Multi-level characterization of host and pathogen in *Helicobacter pylori*-associated gastric carcinogenesis

Kaisa Thorell

Department of Microbiology and Immunology
Institute of Biomedicine
The Sahlgrenska Academy, University of Gothenburg
Göteborg, Sweden 2014

University of Gothenburg

Multi-level characterization of host and pathogen
in *Helicobacter pylori*-associated
gastric carcinogenesis

kaisa.thorell@gu.se

ISBN: 978-91-628-9168-8

Printed in Gothenburg, Sweden 2014
Ineko

Till min mormor Aase för din okuvliga, smittsamma nyfikenhet

ABSTRACT:

Today, more than half of the world's population is infected with *Helicobacter pylori,* and two to three per cent of these will develop gastric cancer associated with this infection. Gastric cancer is today the third largest cause of cancer mortality worldwide, with more than 700 000 deaths annually, a number that is expected to increase. *H. pylori* is usually acquired in childhood, and establish a lifelong infection in the absence of treatment. However, most infected individuals remain asymptomatic; the causal relationship between *H. pylori* and gastric cancer is complex, affected both by bacterial and host factors, as well as environmental factors.

To study this relationship we took a multi-level approach looking both at the host and bacteria in patients during the early stages of gastric cancer development. We studied patients from a low-risk, and a high-risk population for gastric cancer, Sweden and Nicaragua respectively. Altogether, we investigated the human gene expression, *H. pylori* genomic and transcriptomic, features, as well as microbiota composition, all in material from the same individuals. We also made a smaller study of the surface proteome of two *H. pylori* isolates.

We found the Nicaraguan *H. pylori* isolates to carry and express *in vivo,* several of the established virulence factors associated with increased gastric cancer risk, such as CagA and the s1/m1 VacA genotype. We also identified the adhesin BabA to have a South American variant with a specific selection pressure on the BabA protein in this region. This could have effects on the adhesion properties and consequently, strain virulence in these strains. On the host level, we identified the kynurenine pathway of tryptophan degradation to be differentially expressed during the early stages of gastric carcinogenesis, a pathway that has been described to be involved both in immune modulation and in cancer development. We also identified the loss of acidic chitinase (CHIA) expression as a potential biomarker for pre-cancerous gastric lesions.

This is the first study that in a large scale explores both the human and bacterial gene expression in the same tissue, an approach that presents new possibilities for the understanding of gastric cancer development.

ORIGINAL PAPERS

This thesis is based on the following papers, referred to in the text by their assigned Roman numeral (I-V):

I. Nookaew I, Thorell K, Worah K, Wang S, Hibberd ML, Sjövall H, Pettersson S, Nielsen J, Lundin SB. **Transcriptome signatures in *Helicobacter pylori*-infected mucosa identifies acidic mammalian chitinase loss as a corpus atrophy marker.** BMC Med Genomics. 2013 Oct 11;6:41.

II. Thorell K, Hosseini S, Palacios Gonzáles RV, Chaotham C, Graham DY, Paszat L, Rabeneck L, Lundin SB, Nookaew I, and Sjöling Å.
Identification of a Latin American specific babA variant through whole genome sequencing of *Helicobacter pylori* patient isolates from Nicaragua. Submitted.

III. Thorell K, Karlsson R, Hosseini S, Kenny D, Sihlbom C, Karlsson A, and Nookaew I.
Comparative proteomes of two *Helicobacter pylori* strains using genomics and mass spectrometry-based proteomics. Manuscript.

IV. Thorell K, Bengtson-Palme J, Liu O, Nookaew I, Paszat L, Nielsen J, Lundin SB and Sjöling Å.
***In vivo* analysis of the viable microbiota and *Helicobacter pylori* transcriptome in gastric infection and early stages of carcinogenesis.** Manuscript.

V. Thorell K, Andersson Y, Gatto F, Fassan M, El-Zimaity H, Paszat L, Nookaew I, Sjöling Å, Nielsen J and Lundin SB.
Transcriptome analysis reveals differential expression of kynurenine pathway enzymes in *Helicobacter pylori*-induced gastric inflammation. Manuscript.

PAPERS NOT INCLUDED IN THE THESIS

Other publications during the time of the PhD project.

I. Gustafsson JK, Ermund A, Ambort D, Johansson ME, Nilsson HE, Thorell K, Hebert H, Sjövall H, Hansson GC. **Bicarbonate and functional CFTR channel are required for proper mucin secretion and link cystic fibrosis with its mucus phenotype.** J Exp Med. 2012 Jul 2;209(7):1263-72.

II. Geahlen JH, Lapid C, Thorell K, Nikolskiy I, Huh WJ, Oates EL, Lennerz JK, Tian X, Weis VG, Khurana SS, Lundin SB, Templeton AR, Mills JC. **Evolution of the human gastrokine locus and confounding factors regarding the pseudogenicity of GKN3.** Physiol Genomics. 2013 Aug 1;45(15):667-83.

III. Bengtsson-Palme J, Hartmann M, Eriksson KM, Pal C, Thorell K, Larsson DGJ, Nilsson HR. **Metaxa2: Improved Identification and Taxonomic Classification of Small and Large Subunit rRNA in Metagenomic Data.** *Resubmitted for publication* 2014.

Table of Contents

Abbreviations

APC	Antigen presenting cells
BabA	Blood group antigen-binding adhesin
CagA	Cytotoxin-associated gene A
COX-2	Cyclooxygenase-2
DC	Dendritic cells
EBV	Epstein-Barr virus
ECL cell	Enterochromaffin-like cell
EMT	Epithelial to mesenchymal transition
GGT	Gamma-glutamyl transpeptidase
HPV	Human papilloma virus
HtrA	High temperature requirement A
IARC	Interational association for research on cancer
LPS	Lipopolysaccharide
MAPK	Mitogen-activated protein kinase
MPS	Massively parallel sequencing
NapA	Neutrophil-activating protein A
NCGC	Non-cardia gastric cancer
NK cell	Natural killer cell
OMP	Outer membrane protein
PAMP	Pathogen-associated molecular pattern
PMN	Polymorphonuclear cells
PRR	Pattern recognition receptor
RNS	Reactive nitrogen species
ROS	Reactive oxygen species
SES	Socioeconomic status
T4SS	Type four secretion system
Th1	T helper 1 cell
Th17	T helper 17 cell
TLR	Toll-like receptor
TNF-α	Tumor necrosis factor alpha
Treg	T regulatory cell
VacA	Vacuolating cytotoxin A
WGS	Whole genome sequencing

INTRODUCTION

Gastric cancer development is the effect of chronic inflammation that leads to gradual tissue changes over decades of time. The most notable factor that can cause this type of inflammation is *Helicobacter pylori*, a bacterium that lives in the stomach of about half the world's population. The infection is commonly acquired in childhood and persists for the rest of the life span if not treated.

In most infected individuals the bacterial infection never gives rise to any clinical manifestations, but it can also give rise to gastric and duodenal ulcers, and, in a few percent of cases, lead to gastric cancer. The early symptoms of gastric cancer are usually very vague and most cases are not diagnosed until the cancer has reached an advanced stage and the prognosis is very poor.

The development of gastric cancer has been intensely studied and the course of carcinogenesis is established to follow a certain sequence of tissue changes. While the majority of individuals only develop a mild gastric inflammation (*gastritis*) by the infection, some progress into the stage of *atrophy,* where certain gastric cell populations die, which disrupts the tissue organisation and cell behaviour. At this stage, it is still possible to intervene and lower the risk of further progression. However, the atrophy can develop into *metaplasia*, where the tissue in a patchy manner entirely loses its stomach characteristics, which can further advance into *dysplasia* and finally gastric cancer.

In this study, we have focused on the early stages of the carcinogenesis, namely the *gastritis*, *atrophy*, and *metaplasia* stages, in order to try to delineate what are the changes that occur when the clinical manifestations diverge. In this, we have both studied the gene expression patterns of gastric tissue from patients at these stages, as well as studied the genomes and gene expression of *Helicobacter pylori* bacteria in the same patients.

Theoretical Background:

Gastric physiology

Cell types and anatomy

The stomach is histologically and functionally divided into two main parts, the upper corpus or *oxyntic* part and the lower antrum or *pyloric* part, see figure 1. The basic unit of both of these tissue types is the gastric gland, which consists of an outmost *pit* or foveolae region, with a 20-40 µm single layer of columnar surface epithelial mucous cells, the *isthmus* region where the gland closes, the *neck* region and the *base* region, which is located at the bottom of the gland (figure 2).

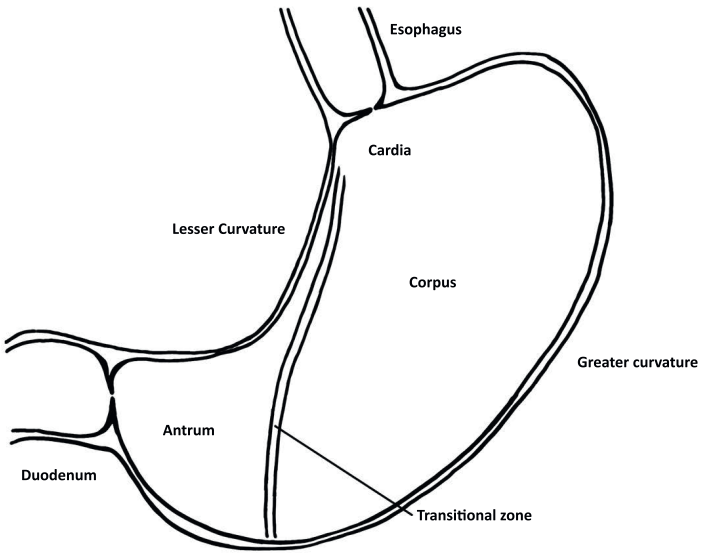


Figure 1: Schematic anatomy of a human stomach.

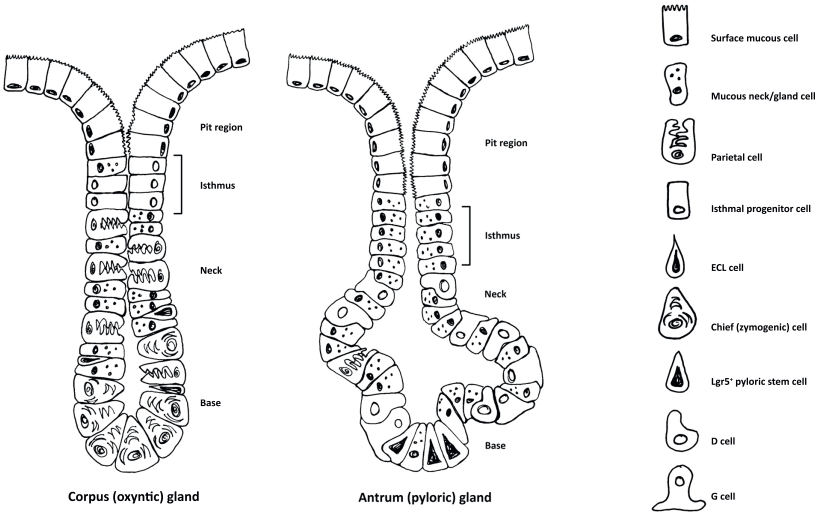


Figure 2: The cell type composition of the oxyntic and pyloric gland

The corpus gland is typically longer and have a small pit region, while the pyloric glands are shorter, with a larger pit region that makes up approximately half on the total gland length [1]. In total the mucosa is 0.5 - 2.5 mm thick, of which the gland part measures approximately 0.65 mm in healthy corpus mucosa, and is slightly shorter in antrum. Each square millimetre of the stomach mucosa comprises approximately 100 glands.

However, there are some major differences in the cell types constituting the glandular regions of the two tissue types. The corpus hosts the majority of the acid secreting parietal cells, which are distributed along much of the length of the gland, with the highest density in the neck and the base. The base of the gland is also rich in chief cells or *zymogenic* cells, which secrete the pro-enzyme pepsinogen. In addition to these two cell types, the corpus glands also contain mucous neck cells and histamine-secreting Enterochromaffin-Like cells (ECL cells), interspersed between the other cells in the neck and base region of the corpus gland (figure 2). The pyloric glands also contain mucous neck cells, and a few parietal and ECL cells, but most importantly it contains the G-cells and D-cells that secrete gastrin and somatostatin, respectively, regulating the corpus acid secretion as well as the response to food and the thought of food (figure 2). In addition, pyloric glands contain endocrine X/A cells producing the hormone ghrelin that

regulate hunger and satiety, and alkaline mucus-producing cells near the bottom of the crypt, which resemble the corpus mucous neck cells in cell-specific marker expression. The pyloric glands are branched and coiled at their basal ends, with longer pits that occupy about half the thickness of the mucosa, and has a higher turnover than the oxyntic mucosa [2].

The border between antrum and corpus is not definite but consists of a hybrid transitional zone where the tissue has characteristics of both tissue types (figure 1).

Tissue regeneration

Another important difference between the mucosa of the antrum and the corpus is the location and characteristics of the stem cells, the progenitor cells that are the origin of all cell types in the gland. In the pyloric gland, similar to intestinal crypts, the progenitor cells that are responsible for the regeneration of healthy mucosa reside in the base of the gland and can be distinguished by their expression of the marker Lgr5 [3]. In the oxyntic gland the progenitor cells reside in the isthmus region, do not express Lgr5, and is dependent on other signalling pathways, as described in detail in a review by Hoffmann in 2013 [1]. The stem cells can be distinguished in the tissue by their high nuclear to cytoplasm ratio, open chromatin and absence of secretory granules [4].

The corpus stem cell regenerates cells that migrate bidirectionally. The progeny cells migrating upwards give rise to presurface cells that populate the surface mucous niche, while the ones migrating downwards give rise to the cell types of the gland. The surface mucous cells of the pit or foveolae are short-lived cells with a life span of 3-5 days in mouse healthy mucosa and thus have a rapid turnover. The parietal and chief cells on the other hand exhibit very different cycling with a life span of months [4]. Unlike parietal cells, chief cells and ECL cells, which are terminally differentiated cell types, the mucous neck cells represent a transitional state between the progenitor cell and the mature, highly secretory chief cell [1].

The gastric stem cell niche is poorly characterised and is believed to be composed of myofibroblasts of the scant mesenchyme between the glandular units as well as endothelial cells or pericytes of the capillary network. The parietal cells are also thought to play a role in maintaining the

niche and guiding the differentiation of the other gastric lineages. Loss of parietal cells, either by chemical ablation or as a consequence of chronic inflammation disrupts the proper differentiation of other lineages such as the chief cell [5].

Helicobacter pylori

General features and habitat

Helicobacter pylori is a spiral-shaped, gram-negative bacterial species that belongs to the family of ε-proteobacteria. It is micro-aerophilic and thus does not thrive in environments with atmospheric oxygen levels. *H. pylori* is trophic for human gastric-type epithelium and preferentially colonises the stomach or metaplastic gastric epithelium of the duodenum.

To cope with the stomach acidity *H. pylori* employs a variety of defence mechanisms, including the production of cytosolic and surface-bound urease, which catalyses the breakdown of urea into carbon dioxide and ammonia, thereby neutralising the surrounding environment. In the absence of urea, *H. pylori* survives at a pH ranging from 4 to 8, while if urea is present, it can survive down to pH levels of 2.5 [6]. However, this mechanism only allows it to cope with high levels of acidity for short periods of time. To overcome this, *H. pylori*'s spiral shape and its bundle of polar flagella also allow it to "swim" in the mucus layer using chemotactic systems. These systems sense among others the pH gradient, and guide the movement away from the acidic lumen towards the epithelium, where the pH is almost neutral. High acidity rapidly impairs the motility of the bacterium, making it crucial for it to quickly penetrate the mucus and establish itself close to the epithelium. To avoid over-alkalinisation of the environment, the uptake of urea into the cell is regulated by a proton-gated channel encoded by *ureI*, which is only active at acidic pH [6]. The alkalinisation also makes the mucus more viscous making it easier for the bacterium to penetrate [7].

H. pylori epidemiology

Helicobacter pylori currently infects half of the world's population with geographical and population-based differences. It is thought that *H. pylori* historically infected virtually all humans; however, its incidence has been

gradually declining in Western and other developed countries, and presently is rare among younger individuals in most Western European countries, North America, Japan, and Oceania [8]. *H. pylori* transmission is believed to occur at a young age through close person-to-person contact, commonly within families or members of communities living close to each other. Intrafamilial transmission is the principal form in developed regions with better hygiene conditions, such as in Japan, where common ways of transmission are mother-to-child or grandmother-to-child [9]. In less developed and rural areas extrafamilial, horizontal transmission is more common [10]. *H. pylori* is not thought to be viable in faeces from healthy individuals but may rather be transmitted via the gastric-oral, oral-oral, or fecal-oral routes in connection to gastrointestinal diseases that cause diarrhoea and/or vomiting [10 11]. However, even though within-family transmission is thought to be the most common mode of transmission, members of the same family can also have totally unrelated *H. pylori* isolates [12].

Even if *H. pylori* can be detected by PCR in samples isolated from sewage and water, viable *H. pylori* bacteria have been harder to retrieve from environmental sources. A handful of studies have managed to culture *H. pylori* from various water sources, such as sea water, wastewater, and lake and river water [13]. However, a study of drinking water and environmental water in a highly endemic area failed to detect *H. pylori* DNA in water [14]. Nevertheless, though the risk of transmission of a fastidious organism like *H. pylori* in water can be debatable, it is clear that the risk of infection by *H. pylori* is linked to socioeconomic status and is particularly prevalent in areas of overcrowding and poor sanitation [15]. Other socioeconomic factors associated with *H. pylori* infection are low family income, low educational level, living in a rural area and having contaminated water sources [9].

Genomic diversity

The genome of *H. pylori* consists of a single chromosome, and based on the 66 isolates with complete genome sequence to date (as of 2014-10-01), approximately one third also carry plasmids. The vast majority of them carry a single plasmid, but isolates carrying 2 plasmids have also been identified. The average genome size among these strains is 1.62 Mbp (1.50-1.72), which is relatively small for a bacterial genome, and they have an average GC content of 38.9%, and 1589 genes.

Helicobacter pylori is naturally competent, meaning it can take up DNA from its environment and integrate it into its own DNA. It shows extensive genomic variation and has the highest mutation and recombination rates of pathogenic bacteria [16]. The main mechanism of variation is natural transformation and subsequent homologous recombination between co-infecting lineages within the same stomach [17], the recombination rate being substantially higher than the mutation rate [18] [12, 19]. Like other bacteria with small genomes, *H. pylori* also has a high frequency of single-nucleotide repeats that are prone to slipped strand mispairing. This commonly affects gene regulatory regions, and thereby provides opportunities for transcriptional variation without a large repertoire of protein transcription factors [20]. Slipped strand mispairing can also occur within genes, which generates inactive transcripts due to out of frame alterations [20].

While point mutations only affect single nucleotides, homologous recombination changes a whole cluster of adjacent base pairs, leading to a much larger effect on allelic diversity. The effect of the high level of variation in *H. pylori,* is that every individual harbours his or her own *H. pylori* population. Within this population, the exchange of genetic material occurs, creating subpopulations with different evolutionary advantages [19]. However, substantial similarities between isolates from individuals of close relationships, support the theory of clonal transmission within families or members of the same communities [12, 21]. Studies of sequential isolates from the same individual and between close relatives show that imports are not randomly distributed over the genome, but tend to appear as groups of imports with stretches of sequence identity in between [19]. A comparison among three isolates from the same family showed that the frequency of imports was significantly higher in the group of outer membrane proteins, specifically of the Hop family, than among other functional categories [21]. In another study comparing Asian isolates, the genes with the highest recombination rates were both metabolic genes such as *trpA*, the gene coding for tryptophan synthase subunit alpha, and L-asparaginase II *ansB*, as well as genes involved in recombination themselves, such as *comE3* and *dprA.* Three genes showing both unusually high diversity and recombination rates were *babA*, *hopZ* and *futB*. The first two are outer membrane proteins involved in adhesion, and *futB* is a fucosyle transferase responsible for variation in surface lipopolysaccharides (LPS). Among the genes showing the lowest recombination rates were genes involved in housekeeping activities

such as translation and transcription, as well as the outer membrane proteins *hopQ* and *hopL* [22]. Recombination rates have also been shown to vary greatly between individuals. In a study on paired corpus and antrum isolates from 52 individuals belonging to two different families in South Africa, it was shown that some of the isolate pairs from the same individual showed large amounts of recombination differences, while others showed weaker signs of recombination [12]. This suggests that there are different evolutionary pressures not only at different genes but also on the bacterial populations within different stomachs.

PHYLOGEOGRAPHICAL ORIGIN

In addition to high genomic variability, *H. pylori* genomes also show clearly traceable phylogeographical features that mirror the migratory paths of human populations throughout history [23]. It has been established that *H. pylori* has accompanied mankind for over 100 000 years and that the genomic traits of *H. pylori* isolates can be used as a proxy to identify the relationship between different human populations. This can be visualised in figure 3A, which depicts human migration routes since humans migrated out of Africa approximately 130 000 years ago [24, 25]. The tight relationship between host and bacterium has lead to the grouping of *H. pylori* isolates into groups based on their ancestral origin, the different groups shown in table 1.

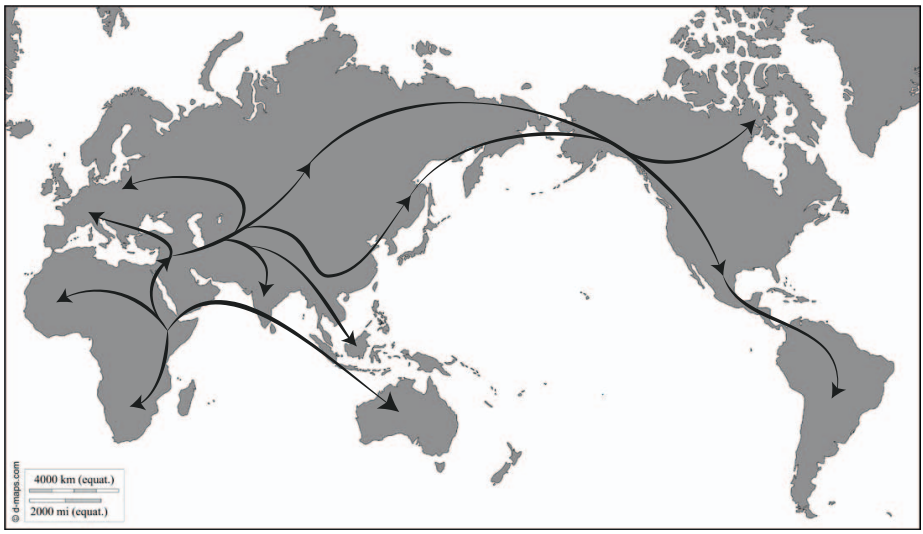
Table 1. Phylogeographical groups of *Helicobacter pylori*

***H. pylori* population**	**Geographical distribution**
hpEurope	Europe, Middle East, India, Iran, Americas
hpAfrica1	
hspWAfrica	Western Africa, Americas
hspSAfrica	South Africa
hpAfrica2	South Africa
hpNEAfrica	Ethiopia, Somalia, Sudan, northern Nigeria
hpEastAsia	
hspEAsia	East Asia
hspAmerind	Native Americans
hspMaori	Taiwan aboriginials, Melanesians, Polynesians
hpAsia2	Northern India, Bangladesh, Thailand, Malaysia
hpSahul	Australian aboriginials, Papua New Guineans

Source: *Correa and Piazuelo 2012*

The phylogenetic relationship between these groups (figure 3B) reflects the ancestry of their carriers. The most notable example is the hpAmerind group of isolates from the indigenous populations of America, which have clear similarities to the hpEastAsia group because indigenous Americans originate from Asian groups migrating over the Bering strait approximately 15 000 years ago [25]. For example, isolates of this origin can be found both in Native North American groups in Alaska, as well as in Mexico,

Guatemala and Peru, where there are still isolated indigenous communities [26]. On the other hand, admixed Latin American populations of *mestizos* and *mulattos,* i.e. people of European and African admixture, tend to carry isolates that show features of the hpEurope and hspWestAfrica groups respectively [26, 27] (figure 3B).



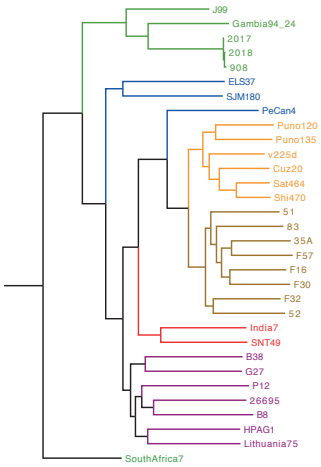


Figure 3. A) World map with arrows marking the main paths of human migration.
B) Phylogenetic tree based on whole-genome SNP comparison.
Colour codes: African strains are shown in *green*, European strains in *purple*, East Asian strains in *brown*, Central Asian strains in *red,* Amerindian strains in *orange,* and urban South American strains are shown in *blue.* SouthAfrica7 is thought to represent the original HpAfrica2 phylogroup. Data on the migratory paths of human populations are taken from [24] and [25]. The background world map is used with permission from http://d-maps.com/m/world/

Interestingly, the origin of the bacterial isolate also seems to affect the risk of severe disease among individuals in the same region, which was demonstrated in a comparison of Colombians with isolates of either African or European ancestry. There, individuals with European strains had more advanced precancerous lesions and higher levels of oxidative damage than those with strains of African ancestry [27].

Virulence factors

Several *Helicobacter pylori* genes have been shown to modulate virulence of bacterial isolates. Virulence factors are bacterial genes that can either enhance the infectivity, persistence, or toxicity of a particular bacterium, as well as the type of immune response it elicits [28].

CagPAI and CagA

The carcinogenic effect of *H. pylori* has been linked to several virulence factors. Most research has focused on the Cag pathogenicity island (*cagPAI*), the factor that appears to have the strongest association to increased cancer risk. *CagPAI* is a genetic element of exogenous origin that has been inserted into the glutamate racemase gene and is made up of 27-31 genes, with a total size of approximately 37 kb [29]. These genes encode structural components of a type four secretion system (T4SS) and an effector protein, the cytotoxicity associated virulence factor CagA. The T4SS pilus is used to inject CagA into the host cells, where it interferes with a series of host proteins and signalling pathways. The presence of this pathogenicity island varies between geographical regions, where almost all East Asian isolates but only approximately 60-70% of Western isolates are *cagPAI* positive [30].

The actions of CagA can be divided to those that are dependent of its phosphorylation inside the host cell, and those that are phosphorylation-independent. The crucial region for the phosphorylation of CagA is a number of EPIYA (glutamic acid-proline-isoleucine-tyrosine-alanine) motifs in the C-terminal region. The tyrosines of these motifs can be phosphorylated by host Src-family kinases, and show a variability that has been linked to strain virulence and geographical origin [29]. On the basis of sequences flanking the EPIYA motifs, 4 different segments, termed EPIYA A, B, C and D, have been described (table 2). Most CagA positive strains have the A and B segments, while EPIYA C is characteristic of strains of European origin and is thus

termed Western CagA. EPIYA D is specific to CagA in East Asian strains and consequently known as East Asian CagA. The East Asian CagA has been shown to have more oncogenic potential than Western CagA, but Western CagA can also vary in the number of C regions, where more repeats are correlated to greater oncogenic potential [31].

Table 2. EPIYA sequence motifs

Motif	Sequence
EPIYA-A	EPIYAKVNKKK(A/T/V/S)GQ
EPIYA-B	EPIY(A/T)(Q/K)VAKKVNAKI
EPIYA-C	EPIYATIDDLG
EPIYA-D	EPIYATIDFDEANQAG

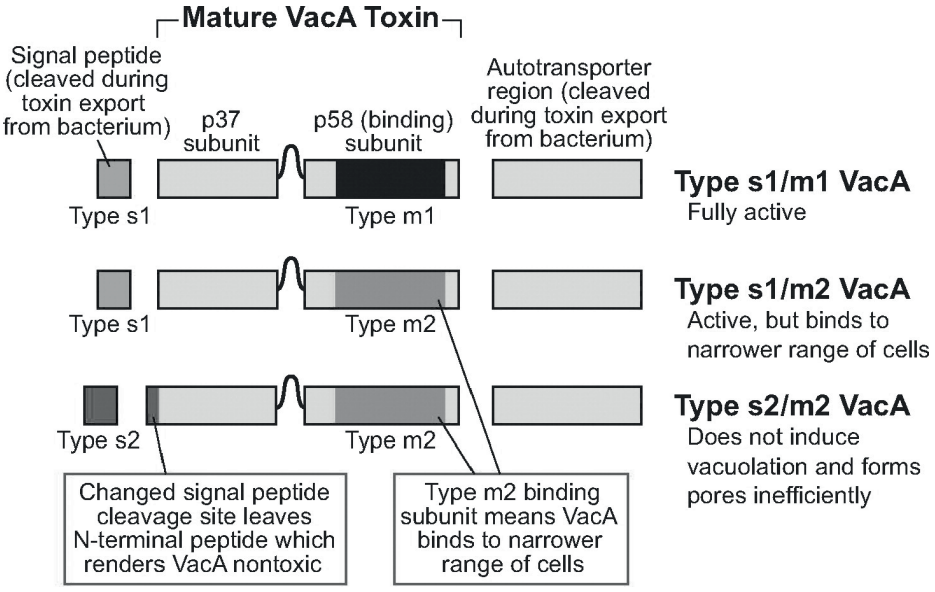
The effects of CagA on host cells include dysregulation of cell-cell adhesion and loss of polarisation in the epithelial cell, cellular elongation that resembles epithelial to mesenchymal transition (EMT), and the activation of Ras-Erk cascade The latter increases interleukin 8 (IL-8) release and the activation of NFκB, which leads to an increase in tumour necrosis factor alpha (TNF-α) and cyclooxygenase 2 (COX-2) pathway, among other effects [32]. Within the cell, phosphorylated CagA activates the SHP-2 phosphatase and the MAPK-ERK signalling cascade, causing effects resembling those induced by growth factor signalling. Reported phosphorylation-independent effects include the interaction with E-cadherin, the c-met receptor, and phospholipase C, which disrupts cell-cell junctions and induces loss of cell polarity, and pro-inflammatory and mitogenic effects [33].

VacA

Along with CagA, the vacuolating cytotoxin (VacA) is the best-studied virulence factor. Unlike the *cagPAI,* the *vacA* gene is present in virtually all *H. pylori* isolates, and most isolates also express the VacA protein [34]. This protein forms hexameric, anion-selective pores through lipid bilayers such as the cytoplasmic and organelle membranes and, as the name implies, induces vacuolisation in the host epithelial cells. Although the gene is ubiquitous, there is a substantial sequence variation within the gene, and variants in several regions have been linked to the severity of disease [35]. These variations are localised to the signal (s)[36], middle (m)[36] and intermediate (i)[37] region of the gene (figure 4). The distribution of these alleles also have geographical differences; s1a is the most common in Northern Europe, s1b predominates in the Iberian peninsula and in Latin America, s1c only is found in South Eastern Asia, and s2 is most common in regions with low risk of gastric cancer such as Australia and North America [30]. The s2 type of VacA has

been shown to be virtually nontoxic [38], while strains with the s1/m1/i1 genotype have been shown to be associated with a higher risk of advanced disease and are found in regions with higher gastric cancer risk. However, this genotype commonly coincides with *cagA* positivity, making the relative contribution of the two virulence factors in these strains hard to evaluate [31].

VacA can be transferred to host epithelial cells either by secretion or by contact-dependent transfer and is cleaved during its transport through the bacterial outer membrane [38]. The precise mechanism of VacA secretion and entry to the host cell is still controversial but both epithelial derived growth factor receptor (EGFR), the receptor tyrosine phosphatase alpha and beta (RPTPa/RPTBb), and sphingomyelin in lipid rafts have been shown to act as interaction partners [39]



Atherton JC. 2006.
Annu. Rev. Pathol. Mech. Dis. 1:63–96

Figure 4. The vacA gene including localisation of the cleavage sites and variable regions. Used with permission from John Atherton**.**

Apart from its capacity to induce vacuolisation by forming pores in endosomes, VacA can also induce apoptosis in epithelial cells and

lymphocytes by interfering with membrane trafficking, leading to loss of mitochondrial membrane potential, mitochondrial instability and the subsequent release of cytochrome C [40]

VacA interacts with epithelial cells as well as with immune cells, including B-cells, T-cells and phagocytes. In phagocytes, VacA can inhibit processing and presentation of antigen peptides to T-cells; however, VacA can also interfere with T-cell function directly by blocking antigen-dependent proliferation or by inhibiting the activation of nuclear factor of activated T-cells (NFAT) [41].

OTHER VIRULENCE FACTORS:

Aside from VacA and the Cag pathogenicity island, several other factors have been proven to play a role in the virulence and severity of *H. pylori* infection.

The neutrophil activating protein of *H. pylori (*NapA, or HP-NAP) is a secreted virulence factor with several functions. It can pass through the epithelial layer into the lamina propria to attract leukocytes, and can cause the recruitment of leukocytes to the site of infection by inducing reactive oxygen species (ROS) release from neutrophils, triggering the production of chemokines [42]. More about HP-NAP is described below in the section about the immune response to *H. pylori* infection.

High temperature requirement A (HtrA) is a serine protease that is well conserved in gram-negative bacteria and has been described to increase the viability of the organism under stressful conditions. *H. pylori* HtrA is secreted into the extracellular space in its active form where it can cleave E-cadherin, a tumour suppressor commonly lost in several cancer settings, that is involved in cell-cell adhesion [43]. The extent to which the cleavage of E-cadherin affects E-cadherin signalling and function is not yet known, but this cleavage has also been shown to allow for bacterial entry into the intercellular space by disrupting the adherence junctions [44].

Gamma-glutamyl transpeptidase (GGT) is a virulence factor that has gained increased attention in recent years. GGT is an enzyme that converts glutamine into glutamate and ammonia, and glutathione into glutamate and cysteinylglycine. This has been shown to lead to glutamine and glutathione consumption in the host cells, which interferes with the oxidative capacity of the cell, resulting in the production of ammonia and ROS. These products affect many central cell functions, inducing cell-cycle arrest, apoptosis, and necrosis in gastric epithelial cells [45]. GGT can also induce immune tolerance

by influencing dendritic cell differentiation and inhibition of T cell-mediated immunity, affecting the efficiency of the immune response towards *H. pylori* [45].

Adherence

Bacteria that are in contact with the host cells are generally thought to be the ones primarily contributing to disease in terms of epithelial cell damage and induction and persistence of the immune response. While a majority of the *H. pylori* bacteria reside in the mucus layer close to the epithelium, approximately 20% can be found adhering to the epithelium [40]. This binding is mediated by several factors and the bacterium seems to have a tropism for cell-cell junctions between epithelial cells [46]. *H. pylori* has also been detected by several investigators inside epithelial cells and in the lamina propria; however, how this is mediated and what role it plays in the pathogenesis or gastric cancer development is not yet established [47].

Despite the relatively small genome of around 1500 genes, *H. pylori has* over 80 genes that encode outer membrane proteins (OMPs), several of which have shown characteristics of adhesins [48, 49]. The outer membrane proteins are hot spots for recombination and frequently contain single-nucleotide repeats, making them prone to high variability due to slipped strand mispairing, as discussed above. Together, these factors highlight the importance of extensive heterogeneity and redundancy in this system, which allows the organism to not be dependent on any single adhesion factor for colonisation. These mechanisms provide the capacity for on/off switching in adhesion repertoire and allow for a large stochastic variation in expression of the different factors, giving the population within a stomach plasticity to adapt to the dynamic environment [50].

BabA

The two most important and well-characterised adhesins in *H. pylori* are the blood group antigen-binding adhesin BabA and the sialic-acid binding adhesin SabA. These two proteins lack homologues in other bacterial species and bind to glycosylated structures of mucins and on the epithelial cell surface.

BabA mediates the binding of *H. pylori* to fucosylated structures including the blood group O antigen Lewis B (Le^b) and the related H1 antigen. These antigens can be found on blood cells and on gastric epithelial cells and

mucins [51]. The binding characteristics of BabA have been shown to vary between different isolates, where some isolates have a specific preference for blood group O antigens, so called "specialist" strains, while other isolates equally well binds to fucosylated blood group A antigens, termed "generalists" [52]. These groups are unevenly distributed geographically; the specialist strains are found predominantly in South American countries, where blood group O has dominated historically in the local population. This suggests that there has been a specific evolutionary pressure on isolates from this region to optimise for binding to O antigens and that there are no evolutionary disadvantage associated with a narrowed specificity in this region, which have most likely been the case in regions where the blood groups are more evenly distributed [52].

Expression of the *babA* gene is regulated by phase variation and recombination with the highly homologous genes *babB* and *babC*. The *babA* gene and regulatory regions are very variable and different strains show different levels of babA RNA expression. In addition, isolates expressing similar levels of the protein may still have very different binding characteristics and binding affinities [52, 53].

SabA

SabA binds to sialylated structures such as the sialyl-Lewis X/A (s-$Le^{x/a}$) antigens also found on mucins and epithelial cells [54, 55]. However, these structures are essentially absent in healthy mucosa but the expression is upregulated during inflammatory states such as the one induced by *H. pylori* infection [54].

SabA also binds to sialylated receptors on neutrophils that invade the tissue upon inflammation, which leads to phagocytosis of the bacterium through non-opsonic activation and subsequent induction of the oxidative burst response [56].

SabA is transcriptionally repressed by the acid-sensitive ArsRS system and is therefore down regulated during acidic conditions [57]. In addition to this direct regulation by transcription factors, expression of the *sabA* gene is regulated through slipped strand mispairing through length variation in the poly-T repeat tract that lies upstream of the promoter element of *sabA* [50, 58]. This variation leads to changes in the DNA structure, that affect the binding of RNA polymerase to the *sabA* promoter, thereby affecting the efficiency of

transcription initiation [58]. Additionally, the gene contains cysteine-thymidine (CT) repeats in the 5' part of the coding sequence, causing on/off phase variation similar to that seen in HopZ (described below). It is also, similarly to BabA, subject to gene conversion through recombination with the homologous gene *sabB* [59]. These dynamic adaptations may allow *H. pylori* to specialise for individual host variation in mucosal glycoprotein sialylation during persistent infection.

Recently, the crystal structure of the extracellular parts of the SabA protein was resolved. The structure is dominated by alpha helices and resembles a club with a handle and a head [60]. The head part contains the ligand-binding cavity, which is constrained by two highly conserved disulphide bridges. One of these pairs of conserved cysteines is also present in BabA, but the major part of the ligand-binding region is not homologous between the two proteins [60].

Other adhesins

In addition to the two major adherence proteins, a number of other OMPs have been shown to function as adhesins, including AlpA, AlpB, HopZ, OipA, SabB, BabB, and HopQ, all of which belong to the major *Hop* family of OMPs. The adherence-associated lipoproteins AlpA (HopC) and AlpB (HopB) are highly related and can be found in basically all *H. pylori* isolates. They are transcribed from the same operon, and their loss impairs bacterial binding to the apical side of human gastric tissue sections, and in animal models. Unlike other OMPs, they are not subjected to phase variation but are expressed in virtually all clinical isolates [39]. The ligand of AlpA and AlpB has not been determined, although extracellular matrix laminins have been proposed as possible candidates, and the understanding of their role in human infection is still incomplete.

The outer inflammatory protein A OipA (HopH) has been proposed to amplify IL-8 secretion and activate β-catenin, in parallel to cagPAI [61]. OipA is phase-variable and can be switched "on" and "off" by slipped strand mispairing during chromosomal replication [62]. OipA expression status is associated with the presence of cagPAI and VacA s1m1 alleles in western isolates, and it has therefore been difficult to assess the separate influence of OipA on clinical manifestations [62]. Like AlpA/AlpB, the host surface receptor/interaction partner of OipA has not yet been identified. However, OipA has been

suggested to induce phosphorylation of focal adhesion kinase (FAK), leading to downstream activation of MAPK/Erk signalling and to be involved in the EGFR – PI3K – PDK1 – Akt signalling cascade activation in host cells [39].

Other Hop proteins that have been implicated in adhesion are HopZ and HopQ. HopZ has been shown to be involved in the early phase of colonisation and is regulated by phase variation of CT repeats in the region encoding for the signal sequence. HopZ ON status have been shown to be strongly selected for during early infection[6], and a majority of individuals experimentally infected with a HopZ OFF strain showed a switch in HopZ status 6 and 10 weeks post-infection. Similar findings were for natural familial transmission [63]. In contrast, HopZ status was very stable during chronic infection [63]. HopQ has also been implicated in binding to epithelial cells, but the binding partner has still not been identified. However, in a recent screening for proteins influencing the T4SS-dependent induction of NF-κB activation, HopQ was identified to be an important accessory factor to the T4SS. The same study showed that HopQ is crucial for the CagA translocation and CagA-associated responses in infection of AGS cells by the strain P12 [64]. Deletion of HopQ also led to a reduction of MAPK signalling and IL-8 secretion, without changes in adherence of the bacteria to infected AGS cells [64].

Stomach Microbiota

Due to its high acidity, the stomach was long thought to be a sterile environment, although the first observations of spiral shaped bacteria was described over a century ago [65]. It was not until in 1983, when Barry Marshall and Robin Warren demonstrated the link between *Helicobacter pylori* and gastric inflammation and ulcers, that this bacterium was recognised as an actual inhabitant of the stomach and not just a contaminant or transient infection [66]. However, this observation only slightly altered the dogma and it was subsequently believed that nothing but *H. pylori* could survive in the stomach.

Nonetheless, even early culturing studies indicated the presence of other bacteria in the stomach, especially in states of reduced acid secretion, whether induced by drugs, premalignant conditions such as atrophy, or gastric cancer [67]. Despite this, it was not until the last decades, with the emergence of sequence-based techniques, that the notion of the gastric

microbiota began to be established and more thoroughly investigated. Even as recently as six years ago, an otherwise very thorough review on *H. pylori* and host-bacterial interaction stated that "*H. pylori* is the only known organism capable of colonising the harsh environment of the human stomach" [46].

During this last decade, interest in the gastrointestinal microbiota have increased greatly, and commensal microbes have been shown to influence host metabolism and the development of immune system. The microbiota can also alter cancer risk by inducing oxidative and nitrosative stress and DNA damage, increasing cell proliferation and by producing mutagenic metabolites [68]. Several studies have investigated the microbiota of stomach tissue and gastric juice of humans and different animals models. This has been done using both culture-dependent methods and culture-independent PCR- and sequencing-based techniques [69]. These studies have revealed a previously unappreciated amount of microbes, with colonisation densities of around 10^1-10^3 colony forming units (cfu)/g; however, this is quite modest compared to the heavily colonised colon, which harbours 10^{12}-10^{14} cfu/g [70]. The most prominent phyla detected are Firmicutes, Actinobacteria, Bacteroidetes, and Proteobacteria [71, 72]. In *H. pylori*-colonised individuals over 90% of the sequence reads consists of *H. pylori*, leading to a lower microbial diversity [68]. However, variability between individuals has been shown to be extensive, complicating the comparison between *H. pylori*-infected and uninfected, as well as among different disease groups such as atrophy, metaplasia, and cancer patients [73]. Nevertheless, some biologically and clinically interesting differences have been found. Two recent studies have investigated the differences in microbiota composition among individuals with chronic gastritis, intestinal metaplasia (IM) and gastric cancer (GC). The first study used PhyloChip microarrays to analyse a total of 15 patients and identified a gradual change of the microbiota from gastritis to IM to GC; bacterial diversity was significantly higher in gastritis than in GC, with IM patients showing intermediate diversity. They also found the genus *Pseudomonas* to be significantly more abundant in GC than in gastritis patients [74].

The other study included 31 patients of the same groups but used 16S rRNA amplification followed by 454 pyrosequencing. This study also identified significant differences among the groups, especially in the *H. pylori*-dominated individuals in each group. The investigators also found the relative

abundance of the *Helicobacteraceae* family to be significantly lower in the GC group compared to the chronic gastritis and IM groups, while finding significant increases in the relative abundance of *Streptococcaceae* family in the GC group [75].

Due to several efforts in investigating the stomach microbiota it is now evident that the interindividual variations are large. Since different methodologies provide information at different levels, the results also are hard to compare. It is hoped that the decreasing cost of sequencing will allow the inclusion of larger sample sizes, which would give much needed power in the estimates of diversity and differences between clinical outcomes, an aspect that is currently lacking.

GASTRIC DISEASE AND ITS CONNECTION TO *H. PYLORI*

It has been estimated that approximately half of the world's population is chronically infected with *H. pylori*. For the absolute majority of individuals, the infection remains asymptomatic for the entire lifespan, where *H. pylori* might be better called a commensal organism. However, the infection also has clinical manifestations, including gastric and duodenal ulcers, which occur in 10 - 15% of infected individuals, as well as gastric cancer, which develops in 1 - 3%.

INFLAMMATION INDUCED BY *H. PYLORI*

Helicobacter pylori infection leads to inflammation through the activation of epithelial cells and immune cells such as macrophages and dendritic cells. These cells either reside in the tissue, or are recruited to the site of infection. In general terms, the inflammatory response to *H.* pylori involves both mononuclear and polymorphonuclear cells of the innate immune system and induces the secretion of pro-inflammatory cytokines such as IL-1β, IL-6, IL-8, TNF-α, and anti-inflammatory transforming growth factor beta (TGF-β) [76]. The adaptive immune response is predominantly Th1-mediated and includes both cellular responses and humoral responses of the IgA, IgM and IgG antibody isotypes [33].

The inflammation is induced by many different mechanisms such as specific interactions between virulence proteins and the host cells, and by general pathogen-associated molecular pattern (PAMP) recognition. The latter mechanism plays an essential role in the activation of the innate immune system; key pattern recognition receptors (PRR) for bacterial infections are the Toll-like receptors (TLR) and intracellular NOD-like receptors (NLR). TLRs have been identified on gastrointestinal epithelial cells; however, under homeostatic conditions they are mainly localised to the basolateral side of the cells. The main TLRs involved in response to *H. pylori* infection are TLR4, which recognises lipopolysaccharide (LPS) on the bacterial cells, TLR2, which recognises for example lipoproteins, lipoteichoic acid, and peptidoglycan, and TLR5, which binds flagellin. The intracellular TLR9 also plays a role in the recognition of *H. pylori* by sensing bacterial DNA after phagocytosis [77]. However, TLR9-signalling in response to *H. pylori* seems to have an immune suppressive role [78]. Along the same lines, *H. pylori* LPS is relatively anergic due to its lipid component and does not activate TLR4 to the same extent as other gram-negative bacteria. Its flagella also do not induce a strong TLR5 response, unlike, for example *Escherichia coli* or *Salmonella* species [41].

Cells of the Innate immune system

A characteristic feature of *H. pylori*-induced inflammation is the infiltration of polymorphonuclear cells (PMN), or neutrophils into the gastric mucosa (figure 5B). These are recruited by host effector cytokines such as CXCL8 (IL-8) and specifically by the bacterially secreted virulence factor HP-NAP, which was described previously. HP-NAP also facilitates neutrophil adhesion to endothelial cells [33]. Although *H. pylori* actively induces the recruitment of neutrophils and stimulates their production of reactive oxygen species (ROS) through HP-NAP, it also evades the PMN-dependent killing by several mechanisms that manipulate phagocytosis. These include delayed phagocytosis by T4SS action and the disruption of PKC-ζ signalling. The glycosylation of surface cholesterol also allows *H. pylori* to evade phagocytosis [79]. Lastly, the bacterium can survive within PMN by interfering with the assembly of NADPH oxidase system, thereby impeding the formation of ROS inside the phagosomes. Instead, the ROS are released into the extracellular space, increasing local tissue damage and the release of essential nutrients [41].

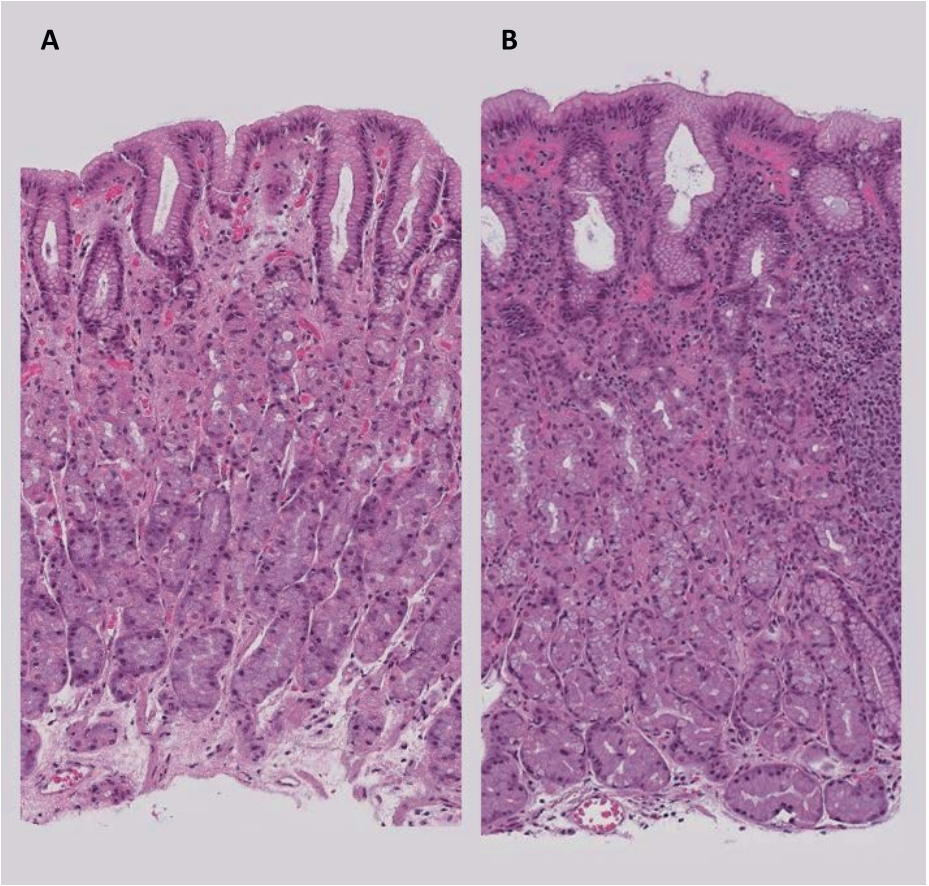


Figure 5. Examples of normal and gastritis tissues stained with haematoxylin and eosin (H&E). A) Normal corpus tissue. B) – Gastritis with infiltrating PMNs and lymphoid aggregates (to the right in the picture).

Macrophages and monocytes are essential in the innate response to *H. pylori*-derived factors, and to signalling from epithelial cells that are in direct contact with the bacteria. Together with dendritic cells (DC), they are important coordinators of the immune response to bacterial infections and are involved in the activation of the adaptive immune system by their release of IL-12 that stimulate Th1 cells. They also amplify the inflammatory response by the secretion of acute phase cytokines IL-1, TNF-α, and IL-6. As seen in PMN, *H. pylori* actively interferes with macrophage effector function, especially the production of bactericidal nitric oxide (NO). This effect is exerted by competing for the substrate for NO formation, L-arginine, by bacterial enzyme expression and bacteria-induced upregulation of host enzymes of

competing pathways. The VacA protein also interferes with macrophage function by disruption of vesicle trafficking; this prevents lysosome from fusing with phagosomes, impairing the capability of the phagocyte to kill engulfed bacteria [33].

Dendritic cells (DC) are pivotal in the immune response to *H. pylori* because they are the major antigen presenting cells (APC) and serve as the bridge between the innate and adaptive immune systems. DC can disrupt cell-cell junctions and sample antigens from the stomach lumen, and infection with *H. pylori* increases the numbers of DC residing in the gastric mucosa. Antigen sampling leads to activation, maturation and migration of DC towards the draining lymph nodes, where they can present the antigen to naïve T-cells. DC also carry PRR sensing conserved bacterial structures such as those mentioned above induces the secretion of the chemokines IL-8 (CXCL8) and CXCL1 which recruit neutrophils [76]. DC will stimulate T-cells that they encounter into different differentiation fates, depending on how it has been activated itself. For example, DC IL-12 secretion at antigen presentation will skew the differentiation into a Th1 phenotype, while IL-23 stimulates Th17 differentiation. Several *H. pylori* antigens have been shown to influence DC in this respect. HP-NAP and HpaA have been shown to increase IL-12 and IL-23 release from macrophages affecting DC maturation and antigen presenting capacity respectively, thereby promoting Th1 differentiation [41]. However, other antigens instead inhibit IL-12 production, inducing a more tolerogenic DC phenotype and promoting regulatory T cell (Treg) induction [76]. In addition, Lewis antigen on *H. pylori* LPS can bind the DC surface molecule DC-SIGN and block Th1 responses [41]. Exposure of DC to *H. pylori* bacteria *in vitro* has been shown to skew the pattern of T-cell differentiation from Th1/Th17 fate to Treg, and the depletion of the systemic DC population actually increases the potency of the immune response to *H. pylori* [80]. The effect on DC maturation is not dependent on the T4SS or CagA and stands in sharp contrast to the pro-inflammatory DC response induced by, for example *E. coli* LPS [80].

Cells of the Adaptive Immune System

The secretion of cytokines and other inflammatory mediators by innate immune cells ultimately leads to an activation of adaptive immunity, including T- and B-cell subsets. T-cells differentiate into diverse subsets upon activation by APC, including T helper 1 (Th1), T helper 2 (Th2), T helper 17 (Th17), and T regulatory (Treg) cells. The fate of these cells depends on many factors; ultimately, however, co-stimulatory molecules and

microenvironmental cytokines combine to activate different transcriptional programs for the different subtypes.

Among the T-cell subpopulations, T helper cells are of particular interest in *H. pylori*-associated inflammation, although cytotoxic $CD8^+$ T-cells also play a role [81]. *H. pylori*-related inflammation has been suggested to be dominated by a Th1 type response, but Th17 and Treg cells provide an important contribution [82, 83]. An efficient Th1 response would be expected to clear the *H. pylori* infection, and a clear correlation has been observed between the Th1-associated inflammatory response, including interferon gamma (IFN-γ) and IL-2 secretion, the concomitant activation of tissue-resident macrophages , and a decrease in the levels of colonising *H. pylori* [76].

Instead, the balance between more detrimental and tissue-damaging Th1 and Th17 responses and the milder but inefficient Treg response that sustains the chronicity of the infection seems to be key in the *H. pylori* pathogenesis. This balance accounts for variation in clinical outcomes and the creation of an inflammatory environment that promotes the initiation and progression of cancer [42].

B-cells are sometimes neglected in descriptions of *H. pylori* induced inflammation but also play an important role. Most interest in B-cells in connection to *H. pylori* has been regarding the development of MALT lymphomas (see below), which present with an aberrant expansion of mucosa-associated lymphoid tissue, largely consisting of immature B-cells [38]. The B-cells are also believed to play a role in the protective immunity against *H. pylori* by secreting IgA and IgG antibodies, but the contribution of this mechanism is not fully understood [33]. However, at least in mice, there is an induction of IL-10-producing regulatory B-cells during early stages of *H. pylori* infection. This further suggests that *H. pylori* may induce several regulatory immune responses in order to survive, not just Treg [84]. IL-10-producing B-cells may also play an important role in the immune suppression at an early stage of infection, prior to the induction of Treg cells.

One of the central transcription factors that are activated by H. pylori infection is NF-κB. Upon release, NF-κB enters the nucleus, where it acts as a transcription regulator of a large number of genes including cytokines,

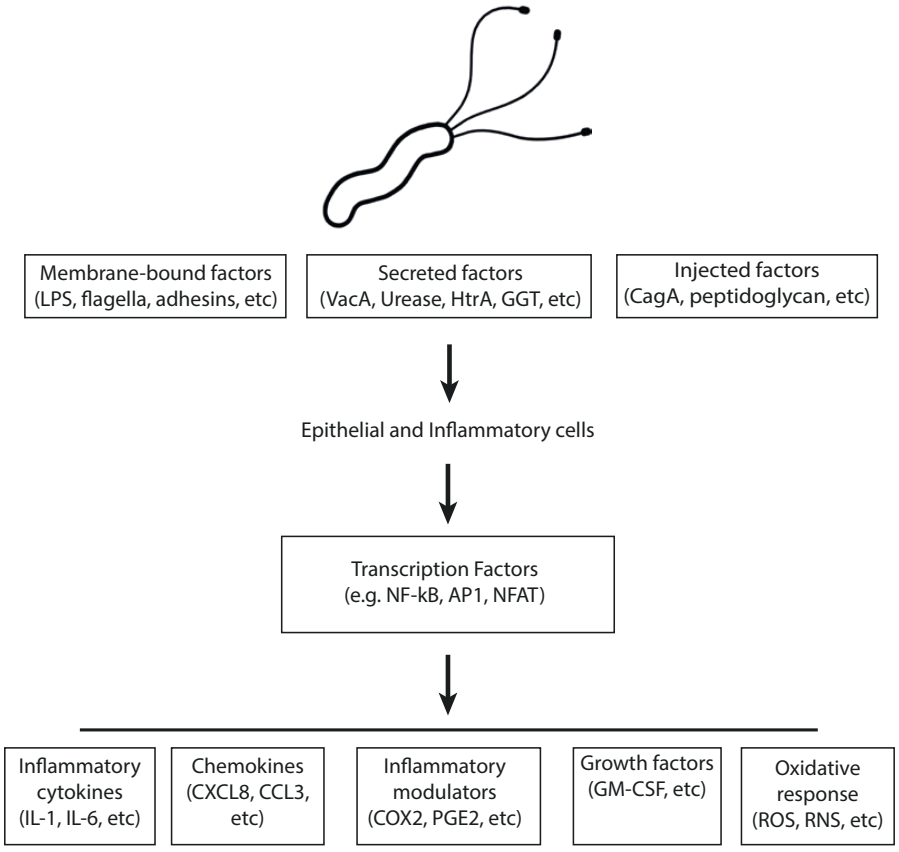


Figure 6. Summary of how *Helicobacter pylori* induces inflammation in the gastric mucosa. Abbreviations; Interleukins IL-1 and IL-6, IL-8 (CXCL8), macrophage inflammatory protein (CCL3), cyclooxygenase (COX) 2, Prostaglandin (PG) E2, Granulocyte-macrophage colony-stimulating factor (GM-CSF), Reactive oxygen- (ROS) and nitrogen (RNS) species.

chemokines, and other genes involved in inflammation, growth and survival of the cell [42]. *H. pylori* can activate NF-κB both the canonical and by the non-canonical pathway, which leads to a different gene regulation pattern. The mode of activation is dependent on the cell type; *H. pylori* activates the canonical pathway in epithelial cells, while it can activate both pathways in immune cells [42]. Contact between *H. pylori* and gastric epithelial cells leads to a rapid induction of NF-κB. This effect is increased in infections with *cagPAI-*

positive strains and has been attributed in part to the effects of CagA, which activates NF-κB through the c-Met-PI3K-AKT pathway. NF-κB can also be induced through the activation of NOD receptors by bacterial peptidoglycan (also delivered by the T4SS), by LPS-mediated activation of TLR4 and TLR2 [42], or by TLR2 activation by HP-NAP [41]. The way *H. pylori* activates the immune response is summarised in figure 6.

Tissue changes

The chronic inflammation described above induces extensive alterations in the microenvironment of the gastric mucosa as illustrated in figure 6. These changes interfere with the molecular signals that orchestrate cell type differentiation from the stem cells in the gastric glands and can lead to alterations in the tissue composition. One hallmark of progressing gastric inflammation is the presence of atrophic gastritis, which marks the first step of the gastric pre-cancerous process. Atrophy of the corpus mucosa is characterised by the loss of the parietal cell population [4]. The progression of atrophic gastritis occurs as an advancing front of the transitional zone that borders to more normal corpus epithelium. This advancing front leaves an increasingly large area of atrophic mucosa, and its presence is associated to an increased risk of cancer development, and is more prevalent in populations with high gastric cancer incidence [85].

Loss of the parietal cells, as an effect of the chronic *H. pylori* infection, or by autoimmune gastritis, disrupts the cell-cell signalling in the glands, leading to loss of organisation of the gland architecture (figure 7A). This mainly affects the differentiation of the chief cells, which require the presence of the parietal cells to reach their terminally differentiated state [86], and results in abnormal differentiation of the tissue, termed metaplasia. Two types of metaplasia is described in the precancerous cascade of the stomach; antralisation of the tissue - so called pseudopyloric or spasmolytic polypeptide expressing metaplasia (SPEM), and the intestinal metaplasia (IM) (figure 7B). While SPEM always develops as a result of the parietal cell loss and has a more diffuse pattern, the intestinal metaplasia also require the presence of inflammation [87] and develop in a more spotty, multifocal manner [88]. The respective premalignant role of SPEM and IM on the cancer development is under debate, the major question being if the dysplastic and malignant changes actually arise from IM tissue or if IM rather is a paramalignant manifestation [89].

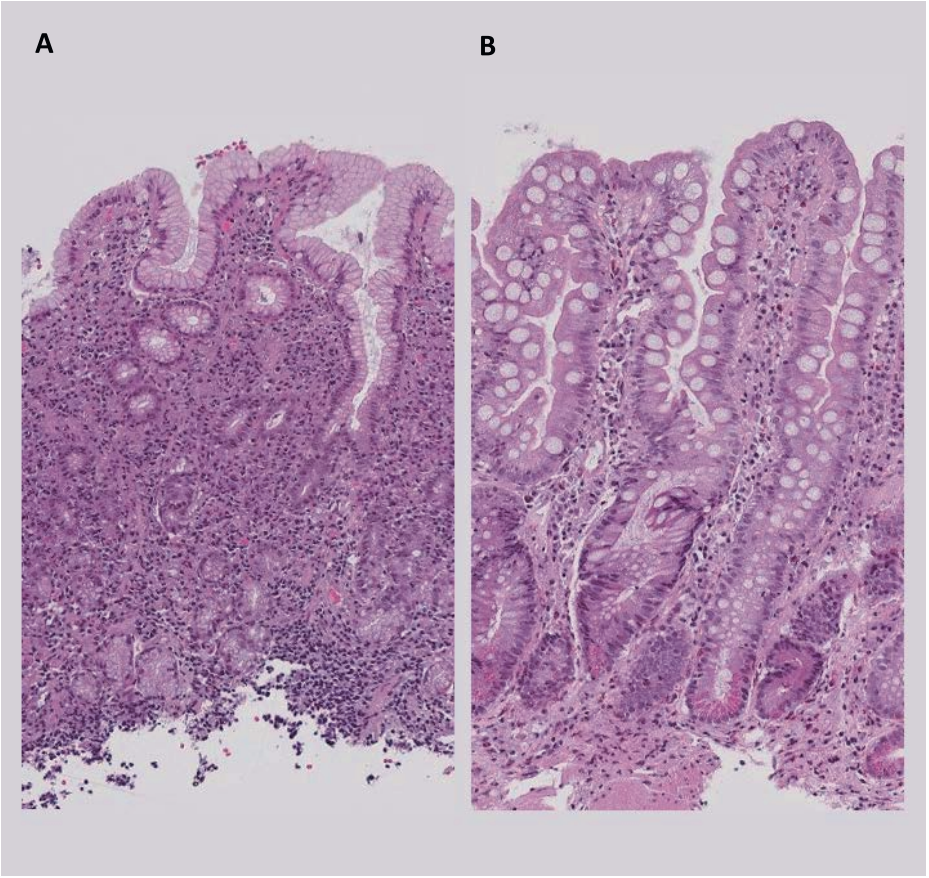


Figure 7. Examples of atrophic and metaplastic tissue stained with H&E. A) Corpus atrophic gastritis. B) Intestinal metaplasia of the corpus with characteristic mucin-filled goblet cells.

Eradication of *H. pylori* leads to a reduction in risk for gastric cancer development [90]. However, once atrophy and metaplasia has developed, the prevention of further progression to cancer by *H. pylori* eradication is uncertain, which has been termed "the point of no return", but exactly when this point occurs has not yet been defined [89].

GASTRIC CANCER

GASTRIC CANCER SUBTYPES

The vast majority (95%) of stomach cancers are gastric adenocarcinomas, meaning that they originate in the glandular epithelium of gastric tissue. These can be coarsely divided by their anatomical location into *cardia* gastric cancer, which affects the proximal part of the stomach (figure 1), and *non-cardia* gastric cancer (NCGC), which affects the more distal parts of the stomach. The most established classification system for non-cardia gastric cancer, the Lauren system, further divides these tumours into intestinal and diffuse types; these subtypes have different aetiologies and different histological manifestations [91].

The intestinal type of gastric cancers is thought to develop through a sequential group of events, which, described more in detail above, commonly are called the Correa cascade (figure 8) [85, 92], and is closely associated to the inflammation induced by chronic infection by *H. pylori.*

The diffuse type of gastric adenocarcinoma does not display glandular structures like the intestinal type, but has a poorly cohesive, unorganized appearance, containing signet-ring cells filled with intracellular mucus. The carcinogenesis of these tumours is much less well-studied than the intestinal type, and does not progress along the Correa cascade. Instead, they develop through signet ring carcinoma *in situ,* into invasive signet ring carcinomas. Loss of heterozygosity of the E-cadherin gene *CDH1* is a hallmark feature, occurring in 30-40% of hereditary cases, and as a somatic mutation in 50-70% of sporadic cases [91]. There are also mixed-type tumours with areas of both appearances. Among gastric adenocarcinomas, the intestinal type accounts for approximately 50%, the diffuse type for 33% and the mixed type for 17% [93].

In addition to adenocarcinomas, a small proportion of gastric MALT lymphomas are etiologically associated with *H. pylori* infection. MALT stands for mucosa-associated lymphoid tissue, a type of secondary lymphoid tissue that does not naturally exist in the gastric mucosa. However, in response to a chronic antigenic immunoinflammatory stimulus, lymphangiogenesis and MALT can be induced. MALT lymphomas are indolent B-cell lymphomas and most commonly develop in the digestive system (predominantly the stomach) and frequently have chromosomal rearrangements involving

immunoglobulin genes. *H. pylori* has been shown to play an important role in the carcinogenesis, and early-stage MALT lymphomas can regress upon *H. pylori* eradication [94].

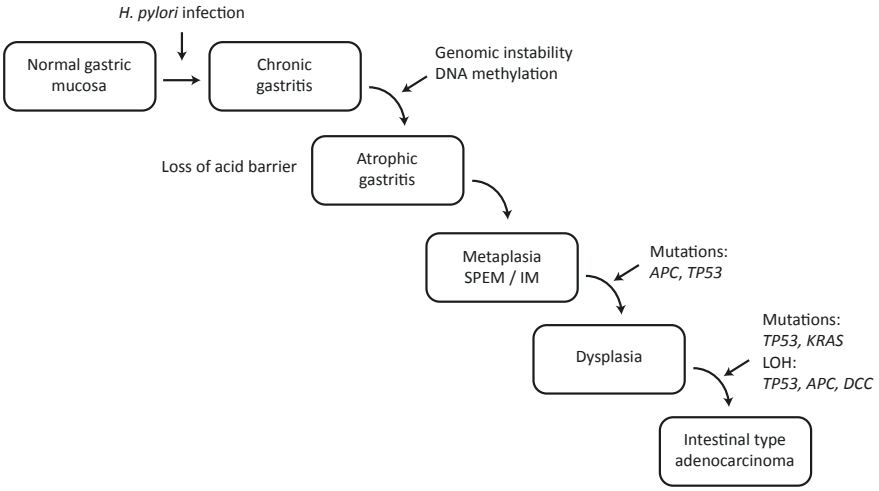


Figure 8. The Correa cascade of intestinal type gastric carcinogenesis.

Epidemiology of Gastric cancer

The most recent worldwide cancer statistics are provided by the GLOBOCAN 2012 study, which was published in 2013 by the International Agency for Research on Cancer (IARC) organ of WHO [95]. This study presents an estimate of cancer deaths in different countries and is based on national registry data from 2012 where this is available. For countries without cancer or mortality registries, it contains estimates extrapolating incidences and mortality rates from neighbouring countries. The GLOBOCAN study estimates that close to one million, 952 000 individuals were diagnosed with gastric cancer worldwide in 2012. In the same year 723 000 individuals died of gastric cancer, making it the third most common cause of cancer death in 2012, after lung cancer and liver cancer. In the 2008 report from the same study group, it was estimated that NCGC accounted for approximately 88% of all gastric cancer cases [96], which would represent 837 760 cases in 2012.

The risk of developing gastric cancer is very unevenly distributed over the world. The highest in incidence is East Asia, including Japan, Korea, and China (figure 9). Another high-risk area is South and Central America. Also in Europe there is a big variation in incidence, with substantially higher rates in Eastern Europe and Russia compared to Western and Northern Europe [15]. There is an approximately 10-fold difference in incidence rates between high- and low-risk countries, with the incidence in Korea and Japan being approximately 60 in 100 000 individuals, compared to below 6 in 100 000 individuals in US whites, Jordan and Saudi Arabia [15]. Gastric cancer rates also differ significantly according to sex, where males have an approximately twice as high risk of developing gastric cancer as females, a pattern that is similar regardless of geographic location [15].

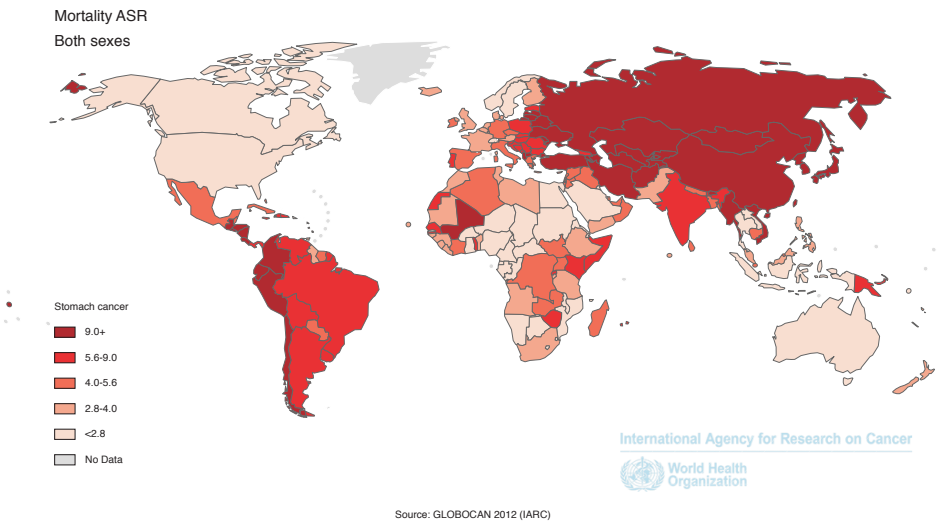


Figure 9. Map of global distribution of mortality from gastric cancer. Reprint with permission from GLOBOCAN 2012, IARC 2013 [97].

In its early stages, gastric cancers are usually asymptomatic or associated with diffuse, non-specific symptoms such as dyspepsia. Advanced stages may be accompanied by persistent abdominal pain, early satiety, weight loss and bleeding from the tumour [98]. Consequently, the mortality of gastric cancer is usually high; 80% of the tumours are not discovered until they have reached advanced stage, leading to a poor prognosis [98]. The five-year survival rate in

patients with early gastric cancer is between 85 and 100% but is only 5-20% for advanced gastric cancer [99]. These figures are relatively equal all over the world, regardless of socioeconomic level and healthcare status (figure 10), making the mortality map appear very similar to the incidence map [15].

Even as late as 1975, gastric cancer was the most common malignancy in the world [100]. However, its rates have been declining, especially in Western countries, since the middle of the last century [15]. Nevertheless, despite this decline, the number of cases is expected to grow due to population increases in countries that still have high incidence rates, and the number of new cases in 2030 is estimated to be 1 060 000 [97].

Risk factors for gastric cancer

The major risk factor for non-cardia gastric cancer (NCGC) development is chronic infection with *Helicobacter pylori* [15]. This was recognized officially in 1994 when *H. pylori* was acknowledged as a Group 1 carcinogen for gastric cancer by IARC [101], and this carcinogen status was confirmed for NCGC in 2009 [102]. Previous estimates of *H. pylori*'s contribution to the risk of developing gastric cancer have largely been based on ELISA measurements of serum anti-*H. pylori* IgG levels to determine the presence and absence of infection, and several of studies have been case-control studies rather than prospective studies. One complicating factor in this regard is that most *H. pylori* infections naturally are declining in early stages of gastric carcinogenesis, such as atrophy and metaplasia. This, together with limited sensitivity of the ELISA method, seems to have led to an underestimation of the prevalence of *H. pylori* infection in patients who later develop cancer [15]. The introduction of the immunoblot method together with an increasing number of prospective studies have clarified this association and in a recent meta-analysis of such studies the relative risk of *H. pylori* on developing gastric cancer is 17.0 (95% CI: 11.6-25.5). Using immunoblot data, the attributable factor of *H. pylori* infection on non-cardia gastric cancer development was shown to be 89.2%; the prevalence of *H. pylori* infection 94.6% in NCGC cases [103]. Not only is the infection of *H. pylori* a risk factor, several virulence factors have been shown to be important in modulating the risk of gastric cancer development. The most pronounced factor is *cagA;* a *cagA* positive infection conferring an approximately doubled risk, compared to a *cagA* negative infection [15].

Other risk factors for the development of gastric cancer are similar in that they are generally associated with socioeconomic status (SES). Gastric cancer is clearly predominant in less developed areas of the world (figure 10) and in both low-risk [104, 105] and high-risk [106] areas, the risk is highest in groups of low SES. This can be illustrated by the association of gastric cancer risk with surrogate markers of low SES, such as low educational status, low income, and low occupational activity, as well as the number of siblings and a crowded living situation. SES may also confound other risk associations such as *H. pylori*, smoking and dietary habits.

Smoking, however, is also a well-established risk factor for gastric cancer, with the level of risk depending on the number of years of smoking and number of cigarettes per day [107]. Dietary habits also influence risk but constitute a complicating story of intertwined factors with small individual contributions. Probable protective dietary factors are fruits and non-starchy vegetables, especially of the allium family. Risk factors with the same level of evidence are salt and salt-preserved food [15].

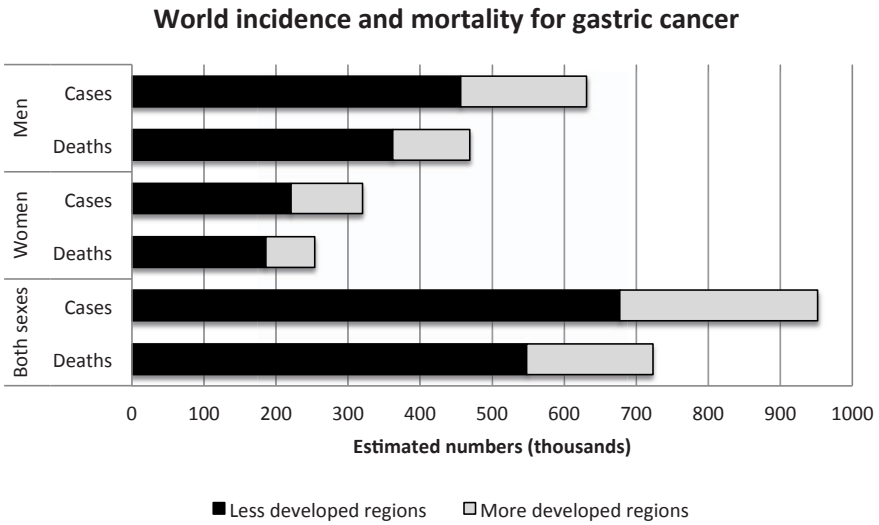


Figure 10. Estimated thousands of cases and deaths in gastric cancer worldwide in 2012. Numbers are separated based on sex and on development status of the region. Source Globocan 2012, IARC/WHO.

The pathogenesis of gastric cancer involves not only environmental factors, and *H. pylori* infection, but also genetic susceptibility. It is estimated that the inherited component contributes to <3% of gastric cancers, while a majority of the of genetic changes associated with gastric cancer are acquired. Familial gastric cancer predisposition loci include the *BRCA1/2* genes, *p53* in Li-Fraumeni syndrome*, APC* in familial adenomatous polyposis, and DNA mismatch repair genes and microsatellite instability in Lynch syndrome [108].

Among candidate genes associated with an increased risk, many are involved in cytokine response or other inflammatory processes. Polymorphisms in the locus encoding for IL1-B and for its receptor, IL1R have been shown to increase the risk, as do alterations in IL-8, IL-10, TNF-a, IL-17A and the toll-like receptor genes for TLR9 and TLR4. Genome wide association studies have also found the *MUC1* mucin gene and the *PSCA* gene to be linked to gastric cancer risk [109].

Location matters

As mentioned previously, most people infected with *H. pylori* are asymptomatic despite a chronic, active gastritis, and one major parameter affecting the clinical outcome from the infection is the extent and location of the infection. Simply put, three gastric phenotypes can be identified, each related to different clinical manifestations [46]. The first and most common is the *simple gastritis phenotype,* with a mild, mixed gastritis that does not significantly affect acid secretion. This occurs in the vast majority of cases and is associated with asymptomatic disease and the lack of serious clinical manifestations.

The second can be termed the *duodenal ulcer phenotype*. The presence and maintenance of a duodenal ulcer requires the stomach to secrete at least 12 moles/h of gastric acid, and this in turn require a normal or near normal gastric corpus, spared from inflammation [110]. This phenotype is instead associated with high antrum inflammatory scores, high gastrin production, and very high acid secretion [46]. This combination contributes to the development of peptic ulcers, especially in the duodenum and pre-pyloric region. Normally, *H. pylori* cannot infect the duodenum because it is inhibited by bile. However, low pH precipitates the bile acids, and duodenal infection can therefore occur when the acid secretion is elevated enough to chronically lower the pH in the duodenal bulb [111].

Despite the urease activity, the bacterium normally cannot survive in gastric pits of the corpus mucosa where parietal cells secrete approximately 160 mmol/L HCl (pH < 1) for transport to the stomach lumen. Initially, in a stomach that still retains normal corpus function, the *H. pylori* density is highest in the non acid-secreting antral part. Consequently, gastritis tends to be most severe in the antrum at early stages of infection. The third and most serious of the phenotypes is the *gastric cancer phenotype,* where the gradual loss of the parietal cells during corpus atrophy allows the gastritis to progress into the corpus area [110]. Physiological hallmarks of this phenotype is low acid secretion despite high gastrin levels, and low pepsinogen I, which leads to low pepsinogen I/II ratio [46]. This corpus-involving pan-gastritis is strongly associated with gastric cancer development, the incidence increasing with the extent and severity of inflammation [112]. Contrary, incidence of duodenal ulcer is associated with a functional corpus and antrum-predominant acid-inducing gastritis. This polarity of disease phenotypes has led to the conception that the presence of duodenal ulcer is mutually exclusive from gastric cancer development. This is however not true, since antrum-predominant gastritis is an initial manifestation of the infection, and can shift into pan-gastritis and atrophic gastritis along the other axis of disease. This means that gastric ulcers, and –cancer, can evolve in individuals with duodenal ulcers but not the opposite way. However, the window where the conditions are favourable for the development of duodenal ulcers might be short in high-risk areas of gastric cancer. Consistent with this theory, scars of healed duodenal ulcers have been observed in 1-7% of gastric cancer cases in populations of high gastric cancer risk [110].

H. pylori and Gastric cancer in Nicaragua

Nicaragua is a country in Central America that, alike several other countries in the region, has a high incidence of gastric cancer. In 2012 the incidence of gastric cancer in Nicaragua was estimated to 274 cases for men and 189 for women, in a population of approximately 6 millions. The mortality was estimated to 253 for men and 174 deaths for women, meaning that gastric cancer is the second leading cause of cancer death in Nicaraguan men and the fourth leading cause for Nicaraguan women [97]. However, these data are according to WHO's GLOBOCAN 2012 report, and as to registry information is

not available in Nicaragua, these numbers are calculated from other sources. In this report, the national incidence rate for the cancers was estimated by modelling, using actual recorded cancer registry data from Cuba, Costa Rica and Puerto Rico, and the estimated national mortality for 2012 (source WHO). This was in turn estimated from national mortality rates from 2001-2010 that were applied to the 2012 population [95]. Costa Rica on the other hand, being a neighbouring country with a functional mortality registry, was shown to have one of the world's highest incidence rates of gastric cancer, excluding Japan and Korea, where the detection rate is higher due to national screening initiatives. In Costa Rica the age-standardized incidence rate from 2000-2004 was 32.1 per 100 000 in males and 16.4 per 100 000 in females, which is exceeded only by Belarus and Russia [15].

In a recent multi-centre study, Porras and collaborators investigated the *H. pylori* prevalence in six Latin American countries and found an average prevalence of 79.4 % [113]. The study site in Nicaragua was Leon, the 4th largest city in Nicaragua, where 240 individuals were tested from a census of households. A prevalence of 83.3 % (78.1 - 88.4%) was found, with the higher prevalence rates in subjects over 50 years of age, and the lowest rates among people between 20-29 years old. Overall, the authors found significant correlations between *H. pylori* positivity, increasing number of siblings, and education status; individuals with more than 12 years of schooling were less likely to be infected. The same was true for occupational status, where participants employed outside of the home were less likely to be *H. pylori* infected. Infection was not, however, associated with BMI, smoking or alcohol use, the use of antibiotics the latest year, or chronic dyspeptic symptoms [113].

It is not entirely known why the risk of gastric cancer development is especially high in South and Central America. However, a recent meta-analysis evaluated a total of 45 studies on risk factors for gastric cancer in Latin America. These studies were published between 1990 and 2011 conducted in Brazil, Colombia, Mexico, Uruguay, Costa Rica, Venezuela, Chile, Peru, and Honduras [114].

Chilli pepper consumption was the only region-specific factor increasing the risk for gastric cancer reported in at least five of these studies. Moderate increases in gastric cancer risk were observed for smoking, alcohol use, high consumption of red meat or processed meat, excessive salt intake and carriage of *IL1RN*2*. Conversely, factors associated with a moderately

decreased risk were higher levels of education, fruit consumption, and total vegetable consumption. However, none of the studies that this meta-analysis was based on were prospective analyses, which makes the assessment of risk factors weaker [114]. Indigenous indian ancestry has been proposed as a risk factor for gastric cancer. However, in a recent analysis in Lima, Peru, a town with mixed mestizo and Amerindian population, this effect was shown to be primarily associated with socioeconomic and nutritional factors rather than ancestry, as determined by genotyping ancestry-informative SNPs in the genome. When ancestry was corrected for socioeconomic parameters, such as educational status, home building materials and sanitary conditions, the association between ancestry and gastric cancer was not longer significant [106].

Aims of the Thesis

The aim of this thesis was to characterize the interplay between a cancer-causing pathogen and its host at three different levels; using genomics, transcriptomics and surface proteomics.

The specific aims in this were:

- How does the human gene expression change over the early stages of gastric cancer development?

- What are the detailed genomic and *in vivo* transcriptional characteristics of *Helicobacter pylori* at the different disease stages?

- Can we identify other microbial species that may contribute to the pathology of gastric cancer development?

MATERIAL AND METHODS

METHODS OVERVIEW

This section aims to be an overview of the methods and patient material used in this thesis project. Rather than focusing on details, the purpose is to try to delineate the different choices of methods and the considerations behind these choices. Since the thesis work has focused on method development and implementation of next generation sequencing techniques, these methods will be the focus also of this section and I will not mention more commonly used molecular and cellular methods such as RT-qPCR and immunohistochemistry.

Since gastric cancer development is a multifactorial process that involves host, bacterial, and environmental factors, we chose to study the patient material with focus on both host and pathogen within the same individuals using different methods.

We aimed to take a three level approach and investigate;

i. the genomic sequence of the bacteria using whole-genome sequencing (WGS) (paper II),
ii. the RNA expression of both host, *Helicobacter pylori*, and other gastric bacteria using initially microarrays and thereafter RNA-sequencing (paper I, IV, and V)
iii. the surface proteome of *H. pylori* isolates (paper III).

The WGS analysis would let us do functional genomics comparisons and phylogenetic analysis of clinical isolates, but also provide reference sequences for the *H. pylori* part of the RNA-seq analysis. To use RNA-seq rather than microarray analysis allows for the analysis of total RNA and is not constrained to a pre-defined probe set of genes. This allows for the analysis not only of the abundances of known transcripts but also uncharacterized transcripts and novel isoforms, as well as nucleotide variation in the gene

sequences. The major advantage of using RNA-seq in the setting of gastric cancer development is, however, that we can map the reads from the sequencing not only to the human genome but also to bacterial reference sequences. Combined with the whole-genome sequencing we could in this case even map the *H. pylori* transcripts to the reference genomes of the patients' individual isolates. This way this method enables the investigation of both the host and the bacterial transcriptomes simultaneously. And apart from mapping to the human and *H. pylori* genomes we would also be able to use the data to identify other bacterial species and characterise the microbial communities in the patients. For the third level, we also made a pilot study of including proteomics data to the characterisation of *H. pylori* strains. In this part of the project, cultured *H. pylori* strains from one subject of the Nicaraguan cohort and one known reference strain were subjected to a novel method for surface proteome analysis that traps live bacteria and allows for trypsin cleavage of membrane-bound proteins selectively for subsequent analysis of the peptides by mass spectrometry.

PATIENT COHORTS

In the projects of this thesis we have used two different patient cohorts; for study I we used patients from a Swedish cohort and for study II-V (figure M1), which constitutes the main thesis project, we used a cohort of Nicaraguan dyspepsia patients. I will, in this section focus mainly on the Nicaraguan cohort and the studies based on the samples from these patients since that has been the focus of my thesis project.

SWEDISH PATIENT MATERIAL

The patient cohort used for paper I was a subset of dyspepsia patients participating in a study called *400-studien*, which included in total 120 patients, enrolled between 2007 and 2009. These patients underwent gastroscopy due to dyspepsia, malabsorption or anemia, and biopsy samples from both antrum and corpus mucosa were sampled for histopathology assessment, *H. pylori* culture, RNA analysis and protein analysis. The patients were then grouped into different analysis groups depending on the histopathology assessment of their tissues, in combination with *H. pylori*

culture and serology results. More details about that cohort can be found in paper I and in a recent publication [115].

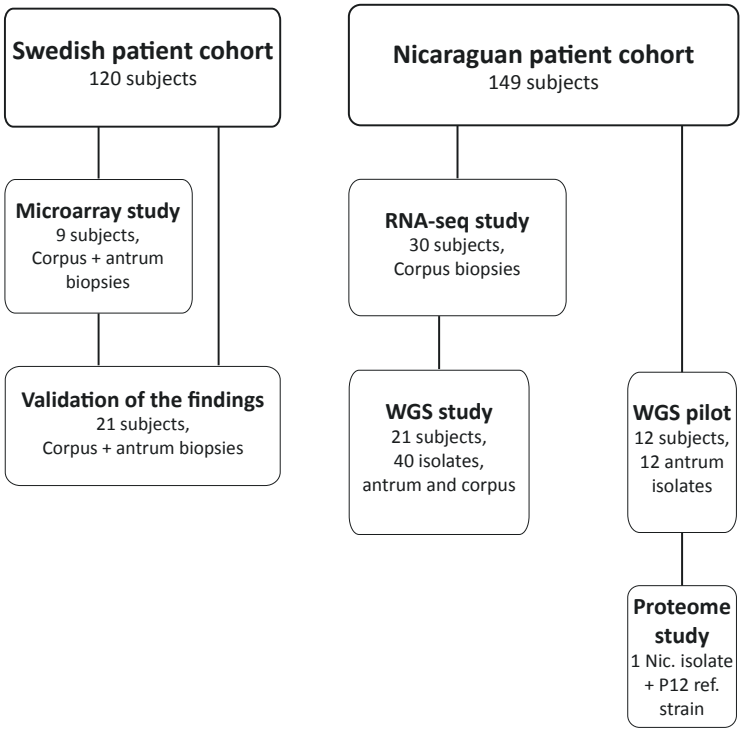


Figure M1. Overview of the patient cohorts and samples included in the different projects.

Nicaraguan patient material

The patient material used for the manuscripts II-V was collected within a large translational collaboration aclled ”Immunological biomarkers for gastric cancer”, involving researchers from Canada, Sweden, Nicaragua, Italy and the US. This project was initiated in 2008 by Dr. Lawrence Paszat, University of Toronto, to map the details of gastric cancer development in a high-risk area. The main sample collection effort was taking place from June to September 2010 at two public hospitals in Managua, Nicaragua; Hospital Escuela Antonio Lenin Fonseca (HEALF) and Hospital Escuela Dr. Roberto Calderón Gutierrez (GHERCG). The subjects were recruited among individuals that had been recommended to undergo endoscopic examination due to dyspeptic

symptoms. In total, 149 subjects were consenting to participate after being given both oral and written information about the study. All subjects were interviewed by the staff about their health, smoking habits, living conditions, and other demographic and socioeconomic factors according to a detailed questionnaire.

Prior to the endoscopy procedure blood samples were taken from which serum, plasma and leukocyte DNA was collected and frozen. During the endoscopy, the internist collected biopsies from both antrum and corpus mucosa, as well as in the vicinity of any observable potentially dysplastic lesion. In total 22 biopsies were collected for different purposes as listed in table M1.

Table M1. Tissue sampling and purpose

Purpose	Antrum	Corpus	Treatment	Storage
Histological assessment for diagnosis	2	2	Formalin	RT
Histological assessment for the study	4	4	Formalin	RT
Helicobacter pylori culturing and resistance testing	1	1	Glycerol	-70°C
RNA extraction	1	1	RNA later	-70°C
Protein extraction	2	2	Fresh frozen	-70°C
Frozen tissue for histology	1	1	Histocon	-20°C
TOTAL	**11**	**11**		

RT = Room Temperature.

The biopsies for *H. pylori* culturing were sent frozen to Houston where Prof. David Graham has been responsible for culture and storage of strains. The remaining biopsies collected for RNA analysis, frozen sectioning, or for later extraction of proteins, and all blood samples, were shipped to our group in Gothenburg for storage.

In addition to the material from the dyspepsia cohort samples we also collected from patients undergoing surgery due to gastric cancer diagnosis. These gastrectomy biopsies were collected in a similar way as described above but a substantially higher number of biopsies was collected, in total around 65 biopsies from each stomach. To be able to get a detailed mapping of the pathological changes, the anatomical origin of each biopsy were denoted on a picture of the stomach, allowing for high-resolution annotation of pathology data just adjacent to each of the frozen biopsy collected for RNA or protein analysis.

PATHOLOGICAL ASSESSMENT AND PATIENT GROUPING

Initially, the first 2 + 2 formalin fixed, paraffin embedded (FFPE) biopsies were read and graded for diagnostic purposes by pathologists in Managua, led by Dra. Reyna Victoria Palacios González. In addition to this 4 + 4 FFPE biopsies were initially graded at University of Toronto. Unfortunately, this grading was not completed and new sections were cut, haematoxylin and eosin (H&E) stained and sent to Dr Matteo Fassan and Prof. Massimo Rugge, Padova, Italy for a collective grading of all the patient sections, both the FFPE tissue and one frozen biopsy from antrum and one from corpus from each patient. The final division of the patients into the respective disease groups used in papers II, III, IV and V was based on gradings using these assessments. Biopsies were graded with regards to immune cell infiltration (lymphocytes resp. neutrophils) and tissue changes (atrophy resp. intestinal metaplasia) in H&E stained FFPE sections (table M2).

Table M2: Criteria for histopathological assessment of FFPE biopsies

Location	Lymphocytes	Neutrophils (PMN)	Atrophy	IM
A = antrum	0 = normal number in the lamina propria	0 = none	0 = none	0 = none
IA = Incisura angularis	1+ = gain in number	1+ = PMN in the lamina propria	1+= 1-33%	1 = present
C = Corpus	2+ = enlarged lamina propria	2+ = infiltrating the glandular epithelia	2+ = 34-66%	
	3+ = follicular gastritis	3+ = glandular microabscesses	3+ > 66%	

FFPE: Formalin-fixed, paraffin embedded tissue. PMN: polymorphonuclear cells. IM: Intestinal metaplasia

DISEASE STAGE GROUPING

Patient grouping is a crucial component of the study design. High throughput methods with an exploratory data-driven approach puts even higher demands on that the groups are highly stratified, to allow for the identification of biologically relevant differences between the groups. In this study we selected 30 subjects for the RNA-seq and WGS analyses based on the initial pathology scores from Toronto. The aim was to have five groups; i) *H. pylori* uninfected controls (n=6), ii) antrum-predominant gastritis (n=6), iii) pan-gastritis (n=6), iv) corpus atrophy without metaplasia (n=7), and v) corpus atrophy with metaplasia (n=5). However, when we got complementing pathology data, the group division between the antrum and pan-gastritis group was not as evident and a pair of the gastritis subjects also had atrophic changes. Therefore we decided to regroup the patients according to the extent of atrophic changes and instead analysed the data in the five groups; i) *H. pylori* uninfected controls (n=5), ii) non-atrophic gastritis (n=6), iii) low-grade corpus atrophy without metaplasia (n=9), iv) extensive

corpus atrophy without metaplasia (n=6), and v) corpus atrophy with metaplasia (n=4). The criteria for the group division are described in table M3.

As described in the background section, atrophy progresses like a lawn from the transitional zone and gradually more distally in the corpus. Intestinal metaplasia on the other hand appears in a more patchy manner. Even if we had four biopsies from each part of the stomach this does not give that the histopathology of the biopsy for RNA extraction is similar to the ones graded by the pathologist. Generally, the more extensive the tissue changes, the more likely the grading will be representative for all biopsies from an area but this is not necessarily the case. Therefore we also performed another group division, based on molecular fingerprints of atrophic changes. We used the eight genes *ATP4A*, *ATP4B*, *GHRL*, *GIF*, *CCKRB*, *PGC*, *PGA4*, and *PGA3*,

Table M3: Criteria for histopathological subgrouping

Group designation	Group description	Criteria
Hp-	*H. pylori* uninfected controls	No *H.p.* infection by any method, corpus atrophy and metaplasia score = 0
Gast	*H. pylori*-associated gastritis	*H. p.* infection by all 3 methods, corpus atrophy and metaplasia score = 0
Atr	Mild corpus atrophy but no IM	Corpus atrophy score 1, corpus metaplasia score = 0
EA	Extensive corpus atrophy but no IM	Corpus atrophy score 2-3, corpus metaplasia score = 0
Met	Extensive corpus atrophy *and* IM	Corpus atrophy score 2-3, corpus IM score 1

IM: intestinal metaplasia

which are tissue-specific genes for areas of atrophic changes, see [116] and paper I. Based on the RNA-seq counts per million reads for these genes each sample got a score on a three-level scale for each gene where score 1 corresponded to expression lower than the 95% confidence interval (CI), 2 within the 95% CI, and 3 > 95% CI. The overall molecular atrophy score was determined using the median value of the scores where a median of 1 was considered Low grade atrophy, $1.5 \leq x \leq 2.5$ as Intermediate atrophy and 3 as High grade atrophy.

MASSIVELY PARALLEL SEQUENCING:

Massively parallel sequencing (MPS), or next generation sequencing (NGS), as it is also called has opened new possibilities for sequence analysis and quantification. It is, compared to the earlier Sanger sequencing technology based on dideoxy termination, allowing for a much higher throughput in a lot less time and to a lower cost. A number of different MPS platforms exist such

as the 454 pyrosequencing, the IonTorrent semiconductor sequencing, the Pacific Biosciences single molecule real time sequencing, and the Illumina/Solexa sequencing-by-synthesis [117].

In this PhD project I have utilised the Illumina sequencing platform that combines single molecule amplification with reversible terminator-based sequencing [117]. In brief, the technology is based on shearing of the input RNA/DNA, ligation of the fragments at both ends with short specific adapter sequences, and attachment of the fragments by primers complementary to the adapters, bound to a glass surface. Since the adapters are ligated to both ends the fragments will form bridges, which are amplified from the primers, creating clusters of identical sequences on the flow cell glass surface. The actual sequencing reaction is then taking place by adding all four dNTPs that are each carrying a base-specific fluorescent dye. Cycle by cycle the single strand bridges in the clusters will be sequenced by the incorporation of one base at a time followed by the excitation of the terminal base probe with fluorescent light. Since the clustering process has multiplied each original fragment on a very limited area, all of these copies will emit the same signal each cycle, which gives enhanced signal intensity and better detection. To assure the incorporation of only one nucleotide per cycle, each nucleotide is blocked at their 3' hydroxyl group and after washing away all unbound nucleotide this group is removed in the end of the cycle, allowing for a new cycle of single incorporation, hence the term "reversible terminator-based sequencing". In this project we have used two different Illumina instruments, the HiScanSQ, for the RNA sequencing (paper IV and V) and the pilot *Helicobacter pylori* whole-genome sequencing (paper II), and the MiSeq instrument for the whole genome sequencing of the following 40 isolates (paper II).

WGS AND DE NOVO ASSEMBLY

To be able to compare the genomic build-up and virulence factor profiles of the strains colonising the individuals of the Nicaraguan patient cohort we decided to perform *H. pylori* whole genome sequencing (WGS). We started with a pilot experiment sequencing one isolate each from the antrum of 12 patients that were not selected due to symptoms but merely were among the first isolates that were sent to us by Professor Graham's group in Houston. The isolates were cultured on plates, genomic DNA was isolated, and sample preparation was performed as described in paper II. All

sequencing was performed using paired-end sequencing meaning that the fragments are sequenced from both ends. By keeping track of which two reads that originate from the same fragment, the performance of downstream alignment or assembly can be increased. The WGS pilot DNA was sequenced on the HiScanSQ instrument with 2 * 100 bp long reads. The second round of WGS were performed on DNA isolated from strains from the gastritis, atrophy and metaplasia patients from the 30 subjects for whom we also performed RNA-seq on corpus biopsies. For these genomes we used the Illumina MiSeq platform with 2 * 250 bp reads. The aim was to sequence one corpus and one antrum isolate from each individual to possibly get a glimpse of intraindividual variation in *H. pylori* populations. However, for two of the metaplasia patients and one of the atrophy patients we were not able to culture any *H. pylori* isolates, from one of the patients we could only retrieve an isolate from the antrum, and from two of the patients we sequenced two corpus and two antrum isolates respectively. For details on the patients included in the sequencing studies, see table M4.

TRANSCRIPTOMICS USING RNA-SEQ

In 2006 the first paper based on RNA-seq was published, using the 454 platform. Since then the field has been constantly expanding and RNA-seq has now taken over the role of being the first choice of methodology to study the transcriptome after almost two decades of dominance by the microarray technology. Quite soon the Illumina sequencing platform started to be preferred with its higher throughput albeit shorter reads and since its introduction in 2007 this platform has increased both its throughput and its read lengths, from the initial 25-40 bp to today's 150-300 bp. The longer reads not only increase the specificity of read alignment, they also allow for better identification of spliced transcripts, as well as increasing the possibilities for *de novo* assembly of transcriptomes in the absence of a reference genome [118].

The reason for us to use RNA sequencing, as opposed to the traditional microarray methodology as we used in paper I, was that this method would let us assay the complete gene expression in the tissue instead of pre-ordaining our targets by selecting the gene-specific probes for our array. As described above, this makes RNA-seq a promising technology to utilise for the investigation of multi-species transcriptomes within the same sample, in our case the gastric biopsy. The concept of utilising RNA-seq for "dual RNA

Table M4a: Detailed pathology scoring of the subjects included in the RNA-seq study.

HEALF* ID	RNA-seq ID	Age	Sex	Location	Lymphocyte	Neutrophils	Atrophy	Intestinal metaplasia	OLGA score†	Isolates
HEALF19432	**Hp-1**	**47**	**M**	A/IA C	0 0	0 0	0 0	0 0	0	n/a
HEALF24260	**Hp-2**	**34**	**F**	A/IA C	0 0	0 0	0 0	0 0	0	n/a
HEALF23836	**Hp-3**	**54**	**F**	A C	1 0	0 0	1 0	0 0	1	n/a
HEALF04977	**Hp-4**	**23**	**M**	A/IA C	0 0	0 0	0 0	0 0	0	n/a
HEALF05581	**Hp-5**	**51**	**F**	A C	1 0	0 0	1 0	0 0	1	n/a
HEALF25546	**Gast1**	**27**	**F**	A IA/C	1 1	2 1	0 0	0 0	0	Nic08_C Nic08_C2
HEALF06010	**Gast2**	**35**	**F**	A/IA C	3 2	3 3	1 0	0 0	1	Nic10_A Nic10_C
HEALF23466	**Gast3**	**30**	**F**	A IA/C	1 1	2 2	0 0	0 0	0	Nic13_A Nic13_C
HEALF16065	**Gast4**	**36**	**F**	A/IA IA/C	2 2	3 2	0 1	0 0	1	Nic14_A Nic14_C
HEALF12846	**Gast5**	**44**	**F**	A C	1 1	1 1	0 0	0 0	0	Nic15_A Nic15_C
HEALF00138	**Gast6**	**24**	**F**	A C	2 2	3 1	1 0	0 0	1	Nic16_A Nic16_C
HEALF19582	**Atr1**	**47**	**F**	A IA/C	3 3	2 3	2 2	0 0	3	Nic20_A Nic20_C
HEALF09107	**Atr2**	**41**	**M**	A IA/C	3 2	2 3	3 1	0 0	3	n/a
HEALF03953	**Atr3**	**42**	**F**	A IA/C	2 1	3 1	1 1	0 0	1	n/a
HEALF27875	**Atr4**	**27**	**M**	A IA/C	2 3	3 1	1 1	0 0	2	Nic07_A Nic07_C
HEALF19868	**Atr5**	**53**	**F**	A IA/C	1 1	2 1	1 1	0 0	1	Nic09_A Nic09_C
HEALF19162	**Atr6**	**58**	**M**	A C	2 2	2 2	1 1	0 0	1	Nic11_A Nic11_C
HEALF08173	**Atr7**	**32**	**F**	A IA/C	1 1	1 1	1 1	0 0	1	Nic12_A Nic12_C
HEALF01245	**Atr8**	**35**	**F**	A IA/C	3 3	**2** **2**	**1** **1**	0 0	1	Nic17_A Nic17_C
HEALF08149	**Atr9**	**27**	**F**	A IA/C	2 3	2 2	1 1	0 0	1	Nic18_A Nic18_C
HEALF24875	**EA1**	**48**	**M**	A IA/C	1 2	1 2	2 2	1 0	3	Nic01_A Nic01_C
HEALF27688	**EA2**	**66**	**F**	A/IA IA/C	1 2	1 2	1 2	0 0	2	Nic02_A
HEALF02414	**EA3**	**29**	**F**	A IA/C	2 2	2 3	2 2	0 0	3	Nic03_A Nic03_C
HEALF24293	**EA4**	**27**	**M**	A IA/C	2 2	3 1	2 3	0 0	4	Nic05_A Nic05_C
HEALF14993	**EA5**	**47**	**F**	A/IA IA/C	3 3	2 2	2 3	0 0	4	Nic06_A Nic06_A2
HEALF00568	**EA6**	**44**	**F**	A C	2 3	2 2	2 2	1 1	3	Nic33_A Nic33_C
HEALF23215	**Met1**	**47**	**M**	A C	2 2	3 2	3 3	0 1	4	Nic19_A Nic19_C
HEALF03191	**Met2**	**38**	**F**	A C	1 1	0 0	3 2	1 1	4	n/a
HEALF02475	**Met3**	**53**	**F**	A IA/C	1 2	1 1	2 2	1 1	3	Nic21_A Nic21_C
HEALF03442	**Met4**	**57**	**F**	A IA/C	1 2	1 1	1 3	0 1	3	n/a

*HEALF= Hospital Escuela Antonio Lenin Fonseca. Grading according to criteria in table M2. †OLGA score according to [24]
Each score is based on a collected assessment of 4 antrum or 4 corpus formaline fixed, paraffin embedded biopsies respectively.

Table M4b: Patient isolates included in the Whole-genome sequencing pilot

HEALF* ID	Age	Sex	Location	Lymphocyte	Neutrophils	Atrophy	Intestinal metaplasia	OLGA score†	Isolates
HEALF02021	26	F	A IA/C	3 2	2 1	1 1	0 0	3	Nic22_A
HEALF12816	55	F	A A/IA/C	3 3	3 2	2 3	1 1	3	Nic23_A
HEALF11221	56	F	A IA	3 3	1 1	0 1	0 0	1	Nic24_A
HEALF04731	58	M	A A/IA	2 3	2 3	2 2	1 0	3	Nic25_A†
HEALF14646	58	F	A C	2 2	3 1	1 0	1 0	1	Nic26_A
HEALF23077	40	F	A A/AI/C	2 2	2 2	0 1	0 0	1	Nic27_A
HEALF10585	18	F	A A/IA	3 3	2 2	1 1	1 0	3	Nic28_A
HEALF19422	60	F	A/IA/C IA/C	3 3	3 3	2 2	0 0	3	Nic29_A
HEALF03699	53	F	A A/IA	3 3	3 2	2 1	0 0	2	Nic30_A
HEALF15615	30	F	A C	3 1	1 1	1 0	0 0	1	Nic31_A
HEALF21906	24	F	A/IA IA/C	2 1	1 1	1 0	0 0	1	Nic32_A

*HEALF= Hospital Escuela Antonio Lenin Fonseca. Grading according to criteria in table M2. †OLGA score according to [24]
Each score is based on a collected assessment of 4 antrum or 4 corpus formaline fixed, paraffin embedded biopsies respectively.
† Isolate Nic25_A was also used in the surface proteomics project

sequencing" of both host and pathogen was highlighted in a very good review in 2012 where the authors state the developing feasibility and promise of this approach [119].

METHODOLOGICAL CONSIDERATIONS

WGS - BIOINFORMATICS ANALYSIS:

The output of the sequencing machine is raw image data from each cycle of the sequencing reaction, which is interpreted by the sequencing machine software. Together with the base called in each cycle, the software also gives a quality (Q) score for each base, signalling with which confidence the base call is determined. The score is given according to a pseudo-phred score corresponding to the probability that this is a correct call and ranges between 0 add 40 where Q=10 for example means that the probability for the correct call is 1/10, Q=20 corresponds to 1/100 and Q=30 to 1/1000, hence a noisy signal will give lower scores. The output files with the raw sequence reads are in FASTQ format and contain, for each sequencing read, both each base of

the sequence and the Q-score for each base. Using this quality information we trimmed the reads with a quality score cut-off of Q30 using the software TrimGalore. We also removed reads shorter than 30 base pairs to ensure that downstream alignments would be specific enough. Finally, we removed reads lacking a pair mate after the trimming to be able to use the paired information. For an overview of the workflow of the bioinformatics analysis of the WGS data, see figure M2.

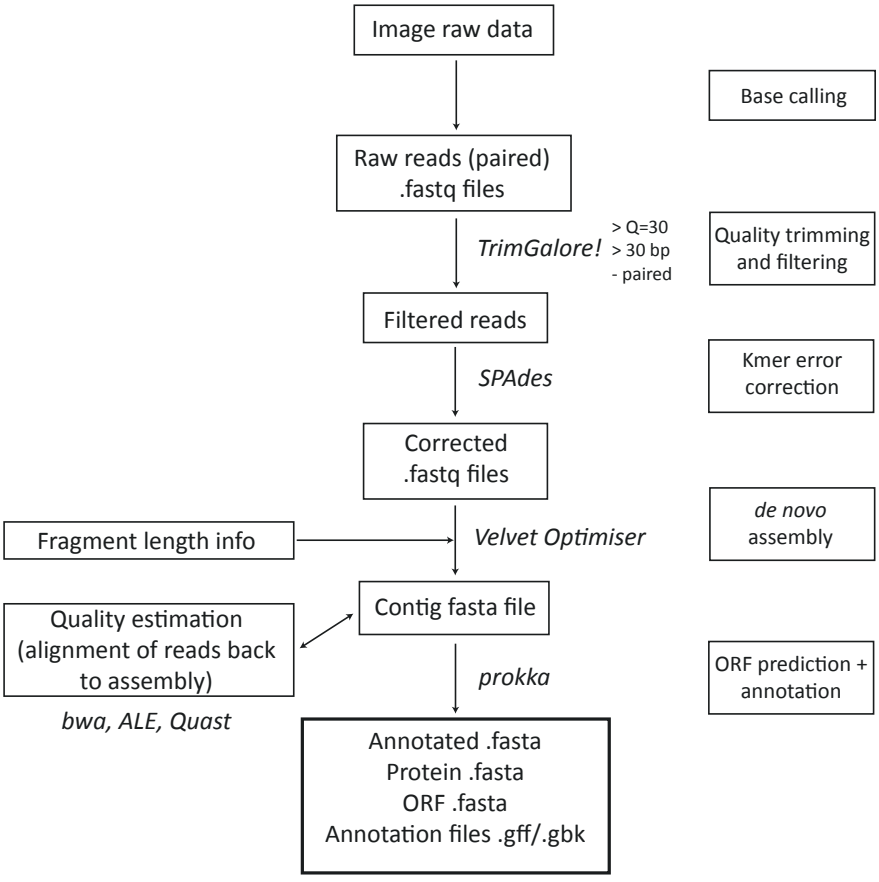


Figure M2. Overview over the de novo assembly pipeline.

Creating draft genomes:

One of the major considerations of this part of the project was to decide on what combination of bioinformatics methods to use to create draft genomes from the raw reads. Two main methods can be used for this, either alignment of the reads to a pre-existing reference genome or *de novo* assembly of the reads. Since the *H. pylori* genome is renowned for being very variable, especially due to recombination events [120], we were concerned that aligning the reads to any of the already existing genomes would introduce untrue constraints when it came to genome arrangement and synteny of the draft genomes. Therefore we decided to pursue to assemble the genomes *de novo*. *De novo* assembly means that one reconstructs the original DNA sequence from the fragment reads alone. Shortly, the algorithms first identify sequencing reads that overlap each other and by this they build graphs, pictures of the read connections (figure M3). The next step is to simplify the graph, which can be complicated due to sequencing errors. The last step is to transverse the graph, building longer paths of contiguous sequences, in a way that the path only visits each node once. The results, called "contigs" are contiguous stretches of assembled sequences that should be unambiguous. This problem is not at all conceptually or computationally trivial and there are several factors complicating the process such as sequence bias, i.e. missing pieces in the puzzle, sequencing errors, repeat regions and highly homologous regions.

Chosing error correction

The jigsaw-puzzle approach of *de novo* assembly builds on finding reads that "fit together" which leads to sensitivity to sequence bias, i.e. missing pieces in the puzzle, and to sequencing errors. Several studies have attempted to evaluate assembly approaches, the two most notable being the GAGE and GAGE-B papers [121, 122]. One of the things highlighted in the GAGE and GAGE-B paper was that the success of an assembly is highly dependent on good input data. By applying an error correction software errors in the reads that have arisen for example by PCR or sequencing errors can be identified so that those won't hamper the assembly process. We decided to use the error correction built into the SPAdes de novo assembly software and that can be used as a stand-alone tool [123] since this was suggested in the GAGE-B paper.

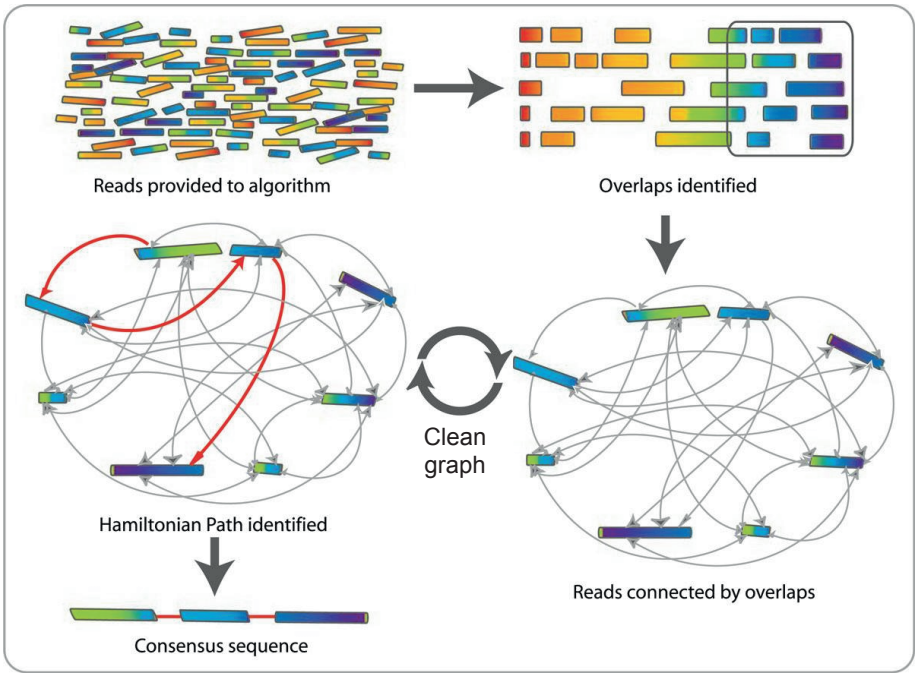


Figure M3. Schematic overview of the *de novo* assembly process.

Chosing assembler

In the effort to chose an assembler we tried several alternatives, namely Velvet, ABYSS, Ray and SPAdes. In the end the decision stood between VelvetOptimiser [124] and SPAdes [123] since both these software allow the iteration over several kmer lengths within an interval, searching for the optimal kmer length for the input data, most importantly depending on the read length and coverage [123].

To evaluate the draft genomes we mapped the input reads back onto the draft genomes using the burrows wheeler alignment (bwa) tool [125], and used the ALE software [126] to calculate an assembly score based on this alignment together with the draft genome contigs. Generally, SPAdes gave slightly higher-scoring assemblies for the MiSeq data of 250 bp reads; however, VelvetOptimiser showed a more marked advantage on the HiScan data with shorter reads (data not shown). Since the difference in score was more

pronounced for the shorter read data, we eventually chose to use VelvetOptimiser to assemble all genomes.

Annotation

Another issue was how to choose approach when it came to genome annotation, i.e. how to find the coding parts of the genome and to name these by their biologically sensible name. *H. pylori* is a species with an open pan-genome, in which each individual isolate contains a distinct set of non-core, and strain-specific genes. While the average genome content is around 1500 genes, only around 900 of these can be found in all sequenced strains when grouping genes in orthologous based on 80% similarity (data not shown). Therefore, it is not optimal to use a single genome as template for annotation, at the same time as one would like to get the annotation as harmonized as possible to facilitate comparative genomics. By using the Prokka genome annotation pipeline developed by Torsten Seemann, we were able to succeed with this. This software uses the Prodigal gene finding software to predict open reading frames (ORFs) [127]. These ORFs are first translated and compared to a primary, trustworthy reference genome using blastp and the annotation of the best significant match is used to annotate the ORF. However, if a good enough match cannot be found in the primary database, Prokka searches a match in UniProt, finished bacterial genomes in RefSeq for a specified genus, and finally Pfam and TIGRFAMs, otherwise it labels the ORF as 'hypothetical protein'. As primary annotation source in Prokka, we used the 26695 genome with the most recent re-annotation [128], and with manually curated outer membrane protein (OMP) annotation according to Alm et al. [48]. As e-value cut-off for blastp we used 10^{-9}. This approach allowed us to get a standardized gene name for most ORFs, which greatly facilitated biological interpretation of both the WGS, transcriptome and proteomics data.

RNA-seq - Sample preparation

The sample preparation for RNA-seq, as described in figure M4, starts with RNA extraction from the sample, in our case a gastric biopsy. After ensuring that the RNA has sufficient quality, i.e. is not too degraded, the RNA is sheared, preferably using an unbiased fragmentation method such as sonication. The fragments are reverse transcribed into complementary DNA

(cDNA) and adapters are ligated to the ends to enable binding of the fragments to the sequencing flow cell.

Prior to clustering on the flow cell the cDNA is amplified using PCR to create the amounts needed for clustering and the quality of the cDNA library is controlled, whereby the sample is ready for sequencing.

Starting this project in 2011, we were among the first to perform RNA sequencing at the Genomics Core Facility at Sahlgrenska Academy, where all sequencing has been developed and performed. Since we on top of that wanted to have the quite unusual approach to be able to study both prokaryotic and eukaryotic RNA we were even more pioneering in the methods development aspect.

One of the ways to optimize for informative output from RNA sequencing is to remove the, in this setting, uninteresting but very abundant ribosomal RNA (rRNA), accounting for up to 80% of the human transcriptome [119]. At the time the most common approach for this was to use poly-dT primers in the cDNA synthesis, thereby selecting for the mRNA. However, since most prokaryotic mRNA is not poly-adenylated this approach would have eliminated the bacterial RNA, and we therefore had to utilise an alternative method. To find the most suitable solution we compared three different techniques for rRNA depletion of human rRNA; mRNA-ONLY (EpiCentre Biotechnologies), MicrobeEnrich (Ambion), and RiboZero (EpiCentre Biotechnologies) by testing them on both a 10:1 mixture of Human:ETEC total RNA and total RNA extracted from gastric biopsies. After evaluating the kit performances by RT-qPCR of both human and bacterial rRNA and mRNA genes (data not shown) we decided to use the RiboZero method. This magnetic bead-based depletion utilise oligonucleotides complementary to the human rRNA molecules 18S and 28S connected to magnetic beads, and one can pull out the rRNA from the mix without losing the product of interest. The next consideration was how to perform the reverse transcription of the RNA into cDNA. To avoid the for our application detrimental bias introduced by polyT-primers as described above, we used random hexamer priming to give all fragments a theoretically equal probability of being reverse transcribed.

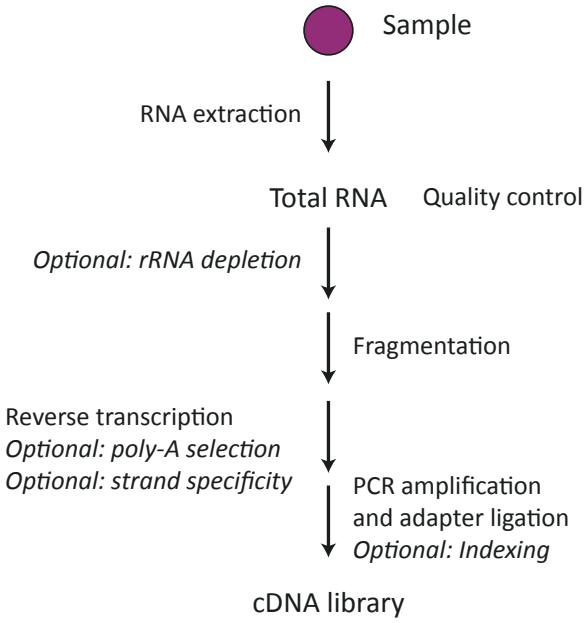


Figure M4. Overview of the sample preparation workflow.

Since prokaryotic organisms to a higher extent have overlapping open reading frames on both strands of the DNA, we wanted to be able to track which strand the RNA had been transcribed from, an information that is lost during the reverse transcription stage unless specific measures are taken. Therefore we decided to work with a strand specific sample preparation kit, the ScriptSeq v2 RNA-Seq sample preparation kit (Epicentre Biotechnologies), which for the cDNA synthesis uses random hexamer primers with a tagging sequence, thereby keeping track of the direction of the RNA.

RNA-seq – Bioinformatics analysis

Alignment to a reference genome

As described in paper V, the alignment of the RNA-seq data to the human genome was performed using RNA-STAR, an alignment software that uses uncompressed suffix arrays and a mapping algorithm similar to those used in large-scale genome alignment tools. It can both utilise splicing information

provided by the user but also executes *de novo* splice site identification [129]. The STAR aligner was also used to map the reads to the *H. pylori* isolate genomes in paper IV. For each of the 20 individuals who had at least one available *H. pylori* reference genome, derived from the same subject, we used this genome to construct individual reference databases for the alignment. A schematic overview over the RNA-seq data gene expression analysis workflow is described in figure M5.

Analysis of microbial composition using rRNA transcripts

In addition to use the data for gene expression quantification by mapping to the host and pathogen reference genomes, we wanted to investigate which other bacteria that could be identified in the biopsy tissue. To do so we used the Metaxa2 software. Metaxa was originally developed as a software tool for automated detection and classification of ribosomal small subunit (SSU) RNA in metagenome datasets. The small rRNA subunits are the 16S rRNA in prokaryotes, 18S in eukaryotes and 12S in mitochondrial genomes. Metaxa identifies and extracts reads corresponding to the SSU rRNA gene from sequencing data, and annotates them based on a database, which consists of SSU rRNA sequences from the Bacteria, Archaea and Eukaryota domains. During the work with this RNA-seq dataset the Metaxa algorithm was further developed to also support classification of the large subunit, (LSU) rRNA gene and to improve the support for short read (100 bp) and paired-end sequences such as Illumina RNA-seq data. Additionally, the classification ability of Metaxa2 is dramatically improved by building the new databases on manually curated entries from SILVA (release 111) and MitoZoa, verified with data from GreenGenes, CRW, and GenBank [130]. The result, Metaxa2 can be found at http://microbiology.se/software/metaxa2/ and was used for analysis of the stomach microbiota in paper IV.

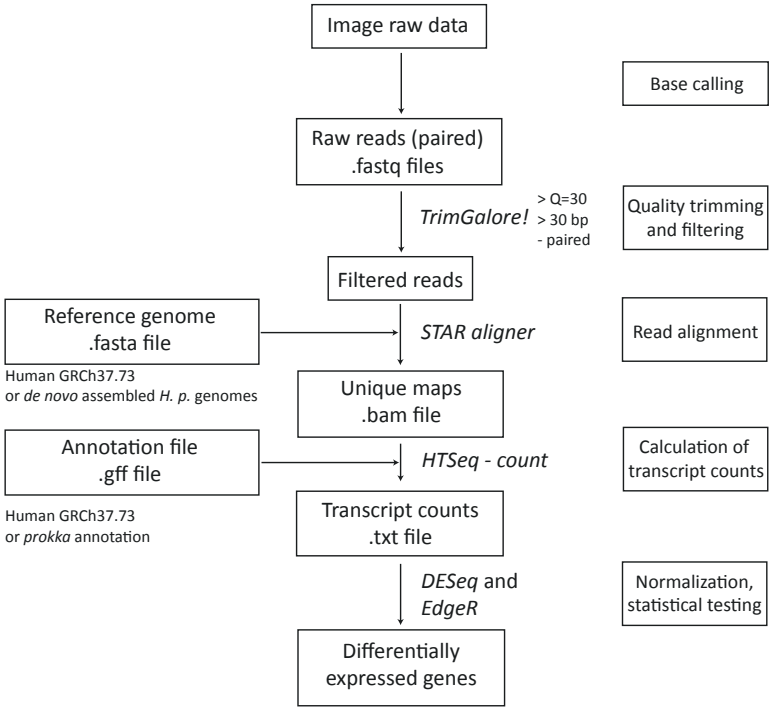


Figure M5. The data analysis workflow for RNA-seq using alignment to a reference genome. The specific software used in our analysis is noted in italics.

Results and Discussion

This thesis consists of five papers using different techniques to focus on one question; how does the interplay between *Helicobacter pylori* and the human gastric tissue gradually lead to the development of gastric cancer? And what when it doesn't, which is true for the majority of cases?

This section of the thesis will mainly focus on results not mentioned in paper I-V with the aim to complete the picture, and these results will be discussed in relation to the findings in the papers.

Nicaraguan sample collection

The initial approach was to sample *H. pylori* positive patients in Sweden (paper I), which was carried out before the start of this doctoral project. However, the collection of samples took considerable time, mainly due to low prevalence of *H. pylori* infection and premalignant changes in the Swedish population. This led to the decision to terminate the Swedish sample collection and instead study individuals from a high-risk population. The Nicaraguan sample collection started in June 2010, just prior to the thesis project. This resulted in a substantially higher number of samples collected and allowed us to proceed with higher pace.

Since gastric disease is a multifactorial disease we aimed to determine not only the infecting strain and the host response but to add other metadata into the studies. For all the patients enrolled in the study a questionnaire containing questions on demographics, living conditions and clinical information was filled in. The sample collection was taking place at the Antonio Lenin Fonseca hospital in Managua, which is a public hospital, meaning it is primarily for the part of the population that does not have health insurance. In Nicaragua, as well as in the rest of the world, gastric cancer is more prevalent in people with lower socioeconomic status (Dr. Omar Eli Morales, personal communication). Performing the sample

collection at a public hospital thereby let us sample the part of the population at highest risk.

Helicobacter pylori in Nicaragua

The prevalence of *H. pylori* infection in Nicaragua has been estimated to around 79.4 % [113]. In our patient cohort we assayed *H. pylori* infection by several different methods; culturing from biopsies, identification by microscopy, the urea breath test (UBT), and IgA and IgG serology for membrane antigens according to [48]. Culture-positivity by itself was enough for a person to be defined as *H. pylori* positive, and in cases of culture-negativity, a person with positive results in at least two of the other methods was still considered as infected. To be considered uninfected, all methods needed to show negative results. Within the Managua cohort of dyspeptic patients, the *H. pylori* prevalence was 68 % and 11% undetermined, i.e. positive in only one method. In another project, we also studied *H. pylori* prevalence in a rural location in Nicaragua, where we recruited subjects at the primary care centre of the village Nueva Guinea. Among 68 patients included in that study, the *H. pylori* infection prevalence, measured by IgA ELISA, was 88 % (unpublished data). Hence the prevalence of *H. pylori* in Nicaragua is high.

Several cancer subtypes are directly associated with chronic infections, both by bacterial agents, viruses and parasites [131]. A common theme is that that the chronic inflammation leads to tissue damage, largely attributed to the host response, which predisposes for neoplastic transformation. Infection-associated cancers are generally more common in developing regions of the world (figure R1), which can be seen for example in Nicaragua, where both gastric, genital and hepatocellular carcinomas have high prevalence rates [97].

The pathogen

Nicaraguan *Helicobacter pylori* isolates

To investigate the diversity and potential virulence of Nicaraguan *H. pylori* isolates from patients at different stages of gastric disease, we performed whole-genome sequencing of in total 52 clinical isolates (Paper II). Much of

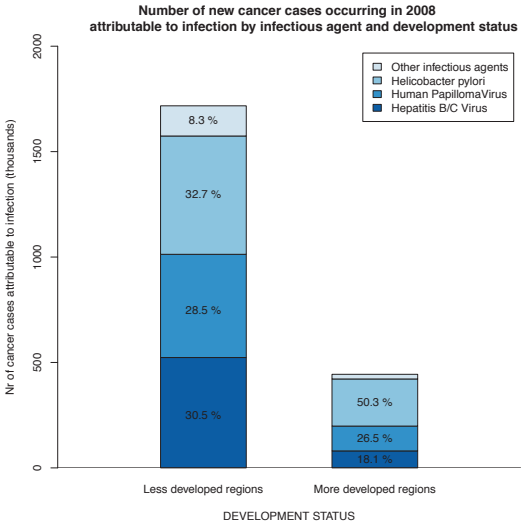


Figure R1. Fraction of infection-associated cancer cases attributable to the major infectious agents according to regional developmental status. Figure kindly provided by Martyn Plummer.

the basic data, including assembly statistics and phylogeny is collected in paper II, with focus on the major virulence factors CagA and VacA, and the adhesins BabA and SabA.

Phylogeographical origin

Phylogenetic analysis was performed on the draft genomes of the isolates using both whole genome comparison and *in silico* multi-locus sequence typing (iMLST) analysis, which is based on sequence variations in seven selected housekeeping genes. Although the iMLST analysis has lower resolution, it allows for comparisons with other datasets and more isolates can be included. The iMLST analysis placed the Nicaraguan isolates within the groups of Western and South American strains. These results fitted well with the structure of the Nicaraguan population, which is ethnically composed of mestizos (mixed Amerindian and European) 69%, Europeans 17%, people of African ancestry 9%, and indigenous Amerindians 5% [132]. However, phylogenetic analyses based on the whole genome allow a much better resolution and the global study of the 52 sequenced patient isolates instead showed a pronounced relationship to strains of West African and North

American origin (figure R2). The subjects sampled in this cohort are mainly from Managua and surrounding areas and would consequently likely be of mostly mestizo origin but unfortunately, in this study we do not know details of the ethnic groups of the subjects.

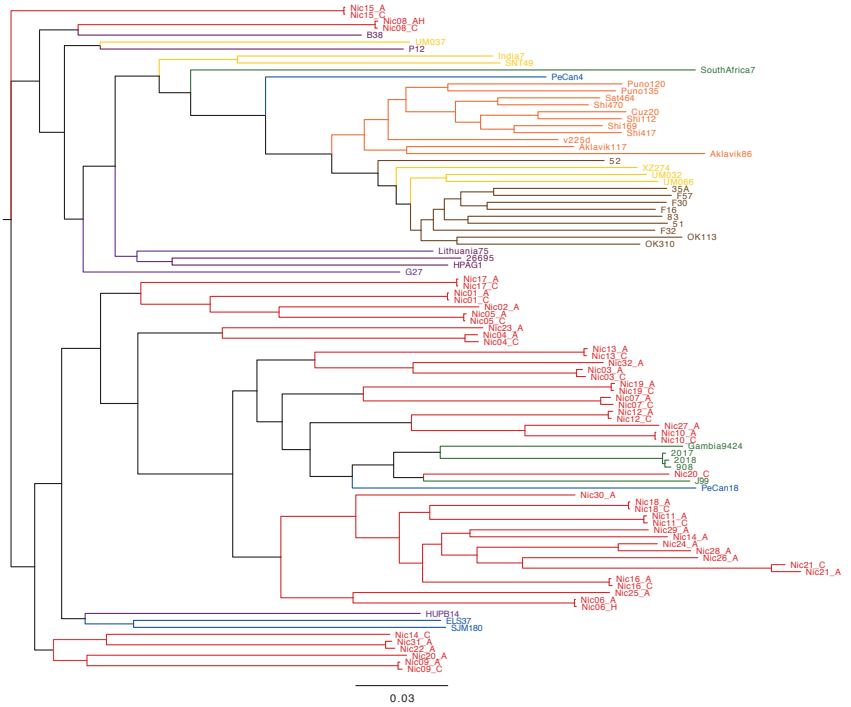


Figure R2. Whole genome SNP tree of the 52 Nicaraguan isolates together with the whole-genome sequenced strains with complete genomes as described in paper II. Nicaraguan isolates are shown in *red*, urban South American isolates are shown in *blue*, African isolates in *green* and European isolates in *purple*, East Asian isolates in *brown*, central Asian isolates in *yellow,* and Amerindian isolates in *orange.*

Other studies have shown that inhabitants of shantytowns and/or indigenous Latin American populations often carry *H. pylori* isolates with an apparent mixed ancestry of European, African, and Amerind sequences [133, 134], which can probably be explained by the high recombination rate observed in *H. pylori* affecting the entire genome including housekeeping genes [134, 135]. Our study did not contain any strains that grouped with the Amerind group but rather revealed an unexpectedly high influence of African genotypes.

Virulence factors in the Nicaraguan isolates

To identify which virulence factors that present in the isolates we searched the genomes for genes annotated as the genes mentioned in the introduction. This analysis showed that all isolates carry the genes encoding the virulence factors VacA, NapA (HP-NAP), HtrA, Ggt, AlpA, AlpB, and OipA. The carriage of the *cagA* gene could be found in 40 out of the 52 isolates (77%), *babA* in 47/52, sabA in 40/52, *hopZ* in 48/52, and *hopQ* in 43/52 of the isolates, see table R1. Presence of several virulence factors in one isolate is often linked to more severe clinical outcomes. Presence of *cagA* was tightly correlated with the more virulent *vacA* genotype (s1/i1/m1) (38/40). However, the *cagA* genotype in the isolates was of the Western type with one-fourth carrying double EPIYA-C motifs, which has been suggested to be associated with more severe outcomes. The results showed that a majority of the isolates carried several of the most common virulence genes of *H. pylori*.

Expression of virulence factors in vivo

The presence of a virulence factor at genomic level does not necessarily equal to the actual expression of the gene. Therefore we also studied the *in vivo* expression of the different genes in the RNA-seq data from patient corpus biopsies (Paper IV). Looking specifically at the virulence factors mentioned in the background we could detect expression of a majority of the virulence factors that were present in the genome of the isolate. For *cagA*, *vacA*, *napA*, *alpA*, and *alpB* expression was detected in all strains carrying the gene (table R1). For *ggt* and *htrA*, expression could not be detected in one individual out of the 20 for which we had strain specific RNA-seq data (see paper IV for method details). For the OMP genes *babA*, *sabA*, *sabB*, *oipA*, *hopZ* and *hopQ,* expression could not be detected in several patients whose isolates did carry the gene. As described in the background, several of the OMP genes are regulated by slipped strand mispairing, where some alleles render the gene transcriptionally inactive or truncated. This ON/OFF switching has been described for example for *sabA*, *oipA*, *hopZ*, and *hopQ* [20], and is a part of the adaptation of the bacteria to the mode of adherence that are favourable under different environmental conditions. This will not be detected in the way we have searched the data in this genome analysis, since

Table R1: Presence and expression of virulence factor genes

Isolate	Patient	cagA	vacA	napA	htrA	ggt	babA	babB	sabA	sabB	alpA	alpB	oipA	hopZ	hopQ
Nic01_A	EA1	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic01_C	EA1	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic02_A	EA2	-	+	+	+	+	+	-	+	-	+	+	+	+	+
Nic03_A	EA3	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic03_C	EA3	+	+	+	+	+	+	-	-	+	+	+	+	+	+
Nic04_A	na	+	+	+	+	+	+	-	+	+	+	+	+	+	+
Nic04_C	na	+	+	+	+	+	+	+	-	+	+	+	+	-	+
Nic05_A	EA4	+	+	+	+	+	-	-	+	-	+	+	+	+	+
Nic05_C	EA4	+	+	+	+	+	-	-	-	+	+	+	+	+	+
Nic06_A	EA5	+	+	+	+	+	+	-	+	+	+	+	+	+	+
Nic06_A2	EA5	+	+	+	+	+	+	-	+	+	+	+	+	+	+
Nic07_A	Atr4	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic07_C	Atr4	+	+	+	+	+	+	-	+	+	+	+	+	-	+
Nic08_C	Gast1	-	+	+	+	+	+	-	+	-	+	+	+	+	-
Nic08_C2	Gast1	-	+	+	+	+	+	-	+	-	+	+	+	+	-
Nic09_A	Atr5	-	+	+	+	+	+	-	+	-	+	+	+	+	+
Nic09_C	Atr5	-	+	+	+	+	-	-	+	-	+	+	+	+	+
Nic10_A	Gast2	+	+	+	+	+	+	+	-	+	+	+	+	+	+
Nic10_C	Gast2	+	+	+	+	+	+	+	+	+	+	+	+	-	+
Nic11_A	Atr6	+	+	+	+	+	+	+	-	+	+	+	+	-	+
Nic11_C	Atr6	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic12_A	Atr7	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic12_C	Atr7	+	+	+	+	+	+	+	-	+	+	+	+	+	+
Nic13_A	Gast3	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic13_C	Gast3	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic14_A	Gast4	+	+	+	+	+	+	-	+	+	+	+	+	+	+
Nic14_C	Gast4	-	+	+	+	+	-	+	+	-	+	+	+	+	-
Nic15_A	Gast5	-	+	+	+	+	+	+	+	-	+	+	+	+	-
Nic15_C	Gast5	-	+	+	+	+	+	+	+	-	+	+	+	+	-
Nic16_A	Gast6	+	+	+	+	+	+	+	-	+	+	+	+	+	+
Nic16_C	Gast6	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic17_A	Atr8	+	+	+	+	+	+	-	+	+	+	+	+	+	+
Nic17_C	Atr8	+	+	+	+	+	+	-	+	+	+	+	+	+	+
Nic18_A	Atr9	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic18_C	Atr9	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic19_A	Met1	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic19_C	Met1	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic20_A	Atr1	-	+	+	+	+	+	-	+	-	+	+	+	+	-
Nic20_C	Atr1	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic21_A	Met3	+	+	+	+	+	+	-	+	+	+	+	+	+	+
Nic21_C	Met3	+	+	+	+	+	+	-	+	+	+	+	+	+	+
Nic22_A	na	-	+	+	+	+	+	-	+	-	+	+	+	+	-
Nic23_A	na	+	+	+	+	+	+	-	+	-	+	+	+	+	+
Nic24_A	na	+	+	+	+	+	-	+	-	-	+	+	+	+	+
Nic25_A	na	+	+	+	+	+	+	-	+	-	+	+	+	+	+
Nic26_A	na	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Nic27_A	na	+	+	+	+	+	+	+	-	-	+	+	+	+	+
Nic28_A	na	+	+	+	+	+	+	-	+	+	+	+	+	+	+
Nic29_A	na	+	+	+	+	+	+	-	-	+	+	+	+	+	+
Nic30_A	na	+	+	+	+	+	+	+	-	-	+	+	+	+	+
Nic31_A	na	-	+	+	+	+	+	-	+	-	+	+	+	+	-
Nic32_A	na	-	+	+	+	+	+	-	-	-	+	+	+	+	-
Total number		40	52	52	52	52	47	27	40	34	52	52	52	48	43

"+" means carriage of the annotated gene, while "-" means that the isolate has no gene annotated with this name.

Green symbol means that this gene is expressed in the RNA-seq data, while *red* means expression not detected. *Black* symbol means no expression data exist for the tissue in question.

For patients EA2 and EA5 no corpus isolate was sequenced and the corpus RNA-seq data is illustrated on the antrum isolate.

an OFF allele would still count as presence of the gene. For *sabB* and *hopZ* however, the transcription levels were in general very low and the failure to detect the expression may therefore be due to insufficient depth of the sequencing.

From a biological point of view the *in vivo* RNA-seq data showed that *H. pylori* express high levels of several of the major virulence factors such as *ureA*, *cagA* and *vacA*, as well as factors involved in pH regulation. This indicates that *H. pylori* in close contact to the epithelium induce virulence expression and probably encounter acidic pH stress. The general view, as discussed in the background, has been that *H. pylori* swim from acidic pH in the stomach lumen and establish close to the epithelium where the pH is more neutral. Our transcriptome data on *H. pylori* probably represent adherent or at least mucus-associated bacteria since the sampling procedure for the biopsies make it unlikely that we have retained luminal content. Hence, the microenvironment close to the epithelium, at least in the patient group studied here, might be more acidic than previously anticipated. However, this needs to be coupled to the host expression levels at antrum and corpus sites to be able to draw any futher conclusions on the response of the respective bacterial strain.

The Blood group antigen-binding adhesin

BabA is the adhesin binding Le^b blood group antigens on the gastric epithelial cells and is one of the primary factors responsible for the adherence of *H. pylori* to host cells. One of the major findings of this thesis is the discovery of a South American-specific BabA variant (paper II). By extracting the sequence of *babA* from the genomes and performing phylogenetic analyses we found that BabA from the Nicaraguan isolates clustered together with isolates from both urban South Americans of mestizo or other mixed heritage, as well as from indigenous individuals carrying the Amerind *H. pylori* type. This specific cluster was not observed for either SabA or CagA and indicated that a specific BabA variant has spread in South America (Paper II).

RNA-seq data confirmed expression of the *babA* gene to varying degree, fifteen samples showed detectable expression while we could not record expression in three samples that carried the genes (table R1). Since we only had individual *H. pylori* reference genomes for 20 of the 30 individuals, we

complemented the RNA-seq by performing RT-qPCR to quantify the *babA* gene expression in the RNA samples from all the 30 corpus biopsies used for RNA-seq. Before the analysis the primers were tested for specificity in all the 52 genomes to avoid this aspect to confound the analysis. Interestingly, there were significantly higher babA expression levels ($p < 0.05$, two-tailed t-test) in the group with low to intermediate atrophy (Atr) compared to the gastritis group (figure R3). These findings might implicate that improved adherence plays a role in disease progression towards atrophy. However, a trend towards lower babA expression was observed in the EA and Met groups indicating that this increased expression of babA might be transient.

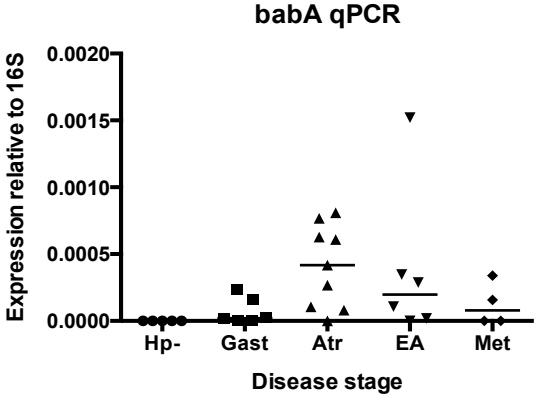


Figure R3. BabA expression measured with RT-qPCR. The lines mark the median expression value within each group.

Outer membrane protein expression (Paper III and IV)

Outer membrane proteins (OMPs) are the first to interact with host epithelial cells and immune system. They are also hot spots for genetic variation and many of them are regulated by phase variation. When the second *H. pylori* genome sequence was released, Alm and colleagues published a hallmark paper on the comparative genomics between the two strains (J99 and 26695), with the focus on genes encoding outer membrane proteins [48]. They identified in total 63 genes with orthologues in both J99 and 26695, one third belonging to the Hop OMP family. Recently, a surface proteomics study made a thorough effort to verify this list and used multiple criteria for surface-exposed outer membrane localisation. They added another 20 proteins to the list, at but could not detect several of the proteins from the other

reference, which may be a result of method and strain choice [49]. We used the union of the lists from these two studies, in total 83 genes, to filter the RNA-seq data , and could detect the expression of 79 of the OMPs. Ten of them could be found among the top 50 most highly expressed genes, most notably *alpA*, *alpB*, *hofC* and *babA*.

THE STOMACH MICROBIOTA

The RNA-seq data used in Paper IV and V was generated after depletion of human rRNA but prokaryotic rRNA was retained. This fact allowed us to extract the rRNA sequences using a software called Metaxa2 [130, 136], which is described more in detail in the methods section. The fraction of RNA-seq reads that mapped to any 16S sequence, encoding the prokaryotic small ribosomal unit, is shown in figure R4. This highlights that the prokaryotic content in the biopsies was very variable among individuals, which could be due to either an actual difference in mucosa-associated microbiota, which is likely the case for the *H. pylori* uninfected individuals, but could also vary due to sampling differences. The Metaxa2 analysis places the 16S rRNA matches at as detailed level as the specificity of the match allows. This means that, if the match is specific for e.g. *H. pylori*, it will be classified as *H. pylori*, while if it is a region that is shared between all *Helicobacteriaceae* or Proteobacteria, it will not be classified beyond this level [136].

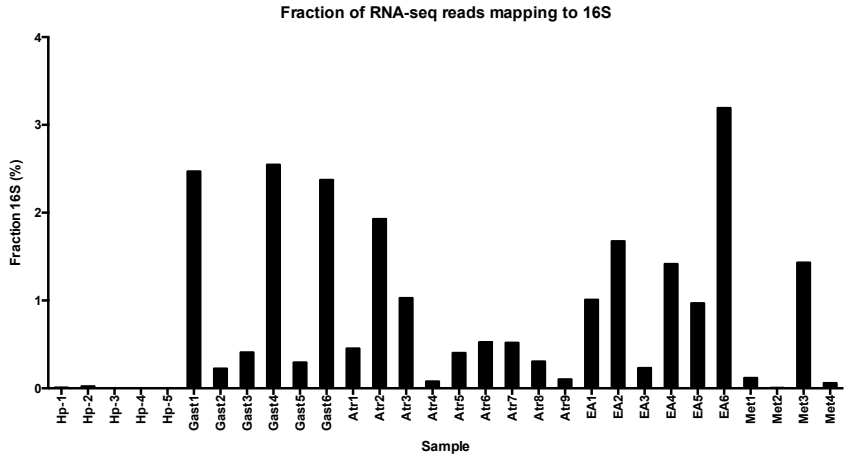


Figure R4: The percent of total reads mapping to bacterial 16S rRNA using Metaxa2.

We could identify 16 bacterial phyla (the highest taxonomic rank) and the eukaryotic kingdom fungi. The most common phylum was Proteobacteria, to which *H. pylori* belongs. In fact the genus Helicobacter was the most common genus in the majority of the samples including samples classified as H. pylori negative by the conventional methods mentioned above (paper IV). We found that presence of the genera *Campylobacter* and *Wolinella*, which together with *Helicobacter* belong to the epsilon subclass of Proteobacteria, was positively correlated to presence of *Helicobacter*. This finding may suggest that these closely related genera could collaborate in the human stomach environment. In *H. pylori* positive subjects. The genus *Helicobacter* dominates completely, however we still find several other genera such as *Escherichia*, *Pseudomonas* and *Acinetobacter*. Other genera are typically associated with the oral microbiota such as *Streptococcus*, *Treponema* and *Porphyromonas*. In conclusion, the stomach microbiota has traditionally been investigated using DNA, which might detect both live and dead prokaryotes while our approach in fact describes the live microbiota in the stomach since we used RNA.

THE HOST

HISTOPATHOLOGICAL AND MOLECULAR CHANGES

In the first paper we studied Swedish patients with different stages of *H. pylori* infection. We found a pronounced antralisation of corpus mucosa in patients with atrophy, which was characterised by loss of a large number of corpus specific genes (paper I), including genes associated with acid production and energy metabolism. A signature gene down-regulated in the atrophic corpus mucosa is *ATP4B*, which codes the beta subunit of the H^+/K^+ ATPase proton pump responsible for gastric acid secretion. A key finding of this study was the loss of chitinase (CHIA) expression in atrophy patients compared to *H. pylori* infected individuals with no premalignant changes (paper I). We could also confirm the loss of CHIA expression in the study of the Nicaraguan patients (Paper V and figure R5).

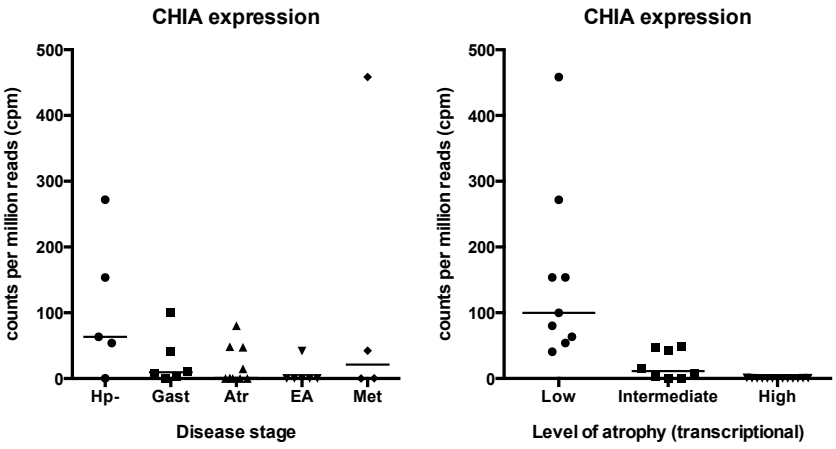


Figure R5: Expression of acidic chitinase (CHIA or AMCase). Expression levels showed both for the histopathological groups (left) and the group division based on atrophy marker genes (right). Lines mark median expression within the groups.

Transcriptome studies are very dependent on accurate patient grouping and sufficient sample size. As discussed in the materials and methods section above, we therefore used eight previously described molecular markers for atrophy (ATP4A, ATP4B, GHRL, GIF, CCKRB, PGC, PGA4, and PGA3) to assess the degree of atrophy based on gene expression data, independently of histological assessment. As illustrated by the expression of ATP4B (figure R6), expressions of the molecular markers were not always consistent with the previous grouping. There were both a number of cases where the histological assessment yielded a lower atrophy score, or a higher atrophy score than the histopathological markers (table R2). A re-analysis of the CHIA expression of the Nicaragua patient samples showed that the reduction of CHIA expression was more strongly associated to increasing molecular atrophy scores than to the original histologically-based patient grouping (figure R5).

Table R2: Comparison between histopathological and molecular atrophy scores

HEALF ID	RNA-seq ID	Location	Atrophy	Intestinal metaplasia	OLGA score	Atrophy score RNA-seq
HEALF19432	**Hp-1**	A/IA C	0 0	0 0	0	1
HEALF24260	**Hp-2**	A/IA C	0 0	0 0	0	1
HEALF23836	**Hp-3**	A C	1 0	0 0	1	**3**
HEALF04977	**Hp-4**	A/IA C	0 0	0 0	0	1
HEALF05581	**Hp-5**	A C	1 0	0 0	1	1
HEALF25546	**Gast1**	A IA/C	0 0	0 0	0	**2**
HEALF06010	**Gast2**	A/IA C	1 0	0 0	1	**2**
HEALF23466	**Gast3**	A IA/C	0 0	0 0	0	**2**
HEALF16065	**Gast4**	A/IA IA/C	0 1	0 0	1	1
HEALF12846	**Gast5**	A C	0 0	0 0	0	1
HEALF00138	**Gast6**	A C	1 0	0 0	1	**3**
HEALF19582	**Atr1**	A IA/C	2 2	0 0	3	3
HEALF09107	**Atr2**	A IA/C	3 1	0 0	3	2
HEALF03953	**Atr3**	A IA/C	1 1	0 0	1	2
HEALF27875	**Atr4**	A IA/C	1 1	0 0	2	2
HEALF19868	**Atr5**	A IA/C	1 1	0 0	1	**3**
HEALF19162	**Atr6**	A C	1 1	0 0	1	**1**
HEALF08173	**Atr7**	A IA/C	1 1	0 0	1	**3**
HEALF01245	**Atr8**	A IA/C	**1** **1**	0 0	1	**3**
HEALF08149	**Atr9**	A IA/C	1 1	0 0	1	**3**
HEALF24875	**EA1**	A IA/C	2 2	1 0	3	3
HEALF27688	**EA2**	A/IA IA/C	1 2	0 0	2	3
HEALF02414	**EA3**	A IA/C	2 2	0 0	3	3
HEALF24293	**EA4**	A IA/C	2 3	0 0	4	3
HEALF14993	**EA5**	A/IA IA/C	2 3	0 0	4	3
HEALF00568	**EA6**	A C	2 2	1 1	3	3
HEALF23215	**Met1**	A C	3 3	0 1	4	3
HEALF03191	**Met2**	A C	3 2	1 1	4	**1**
HEALF02475	**Met3**	A IA/C	2 2	1 1	3	**2**
HEALF03442	**Met4**	A IA/C	1 3	0 1	3	3

Numbers in bold mark where the molecular atrophy score is differing in relation to the histopathological score

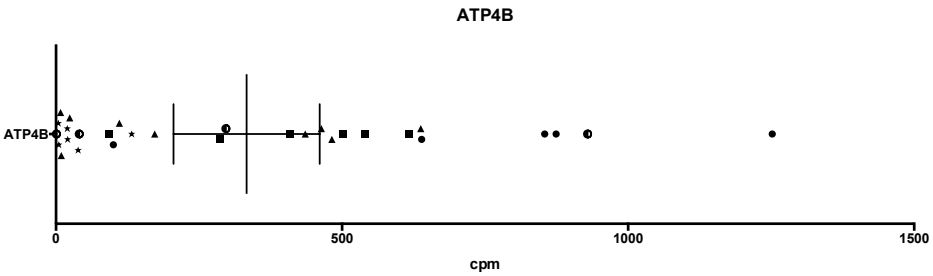


Figure R6. Distribution of ATP4B expression in the RNA-seq data, representative for the eight atrophy signature genes. The Hp- group is shown in filled circles, Gast in squares, Atr in triangles, EA in stars, and Met in split circles. Cpm: counts per million mapped reads.

To study the process of atrophy progression in greater detail, we performed gene set analysis on the RNA-seq data. This focused on metabolic pathways and reporter metabolites that participate in differentially regulated reactions at the different disease stages, and revealed that the kynurenine pathway was significantly enriched in all *H. pylori*-infected groups and most so in patients with extensive atrophy (paper V). This pathway is one of the catabolic pathways of tryptophan, and the depletion of tryptophan as well as the downstream metabolites has been shown to inhibit proliferation of T cells, directly or indirectly via activation of regulatory T (Treg) cells. The pathway has also been shown to be upregulated in several cancer settings.

Association of immune cell types to grade of atrophy

Mouse models of *Helicobacter* infection indicate that development of atrophic gastritis is dependent on presence of T cells, and in particular Th1 cells, in the gastric mucosa [137]. However, there is currently no mouse model that accurately recapitulates the pathogenesis of human *H. pylori* infection. In order to study the progression of disease in humans in relation to different immune cell types, an enrichment analysis was performed using the host RNA-sequencing data of the Nicaraguan patient cohort. The enrichment analysis was based on gene sets recently described by Bindea and colleagues, using the Piano software as described in paper V [138]. In agreement with mouse studies, there was an association with Th1-cell associated genes to gastritis development (figure R7). However, atrophic gastritis was more

strongly associated with cytotoxic cells, follicular helper T cells, central memory T cells, NK cells and activated DCs (figure R7). Thus, this analysis indicates that there is an infiltration of cytotoxic T cells and NK cells in atrophic gastritis as well as a local activation of DCs.

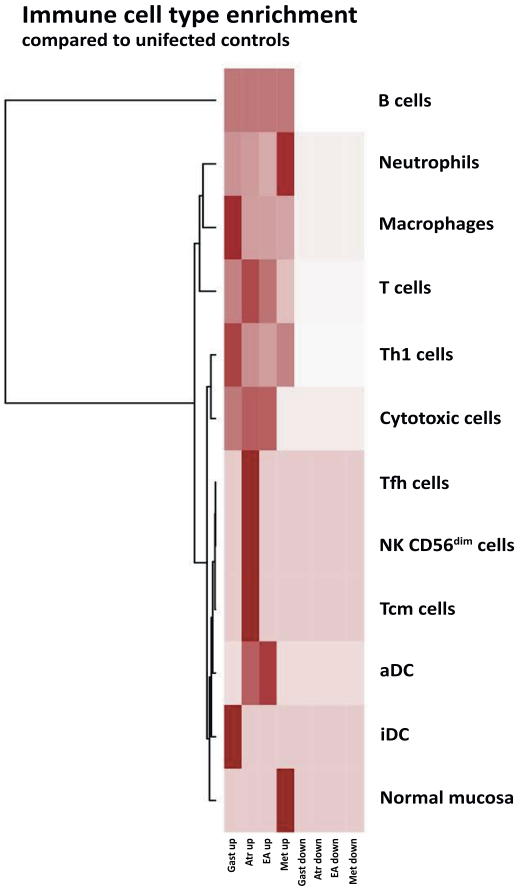


Figure R7. Immune cell types whose cell type-specific genes are significantly enriched at transcriptional level during the different stages compared to uninfected control (Min). Innate immune cells: DC, dendritic cells; iDC, immature DC; aDC, activated DC; macrophages; NK, natural killer cells including NK CD56dim and NK CD56bright cells; and neutrophils. Adaptive immune cells: B cells; Th1, T helper 1; Cytotoxic cells; Tcm, T central memory; and Tfh, T follicular helper cells.

CONCLUSIONS AND FUTURE DIRECTIONS

The wealth of data acquired using whole genome sequencing and RNA-seq is extensive and in the present thesis I have focused on selected aspects of both the pathogen and host response to infection. There are however many more features that can be further explored using the available data. One of the most intriguing is to look at the combined prokaryotic and eukaryotic transcriptomes to identify patterns of co-responses. To include not only the *H. pylori* transcriptome but also other prokaryotic gene signatures would also be illuminating and might reveal additional information about the microenvironment that lead to disease progression. However, this also poses considerable bioinformatics challenges that we have only started to work with. Ideally, one should also want to increase the depth of sequencing to be able to discern differences with higher statistical certainty, since the bacterial RNA counts are currently in a range where random effects are likely to influence comparisons. Nevertheless, an advantage of RNA-seq is that it has very low, if any, background signal because the sequences analysed have been unambiguously mapped to unique regions of the reference genome [139]. This gives the confidence that the transcripts that are detected are very likely to represent the actual presence of gene expression albeit in the cases when the counts are too low to perform differential expression analyses.

We have also only started to validate the findings of these studies in a larger context, the results in the papers II, IV, and V largely builds on observations at the genome and transcriptome levels and in a limited number of individuals. To be able to draw any more general biological conclusions these findings need to be validated in more detail. This would be particularly interesting for the changes in the expression of kynurenine pathway enzymes, where a more detailed investigation on protein and metabolite level is warranted. Also, the observation that the BabA protein is under specific selection pressure in South and Central American isolates, raises questions on how this affects binding and ligand specificity. No data regarding the three-dimensional structure of the BabA protein have yet been

published, and although the protein share partial homology with the SabA protein, the domain involved in ligand specificity is not shared, and the structural information in that region can thus not be transferred. However, such structure is soon to be released and will provide new possibilities for the interpretation of the results.

The work with the dual RNA-seq has been a major part of this study and it has bee very exciting to go from the first conception of the idea through method development, sequencing, bioinformatics method development, and finally biologically interpretable results. Paper IV is, to our knowledge, the first report of global transcriptomics in the stomach, a method that holds large promise for the investigation of the interplay between pathogen and host.

In summary; around one million individuals worldwide are diagnosed with gastric cancer every year, an effect of the chronic inflammation caused by *H. pylori,* and many more suffer from other severe consequences from this infection. The vast majority of gastric cancer cases are diagnosed at an advanced stage, leading to very poor prognosis, clearly implicating that there is a need for methods for early discovery of the disease. Antralisation and atrophy leads to a brusque change in the gastric environmental conditions, which raise the question how *H. pylori* and other bacteria adapt to this new milieu and continue to trigger the progression to gastric cancer. This thesis' characterisation of the host and pathogen at the different phases of the gastric pathogenesis has provided new information on the virulence factors carried and expressed by bacterial isolates in a high-risk population, as well as changes in host gene expression during disease progression including a potential biomarker for the early pre-cancerous changes.

POPULÄRVETENSKAPLIG SAMMANFATTNING

Magsäckscancer är en av de vanligaste anledningarna till att människor dör i cancer i världen idag och omkring en miljon människor får diagnosen varje år. Den vanligaste riskfaktorn för att utveckla magsäckscancer är en bakterie som koloniserar magsäcken hos ungefär halva jordens befolkning, nämligen *Helicobacter pylori*. Bakterieinfektionen och den inflammation den ger upphov till är centrala komponenter i cancerutvecklingen, men hos de allra flesta, ca 85 % av de som är infekterade, ger infektionen inga symptom och drygt 10 % drabbas av magsår. Totalt är det bara några procent av de som är infekterade som under sin livstid utvecklar magsäckscancer, vilket tydligt visar på att bakterieinfektionen i sig bara är en pusselbit.

H. pylori har visat sig vara väldigt variabel och kan bära på olika *virulensfaktorer*. Dessa faktorer har betydelse bland annat för hur bra bakterien kan binda till magsäckcellerna, och hur de påverkar immuncellerna och därigenom styrkan på immunsvaret, vilket har betydelse för cancerrisken. Andra riskfaktorer är olika genetiska faktorer hos den som är infekterad och dessutom miljöfaktorer som rökning samt att äta mycket kött och salt och lite frukt och grönsaker.

I den här avhandlingen har vi studerat patientprover från både Sverige, där risken för att utveckla magsäckscancer är låg, och Nicaragua, som är ett land med hög risk. Vi har jämfört olika stadier av sjukdomsutvecklingen, framför allt de tidiga stadierna i) kronisk men asymptomatisk inflammation (gastrit), ii) atrofiska förändringar i magsäcksvävnaden, och iii) metaplasi, där vävnaden förändras drastiskt. Anledningen att vi fokuserat på dessa stadier är att atrofi är det första stadiet som förknippas med högre risk att utveckla cancer och metaplasi representerar ett stadie där cancerrisken är definitiv och där risken för progression inte längre går att minska genom att behandla *H. pylori*-infektionen, något som går medan man bara har gastrit och atrofi.

Vi har studerat både *H. pylori*-bakterier isolerade från patienterna och vävnad från magsäckens övre del, tagna under gastroskopisk undersökning.

Hos bakterierna har vi tittat på deras arvsmassa, *genom,* för att identifiera vilka isolat som bär på olika virulensfaktorer. Vi har även studerat gen-uttrycket, det vill säga vilka gener som används och hur mycket av dem som det uttrycks. Detta har vi gjort i biopsier från magsäcken och har då tittat på både vilka mänskliga gener som uttrycks och vilka bakteriegener som uttrycks genom en ny metod, RNA-sekvensering. Där har vi sett att de nicaraguanska patienterna är infekterade med bakterier som både bär på och uttrycker en majoritet av de virulensfaktorer som är associerade med hög cancerrisk. De nicaraguanska bakterierna har också en speciell variant av en faktor, BabA, som används för att binda till magsäckscellerna och som visar en stor likhet inom de allra flesta sydamerikanska isolat som har studerats tidigare, oavsett vilken etnisk bakgrund deras bärare har. Det tyder på att det finns någonting i de sydamerikanska individerna eller miljön som gör att bakterierna har en fördel att behålla den varianten, vilket kan ha att samband med att magsäckscancer är väldigt vanligt i syd- och mellanamerika. Vi har också visat att en speciell mänsklig gen, den som kodar för proteinet kitinas, uttrycks i väldigt mycket lägre nivåer när man får atrofi jämfört med när man bara har gastrit, ett mönster som kan ses i både de svenska och de nicaraguanska patienterna vi studerat. Detta protein skulle kunna användas som en markör för att identifiera de som har atrofi och som därmed löper högre risk för att utveckla magsäckscancer.

ACKNOWLEDGEMENTS

I have the great privilege to be surrounded by a whole bunch of very fine people and some of you have been particularly involved in my doctoral period and the making of this thesis, either as contributors and collaborators, or in supporting, cheering, sharing coffee breaks and other necessary aspects of life.

I have also been blessed (?) with an unusual richness of supervisors. *Samuel* - thanks for good discussions and for taking me on as a doctoral student and giving me a lot of freedom. *Intawat* for guiding my first steps in bioinformatics, your support has been very valuable, as so has your laughter and delicious Thai food, thanks a lot! *Jens*; for sharing much needed words of wisdom on the way and, of course, for letting me being a part of your group during these four years. *Åsa*; you have been my rock in many times of despair and are a great friend. I will also in my future career bring with me the two post-it notes "FEL FOKUS!!!" and "VÄRT?".

This thesis is a part of an exciting international collaboration of which I've been very lucky to be a part of. Thank you *Lawrence*, for initiating the project and pursuing the work despite the setbacks, and for sharing reflections and knowledge on the cancer epidemiology field. *David Graham,* for patiently answering my basic questions on the *H. pylori* pathogenesis and sending me literature about South and Central American history. *Jason Mills* for support, good discussions and hospitality and inspiration at WashU. *Matteo Fassan,* for good-humoured replies when we asked you to re-grade all the samples with short notice. ¡Gracias a todos los patólogos e internistas nicaragüenses por su inestimable contribución a este proyecto! And last but not least; *Reyna*, mi madre nicaragüense. Gracias por la hospitalidad, la amistad y el trabajo incesante para esta colaboración.

Big thanks to the staff at Genomics Core Facility for having patience with me despite the fact that I time after another insisted in running methods that nobody had tried before. The same goes for the guys at Proteomics Core Facility, especially *Carina* and *Diarmuid*, thanks for re-running the searches countless times when I came up with new ways of optimising the databases and annotation.

I also want to thank my both groups; all great colleagues at Microbiology and Immunology, and the Sysbio group at Chalmers! *Rahil* - nu är det snart din tur, min fina kontorsgranne som delat våndor och lakrits! *Lotta, Lollo*, *Astrid, Veronica mfl*, tack för bra diskussioner om jobb och liv och många skratt! At the Chalmers side;

Bouke – for sharing so many good times and some worse times, *Amir* – for great discussions both in scientific and more philosophical questions. *Gatto* - for our debates about the holy grail of cancer metabolism and a lot of fun. *Nina*, *Christoph*, *Tobbe*, *Martin, Leif, Petri, Ana*, *Alex*, *Ed*, *Sakda*, *Shaq* and others; big thanks for a lot of fun during these years! *Rahul*; I have really appreciated our discussions of life and science; you have a lot of wisdom! Climbers, Lindy hop dancers, jazz band- and volleyball team mates, your contribution has been very important by reminding me that research is not everything.

My fellow GoBiGgers, thank you for stimulating scientific discussions and many good after works! Also thanks to GiMIICum and BioCare, the GI and cancer research schools that I've been a part of, and all friends there.

Johan - vän, bollplank, samarbetspartner, tack för all feedback och värdefulla och roliga diskussioner. *Anna*; du är grym på så många sätt, fortsätt med det! *Fredrik* - du fick in mig i lakritsträsket men jag gillar dig ändå :) Tack för gästfriheten på Matte under den sista, intensiva skrivperioden.

Jenny, min mentor - du har varit ett värdefullt stöd under doktorandtiden och givit mig en massa bra saker att tänka på inför framtiden. Tack!

Jesper - för gott kompanjonskap! Du var med från första början i funderingarna om det där med doktorerande skulle vara nåt att ha och har varit ett stöd hela vägen in i kaklet, tack!

Andreas - det började med programmering, fortsatte med dans och slutligen komboskap. Men bäst av allt är du som vän. *Yvonne* – detsamma gäller för dig. Även om du även varit kursare, grymt festsällskap och kollega så är du främst en underbar vän och det senaste året hade inte varit detsamma utan dig.

Som grädden på moset, löken på laxen och körsbäret i coctailen – vad hade jag gjort utan min fantastiska familj? Calle, Ellen och Anton; grymmaste syskonen i världshistorien, ni är barr! Jag är så innerligt glad för er och det vi har tillsammans även om (eller just för att) ni inte är riktigt kloka :) Mamma – för din ständiga omtanke och att du sedan jag var liten visat att allt är möjligt och låtit oss prova på allt mellan himmel och jord, alltid uppmuntrande och stöttande trots att vi i vår upptäckslusta och kreativitet ställde till med en hel del kaos. Pappa för korsord och ordstäv, nattliga stjärnhimmelsbeskådningar och att vi alltid fått vara med. Ni har låtit oss prova på och att gå vår egen väg på ett föredömligt vis. Mormor; den här avhandlingen är till dig, din smittsamma nyfikenhet har genomsyrat mitt liv och är en väldigt viktig bidragande faktor till att den här avhandlingen finns till.

REFERENCES:

1. Hoffmann W. Self-renewal of the gastric epithelium from stem and progenitor cells. Front Biosci (Schol Ed) 2013;5:720-31.
2. Hoffmann W. Regeneration of the gastric mucosa and its glands from stem cells. Curr Med Chem 2008;15:3133-44.
3. Singh SR. Gastric cancer stem cells: a novel therapeutic target. Cancer Lett 2013;338:110-9.
4. Mills JC, Shivdasani RA. Gastric epithelial stem cells. Gastroenterology 2011;140:412-24.
5. Nozaki K, Ogawa M, Williams JA, et al. A molecular signature of gastric metaplasia arising in response to acute parietal cell loss. Gastroenterology 2008;134:511-22.
6. Dunne C, Dolan B, Clyne M. Factors that mediate colonization of the human stomach by Helicobacter pylori. World J Gastroenterol 2014;20:5610-24.
7. Celli JP, Turner BS, Afdhal NH, et al. Helicobacter pylori moves through mucus by reducing mucin viscoelasticity. Proc Natl Acad Sci U S A 2009;106:14321-6.
8. Herrera V, Parsonnet J. Helicobacter pylori and gastric adenocarcinoma. Clin Microbiol Infect 2009;15:971-6.
9. Eusebi LH, Zagari RM, Bazzoli F. Epidemiology of Helicobacter pylori Infection. Helicobacter 2014;19 Suppl 1:1-5.
10. Magalhaes Queiroz DM, Luzza F. Epidemiology of Helicobacter pylori infection. Helicobacter 2006;11 Suppl 1:1-5.
11. Janzon A, Bhuiyan T, Lundgren A, et al. Presence of high numbers of transcriptionally active Helicobacter pylori in vomitus from Bangladeshi patients suffering from acute gastroenteritis. Helicobacter 2009;14:237-47.
12. Didelot X, Nell S, Yang I, et al. Genomic evolution and transmission of Helicobacter pylori in two South African families. Proc Natl Acad Sci U S A 2013;110:13880-5.
13. Garcia A, Salas-Jara MJ, Herrera C, et al. Biofilm and Helicobacter pylori: from environment to human host. World J Gastroenterol 2014;20:5632-8.
14. Janzon A, Sjoling A, Lothigius A, et al. Failure to detect Helicobacter pylori DNA in drinking and environmental water in Dhaka, Bangladesh, using highly sensitive real-time PCR assays. Appl Environ Microbiol 2009;75:3039-44.
15. de Martel C, Forman D, Plummer M. Gastric cancer: epidemiology and risk factors. Gastroenterol Clin North Am 2013;42:219-40.
16. Perez-Losada M, Browne EB, Madsen A, et al. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol 2006;6:97-112.
17. Suerbaum S, Josenhans C. Helicobacter pylori evolution and phenotypic diversification in a changing host. Nat Rev Microbiol 2007;5:441-52.
18. Suerbaum S, Smith JM, Bapumia K, et al. Free recombination within Helicobacter pylori. Proc Natl Acad Sci U S A 1998;95:12619-24.
19. Kennemann L, Didelot X, Aebischer T, et al. Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci U S A 2011;108:5033-8.

20. Salaun L, Linz B, Suerbaum S, et al. The diversity within an expanded and redefined repertoire of phase-variable genes in Helicobacter pylori. Microbiology 2004;150:817-30.
21. Krebes J, Didelot X, Kennemann L, et al. Bidirectional genomic exchange between Helicobacter pylori strains from a family in Coventry, United Kingdom. Int J Med Microbiol 2014.
22. Yahara K, Kawai M, Furuta Y, et al. Genome-wide survey of mutual homologous recombination in a highly sexual bacterial species. Genome Biol Evol 2012;4:628-40.
23. Linz B, Balloux F, Moodley Y, et al. An African origin for the intimate association between humans and Helicobacter pylori. Nature 2007;445:915-8.
24. Reyes-Centeno H, Ghirotto S, Detroit F, et al. Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. Proc Natl Acad Sci U S A 2014;111:7248-53.
25. Oppenheimer S. Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. Philos Trans R Soc Lond B Biol Sci 2012;367:770-84.
26. Kersulyte D, Kalia A, Gilman RH, et al. Helicobacter pylori from Peruvian amerindians: traces of human migrations in strains from remote Amazon, and genome sequence of an Amerind strain. PLoS One 2010;5:e15076.
27. de Sablet T, Piazuelo MB, Shaffer CL, et al. Phylogeographic origin of Helicobacter pylori is a determinant of gastric cancer risk. Gut 2011.
28. Finlay BB, Falkow S. Common themes in microbial pathogenicity. Microbiol Rev 1989;53:210-30.
29. Olbermann P, Josenhans C, Moodley Y, et al. A global overview of the genetic and functional diversity in the Helicobacter pylori cag pathogenicity island. PLoS Genet 2010;6:e1001069.
30. Correa P, Piazuelo MB. Evolutionary History of the Helicobacter pylori Genome: Implications for Gastric Carcinogenesis. Gut Liver 2012;6:21-8.
31. Basso D, Zambon CF, Letley DP, et al. Clinical relevance of Helicobacter pylori cagA and vacA gene polymorphisms. Gastroenterology 2008;135:91-9.
32. Polk DB, Peek RM, Jr. Helicobacter pylori: gastric cancer and beyond. Nat Rev Cancer 2010;10:403-14.
33. Wroblewski LE, Peek RM, Jr., Wilson KT. Helicobacter pylori and gastric cancer: factors that modulate disease risk. Clin Microbiol Rev 2010;23:713-39.
34. Cover TL, Tummuru MK, Cao P, et al. Divergence of genetic sequences for the vacuolating cytotoxin among Helicobacter pylori strains. J Biol Chem 1994;269:10566-73.
35. Winter JA, Letley DP, Cook KW, et al. A Role for the Vacuolating Cytotoxin, VacA, in Colonization and Helicobacter pylori-Induced Metaplasia in the Stomach. J Infect Dis 2014.
36. Atherton JC, Cao P, Peek RM, Jr., et al. Mosaicism in vacuolating cytotoxin alleles of Helicobacter pylori. Association of specific vacA types with cytotoxin production and peptic ulceration. J Biol Chem 1995;270:17771-7.

37. Rhead JL, Letley DP, Mohammadi M, et al. A new Helicobacter pylori vacuolating cytotoxin determinant, the intermediate region, is associated with gastric cancer. Gastroenterology 2007;133:926-36.
38. Atherton JC. The pathogenesis of Helicobacter pylori-induced gastro-duodenal diseases. Annu Rev Pathol 2006;1:63-96.
39. Posselt G, Backert S, Wessler S. The functional interplay of Helicobacter pylori factors with gastric epithelial cells induces a multi-step process in pathogenesis. Cell Commun Signal 2013;11:77.
40. Rassow J, Meinecke M. Helicobacter pylori VacA: a new perspective on an invasive chloride channel. Microbes Infect 2012;14:1026-33.
41. Peek RM, Jr., Fiske C, Wilson KT. Role of innate immunity in Helicobacter pylori-induced gastric malignancy. Physiol Rev 2010;90:831-58.
42. Lamb A, Chen LF. Role of the Helicobacter pylori-induced inflammatory response in the development of gastric cancer. J Cell Biochem 2013;114:491-7.
43. Backert S, Clyne M. Pathogenesis of Helicobacter pylori infection. Helicobacter 2011;16 Suppl 1:19-25.
44. Hoy B, Lower M, Weydig C, et al. Helicobacter pylori HtrA is a new secreted virulence factor that cleaves E-cadherin to disrupt intercellular adhesion. EMBO Rep 2010;11:798-804.
45. Ricci V, Giannouli M, Romano M, et al. Helicobacter pylori gamma-glutamyl transpeptidase and its pathogenic role. World J Gastroenterol 2014;20:630-8.
46. Amieva MR, El-Omar EM. Host-bacterial interactions in Helicobacter pylori infection. Gastroenterology 2008;134:306-23.
47. Dubois A, Boren T. Helicobacter pylori is invasive and it may be a facultative intracellular organism. Cell Microbiol 2007;9:1108-16.
48. Alm RA, Bina J, Andrews BM, et al. Comparative genomics of Helicobacter pylori: analysis of the outer membrane protein families. Infect Immun 2000;68:4155-68.
49. Voss BJ, Gaddy JA, McDonald WH, et al. Analysis of surface-exposed outer membrane proteins in Helicobacter pylori. J Bacteriol 2014;196:2455-71.
50. Harvey VC, Acio CR, Bredehoft AK, et al. Repetitive Sequence Variations in the Promoter Region of the Adhesin-Encoding Gene sabA of Helicobacter pylori Affect Transcription. J Bacteriol 2014;196:3421-9.
51. Boren T, Falk P, Roth KA, et al. Attachment of Helicobacter pylori to human gastric epithelium mediated by blood group antigens. Science 1993;262:1892-5.
52. Aspholm-Hurtig M, Dailide G, Lahmann M, et al. Functional adaptation of BabA, the H. pylori ABO blood group antigen binding adhesin. Science 2004;305:519-22.
53. Hennig EE, Mernaugh R, Edl J, et al. Heterogeneity among Helicobacter pylori strains in expression of the outer membrane protein BabA. Infect Immun 2004;72:3429-35.
54. Mahdavi J, Sonden B, Hurtig M, et al. Helicobacter pylori SabA adhesin in persistent infection and chronic inflammation. Science 2002;297:573-8.
55. Aspholm M, Kalia A, Ruhl S, et al. Helicobacter pylori adhesion to carbohydrates. Methods Enzymol 2006;417:293-339.

56. Unemo M, Aspholm-Hurtig M, Ilver D, et al. The sialic acid binding SabA adhesin of Helicobacter pylori is essential for nonopsonic activation of human neutrophils. J Biol Chem 2005;280:15390-7.
57. Goodwin AC, Weinberger DM, Ford CB, et al. Expression of the Helicobacter pylori adhesin SabA is controlled via phase variation and the ArsRS signal transduction system. Microbiology 2008;154:2231-40.
58. Aberg A, Gideonsson P, Vallstrom A, et al. A repetitive DNA element regulates expression of the Helicobacter pylori sialic acid binding adhesin by a rheostat-like mechanism. PLoS Pathog 2014;10:e1004234.
59. Talarico S, Whitefield SE, Fero J, et al. Regulation of Helicobacter pylori adherence by gene conversion. Mol Microbiol 2012;84:1050-61.
60. Pang SS, Nguyen ST, Perry AJ, et al. The three-dimensional structure of the extracellular adhesion domain of the sialic acid-binding adhesin SabA from Helicobacter pylori. J Biol Chem 2013.
61. Yamaoka Y, Kwon DH, Graham DY. A M(r) 34,000 proinflammatory outer membrane protein (oipA) of Helicobacter pylori. Proc Natl Acad Sci U S A 2000;97:7533-8.
62. Yamaoka Y, Ojo O, Fujimoto S, et al. Helicobacter pylori outer membrane proteins and gastroduodenal disease. Gut 2006;55:775-81.
63. Kennemann L, Brenneke B, Andres S, et al. In vivo sequence variation in HopZ, a phase-variable outer membrane protein of Helicobacter pylori. Infect Immun 2012;80:4364-73.
64. Belogolova E, Bauer B, Pompaiah M, et al. Helicobacter pylori outer membrane protein HopQ identified as a novel T4SS-associated virulence factor. Cell Microbiol 2013;15:1896-912.
65. Konturek SJ, Konturek PC, Brzozowski T, et al. From nerves and hormones to bacteria in the stomach; Nobel prize for achievements in gastrology during last century. J Physiol Pharmacol 2005;56:507-30.
66. Marshall BJ, Warren JR. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet 1984;1:1311-5.
67. Sharma BK, Santana IA, Wood EC, et al. Intragastric bacterial activity and nitrosation before, during, and after treatment with omeprazole. Br Med J (Clin Res Ed) 1984;289:717-9.
68. Abreu MT, Peek RM, Jr. Gastrointestinal malignancy and the microbiome. Gastroenterology 2014;146:1534-1546 e3.
69. Yang I, Nell S, Suerbaum S. Survival in hostile territory: the microbiota of the stomach. FEMS Microbiol Rev 2013;37:736-61.
70. Sheh A, Fox JG. The role of the gastrointestinal microbiome in Helicobacter pylori pathogenesis. Gut Microbes 2013;4:505-31.
71. Bik EM, Eckburg PB, Gill SR, et al. Molecular analysis of the bacterial microbiota in the human stomach. Proc Natl Acad Sci U S A 2006;103:732-7.
72. Andersson AF, Lindberg M, Jakobsson H, et al. Comparative analysis of human gut microbiota by barcoded pyrosequencing. PLoS One 2008;3:e2836.

73. Engstrand L, Lindberg M. Helicobacter pylori and the gastric microbiota. Best Pract Res Clin Gastroenterol 2013;27:39-45.
74. Aviles-Jimenez F, Vazquez-Jimenez F, Medrano-Guzman R, et al. Stomach microbiota composition varies between patients with non-atrophic gastritis and patients with intestinal type of gastric cancer. Sci Rep 2014;4:4202.
75. Eun CS, Kim BK, Han DS, et al. Differences in Gastric Mucosal Microbiota Profiling in Patients with Chronic Gastritis, Intestinal Metaplasia, and Gastric Cancer Using Pyrosequencing Methods. Helicobacter 2014.
76. Raghavan S, Quiding-Jarbrink M. Immune modulation by regulatory T cells in Helicobacter pylori-associated diseases. Endocr Metab Immune Disord Drug Targets 2012;12:71-85.
77. Aebischer T, Meyer TF, Andersen LP. Inflammation, immunity, and vaccines for Helicobacter. Helicobacter 2010;15 Suppl 1:21-8.
78. Otani K, Tanigawa T, Watanabe T, et al. Toll-like receptor 9 signaling has anti-inflammatory effects on the early phase of Helicobacter pylori-induced gastritis. Biochem Biophys Res Commun 2012;426:342-9.
79. Wunder C, Churin Y, Winau F, et al. Cholesterol glucosylation promotes immune evasion by Helicobacter pylori. Nat Med 2006;12:1030-8.
80. Oertli M, Muller A. Helicobacter pylori targets dendritic cells to induce immune tolerance, promote persistence and confer protection against allergic asthma. Gut Microbes 2012;3:566-71.
81. Kronsteiner B, Bassaganya-Riera J, Philipson N, et al. Novel insights on the role of CD8+ T cells and cytotoxic responses during Helicobacter pylori infection. Gut Microbes 2014;5:357-62.
82. Flach CF, Ostberg AK, Nilsson AT, et al. Proinflammatory cytokine gene expression in the stomach correlates with vaccine-induced protection against Helicobacter pylori infection in mice: an important role for interleukin-17 during the effector phase. Infect Immun 2011;79:879-86.
83. Kindlund B, Sjoling A, Hansson M, et al. FOXP3-expressing CD4(+) T-cell numbers increase in areas of duodenal gastric metaplasia and are associated to CD4(+) T-cell aggregates in the duodenum of Helicobacter pylori-infected duodenal ulcer patients. Helicobacter 2009;14:192-201.
84. Wei L, Wang J, Liu Y. Prior to Foxp3(+) regulatory T-cell induction, interleukin-10-producing B cells expand after Helicobacter pylori infection. Pathog Dis 2014;72:45-54.
85. Correa P, Piazuelo MB. The gastric precancerous cascade. J Dig Dis 2012;13:2-9.
86. Bredemeyer AJ, Geahlen JH, Weis VG, et al. The gastric epithelial progenitor cell niche and differentiation of the zymogenic (chief) cell lineage. Dev Biol 2009;325:211-24.
87. McDonald SA, Greaves LC, Gutierrez-Gonzalez L, et al. Mechanisms of field cancerization in the human stomach: the expansion and spread of mutated gastric stem cells. Gastroenterology 2008;134:500-10.

88. Goldenring JR, Nam KT, Wang TC, et al. Spasmolytic polypeptide-expressing metaplasia and intestinal metaplasia: time for reevaluation of metaplasias and the origins of gastric cancer. Gastroenterology 2010;138:2207-10, 2210 e1.
89. Bornschein J, Kandulski A, Selgrad M, et al. From gastric inflammation to gastric cancer. Dig Dis 2010;28:609-14.
90. Fuccio L, Zagari RM, Eusebi LH, et al. Meta-analysis: can Helicobacter pylori eradication treatment reduce the risk for gastric cancer? Ann Intern Med 2009;151:121-8.
91. Yakirevich E, Resnick MB. Pathology of gastric cancer and its precursor lesions. Gastroenterol Clin North Am 2013;42:261-84.
92. Correa P. A human model of gastric carcinogenesis. Cancer Res 1988;48:3554-60.
93. Gomceli I, Demiriz B, Tez M. Gastric carcinogenesis. World J Gastroenterol 2012;18:5164-70.
94. Pereira MI, Medeiros JA. Role of Helicobacter pylori in gastric mucosa-associated lymphoid tissue lymphomas. World J Gastroenterol 2014;20:684-98.
95. Ferlay J SI, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11: Lyon, France: International Agency for Research on Cancer, 2013.
96. Ferlay J, Shin HR, Bray F, et al. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010;127:2893-917.
97. Ferlay J. SI, Ervik M., Dikshit R., Eser S., Mathers C., Rebelo M., Parkin D.M., Forman D., Bray, F. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11. Volume 2014: International Agency for Research on Cancer, Lyon, France, 2013.
98. Correa P. Gastric cancer: overview. Gastroenterol Clin North Am 2013;42:211-7.
99. Piazuelo MB, Correa P. Gastric cancer: Overview. Colomb Med (Cali) 2013;44:192-201.
100. Parkin DM, Stjernsward J, Muir CS. Estimates of the worldwide frequency of twelve major cancers. Bull World Health Organ 1984;62:163-82.
101. Schistosomes, liver flukes and Helicobacter pylori. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Lyon, 7-14 June 1994. IARC Monogr Eval Carcinog Risks Hum 1994;61:1-241.
102. Bouvard V, Baan R, Straif K, et al. A review of human carcinogens--Part B: biological agents. Lancet Oncol 2009;10:321-2.
103. Plummer M, Franceschi S, Vignat J, et al. Global burden of gastric cancer attributable to pylori. Int J Cancer 2014.
104. Hemminki K, Zhang H, Czene K. Socioeconomic factors in cancer in Sweden. Int J Cancer 2003;105:692-700.
105. Power C, Hypponen E, Smith GD. Socioeconomic position in childhood and early adult life and risk of mortality: a prospective study of the mothers of the 1958 British birth cohort. Am J Public Health 2005;95:1396-402.

106. Pereira L, Zamudio R, Soares-Souza G, et al. Socioeconomic and nutritional factors account for the association of gastric cancer with Amerindian ancestry in a Latin American admixed population. PLoS One 2012;7:e41200.
107. Ladeiras-Lopes R, Pereira AK, Nogueira A, et al. Smoking and gastric cancer: systematic review and meta-analysis of cohort studies. Cancer Causes Control 2008;19:689-701.
108. Chun N, Ford JM. Genetic testing by cancer site: stomach. Cancer J 2012;18:355-63.
109. McLean MH, El-Omar EM. Genetics of gastric cancer. Nat Rev Gastroenterol Hepatol 2014.
110. Graham DY. History of Helicobacter pylori, duodenal ulcer, gastric ulcer and gastric cancer. World J Gastroenterol 2014;20:5191-204.
111. Shiotani A, Graham DY. Pathogenesis and therapy of gastric and duodenal ulcer disease. Med Clin North Am 2002;86:1447-66, viii.
112. Uemura N, Okamoto S, Yamamoto S, et al. Helicobacter pylori infection and the development of gastric cancer. N Engl J Med 2001;345:784-9.
113. Porras C, Nodora J, Sexton R, et al. Epidemiology of Helicobacter pylori infection in six Latin American countries (SWOG Trial S0701). Cancer Causes Control 2013;24:209-15.
114. Bonequi P, Meneses-Gonzalez F, Correa P, et al. Risk factors for gastric cancer in Latin America: a meta-analysis. Cancer Causes Control 2013;24:217-31.
115. Adamsson J, Lundin SB, Hansson LE, et al. Immune responses against Helicobacter pylori in gastric cancer patients and in risk groups for gastric cancer. Helicobacter 2013;18:73-82.
116. Lee HJ, Nam KT, Park HS, et al. Gene expression profiling of metaplastic lineages identifies CDH17 as a prognostic marker in early stage gastric cancer. Gastroenterology 2010;139:213-25 e3.
117. Mutz KO, Heilkenbrinker A, Lonne M, et al. Transcriptome analysis using next-generation sequencing. Curr Opin Biotechnol 2013;24:22-30.
118. McGettigan PA. Transcriptomics in the RNA-seq era. Curr Opin Chem Biol 2013;17:4-11.
119. Westermann AJ, Gorski SA, Vogel J. Dual RNA-seq of pathogen and host. Nat Rev Microbiol 2012;10:618-30.
120. Yahara K, Furuta Y, Oshima K, et al. Chromosome painting in silico in a bacterial species reveals fine population structure. Mol Biol Evol 2013;30:1454-64.
121. Salzberg SL, Phillippy AM, Zimin A, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res 2012;22:557-67.
122. Magoc T, Pabinger S, Canzar S, et al. GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics 2013;29:1718-25.
123. Bankevich A, Nurk S, Antipov D, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 2012;19:455-77.
124. Gladman S ST. Volume 2012, 2012.
125. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.

126. Clark SC, Egan R, Frazier PI, et al. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. Bioinformatics 2013;29:435-43.
127. Hyatt D, Chen GL, Locascio PF, et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 2010;11:119.
128. Resende T, Correia DM, Rocha M, et al. Re-annotation of the genome sequence of Helicobacter pylori 26695. J Integr Bioinform 2013;10:233.
129. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15-21.
130. Bengtsson-Palme J, Hartmann M, Eriksson KM, et al. Metaxa2: Improved Identification and Taxonomic Classification of Small and Large Subunit rRNA in Metagenomic Data. Submitted 2014:1-14.
131. de Martel C, Ferlay J, Franceschi S, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. Lancet Oncol 2012;13:607-15.
132. CIA. CIA Wold Factbook.
133. Devi SM, Ahmed I, Khan AA, et al. Genomes of Helicobacter pylori from native Peruvians suggest admixture of ancestral and modern lineages and reveal a western type cag-pathogenicity island. BMC Genomics 2006;7:191.
134. Camorlinga-Ponce M, Perez-Perez G, Gonzalez-Valencia G, et al. Helicobacter pylori genotyping from American indigenous groups shows novel Amerindian vacA and cagA alleles and Asian, African and European admixture. PLoS One 2011;6:e27212.
135. Saunders NJ, Boonmee P, Peden JF, et al. Inter-species horizontal transfer resulting in core-genome and niche-adaptive variation within Helicobacter pylori. BMC Genomics 2005;6:9.
136. Bengtsson J, Eriksson KM, Hartmann M, et al. Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. Antonie van Leeuwenhoek 2011;100:471-475.
137. Fox JG, Wang TC. Inflammation, atrophy, and gastric cancer. J Clin Invest 2007;117:60-9.
138. Bindea G, Mlecnik B, Angell HK, et al. The immune landscape of human tumors: Implications for cancer immunotherapy. Oncoimmunology 2014;3:e27456.
139. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009;10:57-63.