were frequently found in the intergenic region between the prophage at the 3′ end and the bacterial gene. Furthermore, prophage genomes present a robust phylogeographic pattern, revealing four distinct clusters: one African, one Asian and two European prophage populations. Evidence of recombination was detected within the genome of some prophages, resulting in genome mosaics composed by different populations, which may yield additional *H. pylori* phenotypes.**

Helicobacter pylori is a major widely distributed human pathogen, with one out of two persons being colonized by this bacterium. Infection by *H. pylori* is associated with gastritis and may progress to more severe conditions, including peptic ulcer and, in rare cases, gastric adenocarcinoma and gastric MALT (mucosa associated lymphoid tissue) lymphoma. *H. pylori* presents a phylogeographic distribution, reflecting a pattern of co-evolution with the human host[1].

Genome rearrangement and high rate of mutation are characteristics of *H. pylori*[2,3], described as a highly genetic diverse[4]. Furthermore, this variability is reinforced by epigenome diversity[5,6]. Among the factors for increased diversity there are mobile genomic elements, including the *cag*-pathogenicity island (PAI)[7], insertion sequences[8], restriction-modification systems[9,10] and prophages[11]. Furthermore, *H. pylori* is among the most recombinogenic known human pathogens[12].

[1]Host-Pathogen Interactions Unit, Research Institute for Medicines (iMed-ULisboa), Faculdade de Farmácia da Universidade de Lisboa, Lisboa, Portugal. [2]Université de Bordeaux, Centre National de Référence des Campylobacters et Hélicobacters, Bordeaux, France. [3]Bioinformatics Unit, Department of Infectious Diseases, National Institute of Health, Lisboa, Portugal. [4]National Reference Laboratory of Gastrointestinal Infections, Department of Infectious Diseases, National Institute of Health, Lisboa, Portugal. [5]Innovation and Tecnhology Unit, Department of Human Genetics, National Institute of Health, Lisboa, Portugal. [6]Department of Biochemistry and Human Biology, Faculdade de Farmácia, Universidade de Lisboa, Lisboa, Portugal. [7]Department of Microbiology, Tumor and Cell Biology, Karolinska Institute, Stockholm, Sweden. [8]The Milner Centre for Evolution, Department of Biology and Biochemistry, University of Bath, Bath, UK. [9]Institute of Life Science, Swansea University Medical School, Swansea, UK. [10]UM Marshall Centre and Department of Medical Microbiology, University of Malaysia, Kuala Lumpur, Malaysia. [11]INSERM U1053, Bordeaux, France. Correspondence and requests for materials should be addressed to F.F.V. (email: vale.filipa@gmail.com or f.vale@ff.ul.pt) or P.L. (email: philippe.lehours@u-bordeaux.fr)

There are about 10^31 phages on the planet, with phages exceeding bacteria in number by tenfold, but less than an estimated 1% have been described[13]. Temperate phages contribute to the evolution of most bacteria, by promoting the transduction of various genes involved in virulence, fitness, and antibiotic resistance[14]. Despite the putative bacterium–phage evolutionary conflict, phages profit from promoting the survival and proliferation of their hosts[15]. Likewise, prophages may harbor cargo genes, or "morons", which while are not essential for the phage, benefits the host. Some very well known lysogenic phages carry genes that enhance the virulence of the bacterial host[16]. In addition, the deletion of prophages from *E. coli* revealed that prophages improved the surviving under adverse environmental conditions, including acid stress or early biofilm formation[17]. Prophages may therefore work as gene reservoirs, many of which benefit pathogens, in ways which are only just beginning to be determined[18]. In a hostile environment like the human stomach, any metabolic advantage or resistance/tolerance mechanism provided by prophages should be important in improving bacterial host competitiveness. Prophage induction may also be used as a weapon for colonizing new niches[19], displacing native strains, although this strategy may be rarely used, first by the creation of lysogens in the susceptible population, second by the cost of cell lysis in a fraction of the population, and third due to the purifying selection of prophages[20]. Taken together, these properties may explain why prophages are more frequent in pathogenic bacteria[21]. Host-prophage driven selection and genetic flux occurs even for prophage genes that do not effect host physiology[20]. Thus, the role of prophages in disease establishment is being progressively acknowledged.

The first descriptions of *H. pylori* phages came from the observation of micrographs where particles compatible with phages are observed[22–26]. The development of the genomic studies, especially using high-throughput genome sequencing led to the first reports of prophages, some remnant[27], others apparently complete and capable of going through a lytic cycle[11,28–32]. Strains carrying prophages do not appear to have a higher pathogenicity or association with particular disease patterns[11,33], but it has been suggested that the presence of phage orthologous genes correlates with the presence of *cagA* and/or *vacA* virulence genes[34]. The population to which prophages belong is determined by prophage sequence typing (PST), which targets two prophage genes (integrase and holin) of *H. pylori* and applies a Bayesian clustering analysis for the identification of distinct genetic populations. Currently there are 4 prophage populations described, hpAfrica1, hpEastAsia, hpNEurope and hpSWEurope[33]. On the other hand, the bacterial population is determined by MultiLocus Sequence Typing (MLST), which is based on the analysis of 7 bacterial housekeeping genes. Presently, there are 7 seven *H. pylori* populations described, hpAfrica1, hpAfrica2, hpNEAfrica, hpSahul, hpAsia2, hpEastAsia and hpEurope[35]. Recently, using the PST method, we determined that *H. pylori* prophage genes, namely integrase and holin genes present a phylogeographic distribution. Furthermore, the European *H. pylori* population (hpEurope), which could not be discriminated using the MLST method, was separated into two different populations (hpNEurope and hpSWEurope) using these two prophage genes[33].

The number of complete phage genomes available in GenBank is low. Despite the recent discovery of the importance of prophages in the diversity of *H. pylori*[11], they remain poorly characterized. The lack of information on bacteriophages of *H. pylori* prompted this study. Based on the presence of the prophage integrase gene we determined that an estimated 20% of *H. pylori* strains carry prophages[11,33]. Based on PCR screening, we compiled a collection of *H. pylori* strains carrying prophages[33]. We therefore undertook a more holistic approach, using the next generation sequencing (NGS) method to study the full genome of strains (Whole-genome sequencing) from this collection as well *H. pylori* strains presenting prophages found in public databases. This information allowed us to identify phage sequences, which were then used for comparative genomics. Our results have increased our knowledge on *H. pylori* prophage genomic organization into syntenic blocks, insertion sites, phylogeography, and diversity. The detailed genomic structure of 28 prophages reported here will provide in the future an important basis to identify the function of prophage genes and to verify if prophages provide advantageous phenotypes.

Results

A summary of *H. pylori* sequenced genomes can be found in Table S1 (Supplementary Information).

Prophage genome characteristics. We were able to close the physical gaps between contigs in over 90% of the prophage genomes using PCR and Sanger sequencing. In most cases the prophage contigs were separated at insertion sequences, repetition zones and/or sequences showing homology with other bacterial genes. A prophage was considered intact if the size was larger than 20 Kb. According to this criterion, prophages were found to be intact in 23 of the 28 genomes (82%) (Table 1). The other five genomes showed remnant prophages (Table S2, Supplementary Information) between 11.6 Kb and 19.8 Kb. Intact prophages were initially divided in several contigs (min 1- max 7) and have an average of 34 predicted genes (min 24, max 39), 28.7 Kb (min 22.6, max 33.0), and 36.7% GC, which is in line with other *H. pylori* prophages described[11,28]. The bacterial average GC percentage was 39.0%, suggesting horizontal gene transfer of the prophage region.

The gene content of intact prophages was similar to phage KHP30, a known complete phage with lytic cycle[30]. The intact prophage genomes had a rather similar sequence (Figures 1 and S1, Supplementary Information) with a reasonably conserved gene order (Tables S3 and S4, Supplementary Information) and in clear contrast with the host *H. pylori*, where the occurrence of genome rearrangement is well known[36]. Genome annotation of prophage genes produced with either RAST[37] or PHAST[38] revealed that most of the open reading frames (ORF) corresponded to hypothetical proteins, disclosing the diversity of prophage genes and the consequent difficulty in the annotation process. The annotation with Phages 1.0 (http://www.phantome.org/PhageSeed/Phage.cgi?page=phast) did not add more information and was not further considered.

The similarity of prophage genomes was also quantified as a heat-map (Figure S2, Supplementary Information). This similarity matrix confirmed the percentages of bases which were identical. Only one prophage genome, strain Pt-4481-G, harbored a rearrangement (Figure S3, Supplementary Information), where the first segment of

Strain	Population		GC%		Insertion Site		Prophage				Accession number
	Phage - PST	MLST	bacteria	prophage	5′	3′	CDS* PHAST	CDS* PHAGES	CDS* RAST	Kb	
UK-EN31-U	hpNEurope	hpEurope	39.0	36.7	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	36	34	36	30.5	KX119174
UK-EN32-U	hpNEurope	hpEurope	38.9	36.7	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	36	34	35	29.9	KX119206
De-M53-M	hpNEurope	hpEurope	38.8	36.2	S-adenosylmethionine synthetase (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	33	32	33	28.1	KX119205
Sw-577-G	hpNEurope	hpEurope	38.9	36.3	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	30	29	32	26.9	KX119204
Sw-A626-G	hpNEurope	hpEurope	38.8	36.6	ND	ND	37	32	37	31.0	KX119177
Pt-B89-G	hpAfrica1	hpEurope	39.0	37.4	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	32	33	32	27.4	KX119203
Pt-1293-U	hpAfrica1	hpEurope	39.0	36.8	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	36	37	36	30.1	KX119202
Fr-ANT170-U	hpAfrica1	hpEurope	39.0	37.2	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	37	33	36	31.2	KX119201
Fr-MEG235-U	hpAfrica1	hpEurope	39.1	37.3	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	37	33	36	31.2	KX119200
Pt-5771-G	hpAfrica1	hpEurope	39.0	36.9	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	34	34	34	29.8	KX119199
Pt-5322-G	hpAfrica1	hpEurope	39.1	36.8	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	31	31	31	28.3	KX119198
Pt-228_99-G	hpAfrica1	hpEurope	39.0	37.2	S-adenosylmethionine synthetase (EC 2.5.1.6) (jhp_0183)	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (EC 2.3.1.191) (jhp_0182)	37	36	38	30.1	KX119175
Pt-1846-U	hpAfrica1	hpEurope	39.0	37.0	GTP cyclohydrolase II/3,4-dihydroxy-2-butanone 4- phosphate synthase (jhp_0740)	ND	32	31	32	28.0	KX119176
Pt-B92-G	hpAfrica1	hpEurope	38.8	36.9	Membrane-associated phospholipid phosphatase (jhp_0787)	ND	39	36	38	30.5	KX119197
Pt-4481-G	hpAfrica1	hpEurope	39.0	36.8	ND	Ribosomal large subunit pseudouridine synthase B (EC 4.2.1.70) (jhp_1353)	32	31	32	25.4	KX119196
Fr-GC43-A	HpEastAsia	hpEurope	39.0	36.3	Competence protein ComGF (jhp_0650)	putative outer membrane protein HomA (jhp_0649)	38	37	39	33.0	KX119195
Fr-G12-G	hpEastAsia	hpEurope	38.9	36.3	Competence protein ComGF (jhp_0650)	putative outer membrane protein (jhp_0649)	36	35	36	28.6	KX119194
Fr-B58-M	hpEastAsia	hpEastAsia	38.8	36.0	Competence protein ComGF (jhp_0650)	putative outer membrane protein (jhp_0649)	26	24	26	22.6	KX119193

Continued

Strain	Population		GC%		Insertion Site		Prophage				Accession number
	Phage - PST	MLST	bacteria	prophage	5′	3′	CDS* PHAST	CDS* PHAGES	CDS* RAST	Kb	
Pt-212-99R-U	hpAfrica1	hpEurope	38.9	37.1	Competence protein ComGF (jhp_0650)	putative outer membrane protein (jhp_0649)	24	24	24	23.0	KX119189
Pt-1918-U	hpSWEurope	hspWAfrica	39.1	36.2	Hypothetical protein (jhp_1347)	Putative outer membrane protein (jhp_1346)	34	33	34	28.7	KX119192
Pt-4497-U	hpSWEurope	hspWAfrica	39.3	36.2	hypothetical protein (jhp_0949)	Putative protein (jhp_0950)	35	34	36	29.4	KX119191
Pt-4472-G	hpSWEurope	hpEurope	38.8	36.6	hypothetical protein (jhp_0191)	hypothetical protein (jhp_0193)	32	30	32	27.6	KX119190
Fr-B41-M	hpSWEurope	hpWAfrica	39.1	35.5	Acetyl-coenzyme A carboxyl transferase alpha chain (EC 6.4.1.2) (jph_0504)	hypothetical protein (jhp_0503)	35	35	36	29.4	KX119188

Table 1. Intact prophage genomes identified after whole genome sequencing. *Number of coding sequences (CDS) detected according to web service used; GC: guanine-cytosine; PST: prophage sequence typing; MLST: multilocus sequence typing.

approximately 10.4 Kb appeared to be inverted. The second segment of about 15 Kb had the same gene order as all of the other prophages.

Regarding remnant prophages (Table S4, Supplementary Information), different scenarios were observed: (i) one phage (Sw-C388-G) has lost the putative DNA primase and helicase, among other proteins of unknown function placed in the first half of the prophage genome; (ii) two phages (Sw-C520-G and Pt-259-G) most likely lost the second half of the prophage sequence; and (iii) another two phages (Is-3180-G and Pt-5303-G) most likely lost specific ORFs. Among the later, only Is-3180-G could be assembled, yielding less than 20 kb.

Insertion Sequences. Insertion sequences (IS), comprised of two ORFs inserted into prophage genomes were found in 39.1% (9/23) of complete prophages (Table S5, Supplementary Information) classified (according to PST typing) as hpNEurope (n = 2), hpAfrica1 (n = 5) and hpEastAsia (n = 2), and in 50% (3/6) of remnant prophages classified as hpNEurope (n = 1), hpAfrica1 (n = 1) and hpSWEurope (n = 1). The complete prophages Uk-EN31-U, Uk-EN32-U, Pt-B92-G and Fr-GC43-A had IS605 inserted once in the first three cases and twice in the last case. Interestingly, in Fr-GC43-A one copy of IS605 was inverted in relation to the other copy (Figure S4, Supplementary Information). IS605 was inverted in Uk-EN31-U, Uk-EN32-U and in one of the IS of Fr-GC43-A. The prophages Pt-228_99-G, Fr-ANT170-U and Fr-MEG235-U had two copies of ISHp608. The IS was inserted in a reverse order in relation to the other copy in Fr-ANT170-U, and twice with the same orientation in Pt-228_99-G. The third IS found was IS607 in genomes Pt-1293-U and Fr-B58-M.

Concerning remnant prophages, Sw-C388-G has the IS606 inserted at its 3′ end and the second ORF is again truncated in two. Finally, Is-3180-G carries ISHp608. The remnant prophage Pt-5303-G could not be completely assembly but ISHp608 was also found in a separate contig. Despite all of our efforts, we were not able to determine if this IS was inserted into the prophage genome or not.

IS were not always found at the same position in the prophage genomes, but prophages from strains of the same country of origin tended to present the same IS at same genome context (Table S5, Supplementary Information). Nevertheless, IS were present in most cases (9/13, 69%) immediately before DNA helicase (2/9), either before or after DNA primase (4/9), after structural protein (2/9), or after holin gene (1/9), which therefore could be considered as hotspots for IS in prophages.

The transposase genes from IS605 were inserted near the lysis cassette, as described for Mu-like phages[39], DNA helicase and DNA primase. IS607 was located adjacent to DNA primase or a structural protein and ISHp608 near DNA primase, portal protein or structural protein. In a few cases IS were inserted into a coding sequence of a structural protein (Pt-1293-U and Pt-228_99-G) or a hypothetical protein (Pt-B92-G, Fr-ANT170-U and Fr-MEG235-U), which may impinge on transcription, and the prophage genes may be non-functional. Accordingly, IS do not appear to be randomly inserted into prophage genome. Our hypothesis is that the presence of IS within the prophage genome may inactivate the lytic cycle, benefiting the host.

Prophage insertion site. Knowledge of the insertion site of prophages provides clues about ancient acquisition and vertical heritage. Accordingly, prophages at similar loci in different genomes can derive from a single ancestral prophage[20]. Furthermore, *H. pylori* prophage insertion sites have not been extensively studied before.

Prophage insertion site was mostly conserved among *H. pylori* PST populations. Interestingly, about 50% of the prophages enrolled in the present study and especially for the populations hpAfrica1 and hpNEurope are inserted between the same two genes, S-adenosylmethionine synthetase (synthesizes S-adenosylmethionine (AdoMet)), and UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (metabolic pathway of lipid A). These two genes are usually contiguous in the *H. pylori* genome. Prophages classified as belonging to the hpEastAsia population, although represented in a very small number, appear to be inserted between genes coding for a competence protein ComGF and a putative outer membrane protein. Phages from hpSWEurope appear to be inserted at random locations (Table 1 and S2, Supplementary Information).

Figure 1. Alignment of 29 complete prophages, using Mauve software (version 2.3.1).

The presence of tandem repeats at the 3′ end of the prophage insertion site was often verified for prophages integrated between S-adenosylmethionine synthetase and UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase (Table S6, Supplementary Information).

Prophage phylogenetic relationships. To get insight into the genetic backbone of the identified prophages and to infer their phylogenetic relationships in the frame of the well-known *H. pylori* geographic distribution, all 23 intact genomic sequences (Table 1) as well as the publicly available complete genomes of six *Helicobacter* phages (India7, Cuz20, 1961 P, KHP30, KHP40 and phiHP33) and the outgroup *H. acinonychis* prophage, were selected for increasing genetic diversity and were analyzed. Figure 2 shows the phylogenetic

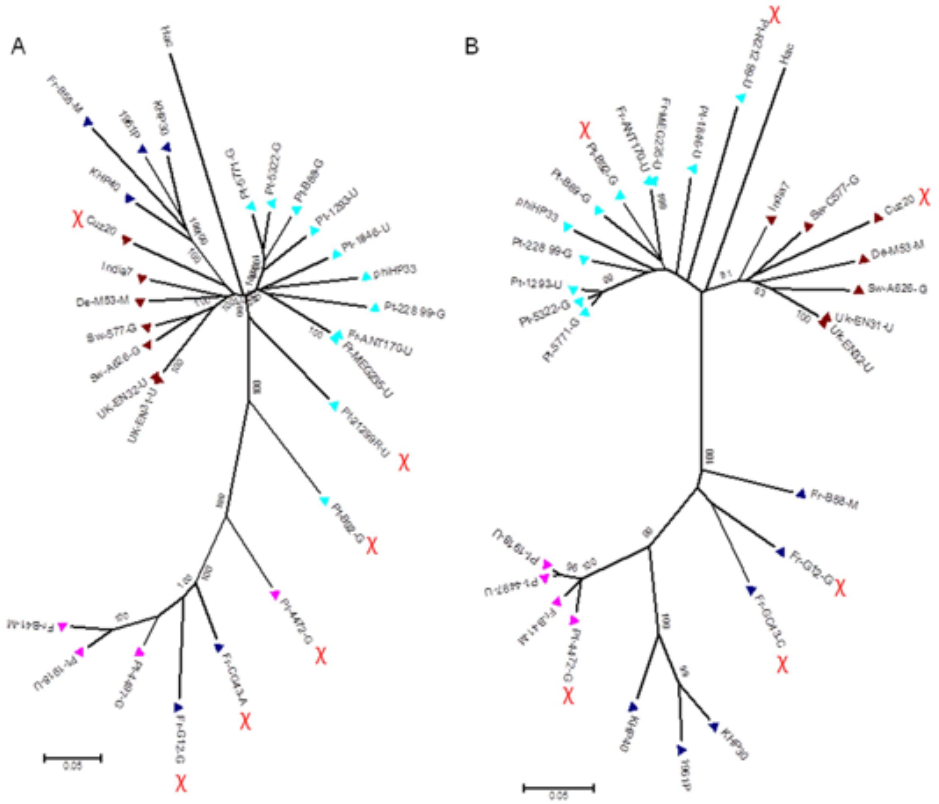


Figure 2. Phylogenetic trees based on (**A**) prophage genomes and (**B**) prophage sequence typing (PST). Neighbour-joining trees, Kimura two-parameter model, complete deletion option and 1000 resampling using MEGA 6.0 software. Phage population: brown triangles: hpNEurope; pink triangles: hpSWEurope; dark-blue triangles: hpEastAsia; light-blue triangles: hpAfrica1. Hac - *Helicobacter acinonychis* prophage. χ - highlights recombinogenic prophage genomes.

inferences found for the complete prophage genome and the concatenated integrase and holin prophage genes (PST). We observed that the majority of the prophages gather by phylogeographic group, clustering accordingly to their population assigned by STRUCTURE[40–42], in a similar fashion to what we described previously for the concatenated integrase and holin genes only[33]. However, evident exceptions were noted for some prophages, namely Pt-4472-G, Fr-G12-G, Fr-CG43-A, Pt-B92-G, Pt-21299R-U and Cuz20, which displayed a discrepant phylogeographic segregation from their PST classification, suggesting the existence of putative recombination events. For instance, Pt-4472-G prophage which, according to STRUCTURE analysis, belongs to hpSWEurope, appears to be a genomic mosaic composed of both hpSWEurope and hpAfrica1 populations. This is clearly evident in Figure 3A, where Pt-4472-G is >90% similar to the latter in the genome central region, whereas the similarity to the hpSWEurope population reached values <50%. Curiously, the regions where the opposite is observed (i.e., >90% similarity to hpSWEurope) encompass both the integrase and holin genes that are used for PST classification. Another clear example of prophage recombination is exhibited by Pt-B92-G, which was PST-classified as hpAfrica1. Although most of its genome appears to be inherited from a hpAfrica1 or hpNEurope population, it displays a small middle region where similarities to the hpSWEurope population reached >95% while is strikingly different from the remainder (Figure 3B). Although less evident, we would also like to highlight two other interesting cases involving mosaicism between hpSWEurope and hpAfrica1 populations, namely Fr-G12-G and Pt-21299R-U. Despite the fact that the former was PST-classified as hpEastAsia, most of its genome was clearly inherited from a hpSWEurope population with the exception of a small 3′-end region which is highly similar to an hpAfrica1 population (data not shown). To the contrary, most of the Pt-21299R-U genome is similar to hpAfrica1, except for its 3′-end which is highly similar (>95%) to an hpSWEurope population (similarity to hpAfrica1 is as low as 40%). Interestingly, the holin gene is absent in this prophage and, in the integrase-involved region, both hpAfrica1 and hpSWEurope populations are almost equally represented (data not shown). Considering the huge genomic diversity observed among all prophage genomes, a precise identification of the location of the breakpoint regions for all of the described recombination events was not possible.

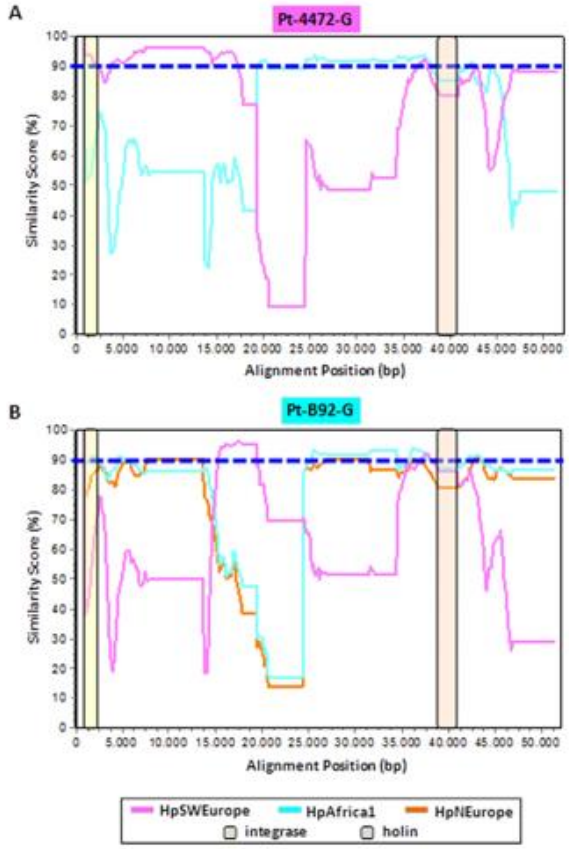


Figure 3. Genomic mosaicism of Pt-44772-G and Pt-B92-G prophages. (**A**) SimPlot showing the genetic similarity of PT-4472-G (PST-classified as hpSWEurope) to both the hpSWEurope and hpAfrica1 populations. (**B**) SimPlot showing the genetic similarity of Pt-B92-G (PST-classified as hpAfrica1) to hpSWEurope, hpAfrica1 and hpNEurope populations. For both plots, the Kimura 2-paramter model was used to calculate nucleotide similarities in a sliding-window of 1500 bp and a step size of 150 bp, with gap strip on. Cut-off of 90% similarity is shown in a blue dashed-line.

Discussion

Most phages identified in the present study, showed a remarkable genetic synteny among themselves (Figure 1, Table S1, Supplementary Information). However, in comparison with phage KHP30, the synteny was punctuated by deletions of certain genes which were replaced by additional IS throughout the prophage genome. When prophages are present, the tendency in *H. pylori* is to have just one prophage per genome, which is in accordance with the small genome size of *H. pylori*, which is expected to have less neutral targets for prophage integration. Furthermore, *H. pylori* has slow bacterial growth, and a population at low density provides few resources for the production of virions, favoring lysogeny[21].

Prophage ORFs were typically found in the same direction, which was opposite to that of the bacterial flanking genes. Concerning annotation most ORFs have an unknown function, as described for other species phages[43]. Although no known virulence gene was found in prophage genomes, the role of prophages in the virulence of *H. pylori* should not be immediately discarded. Frequently phages do not code for toxin genes, as they are not able of directly convert their host into a toxin producer[44], but they can, however, indirectly modulate toxin production, such as TcdA and TcdB in *Clostridium difficille*[45].

Considering the bacterium's ecological niche, *H. pylori*'s persistence might be associated with both its broad genetic variability[46] and its capability of biofilm developing[47,48]. In both cases the presence of extracellular DNA (eDNA) is important, either as a source of DNA taken up by the naturally competent *H. pylori*, promoting recombination or contributing to biofilm development[48]. Apart from outer membrane vesicle shedding, cell lysis via spontaneous prophage induction might be a source of eDNA release, contributing to survival and to the wide genomic variability of *H. pylori*.

The IS found in the present study were previously described in *H. pylori* but outside a prophage context[49]. IS were described to be present in about one-third of a set of 238 independent isolates of *H. pylori*[50]. Bacterial IS of IS200/IS605 and IS607 family often encode a transposase (TnpA) and a protein of unknown function, TnpB,

which were hypothesized to act as a methyltransferase[51]; furthermore, *orfB* is also related to the *Salmonella* virulence gene *gipA*, a *Salmonella* prophage gene which enhances bacterial growth in Peyer's patches[52].

As IS found within prophage sequences showed robust homology with those found in the *H. pylori* genome, it can be hypothesized that prophages mediate the transfer of IS, further contributing to the genome plasticity of *H. pylori*. In contrast, we cannot exclude that the transfer of IS otherwise from bacteria to prophages may also be feasible. Remarkably, IS have been described in other prophages, including cyanophage Ma-LMM01, specifically-infecting *Microcystis aeruginosa* and mediating the transfer of IS607 to the bacterial genome[53]. Besides prophages, IS605 is also associated with the *cag* pathogenicity island, dividing this island into two parts called *cagI* and *cagII* by insertion of one or two copies of IS605, providing intermediate phenotypes[54]. Prophage inactivation should be under selection because lytic cycle induction may kill the cell. Correspondingly, we find five remnant prophages that might result from these evolutionary dynamics, even though defective prophages can still provide an adaptive function to bacteria[20]. Recombination with incoming phages can also imprint a signal for purifying selection. In addition, IS present in prophages have been postulated to play a role in the inactivation and immobilization of incoming phages[55].

We showed that the prophage insertion sites can be diverse in *H. pylori* genomes although with some common traits among *H. pylori* populations, as discussed below. All prophages from hpNEurope from the present study and from *H. pylori* Cuz20 and India7 genomes (available at the NCBI), as well as most prophages from hpAfrica1 populations, have the same genomic context, presenting the bacterial genes S-adenosylmethionine synthetase and UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase at the 5′ end and 3′ end, respectively. Interestingly, the prophages genomes integrated between these two loci usually present tandem repeats at the 3′ end, between the last prophage gene and the first bacterial gene after the prophage (Table S6, Supplementary Information), most often in noncoding regions. DNA tandem repeats or satellite DNA, are inter- or intragenic nucleotide sequences repeated two or more times in a head-to-tail manner. Because these repeat tracts are prone to strand-slippage replication and recombination events causing their copy number to increase or decrease, *loci* containing tandem repeats are hypermutable[56]. Tandem repeats may reversibly shut down or modulate the function of specific genes, allowing them to adapt to changing environments on short evolutionary time scales without an increased overall mutation rate. The environmental adaptability in *H. pylori* depends primarily on tandem repeat variations, which may cause gene phase switching. DNA tandem repeats may modulate gene expression affecting transcription initiation by modifying binding affinity of regulatory proteins (upstream of −35 site) or altering the distance to promoter elements (between −35 and −10 sites), modifying the affinity of regulatory proteins or mRNA stability (between the transcriptional start and an ORF). The most frequent bacterial gene at the 5′ end of prophage codes for S-adenosylmethionine synthetase, which catalyzes the synthesis of AdoMet. AdoMet is an essential metabolic intermediate involved in many biochemical processes, such as a donor of methyl groups that allows DNA methylation (reviewed in ref. 57). Once DNA is methylated it may switch genes[6].

All hpEastAsian prophages either described in the present study or found in the genomes of *H. pylori* YN4–84, UM038, FD430 and UM114 Asian strains (available at the NCBI) were inserted in the same genomic region, including the competence protein ComGF, which plays a role in transformation and DNA binding, at the 5′ end and a putative outer membrane protein at the 3′ end. The gene at the 5′ end is important for genetic variability of *H. pylori*[58], while *H. pylori* outer membrane proteins are known to mediate adherence to gastric epithelium, and ultimately are associated with clinical outcome of the infection[59]. All things considered, the prophage insertion site may not be neutral for *H. pylori* gene expression and further studies are needed to evaluate the impact of prophage insertion on gene expression.

In general, the phylogenetic analysis of intact prophages presents clusters according to prophage population structure (exceptions are discussed below), confirming our previous results obtained by prophage sequence typing[33]. The prophage genomes cluster in four groups corresponding to the hpSWEurope, hpNEurope, hpAfrica1 and hpEastAsia phage populations. The strong phylogeographic signal of prophage genomes is in agreement with a model of co-evolution between the virus and its bacterial host. Indeed, prophages and bacteria are linked by a long history of co-evolution, but the genetic dimension of this co-evolution cannot be defined at present[14]. The phylogeographic clustering was in agreement with integration sites of prophages (discussed above). As suggested by others[60], this could be explained by a vertical transmission of the phage rather than by random insertions which are common to prophages.

Phage evolution is driven by a horizontal exchange of functional modules between more or less related phages, achieved by DNA recombination, explaining the genomic mosaicism among phages[61]. Recombination is a factor of rapid variability in *H. pylori*, which is among the most recombinogenic known pathogens[12]. In parallel, in the present study, phage genomes were shown to be prone to recombination events. Indeed several prophage genome mosaics were detected, involving, for the vast majority of the cases, both hpAfrica1 and hpSWEurope populations. This is not surprising considering that both populations were detected in the same geographic area. Nevertheless, most phage ORFs are of unknown function, so no assumptions can be performed regarding a putative impact of these recombination events on pathogenicity. These mosaic structures also highlight the need for a prudent use of the PST-based classification. In fact, although an agreement is observed for most of the cases, for the studied mosaic structures, for some of the studied mosaic structures only the integrase and/or the holin genes appeared to support the PST-based classification.

The remnant prophages encountered in the present study as well as in other *H. pylori* strains[32,62] and in non-*pylori Helicobacter* species[63] highlight an evolutionary scenario consistent with a prophage decay process during the complex interaction between *H. pylori* and the prophage. However, a model in which *H. pylori* strains from different geographical regions may have been infected by distinct phage lineages after the geographic separation of the bacterial host is also feasible[11], but less likely due to the high genetic synteny between prophages from different geographic areas. Altogether, the integration at the same locus and a gene repertoire relatedness points to

a vertical transmission, suggesting the so called pervasive domestication of prophages by the bacterial host which may drive bacterial adaptation[20]. Remarkably, the most divergent *H. pylori* prophage population (hpSWEurope), presented neither conserved loci for integration site nor IS.

This work not only provides a compendium of novel sequences, but also sets the stage for future studies aimed at better understanding the virus-host relationship. Results of the present study showed that prophages are more common in *H. pylori* than initially expected and that, in most cases, prophages appear to be intact, with a sequence size of over 20 Kb. Remarkably, we show for the first time that for phages classified as hpNEurope, hpAfrica and hpEastAsia, the insertion site appears to be preserved (Table 1). Furthermore, the phylogenetic analysis for a vast majority of phage genomes is similar to the phylogenetic analysis previously presented by our team[33] using two phage genes (integrase and holin), confirming our previous findings and reinforcing the hypothesis of co-evolution between prophages and *H. pylori*. Some recombinant phages were found, suggesting additional genetic diversity that hypothetically may provide *H. pylori* with advantageous phenotypes. Major challenges at present are to identify the function of prophage genes, to understand if the insertion site is neutral for the host and whether prophage presence plays a role in the adaptation of *H. pylori* to its host, or if prophage genes belonging to the lysis cassette are useful for biomedical applications, namely phage therapy.

Material and Methods

Bacteria and cell growth conditions. A total of 28 *H. pylori* strains carrying prophages were analyzed (Table S1, Supplementary Information). These included 15 strains isolated from patients with gastritis, nine from peptic ulcer patients, three from MALT patients and one from gastric cancer patient. The present study included strains from Portugal (n = 14), France (n = 6), Sweden (n = 4), UK (n = 2), Germany (n = 1) and Israel (n = 1). Prior to each assay, bacteria were grown in *H. pylori* selective medium (Biogerm, Portugal) at 37 °C in a microaerophilic environment (Anoxomat®, MART Microbiology BV, The Netherlands) for 24 h to 48 h. The *H. pylori* strains belong to the collection of the French National Reference Centre for Campylobacters and Helicobacters (F. Mégraud and P. Lehours, Bordeaux, France); the Department of Microbiology, Tumor and Cell Biology, Karolinska Institute (Lars Engstrand); the Klinikum Rechts Der Isar II, Medical Department, Technische Universität, Munich, Germany (M. Gerhard); the Department of Infectious Diseases, National Institute of Health, Lisbon, Portugal (M. Oleastro); and the Rabin Medical Center – Beilinson Hospital, Petah Tikva, Israel (T.T. Perets and Y. Niv).

Whole-Genome Sequencing. Genomes were sequenced at the National Institute of Health, Lisbon, Portugal, with exception of four strains (Sw-577-G, Sw-A626-G, Sw-C388-G and Sw-C520-G) that were sequenced at Karolinska Institute, Stockholm, Sweden, and four strains (Fr-ANT170-U, Fr-MEG235-U, Fr-GC43-A and Fr-B41-M) that were sequenced at the Institute of Life Sciences, College of Medicine, Swansea, Wales, UK.

Total DNA was extracted using the QIAmp DNA Mini Kit (Qiagen, UK) according to the manufacturer's instructions.

For genomes sequenced in Portugal and Sweden, the yield and integrity of the purified DNA were then assessed through a Qubit assay (Quanti-it dsDNA Assay Kit, Broad Range; Lifetechnologies, Paisley, CA, USA) and agarose gel electrophoresis (0.7% gel), respectively. High-quality DNA samples were then applied to prepare Nextera XT Illumina paired-end libraries. These were subsequently subjected to cluster generation and paired-end sequencing (2 × 250 bp, 2 × 150 bp and 2 × 100 bp) by using the Illumina MiSeq (Portugal) and HiSeq 2500 (Sweden) platforms (Illumina Inc., San Diego, CA, USA), according to the manufacturer's instructions.

The number of passing filter reads obtained per sample ranged from 0.6–2.7 million reads. The FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and FASTX (http://hannonlab.cshl.edu/fastx_toolkit/) tools were applied to evaluate and improve the quality of the raw sequence data, respectively. Subsequently, high-quality reads were *de novo* assembled using Velvet (version 1.2.10)[64] (several assemblies using different k-mer sizes were run), where the best assembly was assumed as the one with the best cumulative ranks for N50, number of contigs/scaffolds, and length of the largest contig/scaffold. The obtained mean depth of coverage ranged from 135- to 195-fold. The final contigs/scaffolds were visually inspected (using Tablet 1.14.04.10)[65] and corrected.

For genomes sequenced in UK, quantification of DNA was assessed after DNA extraction with a Nanodrop spectrophotometer, as well as the Quant-iT DNA Assay Kit (Life Technologies) prior to sequencing. High-throughput genome sequencing was performed using a HiSeq 2500 machine (Illumina Inc.), and the 100 bp short read paired-end data was *de novo* assembled using Velvet (version 1.2.08)[64]. The VelvetOptimiser script (version 2.2.4) was run for all odd k-mer values from 21 to 99 (several assemblies using different k-mer sizes were run), with all program settings unchanged apart from a minimum output contig size set to 200 bp and the scaffolding option switched off.

All genomes were annotated using the RAST server (http://rast.nmpdr.org/)[37], the NCBI Prokaryotic Genomes Annotation Pipeline version 2.3. and PHAST web server[38]. The respective trimmed reads were submitted to the Sequence Read Archive (SRA).

Assembly of prophage genomes. For prophage identification two strategies were taken. First, the PHAST web server[38] was used to identify putative prophages within contigs of each *H. pylori* genome. Second, MEGABLAST[66] was used to align the genome of *H. pylori* phage KHP30 or phiHP33 with the contigs of each sequenced *H. pylori* genome. PHAST analyses (http://phast.wishartlab.com/) applied over contigs allowed us to check homology, and to identify, annotate and graphically display prophage sequences, providing information on prophage completeness, categorized as either intact, incomplete, questionable or not detected. MEGABLAST was

run using KHP30 or phiHP33 as reference since these prophages genomes were the most commonly found to be similar with the prophages detected by PHAST.

The MEGABLAST analysis results were particularly useful to determine which contigs were from phage origin and the order in which they probably appear. Based on this predicted contig order, primers flanking the contigs were designed, using primer3 v. 0.4.0[67], to bridge gaps in the assembly in order to close the gaps (the gaps were of few bases to about five hundred bases). The PCR mix included Promega (Madison, WI, USA) buffer (1X), dNTPs (0.2 μM), primers (0.5 μM each), GoTaq polymerase (1.5 U), water to complete 25 μl and DNA sample (25 to 50 ng). The PCR cycle was composed of a first cycle at 95 °C for 4 min, 35 cycles at 95 °C for 30 sec, 59 °C for 30 sec and 72 °C for 1 or 2 min. A last cycle at 72 °C for 7 min was applied. The PCR products were purified using MicroSpin S-400 or S-300HR columns (GE Healthcare, Velizy-Villacoublay, France) and directly sequenced on both strands using an external sequencing service provider (Eurofins Genomics, Regensburg, Germany, and Stabvida, Lisbon, Portugal). A multiple sequence alignment[68] was carried out using flanking parts of the contigs and the PCR sequenced product after assembly of the forward and reverse sequences.

The insertion sequences of the prophages were identified whenever the prophage 5′ and 3′ ends were contiguously flanked by bacterial genes in a contig. The last bacterial gene before the prophage sequence and the first bacterial gene after the prophage were identified as well as the homologous locus_tag for the reference genome *H. pylori* J99[36]. The presence of repeated sequences at prophage insertion sites was verified using Tandem Repeat Finder[69] (available at https://tandem.bu.edu/trf/trf.basic.submit.html).

Comparative genomic analyses of prophages. The assembled prophages were analyzed using PHAST to provide a first annotation. The annotation of prophage genomes was carried out further using Phages v. 1.0 (http://www.phantome.org/PhageSeed/Phage.cgi?page=phast), and RAST[37]. The annotation of coding sequences (CDS) found by the three different methods were compared.

The annotation of both *H. pylori* India7 (accession number CP002331) and Cuz20 (CP002076) prophages, as well as that of the *Helicobacter* 1961P (NC_019512.1), KHP30 (NC_019928.1), KHP40 (NC_019931.1), phiHP33 (NC_016568.1) phages, were used for comparative purposes.

The annotated prophages were aligned using the progressive Mauve algorithm software (version 2.3.1)[70], to check the order of the CDS in the prophage genomes and the existence of a consensus sequence. In order to infer phylogenetic relationships among prophages, the intact genomes of the 23 prophages identified in the present study, were aligned using MAFFT version 7[71] together with other six phage *Helicobacter* genomes available at public databases (1961P, KHP30, KHP40, phiHP33, *H. pylori* India7, and *H. pylori* Cuz20) as well as with the *H. acinonychis* (accession number NC_008229.1) prophage used as an outgroup. A nucleotide Neighbour-joining phylogenomic tree was constructed using the MEGA (Molecular Evolutionary Genetics Analysis) 6.0 software[72], with distances estimated using the Kimura two-parameter model[73]. Considering the huge genomic diversity observed among all prophage genomes as well as their different lengths, both complete and pairwise deletion options were used. While the former removes all sites containing missing data or alignment gaps before the distance estimations begin, in the pairwise-deletion, option sites are only removed during the analysis as the need arises. Branching significance was estimated using bootstrap confidence levels by randomly resampling the data 1,000 times with the referred evolutionary distance model.

To determine the population structure of prophages, we use prophage sequence typing (PST), as previously described[33]. Briefly, the multi-fasta file with the alignment of integrase and holin gene sequences was converted to the STRUCTURE 2.3.4[40–42] program input file using xmfa2structure by X. Didelot and D. Falush (http://www.xavierdidelot.xtreemhost.com/clonalframe.htm). STRUCTURE was used to study the number of K populations using the admixture, performing runs in duplicate. In each run, a Markov Chain Monte Carlo (MCMC) of 10,000 iterations and a burn-in period of 10,000 iterations were chosen. The highest mean value of ln likelihood was compared for multiple runs of $2 \leq K \leq 6$.

The existence of putative recombination phenomena within prophage genomes was first evaluated using the Recombination Detection Program version 4 (RDP4)[74] with default settings. RDP4 simultaneously applies different methods for detecting and characterizing individual recombination events that are evident within a sequence alignment without any need for predefined sets of non-recombinant reference sequences. SimPlot software (http://sray.med.som.jhmi.edu/SCRoftware/simplot/) was also used for characterizing with higher detail the genomic mosaicism of the identified recombinant prophages, as previously described for bacterial pathogens[75]. The similarity estimations were performed by using the Kimura two-parameter model with sliding window and step sizes that varied according to each recombinant genome.

Data Availability. The genomes of the prophages are available with the accession numbers KX119174 to KX119206. The trimmed reads were submitted to the Sequence Read Archive (SRA), with the accession numbers SRP064706 to SRP064710, SRP071062, SRP071067, SRP071271, SRP071274, SRP071276 to SRP071280, SRP071282, SRP071284, SRP071289 to SRP071296, and SRP072438 to SRP072441.

References

1. Falush, D. *et al.* Traces of human migrations in *Helicobacter pylori* populations. *Science* **299,** 1582–1585 (2003).
2. Bjorkholm, B. *et al.* Mutation frequency and biological cost of antibiotic resistance in *Helicobacter pylori*. *Proc. Natl. Acad. Sci. USA* **98,** 14607–14612 (2001).
3. Yahara, K. *et al.* Genome-wide survey of mutual homologous recombination in a highly sexual bacterial species. *Genome Biol Evol* **4,** 628–640 (2012).
4. Covacci, A. & Rappuoli, R. *Helicobacter pylori*: molecular evolution of a bacterial quasi-species. *Curr. Opin. Microbiol.* **1,** 96–102 (1998).
5. Furuta, Y. *et al.* Methylome diversification through changes in DNA methyltransferase sequence specificity. *PLoS. Genet.* **10,** e1004272 (2014).

6. Vitoriano, I., Vitor, J. M., Oleastro, M., Roxo-Rosa, M. & Vale, F. F. Proteome variability among *Helicobacter pylori* isolates clustered according to genomic methylation. *J. Appl. Microbiol.* **114,** 1817–1832 (2013).
7. Olbermann, P. *et al.* A global overview of the genetic and functional diversity in the *Helicobacter pylori cag* pathogenicity island. *PLoS. Genet.* **6,** e1001069 (2010).
8. Kersulyte, D. *et al.* Transposable element ISHp608 of *Helicobacter pylori*: nonrandom geographic distribution, functional organization, and insertion specificity. *J. Bacteriol.* **184,** 992–1002 (2002).
9. Kobayashi, I. Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. *Nucleic Acids Res.* **29,** 3742–3756 (2001).
10. Vale, F. F., Encarnacao, P. & Vitor, J. M. A new algorithm for cluster analysis of genomic methylation: the *Helicobacter pylori* case. *Bioinformatics* **24,** 383–388 (2008).
11. Lehours, P. *et al.* Genome sequencing reveals a phage in *Helicobacter pylori*. *MBio.* **2** (2011).
12. Go, M. F., Kapur, V., Graham, D. Y. & Musser, J. M. Population genetic analysis of *Helicobacter pylori* by multilocus enzyme electrophoresis: extensive allelic diversity and recombinational population structure. *J Bacteriol.* **178,** 3934–3938 (1996).
13. Brussow, H. & Kutter, E. *Bacteriophages biology and applications*. Kutter, E. & Sulakvelidze, A. (eds), pp. 129–163 (CRC Press, London, 2005).
14. Brussow, H., Canchaya, C. & Hardt, W. D. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* **68,** 560–602 (2004).
15. Feiner, R. *et al.* A new perspective on lysogeny: prophages as active regulatory switches of bacteria. *Nat. Rev. Microbiol* **13,** 641–650 (2015).
16. Golais, F., Holly, J. & Vitkovska, J. Coevolution of bacteria and their viruses. *Folia Microbiol (Praha)* **58,** 177–186 (2013).
17. Wang, X. *et al.* Cryptic prophages help bacteria cope with adverse environments. *Nat. Commun.* **1,** 147 (2010).
18. Wang, X. & Wood, T. K. Cryptic prophages as targets for drug development. *Drug Resist. Updat.* **27,** 30–38 (2016).
19. Gama, J. A. *et al.* Temperate bacterial viruses as double-edged swords in bacterial warfare. *PLoS. ONE.* **8,** e59043 (2013).
20. Bobay, L. M., Touchon, M. & Rocha, E. P. Pervasive domestication of defective prophages by bacteria. *Proc. Natl. Acad. Sci. USA* **111,** 12127–12132 (2014).
21. Touchon, M., Bernheim, A. & Rocha, E. P. Genetic and life-history traits associated with the distribution of prophages in bacteria. *ISME. J* (2016).
22. Marshall, B. J., Armstrong, J. A., Francis, G. J., Nokes, N. T. & Wee, S. H. Antibacterial action of bismuth in relation to *Campylobacter pyloridis* colonization and gastritis. *Digestion* **37** Suppl 2, 16–30 (1987).
23. Goodwin, C. S., Armstrong, J. A. & Peters, M. *Campylobacter pylori in gastritis and peptic ulcer disease*. Blaser, M. J. (ed.), pp. 25–49 (MD.IGAKU-SHOIN, New York, 1989).
24. Vale, F. F., Alves Matos, A. P., Carvalho, P. & Vitor, J. M. *Helicobacter pylori* phage screening. *Microsc. Microanal.* **14** (supp 3), 150–151 (2008).
25. Schmid, E. N., von, R. G. & Ansorg, R. Bacteriophages in *Helicobacter* (*Campylobacter*) *pylori*. *J. Med. Microbiol.* **32,** 101–104 (1990).
26. Heintschel von, H. E., Nalik, H. P. & Schmid, E. N. Characterisation of a *Helicobacter pylori* phage (HP1). *J. Med. Microbiol.* **38,** 245–249 (1993).
27. Thibergue, J. M. *et al.* Sequence of the first *Helicobacter pylori* strains involved in low-grade Mucosa-Associated Lymphoid Tissue (MALT) Lymphoma. *Helicobacter* **11,** 02.01 (2006).
28. Luo, C. H., Chiou, P. Y., Yang, C. Y. & Lin, N. T. Genome, integration and transduction of a novel temperate phage of *Helicobacter pylori*. *J. Virol.* (2012).
29. Uchiyama, J. *et al.* Complete Genome Sequences of Two *Helicobacter pylori* Bacteriophages Isolated from Japanese Patients. *J. Virol.* **86,** 11400–11401 (2012).
30. Uchiyama, J. *et al.* Characterization of *Helicobacter pylori* bacteriophage KHP30. *Appl Environ. Microbiol* **79,** 3176–3184 (2013).
31. You, Y., He, L., Zhang, M. & Zhang, J. Comparative genomics of a *Helicobacter pylori* isolate from a Chinese Yunnan Naxi ethnic aborigine suggests high genetic divergence and phage insertion. *PLoS. ONE.* **10,** e0120659 (2015).
32. Fan, X., Li, Y., He, R., Li, Q. & He, W. Comparative analysis of prophage-like elements in *Helicobacter* sp. genomes. *PeerJ.* **4,** e2012 (2016).
33. Vale, F. F. *et al.* Dormant phages of *Helicobacter pylori* reveal distinct populations in Europe. *Sci. Rep.* **5,** 14333 (2015).
34. Kyrillos, A., Arora, G., Murray, B. & Rosenwald, A. G. The Presence of Phage Orthologous Genes in *Helicobacter pylori* Correlates with the Presence of the Virulence Factors CagA and VacA. *Helicobacter* (2015).
35. Megraud, F., Lehours, P. & Vale, F. F. The history of *Helicobacter pylori*: from phylogeography to paleomicrobiology. *Clin. Microbiol Infect* (2016).
36. Alm, R. A. *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397,** 176–180 (1999).
37. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC. Genomics* **9,** 75 (2008).
38. Zhou, Y., Liang, Y., Lynch, K. H., Dennis, J. J. & Wishart, D. S. PHAST: a fast phage search tool. *Nucleic Acids Res* **39,** W347–W352 (2011).
39. Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brussow, H. Prophage genomics. *Microbiol Mol. Biol Rev.* **67,** 238–76, table (2003).
40. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155,** 945–959 (2000).
41. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164,** 1567–1587 (2003).
42. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* **7,** 574–578 (2007).
43. Pope, W. H. *et al.* Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *Elife.* **4,** e06416 (2015).
44. Fortier, L. C. & Sekulovic, O. Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence.* **4,** 354–365 (2013).
45. Goh, S., Chang, B. J. & Riley, T. V. Effect of phage infection on toxin production by *Clostridium difficile*. *J Med. Microbiol* **54,** 129–135 (2005).
46. Morelli, G. *et al.* Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. *PLoS. Genet.* **6,** e1001036 (2010).
47. Grande, R. *et al. Helicobacter pylori* ATCC 43629/NCTC 11639 Outer Membrane Vesicles (OMVs) from Biofilm and Planktonic Phase Associated with Extracellular DNA (eDNA). *Front Microbiol* **6,** 1369 (2015).
48. Grande, R. *et al.* Extracellular DNA in *Helicobacter pylori* biofilm: a backstairs rumour. *J Appl Microbiol* **110,** 490–498 (2011).
49. Kalia, A. *et al.* Evolutionary dynamics of insertion sequences in *Helicobacter pylori*. *J Bacteriol.* **186,** 7508–7520 (2004).
50. Kersulyte, D., Akopyants, N. S., Clifton, S. W., Roe, B. A. & Berg, D. E. Novel sequence organization and insertion specificity of IS605 and IS606: chimaeric transposable elements of *Helicobacter pylori*. *Gene* **223,** 175–186 (1998).
51. Bao, W. & Jurka, J. Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. *Mob. DNA* **4,** 12 (2013).

52. Kersulyte, D. *et al.* Sequence organization and insertion specificity of the novel chimeric ISHp609 transposable element of *Helicobacter pylori*. *J Bacteriol.* **186,** 7521–7528 (2004).
53. Kuno, S., Yoshida, T., Kamikawa, R., Hosoda, N. & Sako, Y. The distribution of a phage-related insertion sequence element in the cyanobacterium, *Microcystis aeruginosa*. *Microbes. Environ.* **25,** 295–301 (2010).
54. Censini, S. *et al. cag*, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc. Natl. Acad. Sci. USA* **93,** 14648–14653 (1996).
55. Ooka, T. *et al.* Inference of the impact of insertion sequence (IS) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in *Escherichia coli* O157 genomes. *Genome Res* **19,** 1809–1816 (2009).
56. Zhou, K., Aertsen, A. & Michiels, C. W. The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiol Rev.* **38,** 119–141 (2014).
57. Kozbial, P. Z. & Mushegian, A. R. Natural history of S-adenosylmethionine-binding proteins. *BMC. Struct. Biol* **5,** 19 (2005).
58. Baltrus, D. A., Guillemin, K. & Phillips, P. C. Natural transformation increases the rate of adaptation in the human pathogen *Helicobacter pylori*. *Evolution* **62,** 39–49 (2008).
59. Oleastro, M. *et al.* Disease association with two *Helicobacter pylori* duplicate outer membrane protein genes, homB and homA. *Gut Pathog.* **1,** 12 (2009).
60. Bobay, L. M., Rocha, E. P. & Touchon, M. The adaptation of temperate bacteriophages to their host genomes. *Mol. Biol Evol* **30,** 737–751 (2013).
61. Desiere, F., Lucchini, S. & Brussow, H. Evolution of *Streptococcus thermophilus* bacteriophage genomes by modular exchanges followed by point mutations and small deletions and insertions. *Virology* **241,** 345–356 (1998).
62. Thiberge, J. M. *et al.* From array-based hybridization of *Helicobacter pylori* isolates to the complete genome sequence of an isolate associated with MALT lymphoma. *BMC. Genomics* **11,** 368 (2010).
63. Kersulyte, D., Rossi, M. & Berg, D. E. Sequence divergence and conservation in genomes of *Helicobacter cetorum* strains from a dolphin and a whale. *PLoS. ONE.* **8,** e83177 (2013).
64. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18,** 821–829 (2008).
65. Milne, I. *et al.* Using Tablet for visual exploration of second-generation sequencing data. *Brief. Bioinform.* **14,** 193–202 (2013).
66. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
67. Untergasser, A. *et al.* Primer3–new capabilities and interfaces. *Nucleic Acids Res* **40,** e115 (2012).
68. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res* **16,** 10881–10890 (1988).
69. Gelfand, Y., Rodriguez, A. & Benson, G. TRDB–the Tandem Repeats Database. *Nucleic Acids Res* **35,** D80–D87 (2007).
70. Darling, A. E., Mau, B. & Perna, N. T. Progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS. ONE.* **5,** e11147 (2010).
71. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol Evol.* **30,** 772–780 (2013).
72. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol Evol.* **30,** 2725–2729 (2013).
73. Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16,** 111–120 (1980).
74. Martin, D. P., Murrell, B., Golden, M., Khoosal, A. & Muhire, B. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol.* **1,** vev003 (2015).
75. Gomes, J. P. *et al.* Evolution of Chlamydia trachomatis diversity occurs by widespread interstrain recombination involving hotspots. *Genome Res* **17,** 50–60 (2007).

Acknowledgements

This work was supported by the the Fundação para a Ciência e a Tecnologia (FCT) project grant PTDC/EBB-EBI/119860/2010 and by the University of Malaya-Ministry of Education (UM-MoE) High Impact Research (HIR) Grant UM.C/HIR/MOHE/13/5 (h-50001-00-A000033). F.F.V. is recipient of a postdoctoral fellowship from FCT (SFRH/BPD/95125/2013).

Author Contributions

F.F.V., P.L., J.V. and F.M. contributed to the design of the project. M.O., D.A.S., R.R., J.M.B.V., L.E, B.P., E.B., S.S., and M.D.H. carry out whole genome sequencing. A.N. and J.P.G. did the recombination analysis. F.F.V. did the bioinfomatics and genome anlaysis and wrote the manuscript. All authors contributed to the paper and approved the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Vale, F. F. *et al.* Genomic structure and insertion sites of *Helicobacter pylori* prophages from various geographical origins. *Sci. Rep.* **7**, 42471; doi: 10.1038/srep42471 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
1	27935	Alain Burette	Europe	Gastric Cancer	Gastric Cancer				1
2	28861	Alain Burette	Europe	Gastric Cancer	Gastric Cancer				1
3	29009	Alain Burette	Europe	Normal	Asymptomatic				1
4	29373	Alain Burette	Europe	Normal	Asymptomatic				1
5	31181	Alain Burette	Europe	Normal	Asymptomatic				1
6	33375	Alain Burette	Europe	Normal	Asymptomatic				1
7	34320	Alain Burette	Europe	Gastric Cancer	Gastric Cancer				1
8	3735-2	Francis Megraud	Europe	Gastritis	Asymptomatic				1
9	3843-2	Francis Megraud	Europe	Ulcer	Ulcer	1		1	1
637	26695-Tomb	published	Europe	gastritis	Asymptomatic	1		1	
638	Puno135	published	South America	gastritis	Asymptomatic	1		1	
639	Gambia94-24	published	Africa	inflammation, ulcer	Ulcer	1		1	
640	B45	published	Europe	MALT lymphoma	MALT Lymphoma	1		1	
641	52	published	Asia		unknown	1		1	
643	2018	published	Europe	duodenal ulcer	Ulcer			1	
645	35A	published	Asia		unknown	1		1	
646	51	published	Asia	duodenal ulcer	Ulcer	1		1	
647	83	published	Asia		unknown	1		1	
649	Aklavik117	published	North America		unknown	1		1	
650	Aklavik86	published	North America		unknown			1	
651	B38	published	Europe	MALT lymphoma	MALT Lymphoma	1	1	1	
652	B8	published	Unknown	MALT lymphoma	MALT Lymphoma	1			
653	Cuz20	published	South America		unknown	1		1	
654	EL537	published	North America	gastric cancer	Gastric Cancer	1		1	
655	F16	published	Asia	gastritis	Asymptomatic	1		1	
656	F30	published	Asia	duodenal ulcer	Ulcer	1		1	
657	F32	published	Asia	gastric cancer	Gastric Cancer	1		1	
658	F57	published	Asia	gastric cancer	Gastric Cancer	1		1	
659	G27	published	Europe		unknown	1		1	
660	HPAG1	published	Europe	chronic atrophic gastritis	Atrophic gastritis	1		1	
661	HUP-B14	published	Europe		unknown	1		1	
662	India7	published	Asia		unknown	1		1	
663	J99	published	North America	Duodenal Ulcer	Ulcer	1		1	
664	Lithuania75	published	Europe		unknown	1		1	
665	OK113	published	Asia	duodenal ulcer	Ulcer	1		1	
666	OK310	published	Asia	gastric cancer	Gastric Cancer	1		1	
667	P12	published	Europe	duodenal ulcer	Ulcer	1		1	
668	PeCan18	published	South America	gastric cancer	Gastric Cancer			1	
669	PeCan4	published	South America	gastric cancer	Gastric Cancer	1		1	
670	Puno120	published	South America	gastritis	Asymptomatic	1		1	
673	SJM180	published	South America	gastritis	Asymptomatic	1		1	
674	Santal49	published	Asia	asymptomatic	Asymptomatic	1		1	

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
675	Sat464	published	South America		unknown	1		1	
676	Shi112	published	South America	symptomatic	unknown	1		1	
677	Shi169	published	South America	symptomatic	unknown	1		1	
678	Shi417	published	South America	symptomatic	unknown	1		1	
679	Shi470	published	South America	gastritis	Asymptomatic	1		1	
680	SouthAfrica7	published	Africa		unknown	1		1	
681	XZ274	published	Asia	gastric cancer	Gastric Cancer	1		1	
682	v225d	published	South America	superficial gastritis	Asymptomatic	1		1	
683	FD423	published	Asia	nonulcer dyspepsia	unknown	1		1	
684	FD430	published	Asia	nonulcer dyspepsia	unknown	1		1	
685	FD506	published	Asia	nonulcer dyspepsia	unknown	1		1	
686	FD535	published	Asia	nonulcer dyspepsia	unknown	1		1	
687	FD568	published	Asia	nonulcer dyspepsia	unknown	1		1	
688	FD577	published	Asia	nonulcer dyspepsia	unknown	1		1	
689	FD703	published	Asia	nonulcer dyspepsia	unknown	1		1	
690	FD662	published	Asia	nonulcer dyspepsia	unknown			1	
691	FD719	published	Asia	nonulcer dyspepsia	unknown			1	
692	GC26	published	Asia	gastric cancer	Gastric Cancer			1	
693	8A3	published	Asia		unknown	1		1	
694	98-10	published	Asia	Cancer	Gastric Cancer	1		1	
695	A45	published	Asia	peptic ulcer and chronic gastritis	unknown			1	
696	B128	published	North America	gastric ulcer	Ulcer			1	
698	CCHI33	published	North America		unknown	1		1	
699	CPY1124	published	Asia	gastric ulcer	Ulcer	1		1	
700	CPY1313	published	Asia	duodenal ulcer	Ulcer	1		1	
701	CPY1662	published	Asia	duodenal ulcer	Ulcer	1		1	
702	CPY1962	published	Asia	gastric ulcer	Ulcer	1		1	
703	CPY3281	published	Asia	duodenal ulcer	Ulcer	1		1	
704	CPY6081	published	Asia	gastric cancer	Gastric Cancer	1		1	
705	CPY6261	published	Asia	gastric cancer	Gastric Cancer	1		1	
706	CPY6271	published	Asia	gastric cancer	Gastric Cancer	1		1	
707	CPY6311	published	Asia	gastric cancer	Gastric Cancer	1		1	
708	GAM100Ai	published	Africa		unknown	1		1	
709	GAM101Biv	published	Africa		unknown	1		1	
710	GAM103Bi	published	Africa		unknown	1		1	
711	GAM105Ai	published	Africa		unknown	1		1	
712	GAM112Ai	published	Africa		unknown	1		1	
713	GAM114Ai	published	Africa		unknown	1		1	
714	GAM115Ai	published	Africa		unknown	1		1	
715	GAM118Bi	published	Africa		unknown	1		1	
716	GAM119Bi	published	Africa		unknown	1		1	
717	GAM120Ai	published	Africa		unknown	1		1	

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
718	GAM121Aii	published	Africa		unknown	1		1	
719	GAM201Ai	published	Africa		unknown	1		1	
720	GAM210Bi	published	Africa		unknown			1	
721	GAM231Ai	published	Africa		unknown	1		1	
722	GAM239Bi	published	Africa		unknown	1		1	
723	GAM244Ai	published	Africa		unknown	1		1	
724	GAM245Ai	published	Africa		unknown	1		1	
725	GAM246Ai	published	Africa		unknown	1		1	
726	GAM249T	published	Africa		unknown	1		1	
727	GAM250AFi	published	Africa		unknown			1	
731	GAM254Ai	published	Africa		unknown	1		1	
732	GAM260ASi	published	Africa		unknown	1		1	
733	GAM260BSi	published	Africa		unknown	1		1	
734	GAM260Bi	published	Africa		unknown	1		1	
735	GAM263BFi	published	Africa		unknown	1		1	
736	GAM264Ai	published	Africa		unknown	1		1	
737	GAM265BSii	published	Africa		unknown	1		1	
739	GAM270ASi	published	Africa		unknown	1		1	
741	GAM71Ai	published	Africa		unknown	1		1	
742	GAM80Ai	published	Africa		unknown	1		1	
743	GAM83Bi	published	Africa		unknown	1		1	
745	GAM93Bi	published	Africa		unknown			1	
746	GAM96Ai	published	Africa		unknown	1		1	
747	GAMchjs106B	published	Africa		unknown	1		1	
748	GAMchjs114i	published	Africa		unknown	1		1	
749	GAMchjs117Ai	published	Africa		unknown	1		1	
750	GAMchjs124i	published	Africa		unknown	1		1	
751	GAMchjs136i	published	Africa		unknown	1		1	
752	HLHP193	published	Asia	atrophic gastritis	Asymptomatic	1		1	
753	HLHP253	published	Asia		unknown	1		1	
754	HLHP256	published	Asia	atrophic gastritis	Asymptomatic	1		1	
755	HLHP271	published	Asia	gastric ulcer	Ulcer	1		1	
756	HP116Bi	published	Africa		Asymptomatic	1		1	
774	HPKX-438-AG0C1	published	Europe	atrophic gastritis (subsequent gastric carcinoma)	Atrophic gastritis	1			
776	Hp-A11	published	North America	duodenal ulcer	Ulcer	1		1	
777	Hp-A14	published	North America	gastritis	Asymptomatic	1		1	
778	Hp-A16	published	North America	gastritis	Asymptomatic	1		1	
779	Hp-A17	published	North America	gastric ulcer	Ulcer	1		1	
780	Hp-A20	published	North America	duodenal ulcer	Ulcer	1		1	
781	Hp-A26	published	North America	gastritis	Asymptomatic	1		1	
782	Hp-A27	published	North America	gastritis	Asymptomatic	1		1	
783	Hp-A4	published	North America	duodenal ulcer	Ulcer			1	

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
784	Hp-A5	published	North America	gastric ulcer	Ulcer	1		1	
785	Hp-A6	published	North America	gastritis	Asymptomatic	1		1	
786	Hp-A8	published	North America	gastritis	Asymptomatic	1		1	
787	Hp-A9	published	North America		unknown	1		1	
788	Hp-H1	published	North America	gastritis	Asymptomatic	1		1	
789	Hp-H10	published	North America		unknown	1		1	
790	Hp-H11	published	North America	gastritis	Asymptomatic	1		1	
791	Hp-H16	published	North America	gastric/duodenal ulcer	Ulcer			1	
792	Hp-H18	published	North America		unknown	1		1	
793	Hp-H19	published	North America	gastritis	Asymptomatic	1		1	
794	Hp-H21	published	North America	gastritis	Asymptomatic	1		1	
795	Hp-H23	published	North America	gastritis	Asymptomatic	1		1	
796	Hp-H24	published	North America	gastric ulcer	Ulcer	1		1	
799	Hp-H27	published	North America	gastric ulcer	Ulcer	1		1	
800	Hp-H28	published	North America		unknown	1		1	
801	Hp-H29	published	North America		unknown	1		1	
802	Hp-H3	published	North America	gastritis	Asymptomatic	1		1	
803	Hp-H30	published	North America	gastric ulcer	Ulcer	1		1	
804	Hp-H34	published	North America	gastritis	Asymptomatic	1		1	
805	Hp-H36	published	North America		unknown	1		1	
806	Hp-H4	published	North America	gastritis	Asymptomatic	1		1	
807	Hp-H41	published	North America	duodenal ulcer	Ulcer	1		1	
808	Hp-H42	published	North America		unknown	1		1	
809	Hp-H43	published	North America	duodenal ulcer	Ulcer	1		1	
810	Hp-H44	published	North America		unknown	1		1	
811	Hp-H45	published	North America	duodenal ulcer	Ulcer	1		1	
812	Hp-H5b	published	North America		unknown	1		1	
813	Hp-H6	published	North America	gastritis	Asymptomatic			1	
814	Hp-H9	published	North America	gastritis	Asymptomatic	1		1	
822	Hp-P1	published	North America	gastritis	Asymptomatic	1		1	
823	Hp-P11	published	North America	gastritis	Asymptomatic	1		1	
825	Hp-P13	published	North America	gastritis	Asymptomatic	1		1	
827	Hp-P15	published	North America	gastritis	Asymptomatic	1		1	
829	Hp-P16	published	North America	gastritis	Asymptomatic	1		1	
831	Hp-P2	published	North America	gastritis	Asymptomatic	1		1	
832	Hp-P23	published	North America	gastritis	Asymptomatic	1		1	
833	Hp-P25	published	North America	gastritis	Asymptomatic			1	
836	Hp-P26	published	North America	gastritis	Asymptomatic	1		1	
837	Hp-P28b	published	North America		unknown	1		1	
839	Hp-P3	published	North America	gastritis	Asymptomatic	1		1	
840	Hp-P30	published	North America	gastritis	Asymptomatic	1		1	
842	Hp-P4	published	North America	gastritis	Asymptomatic			1	

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
843	Hp-P41	published	North America	gastritis	Asymptomatic			1	
846	Hp-P62	published	North America	gastritis	Asymptomatic			1	
847	Hp-P74	published	North America	gastritis	Asymptomatic	1		1	
848	Hp-P8	published	North America	gastritis	Asymptomatic			1	
850	N6	published	Asia	gastritis	Asymptomatic	1		1	
851	NAB47	published	Asia	duodenal ulcer	Ulcer	1		1	
852	NAD1	published	Asia	duodenal ulcer	Ulcer			1	
853	NCTC11637	published	Oceania		unknown	1		1	
854	NQ1671	published	South America	multifocal atrophic gastritis	Atrophic gastritis	1		1	
855	NQ1701	published	South America	multifocal atrophic gastritis	Atrophic gastritis	1		1	
862	NQ4044	published	South America	Gastritis/Ulcer	Ulcer	1		1	
863	NQ4053	published	South America	indefinite for dysplasia	unknown	1		1	
865	NQ4076	published	South America	multifocal atrophic gastritis	Atrophic gastritis	1		1	
866	NQ4099	published	South America	indefinite for dysplasia	unknown	1		1	
867	NQ4110	published	South America	multifocal atrophic gastritis w/o metaplasia	Atrophic gastritis	1		1	
868	NQ4161	published	South America	multifocal atrophic gastritis w/o metaplasia	Atrophic gastritis	1		1	
870	NQ4200	published	South America	intestinal metaplasia	Intestinal Metaplasia	1		1	
871	NQ4216	published	South America	indefinite for dysplasia	unknown	1		1	
872	NQ4228	published	South America	intestinal metaplasia	Intestinal Metaplasia	1		1	
874	R018c	published	North America		unknown	1		1	
875	R030b	published	North America	asymptomatic	Asymptomatic			1	
876	R036d	published	North America		unknown	1		1	
877	R037c	published	North America	asymptomatic	Asymptomatic	1		1	
878	R038b	published	North America	asymptomatic	Asymptomatic	1		1	
879	R046Wa	published	North America	asymptomatic	Asymptomatic	1		1	
880	R055a	published	North America		unknown	1		1	
882	R32b	published	North America	asymptomatic	Asymptomatic	1		1	
883	UMB-G1	published	North America		unknown	1		1	
884	UM007	published	Asia		unknown	1		1	
885	UM018	published	Asia		unknown	1		1	
886	UM034	published	Asia		unknown	1		1	
887	UM045	published	Asia		unknown			1	
888	UM054	published	Asia		unknown	1		1	
1341	SouthAfrica20	published	Africa		unknown	1		1	
1342	SouthAfrica50	published	Africa		unknown	1		1	
1343	UM023	published	Asia	Peptic Ulcer Disease	Ulcer	1		1	
1344	UM032	published	Asia	Peptic Ulcer Disease	Ulcer	1		1	
1346	UM037	published	Asia	Stomach fundus tumor	Gastric Cancer			1	
1347	UM038	published	Asia	Nonulcer Dyspepsia	unknown	1		1	
1348	UM065	published	Asia	Peptic Ulcer Disease	Ulcer	1		1	
1349	UM066	published	Asia	Peptic Ulcer Disease	Ulcer	1		1	
1351	UM067	published	Asia	Peptic Ulcer Disease	Ulcer	1		1	

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
1352	UM077	published	Asia	Peptic Ulcer Disease	Ulcer	1		1	
1353	UM084	published	Asia	Peptic Ulcer Disease	Ulcer	1		1	
1354	UM085	published	Asia	Nonulcer Dyspepsia	unknown	1		1	
1355	UM111	published	Asia	Nonulcer Dyspepsia	unknown	1		1	
1356	UM114	published	Asia	Peptic Ulcer Disease	Ulcer	1		1	
1361	PZ5004	published	South America		unknown	1		1	
1362	PZ5024	published	South America		unknown			1	
1363	PZ5026	published	South America		unknown	1		1	
1364	PZ5056	published	South America		unknown	1		1	
1365	PZ5080	published	South America		unknown	1		1	
1366	PZ5086	published	South America		unknown	1		1	
1367	GAM117Ai	published	Africa		unknown	1		1	
1368	NAK7	published	Asia		unknown	1		1	
1369	SA157A	published	Africa		unknown	1		1	
1370	SA144A	published	Africa		unknown	1		1	
1372	SA146A	published	Africa		unknown	1		1	
1374	SA155A	published	Africa		unknown	1		1	
1376	SA156A	published	Africa		unknown	1		1	
1379	SA158A	published	Africa		unknown	1		1	
1381	SA161A	published	Africa		unknown	1		1	
1383	SA162A	published	Africa		unknown	1		1	
1387	SA164A	published	Africa		unknown	1		1	
1391	SA166A	published	Africa		unknown	1		1	
1392	SA168A	published	Africa		unknown	1		1	
1394	SA169C	published	Africa		unknown	1		1	
1396	SA170C	published	Africa		unknown	1		1	
1397	SA171A	published	Africa		unknown	1		1	
1400	SA172C	published	Africa		unknown	1			
1401	SA173A	published	Africa		unknown	1		1	
1403	SA174A	published	Africa		unknown			1	
1404	SA175A	published	Africa		unknown	1		1	
1406	SA194A	published	Africa		unknown	1		1	
1410	SA213A	published	Africa		unknown	1		1	
1411	SA214A	published	Africa		unknown	1		1	
1415	SA220A	published	Africa		unknown	1		1	
1417	SA221A	published	Africa		unknown	1		1	
1419	SA222A	published	Africa		unknown	1		1	
1423	SA226A	published	Africa		unknown	1		1	
1424	SA227A	published	Africa		unknown	1		1	
1425	SA233A	published	Africa		unknown	1		1	
1431	SA253A	published	Africa		unknown	1		1	
1433	SA29A	published	Africa		unknown	1		1	

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
1437	SA301A	published	Africa		unknown	1		1	
1439	SA302A	published	Africa		unknown	1		1	
1441	SA303C	published	Africa		unknown	1		1	
1442	SA30A	published	Africa		unknown	1		1	
1445	SA34A	published	Africa		unknown	1		1	
1447	SA35A	published	Africa		unknown	1		1	
1448	SA36C	published	Africa		unknown	1		1	
1449	SA37A	published	Africa		unknown	1		1	
1451	SA40A	published	Africa		unknown	1		1	
1453	SA45A	published	Africa		unknown	1		1	
1455	SA46C	published	Africa		unknown	1		1	
1456	SA47A	published	Africa		unknown	1		1	
1458	SA160A	published	Africa		unknown	1		1	
3215	Sahul64	published	Oceania	Dyspepsia, Chronic antral gastritis	unknown			1	
3216	E48	published	Europe	chronic gastrtis	Asymptomatic			1	
3217	H13-1	published	Europe	gastric cancer	Gastric Cancer			1	
3222	HPARG63	published	South America	chronic gastritis	Asymptomatic			1	
3223	HPARG8G	published	South America	gastric ulcer disease	Ulcer			1	
3541	wls-5-12	published	Asia	chronic gastritis and nephritis	unknown			1	
3549	wls-5-3	published	Asia	chronic gastritis and nephritis	unknown			1	
3583	B23	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	1
3584	B24	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	1
3585	B25	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	1
3586	B26	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	1
3587	B27	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma				1
3588	B29	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	1
3589	B30	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	1
3590	B31	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	1
3591	B35	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	
3593	B37	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	1
3594	B40	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	1
3595	B41	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	1
3596	B43	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	
3597	B44	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma			1	1
3598	B47	Francis Megraud	Europe	MALT lymphoma	MALT Lymphoma		1	1	1
3599	B47-R1	Francis Megraud	Europe		MALT Lymphoma		1		
3600	B47-R2	Francis Megraud	Europe		MALT Lymphoma		1		
3601	B47-R3	Francis Megraud	Europe		MALT Lymphoma		1		
3602	GC11-HL	Francis Megraud	Europe	Gastric cancer	Gastric Cancer			1	1
3603	GC23-HL	Francis Megraud	Europe	Gastric cancer	Gastric Cancer			1	1
3604	GC26-HL	Francis Megraud	Europe	Gastric cancer	Gastric Cancer			1	1
3605	GC27-HL	Francis Megraud	Europe	Gastric cancer	Gastric Cancer			1	1

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
3606	GC31-B	Francis Megraud	Europe	GIST	Gastric Cancer			1	1
3607	GC34-HL	Francis Megraud	Europe	Gastric cancer	Gastric Cancer			1	1
3608	GC43-HL	Francis Megraud	Europe	Gastric cancer	Gastric Cancer			1	
3609	GC54-HL	Francis Megraud	Europe	Gastric cancer	Gastric Cancer			1	1
3611	GC65-HL	Francis Megraud	Europe	Gastric cancer	Gastric Cancer			1	1
3612	GC67-HL	Francis Megraud	Europe	Gastric cancer	Gastric Cancer			1	
3613	SSR1	Sinead Smith	Europe	Antral ucler with mod chronic gastritis, exten IM	Intestinal Metaplasia			1	
3614	SSR2	Sinead Smith	Europe	Mild chronic gastritis, no evidence of IM	Asymptomatic	1		1	1
3615	SSR3	Sinead Smith	Europe	Mild chronic gastritis, no evidence of IM	Asymptomatic	1		1	
3616	SSR4	Sinead Smith	Europe	Moderate chronic gastritis, no evidence of IM	Asymptomatic	1		1	
3617	SSR5	Sinead Smith	Europe	Moderate chronic gastritis, no evidence of IM	Asymptomatic	1		1	1
3618	SSR7	Sinead Smith	Europe	Moderate chronic gastritis, no evidence of IM	Asymptomatic	1		1	
3619	SSR8	Sinead Smith	Europe	Focal acute and mod chronic inflammation	Asymptomatic	1		1	
3620	SSR9	Sinead Smith	Europe	Moderate chronic gastritis, no evidence of IM	Asymptomatic	1		1	
3622	SSR12	Sinead Smith	Europe	Moderate chronic gastritis, focal IM present	Intestinal Metaplasia	1		1	
3623	SSR13	Sinead Smith	Europe	Moderate chronic gastritis, no evidence of IM	Asymptomatic	1		1	1
3624	SSR14	Sinead Smith	Europe	Antral ucler with mod chronic gastritis, exten IM	Intestinal Metaplasia	1		1	
3625	SSR17	Sinead Smith	Europe	Mild chronic gastritis, no evidence of IM	Asymptomatic	1		1	
3627	SSR20	Sinead Smith	Europe	Mild chronic gastritis, no evidence of IM	Asymptomatic	1		1	
3628	SSR22	Sinead Smith	Europe	Chronic gastritis with mild to mod activity. No IM	Asymptomatic	1		1	
3629	SSR23	Sinead Smith	Europe	Mild chronic gastritis, no evidence of IM	Asymptomatic	1		1	
3630	SSR33	Sinead Smith	Europe	Moderate chronic gastritis, no evidence of IM	Asymptomatic	1		1	
3631	SSR40	Sinead Smith	Europe	Marked acute and chronic inflammtation	unknown	1		1	
3632	SSR43	Sinead Smith	Europe		unknown	1		1	
3634	3774	Francis Megraud	Europe	Ulcer	Ulcer			1	1
3636	ANT170	Francis Megraud	Europe	Ulcer	Ulcer			1	1
3638	GIL237	Francis Megraud	Europe	Ulcer	Ulcer			1	
3639	BON254	Francis Megraud	Europe	Ulcer	Ulcer	1		1	1
3640	CHA185	Francis Megraud	Europe	Ulcer	Ulcer	1		1	1
3641	GRA247	Francis Megraud	Europe	Ulcer	Ulcer	1		1	1
3642	PHI092	Francis Megraud	Europe	Ulcer	Ulcer	1		1	1
3643	GC30-HL	Francis Megraud	Europe	Gastric cancer	Gastric Cancer			1	1
3644	GC52-HL	Francis Megraud	Europe	Gastric Cancer	Gastric Cancer			1	
3645	3755	Francis Megraud	Europe	Gastritis	Asymptomatic			1	1
3646	3770	Francis Megraud	Europe	Gastritis	Asymptomatic			1	1
3647	3800	Francis Megraud	Europe	Gastritis	Asymptomatic	1		1	
3648	3802	Francis Megraud	Europe	Gastritis	Asymptomatic			1	1
3649	3824	Francis Megraud	Europe	Gastritis	Asymptomatic			1	1
3650	TN2GF4	Jane Mikhail	Unknown	Duodenal ulcer	Ulcer			1	
3653	565-99	Jane Mikhail	Unknown		unknown			1	
3658	3754	Francis Megraud	Europe	Gastritis	Asymptomatic	1		1	1
3659	3745	Francis Megraud	Europe	Gastritis	Asymptomatic	1		1	1

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
3662	3697	Francis Megraud	Europe	Gastritis	Asymptomatic	1		1	1
3663	3738	Francis Megraud	Europe	Ulcer	Ulcer	1		1	1
3664	3699	Francis Megraud	Europe	Gastritis	Asymptomatic			1	1
3666	3746	Francis Megraud	Europe	Gastritis	Asymptomatic	1		1	
3667	SW7A	Jane Mikhail	Europe	Moderate active chronic inflammation with IM	Intestinal Metaplasia	1			
3669	SW21A	Jane Mikhail	Europe	Chronic transmural inflammation	unknown	1			
3671	SW23Ai	Jane Mikhail	Europe	Severe inflammation (details in comment)	unknown	1			
3674	GC69-HL	Francis Megraud	Europe	Gastric cancer	Gastric Cancer			1	1
4443	1152-04	published	Europe	Duodenal ulcer	Ulcer			1	
4445	1198-04	published	Europe	Duodenal ulcer	Ulcer			1	
4446	207-99	published	Europe	Nonulcer Dyspepsia	unknown			1	
4447	499-02	published	Europe	Gastric Ulcer	Ulcer			1	
4448	655-99	published	Europe	Nonulcer Dyspepsia	unknown			1	
4449	Yangon244	Yoshio Yamaoka	Asia		unknown	1		1	
4450	Yangon233	Yoshio Yamaoka	Asia		unknown	1		1	
4451	Yangon222	Yoshio Yamaoka	Asia		unknown	1		1	
4452	Yangon202	Yoshio Yamaoka	Asia		unknown	1		1	
4453	Yangon190	Yoshio Yamaoka	Asia		unknown	1		1	
4454	Yangon188	Yoshio Yamaoka	Asia		unknown			1	
4455	Yangon179	Yoshio Yamaoka	Asia		unknown	1		1	
4456	Yangon173	Yoshio Yamaoka	Asia		unknown	1		1	
4457	Yangon159	Yoshio Yamaoka	Asia		unknown	1		1	
4458	Yangon142	Yoshio Yamaoka	Asia		unknown	1		1	
4459	Yangon132	Yoshio Yamaoka	Asia		unknown			1	
4460	oki102	published	Asia	Gastric Atrophy	unknown			1	
4461	NP05	Yoshio Yamaoka	Asia		unknown	1		1	
4462	NP05-282	Yoshio Yamaoka	Asia		unknown			1	
4463	NP05-278	Yoshio Yamaoka	Asia		unknown	1		1	
4464	NP05-272	Yoshio Yamaoka	Asia		unknown	1		1	
4465	NP05-266	Yoshio Yamaoka	Asia		unknown			1	
4466	NP05-261	Yoshio Yamaoka	Asia		unknown			1	
4467	NP05-250	Yoshio Yamaoka	Asia		unknown	1		1	
4468	NP05-234	Yoshio Yamaoka	Asia		unknown			1	
4469	NP05-227	Yoshio Yamaoka	Asia		unknown			1	
4470	NP05-124	Yoshio Yamaoka	Asia		unknown			1	
4471	NP05-121	Yoshio Yamaoka	Asia		unknown	1		1	
4472	NP05-112	Yoshio Yamaoka	Asia		unknown	1		1	
4473	NP05-107	Yoshio Yamaoka	Asia		unknown	1		1	
4474	NP05-105	Yoshio Yamaoka	Asia		unknown	1		1	
4475	NP04	Yoshio Yamaoka	Asia		unknown	1		1	
4476	Myanmar66	Yoshio Yamaoka	Asia		unknown	1		1	
4477	Myanmar52	Yoshio Yamaoka	Asia		unknown	1		1	

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
4478	Myanmar51	Yoshio Yamaoka	Asia		unknown	1		1	
4479	Mandalay60	Yoshio Yamaoka	Asia		unknown			1	
4480	Mandalay46	Yoshio Yamaoka	Asia		unknown			1	
4481	Mandalay38	Yoshio Yamaoka	Asia		unknown	1		1	
4482	Mandalay30	Yoshio Yamaoka	Asia		unknown	1		1	
4483	Mandalay13	Yoshio Yamaoka	Asia		unknown			1	
4484	Mandalay03	Yoshio Yamaoka	Asia		unknown	1		1	
4485	oki112	published	Asia	Gastric Atrophy	unknown			1	
4486	oki128	published	Asia	Gastric Atrophy	unknown			1	
4487	oki154	published	Asia	Duodenal Ulcer	Ulcer			1	
4489	oki673	published	Asia	Gastric Ulcer	Ulcer			1	
4490	oki828	published	Asia	Duodenal Ulcer	Ulcer			1	
4491	oki898	published	Asia	Duodenal Ulcer	Ulcer			1	
4492	J166	published	North America		unknown			1	
4493	BM013A	published	Oceania	Asymptomatic	Asymptomatic			1	
4496	Hp238	published	Asia	Gastric Malt Lymphoma	MALT Lymphoma			1	
4497	BM012A	published	Oceania	Asymptomatic	Asymptomatic			1	
4528	Nic01-A	published	North America	Corpus atrophy but no signs of metaplasia	Atrophic gastritis	1		1	
4530	Nic03-A	published	North America	Corpus atrophy but no signs of metaplasia	Atrophic gastritis	1		1	
4532	Nic04-A	published	North America	Corpus atrophy but no signs of metaplasia	Atrophic gastritis	1		1	
4534	Nic05-A	published	North America	Corpus atrophy but no signs of metaplasia	Atrophic gastritis	1		1	
4536	Nic06-A	published	North America	Corpus atrophy but no signs of metaplasia	Atrophic gastritis	1		1	
4538	Nic07-A	published	North America	Antrum predominant gastritis but no signs of atrop	Asymptomatic	1		1	
4539	Nic07-C	published	North America	Antrum predominant gastritis but no signs of atrop	Asymptomatic			1	
4540	Nic08-C2	published	North America	Antrum predominant gastritis but no signs of atrop	Asymptomatic	1		1	
4542	Nic09-A	published	North America	Antrum predominant gastritis but no signs of atrop	Asymptomatic	1		1	
4544	Nic10-A	published	North America	Antrum predominant gastritis but no signs of atrop	Asymptomatic	1		1	
4546	Nic11-A	published	North America	Antrum predominant gastritis but no signs of atrop	Asymptomatic	1		1	
4548	Nic12-A	published	North America	Antrum predominant gastritis but no signs of atrop	Asymptomatic	1		1	
4549	Nic12-C	published	North America	Antrum predominant gastritis but no signs of atrop	Asymptomatic			1	
4550	Nic13-A	published	North America	Pan-gastritis but no signs of atrophy or metaplasi	Asymptomatic	1		1	
4552	Nic14-A	published	North America	Pan-gastritis but no signs of atrophy or metaplasi	Asymptomatic	1		1	
4553	Nic14-C	published	North America	Pan-gastritis but no signs of atrophy or metaplasi	Asymptomatic	1		1	
4554	Nic15-A	published	North America	Pan-gastritis but no signs of atrophy or metaplasi	Asymptomatic	1		1	
4556	Nic16-A	published	North America	Pan-gastritis but no signs of atrophy or metaplasi	Asymptomatic	1		1	
4558	Nic17-A	published	North America	Pan-gastritis but no signs of atrophy or metaplasi	Asymptomatic	1		1	
4560	Nic18-A	published	North America	Pan-gastritis but no signs of atrophy or metaplasi	Asymptomatic	1		1	
4562	Nic19-A	published	North America	Intestinal metaplasia and atrophy	Intestinal Metaplasia	1		1	
4563	Nic19-C	published	North America	Intestinal metaplasia and atrophy	Intestinal Metaplasia			1	
4564	Nic20-A	published	North America	Intestinal metaplasia and atrophy	Intestinal Metaplasia	1		1	
4565	Nic20-C	published	North America	Intestinal metaplasia and atrophy	Intestinal Metaplasia	1		1	
4566	Nic21-C	published	North America	Intestinal metaplasia and atrophy	Intestinal Metaplasia	1		1	

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
4568	NQ315	published	South America	Multifocal Atrophic gastritis	Atrophic gastritis	1		1	
4569	NQ392	published	South America	Multifocal Atrophic gastritis	Atrophic gastritis	1		1	
4575	YN1-91	published	Asia	Gastritis	Asymptomatic			1	
4577	YN4-84	published	Asia	Gastritis	Asymptomatic			1	
4578	Manado-1	published	Asia		unknown	1		1	
4579	Taiwan-47	published	Asia		unknown	1		1	
6144	Nic02-A	published	North America	atrophic gastritis	Atrophic gastritis			1	
6283	Nic29-A	published	North America		unknown	1		1	
6284	Nic30-A	published	North America	gastritis	Asymptomatic	1		1	
6285	Nic31-A	published	North America		unknown	1		1	
6286	Nic32-A	published	North America	gastritis	Asymptomatic	1		1	
6288	Nic23-A	published	North America	Atrophy and metaplasia	Intestinal Metaplasia	1		1	
6289	Nic24-A	published	North America	gastritis	Asymptomatic	1			
6290	Nic25-A	published	North America	Atrophy and metaplasia	Intestinal Metaplasia	1		1	
6291	Nic26-A	published	North America	gastritis	Asymptomatic	1		1	
6292	Nic27-A	published	North America	gastritis	Asymptomatic	1		1	
6293	Nic28-A	published	North America	gastritis	Asymptomatic	1		1	
8602	21580	Alain Burette	Europe	Gastric Cancer	Gastric Cancer			1	1
8603	30908	Alain Burette	Europe	Control	Asymptomatic			1	1
8604	30950	Alain Burette	Europe	Gastric Cancer	Gastric Cancer			1	1
8605	31235	Alain Burette	Europe	Control	Asymptomatic			1	1
8606	36166	Alain Burette	Europe	Control	Asymptomatic			1	
8607	38185	Alain Burette	Europe	Gastric Cancer	Gastric Cancer			1	1
8608	444	John Atherton	Europe	Ulcer	Ulcer			1	
8609	448	John Atherton	Europe	Normal	Asymptomatic			1	
8610	456	John Atherton	Europe	Normal	Asymptomatic			1	
8611	462	John Atherton	Europe	Normal	Asymptomatic			1	
8612	518	John Atherton	Europe	Normal	Asymptomatic			1	
8614	638	John Atherton	Europe	Ulcer	Ulcer			1	
8615	HE-C1	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8616	HE-C32	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8617	HE-C33	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8618	HE-C34	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8619	HE-C38	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8620	HE-C40	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8621	HE-C50	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8622	HE-C52	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8623	HE-C55	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8624	HE-C57	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8625	HE-C58	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8626	HE-C59	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8627	HE-C66	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
8628	HE-C73	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8629	HE-C9	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8630	HE-C11	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8631	HE-C13	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8632	HE-C18	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8633	HE-C23	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8634	HE-C30	Kaisa Thorell	Europe	Gastric Cancer	Gastric Cancer			1	
8635	HE-NC1-1	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8636	HE-NC13-6	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8637	HE-NC14-2	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8638	HE-NC18-1	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8639	HE-NC18-2	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8640	HE-NC18-4	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8641	HE-NC13-5	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8642	HE-NC19-3	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8643	HE-NC19-5	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8644	HE-NC20-5	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8645	HE-NC1-2	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8646	HE-NC23-2a	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8647	HE-NC24-6	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8648	HE-NC26-4	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8649	HE-NC27-4	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8650	HE-NC29-2	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8651	HE-NC30-2	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8652	HE-NC30-3	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8654	HE-NC32-4	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8655	HE-NC32-5	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8656	HE-NC5-3	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8657	HE-NC36-3	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8659	HE-NC38-2	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8660	HE-NC38-4	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8661	HE-NC38-5	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8662	HE-NC39-3	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8663	HE-NC47-5	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8664	HE-NC55-1	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8665	HE-NC55-2	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8666	HE-NC55-5	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8667	HE-NC60-1	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8668	HE-NC60-3	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8669	HE-NC61-4	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8672	HE-NC89-4	Kaisa Thorell	Europe	Control	Asymptomatic			1	
8673	HE-NC9-1	Kaisa Thorell	Europe	Control	Asymptomatic			1	

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
8674	HE-NCI1-1	Kaisa Thorell	Europe	Control	Atrophic gastritis			1	
8681	2012-26	Javier Torres	North America	Metaplasia	Intestinal Metaplasia	1		1	
8682	22025	Javier Torres	South America	Gastritis	Asymptomatic	1		1	
8683	ms167	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8684	ms1055	Javier Torres	North America	Cancer	Gastric Cancer	1		1	
8685	22402	Javier Torres	South America	Cancer	Gastric Cancer	1		1	
8687	22087	Javier Torres	South America	Gastritis	Asymptomatic	1		1	
8688	ms1063	Javier Torres	North America	Cancer	Gastric Cancer	1		1	
8690	26084	Javier Torres	South America	Cancer	Gastric Cancer	1		1	
8692	2004-20	Javier Torres	North America	Metaplasia	Intestinal Metaplasia	1		1	
8693	2005-98	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8694	ms203	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8695	2005-72	Javier Torres	North America	Metaplasia	Intestinal Metaplasia	1		1	
8696	26093	Javier Torres	South America	Cancer	Gastric Cancer	1		1	
8697	ms1078	Javier Torres	North America	Cancer	Gastric Cancer	1		1	
8698	2006-52	Javier Torres	North America	Cancer	Gastric Cancer	1		1	
8699	2003-98	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8700	2006-407	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8701	2006-103	Javier Torres	North America	Metaplasia	Intestinal Metaplasia	1		1	
8702	22346	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8703	ms15	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8704	22337	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8705	ms23	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8707	ms2	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8709	22341	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8710	2006-56	Javier Torres	North America	Metaplasia	Intestinal Metaplasia	1		1	
8711	22023	Javier Torres	South America	Gastritis	Asymptomatic	1		1	
8712	22327	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8713	ms931	Javier Torres	North America	Cancer	Gastric Cancer	1		1	
8714	2005-100	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8715	ms1080	Javier Torres	North America	Cancer	Gastric Cancer	1		1	
8716	ms13	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8717	26100	Javier Torres	South America	Cancer	Gastric Cancer	1		1	
8718	2003-84	Javier Torres	North America	Duodenal Ulcer	Ulcer	1		1	
8720	22046	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8721	2006-479	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8722	ms1054	Javier Torres	North America	Cancer	Gastric Cancer	1		1	
8723	ms44	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8724	22389	Javier Torres	South America	Gastritis	Asymptomatic	1		1	
8726	22366	Javier Torres	South America	Gastritis	Asymptomatic	1		1	
8728	22013	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8730	22362	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
8731	2005-126	Javier Torres	North America	Metaplasia	Intestinal Metaplasia	1		1	
8732	2003-103	Javier Torres	North America	Metaplasia	Intestinal Metaplasia	1		1	
8734	22367	Javier Torres	South America	Gastritis	Asymptomatic	1		1	
8735	22021	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8736	2003-107	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8737	2006-480	Javier Torres	North America	Metaplasia	Intestinal Metaplasia	1		1	
8738	22385	Javier Torres	South America	Gastritis	Asymptomatic	1		1	
8739	24004	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8740	2006-4	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8741	22370	Javier Torres	South America	Gastritis	Asymptomatic	1		1	
8742	22311	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8743	22390	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8744	22339	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8745	2011-145	Javier Torres	North America	Cancer	Gastric Cancer	1		1	
8746	22312	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8747	22322	Javier Torres	South America	Gastritis	Asymptomatic	1		1	
8749	2004-2	Javier Torres	North America	Duodenal Ulcer	Ulcer	1		1	
8750	22331	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8751	26024	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8752	22368	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8753	22360	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8754	22378	Javier Torres	South America	Gastritis	Asymptomatic	1		1	
8756	22019	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8757	22371	Javier Torres	South America	Gastritis	Asymptomatic	1		1	
8758	22020	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8760	22315	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8761	22335	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8762	ms176	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8763	22393	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8764	22093	Javier Torres	South America	Cancer	Gastric Cancer	1		1	
8766	22095	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8768	22347	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8771	2011-41	Javier Torres	North America	Gastritis	Asymptomatic	1		1	
8773	22388	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8775	22351	Javier Torres	South America	Metaplasia	Intestinal Metaplasia	1		1	
8776	22384	Javier Torres	South America	Gastritis	Asymptomatic	1		1	
8778	24008	Javier Torres	South America	Atrophic Gastritis	Atrophic gastritis	1		1	
8783	ms965	Javier Torres	North America	Cancer	Gastric Cancer	1		1	
9002	B38-6M-1	Philippe Lehours	Europe		MALT Lymphoma		1		
9003	B38-6M-2	Philippe Lehours	Europe		MALT Lymphoma		1		
9004	B38-6M-3	Philippe Lehours	Europe		MALT Lymphoma		1		
9005	B38-6W-1	Philippe Lehours	Europe		MALT Lymphoma		1		

Appendix C: Information Table on all strains used in this thesis

BIGSId	strain name	collection	isolation continent	host pathology		Used in Chapter			
				complete	simplified	3	4	5	6
9006	B38-6W-2	Philippe Lehours	Europe		MALT Lymphoma		1		
9007	B38-6W-3	Philippe Lehours	Europe		MALT Lymphoma		1		
9008	B38-12M-1	Philippe Lehours	Europe		MALT Lymphoma		1		
9009	B38-12M-2	Philippe Lehours	Europe		MALT Lymphoma		1		
9010	B38-12M-3	Philippe Lehours	Europe		MALT Lymphoma		1		
9011	B38-Ri-1	Philippe Lehours	Europe		MALT Lymphoma		1		
9012	B38-Ri-2	Philippe Lehours	Europe		MALT Lymphoma		1		
9013	B47-6M-1	Philippe Lehours	Europe		MALT Lymphoma		1		
9014	B47-6M-2	Philippe Lehours	Europe		MALT Lymphoma		1		
9015	B47-6W-1	Philippe Lehours	Europe		MALT Lymphoma		1		
9016	B47-6W-2	Philippe Lehours	Europe		MALT Lymphoma		1		
9017	B47-12M-1	Philippe Lehours	Europe		MALT Lymphoma		1		
9018	B47-12M-2	Philippe Lehours	Europe		MALT Lymphoma		1		
9583	19027	Alain Burette	Europe	Gastric Cancer	Gastric Cancer				1
9584	GC62-HL-2	Francis Megraud	Europe	Gastric Cancer	Gastric Cancer				1

Appendix D: Published article: Rapid evolution of distinct *Helicobacter pylori* subpopulations in the Americas

This research article was published in the PLOS Genetics Journal in 2017 by Kaisa Thorell and Koji Yahara

My participation in this article was in (i) preparation of the dataset, (ii) analysis of the results.of FineStructure and ChromoPainter, (iii) creation of figure for ChromoPainter results, (iv) design, realisation and analysis of the accessory genome variation.

PLOS | GENETICS

RESEARCH ARTICLE

Rapid evolution of distinct *Helicobacter pylori* subpopulations in the Americas

Kaisa Thorell[1☯], Koji Yahara[2☯], Elvire Berthenet[3], Daniel J. Lawson[4], Jane Mikhail[3], Ikuko Kato[5], Alfonso Mendez[6], Cosmeri Rizzato[7], María Mercedes Bravo[8], Rumiko Suzuki[9], Yoshio Yamaoka[9], Javier Torres[10], Samuel K. Sheppard[11], Daniel Falush[11]*

1 Microbiology, Tumour and Cell Biology, Karolinska Institutet, Stockholm, Sweden, **2** Dept. of Bacteriology II, National Institute of Infectious Diseases, Tokyo, Japan, **3** Medical Microbiology and Infectious Disease group, Swansea University, Swansea, Wales, United Kingdom, **4** Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom, **5** Karmanos Cancer Institute, Wayne State University, Detroit, Michigan, United States of America, **6** Instituto Politecnico Nacional, ENCB, Mexico City, Mexico, **7** Dipartimento di Ricerca Traslazionale e Nuove Tecnologie in Medicina e Chirurgia, Universitá di Pisa, Pisa, Italy, **8** Grupo de Investigación en Biología del Cáncer, Instituto Nacional de Cancerología, Bogota, Colombia, **9** Dept. of Environmental and Preventive Medicine, Oita University Faculty of Medicine, Oita, Japan, **10** Unidad de Investigación en Enfermedades Infecciosas, UMAE Pediatria, IMSS, Mexico City, Mexico, **11** Milner Center for Evolution, Dept. of Biology and Biochemistry, University of Bath, Bath, United Kingdom

☯ These authors contributed equally to this work.
* danielfalush@googlemail.com

OPEN ACCESS

Citation: Thorell K, Yahara K, Berthenet E, Lawson DJ, Mikhail J, Kato I, et al. (2017) Rapid evolution of distinct *Helicobacter pylori* subpopulations in the Americas. PLoS Genet 13(2): e1006546. doi:10.1371/journal.pgen.1006546

Editor: Graham Coop, University of California Davis, UNITED STATES

Received: August 12, 2016

Accepted: December 19, 2016

Published: February 23, 2017

Data availability statement: Both the genome sequences and the alignment are available at the public data repository Dryad (http://datadryad.org/), with doi:10.5061/dryad.8qp4n.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

For the last 500 years, the Americas have been a melting pot both for genetically diverse humans and for the pathogenic and commensal organisms associated with them. One such organism is the stomach-dwelling bacterium *Helicobacter pylori*, which is highly prevalent in Latin America where it is a major current public health challenge because of its strong association with gastric cancer. By analyzing the genome sequence of *H. pylori* isolated in North, Central and South America, we found evidence for admixture between *H. pylori* of European and African origin throughout the Americas, without substantial input from pre-Columbian (hspAmerind) bacteria. In the US, strains of African and European origin have remained genetically distinct, while in Colombia and Nicaragua, bottlenecks and rampant genetic exchange amongst isolates have led to the formation of national gene pools. We found three outer membrane proteins with atypical levels of Asian ancestry in American strains, as well as alleles that were nearly fixed specifically in South American isolates, suggesting a role for the ethnic makeup of hosts in the colonization of incoming strains. Our results show that new *H. pylori* subpopulations can rapidly arise, spread and adapt during times of demographic flux, and suggest that differences in transmission ecology between high and low prevalence areas may substantially affect the composition of bacterial populations.

Author summary

Helicobacter pylori is one of the best studied examples of an intimate association between bacteria and humans, due to its ability to colonize the stomach for decades and to transmit

from generation to generation. A number of studies have sought to link diversity in *H. pylori* to human migrations but there are some discordant signals such as an "out of Africa" dispersal within the last few thousand years that has left a much stronger signal in bacterial genomes than in human ones. In order to understand how such discrepancies arise, we have investigated the evolution of *H. pylori* during the recent colonization of the Americas. We find that bacterial populations evolve quickly and can spread rapidly to people of different ethnicities. Distinct new bacterial subpopulations have formed in Colombia from a European source and in Nicaragua and the US from African sources. Genetic exchange between bacterial populations is rampant within Central and South America but is uncommon within North America, which may reflect differences in prevalence. Our results also suggest that adaptation of bacteria to particular human ethnic groups may be confined to a handful of genes involved in interaction with the immune system.

Introduction

In 1492, Christopher Columbus initiated a rapid colonization of the New World, principally by European migrants and Africans brought as slaves that had catastrophic consequences for the indigenous population. The new migrants brought unfamiliar weapons and pathogens [1], including new populations of the stomach-colonizing bacterium *Helicobacter pylori*. *H. pylori* can persist for decades in the stomach, and is often transmitted vertically from parent to child but can also be acquired from individuals in close proximity. *H. pylori* evolves rapidly by both mutation and homologous recombination with other co-colonizing strains [2].

Studies of the global diversity of *H. pylori* have shown that Europeans, Africans and Native Americans carry genetically distinct populations of bacteria; named hpEurope, hpAfrica1 and hpAfrica2, and hspAmerind, respectively [3]. The relationships between bacterial populations reflect differentiation that has arisen during the complex migration history of humans, with the prefix "hp" indicating a population and "hsp" indicating a subpopulation, which are genetically distinct from each other but less differentiated than populations. hspAmerind bacteria are presumed to be descendants of the strains present in the Americas prior to 1492, and are a subpopulation of hspEAsia, which is found in Asian countries such as China and Japan. However, these strains are rare even within groups with substantial Native American ancestry and may being dying out in competition with other strains, due to low diversity within the population or other factors [4]. hpEurope bacteria are themselves ancient hybrids between two populations, whose close relatives are currently found in unadmixed populations in North East Africa (hpNEAfrica) and central Asia (hpAsia2). The Tyrolean Iceman, Ötzi, who died 5300 years ago in central Europe, was infected by an hpAsia2 strain with little or no African ancestry [5], suggesting that the admixture probably took place within the last few thousand years.

In Latin America, gastric cancer is a leading cause of cancer death, and some countries in the region have among the highest mortality rates worldwide [6]. However, the mortality rates vary in different geographic regions, both between neighboring countries and within nations [6,7]. Several studies have been performed comparing *H. pylori* ancestry in high- and low risk areas and have linked phylogeographic origin of the bacteria, as well as discordant origin of bacteria and host, to increased risk of gastric cancer development [3,8]. However, these studies have been performed using MLST analysis that, being based only on seven housekeeping genes, is limited in its resolution compared to whole-genome comparisons.

To investigate if American *H. pylori* strains have differentiated from those found in the Old World by mixture, genetic drift or natural selection, we combined hundreds of publicly available genomes with over hundred newly sequenced genomes of *H. pylori* sampled in Latin America (Mexico, Nicaragua, and Colombia), Europe, and Central Asia. We show that the American bacterial populations have undergone substantial evolution within 500 years and our results also suggest that *H. pylori* transmission biology has been as important as human migration in determining extant patterns of diversity.

Results

We used the Chromopainter/fineSTRUCTURE pipeline [9,10] to assess the population structure within our global collection of isolates (n = 401, described in S1 Table). Insight into the ancestry of each isolate is obtained by treating it as a "recipient" and using Chromopainter to fit it as a mosaic of DNA chunks, i.e. sets of contiguous SNPs, from a "donor panel" of other genomes. The painting can be interpreted genealogically as described in more detail in [9], namely in the local genealogy for any of the sites within a given chunk, the most recent coalescence involving the recipient individual is with the donor individual for that chunk. Each chunk thus provides information on the most recent clonal relationships and/or genetic exchange between different strains in the sample. In *H. pylori* recombination rates are very high and unless individuals in close proximity are sampled, it is rare to find clear evidence of direct clonal descent [11].

We used two different donor panels. A first consisted only of Old World (European, African and Asian) isolates. Since we expect that almost all of the gene flow historically has been from the Old World to the New World, using an Old World panel allows us to investigate the origins of each New World *H. pylori*, without the complication of determining how the strains are related to each other. Although we are principally interested in gene flow within the last 500 years, hspAmerind strains are excluded from the donor set because many of the strains have undergone post-Columbian admixture with other populations. The DNA in any case originated from the Old World, albeit probably $>$ 10,000 years ago.

The second global painting panel includes all New World strains, apart from the specific individual being painted. Many, although not all of the chunks inferred to be donated by other New World strains in this painting will represent coalescent events that happened in the New World. Therefore using this painting panel allows us to investigate recent demography within the New World.

fineSTRUCTURE uses the output of Chromopainter to assign individuals to populations with distinct ancestry profiles. We applied fineSTRUCTURE to the global painting to identify subpopulations in the dataset (Fig 1). In order to make display and reporting of the results tractable, we merged the most similar populations until 12 distinct populations remained, 5 of which are restricted to the New World. The "palette" of each strain, representing the proportion of chunks that come from each population in the donor panel, was determined for both the Old World (Fig 2A) and the Global (Fig 2B) painting. One of the twelve populations, hspMiscAmerica, contains isolates that are not particularly closely related to each other and should not be thought of as a coherent population (Fig 1). The fineSTRUCTURE results are congruent with those obtained by Principal Component Analysis, PCA, which show differences between the subpopulations within the first 5 Principal Components (S1 Fig) but are easier to interpret.

Increased number of isolates reveals substructures in the Old World populations

Each of the 7 populations found in the Old World has been reported previously with the exception that, with the addition of the large number of isolates in this study, hpEurope isolates

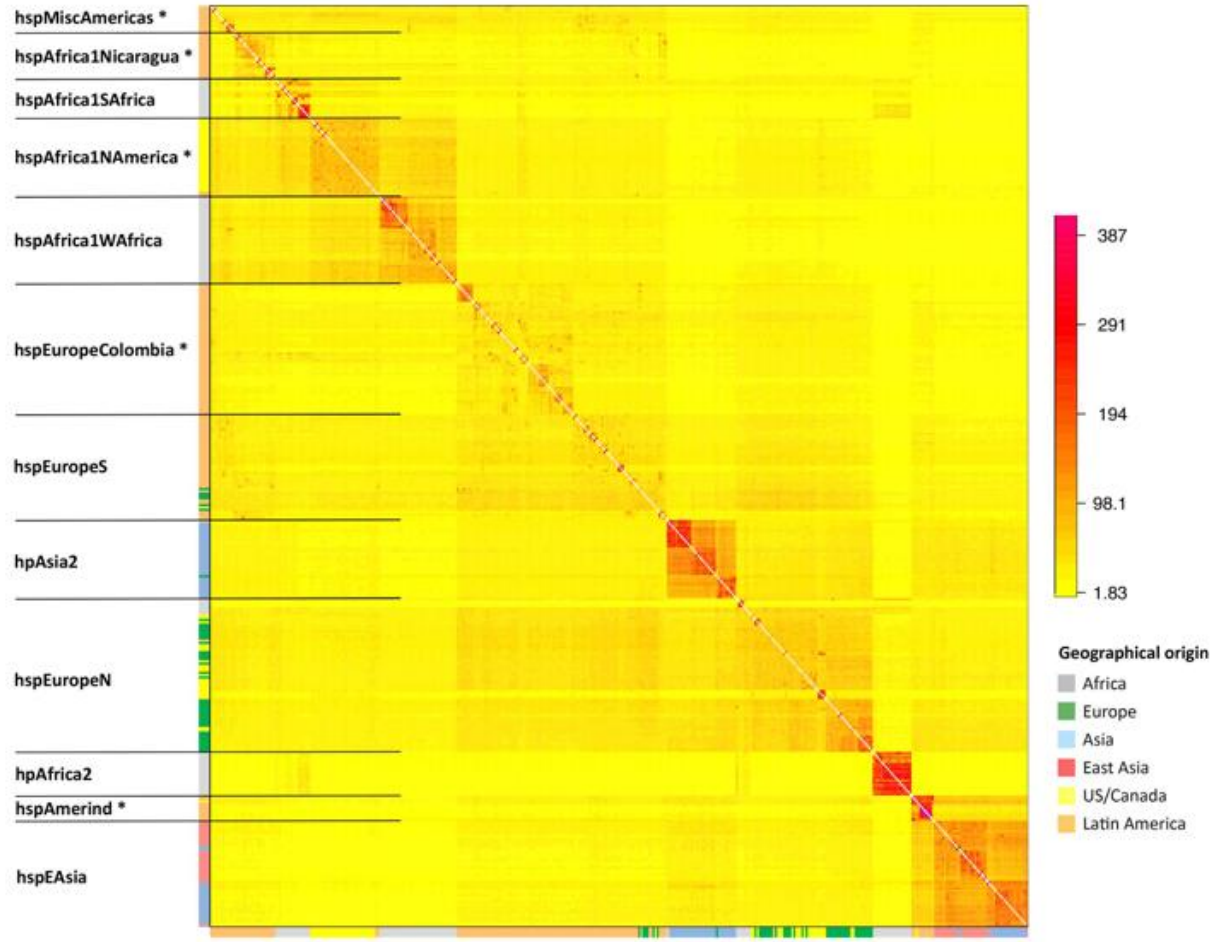


Fig 1. Population structure of global *H. pylori* strains. The colour of each cell of the matrix indicates the expected number of DNA chunks imported from a donor genome (column) to a recipient genome (row). The boundaries between named populations are marked with lines, with New World populations marked with an asterisk. The colour bar on the left indicates the geographical locations where the strains were sampled.

doi:10.1371/journal.pgen.1006546.g001

separated into two distinct groups, which we provisionally label hspEuropeN and hspEuropeS (Fig 1). Our geographical sampling within Europe is limited but this split is likely to reflect the previously observed North to South gene frequency cline [12,13], with the hspEuropeS isolates having a larger fraction of their palette from African populations and hspEuropeN having a higher proportion from hpAsia2. The other five populations, hpAfrica2, hspAfrica1SAfrica, hspAfrica1WAfrica, hpEastAsia and hpAsia2 are highly distinct from each other, each receiving more than half of their palette from their own population in the Old World painting.

Distinct subpopulations of mixed hpEurope and hpAfrica1 ancestry in American *H. pylori*

Among the isolates from the Americas, five additional subpopulations could be distinguished; four have palettes consistent with being European/African hybrids, according to the Old

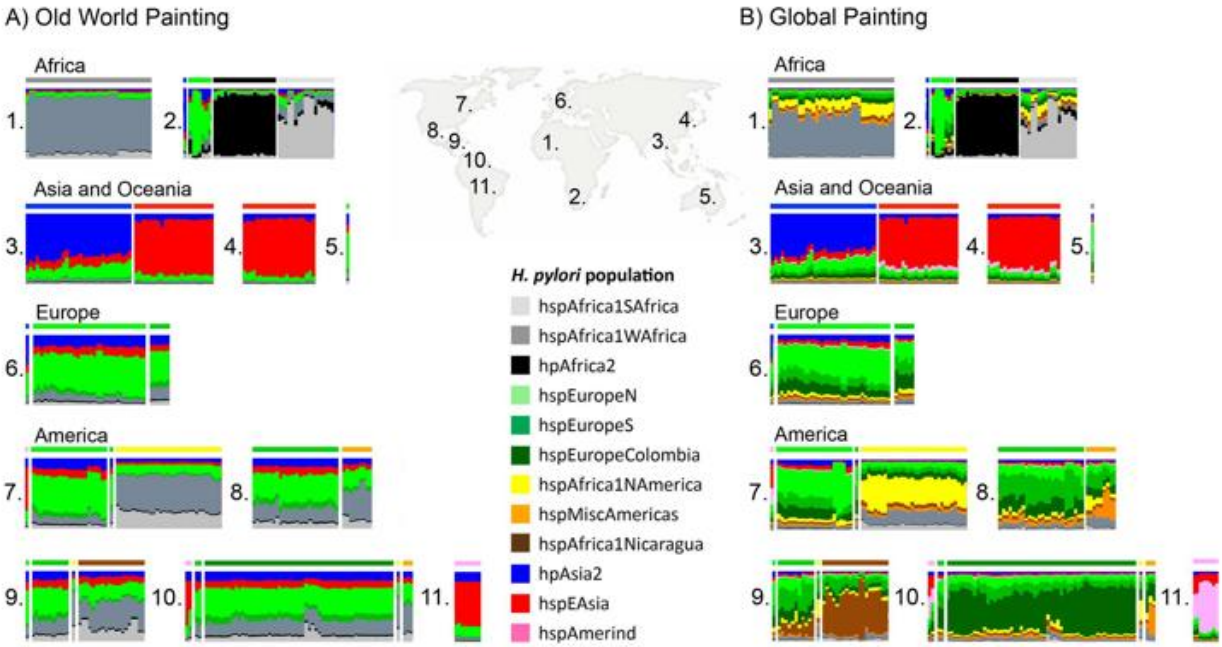


Fig 2. Ancestry of *H. pylori* inferred by chromosome painting. Each vertical bar represents one isolate, which are ordered by geographical origin (1–11). 1: West Africa, 2: South Africa, 3: Central Asia, 4: East Asia, 5: Australia, 6: Europe, 7: US and Canada, 8: Mexico, 9: Central America, 10: Colombia, 11: Peruvian Amazon. The colour composition of each bar indicates each of the subpopulations' contribution to the core genome of that isolate. A) Old world painting where only isolates from Old world areas (1–6 on map) have been used as donors in the chromosome painting. B) Global painting in which all populations have been used as donors.

doi:10.1371/journal.pgen.1006546.g002

World painting (Fig 2A). The population with the highest African ancestry is *hspAfrica1NAmerica*, isolated from 30 individuals in the US, one in Canada, one in Nicaragua, and one in Colombia, followed by *hspAfrica1Nicaragua*, which only contains isolates from Nicaragua; *hspMiscAmerica*, which consists of a number of strains of Mexican and Colombian origin; and *hspEuropeColombia*, which contains most of the Colombian isolates in our data set, and has a palette similar to hspEuropeS (Fig 1). The fifth population, hspAmerind, has a palette similar to hpEastAsia but with more hpEurope ancestry. These results are congruent to those obtained using D statistics (Table 1), which also imply that European and post-Colombian New World subpopulations are hybrids.

Table 1. D-statistics.

Population 1	Population 2	Population 3	Population 4	D-statistic
hpAfrica2	hspAfrica1WAfrica	hpAsia2	hspAfrica1NAmerica	0.538
hpAfrica2	hspAfrica1WAfrica	hpAsia2	hspAfrica1Nicaragua	0.456
hpAfrica2	hspAfrica1WAfrica	hpAsia2	hspMiscAmericas	0.454
hpAfrica2	hspAfrica1WAfrica	hpAsia2	hpEuropeColombia	0.289
hpAfrica2	hspAfrica1WAfrica	hpAsia2	hspEuropeS	0.274
hpAfrica2	hspAfrica1WAfrica	hpAsia2	hspEuropeN	0.102
hpAfrica2	hspAfrica1WAfrica	hpAsia2	hspAmerind	-0.058
hpAfrica2	hspAfrica1WAfrica	hpAsia2	hpEastAsia	-0.072

doi:10.1371/journal.pgen.1006546.t001

In our sample, several isolates from the Americas cluster within the two hpEurope subpopulations (Fig 1). The hpEurope strains from North America largely cluster with hspEuropeN while those from Central and Southern America cluster with hspEuropeS. There was also substantial variation in the proportion of the genomic palette stemming from hspAfrica1WAfrica and hspAfrica1SAfrica, both between and within New World populations. hspAfrica1WAfrica is the major African source in isolates from hspMiscAmerica, hspEuropeColombia as well as hspEuropeS while hspAfrica1SAfrica is a more important source for hspAfrica1NAmerica and hspAfrica1Nicaragua populations. A handful of isolates from both hspEuropeColombia and hspAfrica1Nicaragua populations have elevated hspAfrica1SAfrica proportions, consistent with recent genetic mixture (Fig 2A).

The distinct New World subpopulations show evidence of both drift and mixture

In the global painting, the strains from the New World populations received a large proportion from their palette from within their own subpopulation, meaning that they have differentiated both from the Old World isolates as well as from the other New World subpopulations. The formation of differentiated populations in the Americas is suggestive of recent demographic bottlenecks (see discussion below) but the New World populations have nucleotide diversity as high as or slightly higher than the Old World populations from which they evolved (Fig 3), presumably because the diversity lost in bottlenecks has been replaced by admixture.

Identifying the components of the ancestry of the New World populations that have undergone higher levels of drift provides insight into the process of differentiation. Drift is likely to be caused by the expansion of particular clones or lineages, for example, due to transmission bottlenecks. Specifically, we focused on signatures of DNA that had the most recent coalescent with other members within the same population. We tabulated the proportion of such sites with each distinct ancestry source in the Old World painting that were inferred to instead be derived from other members of their own population in the New World painting (Table 2). Bottlenecks allow small numbers of clones to propagate, leading to high rates of within population coalescence for genomes sampled from the population. This will in turn increase the proportion of sites inferred to have donors within the same population in the New World painting, rather than from Old World or other sources. Diversity acquired by admixture on the other hand, is

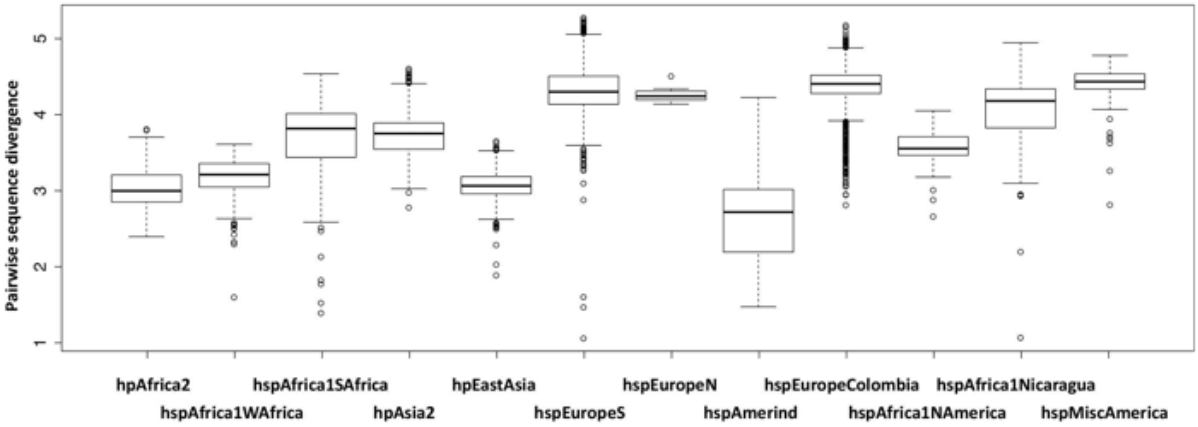


Fig 3. Pairwise sequence divergence within populations. For the two hspEurope populations only the Old World isolates are included.

doi:10.1371/journal.pgen.1006546.g003

Table 2. Proportion of ancestry assigned to each Old World population (columns) in the Old World painting that have a more recent common ancestor within the same subpopulation in the Global Painting.

	hspAfrica1-SAfrica	hspAfrica1-WAfrica	hspEuropeS	hspEuropeN	hpAsia2	hpEastAsia
hspEuropeColombia	0.43	0.44	0.48	0.48	0.46	0.39
hspAfrica1NAmerica	0.42	0.41	0.34	0.27	0.27	0.29
hspAfrica1Nicaragua	0.61	0.66	0.43	0.45	0.47	0.50
hspMiscAmerica	0.11	0.21	0.06	0.06	0.06	0.06
hspAmerind	0.46	0.45	0.44	0.47	0.50	0.53

doi:10.1371/journal.pgen.1006546.t002

more likely to be copied from other populations, unless the admixture sources have themselves been subject to a strong bottleneck.

For hspAfrica1NAmerica and hspAfrica1Nicaragua, the most drifted component is the African component. The level of drift of the African component is significantly higher than that of other components ($p < 10^{-15}$ and $p < 10^{-8}$ by Wilcoxon's rank sum test in hspAfrica1-NAmerica and hspAfrica1Nicaragua, respectively). It suggests that African lineages may have undergone rapid demographic increases during their spread in the Americas and thus that they may have a transmission advantage.

Aside from the isolation of hspAmerind strains from three countries and a single hspAfrica1NAmerica isolate from Colombian and Nicaraguan, there was no indication of sharing of ancestry between North, Central and South American gene pools. There is also no evidence from the palettes of hspAmerind having contributed DNA to any other New World strains. Amongst the Mexican isolates, a few hspMiscAmerica isolates have a substantial hspAfrica1-NAmerica component but there is no sign of elevated ancestry from the Colombian or Nicaraguan populations.

The palettes provide evidence of genetic mixture between populations within countries. The hspEuropeS isolates from Nicaragua have more hspAfrica1Nicaragua in their palette than those from other locations, while Colombian isolates that are not assigned to the hspEuropeColombia have a higher ratio of hspEuropeColombia/hspEuropeS than found elsewhere, which is consistent with recent genetic exchange. Conversely, there is no evidence for elevated hspAfrica1NAmerica ancestry in hspEuropeN isolates from North America. The hspAfrica1-NAmerica population has more hpEurope ancestry than hpAfrica1 isolates from Africa but there is little variation between strains, contrary to what would be expected if there was substantial ongoing gene flow.

Several genes have ancestral origin distinct from the overall core ancestry

The spread of *H. pylori* populations in the Americas provides an opportunity to investigate adaptive introgression as the bacteria encountered new populations of humans, as well as novel diets and environmental conditions. This is of specific interest since *H. pylori* has an outstanding capacity for recombination between co-colonising strains [2,14]. We performed a scan of the core genome for genomic regions with enrichment of specific ancestry components. To this end, we painted the strains from each New World population, using Old World strains as donors and recorded whether the donor was European, African or Asian in origin.

We found several genes where alleles showed significantly higher or lower ancestry from another Old world donor population than would be expected based on the overall ancestry of that isolate ($p < 10^{-8}$, Table 3). Among these were three genes that had ancestry from an unexpected Old World source in more than one of the New World populations. These were the

Table 3. Genes carrying position(s) with enrichment of a specific ancestry components.

Locus tag	Gene	Description	Population showing the enrichment	Enrichment	P-value*
HP0026	*gltA*	type II citrate synthase	hpEuropeColombia	Asia_high	1.4E-10
HP0042		trbI protein	hpEuropeColombia	Africa_high	1.0E-09
HP0099	*tlpA*	methyl-accepting chemotaxis protein	hspAfrica1NAmerica	Africa_low	1.2E-09
HP0160		hypothetical protein	hpEuropeColombia	Europe_high	4.1E-12
HP0216		1-deoxy-D-xylulose 5-phosphate reductoisomerase	hpEuropeColombia	Africa_high	3.9E-10
HP0252		hypothetical protein	hpEuropeColombia	Africa_low	4.3E-09
				Europe_high	2.8E-13
HP0272		hypothetical protein	hpEuropeColombia	Europe_high	3.4E-10
HP0408		hypothetical protein	hpEuropeColombia	Europe_high	2.1E-13
HP0486	*hofC*	outer membrane protein	hpEuropeColombia	Asia_high	7.3E-09
				Europe_high	1.2E-10
			hspAfrica1NAmerica	Asia_high	7.8E-15
			hspAfrica1Nicaragua	Europe_high	8.8E-11
HP0492		hypothetical protein	hpEuropeColombia	Europe_high	9.0E-09
HP0568		hypothetical protein	hpEuropeColombia	Africa_high	2.2E-09
HP0597		penicillin-binding protein 1A (PBP-1A)	hpEuropeColombia	Africa_high	9.5E-10
HP0605		hypothetical protein	hpEuropeColombia	Asia_high	8.8E-11
HP0607	*hefC*	acriflavine resistance protein	hpEuropeColombia	Africa_low	2.4E-09
HP0610		toxin-like outer membrane protein (vacA paralog)	hpEuropeColombia	Europe_high	3.1E-09
HP0667		hypothetical protein	hpEuropeColombia	Africa_high	3.5E-09
HP0872	*phnA*	alkylphosphonate uptake protein	hpEuropeColombia	Asia_high	3.9E-11
HP0913	*alpB/hopB*	outer membrane protein Omp21	hpEuropeColombia	Asia_high	9.8E-15
			hspAfrica1Nicaragua	Asia_high	1.1E-12
HP0953		hypothetical protein	hpEuropeColombia	Europe_high	2.5E-11
HP0978		cell division protein (ftsA) protein	hpEuropeColombia	Africa_low	4.6E-09
HP1055		hypothetical protein	hpEuropeColombia	Africa_high	4.8E-09
HP1086		hemolysin (tly)	hpEuropeColombia	Europe_high	2.8E-10
HP1156		hypothetical protein	hspAfrica1NAmerica	Africa_low	6.8E-09
HP1339	*exbB*	biopolymer transport protein	hpEuropeColombia	Africa_low	5.5E-09
			hpEuropeColombia	Europe_high	8.2E-15
			hspAfrica1Nicaragua	Europe_high	1.4E-12
HP1395		hypothetical protein	hpEuropeColombia	Europe_high	6.9E-10
HP1487		hypothetical protein	hpEuropeColombia	Africa_high	5.5E-09
HP1512	*frpB4*	iron-regulated outer membrane protein	hpEuropeColombia	Asia_high	3.6E-10
			hspAfrica1Nicaragua	Asia_high	3.5E-10
			hspAfrica1Nicaragua	Europe_high	4.1E-13

*the lowest P-value among polymorphic sites in a gene

doi:10.1371/journal.pgen.1006546.t003

genes encoding for AlpB (HP0913), HofC (HP0486), and FrpB4 (HP1512), which notably all are outer membrane proteins (S1 Fig) and all enriched for Asian ancestry in at least one population. The regions in *alpB* (S1A Fig) consist of clusters of 24 and 32 polymorphic sites enriched for Asian ancestry (lowest p-value 9.8 x 10^{-15}) within 49 and 65bp in hspEuropeColombia and hspAfrica1Nicaragua populations, respectively. The regions in *hofC* (S1B Fig) consist of 2 SNPs with interval 171bp and 4 successive SNPs enriched for Asian ancestry (lowest p-value 7.8 x 10^{-15}) in hspEuropeColombia and hspAfrica1NAmerica populations, respectively. The regions in *frpB4* (S1C Fig) consist of 2 successive SNPs and 26 SNPs within 156 bp enriched for Asian

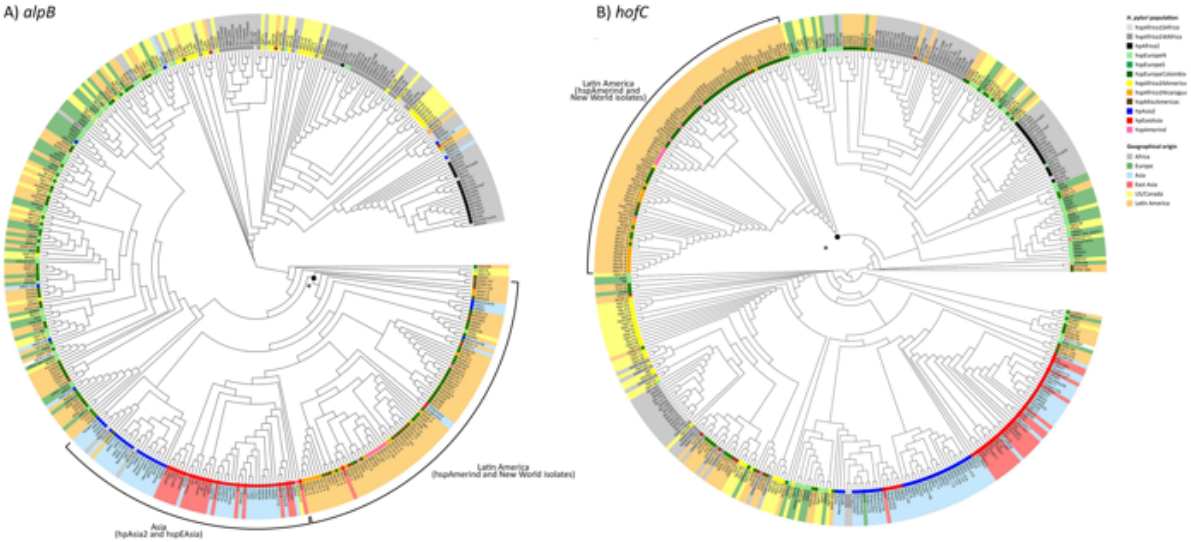


Fig 4. Maximum likelihood phylogenetic trees of alpB and hofC genes. Leaves are shaded according to geographical origin and the *H. pylori* population assignment according to the FineSTRUCTURE analysis is marked at the base of each leaf. A) AlpB. A black dot with an asterisk marks the branch for which the joint Latin American and Asian clade segregate from the others. B) HofC. The black dot with an asterisk marks the branch at which the South American clade segregates from the others.

doi:10.1371/journal.pgen.1006546.g004

ancestry (lowest p-value 3.5 x 10^{-10}) in hspEuropeColombia and hspAfrica1Nicaragua populations, respectively.

To investigate the basis of the low p values in more detail, we first constructed phylogenetic trees of the three genes. Linkage disequilibrium extends over very short distances in *H. pylori* so these trees do not necessarily reflect the genealogy of the gene as a whole. Nevertheless interesting patterns were found in *alpB* and *hofC* trees (Fig 4). For each gene at least one major separate clade of Latin American isolates could be observed, regardless of *H. pylori* population. The tree for *frpB4* can be found in S2 Fig.

For *alpB* there are three major clusters; one predominantly Asian cluster including a majority of the Latin American strains, both Amerind isolates and isolates from the New World subpopulations, one predominantly European cluster, also with a number of Latin American strains, and one African cluster where isolates from Africa group together with isolates the hspAfrica1NAmerica. Notably, in the Asian group the Latin American isolates from multiple *H. pylori* populations cluster together while in the European group they are interspersed with the other isolates (Fig 4A).

For *hofC* there is one clearly distinct South American clade, including all the Amerindian strains except for Aklavik117 and a majority of the strains belonging to the New World subpopulations hspMiscAmericas, hspAfrica1Nicaragua and hspEuropeColombia. The other three main clades are dominated by either: (*i*) hspAfrica1WAfrica, hpAfrica2 and hspAfrica1NAmerica isolates; (*ii*) hspAfrica1SAfrica, European and US/Canadian hpEurope isolates or; (*iii*) Asian isolates, respectively (Fig 4B). Notably, for *hofC* the Mexican isolates did not group within the main South American clade but within clade *i* and *ii*.

Investigating the *hofC* gene alignment in more detail using F_{st} values revealed that the sequence variation strongly contributing to the tree clade structure were nucleotides 826–926 of the gene. We found 10 nucleotide positions with a Fixation index of higher than 0.3 in the

Table 4. The ten core genome positions of highest Fst values in Latin American isolates compared to the rest of the World.

Locus tag	Gene	Description	Position in 26695	Fst
HP0486	hofC	Outer membrane protein HofC	879	0.61
HP1339	exbB	Biopolymer transport protein ExbB	112	0.61
HP0486	hofC	Outer membrane protein HofC	885	0.61
HP0486	hofC	Outer membrane protein HofC	971	0.60
HP0486	hofC	Outer membrane protein HofC	972	0.60
HP0486	hofC	Outer membrane protein HofC	967	0.59
HP0486	hofC	Outer membrane protein HofC	970	0.59
HP0486	hofC	Outer membrane protein HofC	921	0.56
HP0175	ppiC	Putative peptidyl-prolyl cis-trans isomerase PpiC	550	0.44
HP0175	ppiC	Putative peptidyl-prolyl cis-trans isomerase PpiC	636	0.44

doi:10.1371/journal.pgen.1006546.t004

Latin American isolates compared to isolates from rest of the World (S4A Fig), out of which the 8 highest were localized in the above-mentioned region. Notably, these F_{st} values were also among the highest out of all nucleotide positions in the core genome (Table 4).

Within this stretch, several amino acids were completely fixed in the South American clade and were not found in the other isolates (Fig 5). The ones with strongest F_{st} and unique to the South American clade were a Glutamic acid instead of a Glycine at position 278, Asparagine or Aspartic Acid instead of Leucine at position 280, a strong tendency to have Glycine instead of Glutamic Acid at position 292 and a Serine instead of Aspartic Acid at position 309 (Fig 5). These changes, which in most of the cases entirely changes the residue characteristics have spread to a large proportion of isolates in all of the populations found in South America, suggesting they confer an adaptive advantage, and stand out strongly in the F_{st} analyses even though this includes all Latin American isolates and not only the specific clade in the tree.

Accessory genome analysis shows similar but not identical ancestral patterns to the core genome

Our collection of multiple genomes from each population allowed us to examine patterns of gene presence and absence. A neighbour-joining tree based on gene sharing distance between isolates largely recovered the populations and sub-populations identified based on core genome sequence, but with distinct clusters for isolates carrying the Cag Pathogenicity Island (cagPAI) positive and for cagPAI negative isolates respectively (S5 Fig). The *cagPAI* is an approximately 40 kb cluster of genes encoding for a Type IV secretion system. This secretion system is translocating the CagA protein into host cells and has been shown to be of high importance for bacterial virulence [15,16].

In order to assess whether the pan genome evolved by the same processes of clonal descent and genetic exchange as the core genome, we examined the frequency of different pan genome elements in different populations. Specifically, we jointly analysed the frequency genes of triplets of populations, two of which are close representatives of the presumed ancestral source population and a third putative hybrid, with projections of the resulting 3D plots shown in Fig 6. Fig 6A shows the expectations if the pan genome of the descendent population had identical gene frequencies to either source or a 50–50 hybrid.

It has been previously shown that for the core genome, hpEurope bacteria are hybrids between hpAsia2 and hpNEAfrica (which is related to hpAfrica1), with higher hpAsia2 ancestry proportions in Northern Europe [12,17]. The same pattern for the pan genome could also be observed in our analysis, where the hpEurope population has a profile that is intermediate

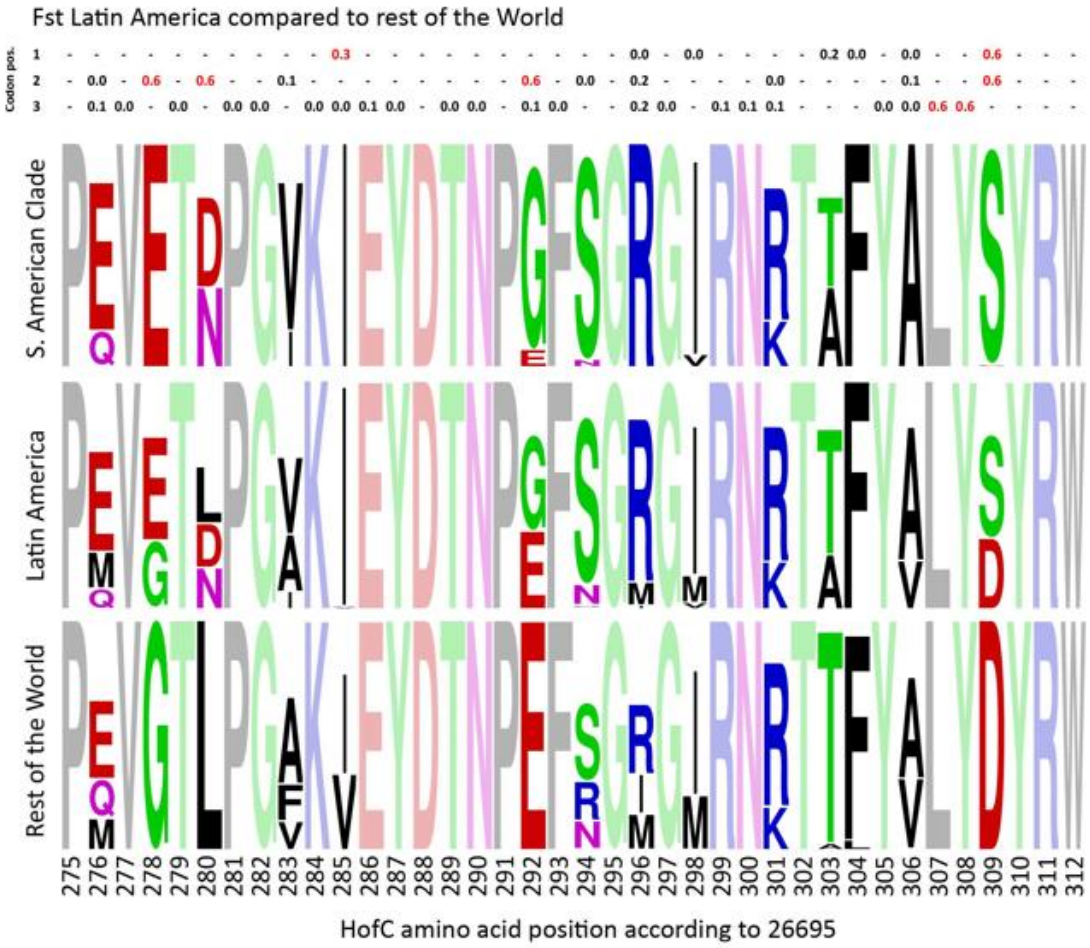


Fig 5. Fst values and WebLogo representations for amino acids 275–312 of HofC. The upper rows are Fst values for the triplets of nucleotides in each codon, calculated for isolates originating from Latin America compared to isolates from the rest of the world. WebLogo representations of the region are showed for i) the South American clade in Fig 4B, ii) All isolates from Latin America (for which the Fst was calculated), and iii) Isolates from the rest of the world. Shaded residues are synonymous in all three populations.

doi:10.1371/journal.pgen.1006546.g005

between that of hpAsia2 and hpAfrica1, but with considerable variation in the pattern amongst genes, consistent with genetic drift in the thousands of years since hybridization (Fig 6B, S1 Movie). We confirmed this visual impression using an ANOVA (S3 Table). Specifically, we tabulated the genes that differed in frequency amongst the three populations and found that the average deviations from equality were largest for genes with pattern showing either hpEurope being similar in frequency to hpAsia2 or hpEurope being similar in frequency to hpAfrica1.

For the New World populations, hspEuropeColombia has a profile that is intermediate between Africa1 and European isolates (Fig 6D, S3 Movie), with the ANOVA implying that gene frequencies are more similar to hpEurope than to hpAfrica1 (S3 Table).

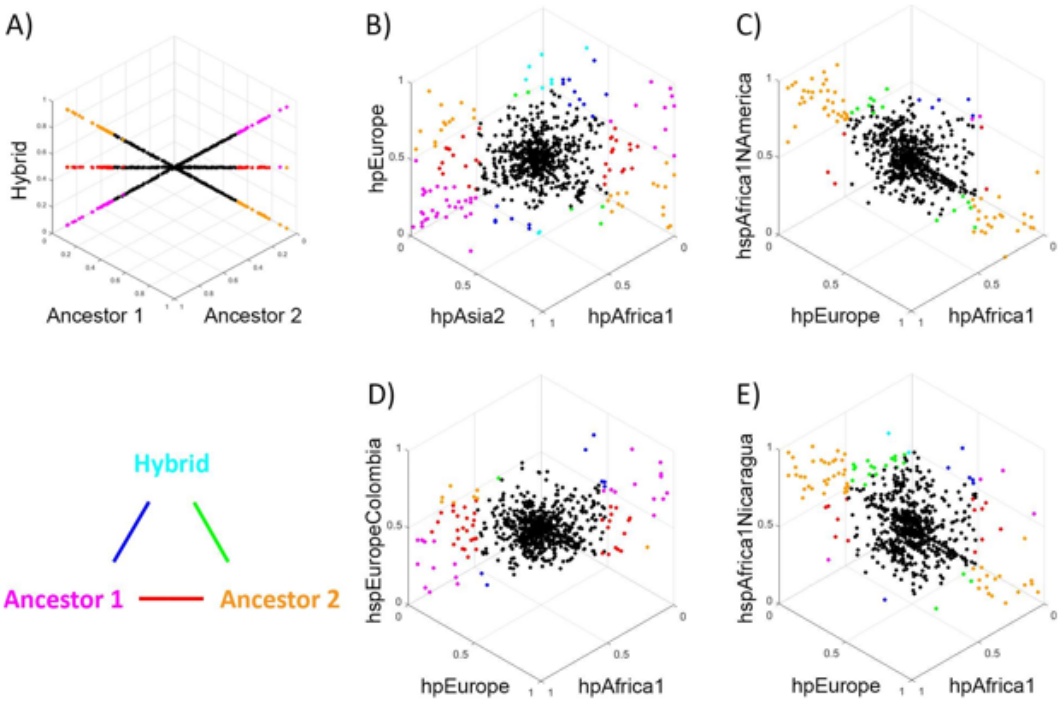


Fig 6. Comparison of accessory gene frequency in a hybrid population with the frequencies in its putative ancestors. Each dot shows the frequencies of an accessory gene in three populations, with the graphs orientated such that genes with identical frequencies in all three appear at the centre of the plot. Genes with large frequency differences between populations are labelled in colours, according to the triangular legend. Colours shown on the vertices indicate genes that differ substantially between one population and the other two (according to the criteria that X is considered substantially bigger than Y if X–Y > = 0.5, X > = 0.5 and Y < 0.1, or X > 0.9 and Y < = 0.5), while colours on the edges indicate genes where the two populations on the vertices differ substantially in frequency, with the third population having an intermediate frequency. A) Plot showing results obtained if the frequency of genes in the hybrid population is either identical to Ancestor 1 (line ending in magenta), to Ancestor 2 (line ending in orange) or a 50–50 hybrid (line ending in red). B) Comparison between Old world populations hpEurope, hpAsia2 and hpAfrica1, C) Comparison of hspEuropeColombia to hpEurope and hpAfrica, D) Comparison of hspAfrica1Nicaragua to hpEurope and hpAfrica, E) Comparison of hspAfrica1NAmerica to hpEurope and hpAfrica.

doi:10.1371/journal.pgen.1006546.g006

hspAfrica1Nicaragua and hspAfrica1NAmerica have pan genomes that are more similar to those of hpAfrica1 than hpEurope (Fig 6C–6E, S2 and S4 Movies, S3 Table).

Discussion

Millions of people from diverse geographical and ethnic backgrounds have migrated from the Old World to the Americas in the last 500 years and it is likely that a majority carried *H. pylori*. Transmission between ethnicities and DNA exchange between strains might be expected to scramble the relationship between bacterial and human ancestry at the individual level, but in the absence of selection or bottlenecks, overall *H. pylori* ancestry should largely recapitulate the ancestry found in humans [12,17,18]. Consistent with this expectation, we find diverse populations of hpEurope bacteria in Northern and Latin America, with chromosome painting profiles comparable to those found in European isolates. We find a broad North-South divide amongst hpEurope isolates, both in the New and Old World, with higher relatedness to hpAfrica1 DNA in the southern populations. This is consistent with the gene frequency cline

already observed in Europe and known differences in the colonization history of North and South America [19].

However, *H. pylori* genomic variation does not necessarily recapitulate patterns found in humans. The Americas constituted both a new physical and dietary environment and a new ethnic mix of hosts. Particular bacterial lineages may have had, or acquired, traits that adapted them to these new conditions. In extreme cases, human migrations that have little or no effect on human ancestry might precipitate substantial changes in *H. pylori* populations. For example, hspAmerind strains are rare even in populations with substantial Native American ancestry [3]. This suggests that after more than 10,000 years of independent evolution, resident *H. pylori* lineages were poorly equipped to compete with incoming lineages or with changes in the environment caused by the new settlers. We also found evidence of substantial differentiation of New World *H. pylori* populations from their ancestors, which suggests that there have been bottlenecks with particular lineages contributing disproportionately to extant populations. These bottlenecks have most strongly affected African components of ancestry (Table 2), suggesting that bacteria of African origin may have been particularly effective in colonizing the new continent.

We identified three differentiated populations in the Americas, in addition to hspAmerind. The hspAfrica1NAmerica population includes the non-European isolates from the US, also found in single Canadian, Colombian and Nicaraguan isolates. This population has an ancestry profile consistent with it being a mix of West African, South African and European sources. However, our global chromosome painting results (Fig 2B) show that within genomic regions of African origin, the DNA of hspAfrica1NAmerica is distinct from that found in modern Gambian and South African populations. Differentiation at the DNA sequence level is also found in the hspEuropeColombia and hspAfrica1Nicaragua populations, whose gene pools are distinct from each other and from those in Mexico and Europe.

The three larger groups of samples, from Mexico (Mexico City), Nicaragua (Managua) and Colombia (Bogotá) respectively, were all collected at hospitals that are tertiary referral centres for endoscopy with large catchment areas, while all but one of the US isolates came from a hospital in Cleveland, a cosmopolitan city. Therefore, our findings likely reflect broad patterns of diversity within large geographic regions. Within our sample, there are regional differences in the proportions of European, African and Amerind ancestry and wider sampling might have differentiated the picture further. Nevertheless, the distinct patterns of *H. pylori* ancestry in the four countries indicate that recent population movements have been strongly influenced by national boundaries.

H. pylori can undergo high levels of recombination during mixed infection. Over time, this might lead to bacteria acquiring an ancestry profile that reflects their local gene pool rather than their continent of origin. Recombination has not proceeded this far anywhere in the Americas and multiple populations with distinct ancestry profiles are found in most locations. hspAmerind strains have not contributed substantially to the ancestry of bacteria from any other population, but do appear to have acquired hpEurope DNA themselves. In Nicaragua and Colombia, recombination has transmitted distinctive DNA between populations, e.g. the brown shaded component in the hspEuropeS isolates from Nicaragua (Fig 2B), leading to what can informally be thought of as a national signature in the *H. pylori* DNA. There is no equivalent signal of hspAfrica1NAmerica DNA amongst the hpEurope bacteria from the US, indicating that recombination between these populations has been less extensive, and there is also no evidence within our sample of a distinctive population of hpEurope bacteria evolving within the US. Similar patterns of higher admixture in African American and Hispanic American individuals than in American individuals of European descent have been observed also on human genomic level [19].

The differences in the extent of admixture in the New World populations can have several explanations including differences in dates of colonization and extent of European and African influx/admixture in Latin America compared to the US. Another important factor can be the prevalence of infection in different areas. The prevalence of *H. pylori* infection remains high in Latin American countries, ranging from 70.1% to 84.7% of adults in a recent multi-country study [20]. In the US, the prevalence has been declining from high levels and according to data from the end of the 1990's, is around 32.5% [21]. The prevalence was different between the ethnic groups: 52.7% in non-Hispanic blacks; 61.6% in Mexican Americans and; 26.2% in non-Hispanic whites [21]. High prevalence likely entails higher occurrence of horizontal transmission and mixed infections and thus the possibility of recombination between distantly related strains [22] [23].

Our sample of Old World sources is incomplete, both in Africa and Europe, and therefore it is likely that Old World sub-populations exist that are more closely related to the New World populations than those in our sample, one such area being the Iberian peninsula. Also, even if we sample extensively in modern human groups, this may not fully reflect structure 500 years ago. The absence of sampling of close surrogates of the true ancestral subpopulations may alter our conclusions about selection or drift, which we have interpreted to have taken place in the New World rather than in the Old World. Sampling limitations for example make it unclear how much of the extensive mixture between African and European DNA observed in many Central and Southern American isolates actually took place in the Americas. Nevertheless, it is difficult to explain the local affinities within the diverse gene pools in both Nicaragua and Colombia, except by local genetic exchange. The hspAfrica1NAmerica isolates are homogeneous in their ancestry profile, suggesting that they also form a distinct gene pool that has acquired its characteristics through substantial evolution within the USA, although some of this evolution may have happened in an as yet unsampled subpopulation in Africa.

hspAfrica1NAmerica appears to be an approximately panmictic population. For example, all isolates have approximately the same level of hpEurope ancestry in Fig 1. This feature is difficult to reconcile with the low levels of genetic exchange observed with hpEurope isolates from the US. Since it has been shown that *H. pylori* from the same population (hpEastAsia) can exchange 10% of their genome during a single four year mixed infection in human [24], the ancestral pattern in US *H. pylori* implies barriers to recombination between the two populations. Such barriers may be the result of ethnic segregation and thus less diverse co-infections, of differential uptake or incorporation of DNA from different populations, or of efficient competitive exclusion of bacteria from one population by bacteria from the other within individual stomachs.

In the New World populations, four genes encoding for outer membrane proteins have sequence with ancestry that differed from that inferred for the overall core genome in more than one of the New World population. Interestingly, several of these variants were common for Latin American isolates regardless of which ancestral population they belonged to. AlpB is an adhesin binding to laminins in the extracellular matrix [25] that is present in all *H. pylori* strains [26]. Together with AlpA, it is required for colonization in experimental models and for efficient adhesion to gastric epithelial cells [27]. The HofC protein is also required for *H. pylori* colonization in mice and gerbils [28,29] but is not well characterized and little is known about its function. FrpB4 is important in the bacterial adaptation to variation in the microenvironment. FrpB4 is regulated by the levels of nickel, a micronutrient essential for *H. pylori* survival, growth and expression of virulence factors in the human stomach [30–32].

The enrichment pattern in *hofC* in a high number of the South American isolates was largely explained by the positions in region 276–309 of the 528 amino acid protein. The variants were found in all the South American Amerindian strains as well as almost all of the

hspAfrica1Nicaragua and a majority of hspEuropeColombia strains together with strains from Peru and El Salvador. No Mexican strains were found in this clade. Since the HofC protein structure and function are not characterised in detail, we are unfortunately unable to predict how these alleles contribute to the function or specificity of the protein. Interestingly, also in FrpB4 there were several positions of high Fst in Latin America compared to the rest of the world (S4 Fig) but nor in this case we are able speculate in the functional impact of these specific positions. Nevertheless, the very pronounced enrichment pattern, as well as that in the other genes, is consistent with the New World *H. pylori* having adapted to their respective human populations, allowing certain traits to propagate relative to the overall genetic background. This could be important in understanding the differences in pathogenicity in different areas and different host/bacterial interactions, suggesting a need for further investigation of the function of these proteins.

Our analysis of the accessory genome shows that *H. pylori* gene content, as well as nucleotide composition, is mixed during admixture between host populations. For example, the gene content of hpEurope is intermediate between that of hpAfrica1 and hpAsia2, but with substantial variation that may reflect the large time that has elapsed since admixture. hspEuropeColombia is more African in genome content than the average hpEurope bacteria from Europe, as would be expected because of its higher African ancestry at the nucleotide level. However, the genome content of strains from the hspAfrica1Nicaragua population is more African than would be expected given its substantial co-ancestry with hpEurope within the core genome. This observation is concordant with recent observations showing that restriction modification inhibits non homologous but not homologous recombination [33], suggesting that core genome ancestry may mix more readily between populations than accessory elements if restriction modification is an important barrier to exchange.

Our results on the population structure in the Americas sheds new light on the relationship between human migration and *H. pylori* diversity. In particular, we show that at least during human population upheavals, evolution within geographic locations is far more dynamic than the broad correlation with human genetic variation would suggest and that novel subpopulations can arise by a combination of genetic drift and admixture within hundreds of years.

Materials and methods

Helicobacter pylori whole genome sequencing data

We used both publicly available and newly sequenced genomes of *H. pylori* isolates, 401 in total (S1 Table). Nicaraguan isolates were collected at Hospital Escuela Antonio Lenin Fonseca (HEALF) in Managua, within the international collaboration "Immunological Biomarkers in Gastric Cancer development" and previously described in [34]. Colombian isolates that are not previously described were collected at the Oncology hospital (INCAN) in Bogota, and the Mexican isolates were collected at the Oncology and General Hospital in Mexico City. All three hospitals are tertiary referral centres for endoscopy and patients may thus come from other locations within the countries. For the cases we had more detailed data on the origin of the individuals, this is noted in S1 Table.

The publicly available Colombian and North American genomes were those reported in preceding studies, i.e [35–37].

Data preparation

All of the genome sequences were imported into the Bacterial Isolate Genome sequence database (BIGSdb) [38]. After this, a gene-by-gene alignment was performed using CDS sequences of the *H. pylori* 26695 strains as reference, and the alignments were exported from the

database. Both the genome sequences and the alignment are available at the public data repository Dryad (http://datadryad.org/), with doi doi:10.5061/dryad.8qp4n. We conducted SNP calling for each alignment, and imputation for polymorphic sites with missing frequency $< 1\%$ using BEAGLE [39] as our preceding study [40]. We combined in total 401350 SNPs in 1232 genes while preserving information of SNP positions in the reference genome, to prepare genome-wide haplotype data.

Population structure analysis

We inferred population structure among the strains from the genome-wide haplotype data by using the chromosome painting and fineSTRUCTURE [9], according to a procedure of our preceding study that applied them to *H. pylori* genomes [10]. Briefly, we used ChromoPainter (version 0.04) to infer chunks of DNA donated from a donor to a recipient for each recipient haplotype, and summarized the results into a "co-ancestry matrix" which contains the number of recombination-derived chunks from each donor to each recipient individual. We then ran fineSTRUCTURE (version 0.02) for 100,000 iterations of both the burn-in and Markov chain Monte Carlo (MCMC) chain in order to conduct clustering of individuals based on the co-ancestry matrix.

Principal Component Analysis was performed by applying the standard PCA implemented in Eigensoft to our data (more precisely, all biallelic data after pruning of SNPs with $r2 > 0.7$).

D-statistics were calculated by using popstats (https://github.com/pontussk/popstats) and specifying POP1 as hpAfrica2, POP2 as hspAfrica1WAfrica, POP3 as hpAsia2, and POP4 as either of the remaining 9 populations, respectively.

Stratified chromosome painting

We conducted two types of chromosome painting; "Old World chromosome painting" using only Old world isolates as donors, and "Global chromosome painting" in which each isolate is painted using all of the others. For this purpose, we used ChromoPainterV2 software [9].

To identify genomic regions with enrichment of unexpected ancestry components in the New World populations hspAfrica1NAmerica, hspAfrica1Nicaragua, and hspEuropeColombia, we conducted a novel statistical test for each of the 401350 SNPs. This was done using the Old world strains as donors, grouped into African, Asian and European geographic origin respectively.

We aim to count the number of recipient haplotypes from a certain donor population at each SNP. However, we do not observe whether a recipient i uses a particular donor population a, but instead the probability that it does at each locus l. The distribution of the total number of isolates at locus l from donor population a is ~ Poisson-Binomial(p_{lia}). If we let the genome-wide painting probability be $p_{ia} = (\sum_{l=1}^{L} p_{lia})/L$, then the distribution expected under the null that there is no local structure to the painting donors is ~ Poisson-Binomial(p_{ia}). We therefore report the p-values to test whether locus i has significantly enriched for donor a (and likewise to test for de-enrichment). We used $P<10^{-8}$ as a significance level, which corresponds to $P<0.05$ after Bonferroni correction.

Because a) the variance of a Poisson-Binomial is highest when is close to 0.5, and b) the distribution is discrete, this statistic has less power to detect high ancestry contributions from components that have high genome-wide ancestry, especially when sample size is small. In practice this has limited our power to detect regions that have an excess of African ancestry.

Phylogenetic analysis of genes with enriched ancestry

Multiple alignments of the genes were performed using MUSCLE [41] and the alignment manually inspected to remove sequences with incomplete coverage before a PhyML maximum likelihood tree was created using the SeaView software [42]. All trees were visualized using Evolview [43].

Fixation index (Fst) analysis

Fixation index (Fst) analysis was performed using the R package PopGenome [44]. For all the 1232 core-genome multiple alignments were converted to VCF format using SNP-sites [45] and site-wise Fst was calculated over all biallelic sites for the subpopulation consisting of all isolates that were geographically originating from Latin America. In total 164 358 positions in 933 of the genes were eligible for the analysis. Of those 187 positions had an Fst of more than 0.25 in the Latin American isolates compared to strains from the rest of the World (S2 Table). WebLogo plots were generated using [46].

Analysis of gene presence/absence and accessory genome

A pan-genome was constructed with all loci present in at least one of our 401 strains to examine presence/absence of all *H. pylori* genes. This pan-genome list of 2462 genes was used as queries of BLASTN against each genome analysed in this study through the BIGSdb Genome Comparator pipeline [38]. Gene presence was judged by a BLASTN match of $\geq$70% identity over $\geq$50% of the locus length [47].

Accessory presence/absence tree

The Genome Comparator Output matrix obtained with BIGSdb was used to build a distance matrix (MATLAB R2015a, The MathWorks, Inc., Natick, Massachusetts, United States). A tree was obtained using SplitsTree4 [48] and was visualised with Evolview [43].

Supporting information

S1 Fig. PCA plots describing the relationships between populations.
(TIF)

S2 Fig. P-values for enrichment of European and Asian ancestry over genes. Each dot corresponds to a polymorphic site that was tested statistically. The three genes in Table 3 satisfying significance level $p < 10^{-8}$ ($p < 0.05$ after Bonferroni correction) in more than one of the New World populations are shown. Blue symbols indicate the strength of statistical evidence for Asian enrichment and green European enrichment. Gaps represent sites where the missing frequency > 1% and sites in non-coding regions. A) alpB, B) hofC, and C) frpB4.
(TIF)

S3 Fig. Maximum likelihood phylogenetic trees of the frpB4 gene. Leaves are shaded according to geographical origin and the *H. pylori* population assignment to according to the FineSTRUCTURE analysis is marked at the base of each leaf.
(TIF)

S4 Fig. Fst over the hofC and frpB4 genes. Each dot represents a nucleotide position. For positions with Fst > 0.25 the nucleotide position in 26695 is denoted.
(TIF)

S5 Fig. Accessory genome tree. Neighbour-joining tree based on gene sharing distance (absence and presence of genes). The outer circle shows the Old World chromosome painting as in Fig 2A. Circles denote geographical origin and squares the *H. pylori* population assignment according to the FineSTRUCTURE analysis. Red stars are marking strains without the Cag Pathogenicity Island (CagPAI)
(TIF)

S1 Table. Detailed information of isolates included in the study.
(XLSX)

S2 Table. Fst values of over 0.25 in comparison of Latin American isolates with those from the rest of the World.
(XLSX)

S3 Table. Comparisons between scenarios in Fig 5B, based on hpEurope as a hybrid between hpAsia2 and hpAfrica1.
(XLSX)

S1 Movie. Comparison between Old world populations hpEurope, hpAsia2 and hpAfrica1.
(AVI)

S2 Movie. Comparison of hspEuropeColombia to hpEurope and hpAfrica.
(AVI)

S3 Movie. Comparison of hspAfrica1Nicaragua to hpEurope and hpAfrica.
(AVI)

S4 Movie. Comparison of hspAfrica1NAmerica to hpEurope and hpAfrica.
(AVI)

Acknowledgments

We thank all the researchers worldwide that have whole-genome sequenced *Helicobacter pylori* isolates and made their data available to us, either by personal connections or by making the data publicly available.

The computational calculations were done at HPC Wales, at UPPMAX (Uppsala Multidisciplinary Center for Advanced Computational Science), Sweden, and at the Human Genome Center at the Institute of Medical Science (the University of Tokyo).

Author contributions

Conceptualization: KT KY SKS DF.

Data curation: EB KT JM.

Formal analysis: KT KY EB AM.

Investigation: KT KY EB.

Methodology: DJL KY DF KT.

Project administration: KT JT.

Resources: KT IK AM MMB RS YY JT SKS CR.

Supervision: SKS DF.

Writing – original draft: KT KY EB DF.

Writing – review & editing: KT KY EB IK JT DF CR SKS.

References

1. Bianchine PJ, Russo TA (1992) The Role of Epidemic Infectious-Diseases in the Discovery of America. Allergy Proceedings 13: 225–232. PMID: 1483570
2. Suerbaum S, Josenhans C (2007) Helicobacter pylori evolution and phenotypic diversification in a changing host. Nat Rev Microbiol 5: 441–452. doi: 10.1038/nrmicro1658 PMID: 17505524
3. Kodaman N, Pazos A, Schneider BG, Piazuelo MB, Mera R, et al. (2014) Human and Helicobacter pylori coevolution shapes the risk of gastric disease. Proc Natl Acad Sci U S A 111: 1455–1460. doi: 10.1073/pnas.1318093111 PMID: 24474772
4. Montano V, Didelot X, Foll M, Linz B, Reinhardt R, et al. (2015) Worldwide Population Structure, Long-Term Demography, and Local Adaptation of Helicobacter pylori. Genetics 200: 947–963. doi: 10.1534/genetics.115.176404 PMID: 25995212
5. Maixner F, Krause-Kyora B, Turaev D, Herbig A, Hoopmann MR, et al. (2016) The 5300-year-old Helicobacter pylori genome of the Iceman. Science 351: 162–165. doi: 10.1126/science.aad2545 PMID: 26744403
6. Ferlay J. SI, Ervik M., Dikshit R., Eser S., Mathers C., Rebelo M., Parkin D.M., Forman D., Bray, F. (2013) GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. International Agency for Research on Cancer, Lyon, France.
7. Torres J, Correa P, Ferreccio C, Hernandez-Suarez G, Herrero R, et al. (2013) Gastric cancer incidence and mortality is associated with altitude in the mountainous regions of Pacific Latin America. Cancer Causes Control 24: 249–256. doi: 10.1007/s10552-012-0114-8 PMID: 23224271
8. de Sablet T, Piazuelo MB, Shaffer CL, Schneider BG, Asim M, et al. (2011) Phylogeographic origin of Helicobacter pylori is a determinant of gastric cancer risk. Gut 60: 1189–1195. doi: 10.1136/gut.2010.234468 PMID: 21357593
9. Lawson DJ, Hellenthal G, Myers S, Falush D (2012) Inference of population structure using dense haplotype data. PLoS Genet 8: e1002453. doi: 10.1371/journal.pgen.1002453 PMID: 22291602
10. Yahara K, Furuta Y, Oshima K, Yoshida M, Azuma T, et al. (2013) Chromosome painting in silico in a bacterial species reveals fine population structure. Mol Biol Evol 30: 1454–1464. doi: 10.1093/molbev/mst055 PMID: 23505045
11. Didelot X, Nell S, Yang I, Woltemate S, van der Merwe S, et al. (2013) Genomic evolution and transmission of Helicobacter pylori in two South African families. Proc Natl Acad Sci U S A 110: 13880–13885. doi: 10.1073/pnas.1304681110 PMID: 23898187
12. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, et al. (2003) Traces of human migrations in Helicobacter pylori populations. Science 299: 1582–1585. doi: 10.1126/science.1080857 PMID: 12624269
13. Moodley Y, Linz B, Bond RP, Nieuwoudt M, Soodyall H, et al. (2012) Age of the association between Helicobacter pylori and man. PLoS Pathog 8: e1002693. doi: 10.1371/journal.ppat.1002693 PMID: 22589724
14. Perez-Losada M, Browne EB, Madsen A, Wirth T, Viscidi RP, et al. (2006) Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol 6: 97–112. doi: 10.1016/j.meegid.2005.02.003 PMID: 16503511
15. Censini S, Lange C, Xiang Z, Crabtree JE, Ghiara P, et al. (1996) cag, a pathogenicity island of Helicobacter pylori, encodes type I-specific and disease-associated virulence factors. Proc Natl Acad Sci U S A 93: 14648–14653. PMID: 8962108
16. Tegtmeyer N, Wessler S, Backert S (2011) Role of the cag-pathogenicity island encoded type IV secretion system in Helicobacter pylori pathogenesis. FEBS J 278: 1190–1202. doi: 10.1111/j.1742-4658.2011.08035.x PMID: 21352489
17. Moodley Y, Linz B, Bond RP, Nieuwoudt M, Soodyall H, et al. (2012) Age of the association between Helicobacter pylori and man. PLoS Pathog 8: e1002693. Epub. doi: 10.1371/journal.ppat.1002693 PMID: 22589724
18. Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, et al. (2009) The peopling of the Pacific from a bacterial perspective. Science 323: 527–530. doi: 10.1126/science.1166083 PMID: 19164753
19. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL (2015) The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am J Hum Genet 96: 37–53. doi: 10.1016/j.ajhg.2014.11.010 PMID: 25529636

20. Porras C, Nodora J, Sexton R, Ferreccio C, Jimenez S, et al. (2013) Epidemiology of Helicobacter pylori infection in six Latin American countries (SWOG Trial S0701). Cancer Causes Control 24: 209–215. doi: 10.1007/s10552-012-0117-5 PMID: 23263777

21. Everhart JE, Kruszon-Moran D, Perez-Perez GI, Tralka TS, McQuillan G (2000) Seroprevalence and ethnic differences in Helicobacter pylori infection among adults in the United States. J Infect Dis 181: 1359–1363. doi: 10.1086/315384 PMID: 10762567

22. Schwarz S, Morelli G, Kusecek B, Manica A, Balloux F, et al. (2008) Horizontal versus familial transmission of Helicobacter pylori. PLoS Pathog 4: e1000180. doi: 10.1371/journal.ppat.1000180 PMID: 18949030

23. Ghose C, Perez-Perez GI, van Doorn LJ, Dominguez-Bello MG, Blaser MJ (2005) High frequency of gastric colonization with multiple Helicobacter pylori strains in Venezuelan subjects. J Clin Microbiol 43: 2635–2641. doi: 10.1128/JCM.43.6.2635-2641.2005 PMID: 15956377

24. Cao Q, Didelot X, Wu Z, Li Z, He L, et al. (2014) Progressive genomic convergence of two Helicobacter pylori strains during mixed infection of a patient with chronic gastritis. Gut 0: 1–8.

25. Senkovich OA, Yin J, Ekshyyan V, Conant C, Traylor J, et al. (2011) Helicobacter pylori AlpA and AlpB bind host laminin and influence gastric inflammation in gerbils. Infect Immun 79: 3106–3116. doi: 10.1128/IAI.01275-10 PMID: 21576328

26. Odenbreit S, Swoboda K, Barwig I, Ruhl S, Boren T, et al. (2009) Outer membrane protein expression profile in Helicobacter pylori clinical isolates. Infect Immun 77: 3782–3790. doi: 10.1128/IAI.00364-09 PMID: 19546190

27. Odenbreit S, Till M, Hofreuter D, Faller G, Haas R (1999) Genetic and functional characterization of the alpAB gene locus essential for the adhesion of Helicobacter pylori to human gastric tissue. Mol Microbiol 31: 1537–1548. PMID: 10200971

28. Baldwin DN, Shepherd B, Kraemer P, Hall MK, Sycuro LK, et al. (2007) Identification of Helicobacter pylori genes that contribute to stomach colonization. Infect Immun 75: 1005–1016. doi: 10.1128/IAI.01176-06 PMID: 17101654

29. Kavermann H, Burns BP, Angermuller K, Odenbreit S, Fischer W, et al. (2003) Identification and characterization of Helicobacter pylori genes essential for gastric colonization. J Exp Med 197: 813–822. doi: 10.1084/jem.20021531 PMID: 12668646

30. Davis GS, Flannery EL, Mobley HL (2006) Helicobacter pylori HP1512 is a nickel-responsive NikR-regulated outer membrane protein. Infect Immun 74: 6811–6820. doi: 10.1128/IAI.01188-06 PMID: 17030579

31. Ernst FD, Stoof J, Horrevoets WM, Kuipers EJ, Kusters JG, et al. (2006) NikR mediates nickel-responsive transcriptional repression of the Helicobacter pylori outer membrane proteins FecA3 (HP1400) and FrpB4 (HP1512). Infect Immun 74: 6821–6828. doi: 10.1128/IAI.01196-06 PMID: 17015456

32. Schauer K, Gouget B, Carriere M, Labigne A, de Reuse H (2007) Novel nickel transport mechanism across the bacterial outer membrane energized by the TonB/ExbB/ExbD machinery. Mol Microbiol 63: 1054–1068. doi: 10.1111/j.1365-2958.2006.05578.x PMID: 17238922

33. Bubendorfer S, Krebes J, Yang I, Hage E, Schulz TF, et al. (2016) Genome-wide analysis of chromosomal import patterns after natural transformation of Helicobacter pylori. Nat Commun 7: 11995. doi: 10.1038/ncomms11995 PMID: 27329939

34. Thorell K, Hosseini S, Palacios Gonzales RV, Chaotham C, Graham DY, et al. (2016) Identification of a Latin American-specific BabA adhesin variant through whole genome sequencing of Helicobacter pylori patient isolates from Nicaragua. BMC Evol Biol 16: 53. doi: 10.1186/s12862-016-0619-y PMID: 26928576

35. Blanchard TG, Czinn SJ, Correa P, Nakazawa T, Keelan M, et al. (2013) Genome sequences of 65 Helicobacter pylori strains isolated from asymptomatic individuals and patients with gastric cancer, peptic ulcer disease, or gastritis. Pathog Dis 68: 39–43. doi: 10.1111/2049-632X.12045 PMID: 23661595

36. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, et al. (2011) Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci U S A 108: 5033–5038. doi: 10.1073/pnas.1018444108 PMID: 21383187

37. Sheh A, Piazuelo MB, Wilson KT, Correa P, Fox JG (2013) Draft Genome Sequences of Helicobacter pylori Strains Isolated from Regions of Low and High Gastric Cancer Risk in Colombia. Genome Announc 1.

38. Jolley KA, Maiden MC (2010) BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11: 595. doi: 10.1186/1471-2105-11-595 PMID: 21143983

39. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84: 210–223. doi: 10.1016/j.ajhg.2009.01.005 PMID: 19200528

40. Yahara K, Didelot X, Ansari MA, Sheppard SK, Falush D (2014) Efficient inference of recombination hot regions in bacterial genomes. Mol Biol Evol 31: 1593–1605. doi: 10.1093/molbev/msu082 PMID: 24586045

41. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797. doi: 10.1093/nar/gkh340 PMID: 15034147

42. Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27: 221–224. doi: 10.1093/molbev/msp259 PMID: 19854763

43. He Z, Zhang H, Gao S, Lercher MJ, Chen WH, et al. (2016) Evolview v2: an online visualization and management tool for customized and annotated phylogenetic trees. Nucleic Acids Res.

44. Pfeifer B, Wittelsburger U, Ramos-Onsins SE, Lercher MJ (2014) PopGenome: an efficient Swiss army knife for population genomic analyses in R. Mol Biol Evol 31: 1929–1936. doi: 10.1093/molbev/msu136 PMID: 24739305

45. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, et al. (2016) SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genomics 2.

46. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. Genome Res 14: 1188–1190. doi: 10.1101/gr.849004 PMID: 15173120

47. Meric G, Yahara K, Mageiros L, Pascoe B, Maiden MC, et al. (2014) A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic Campylobacter. PLoS One 9: e92798. doi: 10.1371/journal.pone.0092798 PMID: 24676150

48. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Mol Biol Evol 23: 254–267. doi: 10.1093/molbev/msj030 PMID: 16221896

Appendix E: Table of information for strains used in Chapter 5

BIGSid	isolate name	host pathology	ClonalFrame GWAS group	bugwas GWAS group	isolation country	isolation continent	isolation city or region	hp population
637	26695_Tomb	gastritis		Non Atrophic Gastritis	UK	Europe		hspEuropeNEurope
654	ELS37	gastric cancer	Cancer duplicate 2	Gastric Cancer	El Salvador	North America		hspEuropeS
660	HPAG1	atrophic gastritis	Cancer duplicate 1	Progressive towards Cancer	Sweden	Europe	Kalixanda	hspEuropeNEurope
673	SJM180	gastritis		Non Atrophic Gastritis	Peru	South America	Lima	hspEuropeS
777	Hp_A-14	gastritis		Non Atrophic Gastritis	USA	North America	Cleveland, Ohio	hspEuropeN
781	Hp_A-26	gastritis		Non Atrophic Gastritis	USA	North America	Cleveland, Ohio	hspEuropeSEurope
782	Hp_A-27	gastritis	Non Cancer duplicate 2	Non Atrophic Gastritis	USA	North America	Cleveland, Ohio	hspEuropeN
790	Hp_H-11	gastritis		Non Atrophic Gastritis	USA	North America	Cleveland, Ohio	hspEuropeN
814	Hp_H-9	gastritis	Non Cancer duplicate 1	Non Atrophic Gastritis	USA	North America	Cleveland, Ohio	hspEuropeN
827	Hp_P-15	gastritis		Non Atrophic Gastritis	USA	North America	Cleveland, Ohio	hspEuropeN
829	Hp_P-16	gastritis	Non Cancer duplicate 2	Non Atrophic Gastritis	USA	North America	Cleveland, Ohio	hspEuropeN
832	Hp_P-23	gastritis		Non Atrophic Gastritis	USA	North America	Cleveland, Ohio	hspEuropeN
840	Hp_P-30	gastritis		Non Atrophic Gastritis	USA	North America	Cleveland, Ohio	hspEuropeN
847	Hp_P-74	gastritis		Non Atrophic Gastritis	USA	North America	Cleveland, Ohio	hspEuropeN
850	N6	gastritis	Non Cancer duplicate 2	Non Atrophic Gastritis	Burma	Asia		hspEuropeN
870	NQ4200	intestinal metaplasia	Cancer duplicate 1	Progressive towards Cancer	Colombia	South America		hspEuropeColombia
872	NQ4228	intestinal metaplasia	Cancer duplicate 2	Progressive towards Cancer	Colombia	South America		hspEuropeColombia
877	R037c	asymptomatic	Non Cancer duplicate 2	Non Atrophic Gastritis	Canada	North America	Alberta	hspEuropeN
878	R038b	asymptomatic		Non Atrophic Gastritis	Canada	North America	Alberta	hspEuropeN
879	R046Wa	asymptomatic	Non Cancer duplicate 1	Non Atrophic Gastritis	Canada	North America	Alberta	hspEuropeN
882	R32b	asymptomatic	Non Cancer duplicate 2	Non Atrophic Gastritis	Canada	North America	Alberta	hspEuropeNEurope
1346	UM037	Stomach fundus tumor	Cancer duplicate 1	Gastric Cancer	Malaysia	Asia	Kuala Lumpur	hspEuropeN
3222	HPARG63	chronic gastritis	Non Cancer duplicate 2		Argentina	South America		hspEuropeS
3602	GC11-HL	Gastric cancer	Cancer duplicate 1	Gastric Cancer	France	Europe	Bordeaux	hspEuropeN
3603	GC23-HL	Gastric cancer	Cancer duplicate 2	Gastric Cancer	France	Europe	Bordeaux	hspEuropeN
3604	GC26-HL	Gastric cancer		Gastric Cancer	France	Europe	Bordeaux	hspEuropeN
3605	GC27-HL	Gastric cancer	Cancer duplicate 1	Gastric Cancer	France	Europe	Bordeaux	hspEuropeN
3606	GC31-B	GIST	Cancer duplicate 2	Gastric Cancer	France	Europe	Bordeaux	hspEuropeN
3607	GC34-HL	Gastric cancer		Gastric Cancer	France	Europe	Bordeaux	hspEuropeN
3608	GC43-HL	Gastric cancer		Gastric Cancer	France	Europe	Bordeaux	hspEuropeN
3609	GC54-HL	Gastric cancer	Cancer duplicate 1	Gastric Cancer	France	Europe	Bordeaux	hspEuropeN
3611	GC65-HL	Gastric cancer	Cancer duplicate 2	Gastric Cancer	France	Europe	Bordeaux	hspEuropeN
3612	GC67-HL	Gastric cancer		Gastric Cancer	France	Europe	Bordeaux	hspEuropeSEurope
3613	SSR1	Antral ulcer with moderate chronic gastritis, extended intestinal metaplasia	Cancer duplicate 2		Ireland	Europe	Dublin	hspEuropeNEurope
3614	SSR2	Mild chronic gastritis, no evidence of intestinal metaplasia	Non Cancer duplicate 1		Ireland	Europe	Dublin	hspEuropeNEurope
3616	SSR4	Moderate chronic gastritis, no evidence of intestinal metaplasia	Non Cancer duplicate 2		Ireland	Europe	Dublin	hspEuropeNEurope
3617	SSR5	Moderate chronic gastritis, no evidence of intestinal metaplasia	Non Cancer duplicate 2		Ireland	Europe	Dublin	hspEuropeNEurope
3619	SSR8	Focal acute and moderate chronic inflammation	Non Cancer duplicate 2		Ireland	Europe	Dublin	hspEuropeNEurope
3620	SSR9	Moderate chronic gastritis, no evidence of intestinal metaplasia	Non Cancer duplicate 2		Ireland	Europe	Dublin	hspEuropeNEurope
3622	SSR12	Moderate chronic gastritis, focal Intestinal Metaplasia present	Cancer duplicate 2		Ireland	Europe	Dublin	hspEuropeNEurope
3623	SSR13	Moderate chronic gastritis, no evidence of intestinal metaplasia	Non Cancer duplicate 1		Ireland	Europe	Dublin	hspEuropeNEurope

BIGSid	isolate name	host pathology	ClonalFrame GWAS group	bugwas GWAS group	isolation country	isolation continent	isolation city or region	hp population
3624	SSR14	Antral ulcer with moderate chronic gastritis, extended intestinal metaplasia	Cancer duplicate 2		Ireland	Europe	Dublin	hspEuropeNEurope
3643	GC30-HL	Gastric cancer		Gastric Cancer	France	Europe	Bordeaux	hspEuropeN
3644	GC52-HL	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	France	Europe	Bordeaux	hspEuropeSEurope
3647	3800	Gastritis		Non Atrophic Gastritis	France	Europe	Bordeaux	hspEuropeSEurope
3659	3745	Gastritis		Non Atrophic Gastritis	France	Europe	Bordeaux	hspEuropeN
3662	3697	Gastritis		Non Atrophic Gastritis	France	Europe	Bordeaux	hspEuropeSEurope
3664	3699	Gastritis	Non Cancer duplicate 1	Non Atrophic Gastritis	France	Europe	Bordeaux	hspEuropeSEurope
3666	3746	Gastritis	Non Cancer duplicate 1	Non Atrophic Gastritis	France	Europe	Bordeaux	hspEuropeN
3674	GC69-HL	Gastric cancer	Cancer duplicate 2	Gastric Cancer	France	Europe	Bordeaux	hspEuropeN
4493	BM013A	Asymptomatic		Non Atrophic Gastritis	Australia	Oceania	Perth	hspEuropeN
4497	BM012A	Asymptomatic		Non Atrophic Gastritis	Australia	Oceania	Perth	hspEuropeNEurope
4542	Nic09_A	Antrum predominant gastritis but no signs of	Non Cancer duplicate 2		Nicaragua	North America	Managua	hspEuropeS
4564	Nic20_A	Intestinal metaplasia and atrophy		Progressive towards Cancer	Nicaragua	North America	Managua	hspEuropeS
8602	21580	Gastric Cancer		Gastric Cancer	Belgium	Europe		hspEuropeN
8603	30908	Control	Non Cancer duplicate 1	Non Atrophic Gastritis	Belgium	Europe		hspEuropeN
8604	30950	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	Belgium	Europe		hspEuropeN
8605	31235	Control	Non Cancer duplicate 2	Non Atrophic Gastritis	Belgium	Europe		hspEuropeN
8606	36166	Control		Non Atrophic Gastritis	Belgium	Europe		hspEuropeN
8607	38185	Gastric Cancer	Cancer duplicate 2	Gastric Cancer	Belgium	Europe		hspEuropeSEurope
8609	448	Normal		Non Atrophic Gastritis	UK	Europe	Nottingham	hspEuropeN
8610	456	Normal		Non Atrophic Gastritis	UK	Europe	Nottingham	hspEuropeN
8612	518	Normal		Non Atrophic Gastritis	UK	Europe	Nottingham	hspEuropeN
8615	HE_C1	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8616	HE_C32	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8617	HE_C33	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8618	HE_C34	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8619	HE_C38	Gastric Cancer	Cancer duplicate 2	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8620	HE_C40	Gastric Cancer	Cancer duplicate 2	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8621	HE_C50	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8622	HE_C52	Gastric Cancer	Cancer duplicate 2	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8623	HE_C55	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8624	HE_C57	Gastric Cancer	Cancer duplicate 2	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8625	HE_C58	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8626	HE_C59	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8627	HE_C66	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8628	HE_C73	Gastric Cancer	Cancer duplicate 2	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8629	HE_C9	Gastric Cancer	Cancer duplicate 2	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8630	HE_C11	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8631	HE_C13	Gastric Cancer	Cancer duplicate 2	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8632	HE_C18	Gastric Cancer	Cancer duplicate 2	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8633	HE_C23	Gastric Cancer	Cancer duplicate 2	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8634	HE_C30	Gastric Cancer	Cancer duplicate 1	Gastric Cancer	Sweden	Europe		hspEuropeNEurope
8635	HE_NC1-1	Control		Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8636	HE_NC13-6	Control		Non Atrophic Gastritis	Sweden	Europe		hspEuropeN
8637	HE_NC14-2	Control	Non Cancer duplicate 2	Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8638	HE_NC18-1	Control	Non Cancer duplicate 1	Progressive towards Cancer	Sweden	Europe		hspEuropeN
8639	HE_NC18-2	Control		Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8640	HE_NC18-4	Control	Non Cancer duplicate 1	Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8641	HE_NC13-5	Control	Non Cancer duplicate 1	Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope

Appendix E: Table of information for strains used in Chapter 5

BIGSid	isolate name	host pathology	ClonalFrame GWAS group	bugwas GWAS group	isolation country	isolation continent	isolation city or region	hp population
8642	HE_NC19-3	Control	Non Cancer duplicate 2	Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8643	HE_NC19-5	Control	Non Cancer duplicate 2	Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8644	HE_NC20-5	Control		Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8645	HE_NC1-2	Control	Non Cancer duplicate 1	Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8646	HE_NC23-2a	Control	Non Cancer duplicate 1	Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8647	HE_NC24-6	Control		Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8648	HE_NC26-4	Control	Non Cancer duplicate 1	Non Atrophic Gastritis	Sweden	Europe		hspEuropeN
8649	HE_NC27-4	Control		Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8650	HE_NC29-2	Control		Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8651	HE_NC30-2	Control		Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8652	HE_NC30-3	Control	Non Cancer duplicate 1	Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8654	HE_NC32-4	Control		Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8655	HE_NC32-5	Control		Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8656	HE_NC5-3	Control		Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8657	HE_NC36-3	Control	Non Cancer duplicate 1	Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8659	HE_NC38-2	Control	Non Cancer duplicate 1	Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8660	HE_NC38-4	Control		Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8661	HE_NC38-5	Control		Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8662	HE_NC39-3	Control		Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8663	HE_NC47-5	Control	Non Cancer duplicate 1	Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8664	HE_NC55-1	Control	Non Cancer duplicate 2	Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8665	HE_NC55-2	Control		Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8666	HE_NC55-5	Control	Non Cancer duplicate 2	Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8667	HE_NC60-1	Control		Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8668	HE_NC60-3	Control		Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8669	HE_NC61-4	Control		Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8672	HE_NC89-4	Control		Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8673	HE_NC9-1	Control	Non Cancer duplicate 2	Non Atrophic Gastritis	Sweden	Europe		hspEuropeNEurope
8674	HE_NC11-1	Control	Non Cancer duplicate 1	Progressive towards Cancer	Sweden	Europe		hspEuropeNEurope
8681	2012-26	Metaplasia	Cancer duplicate 2	Progressive towards Cancer	Mexico	North America	Mexico City	hspEuropeS
8682	22025	Gastritis	Non Cancer duplicate 1	Non Atrophic Gastritis	Colombia	South America	Boyacá, inland	hspEuropeColombia
8684	ms1055	Cancer		Gastric Cancer	Mexico	North America	Mexico City	hspEuropeS
8685	22402	Cancer	Cancer duplicate 2	Gastric Cancer	Colombia	South America	Cundinamarca, inland	hspEuropeColombia
8687	22087	Gastritis	Non Cancer duplicate 2	Non Atrophic Gastritis	Colombia	South America	Boyacá, inland	hspEuropeColombia
8690	26084	Cancer	Cancer duplicate 1	Gastric Cancer	Colombia	South America	Cundinamarca, inland	hspEuropeColombia
8692	2004-20	Metaplasia	Cancer duplicate 1	Progressive towards Cancer	Mexico	North America	Mexico City	hspEuropeS
8693	2005-98	Gastritis	Non Cancer duplicate 1	Non Atrophic Gastritis	Mexico	North America	Mexico City	hspEuropeS
8694	ms203	Gastritis	Non Cancer duplicate 1	Non Atrophic Gastritis	Mexico	North America	Mexico City	hspEuropeS
8696	26093	Cancer	Cancer duplicate 1	Gastric Cancer	Colombia	South America	Cundinamarca, inland	hspEuropeColombia
8697	ms1078	Cancer	Cancer duplicate 2	Gastric Cancer	Mexico	North America	Mexico City	hspEuropeS
8698	2006-52	Cancer	Cancer duplicate 1	Gastric Cancer	Mexico	North America	Mexico City	hspEuropeS
8700	2006-407	Gastritis	Non Cancer duplicate 2	Non Atrophic Gastritis	Mexico	North America	Mexico City	hspEuropeS
8702	22346	Metaplasia	Cancer duplicate 2	Progressive towards Cancer	Colombia	South America	Cundinamarca, inland	hspEuropeColombia
8703	ms15	Gastritis	Non Cancer duplicate 2	Non Atrophic Gastritis	Mexico	North America	Mexico City	hspEuropeS
8704	22337	Atrophic Gastritis		Progressive towards Cancer	Colombia	South America	Cundinamarca, inland	hspEuropeColombia
8705	ms23	Gastritis	Non Cancer duplicate 1	Non Atrophic Gastritis	Mexico	North America	Mexico City	hspEuropeS
8707	ms2	Gastritis		Non Atrophic Gastritis	Mexico	North America	Mexico City	hspEuropeS
8709	22341	Metaplasia	Cancer duplicate 1	Progressive towards Cancer	Colombia	South America	Cundinamarca, inland	hspEuropeColombia
8710	2006-56	Metaplasia		Progressive towards Cancer	Mexico	North America	Mexico City	hspEuropeS
8711	22023	Gastritis	Non Cancer duplicate 2	Non Atrophic Gastritis	Colombia	South America	Boyacá, inland	hspEuropeColombia

Appendix E: Table of information for strains used in Chapter 5

BIGSid	isolate name	host pathology	ClonalFrame GWAS group	bugwas GWAS group	isolation country	isolation continent	isolation city or region	hp population
8712	22327	Atrophic Gastritis	Non Cancer duplicate 2	Progressive towards Cancer	Colombia	South America	Bogota D.C.	hspEuropeColombia
8713	ms931	Cancer		Gastric Cancer	Mexico	North America	Mexico City	hspEuropeS
8714	2005-100	Gastritis		Non Atrophic Gastritis	Mexico	North America	Mexico City	hspEuropeS
8715	ms1080	Cancer	Cancer duplicate 2	Gastric Cancer	Mexico	North America	Mexico City	hspEuropeS
8716	ms13	Gastritis	Non Cancer duplicate 2	Non Atrophic Gastritis	Mexico	North America	Mexico City	hspEuropeS
8717	26100	Cancer	Cancer duplicate 1	Gastric Cancer	Colombia	South America	Caquetá, inland	hspEuropeColombia
8720	22046	Metaplasia		Progressive towards Cancer	Colombia	South America	Boyacá, inland	hspEuropeColombia
8721	2006-479	Gastritis	Non Cancer duplicate 2	Non Atrophic Gastritis	Mexico	North America	Mexico City	hspEuropeS
8724	22389	Gastritis		Non Atrophic Gastritis	Colombia	South America	Cundinamarca, inland	hspEuropeColombia
8728	22013	Metaplasia		Progressive towards Cancer	Colombia	South America	Boyacá, inland	hspEuropeColombia
8730	22362	Atrophic Gastritis		Progressive towards Cancer	Colombia	South America	Tolima, inland	hspEuropeColombia
8731	2005-126	Metaplasia	Cancer duplicate 1	Progressive towards Cancer	Mexico	North America	Mexico City	hspEuropeS
8732	2003-103	Metaplasia	Cancer duplicate 2	Progressive towards Cancer	Mexico	North America	Mexico City	hspEuropeS
8734	22367	Gastritis	Non Cancer duplicate 2	Non Atrophic Gastritis	Colombia	South America	Boyacá, inland	hspEuropeColombia
8735	22021	Metaplasia	Cancer duplicate 2	Progressive towards Cancer	Colombia	South America	Boyacá, inland	hspEuropeColombia
8737	2006-480	Metaplasia	Cancer duplicate 1	Progressive towards Cancer	Mexico	North America	Mexico City	hspEuropeS
8738	22385	Gastritis	Non Cancer duplicate 1	Non Atrophic Gastritis	Colombia	South America	Santander, inland	hspEuropeColombia
8740	2006-4	Gastritis		Non Atrophic Gastritis	Mexico	North America	Mexico City	hspEuropeSEurope
8741	22370	Gastritis	Non Cancer duplicate 1	Non Atrophic Gastritis	Colombia	South America	Boyacá, inland	hspEuropeColombia
8742	22311	Atrophic Gastritis	Non Cancer duplicate 1	Progressive towards Cancer	Colombia	South America	Bogota D.C.	hspEuropeColombia
8743	22390	Metaplasia	Cancer duplicate 2	Progressive towards Cancer	Colombia	South America	Caldas, inland	hspEuropeColombia
8744	22339	Atrophic Gastritis	Non Cancer duplicate 2	Progressive towards Cancer	Colombia	South America	Bogota D.C.	hspEuropeColombia
8746	22312	Atrophic Gastritis	Non Cancer duplicate 1	Progressive towards Cancer	Colombia	South America	Bogota D.C.	hspEuropeColombia
8747	22322	Gastritis		Non Atrophic Gastritis	Colombia	South America	Cundinamarca, inland	hspEuropeColombia
8750	22331	Metaplasia	Cancer duplicate 1	Progressive towards Cancer	Colombia	South America	Boyacá, inland	hspEuropeColombia
8751	26024	Atrophic Gastritis	Non Cancer duplicate 1	Progressive towards Cancer	Colombia	South America	Meta, inland	hspEuropeColombia
8752	22368	Metaplasia		Progressive towards Cancer	Colombia	South America	Boyacá, inland	hspEuropeColombia
8753	22360	Metaplasia		Progressive towards Cancer	Colombia	South America	Bogota D.C.	hspEuropeColombia
8754	22378	Gastritis	Non Cancer duplicate 2	Non Atrophic Gastritis	Colombia	South America	Santander, inland	hspEuropeColombia
8756	22019	Atrophic Gastritis		Progressive towards Cancer	Colombia	South America	Santander, inland	hspEuropeColombia
8758	22020	Atrophic Gastritis	Non Cancer duplicate 1	Progressive towards Cancer	Colombia	South America	Boyacá, inland	hspEuropeColombia
8760	22315	Metaplasia	Cancer duplicate 1	Progressive towards Cancer	Colombia	South America	Boyacá, inland	hspEuropeColombia
8761	22335	Atrophic Gastritis	Non Cancer duplicate 2	Progressive towards Cancer	Colombia	South America	Boyacá, inland	hspEuropeColombia
8762	ms176	Gastritis	Non Cancer duplicate 1	Non Atrophic Gastritis	Mexico	North America	Mexico City	hspEuropeS
8763	22393	Atrophic Gastritis	Non Cancer duplicate 1	Progressive towards Cancer	Colombia	South America	Cundinamarca, inland	hspEuropeColombia
8764	22093	Cancer	Cancer duplicate 2	Gastric Cancer	Colombia	South America	Boyacá, inland	hspEuropeColombia
8766	22095	Atrophic Gastritis		Progressive towards Cancer	Colombia	South America	Santander, inland	hspEuropeColombia
8768	22347	Metaplasia	Cancer duplicate 1	Progressive towards Cancer	Colombia	South America	Boyacá, inland	hspEuropeColombia
8771	2011-41	Gastritis	Non Cancer duplicate 1	Non Atrophic Gastritis	Mexico	North America	Mexico City	hspEuropeS
8773	22388	Metaplasia	Cancer duplicate 2	Progressive towards Cancer	Colombia	South America	Cundinamarca, inland	hspEuropeColombia
8775	22351	Metaplasia	Cancer duplicate 1	Progressive towards Cancer	Colombia	South America	Bogota D.C.	hspEuropeColombia
8776	22384	Gastritis	Non Cancer duplicate 2	Non Atrophic Gastritis	Colombia	South America	Boyacá, inland	hspEuropeColombia
8778	24008	Atrophic Gastritis		Progressive towards Cancer	Colombia	South America	Bolivar, caribean coast	hspEuropeColombia
8783	ms965	Cancer	Cancer duplicate 2	Gastric Cancer	Mexico	North America	Mexico City	hspEuropeS

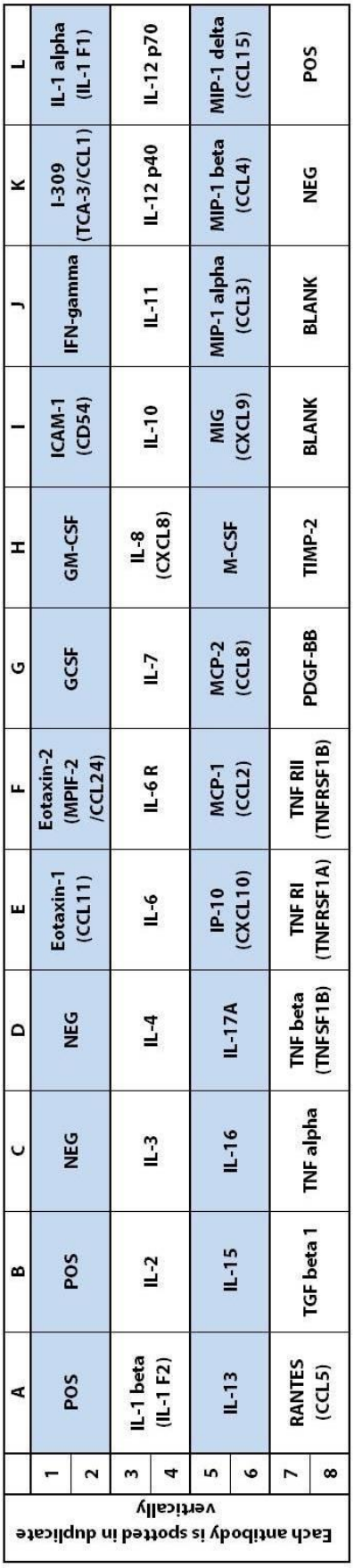
Appendix F: Table referencing all gene hits from ClonalFrame based GWAS with an association score of more than 24

Gene Tag	Gene Name	Average association score	Hit in the VFdb	Function	Functional Group associated
HP1241	*alaS*	29		Alanyl-tRNA synthetase (EC 6.1.1.7)	
HP1392		29		Fibronectin/fibrinogen-binding protein	
HP1588		27		FIG00711288: hypothetical protein	
HP1294	*rpsD*	27		SSU ribosomal protein S4p (S9e)	
HP0295	*flgL*	26	*flgL*	Flagellar hook-associated protein FlgL	motility
HP0572	*apt*	26		Adenine phosphoribosyltransferase (EC 2.4.2.7)	
HP0701	*gyrA*	26		DNA gyrase subunit A (EC 5.99.1.3)	
HP0876	*frpB*	26		putative IRON-REGULATED OUTER MEMBRANE PROTEIN	
HP0920		26		Integral membrane protein	membrane
HP1156	*hopI*	26		putative Outer membrane protein	outer-membrane protein
HP1177	*hopQ*	26	*sabB/hopO + sabA/hopP + hopZ + babA/hopS + babB/hopT*	putative Outer membrane protein	outer-membrane protein
HP1252	*oppA*	26		Oligopeptide ABC transporter, periplasmic oligopeptide-binding protein OppA (TC 3.A.1.5.1)	
HP1403	*hsdM*	26		Type I restriction-modification system, DNA-methyltransferase subunit M (EC 2.1.1.72)	
HP0116	*topA*	25		DNA topoisomerase I (EC 5.99.1.2)	
HP0289		25		toxin-like outer membrane protein	outer-membrane protein
HP0354	*dxs*	25		1-deoxy-D-xylulose 5-phosphate synthase (EC 2.2.1.7)	
HP1091	*kgtP*	25		dicarboxylic acid transporter PcaT	
HP1045	*acoE*	25		Acetyl-coenzyme A synthetase (EC 6.2.1.1)	
HP0597	*pbp1A*	25		Multimodular transpeptidase-transglycosylase (EC 2.4.1.129) (EC 3.4.-.-)	
HP0605	*hefA*	25		Probable outer membrane component of multidrug efflux pump	outer-membrane protein
HP0675	*xerC*	25		integrase/recombinase XerD	
HP0696		25		Acetone carboxylase, alpha subunit (EC 6.4.1.6)	
HP1116		25		FIG00710333: hypothetical protein	
HP0922		25		putative vacuolating cytotoxin (VacA) paralog	VacA paralog
HP1114	*uvrB*	25		Excinuclease ABC subunit B	
HP1271	*nqo12*	25		NADH-ubiquinone oxidoreductase chain L (EC 1.6.5.3)	
HP1335	*mnmA*	25		tRNA-specific 2-thiouridylase MnmA	
HP1363		25		NAD(P)HX epimerase / NAD(P)HX dehydratase	
HP1382	*nucG*	25		putative endonuclease G(EC:3.1.30.-)	
HP1503	*copA*	25		Lead, cadmium, zinc and mercury transporting ATPase (EC 3.6.3.3) (EC 3.6.3.5); Copper-translocating P-type ATPase (EC 3.6.3.4)	
HP1494	*murE*	25		UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase (EC 6.3.2.13)	
HP0068	*ureG*	24	*ureG*	Urease accessory protein UreG	acid resistance
HP0075	*glmM*	24		Phosphoglucosamine mutase (EC 5.4.2.10)	
HP0077	*prfA*	24		Peptide chain release factor 1	
HP0099	*tlpA*	24	*tlpA*	methyl-accepting chemotaxis protein (tlpA)	chemotaxis

Appendix F: Table referencing all gene hits from ClonalFrame based GWAS with an association score of more than 24

Gene Tag	Gene Name	Average association score	Hit in the VFdb	Function	Functional Group associated
HP0124	*infC*	24		Translation initiation factor 3	
HP0147	*fixP*	24		Cytochrome c oxidase subunit CcoP (EC 1.9.3.1)	
HP0234		24		INTEGRAL MEMBRANE PROTEIN (Rhomboid family)	
HP1031	*fliM*	24	*fliM*	Flagellar motor switch protein FliM	motility
HP1027	*fur*	24		Ferric uptake regulation protein FUR	acid resistance
HP0468		24		FIG00710596: hypothetical protein	
HP0503		24		FIG00710187: hypothetical protein	
HP0529		24	*virB6/cagW*	cag pathogenicity island protein (cag9)	CagPAI and type IV secretion system
HP0544	*cagE / cag23*	24	*virB4/cagE*	CAG pathogenicity island protein 23 (Protein picB)	CagPAI and type IV secretion system
HP0547	*cagA / cag26*	24	*cagA*	cag island protein, CYTOTOXICITY ASSOCIATED IMMUNODOMINANT ANTIGEN	CagPAI and type IV secretion system
HP0610		24		putative vacuolating cytotoxin (VacA) paralog	VacA paralog
HP0623	*murC*	24		UDP-N-acetylmuramate--alanine ligase (EC 6.3.2.8)	
HP0655		24		Outer membrane protein assembly factor YaeT precursor	
HP0685	*fliP*	24	*fliP*	Flagellar biosynthesis protein FliP	motility
HP0759	*dinF*	24		FIG00710308: hypothetical protein	
HP0778		24		Menaquinone via 6-amino-6-deoxyfutalosine step 1	
HP0831	*coaE*	24		Dephospho-CoA kinase (EC 2.7.1.24)	
HP0867	*lpxB*	24	*lpxB*	Lipid-A-disaccharide synthase (EC 2.4.1.182)	LPS
HP1243	*babA*	24	*babA/hopS + babB/hopT*	putative Outer membrane protein	outer-membrane protein
HP0940	*ykkK*	24		putative AMINO ACID ABC TRANSPORTER, BINDING PROTEIN PRECURSOR	
HP0946		24		FIG00711099: hypothetical protein	
HP0390	*tpx*	24		Thiol peroxidase, Tpx-type (EC 1.11.1.15)	
HP1119	*flgK*	24	*flgK*	Flagellar hook-associated protein FlgK	motility
HP1129	*exbD*	24		Biopolymer transport protein ExbD/TolR	
HP1145		24		FIG00710802: hypothetical protein	
HP1175		24		Guanine-hypoxanthine permease	
HP1179	*deoB*	24		Phosphopentomutase (EC 5.4.2.7)	
HP1329	*czcA*	24		Cobalt-zinc-cadmium resistance protein CzcA; Cation efflux system protein CusA	
HP1414		24		lojap protein	
HP1521	*res*	24		Type III restriction-modification system restriction subunit (EC 3.1.21.5)	
HP1478	*rep*	24		ATP-dependent DNA helicase UvrD/PcrA/Rep, epsilon proteobacterial type 2	
idRefC_660_1536		24		restriction enzyme BcgI alpha chain-like protein(EC:2.1.1.72)	
HP1550	*secD*	24		Protein-export membrane protein SecD (TC 3.A.5.1.1)	
HP1556	*ftsI*	24		Cell division protein FtsI [Peptidoglycan synthetase] (EC 2.4.1.129)	
HP1562	*ceuE*	24		iron(III) ABC transporter, periplasmic iron-binding protein (ceuE)	
HP1582	*pdxJ*	24	*pdxJ*	Pyridoxine 5'-phosphate synthase (EC 2.6.99.2)	motility

Appendix G: Map of the Cytokine Array used on supernatants in Chapter 6



Appendix H: Table referencing all genes highlighted in at least one chapter of this thesis

Gene Tag	Gene Name (NCBI)	description (NCBI)	Highlighted in Chapter				Reason of interest
			3	4	5	6	
HP0030		hypothetical protein				1	higher prevalence in high motility strains (21% difference)
HP0052		putative TYPE II DNA MODIFICATION ENZYME (METHYLTRANSFERASE) (Source: PATRIC)				1	higher prevalence in high motility strains (32% difference)
HP0053		hypothetical protein				1	higher prevalence in high motility strains (42% difference)
HP0054		adenine/cytosine DNA methyltransferase				1	higher prevalence in high motility strains (42% difference)
HP0068	ureG	urease accessory protein UreG			2		Hit in ClonalFrame GWAS with association score of 24 + Hit in bugwas GWAS with a p-value of 1.42E-6
HP0075	glmM	phosphoglucosamine mutase			1		Hit in ClonalFrame GWAS with association score of 24
HP0077	prfA	peptide chain release factor 1			1		Hit in ClonalFrame GWAS with association score of 24
HP0079		membrane protein				1	Lower prevalence in high IL8 AGS (60% difference)
HP0099	tlpA	methyl-accepting chemotaxis protein TlpA			1		Hit in ClonalFrame GWAS with association score of 24
HP0102		glycosyltransferase			1		Hit in bugwas GWAS with a p-value of 4.49E-6
HP0116	topA	DNA topoisomerase I			1		Hit in ClonalFrame GWAS with association score of 25
HP0124	infC	translation initiation factor IF-3			1		Hit in ClonalFrame GWAS with association score of 24
HP0147	fixP	cbb3-type cytochrome c oxidase subunit III			1		Hit in ClonalFrame GWAS with association score of 24
HP0217	cgtA	beta-1,4-N-acetylgalactosamyltransferase		1			1 PV in B38 during change of host
HP0234		membrane protein			1		Hit in ClonalFrame GWAS with association score of 24
HP0251	oppC	ABC transporter permease		1			1 PV in B38 during change of host
HP0269		tRNA-2-methylthio-N(6)-dimethylallyladenosine synthase			1		Hit in bugwas GWAS with a p-value of 5.90E-6
HP0289		toxin-like outer membrane protein			1		Hit in ClonalFrame GWAS with association score of 25
HP0290	lysA	diaminopimelate decarboxylase			1		Hit in bugwas GWAS with a p-value of 5.88E-6
HP0295	flgL	flagellar hook-associated protein FlgL			1		Hit in ClonalFrameGWAS method with average score of 26
HP0354	dxs	1-deoxy-D-xylulose-5-phosphate synthase			1		Hit in ClonalFrame GWAS with association score of 25
HP0356		hypothetical protein				1	Higher prevalence in high IL8 THP1 (60% difference)
HP0379	fucU	fucosyltransferase		1			1 SNP + small sequence repetition variation in B38 during long-term colonisation
HP0390	tpx	2-Cys peroxiredoxin			1		Hit in ClonalFrame GWAS with association score of 24
HP0437		IS605 transposase TnpA				1	Lower prevalence in high IL8 AGS (60% difference)
HP0438		IS605 transposase TnpB				1	Lower prevalence in high IL8 AGS (60% difference)
HP0462	hsdS	type I restriction-modification system specificity protein				2	Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (60% difference)
HP0464	hsdR	type I restriction-modification system endonuclease		1			1 PV and 1 SNP in B38 during change of host

Appendix H: Table referencing all genes highlighted in at least one chapter of the thesis

Gene Tag	Gene Name (NCBI)	description (NCBI)	Highlighted in Chapter				Reason of interest
			3	4	5	6	
HP0468		hypothetical protein			2		Hit in ClonalFrame GWAS with association score of 24 + Hit in bugwas GWAS with a p-value of 4.59E-7
HP0499	pldA	phospholipase A1		1			1 PV in B38 during change of host
HP0503		hypothetical protein			1	1	Hit in ClonalFrame GWAS with association score of 24 + Higher prevalence in high motility strains (21% difference)
HP0504		hypothetical protein				1	Higher prevalence in high motility strains (32% difference)
HP0520	cag1	cag pathogenicity island protein cag1				2	Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (80% Difference)
HP0522		cag pathogenicity island protein cag3				2	Higher prevalence in high motility strains (21% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0523		cag pathogenicity island protein cag4				2	Higher prevalence in high motility strains (21% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0524	virD4/ cag5	type IV secretion system protein			1	2	Hit in bugwas GWAS with a p-value of 2.53E-6 + Higher prevalence in high motility strains (21% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0525		type IV secretion system ATPase				2	Higher prevalence in high motility strains (21% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0526		Cag-specific translocation protein CagZ				2	Higher prevalence in high motility strains (21% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0527	cag7	cag pathogenicity island protein cag7			1	1	Hit in bugwas GWAS with a p-value of 2.34E-6 + Higher prevalence in high IL8 AGS (60% difference)
HP0528	cag8	cag pathogenicity island protein cag8			1	2	Hit in bugwas GWAS with a p-value of 4.54E-6 + Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0529		cag pathogenicity island protein cag9			1	2	Hit in ClonalFrame GWAS with association score of 24 + Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0530		type IV secretion system protein				2	Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% Difference)
HP0531		cag pathogenicity island protein cag11			1	2	Hit in bugwas GWAS with a p-value of 5.40E-7 + Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0532	cagT / cag12	cag pathogenicity island protein cag12			1	2	Hit in bugwas GWAS with a p-value of 3.62E-7 + Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0534		cag pathogenicity island protein cag13				2	Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% Difference)
HP0535	cag14	cag pathogenicity island protein cag14				1	Higher prevalence in high motility strains (32% difference)

Appendix H: Table referencing all genes highlighted in at least one chapter of the thesis

Gene Tag	Gene Name (NCBI)	description (NCBI)	Highlighted in Chapter				Reason of interest
			3	4	5	6	
HP0537		cag pathogenicity island protein cag16				2	Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% Difference)
HP0538		cag pathogenicity island protein cag17				2	Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% Difference)
HP0539		cag pathogenicity island protein cag18				2	Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% Difference)
HP0540	cagI / cag19	cag pathogenicity island protein cag19			1	2	Hit in bugwas GWAS with a p-value of 2.33E-6 + Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0541	cagH / cag 20	cag pathogenicity island protein cag20			1	2	Hit in bugwas GWAS with a p-value of 6.60E-7 + Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0542		cag pathogenicity island protein cag21				2	Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% Difference)
HP0543		cag pathogenicity island protein cag22				2	Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% Difference)
HP0544	cagE / cag23	type IV secretion/conjugal transfer ATPase			2	2	Hit in ClonalFrame GWAS with association score of 24 + Hit in bugwas GWAS with a p-value of 7.92E-6 + Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0545		cag pathogenicity island protein cag24				2	Higher prevalence in high motility strains (21% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0546		cag pathogenicity island protein cag25				2	Higher prevalence in high motility strains (21% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0547	cagA / cag26	cytotoxicity-associated immunodominant antigen			1	2	Hit in ClonalFrame GWAS with association score of 24 + Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (100% difference)
HP0555		membrane protein			1		Hit in bugwas GWAS with a p-value of 5.58E-8
HP0569	ychF	GTP-binding protein			1		Hit in bugwas GWAS with a p-valie of 8.86E-6
HP0572	apt	adenine phosphoribosyltransferase			1		Hit in ClonalFrame GWAS method with average score of 26
HP0593		pseudo				1	Lower prevalence in high IL8 AGS (60% difference)
HP0597	pbp1A	penicillin-binding protein 1A			1		Hit in ClonalFrame GWAS with association score of 25
HP0605	hefA	membrane protein			1		Hit in ClonalFrame GWAS with association score of 25
HP0610		toxin-like outer membrane protein			1		Hit in ClonalFrame GWAS with association score of 24
HP0615	ligA	NAD-dependent DNA ligase LigA			1		Hit in bugwas GWAS with a p-value of 4.94E-6
HP0623	murC	UDP-N-acetylmuramate--L-alanine ligase			1		Hit in ClonalFrame GWAS with association score of 24
HP0629		hypothetical protein		1			1 SNP in B47 during long-term colonisation
HP0651	fucT	fucosyltransferase		1			1 SNP + small sequence repetition variation in B38 during long-term colonisation

Gene Tag	Gene Name (NCBI)	description (NCBI)	Highlighted in Chapter				Reason of interest
			3	4	5	6	
HP1031	fliM	flagellar motor switch protein FliM			1		Hit in ClonalFrame GWAS with association score of 24
HP1041	flhA	flagellar biosynthesis protein FlhA		1			1 SNP in B47 during long-term colonisation
HP1045	acoE	acetyl-CoA synthetase			1		Hit in ClonalFrame GWAS with association score of 25
HP1046		ribosome maturation factor RimP			1		Hit in bugwas GWAS with a p-value of 2.27E-6
HP1054		hypothetical protein		1			1 SNP in B38 during change of host
HP1055		membrane protein			1		Hit in bugwas GWAS with p-value of 1.4E-9
HP1079		ATP/GTP phosphatase				1	Higher prevalence in high motility strains (21% difference)
HP1088	tktA	transketolase		1			1 SNP in B38 during change of host
HP1091	kgtP	alpha-ketoglutarate permease			1		Hit in ClonalFrame GWAS with association score of 25
HP1095		IS605 transposase TnpB					Lower prevalence in high IL8 AGS (60% difference)
HP1096		IS605 transposase TnpA					Lower prevalence in high IL8 AGS (60% difference)
HP1114	uvrB	excinuclease ABC subunit B			1		Hit in ClonalFrame GWAS with association score of 25
HP1116		hypothetical protein			1		Hit in ClonalFrame GWAS with association score of 25
HP1119	flgK	flagellar hook-associated protein FlgK			1		Hit in ClonalFrame GWAS with association score of 24
HP1129	exbD	biopolymer transport protein ExbD			1		Hit in ClonalFrame GWAS with association score of 24
HP1145		pseudo			1		Hit in ClonalFrame GWAS with association score of 24
HP1149	rimM	ribosome maturation factor RimM			1		Hit in bugwas GWAS with a p-value of 3.46E-6
HP1156	hopI	membrane protein			1		Hit in ClonalFrame GWAS method with average score of 26
HP1175		guanine permease			1		Hit in ClonalFrame GWAS method with average score of 24
HP1177	hopQ	membrane protein			2		Hit in ClonalFrame GWAS method with average score of 26 + Hit in bugwas GWAS with a p-value of 7.48E-6
HP1179	deoB	phosphopentomutase			1		Hit in ClonalFrame GWAS method with average score of 24
HP1184		multidrug transporter			1		Hit in bugwas GWAS with a p-value of 8.65E-6
HP1192		hypothetical protein				1	Higher prevalence in high motility strains (21% difference)
HP1237	pyrAa	carbamoyl phosphate synthase small subunit		1			1 SNP in B38 during change of host
HP1241	alaS	alanine--tRNA ligase			1		Hit in ClonalFrame GWAS method with average score of 29
HP1243	babA	membrane protein		1	2	2	3 SNP and 1 deletion in B38 during change of host + Hit in ClonalFrame GWAS with average score of 24 + Hit in bugwas GWAS with p-value of 3.99E-8 + Higher prevalence in high motility strains (21% prevalence) + Higher prevalence in high IL8 AGS (100% difference)
HP1251	oppB	oligopeptide ABC transporter permease OppB		1			1PV in B47 during change of host
HP1252	oppA	oligopeptide ABC transporter substrate-binding protein OppA		1	1		1 PV in B38 during change of host + Hit in ClonalFrame GWAS with association score of 26
HP1271	nqo12	NADH-quinone oxidoreductase subunit L			1		Hit in ClonalFrame GWAS method with average score of 25
HP1276		hypothetical protein				1	Higher prevalence in high IL8 AGS (60% difference)
HP1294	rpsD	30S ribosomal protein S4			1		Hit in ClonalFrame GWAS method with average score of 27

Gene Tag	Gene Name (NCBI)	description (NCBI)	Highlighted in Chapter				Reason of interest
			3	4	5	6	
HP0655		outer membrane protein assembly factor BamA			1		Hit in ClonalFrame GWAS with association score of 24
HP0675	xerC	integrase/recombinase			1		Hit in ClonalFrame GWAS with association score of 25
HP0685	fliP	flagellar biosynthesis protein FliP		1	1		1 PV in B47 during change of host + Hit in ClonalFrame GWAS with association score of 24
HP0696		acetone carboxylase subunit alpha			1		Hit in ClonalFrame GWAS with association score of 25
HP0701	gyrA	DNA gyrase subunit A			1		Hit in ClonalFrame GWAS method with average score of 26
HP0709		S-adenosyl-l-methionine hydroxide adenosyltransferase			1		Hit in bugwas GWAS with a p-value of 2.13E-7
HP0747	trmB	tRNA (guanine-N(7)-)-methyltransferase			1		Hit in bugwas GWAS with a p-value of 1.69E-7
HP0759	dinF	membrane protein			1		Hit in ClonalFrame GWAS method with average score of 24
HP0778		hypothetical protein			1		Hit in ClonalFrame GWAS method with average score of 24
HP0797		neuraminyllactose-binding hemagglutinin			1		Hit in bugwas GWAS with p-value of 2.24E-8
HP0831	coaE	dephospho-CoA kinase			1		Hit in ClonalFrame GWAS method with average score of 24
HP0855		peptidoglycan O-acetyltransferase		2			PV variation in B38 and B47 during long-term colonisation
HP0867	lpxB	lipid-A-disaccharide synthase			1		Hit in ClonalFrame GWAS method with average score of 24
HP0876	frpB	outer membrane protein			1		Hit in ClonalFrame GWAS method with average score of 26
HP0892		addiction module toxin				1	Higher prevalence in high motility strains (21% difference)
HP0893		hypothetical protein				1	Higher prevalence in high motility strains (26% difference)
HP0906		hypothetical protein			1		Hit in bugwas GWAS with a p-value of 7.08E-7
HP0920		membrane protein			1		Hit in ClonalFrame GWAS method with average score of 26
HP0922		hypothetical protein			1		Hit in ClonalFrame GWAS method with average score of 25
HP0936	proP	proline/betaine transporter ProP			1		Hit in bugwas GWAS with a p-value of 7.24E-6
HP0940	yckK	amino acid ABC transporter substrate-binding protein			1		Hit in ClonalFrame GWAS method with average score of 24
HP0946		sodium:proton antiporter			1		Hit in ClonalFrame GWAS method with average score of 24
HP0962		acyl carrier protein				1	Higher prevalence in high IL8 THP1 (60% difference)
HP0988		IS605 transposase TnpA				1	Lower prevalence in high IL8 AGS (60% difference)
HP0989		IS605 transposase TnpB				1	Lower prevalence in high IL8 AGS (60% difference)
HP0990		hypothetical protein				1	Higher prevalence in high motility strains (21% difference)
HP0997		IS605 transposase TnpB				1	Lower prevalence in high IL8 AGS (60% difference)
HP0998		IS605 transposase TnpA				1	Lower prevalence in high IL8 AGS (60% difference)
HP1003		pseudo				1	Higher prevalence in high CCL4 THP1 (60% difference)
HP1004		hypothetical protein			1	1	Hit in bugwas GWAS with a p-value of 2.73E-7 + Higher prevalence in high CCL4 THP1 (60% difference)
HP1005		hypothetical protein				1	Higher prevalence in high CCL4 THP1 (60% difference)
HP1006		conjugal transfer protein TraG				1	Higher prevalence in high CCL4 THP1 (60% difference)
HP1027	fur	ferric uptake regulation protein			1		Hit in ClonalFrame GWAS with association score of 24

Appendix H: Table referencing all genes highlighted in at least one chapter of the thesis

Gene Tag	Gene Name (NCBI)	description (NCBI)	Highlighted in Chapter				Reason of interest
			3	4	5	6	
HP1329	czcA	cation efflux system protein CzCA			1		Hit in ClonalFrame GWAS method with average score of 24
HP1331	azlC	membrane protein			1		Hit in bugwas GWAS with a p-value of 1.06E-6
HP1335	mnmA	tRNA-specific 2-thiouridylase MnmA			1		Hit in ClonalFrame GWAS method with average score of 25
HP1363		bifunctional ADP-dependent (S)-NAD(P)H-hydrate dehydratase/NAD(P)H-hydrate epimerase			1		Hit in ClonalFrame GWAS method with average score of 25
HP1365		response regulator		1			1 SNP in B38 during change of host
HP1366		type IIS restriction-modification system endonuclease				1	Higher prevalence in high motility strains (37% difference)
HP1367		type IIS restriction-modification system methyltransferase				1	Higher prevalence in high motility strains (26% difference)
HP1368		type IIS restriction-modification system methyltransferase				1	Higher prevalence in high motility strains (32% difference)
HP1382	nucG	endonuclease G			1		Hit in ClonalFrame GWAS method with average score of 25
HP1383		hypothetical protein				1	Higher prevalence in high motility strains (21% difference)
HP1392		hypothetical protein			1		Hit in ClonalFrame GWAS method with average score of 29
HP1403	hsdM	type I restriction-modification system methyltransferase			1		Hit in ClonalFrame GWAS with association score of 26
HP1414		hypothetical protein			1		Hit in ClonalFrame GWAS with association score of 24
HP1421	virB11_2	type IV secretion system ATPase			1		Hit in bugwas GWAS with a p-value of 2.26E-6
HP1433		hypothetical protein				1	Higher prevalence in high motility strains (21% difference)
HP1438		hypothetical protein				1	Higher prevalence in high motility strains (32% difference)
HP1460	dnaE	DNA polymerase III subunit alpha			1		Hit in bugwas GWAS with a p-value of 9.73E-6
HP1478	rep	DNA helicase II UvrD			1		Hit in ClonalFrame GWAS with association score of 24
HP1494	murE	UDP-N-acetylmuramoylalanyl-D-glutamate--2,6-diaminopimelate ligase			1		Hit in ClonalFrame GWAS with association score of 25
HP1499		restriction endonuclease				1	Lower prevalence in high IL8 AGS (60% difference)
HP1503	copA	cation-transporting ATPase			1		Hit in ClonalFrame GWAS with association score of 25
HP1517		hypothetical protein				1	Lower prevalence in high CCL4 THP1 (60% difference)
HP1518		hypothetical protein				1	Higher prevalence in high CCL4 THP1 (60% difference)
HP1519		pseudo				1	Higher prevalence in high IL8 AGS (60% difference)
HP1520		hypothetical protein				1	Higher prevalence in high IL8 AGS (60% difference)
HP1521	res	type III restriction-modification system endonuclease			1		Hit in ClonalFrame GWAS with association score of 24
HP1534		IS605 transposase TnpB				1	Lower prevalence in high IL8 AGS (60% difference)
HP1535		IS605 transposase TnpA				1	Lower prevalence in high IL8 AGS (60% difference)
HP1550	secD	preprotein translocase subunit SecD			1		Hit in ClonalFrame GWAS with association score of 24
HP1556	ftsI	cell division protein FtsI			1		Hit in ClonalFrame GWAS with association score of 24

Appendix H: Table referencing all genes highlighted in at least one chapter of the thesis

Gene Tag	Gene Name (NCBI)	description (NCBI)	Highlighted in Chapter				Reason of interest
			3	4	5	6	
HP1562	ceuE	iron(III) ABC transporter substrate-binding protein CeuE			1		Hit in ClonalFrame GWAS with association score of 24
HP1572	dniR	lytic transglycosylase			1		Hit in bugwas GWAS with a p-value of 5.90E-6
HP1582	pdxJ	pyridoxine 5'-phosphate synthase			1		Hit in ClonalFrame GWAS with association score of 24
HP1588		hypothetical protein			1		Hit in ClonalFrame GWAS method with average score of 27
0010_8940_0104				1			1 PV in B38 during long-term colonisation
idRefC_660_1536					1		Hit in ClonalFrame GWAS method with average score of 24
009_8_0803						1	Higher prevalence in high motility strains (47% difference)
004_3_0002						1	Higher prevalence in high motility strains (26% difference)
003_2_0285						1	Higher prevalence in high motility strains (26% difference)
003_2_0286						1	Higher prevalence in high motility strains (26% difference)
005_4_1517						1	Higher prevalence in high motility strains (26% difference)
009_8_1030						1	Higher prevalence in high motility strains (26% difference)
0010_9_0525						2	Higher prevalence in high motility strains (26% difference) + Higher prevalence in high IL8 AGS (60% difference)
002_1_0820						1	Higher prevalence in high motility strains (21% difference)
0056_9583_1019						1	Higher prevalence in high motility strains (21% difference)
0041_3645_0874						1	Higher prevalence in high motility strains (21% difference)
003_2_1052						1	Higher prevalence in high motility strains (21% difference)
003_2_1053						1	Higher prevalence in high motility strains (21% difference)
004_3_0003						1	Higher prevalence in high motility strains (21% difference)
004_3_1156						1	Higher prevalence in high motility strains (21% difference)
004_3_0229						1	Higher prevalence in high motility strains (21% difference)
002_1_0516						1	Higher prevalence in high IL8 AGS (60% difference)
008_7_0302						1	Higher prevalence in high IL8 AGS (60% difference)
009_8_0878						1	Higher prevalence in high IL8 AGS (60% difference)
006_5_0442						1	Higher prevalence in high IL8 AGS (60% difference)
009_8_0877						1	Higher prevalence in high IL8 AGS (60% difference)
006_5_0441						1	Higher prevalence in high IL8 AGS (60% difference)
006_5_0440						1	Higher prevalence in high IL8 AGS (60% difference)
0057_9584_0148						1	Higher prevalence in high IL8 AGS (60% difference)
002_1_0250						1	Lower prevalence in high IL8 AGS (80% difference)
0049_3664_1623						1	Lower prevalence in high IL8 AGS (60% difference)
005_4_0396						1	Higher prevalence in high IL8 THP1 (80% difference)
0030_3611_0746						1	Higher prevalence in high IL8 THP1 (60% difference)
0023_3598_0948						1	Higher prevalence in high IL8 THP1 (60% difference)
006_5_1019						1	Higher prevalence in high IL8 THP1 (60% difference)
003_2_1164						1	Higher prevalence in high CCL4 THP1 (80% difference)
0032_3617_1050						1	Higher prevalence in high CCL4 THP1 (60% difference)
002_1_1035						1	Higher prevalence in high CCL4 THP1 (60% difference)

Appendix H: Table referencing all genes highlighted in at least one chapter of the thesis

Gene Tag	Gene Name (NCBI)	description (NCBI)	Highlighted in Chapter				Reason of interest
			3	4	5	6	
004_3_0774						1	Higher prevalence in high CCL4 THP1 (60% difference)
0035_3636_0533						1	Lower prevalence in low CCL4 THP1 (60% difference)
0026_3605_0564						1	Lower prevalence in low CCL4 THP1 (60% difference)
0054_8605_1069						1	Lower prevalence in low CCL4 THP1 (60% difference)
0015_3587_0394						1	Lower prevalence in low CCL4 THP1 (60% difference)
0015_3587_0124						1	Lower prevalence in low CCL4 THP1 (60% difference)
009_8_0730						1	Lower prevalence in low CCL4 THP1 (60% difference)
009_8_0741						1	Lower prevalence in low CCL4 THP1 (60% difference)
0015_3587_0306						1	Lower prevalence in low CCL4 THP1 (60% difference)
009_8_0737						1	Lower prevalence in low CCL4 THP1 (60% difference)
009_8_0732						1	Lower prevalence in low CCL4 THP1 (60% difference)
0015_3587_1081						1	Lower prevalence in low CCL4 THP1 (60% difference)
0017_3589_1266						1	Lower prevalence in low CCL4 THP1 (60% difference)
0015_3587_0433						1	Lower prevalence in low CCL4 THP1 (60% difference)
0054_8605_0484						1	Lower prevalence in low CCL4 THP1 (60% difference)

References

"A Clinical Evaluation of the International Lymphoma Study Group Classification of Non-Hodgkin's Lymphoma. The Non-Hodgkin's Lymphoma Classification Project." 1997. *Blood* 89 (11):3909–18. http://www.ncbi.nlm.nih.gov/pubmed/9166827.

Abadi, Amin Talebi Bezmin, Ashraf Mohhabati Mobarez, Marc JM Bonten, Jaap A Wagenaar, and Johannes G Kusters. 2014. "Clinical Relevance of the cagA, tnpA and tnpB Genes in Helicobacter Pylori." *BMC Gastroenterology* 14 (1):33. https://doi.org/10.1186/1471-230X-14-33.

Achtman, Mark, Takeshi Azuma, Douglas E. Berg, Yoshiyuki Ito, Giovanna Morelli, Zhi-Jun Pan, Sebastian Suerbaum, Stuart A. Thompson, Arie van der Ende, and Leen-Jan van Doorn. 1999. "Recombination and Clonal Groupings within Helicobacter Pylori from Different Geographical Regions." *Molecular Microbiology* 32 (3):459–70. https://doi.org/10.1046/j.1365-2958.1999.01382.x.

Ahmadzadeh, A., H. Ghalehnoei, N. Farzi, A. Yadegar, M. Alebouyeh, H.A. Aghdaei, M. Molaei, M.R. Zali, and M.A. pour Hossein Gholi. 2015. "Association of Cag PAI Integrity with Severeness of Helicobacter Pylori Infection in Patients with Gastritis." *Pathologie Biologie* 63 (6):252–57. https://doi.org/10.1016/j.patbio.2015.09.004.

Akopyants, N S, S W Clifton, D Kersulyte, J E Crabtree, B E Youree, C A Reece, N O Bukanov, E S Drazek, B A Roe, and D E Berg. 1998. "Analyses of the Cag Pathogenicity Island of Helicobacter Pylori." *Molecular Microbiology* 28 (1):37–53. http://www.ncbi.nlm.nih.gov/pubmed/9593295.

Alam, Md Tauqeer, Robert A Petit, Emily K Crispell, Timothy A Thornton, Karen N Conneely, Yunxuan Jiang, Sarah W Satola, and Timothy D Read. 2014. "Dissecting Vancomycin-Intermediate Resistance in Staphylococcus Aureus Using Genome-Wide Association." *Genome Biology and Evolution* 6 (5):1174–85. https://doi.org/10.1093/gbe/evu092.

Algood, Holly M Scott, and Timothy L Cover. 2006. "Helicobacter Pylori Persistence: An Overview of Interactions between H. Pylori and Host Immune Defenses." *Clinical Microbiology Reviews* 19 (4). American Society for Microbiology (ASM):597–613. https://doi.org/10.1128/CMR.00006-06.

Alm, R A, and T J Trust. 1999. “Analysis of the Genetic Diversity of Helicobacter Pylori: The Tale of Two Genomes.” *Journal of Molecular Medicine (Berlin, Germany)* 77 (12):834–46. http://www.ncbi.nlm.nih.gov/pubmed/10682319.

Altschul, S F, W Gish, W Miller, E W Myers, and D J Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Ameri Shah Reza, Mahdieh, Seyed Latif Mousavi Gargari, Iraj Rasooli, Mohammadreza Jalali Nadoushan, and Walead Ebrahimizadeh. 2012. “Inhibition of H. Pylori Colonization and Prevention of Gastritis in Murine Model.” *World Journal of Microbiology & Biotechnology* 28 (7):2513–19. https://doi.org/10.1007/s11274-012-1059-5.

Andersson, Anders F., Mathilda Lindberg, Hedvig Jakobsson, Fredrik Bäckhed, Pål Nyrén, and Lars Engstrand. 2008. “Comparative Analysis of Human Gut Microbiota by Barcoded Pyrosequencing.” Edited by Niyaz Ahmed. *PLoS ONE* 3 (7). Public Library of Science:e2836. https://doi.org/10.1371/journal.pone.0002836.

Appelmelk, B J, S L Martin, M A Monteiro, C A Clayton, A A McColm, P Zheng, T Verboom, et al. 1999. “Phase Variation in Helicobacter Pylori Lipopolysaccharide due to Changes in the Lengths of poly(C) Tracts in alpha3-Fucosyltransferase Genes.” *Infection and Immunity* 67 (10):5361–66. http://www.ncbi.nlm.nih.gov/pubmed/10496917.

Arnim, U. von, T. Wex, A. Link, M. Messerschmidt, M. Venerito, S. Miehlke, and P. Malfertheiner. 2016. “*Helicobacter Pylori* Infection Is Associated with a Reduced Risk of Developing Eosinophilic Oesophagitis.” *Alimentary Pharmacology & Therapeutics* 43 (7):825–30. https://doi.org/10.1111/apt.13560.

Arnold, Isabelle C., Nina Dehzad, Sebastian Reuter, Helen Martin, Burkhard Becher, Christian Taube, and Anne Müller. 2011. “Helicobacter Pylori Infection Prevents Allergic Asthma in Mouse Models through the Induction of Regulatory T Cells.” *Journal of Clinical Investigation* 121 (8):3088–93. https://doi.org/10.1172/JCI45041.

Asano, Naoki, Katsunori Iijima, Tomoyuki Koike, Akira Imatani, and Tooru Shimosegawa. 2015. “*Helicobacter Pylori* -Negative Gastric Mucosa-Associated Lymphoid Tissue Lymphomas: A Review.” *World Journal of Gastroenterology* 21 (26):8014. https://doi.org/10.3748/wjg.v21.i26.8014.

Asim, Mohammad, Surendra K Chikara, Arpita Ghosh, Srinivas Vudathala, Judith Romero-Gallo, Uma S Krishna, Keith T Wilson, Dawn A Israel, Richard M Peek, and Rupesh Chaturvedi. 2015. "Draft Genome Sequence of Gerbil-Adapted Carcinogenic Helicobacter Pylori Strain 7.13." *Genome Announcements* 3 (3). https://doi.org/10.1128/genomeA.00641-15.

Aspholm-Hurtig, Marina, Giedrius Dailide, Martina Lahmann, Awdhesh Kalia, Dag Ilver, Niamh Roche, Susanne Vikström, et al. 2004. "Functional Adaptation of BabA, the H. Pylori ABO Blood Group Antigen Binding Adhesin." *Science (New York, N.Y.)* 305 (5683):519–22. https://doi.org/10.1126/science.1098801.

Aspinall, Gerald O., and Mario A. Monteiro. 1996. "Lipopolysaccharides of *Helicobacter Pylori* Strains P466 and MO19: Structures of the O Antigen and Core Oligosaccharide Regions †." *Biochemistry* 35 (7):2498–2504. https://doi.org/10.1021/bi951853k.

Atherton, J C, P Cao, R M Peek, M K Tummuru, M J Blaser, and T L Cover. 1995. "Mosaicism in Vacuolating Cytotoxin Alleles of Helicobacter Pylori. Association of Specific vacA Types with Cytotoxin Production and Peptic Ulceration." *The Journal of Biological Chemistry* 270 (30):17771–77. http://www.ncbi.nlm.nih.gov/pubmed/7629077.

Atherton, John C. 2006. "THE PATHOGENESIS OF *HELICOBACTER PYLORI* – INDUCED GASTRO-DUODENAL DISEASES." *Annual Review of Pathology: Mechanisms of Disease* 1 (1):63–96. https://doi.org/10.1146/annurev.pathol.1.110304.100125.

Avasthi, Tiruvayipati Suma, Singamaneni Haritha Devi, Todd D Taylor, Narender Kumar, Ramani Baddam, Shinji Kondo, Yutaka Suzuki, Hervé Lamouliatte, Francis Mégraud, and Niyaz Ahmed. 2011. "Genomes of Two Chronological Isolates (Helicobacter Pylori 2017 and 2018) of the West African Helicobacter Pylori Strain 908 Obtained from a Single Patient." *Journal of Bacteriology* 193 (13):3385–86. https://doi.org/10.1128/JB.05006-11.

Banerjee, Arun, and Desirazu N Rao. 2011. "Functional Analysis of an Acid Adaptive DNA Adenine Methyltransferase from Helicobacter Pylori 26695." Edited by Shuang-yong Xu. *PloS One* 6 (2):e16810. https://doi.org/10.1371/journal.pone.0016810.

Bauer, Bianca, Ervinna Pang, Carsten Holland, Mirjana Kessler, Sina Bartfeld, and Thomas F. Meyer. 2012. "The Helicobacter Pylori Virulence Effector CagA

Abrogates Human β-Defensin 3 Expression via Inactivation of EGFR Signaling." *Cell Host & Microbe* 11 (6):576–86. https://doi.org/10.1016/j.chom.2012.04.013.

Becker, Karl-Friedrich, Michael J. Atkinson, Ulrike Reich, Ingrid Becker, Hjalmar Nekarda, Jörg R. Siewert, and Heinz Höfler. 1994. "E-Cadherin Gene Mutations Provide Clues to Diffuse Type Gastric Carcinomas." *Cancer Research* 54 (14).

Bergman, Mathijs, Gianfranco Del Prete, Yvette van Kooyk, and Ben Appelmelk. 2006. "Helicobacter Pylori Phase Variation, Immune Modulation and Gastric Autoimmunity." *Nature Reviews Microbiology* 4 (2):151–59. https://doi.org/10.1038/nrmicro1344.

Berthenet, Elvire, Sam Sheppard, and Filipa F. Vale. 2016. "Recent 'omics' advances in Helicobacter Pylori." *Helicobacter* 21 (September):14–18. https://doi.org/10.1111/hel.12334.

Bessède, E, P Dubus, F Mégraud, and C Varon. 2015. "Helicobacter Pylori Infection and Stem Cells at the Origin of Gastric Cancer." *Oncogene* 34 (20):2547–55. https://doi.org/10.1038/onc.2014.187.

Bik, Elisabeth M, Paul B Eckburg, Steven R Gill, Karen E Nelson, Elizabeth A Purdom, Fritz Francois, Guillermo Perez-Perez, Martin J Blaser, and David A Relman. 2006. "Molecular Analysis of the Bacterial Microbiota in the Human Stomach." *Proceedings of the National Academy of Sciences of the United States of America* 103 (3):732–37. https://doi.org/10.1073/pnas.0506655103.

Binh, Tran Thanh, Rumiko Suzuki, Tran Thi Huyen Trang, Dong Hyeon Kwon, and Yoshio Yamaoka. 2015. "Search for Novel Candidate Mutations for Metronidazole Resistance in Helicobacter Pylori Using next-Generation Sequencing." *Antimicrobial Agents and Chemotherapy* 59 (4):2343–48. https://doi.org/10.1128/AAC.04852-14.

Blanchard, Thomas G, and John G Nedrud. 2012. "Laboratory Maintenance of Helicobacter Species." *Current Protocols in Microbiology* Chapter 8 (February):Unit8B.1. https://doi.org/10.1002/9780471729259.mc08b01s24.

Blaser, Martin J., and John C. Atherton. 2004. "Helicobacter Pylori Persistence: Biology and Disease." *Journal of Clinical Investigation* 113 (3):321–33. https://doi.org/10.1172/JCI20925.

Boncristiano, Marianna, Silvia Rossi Paccani, Silvia Barone, Cristina Ulivieri, Laura Patrussi, Dag Ilver, Amedeo Amedei, Mario Milco D'Elios, John L. Telford, and

Cosima T. Baldari. 2003. “The *Helicobacter Pylori* Vacuolating Toxin Inhibits T Cell Activation by Two Independent Mechanisms.” *The Journal of Experimental Medicine* 198 (12):1887–97. https://doi.org/10.1084/jem.20030621.

Bradford Hill, Austin. 1965. “The Environment and Disease: Association or Causation?” *Proceedings of the Royal Society of Medicine* 58 (5). Royal Society of Medicine Press:295–300. http://www.ncbi.nlm.nih.gov/pubmed/14283879.

Brandt, S., T. Kwok, R. Hartig, W. Konig, and S. Backert. 2005. “NF- B Activation and Potentiation of Proinflammatory Responses by the Helicobacter Pylori CagA Protein.” *Proceedings of the National Academy of Sciences* 102 (26):9300–9305. https://doi.org/10.1073/pnas.0409873102.

Breed, Robert S, and W D Dotterrer. 1916. “THE NUMBER OF COLONIES ALLOWABLE ON SATISFACTORY AGAR PLATES.” https://www.ncbi.nlm.nih.gov/pmc/articles/PMC378655/pdf/jbacter01098-0078.pdf.

Brückner, Markus, Philipp Lenz, Tobias M Nowacki, Friederike Pott, Dirk Foell, and Dominik Bettenworth. 2014. “Murine Endoscopy for in Vivo Multimodal Imaging of Carcinogenesis and Assessment of Intestinal Wound Healing and Inflammation.” *Journal of Visualized Experiments : JoVE*, no. 90(August). MyJoVE Corporation. https://doi.org/10.3791/51875.

Buffart, TE, B Carvalho, E Hopmans, V Brehm, E Klein Kranenbarg, TBM Schaaij-Visser, PP Eijk, et al. 2007. “Gastric Cancers in Young and Elderly Patients Show Different Genomic Profiles.” *The Journal of Pathology* 211 (1):45–51. https://doi.org/10.1002/path.2085.

Buffart, Tineke E, Melanie Louw, Nicole CT van Grieken, Marianne Tijssen, Beatriz Carvalho, Bauke Ylstra, Heike Grabsch, et al. 2011. “Gastric Cancers of Western European and African Patients Show Different Patterns of Genomic Instability.” *BMC Medical Genomics* 4 (1):7. https://doi.org/10.1186/1755-8794-4-7.

Calam, J. 1998. “Clinical Science of Helicobacter Pylori Infection: Ulcers and NSAIDs.” *British Medical Bulletin* 54 (1):55–62. http://www.ncbi.nlm.nih.gov/pubmed/9604430.

Camargo, M Constanza, Kyoung-Mee Kim, Keitaro Matsuo, Javier Torres, Linda M Liao, Douglas R Morgan, Angelika Michel, et al. 2016. “Anti-Helicobacter Pylori Antibody Profiles in Epstein-Barr Virus (EBV)-Positive and EBV-Negative Gastric Cancer.” *Helicobacter* 21 (2):153–57.

https://doi.org/10.1111/hel.12249.

"Canadian Cancer Society." 2017. "Anatomy and Physiology of the Stomach - Canadian Cancer Society." 2017. http://www.cancer.ca/en/cancer-information/cancer-type/stomach/stomach-cancer/the-stomach/?region=on.

"Cancer Research UK." 2017a. "Stages of Stomach Cancer | Cancer Research UK." 2017. http://www.cancerresearchuk.org/about-cancer/stomach-cancer/stages.

Cancer Research UK. 2017b. "Stomach Cancer Statistics | Cancer Research UK." 2017. http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/stomach-cancer#heading-One.

"Cancer Research UK." 2017c. "Survival for Stomach Cancer | Cancer Research UK." 2017. http://www.cancerresearchuk.org/about-cancer/stomach-cancer/survival.

Cao, Qizhi, Xavier Didelot, Zhongbiao Wu, Zongwei Li, Lihua He, Yunsheng Li, Ming Ni, et al. 2015. "Progressive Genomic Convergence of Two Helicobacter Pylori Strains during Mixed Infection of a Patient with Chronic Gastritis." *Gut* 64 (4):554–61. https://doi.org/10.1136/gutjnl-2014-307345.

Capelle, Lisette G., Annemarie C. de Vries, Jelle Haringsma, Frank Ter Borg, Richard A. de Vries, Marco J. Bruno, Herman van Dekken, Jos Meijer, Nicole C.T. van Grieken, and Ernst J. Kuipers. 2010. "The Staging of Gastritis with the OLGA System by Using Intestinal Metaplasia as an Accurate Alternative for Atrophic Gastritis." *Gastrointestinal Endoscopy* 71 (7):1150–58. https://doi.org/10.1016/j.gie.2009.12.029.

Carlsohn, Elisabet, Johanna Nyström, Ingrid Bölin, Carol L Nilsson, and Ann-Mari Svennerholm. 2006. "HpaA Is Essential for Helicobacter Pylori Colonization in Mice." *Infection and Immunity* 74 (2):920–26. https://doi.org/10.1128/IAI.74.2.920-926.2006.

Carneiro, F. 2012. "Hereditary Gastric Cancer." *Der Pathologe* 33 (S2):231–34. https://doi.org/10.1007/s00292-012-1677-6.

Carraher, Sally, Hsiu-Ju Chang, Rachel Munday, Karen J Goodman, and the CaNHelp Working CANHelp Working Group. 2013. "Helicobacter Pylori Incidence and Re-Infection in the Aklavik H. Pylori Project." *International Journal of Circumpolar Health* 72. Taylor & Francis. https://doi.org/10.3402/ijch.v72i0.21594.

Censini, S, C Lange, Z Xiang, J E Crabtree, P Ghiara, M Borodovsky, R Rappuoli,

and A Covacci. 1996. “Cag, a Pathogenicity Island of Helicobacter Pylori, Encodes Type I-Specific and Disease-Associated Virulence Factors.” *Proceedings of the National Academy of Sciences of the United States of America* 93 (25):14648–53. http://www.ncbi.nlm.nih.gov/pubmed/8962108.

Chamberlain, C E, and D A Peura. 1990. “Campylobacter (Helicobacter) Pylori. Is Peptic Disease a Bacterial Infection?” *Archives of Internal Medicine* 150 (5):951–55. http://www.ncbi.nlm.nih.gov/pubmed/2184791.

Chen, Peter E, and B Jesse Shapiro. 2015. “The Advent of Genome-Wide Association Studies for Bacteria.” *Current Opinion in Microbiology* 25 (June):17–24. https://doi.org/10.1016/j.mib.2015.03.002.

Chen, Wanqing, Rongshou Zheng, Peter D Baade, Siwei Zhang, Hongmei Zeng, Freddie Bray, Ahmedin Jemal, Xue Qin Yu, and Jie He. 2016. “Cancer Statistics in China, 2015.” *CA: A Cancer Journal for Clinicians* 66 (2):115–32. https://doi.org/10.3322/caac.21338.

Chen, Yu, and Martin J. Blaser. 2008. “*Helicobacter Pylori* Colonization Is Inversely Associated with Childhood Asthma.” *The Journal of Infectious Diseases* 198 (4):553–60. https://doi.org/10.1086/590158.

Chmiela, Magdalena, Eliza Miszczyk, and Karolina Rudnicka. 2014. “Structural Modifications of Helicobacter Pylori Lipopolysaccharide: An Idea for How to Live in Peace.” *World Journal of Gastroenterology* 20 (29):9882. https://doi.org/10.3748/wjg.v20.i29.9882.

Chrisment, Delphine, Pierre Dubus, Lucie Chambonnier, Anaïs Hocès de la Guardia, Elodie Sifré, Alban Giese, Myriam Capone, et al. 2014. “Neonatal Thymectomy Favors Helicobacter Pylori-Promoted Gastric Mucosa-Associated Lymphoid Tissue Lymphoma Lesions in BALB/c Mice.” *The American Journal of Pathology* 184 (8):2174–84. https://doi.org/10.1016/j.ajpath.2014.04.008.

Christie, J, N A Shepherd, B W Codling, and R M Valori. 1997. “Gastric Cancer below the Age of 55: Implications for Screening Patients with Uncomplicated Dyspepsia.” *Gut* 41 (4):513–17. http://www.ncbi.nlm.nih.gov/pubmed/9391251.

Chung, Daniel, Jonathan Glickman, Martin Carey, and Raymond Chung. 2005. “HST.121 Gastroenterology. Fall 2005. Massachusetts Institute of Technology: MIT OpenCourseWare.” 2005. https://ocw.mit.edu/courses/health-sciences-and-technology/hst-121-gastroenterology-fall-2005/#.

Cohen, Seth M, Magdalena Petryk, Mala Varma, Peter S Kozuch, Elizabeth D Ames,

and Michael L Grossbard. 2006. “Non-Hodgkin’s Lymphoma of Mucosa-Associated Lymphoid Tissue.” *The Oncologist* 11 (10). AlphaMed Press:1100–1117. https://doi.org/10.1634/theoncologist.11-10-1100.

Collins, Caitlin, and Xavier Didelot. 2018. “A Phylogenetic Method to Perform Genome-Wide Association Studies in Microbes That Accounts for Population Structure and Recombination.” Edited by Alice Carolyn McHardy. *PLOS Computational Biology* 14 (2):e1005958. https://doi.org/10.1371/journal.pcbi.1005958.

Correa, P. 1988. “Chronic Gastritis: A Clinico-Pathological Classification.” *The American Journal of Gastroenterology* 83 (5):504–9. http://www.ncbi.nlm.nih.gov/pubmed/3364410.

Cosgun, Yasemin, Abdullah Yildirim, Mihriban Yucel, Ayse Esra Karakoc, Gokhan Koca, Alpaslan Gonultas, Gul Gursoy, Huseyin Ustun, and Meliha Korkmaz. 2016. “Evaluation of Invasive and Noninvasive Methods for the Diagnosis of Helicobacter Pylori Infection.” *Asian Pacific Journal of Cancer Prevention : APJCP* 17 (12):6165–72. https://doi.org/10.22034/APJCP.2016.17.12.6165.

Costa, Débora Menezes da, Eliane dos Santos Pereira, and Silvia Helena Barem Rabenhorst. 2015. “What Exists beyond cagA and vacA? Helicobacter Pylori Genes in Gastric Diseases.” *World Journal of Gastroenterology* 21 (37):10563–72. https://doi.org/10.3748/wjg.v21.i37.10563.

Cover, T L, and M J Blaser. 1992. “Purification and Characterization of the Vacuolating Toxin from Helicobacter Pylori.” *The Journal of Biological Chemistry* 267 (15):10570–75. http://www.ncbi.nlm.nih.gov/pubmed/1587837.

Cristescu, Razvan, Jeeyun Lee, Michael Nebozhyn, Kyoung-Mee Kim, Jason C Ting, Swee Seong Wong, Jiangang Liu, et al. 2015. “Molecular Analysis of Gastric Cancer Identifies Subtypes Associated with Distinct Clinical Outcomes.” *Nature Medicine* 21 (5):449–56. https://doi.org/10.1038/nm.3850.

Crooks, G. E., Gary Hon, John-Marc Chandonia, and Steven E Brenner. 2004. “WebLogo: A Sequence Logo Generator.” *Genome Research* 14 (6):1188–90. https://doi.org/10.1101/gr.849004.

Cullen, Thomas W., David K. Giles, Lindsey N. Wolf, Chantal Ecobichon, Ivo G. Boneca, and M. Stephen Trent. 2011. “Helicobacter Pylori versus the Host: Remodeling of the Bacterial Outer Membrane Is Required for Survival in the Gastric Mucosa.” Edited by Nina Salama. *PLoS Pathogens* 7 (12):e1002454.

https://doi.org/10.1371/journal.ppat.1002454.

D'Elia, Lanfranco, Giovanni Rossi, Renato Ippolito, Francesco P. Cappuccio, and Pasquale Strazzullo. 2012. "Habitual Salt Intake and Risk of Gastric Cancer: A Meta-Analysis of Prospective Studies." *Clinical Nutrition* 31 (4):489–98. https://doi.org/10.1016/j.clnu.2012.01.003.

D'Elios, Mario M, and Marina de Bernard. 2010. "To Treat or Not to Treat *Helicobacter Pylori* to Benefit Asthma Patients." *Expert Review of Respiratory Medicine* 4 (2):147–50. https://doi.org/10.1586/ers.10.9.

Deng, Hai, and David O'Hagan. 2008. "The Fluorinase, the Chlorinase and the Duf-62 Enzymes." *Current Opinion in Chemical Biology* 12 (5):582–92. https://doi.org/10.1016/j.cbpa.2008.06.036.

Devi, Savita, Eerappa Rajakumara, and Niyaz Ahmed. 2015. "Induction of Mincle by Helicobacter Pylori and Consequent Anti-Inflammatory Signaling Denote a Bacterial Survival Strategy." *Scientific Reports* 5 (1):15049. https://doi.org/10.1038/srep15049.

Diaz, M I, A Valdivia, P Martinez, J L Palacios, P Harris, J Novales, E Garrido, et al. 2005. "Helicobacter Pylori vacA s1a and s1b Alleles from Clinical Isolates from Different Regions of Chile Show a Distinct Geographic Distribution." *World Journal of Gastroenterology* 11 (40):6366–72. http://www.ncbi.nlm.nih.gov/pubmed/16419167.

Didelot, Xavier, and Daniel Falush. 2007. "Inference of Bacterial Microevolution Using Multilocus Sequence Data." *Genetics* 175 (3):1251–66. https://doi.org/10.1534/genetics.106.063305.

Domínguez-Bello, Maria G., Maria E. Pérez, Maria C. Bortolini, Francisco M. Salzano, Luis R. Pericchi, Orlisbeth Zambrano-Guzmán, and Bodo Linz. 2008. "Amerindian Helicobacter Pylori Strains Go Extinct, as European Strains Expand Their Host Range." Edited by Angus Buckling. *PLoS ONE* 3 (10):e3307. https://doi.org/10.1371/journal.pone.0003307.

Dong, Quan-Jiang, Li-Li Wang, Zi-Bing Tian, Xin-Jun Yu, Sheng-Jiao Jia, and Shi-Ying Xuan. 2014. "Reduced Genome Size of *Helicobacter Pylori* Originating from East Asia." *World Journal of Gastroenterology* 20 (19):5666. https://doi.org/10.3748/wjg.v20.i19.5666.

Dooley, Cornelius P., Hartley Cohen, Patrick L. Fitzgibbons, Madeline Bauer, Maria D. Appleman, Guillermo I. Perez-Perez, and Martin J. Blaser. 1989. "Prevalence

of *Helicobacter Pylori* Infection and Histologic Gastritis in Asymptomatic Persons." *New England Journal of Medicine* 321 (23):1562–66. https://doi.org/10.1056/NEJM198912073212302.

Doorn, L J Van, C Figueiredo, F Mégraud, S Pena, P Midolo, D M Queiroz, F Carneiro, et al. 1999. "Geographic Distribution of vacA Allelic Types of Helicobacter Pylori." *Gastroenterology* 116 (4):823–30. http://www.ncbi.nlm.nih.gov/pubmed/10092304.

Dorer, Marion S, Tate H Sessler, and Nina R Salama. 2011. "Recombination and DNA Repair in Helicobacter Pylori." *Annual Review of Microbiology* 65. NIH Public Access:329–48. https://doi.org/10.1146/annurev-micro-090110-102931.

Draper, Jenny L, Lori M Hansen, David L Bernick, Samar Abedrabbo, Jason G Underwood, Nguyet Kong, Bihua C Huang, et al. 2017. "Fallacy of the Unique Genome: Sequence Diversity within Single Helicobacter Pylori Strains." Edited by Claire M. Fraser. *mBio* 8 (1):e02321-16. https://doi.org/10.1128/mBio.02321-16.

Earle, Sarah G, Chieh-Hsi Wu, Jane Charlesworth, Nicole Stoesser, N Claire Gordon, Timothy M Walker, Chris C A Spencer, et al. 2016. "Identifying Lineage Effects When Controlling for Population Structure Improves Power in Bacterial Association Studies." *Nature Microbiology* 1 (5):16041. https://doi.org/10.1038/nmicrobiol.2016.41.

Eaton, K A, and S Krakowka. 1994. "Effect of Gastric pH on Urease-Dependent Colonization of Gnotobiotic Piglets by Helicobacter Pylori." *Infection and Immunity* 62 (9):3604–7. http://www.ncbi.nlm.nih.gov/pubmed/8063376.

Eaton, K A, S Suerbaum, C Josenhans, and S Krakowka. 1996. "Colonization of Gnotobiotic Piglets by Helicobacter Pylori Deficient in Two Flagellin Genes." *Infection and Immunity* 64 (7):2445–48. http://www.ncbi.nlm.nih.gov/pubmed/8698465.

El-Omar, E M. 2001. "The Importance of Interleukin 1beta in Helicobacter Pylori Associated Disease." *Gut* 48 (6):743–47. http://www.ncbi.nlm.nih.gov/pubmed/11358884.

El-Omar, Emad M, Charles S Rabkin, Marilie D Gammon, Thomas L Vaughan, Harvey A Risch, Janet B Schoenberg, Janet L Stanford, et al. 2003. "Increased Risk of Noncardia Gastric Cancer Associated with Proinflammatory Cytokine Gene Polymorphisms." *Gastroenterology* 124 (5):1193–1201.

http://www.ncbi.nlm.nih.gov/pubmed/12730860.

Engstrand, Lars, and Mathilda Lindberg. 2013. “Helicobacter Pylori and the Gastric Microbiota.” *Best Practice & Research. Clinical Gastroenterology* 27 (1):39–45. https://doi.org/10.1016/j.bpg.2013.03.016.

Enroth, H, W Kraaz, L Engstrand, O Nyrén, and T Rohan. 2000. “Helicobacter Pylori Strain Types and Risk of Gastric Cancer: A Case-Control Study.” *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 9 (9):981–85. http://www.ncbi.nlm.nih.gov/pubmed/11008919.

Evans, D G, T K Karjalainen, D J Evans, D Y Graham, and C H Lee. 1993. “Cloning, Nucleotide Sequence, and Expression of a Gene Encoding an Adhesin Subunit Protein of Helicobacter Pylori.” *Journal of Bacteriology* 175 (3):674–83. http://www.ncbi.nlm.nih.gov/pubmed/7678592.

Evans, D J, D G Evans, T Takemura, H Nakano, H C Lampert, D Y Graham, D N Granger, and P R Kvietys. 1995. “Characterization of a Helicobacter Pylori Neutrophil-Activating Protein.” *Infection and Immunity* 63 (6):2213–20. http://www.ncbi.nlm.nih.gov/pubmed/7768601.

Falush, Daniel, Thierry Wirth, Bodo Linz, Jonathan K Pritchard, Matthew Stephens, Mark Kidd, Martin J Blaser, et al. 2003. “Traces of Human Migrations in Helicobacter Pylori Populations.” *Science (New York, N.Y.)* 299 (5612):1582–85. https://doi.org/10.1126/science.1080857.

Ferlay, Jacques, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. 2015. “Cancer Incidence and Mortality Worldwide: Sources, Methods and Major Patterns in GLOBOCAN 2012.” *International Journal of Cancer* 136 (5):E359-86. https://doi.org/10.1002/ijc.29210.

Fernandez-Gonzalez, Esther, and Steffen Backert. 2014. “DNA Transfer in the Gastric Pathogen Helicobacter Pylori.” *Journal of Gastroenterology* 49 (4):594–604. https://doi.org/10.1007/s00535-014-0938-y.

Figura, Natale, Luigi Marano, Elena Moretti, and Antonio Ponzetto. 2016. “Helicobacter Pylori Infection and Gastric Carcinoma: Not All the Strains and Patients Are Alike.” *World Journal of Gastrointestinal Oncology* 8 (1):40–54. https://doi.org/10.4251/wjgo.v8.i1.40.

Fischer, W, J Püls, R Buhrdorf, B Gebert, S Odenbreit, and R Haas. 2001. "Systematic Mutagenesis of the Helicobacter Pylori Cag Pathogenicity Island: Essential Genes for CagA Translocation in Host Cells and Induction of Interleukin-8." *Molecular Microbiology* 42 (5):1337–48. http://www.ncbi.nlm.nih.gov/pubmed/11886563.

Ford, A. C., D. Forman, R. H. Hunt, Y. Yuan, and P. Moayyedi. 2014. "Helicobacter Pylori Eradication Therapy to Prevent Gastric Cancer in Healthy Asymptomatic Infected Individuals: Systematic Review and Meta-Analysis of Randomised Controlled Trials." *BMJ* 348 (may20 1):g3174–g3174. https://doi.org/10.1136/bmj.g3174.

Fox, Sarah, Kieran A. Ryan, Alice H. Berger, Katie Petro, Soumita Das, Sheila E. Crowe, and Peter B. Ernst. 2015. "The Role of C1q in Recognition of Apoptotic Epithelial Cells and Inflammatory Cytokine Production by Phagocytes during Helicobacter Pylori Infection." *Journal of Inflammation* 12 (1):51. https://doi.org/10.1186/s12950-015-0098-8.

Furuta, Yoshikazu, Mutsuko Konno, Takako Osaki, Hideo Yonezawa, Taichiro Ishige, Misaki Imai, Yuh Shiwa, et al. 2015. "Microevolution of Virulence-Related Genes in Helicobacter Pylori Familial Infection." *PloS One* 10 (5):e0127197. https://doi.org/10.1371/journal.pone.0127197.

Gewirtz, Andrew T., Yimin Yu, Uma S. Krishna, Dawn A. Israel, Sean L. Lyons, and Richard M. Peek, Jr. 2004. "*Helicobacter Pylori* Flagellin Evades Toll-Like Receptor 5–Mediated Innate Immunity." *The Journal of Infectious Diseases* 189 (10):1914–20. https://doi.org/10.1086/386289.

Globocan. 2012. "Fact Sheets by Cancer." 2012. http://globocan.iarc.fr/Pages/fact_sheets_cancer.aspx.

Go, M F, V Kapur, D Y Graham, and J M Musser. 1996. "Population Genetic Analysis of Helicobacter Pylori by Multilocus Enzyme Electrophoresis: Extensive Allelic Diversity and Recombinational Population Structure." *Journal of Bacteriology* 178 (13):3934–38. http://www.ncbi.nlm.nih.gov/pubmed/8682800.

Guilford, Parry, Justin Hopkins, James Harraway, Maybelle McLeod, Ngahiraka McLeod, Pauline Harawira, Huriana Taite, Robin Scoular, Andrew Miller, and Anthony E. Reeve. 1998. "E-Cadherin Germline Mutations in Familial Gastric Cancer." *Nature* 392 (6674):402–5. https://doi.org/10.1038/32918.

Guo, Le, Runting Yin, Kunmei Liu, Xiaobo Lv, Yonghong Li, Xiangguo Duan, Yuankui Chu, Tao Xi, and Yingying Xing. 2014. “Immunological Features and Efficacy of a Multi-Epitope Vaccine CTB-UE against H. Pylori in BALB/c Mice Model.” *Applied Microbiology and Biotechnology* 98 (8):3495–3507. https://doi.org/10.1007/s00253-013-5408-6.

Guttman, D S, and D E Dykhuizen. 1994. “Clonal Divergence in Escherichia Coli as a Result of Recombination, Not Mutation.” *Science (New York, N.Y.)* 266 (5189):1380–83. http://www.ncbi.nlm.nih.gov/pubmed/7973728.

Hacker, Jörg, and James B. Kaper. 2000. “Pathogenicity Islands and the Evolution of Microbes.” *Annual Review of Microbiology* 54 (1):641–79. https://doi.org/10.1146/annurev.micro.54.1.641.

Hammond, Charles E, Craig Beeson, Giovanni Suarez, Richard M Peek, Steffen Backert, and Adam J Smolka. 2015. “Helicobacter Pylori Virulence Factors Affecting Gastric Proton Pump Expression and Acid Secretion.” *American Journal of Physiology. Gastrointestinal and Liver Physiology* 309 (3):G193-201. https://doi.org/10.1152/ajpgi.00099.2015.

HATAKEYAMA, Masanori. 2017. “Structure and Function of <i>Helicobacter Pylori</i> CagA, the First-Identified Bacterial Protein Involved in Human Cancer.” *Proceedings of the Japan Academy, Series B* 93 (4):196–219. https://doi.org/10.2183/pjab.93.013.

He, Zilong, Huangkai Zhang, Shenghan Gao, Martin J. Lercher, Wei-Hua Chen, and Songnian Hu. 2016. “Evolview v2: An Online Visualization and Management Tool for Customized and Annotated Phylogenetic Trees.” *Nucleic Acids Research* 44 (W1):W236–41. https://doi.org/10.1093/nar/gkw370.

Holcombe, C. 1992. “Helicobacter Pylori: The African Enigma.” *Gut* 33 (4):429–31. http://www.ncbi.nlm.nih.gov/pubmed/1582581.

HPC Wales. 2017. “HPC Wales Portal.” 2017. https://portal.hpcwales.co.uk/wordpress/.

Hu, Bing, Nassim El Hajj, Scott Sittler, Nancy Lammert, Robert Barnes, and Aurelia Meloni-Ehrig. 2012. “Gastric Cancer: Classification, Histology and Application of Molecular Pathology.” *Journal of Gastrointestinal Oncology* 3 (3). AME Publications:251–61. https://doi.org/10.3978/j.issn.2078-6891.2012.021.

Huang, Chih-Chieh, Kuo-Wang Tsai, Tzung-Jiun Tsai, and Ping-I Hsu. 2017. “Update on the First-Line Treatment for Helicobacter Pylori Infection - a

Continuing Challenge from an Old Enemy." *Biomarker Research* 5. BioMed Central:23. https://doi.org/10.1186/s40364-017-0103-x.

Huang, Jia Qing, Ge Fan Zheng, Katica Sumanac, E Jan Irvine, and Richard H Hunt. 2003. "Meta-Analysis of the Relationship between cagA Seropositivity and Gastric Cancer." *Gastroenterology* 125 (6):1636–44. http://www.ncbi.nlm.nih.gov/pubmed/14724815.

Hussain, Khiyam, Darren P. Letley, A. Borgel Greenaway, Rupert Kenefeck, Jody A. Winter, William Tomlinson, Joanne Rhead, et al. 2016. "Helicobacter Pylori-Mediated Protection from Allergy Is Associated with IL-10-Secreting Peripheral Blood Regulatory T Cells." *Frontiers in Immunology* 7 (March):71. https://doi.org/10.3389/fimmu.2016.00071.

Hwang, Jae Jin, Dong Ho Lee, Ae-Ra Lee, Hyuk Yoon, Cheol Min Shin, Young Soo Park, and Nayoung Kim. 2015. "Characteristics of Gastric Cancer in Peptic Ulcer Patients with Helicobacter Pylori Infection." *World Journal of Gastroenterology* 21 (16). Baishideng Publishing Group Inc:4954–60. https://doi.org/10.3748/wjg.v21.i16.4954.

IARC. 1994. "Schistosomes, Liver Flukes and Helicobacter Pylori. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans." In *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans / World Health Organization, International Agency for Research on Cancer*, 61:1–241. http://www.ncbi.nlm.nih.gov/pubmed/7715068.

Ierardi, Enzo, Floriana Giorgio, Giuseppe Losurdo, Alfredo Di Leo, and Mariabeatrice Principi. 2013. "How Antibiotic Resistances Could Change Helicobacter Pylori Treatment: A Matter of Geography?" *World Journal of Gastroenterology* 19 (45). Baishideng Publishing Group Inc:8168–80. https://doi.org/10.3748/wjg.v19.i45.8168.

Iizasa, Hisashi, Asuka Nanbo, Jun Nishikawa, Masahisa Jinushi, and Hironori Yoshiyama. 2012. "Epstein-Barr Virus (EBV)-Associated Gastric Carcinoma." *Viruses* 4 (12). Multidisciplinary Digital Publishing Institute (MDPI):3420–39. https://doi.org/10.3390/V4123420.

Isaacson, P, and D H Wright. 1983. "Malignant Lymphoma of Mucosa-Associated Lymphoid Tissue. A Distinctive Type of B-Cell Lymphoma." *Cancer* 52 (8):1410–16. http://www.ncbi.nlm.nih.gov/pubmed/6193858.

Ishijima, Nozomi, Masato Suzuki, Hiroshi Ashida, Yusuke Ichikawa, Yumi Kanegae,

Izumu Saito, Thomas Borén, Rainer Haas, Chihiro Sasakawa, and Hitomi Mimuro. 2011. “BabA-Mediated Adherence Is a Potentiator of the *Helicobacter Pylori* Type IV Secretion System Activity.” *Journal of Biological Chemistry* 286 (28):25256–64. https://doi.org/10.1074/jbc.M111.233601.

Israel, D A, N Salama, U Krishna, U M Rieger, J C Atherton, S Falkow, and R M Peek. 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). National Academy of Sciences:14625–30. https://doi.org/10.1073/pnas.251551698.

Iwamoto, H, D M Czajkowsky, T L Cover, G Szabo, and Z Shao. 1999. “VacA from Helicobacter Pylori: A Hexameric Chloride Channel.” *FEBS Letters* 450 (1–2):101–4. http://www.ncbi.nlm.nih.gov/pubmed/10350065.

Jee, Justin, Aviram Rasouly, Ilya Shamovsky, Yonatan Akivis, Susan R. Steinman, Bud Mishra, and Evgeny Nudler. 2016. “Rates and Mechanisms of Bacterial Mutagenesis from Maximum-Depth Sequencing.” *Nature* 534 (7609):693–96. https://doi.org/10.1038/nature18313.

Jemal, A., R. Siegel, J. Xu, and E. Ward. 2010. “Cancer Statistics, 2010.” *CA: A Cancer Journal for Clinicians* 60 (5). Wiley Subscription Services, Inc., A Wiley Company:277–300. https://doi.org/10.3322/caac.20073.

Jenks, P J, S Foynes, S J Ward, C Constantinidou, C W Penn, and B W Wren. 1997. “A Flagellar-Specific ATPase (FliI) Is Necessary for Flagellar Export in Helicobacter Pylori.” *FEMS Microbiology Letters* 152 (2):205–11. http://www.ncbi.nlm.nih.gov/pubmed/9231413.

Jolley, Keith. 2017. “The Helicobacter Pylori Multi Locus Sequence Typing Website.” 2017. https://pubmlst.org/helicobacter/info/primers.shtml.

Jolley, Keith A, and Martin C J Maiden. 2010. “BIGSdb: Scalable Analysis of Bacterial Genome Variation at the Population Level.” *BMC Bioinformatics* 11 (1):595. https://doi.org/10.1186/1471-2105-11-595.

Jones, A C, R P Logan, S Foynes, A Cockayne, B W Wren, and C W Penn. 1997. “A Flagellar Sheath Protein of Helicobacter Pylori Is Identical to HpaA, a Putative N-Acetylneuraminyllactose-Binding Hemagglutinin, but Is Not an Adhesin for AGS Cells.” *Journal of Bacteriology* 179 (17):5643–47. http://www.ncbi.nlm.nih.gov/pubmed/9287032.

Jones, M D, I Ademi, X Yin, Y Gong, and D B Zamble. 2015. “Nickel-Responsive

Regulation of Two Novel Helicobacter Pylori NikR-Targeted Genes." *Metallomics : Integrated Biometal Science* 7 (4):662–73. https://doi.org/10.1039/c4mt00210e.

Jönsson, Klas, Betty P. Guo, Hans-Jürg Monstein, John J. Mekalanos, and Göran Kronvall. 2004. "Molecular Cloning and Characterization of Two *Helicobacter Pylori* Genes Coding for Plasminogen-Binding Proteins." *Proceedings of the National Academy of Sciences* 101 (7):1852–57. https://doi.org/10.1073/pnas.0307329101.

Joossens, J V, M J Hill, P Elliott, R Stamler, E Lesaffre, A Dyer, R Nichols, and H Kesteloot. 1996. "Dietary Salt, Nitrate and Stomach Cancer Mortality in 24 Countries. European Cancer Prevention (ECP) and the INTERSALT Cooperative Research Group." *International Journal of Epidemiology* 25 (3):494–504. http://www.ncbi.nlm.nih.gov/pubmed/8671549.

Junaid, Muhammad, Aung Khine Linn, Mohammad Bagher Javadi, Sarbast Al-Gubare, Niaz Ali, and Gerd Katzenmeier. 2016. "Vacuolating Cytotoxin A (VacA) – A Multi-Talented Pore-Forming Toxin from Helicobacter Pylori." *Toxicon* 118 (August):27–35. https://doi.org/10.1016/j.toxicon.2016.04.037.

Kao, Cheng-Yen, Bor-Shyang Sheu, Shew-Meei Sheu, Hsiao-Bai Yang, Wei-Lun Chang, Hsiu-Chi Cheng, and Jiunn-Jong Wu. 2012. "Higher Motility Enhances Bacterial Density and Inflammatory Response in Dyspeptic Patients Infected with *Helicobacter Pylori*." *Helicobacter* 17 (6):411–16. https://doi.org/10.1111/j.1523-5378.2012.00974.x.

Kao, Cheng-Yen, Bor-Shyang Sheu, and Jiunn-Jong Wu. 2014. "CsrA Regulates *Helicobacter Pylori* J99 Motility and Adhesion by Controlling Flagella Formation." *Helicobacter* 19 (6):443–54. https://doi.org/10.1111/hel.12148.

Kao, John Y., Min Zhang, Mark J. Miller, Jason C. Mills, Baomei Wang, Maochang Liu, Kathyn A. Eaton, et al. 2010. "Helicobacter Pylori Immune Escape Is Mediated by Dendritic Cell–Induced Treg Skewing and Th17 Suppression in Mice." *Gastroenterology* 138 (3):1046–54. https://doi.org/10.1053/j.gastro.2009.11.043.

Kelley, Jon R., and John M. Duggan. 2003. "Gastric Cancer Epidemiology and Risk Factors." *Journal of Clinical Epidemiology* 56 (1):1–9. https://doi.org/10.1016/S0895-4356(02)00534-6.

Kennemann, Lynn, Xavier Didelot, Toni Aebischer, Stefanie Kuhn, Bernd Drescher,

Marcus Droege, Richard Reinhardt, et al. 2011. “Helicobacter Pylori Genome Evolution during Human Infection.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (12). National Academy of Sciences:5033–38. https://doi.org/10.1073/pnas.1018444108.

Kersulyte, Dangeruta, M Teresita Bertoli, Sravya Tamma, Monika Keelan, Rachel Munday, Janis Geary, Sander Veldhuyzen van Zanten, Karen J Goodman, and Douglas E Berg. 2015. “Complete Genome Sequences of Two Helicobacter Pylori Strains from a Canadian Arctic Aboriginal Community.” *Genome Announcements* 3 (2). https://doi.org/10.1128/genomeA.00209-15.

Kersulyte, Dangeruta, Henrikas Chalkauskas, and Douglas E. Berg. 1999. “Emergence of Recombinant Strains of Helicobacter Pylori during Human Infection.” *Molecular Microbiology* 31 (1). Blackwell Science Ltd:31–43. https://doi.org/10.1046/j.1365-2958.1999.01140.x.

Khatoon, J., K. N. Prasad, R. Prakash Rai, U. C. Ghoshal, and N. Krishnani. 2017. “Association of Heterogenicity of *Helicobacter Pylori* Cag Pathogenicity Island with Peptic Ulcer Diseases and Gastric Cancer.” *British Journal of Biomedical Science* 74 (3):121–26. https://doi.org/10.1080/09674845.2017.1278887.

Kibria, Khandoker Mohammad K., Md Enayet Hossain, Jinath Sultana, Shafiqul A. Sarker, Pradip Kumar Bardhan, Motiur Rahman, and Shamsun Nahar. 2015. “The Prevalence of Mixed *Helicobacter Pylori* Infections in Symptomatic and Asymptomatic Subjects in Dhaka, Bangladesh.” *Helicobacter* 20 (5):397–404. https://doi.org/10.1111/hel.12213.

Kidd, M, A J Lastovica, J C Atherton, and J A Louw. 1999. “Heterogeneity in the Helicobacter Pylori vacA and cagA Genes: Association with Gastroduodenal Disease in South Africa?” *Gut* 45 (4):499–502. http://www.ncbi.nlm.nih.gov/pubmed/10486355.

Kim, D H., S W Kim, Y. J. Song, T. Y. Oh, S. U. Han, Y B Kim, H. J. Joo, et al. 2003. “Long-Term Evaluation of Mice Model Infected with Helicobacter Pylori: Focus on Gastric Pathology Including Gastric Cancer.” *Alimentary Pharmacology & Therapeutics* 18 Suppl 1 (s1). Blackwell Publishing Ltd:14–23. https://doi.org/10.1046/j.1365-2036.18.s1.4.x.

Kim, J. M., J. S. Kim, D. Y. Yoo, S. H. Ko, N. Kim, H. Kim, and Y.-J. Kim. 2011. “Stimulation of Dendritic Cells with Helicobacter Pylori Vacuolating Cytotoxin Negatively Regulates Their Maturation via the Restoration of E2F1.” *Clinical &*

Experimental Immunology 166 (1):34–45. https://doi.org/10.1111/j.1365-2249.2011.04447.x.

Kim, J S, J H Chang, S I Chung, and J S Yum. 1999. “Molecular Cloning and Characterization of the Helicobacter Pylori fliD Gene, an Essential Factor in Flagellar Structure and Motility.” *Journal of Bacteriology* 181 (22):6969–76. http://www.ncbi.nlm.nih.gov/pubmed/10559162.

Kim, Jeong Wook, Jae Gyu Kim, Seok Lae Chae, Young Joo Cha, and Sill Moo Park. 2004. “High Prevalence of Multiple Strain Colonization of Helicobacter Pylori in Korean Patients: DNA Diversity among Clinical Isolates from the Gastric Corpus, Antrum and Duodenum.” *The Korean Journal of Internal Medicine* 19 (1):1–9. http://www.ncbi.nlm.nih.gov/pubmed/15053036.

Kim, Seok Yong, Chan Won Woo, Young Min Lee, Bo Ra Son, Ji Won Kim, Hee Bok Chae, Sei Jin Youn, and Seon Mee Park. 2001. “Genotyping *CagA* , *VacA* Subtype, *IceA1* , and *BabA* of *Helicobacter Pylori* Isolates from Korean Patients, and Their Association with Gastroduodenal Diseases.” *Journal of Korean Medical Science* 16 (5):579. https://doi.org/10.3346/jkms.2001.16.5.579.

Kimura, Motoo. 1967. “On the Evolutionary Adjustment of Spontaneous Mutation Rates.” *Genetical Research* 9 (1). Cambridge University Press:23. https://doi.org/10.1017/S0016672300010284.

Kivi, Mårten, Sandra Rodin, Ilya Kupershmidt, Annelie Lundin, Ylva Tindberg, Marta Granström, and Lars Engstrand. 2007. “Helicobacter Pylori Genome Variability in a Framework of Familial Transmission.” *BMC Microbiology* 7 (1). BioMed Central:54. https://doi.org/10.1186/1471-2180-7-54.

Kodama, Masaaki, Kazunari Murakami, Ryugo Sato, Tadayoshi Okimoto, Akira Nishizono, and Toshio Fujioka. 2005. “Helicobacter Pylori-Infected Animal Models Are Extremely Suitable for the Investigation of Gastric Carcinogenesis.” *World Journal of Gastroenterology* 11 (45). Baishideng Publishing Group Inc:7063–71. https://doi.org/10.3748/WJG.V11.I45.7063.

Kodaman, Nuri, Alvaro Pazos, Barbara G Schneider, M Blanca Piazuelo, Robertino Mera, Rafal S Sobota, Liviu A Sicinschi, et al. 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). National Academy of Sciences:1455–60. https://doi.org/10.1073/pnas.1318093111.

Kuhns, Lisa G, Stéphane L Benoit, Krishnareddy Bayyareddy, Darryl Johnson, Ron

Orlando, Alexandra L Evans, Grover L Waldrop, and Robert J Maier. 2016. “Carbon Fixation Driven by Molecular Hydrogen Results in Chemolithoautotrophically Enhanced Growth of Helicobacter Pylori.” Edited by P. J. Christie. *Journal of Bacteriology* 198 (9):1423–28. https://doi.org/10.1128/JB.00041-16.

Kuipers, E J, D A Israel, J G Kusters, M M Gerrits, J Weel, A van Der Ende, R W van Der Hulst, et al. 2000. “Quasispecies Development of Helicobacter Pylori Observed in Paired Isolates Obtained Years apart from the Same Host.” *The Journal of Infectious Diseases* 181 (1). NIH Public Access:273–82. https://doi.org/10.1086/315173.

Kumar, S., A. Kumar, and V. K. Dixit. 2010. “Diversity in the Cag Pathogenicity Island of Helicobacter Pylori Isolates in Populations from North and South India.” *Journal of Medical Microbiology* 59 (1):32–40. https://doi.org/10.1099/jmm.0.013763-0.

Kumar Pachathundikandi, Suneesh, Sabine Brandt, Joseph Madassery, and Steffen Backert. 2011. “Induction of TLR-2 and TLR-5 Expression by Helicobacter Pylori Switches cagPAI-Dependent Signalling Leading to the Secretion of IL-8 and TNF-α.” Edited by Yoshio Yamaoka. *PLoS ONE* 6 (5):e19614. https://doi.org/10.1371/journal.pone.0019614.

Lanas, Angel, and Francis K L Chan. 2017. “Peptic Ulcer Disease.” *Lancet (London, England)* 390 (10094). Elsevier:613–24. https://doi.org/10.1016/S0140-6736(16)32404-7.

Lawson, Daniel John, Garrett Hellenthal, Simon Myers, and Daniel Falush. 2012. “Inference of Population Structure Using Dense Haplotype Data.” Edited by Gregory P. Copenhaver. *PLoS Genetics* 8 (1):e1002453. https://doi.org/10.1371/journal.pgen.1002453.

Letunic, Ivica, and Peer Bork. 2016. “Interactive Tree of Life (iTOL) v3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees.” *Nucleic Acids Research* 44 (W1):W242–45. https://doi.org/10.1093/nar/gkw290.

Li, S D, D Kersulyte, I J Lindley, B Neelam, D E Berg, and J E Crabtree. 1999. “Multiple Genes in the Left Half of the Cag Pathogenicity Island of Helicobacter Pylori Are Required for Tyrosine Kinase-Dependent Transcription of Interleukin-8 in Gastric Epithelial Cells.” *Infection and Immunity* 67 (8):3893–99. http://www.ncbi.nlm.nih.gov/pubmed/10417153.

LI, W, M MINOHARA, J SU, T MATSUOKA, M OSOEGAWA, T ISHIZU, and J KIRA. 2007. “Helicobacter Pylori Infection Is a Potential Protective Factor against Conventional Multiple Sclerosis in the Japanese Population.” *Journal of Neuroimmunology* 184 (1–2):227–31. https://doi.org/10.1016/j.jneuroim.2006.12.010.

Lina, Taslima T, Shatha Alzahrani, Jazmin Gonzalez, Irina V Pinchuk, Ellen J Beswick, and Victor E Reyes. 2014. “Immune Evasion Strategies Used by Helicobacter Pylori.” *World Journal of Gastroenterology* 20 (36). Baishideng Publishing Group Inc:12753–66. https://doi.org/10.3748/wjg.v20.i36.12753.

Ling, T K, A F Cheng, J J Sung, P Y Yiu, and S S Chung. 1996. “An Increase in Helicobacter Pylori Strains Resistant to Metronidazole: A Five-Year Study.” *Helicobacter* 1 (1):57–61. http://www.ncbi.nlm.nih.gov/pubmed/9398914.

Linz, Bodo, François Balloux, Yoshan Moodley, Andrea Manica, Hua Liu, Philippe Roumagnac, Daniel Falush, et al. 2007. “An African Origin for the Intimate Association between Humans and Helicobacter Pylori.” *Nature* 445 (7130):915–18. https://doi.org/10.1038/nature05562.

Lionetti, Elena, Salvatore Leonardi, Angela Lanzafame, Maria Teresa Garozzo, Martina Filippelli, Stefania Tomarchio, Viviana Ferrara, et al. 2014. “*Helicobacter Pylori* Infection and Atopic Diseases: Is There a Relationship? A Systematic Review and Meta-Analysis.” *World Journal of Gastroenterology* 20 (46):17635. https://doi.org/10.3748/wjg.v20.i46.17635.

Liu, Hui, Jutta B Fero, Melissa Mendez, Beth M Carpenter, Stephanie L Servetas, Arifur Rahman, Matthew D Goldman, et al. 2015. “Analysis of a Single Helicobacter Pylori Strain over a 10-Year Period in a Primate Model.” *International Journal of Medical Microbiology : IJMM* 305 (3):392–403. https://doi.org/10.1016/j.ijmm.2015.03.002.

Loffeld, R. J. L. F., E. Stobberingh, J. A. Flendrig, and J. W. Arends. 1991. “Helicobacter Pylori in Gastric Biopsy Specimens. Comparison of Culture, Modified Giemsa Stain, and Immunohistochemistry. A Retrospective Study.” *The Journal of Pathology* 165 (1):69–73. https://doi.org/10.1002/path.1711650111.

Loman, Nicholas J., Chrystala Constantinidou, Jacqueline Z. M. Chan, Mihail Halachev, Martin Sergeant, Charles W. Penn, Esther R. Robinson, and Mark J. Pallen. 2012. “High-Throughput Bacterial Genome Sequencing: An

Embarrassment of Choice, a World of Opportunity." *Nature Reviews Microbiology* 10 (9):599–606. https://doi.org/10.1038/nrmicro2850.

López-Vidal, Yolanda, Sergio Ponce-de-León, Gonzalo Castillo-Rojas, Rafael Barreto-Zúñiga, and Aldo Torre-Delgadillo. 2008. "High Diversity of vacA and cagA Helicobacter Pylori Genotypes in Patients with and without Gastric Cancer." *PloS One* 3 (12). Public Library of Science:e3849. https://doi.org/10.1371/journal.pone.0003849.

Lu, Hengyun, Francesca Giordano, and Zemin Ning. 2016. "Oxford Nanopore MinION Sequencing and Genome Assembly." *Genomics, Proteomics & Bioinformatics* 14 (5):265–79. https://doi.org/10.1016/j.gpb.2016.05.004.

Lugli, Alessandro, Inti Zlobec, Gad Singer, Andrea Kopp Lugli, Luigi M Terracciano, and Robert M Genta. 2007. "Napoleon Bonaparte's Gastric Cancer: A Clinicopathologic Approach to Staging, Pathogenesis, and Etiology." *Nature Clinical Practice Gastroenterology & Hepatology* 4 (1):52–57. https://doi.org/10.1038/ncpgasthep0684.

Ma, Ke, Zulqarnain Baloch, Ting-Ting He, and Xueshan Xia. 2017. "Alcohol Consumption and Gastric Cancer Risk: A Meta-Analysis." *Medical Science Monitor : International Medical Journal of Experimental and Clinical Research* 23 (January):238–46. http://www.ncbi.nlm.nih.gov/pubmed/28087989.

Maeda, S, K Ogura, H Yoshida, F Kanai, T Ikenoue, N Kato, Y Shiratori, and M Omata. 1998. "Major Virulence Factors, VacA and CagA, Are Commonly Positive in Helicobacter Pylori Isolates in Japan." *Gut* 42 (3):338–43. http://www.ncbi.nlm.nih.gov/pubmed/9577338.

Maixner, F., B. Krause-Kyora, D. Turaev, A. Herbig, M. R. Hoopmann, J. L. Hallows, U. Kusebauch, et al. 2016. "The 5300-Year-Old Helicobacter Pylori Genome of the Iceman." *Science* 351 (6269):162–65. https://doi.org/10.1126/science.aad2545.

Malfertheiner, P, F Megraud, C A O'Morain, J P Gisbert, E J Kuipers, A T Axon, F Bazzoli, et al. 2017. "Management of *Helicobacter Pylori* Infection?the Maastricht V/Florence Consensus Report." *Gut* 66 (1):6–30. https://doi.org/10.1136/gutjnl-2016-312288.

Mansour, Khansa Ben, Chedlia Fendri, Hajer Battikh, Martine Garnier, Meriem Zribi, Asma Jlizi, and Christophe Burucoa. 2016. "Multiple and Mixed Helicobacter Pylori Infections: Comparison of Two Epidemiological Situations in Tunisia and

France." *Infection, Genetics and Evolution* 37 (January):43–48. https://doi.org/10.1016/j.meegid.2015.10.028.

Marangoni, Aurelio, David Caramelli, and Giorgio Manzi. 2014. "Homo Sapiens in the Americas. Overview of the Earliest Human Expansion in the New World." *Journal of Anthropological Sciences = Rivista Di Antropologia : JASS* 92:79–97. https://doi.org/10.4436/jass.91002.

Marshall, B J. n.d. "Campylobacter Pylori: Its Link to Gastritis and Peptic Ulcer Disease." *Reviews of Infectious Diseases* 12 Suppl 1:S87-93. Accessed November 15, 2017. http://www.ncbi.nlm.nih.gov/pubmed/2406862.

Marshall, BarryJ, and J.Robin Warren. 1984. "UNIDENTIFIED CURVED BACILLI IN THE STOMACH OF PATIENTS WITH GASTRITIS AND PEPTIC ULCERATION." *The Lancet* 323 (8390):1311–15. https://doi.org/10.1016/S0140-6736(84)91816-6.

Mattar, Rejane, Maria S Monteiro, Sergio B Marques, Bruno Zilberstein, Cláudio L Hashimoto, and Flair J Carrilho. 2010. "Association of LEC and tnpA Helicobacter Pylori Genes with Gastric Cancer in a Brazilian Population." *Infectious Agents and Cancer* 5 (January). BioMed Central:1. https://doi.org/10.1186/1750-9378-5-1.

Méric, Guillaume, Koji Yahara, Leonardos Mageiros, Ben Pascoe, Martin C J Maiden, Keith A Jolley, and Samuel K Sheppard. 2014. "A Reference Pan-Genome Approach to Comparative Bacterial Genomics: Identification of Novel Epidemiological Markers in Pathogenic Campylobacter." Edited by Stefan Bereswill. *PloS One* 9 (3):e92798. https://doi.org/10.1371/journal.pone.0092798.

Miehlke, S, C Kirsch, K Agha-Amiri, T Günther, N Lehn, P Malfertheiner, M Stolte, G Ehninger, and E Bayerdörffer. 2000. "The Helicobacter Pylori vacA s1, m1 Genotype and cagA Is Associated with Gastric Carcinoma in Germany." *International Journal of Cancer* 87 (3):322–27. http://www.ncbi.nlm.nih.gov/pubmed/10897035.

Miftahussurur, Muhammad, and Yoshio Yamaoka. 2015. "*Helicobacter Pylori* Virulence Genes and Host Genetic Polymorphisms as Risk Factors for Peptic Ulcer Disease." *Expert Review of Gastroenterology & Hepatology* 9 (12):1535–47. https://doi.org/10.1586/17474124.2015.1095089.

Mitsuno, Y, H Yoshida, S Maeda, K Ogura, Y Hirata, T Kawabe, Y Shiratori, and M Omata. 2001. "Helicobacter Pylori Induced Transactivation of SRE and AP-1

through the ERK Signalling Pathway in Gastric Cancer Cells." *Gut* 49 (1):18–22. http://www.ncbi.nlm.nih.gov/pubmed/11413105.

Mobley, H L, L T Hu, and P A Foxal. 1991. "Helicobacter Pylori Urease: Properties and Role in Pathogenesis." *Scandinavian Journal of Gastroenterology. Supplement* 187:39–46. http://www.ncbi.nlm.nih.gov/pubmed/1775923.

Montano, Valeria, Xavier Didelot, Matthieu Foll, Bodo Linz, Richard Reinhardt, Sebastian Suerbaum, Yoshan Moodley, and Jeffrey D Jensen. 2015. "Worldwide Population Structure, Long-Term Demography, and Local Adaptation of Helicobacter Pylori." *Genetics* 200 (3):947–63. https://doi.org/10.1534/genetics.115.176404.

Monteil, Caroline L, Koji Yahara, David J Studholme, Leonardos Mageiros, Guillaume Méric, Bryan Swingle, Cindy E Morris, Boris A Vinatzer, and Samuel K Sheppard. 2016. "Population-Genomic Insights into Emergence, Crop Adaptation and Dissemination of Pseudomonas Syringae Pathogens." *Microbial Genomics* 2 (10):e000089. https://doi.org/10.1099/mgen.0.000089.

Monteiro, M A, K H Chan, D A Rasko, D E Taylor, P Y Zheng, B J Appelmelk, H P Wirth, et al. 1998. "Simultaneous Expression of Type 1 and Type 2 Lewis Blood Group Antigens by Helicobacter Pylori Lipopolysaccharides. Molecular Mimicry between H. Pylori Lipopolysaccharides and Human Gastric Epithelial Cell Surface Glycoforms." *The Journal of Biological Chemistry* 273 (19):11533–43. http://www.ncbi.nlm.nih.gov/pubmed/9565568.

Moodley, Yoshan, Bodo Linz, Robert P Bond, Martin Nieuwoudt, Himla Soodyall, Carina M Schlebusch, Steffi Bernhöft, et al. 2012. "Age of the Association between Helicobacter Pylori and Man." *PLoS Pathogens* 8 (5). Public Library of Science:e1002693. https://doi.org/10.1371/journal.ppat.1002693.

Moran, A P, and G O Aspinall. 1998. "Unique Structural and Biological Features of Helicobacter Pylori Lipopolysaccharides." *Progress in Clinical and Biological Research* 397:37–49. http://www.ncbi.nlm.nih.gov/pubmed/9575546.

Mukherjee, Supratim, Marcel Huntemann, Natalia Ivanova, Nikos C Kyrpides, and Amrita Pati. 2015. "Large-Scale Contamination of Microbial Isolate Genomes by Illumina PhiX Control." *Standards in Genomic Sciences* 10 (1). BioMed Central:18. https://doi.org/10.1186/1944-3277-10-18.

Muotiala, A, I M Helander, L Pyhälä, T U Kosunen, and A P Moran. 1992. "Low Biological Activity of Helicobacter Pylori Lipopolysaccharide." *Infection and*

Immunity 60 (4). American Society for Microbiology (ASM):1714–16. http://www.ncbi.nlm.nih.gov/pubmed/1548097.

NCBI. 2017. "Genome - Helicobacter Pylori - NCBI." 2017. https://www.ncbi.nlm.nih.gov/genome/genomes/169#.

Nell, Sandra, Daniel Eibach, Valeria Montano, Ayas Maady, Armand Nkwescheu, Jose Siri, Wael F Elamin, et al. 2013. "Recent Acquisition of Helicobacter Pylori by Baka Pygmies." *PLoS Genetics* 9 (9):e1003775. https://doi.org/10.1371/journal.pgen.1003775.

Nilsson, Christina, Anna Sillén, Lena Eriksson, Mona-Lisa Strand, Helena Enroth, Staffan Normark, Per Falk, and Lars Engstrand. 2003. "Correlation between Cag Pathogenicity Island Composition and Helicobacter Pylori-Associated Gastroduodenal Disease." *Infection and Immunity* 71 (11). American Society for Microbiology:6573–81. https://doi.org/10.1128/IAI.71.11.6573-6581.2003.

Nishikawa, K, T Sugiyama, M Kato, J Ishizuka, H Kagaya, K Hokari, and M Asaka. 2000. "A Prospective Evaluation of New Rapid Urease Tests before and after Eradication Treatment of Helicobacter Pylori, in Comparison with Histology, Culture and 13C-Urea Breath Test." *Gastrointestinal Endoscopy* 51 (2):164–68. http://www.ncbi.nlm.nih.gov/pubmed/10650258.

Nordenstedt, Helena, David Y Graham, Jennifer R Kramer, Massimo Rugge, Gordana Verstovsek, Stephanie Fitzgerald, Abeer Alsarraj, et al. 2013. "Helicobacter Pylori-Negative Gastritis: Prevalence and Risk Factors." *The American Journal of Gastroenterology* 108 (1). NIH Public Access:65–71. https://doi.org/10.1038/ajg.2012.372.

O'Connor, Anthony, Colm A. O'Morain, and Alexander C. Ford. 2017. "Population Screening and Treatment of Helicobacter Pylori Infection." *Nature Reviews Gastroenterology & Hepatology*, January. https://doi.org/10.1038/nrgastro.2016.195.

O'Ryan, Miguel L., Yalda Lucero, Marcela Rabello, Nora Mamani, Ana María Salinas, Alfredo Peña, Juan Pablo Torres-Torreti, et al. 2015. "Persistent and Transient *Helicobacter Pylori* Infections in Early Childhood." *Clinical Infectious Diseases* 61 (2):211–18. https://doi.org/10.1093/cid/civ256.

O'Toole, P W, L Janzon, P Doig, J Huang, M Kostrzynska, and T J Trust. 1995. "The Putative Neuraminyllactose-Binding Hemagglutinin HpaA of Helicobacter Pylori CCUG 17874 Is a Lipoprotein." *Journal of Bacteriology* 177 (21):6049–

57. http://www.ncbi.nlm.nih.gov/pubmed/7592366.

O'Toole, P W, M Kostrzynska, and T J Trust. 1994. "Non-Motile Mutants of Helicobacter Pylori and Helicobacter Mustelae Defective in Flagellar Hook Production." *Molecular Microbiology* 14 (4):691–703. http://www.ncbi.nlm.nih.gov/pubmed/7891557.

O'Toole, Paul W, Michael C Lane, and Steffen Porwollik. 2000. "Helicobacter Pylori Motility." *Microbes and Infection* 2 (10):1207–14. https://doi.org/10.1016/S1286-4579(00)01274-0.

Odenbreit, S, J Püls, B Sedlmaier, E Gerland, W Fischer, and R Haas. 2000. "Translocation of Helicobacter Pylori CagA into Gastric Epithelial Cells by Type IV Secretion." *Science (New York, N.Y.)* 287 (5457):1497–1500. http://www.ncbi.nlm.nih.gov/pubmed/10688800.

Odenbreit, S, B Wieland, and R Haas. 1996. "Cloning and Genetic Characterization of Helicobacter Pylori Catalase and Construction of a Catalase-Deficient Mutant Strain." *Journal of Bacteriology* 178 (23). American Society for Microbiology (ASM):6960–67. http://www.ncbi.nlm.nih.gov/pubmed/8955320.

Odenbreit, Stefan, Gerhard Faller, and Rainer Haas. 2002. "Role of the alpAB Proteins and Lipopolysaccharide in Adhesion of Helicobacter Pylori to Human Gastric Tissue." *International Journal of Medical Microbiology : IJMM* 292 (3–4):247–56. https://doi.org/10.1078/1438-4221-00204.

Oderda, G, M Forni, D Dell'Olio, and N Ansaldi. 1990. "Cure of Peptic Ulcer Associated with Eradication of Helicobacter Pylori." *Lancet (London, England)* 335 (8705):1599. http://www.ncbi.nlm.nih.gov/pubmed/1972523.

Oertli, M., M. Noben, D. B. Engler, R. P. Semper, S. Reuter, J. Maxeiner, M. Gerhard, C. Taube, and A. Muller. 2013. "Helicobacter Pylori -Glutamyl Transpeptidase and Vacuolating Cytotoxin Promote Gastric Persistence and Immune Tolerance." *Proceedings of the National Academy of Sciences* 110 (8):3047–52. https://doi.org/10.1073/pnas.1211248110.

Oertli, Mathias, and Anne Müller. 2012. "*Helicobacter Pylori* Targets Dendritic Cells to Induce Immune Tolerance, Promote Persistence and Confer Protection against Allergic Asthma." *Gut Microbes* 3 (6):566–71. https://doi.org/10.4161/gmic.21750.

Olbermann, Patrick, Christine Josenhans, Yoshan Moodley, Markus Uhr, Christiana Stamer, Marc Vauterin, Sebastian Suerbaum, Mark Achtman, and Bodo Linz.

2010. “A Global Overview of the Genetic and Functional Diversity in the Helicobacter Pylori Cag Pathogenicity Island.” Edited by Harmit S. Malik. *PLoS Genetics* 6 (8):e1001069. https://doi.org/10.1371/journal.pgen.1001069.

Osaki, T., M. Konno, H. Yonezawa, F. Hojo, C. Zaman, M. Takahashi, S. Fujiwara, and S. Kamiya. 2015. “Analysis of Intra-Familial Transmission of Helicobacter Pylori in Japanese Families.” *Journal of Medical Microbiology* 64 (Pt_1):67–73. https://doi.org/10.1099/jmm.0.080507-0.

Overbeek, Ross, Robert Olson, Gordon D. Pusch, Gary J. Olsen, James J. Davis, Terry Disz, Robert A. Edwards, et al. 2014. “The SEED and the Rapid Annotation of Microbial Genomes Using Subsystems Technology (RAST).” *Nucleic Acids Research* 42 (D1):D206–14. https://doi.org/10.1093/nar/gkt1226.

Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. “Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis.” *Bioinformatics* 31 (22):3691–93. https://doi.org/10.1093/bioinformatics/btv421.

Pallen, Mark J, Nicholas J Loman, and Charles W Penn. 2010. “High-Throughput Sequencing and Clinical Microbiology: Progress, Opportunities and Challenges.” *Current Opinion in Microbiology* 13 (5):625–31. https://doi.org/10.1016/j.mib.2010.08.003.

Park, E. K., H. S. Jung, H. I. Yang, M. C. Yoo, C. Kim, and K. S. Kim. 2007. “Optimized THP-1 Differentiation Is Required for the Detection of Responses to Weak Stimuli.” *Inflammation Research* 56 (1):45–50. https://doi.org/10.1007/s00011-007-6115-5.

Parsonnet, J. 1998. “Helicobacter Pylori.” *Infectious Disease Clinics of North America* 12 (1):185–97. http://www.ncbi.nlm.nih.gov/pubmed/9494838.

Parsonnet, J, G D Friedman, N Orentreich, and H Vogelman. 1997. “Risk for Gastric Cancer in People with CagA Positive or CagA Negative Helicobacter Pylori Infection.” *Gut* 40 (3):297–301. http://www.ncbi.nlm.nih.gov/pubmed/9135515.

Patel, S. R., K. Smith, D. P. Letley, K. W. Cook, A. A. Memon, R. J. M. Ingram, E. Staples, et al. 2013. “*H Elicobacter Pylori* Downregulates Expression of Human β-Defensin 1 in the Gastric Mucosa in a Type IV Secretion-Dependent Fashion.” *Cellular Microbiology* 15 (12):2080–92. https://doi.org/10.1111/cmi.12174.

Patel, Saurabh Kumar, Chandra Bhan Pratap, Ashok Kumar Jain, Anil Kumar Gulati,

and Gopal Nath. 2014. “Diagnosis of *Helicobacter Pylori* : What Should Be the Gold Standard?” *World Journal of Gastroenterology* 20 (36):12847. https://doi.org/10.3748/wjg.v20.i36.12847.

PatricdB. 2017a. “Helicobacter pylori::Taxonomy Genomes.” 2017. https://www.patricbrc.org/view/Taxonomy/210#view_tab=genomes.

———. 2017b. “Helicobacter Pylori 26695::Genome Overview.” 2017. https://www.patricbrc.org/view/Genome/85962.8#view_tab=overview.

Peck, B, M Ortkamp, K D Diehl, E Hundt, and B Knapp. 1999. “Conservation, Localization and Expression of HopZ, a Protein Involved in Adhesion of Helicobacter Pylori.” *Nucleic Acids Research* 27 (16):3325–33. http://www.ncbi.nlm.nih.gov/pubmed/10454640.

Peek, Richard M., and Martin J. Blaser. 2002. “HELICOBACTER PYLORI AND GASTROINTESTINAL TRACT ADENOCARCINOMAS.” *Nature Reviews Cancer* 2 (1):28–37. https://doi.org/10.1038/nrc703.

Pérez-Pérez, G I, V L Shepherd, J D Morrow, and M J Blaser. 1995. “Activation of Human THP-1 Cells and Rat Bone Marrow-Derived Macrophages by Helicobacter Pylori Lipopolysaccharide.” *Infection and Immunity* 63 (4):1183–87. http://www.ncbi.nlm.nih.gov/pubmed/7890370.

Pettersson, Erik, Joakim Lundeberg, and Afshin Ahmadian. 2009. “Generations of Sequencing Technologies.” *Genomics* 93 (2):105–11. https://doi.org/10.1016/j.ygeno.2008.10.003.

Pflock, Michael, Melanie Bathon, Jennifer Schär, Stefanie Müller, Hans Mollenkopf, Thomas F Meyer, and Dagmar Beier. 2007. “The Orphan Response Regulator HP1021 of Helicobacter Pylori Regulates Transcription of a Gene Cluster Presumably Involved in Acetone Metabolism.” *Journal of Bacteriology* 189 (6):2339–49. https://doi.org/10.1128/JB.01827-06.

Piazuelo, M Blanca, and Pelayo Correa. 2013. “Gastric Cáncer: Overview.” *Colombia Medica (Cali, Colombia)* 44 (3):192–201. http://www.ncbi.nlm.nih.gov/pubmed/24892619.

Plummer, Martyn, Silvia Franceschi, Jérôme Vignat, David Forman, and Catherine de Martel. 2015. “Global Burden of Gastric Cancer Attributable to Helicobacter Pylori.” *International Journal of Cancer* 136 (2):487–90. https://doi.org/10.1002/ijc.28999.

Power, Robert A., Julian Parkhill, and Tulio de Oliveira. 2016. “Microbial Genome-

Wide Association Studies: Lessons from Human GWAS." *Nature Reviews Genetics* 18 (1):41–50. https://doi.org/10.1038/nrg.2016.132.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin. 2010. "FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments." Edited by Art F. Y. Poon. *PLoS ONE* 5 (3):e9490. https://doi.org/10.1371/journal.pone.0009490.

Quigley, E M, and L A Turnberg. 1987. "pH of the Microclimate Lining Human Gastric and Duodenal Mucosa in Vivo. Studies in Control Subjects and in Duodenal Ulcer Patients." *Gastroenterology* 92 (6):1876–84. http://www.ncbi.nlm.nih.gov/pubmed/3569763.

Rad, Roland, Markus Gerhard, Roland Lang, Martin Schöniger, Thomas Rösch, Wolfgang Schepp, Ingrid Becker, Hermann Wagner, and Christian Prinz. 2002. "The Helicobacter Pylori Blood Group Antigen-Binding Adhesin Facilitates Bacterial Colonization and Augments a Nonspecific Immune Response." *Journal of Immunology (Baltimore, Md. : 1950)* 168 (6):3033–41. http://www.ncbi.nlm.nih.gov/pubmed/11884476.

Ramarao, N, S D Gray-Owen, S Backert, and T F Meyer. 2000. "Helicobacter Pylori Inhibits Phagocytosis by Professional Phagocytes Involving Type IV Secretion Components." *Molecular Microbiology* 37 (6):1389–1404. http://www.ncbi.nlm.nih.gov/pubmed/10998171.

Ramarao, N, S D Gray-Owen, and T F Meyer. 2000. "Helicobacter Pylori Induces but Survives the Extracellular Release of Oxygen Radicals from Professional Phagocytes Using Its Catalase Activity." *Molecular Microbiology* 38 (1):103–13. http://www.ncbi.nlm.nih.gov/pubmed/11029693.

Ramarao, N, and T F Meyer. 2001. "Helicobacter Pylori Resists Phagocytosis by Macrophages: Quantitative Assessment by Confocal Microscopy and Fluorescence-Activated Cell Sorting." *Infection and Immunity* 69 (4). American Society for Microbiology (ASM):2604–11. https://doi.org/10.1128/IAI.69.4.2604-2611.2001.

Rao, P, A Sarkar, P G Shivananda, and G Pai. 2001. "Comparison of ELISA for Antibody Detection and Biopsy Urease Test against H. Pylori in Cases of Gastoduodenal Disorders." *Indian Journal of Medical Sciences* 55 (7):366–70. http://www.ncbi.nlm.nih.gov/pubmed/11883335.

Rasko, D A, G Wang, M M Palcic, and D E Taylor. 2000. "Cloning and

Characterization of the alpha(1,3/4) Fucosyltransferase of Helicobacter Pylori." *The Journal of Biological Chemistry* 275 (7). American Society for Biochemistry and Molecular Biology:4988–94. https://doi.org/10.1074/JBC.275.7.4988.

Raymond, Josette, Jean-Michel Thiberge, Catherine Chevalier, Nicolas Kalach, Michel Bergeret, Agnès Labigne, and Catherine Dauga. 2004. "Genetic and Transmission Analysis of *Helicobacter Pylori* Strains within a Family1." *Emerging Infectious Diseases* 10 (10):1816–21. https://doi.org/10.3201/eid1010.040042.

Raymond, Josette, Jean-Michel Thiberge, Nicolas Kalach, Michel Bergeret, Christophe Dupont, Agnès Labigne, and Catherine Dauga. 2008. "Using Macro-Arrays to Study Routes of Infection of Helicobacter Pylori in Three Families." Edited by Martin Blaser. *PLoS ONE* 3 (5). Public Library of Science:e2259. https://doi.org/10.1371/journal.pone.0002259.

Rhee, Kwang-Ho, Jin-Sik Park, and Myung-Je Cho. 2014. "Helicobacter Pylori: Bacterial Strategy for Incipient Stage and Persistent Colonization in Human Gastric Niches." *Yonsei Medical Journal* 55 (6). Yonsei University College of Medicine:1453–66. https://doi.org/10.3349/ymj.2014.55.6.1453.

Rizzato, Cosmeri, Javier Torres, Martyn Plummer, Nubia Muñoz, Silvia Franceschi, Margarita Camorlinga-Ponce, Ezequiel M. Fuentes-Pananá, Federico Canzian, and Ikuko Kato. 2012. "Variations in Helicobacter Pylori Cytotoxin-Associated Genes and Their Influence in Progression to Gastric Cancer: Implications for Prevention." Edited by Masaru Katoh. *PLoS ONE* 7 (1):e29605. https://doi.org/10.1371/journal.pone.0029605.

Romo-González, Carolina, Nina R Salama, Juan Burgeño-Ferreira, Veronica Ponce-Castañeda, Eduardo Lazcano-Ponce, Margarita Camorlinga-Ponce, and Javier Torres. 2009. "Differences in Genome Content among Helicobacter Pylori Isolates from Patients with Gastritis, Duodenal Ulcer, or Gastric Cancer Reveal Novel Disease-Associated Genes." *Infection and Immunity* 77 (5). American Society for Microbiology:2201–11. https://doi.org/10.1128/IAI.01284-08.

Rugge, Massimo, Alberto Meggio, Gianmaria Pennelli, Francesco Piscioli, Luciano Giacomelli, Giovanni De Pretis, and David Y Graham. 2007. "Gastritis Staging in Clinical Practice: The OLGA Staging System." *Gut* 56 (5). BMJ Group:631–36. https://doi.org/10.1136/gut.2006.106666.

Sablet, Thibaut de, M Blanca Piazuelo, Carrie L Shaffer, Barbara G Schneider,

Mohammad Asim, Rupesh Chaturvedi, Luis E Bravo, et al. 2011. "Phylogeographic Origin of Helicobacter Pylori Is a Determinant of Gastric Cancer Risk." *Gut* 60 (9):1189–95. https://doi.org/10.1136/gut.2010.234468.

Salama, N. R., B. Shepherd, and S. Falkow. 2004. "Global Transposon Mutagenesis and Essential Gene Analysis of Helicobacter Pylori." *Journal of Bacteriology* 186 (23):7926–35. https://doi.org/10.1128/JB.186.23.7926-7935.2004.

Sanger, F, S Nicklen, and A R Coulson. 1977. "DNA Sequencing with Chain-Terminating Inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74 (12):5463–67. http://www.ncbi.nlm.nih.gov/pubmed/271968.

Satin, B, G Del Giudice, V Della Bianca, S Dusi, C Laudanna, F Tonello, D Kelleher, R Rappuoli, C Montecucco, and F Rossi. 2000. "The Neutrophil-Activating Protein (HP-NAP) of Helicobacter Pylori Is a Protective Antigen and a Major Virulence Factor." *The Journal of Experimental Medicine* 191 (9):1467–76. http://www.ncbi.nlm.nih.gov/pubmed/10790422.

Schmitz, A, C Josenhans, and S Suerbaum. 1997. "Cloning and Characterization of the Helicobacter Pylori flbA Gene, Which Codes for a Membrane Protein Involved in Coordinated Expression of Flagellar Genes." *Journal of Bacteriology* 179 (4):987–97. http://www.ncbi.nlm.nih.gov/pubmed/9023175.

Schröder, Gunnar, Sabine Krause, Ellen L Zechner, Beth Traxler, Hye-Jeong Yeo, Rudi Lurz, Gabriel Waksman, and Erich Lanka. 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):2767–79. http://www.ncbi.nlm.nih.gov/pubmed/11976307.

Segal, E D, C Lange, A Covacci, L S Tompkins, and S Falkow. 1997. "Induction of Host Signal Transduction Pathways by Helicobacter Pylori." *Proceedings of the National Academy of Sciences of the United States of America* 94 (14):7595–99. http://www.ncbi.nlm.nih.gov/pubmed/9207137.

Senkovich, O. A., J. Yin, V. Ekshyyan, C. Conant, J. Traylor, P. Adegboyega, D. J. McGee, R. E. Rhoads, S. Slepenkov, and T. L. Testerman. 2011. "Helicobacter Pylori AlpA and AlpB Bind Host Laminin and Influence Gastric Inflammation in Gerbils." *Infection and Immunity* 79 (8):3106–16. https://doi.org/10.1128/IAI.01275-10.

Seyler, R. W., J. W. Olson, and R. J. Maier. 2001. “Superoxide Dismutase-Deficient Mutants of Helicobacter Pylori Are Hypersensitive to Oxidative Stress and Defective in Host Colonization.” *Infection and Immunity* 69 (6):4034–40. https://doi.org/10.1128/IAI.69.6.4034-4040.2001.

Sgouros, S N, and C Bergele. 2006. “Clinical Outcome of Patients with Helicobacter Pylori Infection: The Bug, the Host, or the Environment?” *Postgraduate Medical Journal* 82 (967). BMJ Publishing Group:338–42. https://doi.org/10.1136/pgmj.2005.038273.

She, Fei Fei, Jian Yin Lin, Jun Yan Liu, Cheng Huang, and Dong Hui Su. 2003. “Virulence of Water-Induced Coccoid Helicobacter Pylori and Its Experimental Infection in Mice.” *World Journal of Gastroenterology* 9 (3). Baishideng Publishing Group Inc:516–20. https://doi.org/10.3748/wjg.v9.i3.516.

Sheh, A., R. Chaturvedi, D. S. Merrell, P. Correa, K. T. Wilson, and J. G. Fox. 2013. “Phylogeographic Origin of Helicobacter Pylori Determines Host-Adaptive Responses upon Coculture with Gastric Epithelial Cells.” *Infection and Immunity* 81 (7):2468–77. https://doi.org/10.1128/IAI.01182-12.

Sheppard, Samuel K, Xavier Didelot, Guillaume Meric, Alicia Torralbo, Keith A Jolley, David J Kelly, Stephen D Bentley, Martin C J Maiden, Julian Parkhill, and Daniel Falush. 2013. “Genome-Wide Association Study Identifies Vitamin B5 Biosynthesis as a Host Specificity Factor in Campylobacter.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (29):11923–27. https://doi.org/10.1073/pnas.1305559110.

Shi, Wen-Jia, and Jin-Bo Gao. 2016. “Molecular Mechanisms of Chemoresistance in Gastric Cancer.” *World Journal of Gastrointestinal Oncology* 8 (9):673. https://doi.org/10.4251/wjgo.v8.i9.673.

Sipponen, Pentti, and Ashley B Price. 2011. “The Sydney System for Classification of Gastritis 20 Years Ago.” *Journal of Gastroenterology and Hepatology* 26 (January):31–34. https://doi.org/10.1111/j.1440-1746.2010.06536.x.

Smith, A C. 1989. “Duodenal Ulcer Disease: What Role Does Campylobacter Pylori Play?” *Scandinavian Journal of Gastroenterology. Supplement* 160:14–18. http://www.ncbi.nlm.nih.gov/pubmed/2683020.

Smith, Malcolm-G, Georgina-L Hold, Eiichi Tahara, and Emad-M El-Omar. 2006. “Cellular and Molecular Aspects of Gastric Cancer.” *World Journal of Gastroenterology* 12 (19):2979–90.

http://www.ncbi.nlm.nih.gov/pubmed/16718776.

Smolka, Adam J., and Mitchell L. Schubert. 2017. “Helicobacter Pylori-Induced Changes in Gastric Acid Secretion and Upper Gastrointestinal Disease.” In , 227–52. https://doi.org/10.1007/978-3-319-50520-6_10.

Sokoloff, Boris. 1938. “Predisposition to Cancer in the Bonaparte Family.” *The American Journal of Surgery* 40 (3):673–78. https://doi.org/10.1016/S0002-9610(38)90653-1.

Son, Seok Hyun, Byung Ock Choi, Gi Won Kim, Suk Woo Yang, Young Seon Hong, Ihl Bohng Choi, and Yeon Sil Kim. 2010. “Primary Radiation Therapy in Patients With Localized Orbital Marginal Zone B-Cell Lymphoma of Mucosa-Associated Lymphoid Tissue (MALT Lymphoma).” *International Journal of Radiation Oncology*Biology*Physics* 77 (1):86–91. https://doi.org/10.1016/j.ijrobp.2009.04.018.

Soybel, David I. 2005. “Anatomy and Physiology of the Stomach.” *Surgical Clinics of North America* 85 (5):875–94. https://doi.org/10.1016/j.suc.2005.05.009.

Spiegelhalder, C, B Gerstenecker, A Kersten, E Schiltz, and M Kist. 1993. “Purification of Helicobacter Pylori Superoxide Dismutase and Cloning and Sequencing of the Gene.” *Infection and Immunity* 61 (12):5315–25. http://www.ncbi.nlm.nih.gov/pubmed/8225605.

Stewart, Bernard W., and Christopher P. Wild. 2014. *World Cancer Report 2014*. http://publications.iarc.fr/Non-Series-Publications/World-Cancer-Reports/World-Cancer-Report-2014.

Suerbaum, Sebastian, and Christine Josenhans. 2007. “Helicobacter Pylori Evolution and Phenotypic Diversification in a Changing Host.” *Nature Reviews. Microbiology* 5 (6):441–52. https://doi.org/10.1038/nrmicro1658.

Sugano, Kentaro, Jan Tack, Ernst J Kuipers, David Y Graham, Emad M El-Omar, Soichiro Miura, Ken Haruma, et al. 2015. “Kyoto Global Consensus Report on *Helicobacter Pylori* Gastritis.” *Gut* 64 (9):1353–67. https://doi.org/10.1136/gutjnl-2015-309252.

Supajatura, Volaluck, Hiroko Ushio, Akihiro Wada, Kinnosuke Yahiro, Ko Okumura, Hideoki Ogawa, Toshiya Hirayama, and Chisei Ra. 2002. “Cutting Edge: VacA, a Vacuolating Cytotoxin of Helicobacter Pylori, Directly Activates Mast Cells for Migration and Production of Proinflammatory Cytokines.” *Journal of Immunology (Baltimore, Md. : 1950)* 168 (6):2603–7.

http://www.ncbi.nlm.nih.gov/pubmed/11884423.

Sutton, Philip, and Joanne M. Boag. 2018. “Status of Vaccine Research and Development for Helicobacter Pylori.” *Vaccine*, April. https://doi.org/10.1016/j.vaccine.2018.01.001.

Sutton, Philip, Christopher Doidge, Gideon Pinczower, John Wilson, Stacey Harbour, Agnieszka Swierczak, and Adrian Lee. 2007. “Effectiveness of Vaccination with Recombinant HpaA from Helicobacter Pylori Is Influenced by Host Genetic Background.” *FEMS Immunology and Medical Microbiology* 50 (2):213–19. https://doi.org/10.1111/j.1574-695X.2006.00206.x.

Sutton, Scott. 2006. “Microbiology Network - Counting Colonies | The Microbiology Network: Dedicated to the Improvement of Regulatory Science and Compliance.” 2006. http://www.microbiol.org/resources/monographswhite-papers/counting-colonies/.

Suzuki, Masato, Keigo Shibayama, Koji Yahara, K. M. Papp-Wallace, A. Endimiani, M. A. Taracila, R. A. Bonomo, et al. 2016. “A Genome-Wide Association Study Identifies a Horizontally Transferred Bacterial Surface Adhesin Gene Associated with Antimicrobial Resistant Strains.” *Scientific Reports* 6 (November). Nature Publishing Group:37811. https://doi.org/10.1038/srep37811.

“TabletsManual.com.” 2017. “Gastritis and Gastric Ulcer.” 2017. http://www.tabletsmanual.com/wiki/read/gastritis.

Takashima, M, T Furuta, H Hanai, H Sugimura, and E Kaneko. 2001. “Effects of Helicobacter Pylori Infection on Gastric Acid Secretion and Serum Gastrin Levels in Mongolian Gerbils.” *Gut* 48 (6). BMJ Publishing Group:765–73. https://doi.org/10.1136/GUT.48.6.765.

Talley, N J, J E Ormand, C A Frie, and A R Zinsmeister. 1992. “Stability of pH Gradients in Vivo across the Stomach in Helicobacter Pylori Gastritis, Dyspepsia, and Health.” *The American Journal of Gastroenterology* 87 (5):590–94. http://www.ncbi.nlm.nih.gov/pubmed/1595645.

Taylor, D N, and M J Blaser. 1991. “The Epidemiology of Helicobacter Pylori Infection.” *Epidemiologic Reviews* 13:42–59. http://www.ncbi.nlm.nih.gov/pubmed/1765119.

Telford, J. L. 1994. “Gene Structure of the Helicobacter Pylori Cytotoxin and Evidence of Its Key Role in Gastric Disease.” *Journal of Experimental Medicine* 179 (5):1653–58. https://doi.org/10.1084/jem.179.5.1653.

Testerman, Traci L, and James Morris. 2014. “Beyond the Stomach: An Updated View of Helicobacter Pylori Pathogenesis, Diagnosis, and Treatment.” *World Journal of Gastroenterology* 20 (36). Baishideng Publishing Group Inc:12781–808. https://doi.org/10.3748/wjg.v20.i36.12781.

Thorell, Kaisa, Shaghayegh Hosseini, Reyna Victoria Palacios Palacios Gonzáles, Chatchai Chaotham, David Y Graham, Lawrence Paszat, Linda Rabeneck, Samuel B Lundin, Intawat Nookaew, and Åsa Sjöling. 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):53. https://doi.org/10.1186/s12862-016-0619-y.

Thorell, Kaisa, Koji Yahara, Elvire Berthenet, Daniel J. Lawson, Jane Mikhail, Ikuko Kato, Alfonso Mendez, et al. 2017. “Rapid Evolution of Distinct Helicobacter Pylori Subpopulations in the Americas.” Edited by Graham Coop. *PLOS Genetics* 13 (2):e1006546. https://doi.org/10.1371/journal.pgen.1006546.

Thrift, Aaron P., Nirmala Pandeya, Kylie J. Smith, Adèle C. Green, Nicholas K. Hayward, Penelope M. Webb, and David C. Whiteman. 2012. “Helicobacter Pylori Infection and the Risks of Barrett’s Oesophagus: A Population-Based Case-Control Study.” *International Journal of Cancer* 130 (10):2407–16. https://doi.org/10.1002/ijc.26242.

Tobias, Joshua, Michael Lebens, Sun Nyunt Wai, Jan Holmgren, and Ann-Mari Svennerholm. 2017. “Surface Expression of Helicobacter Pylori HpaA Adhesion Antigen on Vibrio Cholerae , Enhanced by Co-Expressed Enterotoxigenic Escherichia Coli Fimbrial Antigens.” *Microbial Pathogenesis* 105 (April):177–84. https://doi.org/10.1016/j.micpath.2017.02.021.

Tomasiewicz, Diane M. 1980. “The Most Suitable Number of Colonies on Plates for Counting 1” 43 (4):282–86. http://jfoodprotection.org/doi/pdf/10.4315/0362-028X-43.4.282?code=fopr-site.

Tomb, Jean-F., Owen White, Anthony R. Kerlavage, Rebecca A. Clayton, Granger G. Sutton, Robert D. Fleischmann, Karen A. Ketchum, et al. 1997. “The Complete Genome Sequence of the Gastric Pathogen Helicobacter Pylori.” *Nature* 388 (6642):539–47. https://doi.org/10.1038/41483.

Tombola, Francesco, Laura Morbiato, Giuseppe Del Giudice, Rino Rappuoli, Mario Zoratti, and Emanuele Papini. 2001. “The Helicobacter Pylori VacA Toxin Is a Urea Permease That Promotes Urea Diffusion across Epithelia.” *Journal of*

Clinical Investigation 108 (6):929–37. https://doi.org/10.1172/JCI13045.

Torre, Giuseppe La, Giacomina Chiaradia, Francesco Gianfagna, Angelo De Lauretis, Stefania Boccia, Alice Mannocci, and Walter Ricciardi. 2009. "Smoking Status and Gastric Cancer Risk: An Updated Meta-Analysis of Case-Control Studies Published in the Past Ten Years." *Tumori* 95 (1):13–22. http://www.ncbi.nlm.nih.gov/pubmed/19366050.

Torres, Javier, Pelayo Correa, Catterina Ferreccio, Gustavo Hernandez-Suarez, Rolando Herrero, Maria Cavazza-Porro, Ricardo Dominguez, and Douglas Morgan. 2013. "Gastric Cancer Incidence and Mortality Is Associated with Altitude in the Mountainous Regions of Pacific Latin America." *Cancer Causes & Control : CCC* 24 (2):249–56. https://doi.org/10.1007/s10552-012-0114-8.

Tsang, Jennifer, Takanori Hirano, Timothy R Hoover, and Jonathan L McMurry. 2015. "Helicobacter Pylori FlhA Binds the Sensor Kinase and Flagellar Gene Regulatory Protein FlgS with High Affinity." *Journal of Bacteriology* 197 (11):1886–92. https://doi.org/10.1128/JB.02610-14.

Tsang, Jennifer, Todd G Smith, Lara E Pereira, and Timothy R Hoover. 2013. "Insertion Mutations in Helicobacter Pylori flhA Reveal Strain Differences in RpoN-Dependent Gene Expression." *Microbiology (Reading, England)* 159 (Pt 1):58–67. https://doi.org/10.1099/mic.0.059063-0.

Tsao, Ming-Yang, Tzu-Lung Lin, Pei-Fang Hsieh, and Jin-Town Wang. 2009. "The 3'-to-5' Exoribonuclease (Encoded by HP1248) of Helicobacter Pylori Regulates Motility and Apoptosis-Inducing Genes." *Journal of Bacteriology* 191 (8):2691–2702. https://doi.org/10.1128/JB.01182-08.

Uchiyama, Ikuo, Jacob Albritton, Masaki Fukuyo, Kenji K Kojima, Koji Yahara, and Ichizo Kobayashi. 2016. "A Novel Approach to Helicobacter Pylori Pan-Genome Analysis for Identification of Genomic Islands." *PloS One* 11 (8). Public Library of Science:e0159419. https://doi.org/10.1371/journal.pone.0159419.

Unemo, M., M. Aspholm-Hurtig, D. Ilver, J. Bergstrom, T. Boren, D. Danielsson, and S. Teneberg. 2005. "The Sialic Acid Binding SabA Adhesin of Helicobacter Pylori Is Essential for Nonopsonic Activation of Human Neutrophils." *Journal of Biological Chemistry* 280 (15):15390–97. https://doi.org/10.1074/jbc.M412725200.

Vale, Filipa F., Alexandra Nunes, Mónica Oleastro, João P. Gomes, Daniel A.

Sampaio, Raquel Rocha, Jorge M. B. Vítor, et al. 2017. “Genomic Structure and Insertion Sites of Helicobacter Pylori Prophages from Various Geographical Origins.” *Scientific Reports* 7 (February):42471. https://doi.org/10.1038/srep42471.

Vannini, Andrea, Davide Roncarati, and Alberto Danielli. 2016. “The Cag-Pathogenicity Island Encoded CncR1 sRNA Oppositely Modulates Helicobacter Pylori Motility and Adhesion to Host Cells.” *Cellular and Molecular Life Sciences* 73 (16):3151–68. https://doi.org/10.1007/s00018-016-2151-z.

VFDB. 2017. “VFDB - Helicobacter.” 2017. http://www.mgc.ac.cn/cgi-bin/VFs/genus.cgi?Genus=Helicobacter.

Vincent, Caroline, David A Stephens, Vivian G Loo, Thaddeus J Edens, Marcel A Behr, Ken Dewar, and Amee R Manges. 2013. “Reductions in Intestinal Clostridiales Precede the Development of Nosocomial Clostridium Difficile Infection.” *Microbiome* 1 (1):18. https://doi.org/10.1186/2049-2618-1-18.

Vos, Michiel, and Xavier Didelot. 2009. “A Comparison of Homologous Recombination Rates in Bacteria and Archaea.” *The ISME Journal* 3 (2):199–208. https://doi.org/10.1038/ismej.2008.93.

Walter, Mathias C., Katrin Zwirglmaier, Philipp Vette, Scott A. Holowachuk, Kilian Stoecker, Gelimer H. Genzel, and Markus H. Antwerpen. 2017. “MinION as Part of a Biomedical Rapidly Deployable Laboratory.” *Journal of Biotechnology* 250 (May):16–22. https://doi.org/10.1016/j.jbiotec.2016.12.006.

Wang, G, Z Ge, D A Rasko, and D E Taylor. 2000. “Lewis Antigens in Helicobacter Pylori: Biosynthesis and Phase Variation.” *Molecular Microbiology* 36 (6):1187–96. http://www.ncbi.nlm.nih.gov/pubmed/10931272.

Wang, Hsuan-Chen, Feng-Chi Cheng, Ming-Shiang Wu, Hung-Yu Shu, H Sunny Sun, Yu-Chun Wang, Ih-Jen Su, and Chi-Jung Wu. 2015. “Genome Sequences of Three Helicobacter Pylori Strains from Patients with Gastric Mucosa-Associated Lymphoid Tissue Lymphoma.” *Genome Announcements* 3 (2). https://doi.org/10.1128/genomeA.00229-15.

Whalen, Michael B, and Orietta Massidda. 2015. “Helicobacter Pylori: Enemy, Commensal Or, Sometimes, Friend?” *Journal of Infection in Developing Countries* 9 (6):674–78. http://www.ncbi.nlm.nih.gov/pubmed/26142681.

WHO. 1948. “Preamble to the Constitution of the World Health Organization.” 1948. http://who.int/about/definition/en/print.html.

Wijck, Yolanda van, Stan de Kleijn, Gerrit John-Schuster, Tinne C. J. Mertens, Pieter S. Hiemstra, Anne Müller, Hermelijn H. Smits, and Christian Taube. 2018. "Therapeutic Application of an Extract of *Helicobacter Pylori* Ameliorates the Development of Allergic Airway Disease." *The Journal of Immunology* 200 (5):ji1700987. https://doi.org/10.4049/jimmunol.1700987.

Woolf, C M, and E A Isaacson. 1961. "An Analysis of 5 Stomach Cancer Families; in the State of Utah." *Cancer* 14:1005–16. http://www.ncbi.nlm.nih.gov/pubmed/13786627.

World Gastroenterology Organisation Global Guidelines. 2010. "Helicobacter Pylori in Developing Countries." 2010. http://www.worldgastroenterology.org/guidelines/global-guidelines/helicobacter-pylori-in-developing-countries/helicobacter-pylori-in-developing-countries-english.

Wroblewski, Lydia E., and Richard M. Peek. 2016. "Helicobacter Pylori, Cancer, and the Gastric Microbiota." In , 393–408. https://doi.org/10.1007/978-3-319-41388-4_19.

Xia, Youlin, Yoshio Yamaoka, Qi Zhu, Ivan Matha, and Xiaolian Gao. 2009. "A Comprehensive Sequence and Disease Correlation Analyses for the C-Terminal Region of CagA Protein of Helicobacter Pylori." Edited by Niyaz Ahmed. *PLoS ONE* 4 (11):e7736. https://doi.org/10.1371/journal.pone.0007736.

Xie, Feng, Daria O'Reilly, Ilia L Ferrusi, Gord Blackhouse, James M Bowen, Jean-Eric Tarride, and Ron Goeree. 2009. "Illustrating Economic Evaluation of Diagnostic Technologies: Comparing Helicobacter Pylori Screening Strategies in Prevention of Gastric Cancer in Canada." *Journal of the American College of Radiology : JACR* 6 (5):317–23. https://doi.org/10.1016/j.jacr.2009.01.022.

Yahara, K., Y. Furuta, K. Oshima, M. Yoshida, T. Azuma, M. Hattori, I. Uchiyama, and I. Kobayashi. 2013. "Chromosome Painting In Silico in a Bacterial Species Reveals Fine Population Structure." *Molecular Biology and Evolution* 30 (6):1454–64. https://doi.org/10.1093/molbev/mst055.

Yahara, Koji, Yoshikazu Furuta, Shinpei Morimoto, Chie Kikutake, Sho Komukai, Dorota Matelska, Stanisław Dunin-Horkawicz, Janusz M Bujnicki, Ikuo Uchiyama, and Ichizo Kobayashi. 2016. "Genome-Wide Survey of Codons under Diversifying Selection in a Highly Recombining Bacterial Species, Helicobacter Pylori." *DNA Research : An International Journal for Rapid*

Publication of Reports on Genes and Genomes, March. https://doi.org/10.1093/dnares/dsw003.

Yakoob, J, W Jafri, Z Abbas, S Abid, R Khan, N Jafri, and Z Ahmad. 2009. "Low Prevalence of the Intact Cag Pathogenicity Island in Clinical Isolates of Helicobacter Pylori in Karachi, Pakistan." *British Journal of Biomedical Science* 66 (3):137–42. http://www.ncbi.nlm.nih.gov/pubmed/19839224.

Yamamoto, Yorimasa, Junko Fujisaki, Masami Omae, Toshiaki Hirasawa, and Masahiro Igarashi. 2015. "*Helicobacter Pylori* -Negative Gastric Cancer: Characteristics and Endoscopic Findings." *Digestive Endoscopy* 27 (5):551–61. https://doi.org/10.1111/den.12471.

Yamaoka, Y, M S Osato, A R Sepulveda, O Gutierrez, N Figura, J G Kim, T Kodama, K Kashima, and D Y Graham. 2000. "Molecular Epidemiology of Helicobacter Pylori: Separation of H. Pylori from East Asian and Non-Asian Countries." *Epidemiology and Infection* 124 (1):91–96. http://www.ncbi.nlm.nih.gov/pubmed/10722135.

Yamazaki, S., A. Yamakawa, T. Okuda, M. Ohtani, H. Suto, Y. Ito, Y. Yamazaki, et al. 2005. "Distinct Diversity of vacA, cagA, and cagE Genes of Helicobacter Pylori Associated with Peptic Ulcer in Japan." *Journal of Clinical Microbiology* 43 (8):3906–16. https://doi.org/10.1128/JCM.43.8.3906-3916.2005.

Yang, Jae Jeong, Kwang-Pil Ko, Lisa Y Cho, Aesun Shin, Jin Gwack, Soung-Hoon Chang, Hai-Rim Shin, Keun-Young Yoo, Daehee Kang, and Sue K Park. 2009. "The Role of TNFgenetic Variants and the Interaction with Cigarette Smoking for Gastric Cancer Risk: A Nested Case-Control Study." *BMC Cancer* 9 (1):238. https://doi.org/10.1186/1471-2407-9-238.

Yuan, Xiao-yan, Jin-Jun Yan, Ya-chao Yang, Chun-mei Wu, Yan Hu, and Jian-li Geng. 2017. "Helicobacter Pylori with East Asian-Type cagPAI Genes Is More Virulent than Strains with Western-Type in Some cagPAI Genes." *Brazilian Journal of Microbiology* 48 (2). Elsevier:218–24. https://doi.org/10.1016/J.BJM.2016.12.004.

Zeng, Ming, Xu-Hu Mao, Jing-Xin Li, Wen-De Tong, Bin Wang, Yi-Ju Zhang, Gang Guo, et al. 2015. "Efficacy, Safety, and Immunogenicity of an Oral Recombinant Helicobacter Pylori Vaccine in Children in China: A Randomised, Double-Blind, Placebo-Controlled, Phase 3 Trial." *Lancet (London, England)* 386 (10002):1457–64. https://doi.org/10.1016/S0140-6736(15)60310-5.

Zhang, Rongguang, Guangcai Duan, Qingfeng Shi, Shuaiyin Chen, Qingtang Fan, Nan Sun, and Yuanlin Xi. 2016. “Construction of a Recombinant Lactococcus Lactis Strain Expressing a Fusion Protein of Omp22 and HpaA from Helicobacter Pylori for Oral Vaccine Development.” *Biotechnology Letters* 38 (11):1911–16. https://doi.org/10.1007/s10529-016-2173-5.

Zhang, Songhua, Dong Soo Lee, Rhiannon Morrissey, Jose R Aponte-Pieras, Arlin B Rogers, and Steven F Moss. 2014. “Early or Late Antibiotic Intervention Prevents Helicobacter Pylori-Induced Gastric Cancer in a Mouse Model.” *Cancer Letters* 355 (1). NIH Public Access:106–12. https://doi.org/10.1016/j.canlet.2014.09.010.

Zheng, Yu, Richard J Roberts, and Simon Kasif. 2004. “Identification of Genes with Fast-Evolving Regions in Microbial Genomes.” *Nucleic Acids Research* 32 (21):6347–57. https://doi.org/10.1093/nar/gkh935.