Inaugural dissertation

for

obtaining the doctoral degree

of the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of the

Ruprecht - Karls - University Heidelberg

presented by

Rima Jeske, M.Sc.

born in Woswyschenka, Kazakhstan

Oral examination: 29.07.2022

Identification of novel serological biomarkers for non-cardia gastric cancer using *Helicobacter pylori* multi-strain microarrays

Referees:

apl. Prof. Dr. Martin Müller

PD Dr. Ralf Bischoff

Acknowledgements

The here presented project was performed under the supervision of Dr. Tim Waterboer in the Division of Infections and Cancer Epidemiology at the German Cancer Research Center in Heidelberg, Germany, from October 2018 until May 2022.

I would like to thank Prof. Dr. Martin Müller and PD Dr. Ralf Bischoff for reviewing this dissertation and for accompanying the project during the last years. A big thank you also goes to the other members of my thesis advisory committee, Prof. Dr. Paul Schnitzler and Dr. Eli Zamir, for contributing their knowledge and time.

My special thanks go to Dr. Tim Waterboer for welcoming me into his division, supporting me in stressful situations and for always supplying coffee.
A further thank you goes to all current and former members of the division for the support inside and outside of the lab. I am very grateful that I was given the opportunity to join the team. Thank you to Claudia Brandel, Moni Oppenländer and Ute Koch for managing experiments and staying calm throughout stressful experiments. Thank you to Dr. Noemi Bender and Dr. Nicole Brenner for their organizational, scientific and emotional support whenever I needed it.
A huge thank you goes to Dr. Julia Butt for supporting me with her knowledge, time and experience. I am grateful for the motivation and encouragement, particularly in the last couple of weeks.

I would also like to thank my students who joined me in the lab and made work more enjoyable despite an ongoing pandemic, Thank you to Amber, Leander, Anna, Domi, Stella and Nina.

I would also like to thank all my collaborators who shared their expertise, samples and time: Dr. Christian Abnet and Dr. Yingxi Chen from the NIT and Shanxi study; Dr. Constanza Camargo and Dr. Charles Rabkin from the *Helicobacter pylori* genome project; Dr. Nerea Fernández de Larrea-Baz and Nuria Aragonés from the MCC Spain study.

Zuletzt möchte ich meinen Freunden und meiner Familie für ihre bedingungslose Unterstützung in allen Lebenslagen danken.

Zusammenfassung

Infektionen mit dem Bakterium *Helicobacter pylori* (*H. pylori*) sind die Hauptursache für die Entwicklung von Magenkrebs, welche weltweit die fünfthäufigste Tumorerkrankung und vierthäufigsten Todesursache durch Krebs darstellt. Obwohl etwa die Hälfte der weltweiten Bevölkerung mit *H. pylori* infiziert ist, kommt es in nur 3-5% der Infizierten zu einer Entwicklung von Magenkrebs. Daher empfiehlt es sich eine Risikogruppe zu definieren, um eine effiziente Vorsorge zu ermöglichen, beispielsweise durch den Nachweis von Antikörpern gegen *H. pylori*. Bisher sind hauptsächlich Antikörperantworten gegen ein Gesamt-Proteinextrakt des Erregers oder gegen bekannte Virulenz Faktoren beschrieben, doch die Assoziationen mit Magenkrebs sind nicht ausreichend für eine effiziente Stratifizierung des Krebsrisikos. Ziel dieser Arbeit war es, *H. pylori* Antigene *de novo* zu identifizieren, die als Risikomarker verwendet werden können. Hierfür habe ich einen *H. pylori* Microarray generiert, der eine Vielzahl verschiedener Antigene parallel präsentiert. In einer Pilotstudie habe ich damit Antikörperantworten gegen 245 *H. pylori* Antigene in Seren von Patienten mit Magenkrebs und gesunden Kontrollen gemessen. Dadurch konnte ich eine Assoziation mit Antikörpern gegen das Cytotoxin-associated gene A (CagA) bestätigen, sowie eine Assoziation mit Antikörpern gegen ein neu-identifiziertes Antigen, das *Helicobacter* outer membrane porin A (HopA), beschreiben.

In einem nächsten Schritt habe ich die Anzahl der *H. pylori* Antigene pro Microarray auf 1,833 vergrößert, um die genetische Varianz des Bakteriums zu adressieren. Dazu habe ich *H. pylori* Stämme aus verschiedenen Ländern verwendet.

Diese *H. pylori* ‚multi-strain' Microarrays wurden mit chinesischen Seren aus der Shanxi Querschnittsstudie (58 Fälle und 58 Kontrollen) und aus dem prospektiven ‚Linxian General Population Nutrition Intervention Trial' (119 Fälle und 119 Kontrollen) beprobt. Anschließend wurden 22 *H. pylori* Antigene zur weiteren Validierung mittels Multiplex Serologie, einer Hochdurchsatz-Plattform, ausgewählt. Diese wurden als GST-X-TAG Fusionsproteine exprimiert und entsprechende Antikörperantworten in >1,600 Teilnehmern beider chinesischer Studien gemessen.

Assoziationen zwischen Magenkrebs und Antikörpern gegen die jeweiligen *H. pylori* Antigene wurden mittels logistischer Regression ermittelt. HopA konnte als neuer serologischer Marker für Magenkrebs bestätigt werden, doch Antikörper gegen die restlichen *H. pylori* Antigene zeigten einzeln keine konsistente Assoziation mit Magenkrebs. Ob das hier identifizierte HopA mit weiteren klinischen Parametern zur Ermittlung des Magenkrebs-Risikos und somit zur Verbesserung der Magenkrebs-Vorsorge beitragen kann, muss in weiteren Studien geprüft werden.

Summary

Infection with the bacterium *Helicobacter pylori* (*H. pylori*) is the main cause for gastric cancer (GC) and although approximately half of the world population is infected, only 3-5% develop the neoplasia. Globally, this makes GC the fifth most commonly diagnosed cancer and fourth most common cause of death from cancer. In order to enable efficient screening for GC, identifying individuals at risk is desirable, which could be achieved by measuring antibody responses against *H. pylori*. So far, serological associations between GC and *H. pylori* have mainly been described for an overall infection or for well-known virulence factors, but did not show sufficiently strong associations to achieve a relevant stratification for GC risk.

The aim of this thesis was to identify further *H. pylori* antigens, which could be used to enhance GC risk stratification. For this purpose, I generated *H. pylori* microarrays presenting a multitude of antigens in parallel. A pilot study was conducted probing 245 different *H. pylori* antigens with sera from non-cardia gastric cancer (NCGC) patients and healthy controls. I was able to confirm the well-known association between NCGC and antibodies against the Cytotoxin-associated gene A (CagA) and described a new association between NCGC and antibodies against the *Helicobacter* outer membrane porin A (HopA).

In a next step, I expanded the antigen repertoire of the *H. pylori* microarrays to a total of 1,833 non-redundant antigens by including additional *H. pylori* strains from different countries to account for the genetic variety of the bacterium.

H. pylori multi-strain microarrays were probed with sera from the cross-sectional Shanxi study (58 cases and 58 controls) and the prospective Linxian General Population Nutrition Intervention Trial (119 cases and 119 controls), both samples in Northern China. Subsequently, I selected 22 *H. pylori* antigens to be further validated using high-throughput multiplex serology. To enable this, I expressed the selected *H. pylori* antigens as GST-X-TAG fusion proteins and measured respective antibody responses in >1,600 samples derived from participants of two Chinese studies.

Associations between NCGC and antibodies against individual *H. pylori* proteins were examined by logistic regression analyses. Thereby, I confirmed HopA as new serological NCGC marker which performed comparable to the widely used CagA. The other new *H. pylori* antigens individually did not show a consistent association with NCGC between the analyzed studies. However, a combinational approach could potentially enable the generation of enhanced GC risk models. Future studies should address the clinical applicability of the newly identified markers to potentially promote secondary prevention of GC.

Contents

1 Introduction

1.1 Gastric Cancer

1.1.1 Epidemiology and classification

GC represents a major health burden and is currently the fifth most commonly diagnosed cancer. With a global age-standardized incidence rate of 15.8 per 100,000 males and 7.0 per 100,000 females, GC accounted for 5.6% of all cancer diagnoses in 2020.[1] These incidence rates show large geographical variation and are particularly high in East Asia, Eastern Europe and South America (Figure 1). In Iran, Afghanistan and Turkmenistan GC is the leading type of cancer among men, but out of 1.2 million newly diagnosed GC cases in 2020, approximately 40% occurred in China. [1, 2]
Until 1940, GC was the most common cancer type among males globally. Since then, incidence rates have been steadily declining, but absolute numbers will rise due to overaging and population growth.[1,3] By 2035, e.g., GC diagnoses in China are expected to increase by 50%.[4] This means that GC will remain a major global health burden unless prevention or early detection measures are implemented.

With a 5-year survival rate of 31% in the US, 26% in Europe and 27% in China, the prognosis of GC is poor, because neoplasms are usually diagnosed at an advanced, metastatic stage.[5, 6] Globally, GC ranks fourth for mortality being responsible for 7.7% of all cancer-related deaths.[1] Japan and Korea, which are among the countries with the highest incidence rates for GC worldwide, initiated screening programs in 1983 and 1999, respectively.[7, 8] This led to a substantially increased 5-year survival rate of 60% and 69%, respectively, supporting the impact of early detection (Figure 1).[9]

1.1.2 Classification and pathology

About 90% of all GCs are adenocarcinomas arising in the gastric epithelial tissue. The remaining 10% comprise mucosa-associated lymphoid tissue (MALT) lymphomas and leiomyosarcomas.[10] Adenocarcinomas are further stratified by their anatomical site: 73% arise in the body of the stomach as ‘non-cardia gastric cancer’ (NCGC). In East Asia, a higher proportion of up to 90% has been reported. The remaining neoplasms are found in the fundus of the stomach, in proximity to the esophagus. These are referred to as ‘cardia gastric cancer’ (CGC). NCGC and CGC differ in clinical, pathological and epidemiological characteristics (described below).[11 12]

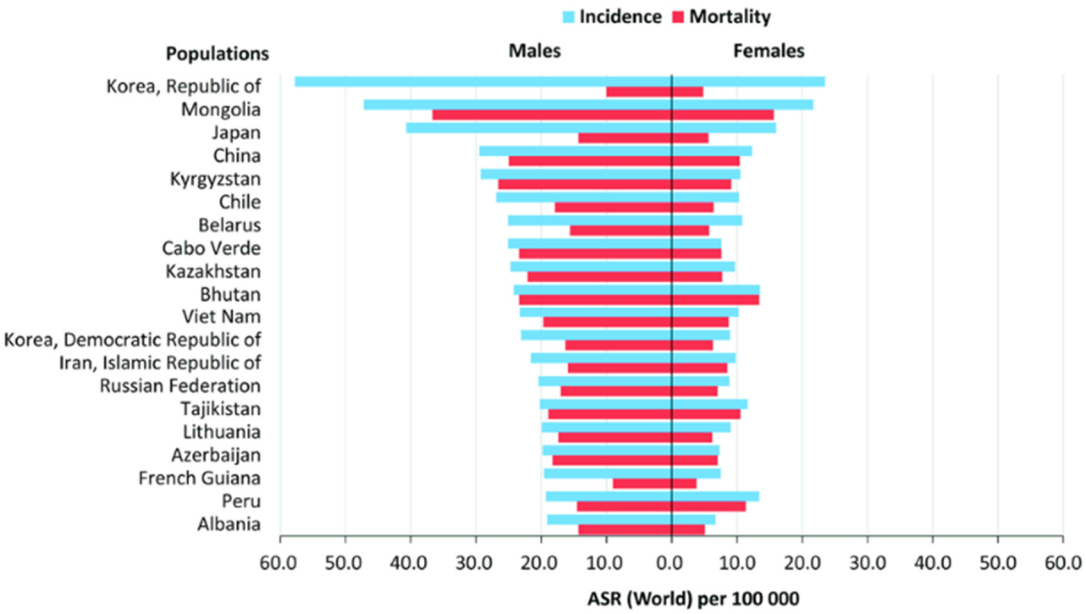


Figure 1: GC rates
Age-standardized incidence and mortality rates of GC for males and females in selected countries. Adapted from Liou et al.[13]

GCs can be further stratified by their pathohistological subtype into 'intestinal-type' or 'diffuse-type' according to Lauren's classification.[14] Countries with high GC incidences are usually dominated by intestinal-type GCs.[11] They are the outcome of sequential pathological changes in the gastric mucosa referred to as 'Correa's cascade' (Figure 2).[15, 16] This cascade is usually initiated through colonization with the bacterium *H. pylori*. Gastric cells accumulate genetic and epigenetic alterations until a 'point of no return' is reached when *H. pylori* no longer drives carcinogenesis. The exact time point, however, has yet to be determined.[17]
The transition from healthy tissue to precancerous states and adenocarcinoma has a latency time of many years and resembles the Vogelstein model for colorectal cancer.[17, 18] It is furthermore hypothesized that the bacterium disappears due to an unfavorable milieu that accompanies gastric carcinogenesis, and can no longer be detected at time of GC diagnosis. This observation was described as 'hit-and-run' mechanism.[19, 20]

After the initial colonization of the gastric epithelial mucosa by *H. pylori*, chronic inflammation manifests. This state is hallmarked by increased infiltration of neutrophilic and mononuclear cells into the gastric tissue and can induce oxidative stress and DNA methylation.[16, 21] Still, most infected individuals (~90%) remain asymptomatic, although approximately half of them, eventually, experience a loss of normal gastric architecture with glandular structures being replaced by fibrosis and intestinal-type epithelium. [11, 16]

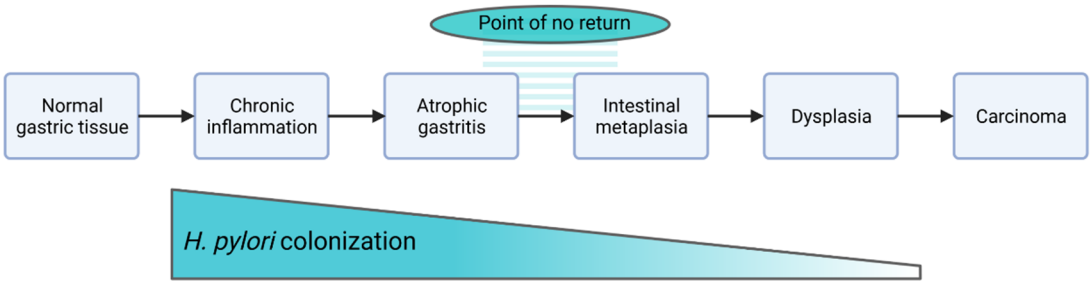


Figure 2: Correa's cascade
Intestinal-type adenocarcinomas are the outcome of a sequential pathological change termed 'Correa's cascade'. H. pylori *usually initiates the transition from normal epithelial tissue to precancerous lesions and carcinoma. Eventually, a 'point of no return' is reached at which* H. pylori *is no longer driving the transition but accumulated genetic and epigenetic aberrations. Adapted from Uno* et al.*, created with BioRender.com.*[17]

About 8% of the individuals who exhibit intestinal metaplasia, develop noninvasive neoplasia (=dysplasia). Their respective epithelium shows a neoplastic phenotype with enlarged and crowded nuclei.[16] The final stage of the cascade is the transition to carcinomas which occurs in only 3-5% of *H. pylori* infected individuals.[11] It is characterized by the penetration of neoplastic cells into the surrounding stroma.[16] Overall, morphological changes are accompanied by sporadic genetic and epigenetic alterations which accumulate during carcinogenesis. They particularly affect regulatory genes, e.g. mediators of the p53 pathway, the RB1 pathway, the TGFβ pathway or the APC/β-catenin pathway.[22]

Diffuse-type adenocarcinomas are less well-characterized, but were also found associated with a prior *H. pylori* infection, although the exact mechanism is not known, so far. They usually affect younger patients and progress more rapidly to a metastatic state.[23, 24] Furthermore, they do not establish precancerous lesions, but are generally characterized by the lack of glandular organization and a poorly-differentiated cell morphology. This aggravates early detection and contributes to a poor prognosis.[24] Generally, diffuse-type GC shows a stronger genetic component than intestinal-type GC. About 5-10% of them are considered hereditary, e.g. caused by germline mutations in the tumor suppressor gene *Cadherin 1* (*CDH1*) which encodes E-cadherin[25, 26] This protein is essential to form cell-cell adhesions and tight junction complexes that are crucial for cell differentiation and proliferation.[24] Most cases of diffuse-type GC, however, occur sporadically and exhibit similar somatic mutations as intestinal-type GC.[22]

1.1.3 Risk factors

GC is a multifactorial disease and the result of an unfavorable interplay between *H. pylori* infection, environmental factors and genetic predispositions. Specific risk factors vary depending on the anatomic subsite and pathohistology of the neoplasm.

Risk factors which are similar for CGC and NCGC are older age, male gender, a low socioeconomic status and smoking. NCGC is further associated with a diet high in salt and processed meat, but low in fruit. The major risk factor, however, is *H. pylori* with almost 90% of NCGC attributable to the infection, irrespective of the pathohistological subtype (intestinal-type and diffuse-type).[1, 27, 28]

The role of *H. pylori* in the development of CGC is not as clear and results have been contradictory. Modest, but significant, associations were described in East Asian countries, while associations in Western countries were not significant.[29, 30] Recent studies proposed a dual etiology for CGC, with some CGC cases attributable to an *H. pylori* infection and others caused by obesity and gastroesophageal reflux disease injury. This could partially explain geographical differences, as the proportion of CGC cancers compared to NCGC cases is higher in countries with low *H. pylori* incidences.[31] Noteworthy, an increasing rate of CGC among young adults (<50 years) has been noticed over the past years in countries with high and low GC incidence rates.[4]

A further infectious agent which is considered to provoke GC is Epstein-Barr virus (EBV). Approximately 10% of GC biopsies show traces of clonal EBV-encoded small RNA (EBER), indicating a causative role of EBV.[32-34] These neoplasms can be CGC (58%) or NCGC (42%) and form a molecular subtype, which was not found associated with *H. pylori* infection.[35, 36] A potential role of co-infections is, however, not well understood.[37]

Additionally, genetic predispositions can play a role in the development of GC, e.g. germline mutations in the *CDH1* gene causing hereditary, diffuse-type CGC.[26]
Allelic variations of other genes, particularly genes involved in inflammatory response and other interleukins, were found to increase susceptibility for GC.[38] Examples include *IL1B, IL1RN, TNF* and *HLA*. Meta-analyses, however, revealed that many genetic variations could not be identified consistently across different populations and associations with GC were of only moderate strength.[39]

1.2 *Helicobacter pylori*

1.2.1 Epidemiology and phylogeographic variations

The Gram-negative bacterium *H. pylori* colonizes the human gastric mucosa. It was first described in 1982 by two Australian physicians, Dr. Barry Marshall and Dr. John Robin Warren, who recognized and described the pathological link between *H. pylori* infection, gastritis and peptic ulcers.[40, 41] This was honored by the Nobel committee in 2005. A few years later epidemiological studies further revealed the causative role in GC. So far, *H. pylori* is the only bacterium classified as a type I carcinogen.[42]

H. pylori is one of the most widespread human pathogens with a global prevalence of approximately 50%.[43] Geographical differences have been reported with 79% of the general adult population infected in Africa, 63% in Latin America and the Caribbean, 55% in Asia and 47% in Europe.[43] Generally, prevalence rates are higher in developing countries than in industrialized countries, but a declining trend has been observed worldwide. Infections are usually transmitted in early childhood within families and accompany humans throughout life.[42]

The bacterium has been inhabiting human stomachs for at least 50,000 years and co-evolving together with its host. Eventually, early human migration (e.g. prehistoric colonization of Polynesia and America) led to the segregation of the species *H. pylori.* This resulted in a phylogeny, which closely corresponds to the human migration and shows a 'phylogeographic pattern'.[44, 45] This means that *H. pylori* strains which are isolated in geographical proximity are closer related than isolates from two distinct geographical origins.

As a result, *H. pylori* strains can nowadays be assigned one of eight clades: hpEurope, hpAmerind, hpAfrica, hpAfrica2, hpNEAsia, hpAsia2, hpEastAsia or hspMaori, based on single nucleotide polymorphisms of seven housekeeping genes (Figure 3).[44, 45] However, in the last centuries, the phylogeographic pattern of *H. pylori* has been blurring due to colonization and recent migration waves. Strains of the clade hpAfrica and hpNEAfrica can be found on the American continent, while strains of the clade hpEastAsia have spread to Oceania. Furthermore, strains of the clade hpEurope became common on the whole continent of America and in Oceania. [46]

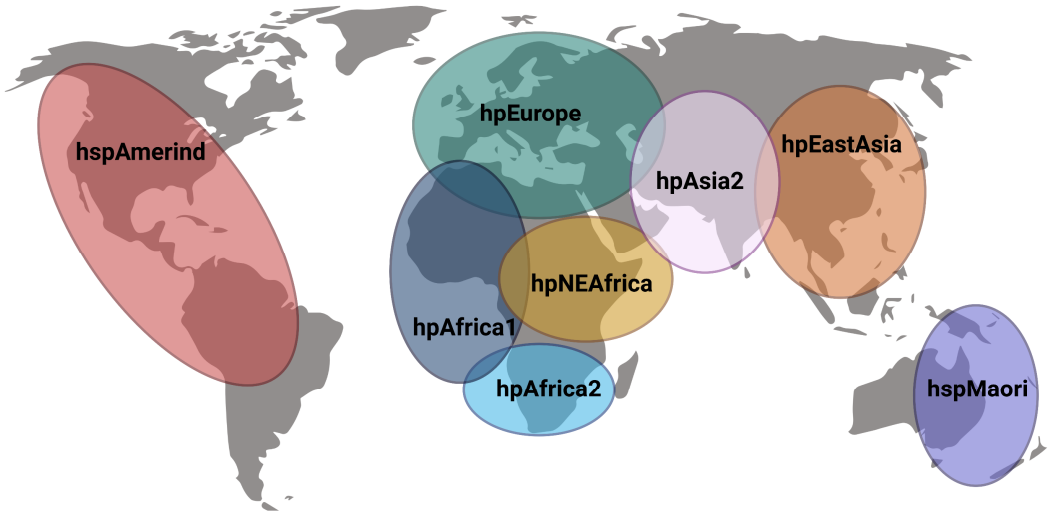


Figure 3:* H. pylori *clades
Geographical distribution of eight H. pylori *clades before the era of colonization and modern migration. Adapted from Yamaoka* et al.*, created with BioRender.com.*[46]

1.2.2 Pathology and virulence

Although *H. pylori* is considered the main risk factor for GC development, many people tolerate the infection and do not show any symptoms throughout their life. In only 3-5% of them, the infection initiates a pathological change which leads to GC.

Numerous studies have been investigating the genetic diversity of *H. pylori* trying to find features which explain why some people develop GC and others not.[47] A special focus was put on *H. pylori* virulence factors. Among them, by far the most investigated is CagA, which is encoded on the *cag* pathogenicity island (*cag* PAI) alongside components of a bacterial type IV secretion system (T4SS).[48]
This secretion system directly injects CagA into gastric epithelial cells, where it gets phosphorylated at its Glu-Pre-Ile-Tyr-Ala (EPIYA) motif. Phosphorylated CagA then interacts with numerous host signaling molecules deregulating different pathways, e.g., the Wnt or PI3K-AKT pathway. Eventually, transformation of gastric epithelial cells can be induced which then may lead to the development of GC.[49, 50] The efficacy of this transformation was found associated to allelic variations of CagA, which mainly affect the EPIYA motif. [51] CagA can be found in all *H. pylori* clades, except for hpAfrica2. Virtually every East Asian *H. pylori* strain expressed CagA, while only 60% of the European strains encode this protein.[52]

Another well-characterized risk factor is the Vacuolating cytotoxin autotransporter A (VacA). It was found to induce disruption of mitochondrial functions, stimulate apoptosis and block T-cell proliferation.[53] As for CagA, allelic variations exist that increase the virulence of an *H. pylori* strain.[46]

Outer membrane proteins (OMPs) comprise another group of virulence factors. They have essential functions, including structural maintenance or transportation through the bacterial membrane. Furthermore they have direct contact with host cells, which predisposes them as potential virulence mediators.[54]
Most OMPs belong to the Hop (*Helicobacter* outer membrane porins) family or Hor (Hop-related proteins) family. Multiple members of these families have been described associated with GC development. These include the blood-group-antigen-binding adhesin BabA (HopS), the sialic-acid-binding adhesin SabA (HopP), the outer inflammatory protein A OipA (HopH), HopQ and HopZ.[55-60] BabA, SabA and HopQ directly interact with the T4SS and enhance translocation of CagA into gastric epithelial cells. OipA and HopZ mediate adhesion to host cells, which indirectly influences the efficiency of the T4SS.[54]

Numerous other risk factors have been described in context of GC and summarized in multiple review articles.[61, 62] Still, no conclusive explanation was found why some people develop GC, while others remain asymptomatic despite an *H. pylori* infection. Furthermore, associations with cancer status often vary across different populations.

1.3 Prevention of gastric cancer

With a lead time of several years, high incidence rates and potentially detectable precancerous lesions, GC is a suitable target for primary and secondary prevention.

Japan and South Korea introduced screening programs based on endoscopy of the upper gastrointestinal tract to early detect precancerous lesions. Although participation rates only reach 10% in Japan and 44% in Korea, both countries have been observing a reduction of GC-associated mortality by approximately 30% and 50%.[7, 8, 63-65] Noteworthy, endoscopy is often performed outside of screening programs, especially in Japan, which means that participation rates underestimate people getting examined for potential GC precursor lesions.[65]

Despite confirmed successful, secondary prevention by population-based endoscopic screening is not feasibly in every country. E.g., China has a much larger population and not the resources for laborious and expensive endoscopies on a large scale.[66] Non-invasive or minimally-invasive tests, which can be performed rapidly in a high-throughput manner, would be considered superior to detect GC early.[67] Current laboratory assays, however, lack sensitivity and specificity to meet this need.

An alternative solution is stratifying individuals by their risk to develop GC and include those at high-risk into respective screening programs, e.g. based on endoscopy. Thereby, overall costs could be decreased and scarce resources could be used more efficiently.
For this purpose, the 'ABC' method was developed.[68] It combined an *H. pylori* assay detecting infection, with a pepsinogen assay detecting gastric atrophy. The transition from healthy gastric tissue to atrophic gastritis and intestinal metaplasia is marked by a decreased secretion of the enzyme pepsinogen I (PGI) and an increased secretion of pepsinogen II (PGII). The ratio of these enzymes has been proposed as a biomarker to measure atrophy.[69]
The ABC method assigns patients to one of four risk groups based on the results of these assays and proposes adapted interventions, respectively.[69, 70] Although this risk stratification showed promise in some populations, it did not show good discrimination in populations with generally high *H. pylori* incidences, e.g. in China.[71, 72] In order to refine risk estimations in these countries, addressing *H. pylori* virulence factors has been proposed.[73, 74]

The association between CagA and GC is well known and could be used for this kind of risk stratification, e.g. by indirectly detecting serum antibodies against CagA. This approach seems practical in Western countries, but barely effective in East Asian countries. In those populations, *H. pylori* strains of the clade hpEastAsia are common and virtually every strain of this clade carries a *cag* PAI, which encodes CagA. This means that anti-CagA serum antibodies would rather be infection than risk markers, especially as common serological assays cannot address genetic variants of CagA. As a consequence, the strength of the association between CagA serum antibodies and GC is reduced in East Asian compared to Western populations.
Hence, the need for additional markers is high, especially markers which could be measured serologically by detecting antibodies. Blood draw is only minimally invasive and respective assays could be performed in low-resource settings and a high-throughput fashion.

Using *H. pylori* antibodies to stratify infected individuals into risk groups has the further advantage that it could be measured even before pathological changes in the gastric mucosa become visible and potentially irreversible. This would potentially enable primary prevention of GC, because individuals at risk could be preventatively treated for the infection.

There is clear evidence that eradication of *H. pylori* can reduce GC mortality rates and an evaluation by the International Agency for Research on Cancer (IARC) of the World Health Organization found the approach feasible and cost-effective.[75, 76] Yet, no country has implemented large-scale eradication programs due to uncertainty about potential adverse effects. Most importantly, widespread eradication programs would give rise to antibiotic resistances, e.g., clarithromycin and levofloxacin resistances increased by 21% and 27%, respectively, between 2000 and 2015.[13]

Altogether, this means that risk stratification is applicable to support secondary and primary prevention of GC, as individuals at risk could be closer monitored for precancerous lesions or preventatively treated for *H. pylori* infection. Especially serological assays enabling risk stratification are desired due to their eligibility for large-scale screening scenarios.

1.4 *H. pylori* serology

Serological assays detect antibodies in peripheral blood, serum or plasma. Antibodies are part of the immunological memory of an individual and can be used to detect prior contact with an infectious agent. This means that serological assays can usually be used even after an infection was cured, either spontaneously or by eradication therapy. On a population-level, measuring antibodies enables seroepidemiological research investigating potential associations between exposures and outcomes.
In order to describe the strength of an association, e.g. seropositivity to *H. pylori* (exposure) and GC (outcome), odds ratios (OR) can be calculated. The odds for an outcome at the presence of exposure is divided through the odds for an outcome at the absence of exposure. OR above the value 1 indicate a positive association between exposure and outcome, while ORs below the value 1 indicate an inverse association. However, only if 95% confidence intervals (95% CI) do not include the value 1, associations can be considered significant.

For *H. pylori* infection and NCGC, ORs between 3.0 (95% CI: 2.3 - 3.8) and 21.4 (95% CI: 7.1 – 64.4) were reported, which corresponds to a 3.0- to 21.4-fold increase for NCGC given an *H. pylori* infection. Exact point estimates depend on the population and the diagnostic assay.[77]

1.4.1 *H. pylori* multiplex serology

Multiplex serology is a bead-based high-throughput assay, which enables measuring antibody responses against up to 100 antigens in a single reaction, while only requiring few microliters of serum or plasma.[78, 79] Respective antigens are fused to an N-terminal glutathione-S-transferase (GST) and a C-terminal peptide comprising the last seven amino acids of the SV40 large T antigen (TAG). This enables loading GST-X-TAG fusion proteins, X being the antigen, onto glutathione-casein coupled fluorescent polystyrene beads. Each antigen is assigned to a spectrally distinguishable bead type. To quantify serum antibodies, secondary biotinylated anti-human antibodies are used in combination with the fluorophore streptavidin-R-phycoerythrin (strep-PE) (Figure 4). Output signals are generated using a Luminex 200 analyzer which identifies the respective bead type and simultaneously translates strep-PE signals into median fluorescence intensity (MFI) values.

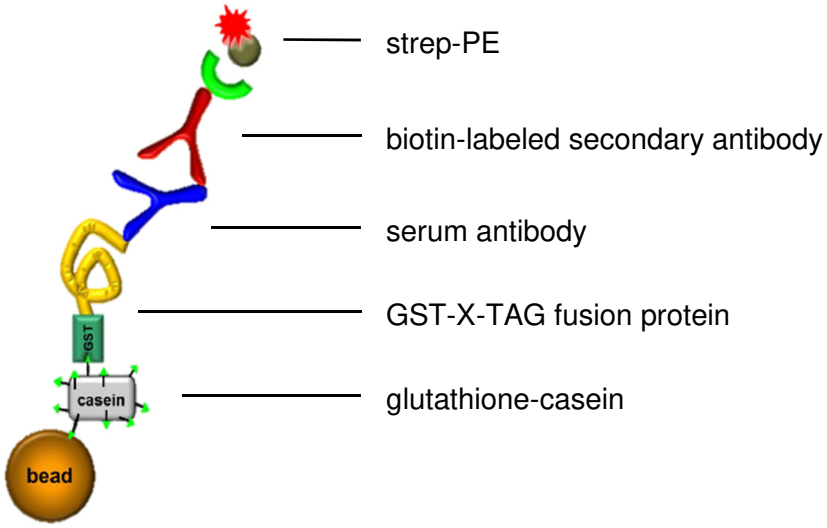


Figure 4: Assay principle of multiplex serology
Antigens are expressed as GST-X-TAG fusion proteins and coupled to casein-glutathione covered beads. As these beads are spectrally distinguishable, antibodies to different antigens can be measured in a single reaction. Bound serum antibodies are quantified using secondary biotinylated anti-human antibodies and Strep-PE. Adapted from Waterboer et al.[78]

Multiplex serology has been used in multiple large seroepidemiological studies, e.g. to validate serological risk markers.[80-82] It is a versatile platform and currently comprises antigens from different viruses, bacteria and protozoans, including *H. pylori*.[83] The selection of antigens is usually knowledge-driven and based on reported immunogenicity or association with certain outcomes, e.g. GC.[84-86] The established *H. pylori* panel currently covers 15 antigens (Table 1).[83]
All of these antigens derive from the common reference strain *H. pylori* 26695, except for GroEL which derives from *H. pylori* G27. The assay was compared to a commercial diagnostic assay using a German reference population comprising individuals who were seropositive for *H. pylori* (had detectable antibodies) and individuals who were seronegative for *H. pylori* (did not have detectable antibodies). Exhibiting antibodies against at least four specific *H. pylori* proteins was found to reach a specificity of 89% and sensitivity of 82% compared to the commercial assay.[83]

H. pylori multiplex serology has been used in different studies investigating the association between *H. pylori* and diverse clinical outcomes, including but not limited to GC. [87] Antigens which have been found informative in East Asian populations were e.g., HP1564 and HP0305. Seropositivity to these antigens was associated with NCGC exhibiting ORs of 2.84 (95% CI: 2.35 – 3.44) and 2.00 (95% CI: 1.74 – 2.30), respectively.[88] Associations to precancerous lesions were even stronger with OR of

5.37 (95% CI: 4.20–6.89) and 3.85 (95%CI: 3.04–4.88), respectively. Double-seropositivity to HP1564 and HP0305 markers increased the OR to 7.43 (95% CI: 5.59–9.88).[89] Recently, Murphy *et al.* showed that including HP1564 and HP0305 into a predictive model for GC development, significantly improved accuracy compared to the ABC method.[90] This demonstrates that alternative targets for serological risk stratification of *H. pylori* infected individuals can show promise.

All of the *H. pylori* proteins which are part of the multiplex serology panel, were initially found immunogenic on 2D immunoblots.[84, 85] However, this classical approach to identify immunogenic proteins has some limitations, e.g. denatured antigen presentation and low separation of similar sized antigens.
Furthermore, only antibodies against a single *H. pylori* strain (usually *H. pylori* 26695) are detected and only few patient samples are included. It is well possible that further undiscovered risk markers exist which were not yet identified using classical methods.

Table 1: Established H. pylori *multiplex serology panel*

Locus tag	Trivial name	Protein
HP0010	GroEL	Chaperonin
HP0073	UreA	Urease alpha subunit
HP0231	-	Uncharacterized
HP0243	NapA	Neutrophil-activating protein A (bacterioferritin)
HP0305	-	Uncharacterized
HP0410	HpaA	*H. pylori* adhesin A
HP0547N	CagAN	N-terminus of the Cytotoxin-associated antigen A
HP0547C	CagAC	C-terminus of the Cytotoxin-associated antigen A
HP0695N	HyuAN	N-terminus of the hydantoin utilization protein A
HP0695C	HyuAC	C-terminus of the hydantoin utilization protein A
HP0875	Kat	Catalase
HP0887C	VacAC	C-terminus of the Vacuolating cytotoxin A
HP1098	HcpC	*Helicobacter* cysteine-rich protein C; putative beta-lactamase
HP1104	Cad	Cinnamyl alcohol dehydrogenase ELI3-2
HP1564	-	Lipoprotein

1.4.2 Antigen microarrays

An alternative way to identify antibodies against specific antigens are microarrays. These miniaturized test systems can represent numerous antigens at once in a high-density fashion. Thereby, a virtually unbiased screening of antigens is possible, without a knowledge-based or hypothesis-driven pre-selection.

A technique to generate bacterial whole-proteome microarrays was recently developed in our laboratory. This technique was based on a double spotting technique enabling cell-free *on-chip* expression of linear DNA constructs (Figure 5).[91]

DNA expression constructs were generated by two successive polymerase chain reactions (PCR). First, antigens of interest were amplified using gene-specific primer pairs, which added a shared sequence to the 5' and 3' ends of the amplicons. These shared sequences were then used as a template to add transcriptional and translational elements, as well as sequences coding for a V5- and a 6xHis-tag, in a second PCR.

The products of these second PCRs were spotted onto epoxysilane coated glass slides and subsequently overlaid with a second spot containing cell-free expression kit. During a subsequent incubation period, DNA expression constructs were transcribed and translated into corresponding proteins, flanked by two tags, directly on the slide.[91]

By probing these microarrays with serum from GC patients and healthy individuals, an association between seropositive antigens and case-control status can be estimated.

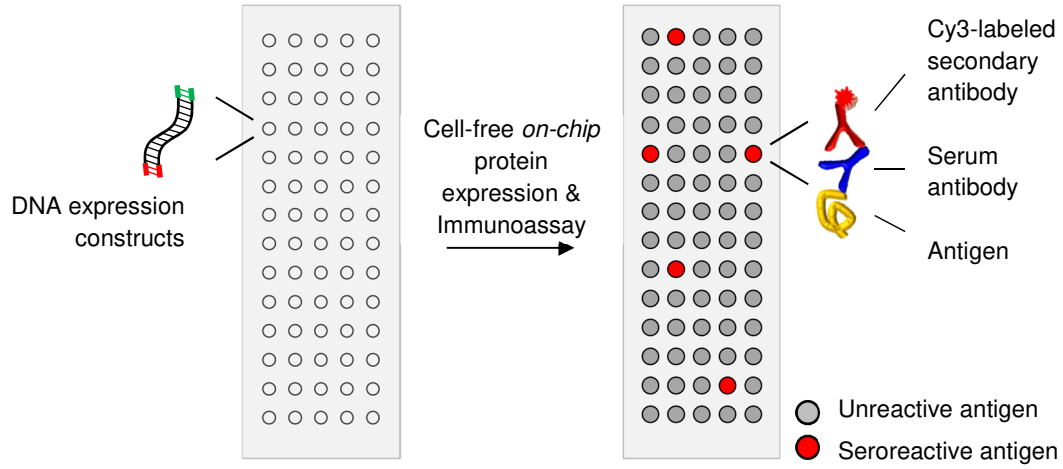


Figure 5: Generation of antigen microarrays

DNA expression constructs that encode the gene of interest flanked by two terminal tags, are spotted onto epoxysilane coated glass slides and overlaid with a second spot containing cell-free expression kit. In a following incubation step, DNA expression constructs are transcribed and translated to generate the corresponding antigen. By probing antigen microarrays with serum, specific serum antibodies can bind. They are subsequently visualized using a fluorescence-labeled secondary antibody.

1.5 Aim of the thesis

Infection with *H. pylori* is the main cause for GC, particularly NCGC. In order to enhance effective primary and secondary prevention, a stratification of infected individuals into risk groups has shown promise. This could be achieved by detecting antibodies against specific *H. pylori* antigens which are associated with GC, e.g. the virulence factor CagA. However, in countries like China, anti-CagA antibodies are barely useful to stratify risk, as the baseline prevalence is overall high. Here, alternative targets are highly desired.

The first aim of this project was to generate a platform which enabled an unbiased *de novo* discovery of novel GC-associated antibodies against specific *H. pylori* antigens. For this purpose, I adapted a recently developed technique, in order to generate antigen microarray representing *H. pylori* 26695. For a pilot experiment, sera from NCGC cases and controls, derived from the MCC Spain study, were used.
In a next step, I included antigens from further *H. pylori* strains that were isolated in other countries, in order to address the high genetic variety of the bacterium and increase the chance to identify informative antibodies. By probing these *H. pylori* multi-strain microarrays with sera from Chinese NCGC patients and healthy controls, I aimed to identify potential candidate antigens associated with NCGC in a Chinese population.
In a last phase, I transferred these selected *H. pylori* antigens to multiplex serology, a high-throughput platform. Measuring antibody responses in a larger number of Chinese NCGC cases and healthy controls, allowed me to determine the association between NCGC and seropositivity to the selected *H. pylori* antigens with a higher statistical power. Finally, I compared new *H. pylori* antigens with other approaches for risk stratification and hypothesized how they could support the prevention of GC.

2 Material & Methods

2.1 Material

2.1.1 Chemicals

Table 2: List of chemicals

Product	Manufacturer
Acetic acid	Merck (Darmstadt)
Acrylamide/Bisacrylamide-solutions	Carl Roth (Karlsruhe)
Ampicillin	Roche (Basel, CHE)
Ammoniumperoxodisulfate (APS)	Carl Roth (Karlsruhe)
β-mercaptoethanol	Merck (Darmstadt)
Bacto agar	DIFCO Becton Dickinson (Sparks, MD, USA)
Bacto tryptone	DIFCO Becton Dickinson (Sparks, MD, USA)
Betaine monohydrate	Merck (Darmstadt)
Biozym LE Agarose	Biozym (Hessisch Oldendorf)
Bromphenol blue	Merck (Darmstadt)
Calcium chloride ($CaCl_2$)	Merck (Darmstadt)
Casein	Merck (Darmstadt)
CBS-K (superchemiblock)	Chemicon (Temecula, CA, USA)
Di-sodiumhydrogenphosphate (Na_2HPO_4)	Merck (Darmstadt)
DL-dithiothreitol (DTT)	Thermo Fisher Scientific (Karlsruhe)
Ethanol, absolute	Merck (Darmstadt)
Ethylendiamintetraacetat (EDTA)	GIBCO, Invitrogen (Karlsruhe)
Glacial acetic acid	Merck (Darmstadt)
Glutathione	Merck (Darmstadt)
Glutathione-casein	*In-house* produced[92]
Glycerol (100%)	Carl Roth (Karlsruhe)
Glycine	Gerbu Biotechnik (Heidelberg)
Isopropanol	Merck (Darmstadt)
Isopropyl β-D-1-thiogalactopyranoside (IPTG)	Carl Roth (Karlsruhe)
Methanol	Carl Roth (Karlsruhe)
Milk powder	Carl Roth (Karlsruhe)
Nuclease-free water	Life Technologies (Carlsbad, CA, USA)
Phosphate buffered saline (PBS)	Pan Biotech (Aidenbach)
Polyvinylalcohol (PVA)	Merck (Darmstadt)
Polyvinylpyrrolidon (PVP)	Merck (Darmstadt)
Potassium chloride (KCl)	Merck (Darmstadt)
Potassiumdihydrogenphosphate (KH_2PO_4)	Merck (Darmstadt)
Sekusept aktiv	Ecolab (Monheim am Rhein)
Sodium chloride (NaCl)	Merck (Darmstadt)
Sodium-dodecyl-sulfate (SDS)	Gerbu (Gaiberg)
Sodium hydrogen carbonate ($NaHCO_3$)	Merck (Darmstadt)
Sodium-azide (NaN_3)	Merck (Darmstadt)
Sodium-carbonate (Na_2CO_3)	Carl Roth (Karlsruhe)

Streptavidin-R-Phycoerythrin (Strep-PE)	Moss Inc. (Pasadena, Maryland, USA)
Sucrose	Merck (Darmstadt)
Sulfuric acid (H2SO4), 95-97 %	AppliChem (Darmstadt)
Tetramethylbencidine (TMB)	Sigma-Aldrich (Taufkirchen)
Tris(hydroxymethyl)-aminoethan (Tris)	Sigma-Aldrich (Steinheim)
Tween-20	Merck (Darmstadt)
Yeast extract	Merck (Darmstadt)

2.1.2 Equipment

Table 3: List of equipment

Product	**Manufacturer**
Agarose gel electrophorese system	Renner GmbH (Darmstadt)
Benchtop centrifuge (5415D)	Eppendorf (Hamburg)
Centrifuge Sorvall RC6+	Thermo Fisher Scientific (Karlsruhe)
Centrifuge Sorvall Lynx 6000	Thermo Fisher Scientific (Karlsruhe)
ChemiDoc MP Imager	BioRad (Munich)
GelDoc EZ Imager	BioRad (Munich)
Gene Pulser + Pulse Controller	BioRad (Munich)
Heating Block	Eppendorf (Hamburg)
Luminex 200 analyzer	Luminex Corp. (Austin, TX, USA)
Luminex SD sheath fluid delivery system	Luminex Corp. (Austin, TX, USA)
Luminex XYP plate handler	Luminex Corp. (Austin, TX, USA)
Microarray hybridization cassettes	Arrayit Corp (Sunnyvale, CA, USA)
Microwave oven	AEG (Nürnberg)
Mini Trans-Blot Electrophoretic Transfer Cell	BioRad (Munich)
Multiskan™ microplate spectrophotometer	Thermo Fisher Scientific (Karlsruhe)
Nalgene™ PPCO Centrifuge Bottles	Thermo Fisher Scientific (Karlsruhe)
Nanodrop Spectrophotometer MD-1000	peqLab, VWR (Erlangen)
Nano-Plotter NP2.1	GeSIM (Radeberg)
Orbital Shaker	Edmund Bühler GmbH (Bodelshausen)
pH/mV benchtop meter	InoLab, VWR (Erlangen)
Power Pac 300	BioRad (Munich)
Power Scanner	Tecan Trading AG (Männedorf, CHE)
Pressure homogenizer EmulsiFlex-C5	Avestin (Mannheim)
Rotor Sorvall SA-600	Thermo Fisher Scientific (Karlsruhe)
Rotor ThermoScientific F12-6x500 LEX	Thermo Fisher Scientific (Karlsruhe)
Single chamber frames	Grace Bio-Labs (Bend, OR, USA)
Slide staining and storage systems	Merck (Darmstadt)
Thermocycler	BioER Technology (Hangzhou, CHN)
TKA MilliQwater supply	Merck (Darmstadt)
Ultrasonic bath (Sonorex)	Bandelin (Berlin)
Vacuum manifold	Merck (Darmstadt)
Vacuum pump (Millivac™)	Merck (Darmstadt)
Ventilated Incubator	Benchmark Scientific (Sayreville, NJ, USA)
Vortex Mixer	Neolab (Heidelberg)
QIAxcel Advanced system	Qiagen (Hilden)

2.1.3 Consumables

Table 4: List of consumables

Product	Manufacturer
384-well Whatman® Uniplate microplates	GE Healthcare (Freiburg)
96-well polysorb plates	Nunc (Wiesbaden)
96-well filter plates	Merck (Darmstadt)
96-well thin wall PCR plate	peqLab, VWR (Erlangen)
96-well polystyrene flat-bottom plates	Greiner bio-one (Frickenhausen)
96-well polystyrene V-bottom plates	Greiner bio-one (Frickenhausen)
Conical bottom falcon tubes (15 ml/50 ml)	Greiner bio-one (Frickenhausen)
Electroporation cuvettes (2 mm)	peqLab, VWR (Erlangen)
Epoxysilane coated slides	Schott (Mainz)
Nalgene™ cryogenic vials (1.5 ml)	Thermo Fisher Scientific (Karlsruhe)
Nitrocellulose membrane PROTRAN	Schleicher & Schuell (Dassel)
Parafilm™	LaboModerne (Gennevilliers, FRA)
PCR plate sealing foil	Steinbrenner (Wiesenbach)
PCR tubes, 8-strip (0.2ml)	Life Technologies (Carlsbad, CA, USA)
Reagent reservoir (50 ml)	Corning (Kaiserslautern)
SeroMAP™ Microspheres (Fluorescent polysterene beads)	Luminex Corp. (Austin, TX, USA)
Chromatography paper (3mm)	Whatman (Maidstone, UK)

2.1.4 Antibodies

Table 5: List of antibodies

Antibody	Manufacturer
Goat anti-human IgA/IgG/IgM Alexa Fluor 647-conjugated	Jackson Immuno Research (Ely, UK)
Goat anti-human IgA/IgG/IgM biotin-conjugated	Dianova (Hamburg)
Goat anti-mouse IgG HRP-conjugated	Dianova (Hamburg)
Goat anti-rabbit IgG HRP-conjugated	Dianova (Hamburg)
Mouse anti-6XHis [AD1.1.10] DyLight 650-conjugated	Abcam (Cambridge, UK)
Mouse anti-V5-Cy3	Merck (Darmstadt)
Mouse anti-TAG (KT3 hybridoma cell supernatant)	*In-house* produced[93]
Rabbit anti-GST	Merck (Darmstadt)

2.1.5 Kits, ready-to-use reagents and enzymes

Table 6: List of kits and enzymes

Product	Manufacturer
S30 T7 High-Yield Protein Expression Kit	Promega (Walldorf)
Bradford reagent (Roti-Quant)	Carl Roth (Karlsruhe)
Clarity Western ECL Substrate	BioRad (Munich)
CutSmart Buffer	New England Biolabs (Frankfurt a. M.)
DNA loading dye (6x)	New England Biolabs (Frankfurt a. M.)

dNTP Set, PCR Grade	Qiagen (Hilden)
ECL™ Western blotting Detection reagents	GE Healthcare (Freiburg)
GelCode™ Blue Stain Reagent (Coomassie)	Thermo Fisher Scientific (Karlsruhe)
peqGreen	peqLab, VWR (Erlangen)
Prestained protein ladder (broad range)	New England Biolabs (Frankfurt a. M.)
Protease inhibitor complete	Roche (Basel, CHE)
Q5 High-Fidelity DNA Polymerase	New England Biolabs (Frankfurt a. M.)
QIAgen Plasmid Midi Kit (25)	Qiagen (Hilden)
QIAquick Gel Extraction Kit	Qiagen (Hilden)
QIAquick PCR Purification Kit	Qiagen (Hilden)
QIAprep Spin Miniprep Kit	Qiagen (Hilden)
QIAxcel screening cartridge	Qiagen (Hilden)
Smart Ladder (DNA Size Marker)	Eurogentec (Seraing, BEL)
T4 DNA Ligase	New England Biolabs (Frankfurt a. M.)
Taq DNA Polymerase (1000 U)	Qiagen (Hilden)
Quick CIP	New England Biolabs (Frankfurt a. M.)
xMAP™ Sheath fluid	Luminex Corp. (Austin, TX, USA)

Restriction enzymes (BamHI-HF, SalI-HF, AvaI, EcoRV-HF, BsmI, PstI-HF, EcoNI, AlwNI, BtgI-HF, BssHII, StuI, MluI-HF, BstEII-HF, BsaBI, PsI) were purchased from New England Biolabs (Frankfurt a. M.)

2.1.6 Bacterial strains

Table 7: List of bacterial strains

Organism	**Provider**
Escherichia coli BL21	Thermo Fisher Scientific (Karlsruhe)
Helicobacter pylori 26695	DSMZ (Braunschweig)
Helicobacter pylori MAL-007	*Helicobacter Pylori* Genome Project
Helicobacter pylori LAT-022	*Helicobacter Pylori* Genome Project
Helicobacter pylori KOR-037	*Helicobacter Pylori* Genome Project
Helicobacter pylori CHI-112	*Helicobacter Pylori* Genome Project

2.1.7 Buffers and media

Table 8: List and composition of buffers and media

Name	**Ingredients**
LB medium	10 g Bacto tryptone 5 g NaCl 5 g yeast extract ddH_2O to 1,000 ml pH to 7.4
LB_{Amp} medium	10 µl/ml ampicillin (stock: 1 mg/ml) LB medium to 1,000 ml

LB_{Amp} agar	1.5% (w/v) Bacto agar 2.5 µl/ml ampicillin (stock: 1 mg/ml) LB medium to 250 ml
PBS (10x)	500 g PBS ddH_20 to 5,000 ml
PBS-T	100 ml 10x PBS 2.5 ml 20% (w/v) Tween-20 ddH_2O to 1,000 ml
ELISA blocking buffer	2 mg/ml casein in PBS-T
ELISA coating buffer	2 mg/ml glutathione-casein 50 mM Na_2CO_3 50 mM $NaHCO_3$ pH 9.6
ELISA substrate buffer	100 mM NaAc 100 µg/ml TMB 0.015% (w/v) H_2O_2 pH 6.0
Multiplex serology blocking buffer	1 mg/ml casein in PBS
Multiplex serology pre-incubation buffer	2 mg/ml GST-TAG lysate 0.5% (w/v) PVA, 0.8% (w/v) PVP 2.5% (w/v) CBS-K in multiplex serology blocking buffer
Multiplex serology storage buffer	0.05% (w/v) NaN_3 in multiplex serology blocking buffer
SDS-PAGE sample buffer (4x)	2 ml Tris (1M), pH 6.8 4 ml glycerol 0.8 g SDS 2 ml β-mercaptoethanol 2 ml ddH_2O bromophenol blue
SDS-PAGE running buffer (10x)	36.3 g Tris 144 g glycine 100 ml SDS (10%) ddH_2O to 1,000 ml
TAE buffer (50x)	242 g Tris 57.1 ml glacial acetic acid 100 ml 0.5 M EDTA ddH_2O to 1,000 m

2.1.8 Software and websites

Table 9: List of software and websites

Software / Tool	**Accession / Developer**
Biopython	https://biopython.org/
BlastP	https://blast.ncbi.nlm.nih.gov/Blast.cgi
Codon Optimization Tool	https://eu.idtdna.com/CodonOpt
DNA Oligo Analyzer	https://eu.idtdna.com/calc/analyzer
Docker	https://www.docker.com/
EMBOSS Backtranseq	https://www.ebi.ac.uk/Tools/st/emboss_backtranseq/
Genepix Pro 6.0	Molecular Devices (Sunnyvale, CA, USA)
GraphPad Prism 9.0	GraphPad Software (La Jolla, CA, USA)
Jupyter lab (python)	https://jupyter.org/
Luminex 100 IS 2.2 SP1 Software	Luminex Corp. (Austin, TX, USA)
Molecular Evolutionary Genetics Analysis (MEGA) 10.0.5	https://www.megasoftware.net/
Microsoft Windows 10	Microsoft Corp. (Unterschleißheim)
Microsoft Office 2016	Microsoft Corp. (Unterschleißheim)
MMseqs2	https://github.com/soedinglab/MMseqs2
NPC16 (Spotter)	GeSiM (Radeberg)
OpenStack	https://www.openstack.org/
Prokaryotic Genome Annotation Pipeline	https://github.com/ncbi/pgap
PowerScanner V1.2	Tecan Trading AG (Männedorf, CHE)
PubMed	https://www.ncbi.nlm.nih.gov/pubmed
RefSeq	https://www.ncbi.nlm.nih.gov/refseq/
R Studio (R version 3.3.2)	https://www.rstudio.com/
UniProt	https://www.uniprot.org

2.1.9 Primer

Table 10: List of primers

Name	**Sequence (5'→3')**
forward expression primer	GAAATTAATACGACTCACTATAGGGAGACCACAACGGTTTCCC TCTAGAAATAATTTTGTTTAAGAAGGAGATATACATATGCATCA TCATCATCATCATATGCACCAAACCCAA
reverse expression primer	CTGGAATTCGCCCTTTTATTACGTAGAATCGAGACCGAGGAG AGGGTTAGGGATAGGCTTACCCGCACTGGCATCATC
pGEX-T7	TAATACGACTCACTATAGGGtccaaaatcggatctggttccgcgtgga
pGEX-T3	AATTAACCCTCACTAAAGGGgatgcggccgcttatgtttcaggttcaggg

Primer syntheses were ordered from biomer.net (Ulm), except for the gene-specific primer pairs for *H. pylori* MAL-007, LAT-022, KOR-037 and CHI-112 which were ordered from Merck (Darmstadt). Primer sequences for all *H. pylori* antigens are listed in the Supplementary Table S1.

2.2 Antigen microarrays

2.2.1 Annotation of genomes

In order to identify protein-coding sequences in raw genomic DNA, the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) was used.[94] It was installed as a stand-alone software package on a virtual machine (OpenStack) in a docker container as described by the developers (https://github.com/ncbi/pgap/wiki/). The virtual machine was operated with the Linux distribution CentOS7.7, had four virtual centralized processing unit (VCPUs), 20GB Disk and 32 random access memory (RAM).
Genomic DNA (fasta format) and meta data (yaml format including author information, topology of the DNA and organism information in yaml format) were provided as input for the software. The program was executed in the respective docker container running pgap.py. A GenBank flat file was created as output which contained protein-coding sequences and a functional annotation.

2.2.2 Clustering proteins

Proteins were clustered by their amino acid identity using the GPL-licensed software MMSeqs2.[95] It was installed on the same virtual machine as described in 2.2.1 and executed according to the developers' instructions.[95]
To provide an input file, all amino acid sequences were merged into a single fasta file. The minimal sequence identity of proteins to be assigned into the same cluster was set to 0.7. The output file contained cluster accessions followed by amino acid sequences of the respective members, also in fasta format. It was subsequently processed for primer design with biophyto.[96]

2.2.3 Design of gene-specific primers

DNA expression constructs enabling cell-free protein expression on the microarrays, required two successive PCRs. For the first, gene-specific primers linked to a common overhang sequence were used. Primers for *H. pylori* 26695 genes were available from a previous project and designed by the DKFZ bioinformatics core facility Heidelberg Unix Sequence Analysis Resources (HUSAR) as described by Hufnagel *et al*.[91]
Primers for accessory *H. pylori* strains (MAL-007, LAT-022, KOR-037 and CHI-112) were design with biopython.[96] Melting temperatures were calculated with the module SeqUtils.MeltingTemp using nearest neighbor thermodynamics.

Gene sequences were obtained from GenBank files (2.2.1), sense and antisense primers were designed complementary to the template strands. Sense primers started at the first base, while antisense primers started at the fourth base to omit the stop codon and thereby enable generating fusion proteins later on. Each primer had a minimal length of 15 bp and was extended until a melting temperature between 45°C and 48°C was reached. For 64 out of 892 primers (446 forward and 446 backward), this was not achievable. Hence, their starting position was moved by one or two triplets. Primer overhangs (forward: ATGCACCAAACCCAA; reverse: CGCACTGGCATCATC) were added to the final primer sequences before ordering their synthesis from Merck (Darmstadt).

2.2.4 Generation of DNA expression constructs by PCR

For *on-chip* expression of *H. pylori* antigens on the microarray, DNA expression constructs were generated. An expression construct encoded the gene of interest flanked by N-terminal and C-terminal sequences required for transcription and translation (Figure 6). Accordingly, an expression construct for the EBV viral capsid antigen p18 (VCA p18) was included, which served as a positive control.
The first PCR required gene-specific primers (2.2.3), genomic *H. pylori* DNA and used a Q5 polymerase (Table 11). The PCR product served as a DNA template for the second PCR that used common expression primers and a Taq polymerase.

Table 11: PCR schemes to generate gene expression constructs

1st PCR: Q5 (25 µl)		**2nd PCR: Taq (50 µl)**	
Reaction buffer	5 µl	Reaction buffer	5 µl
High GC enhancer	5 µl	Q solution	10 µl
dNTPs [10 µM]	2 µl	$MgCl_2$ [25mM]	2 µl
Q5 Polymerase	0.25 µl	dNTPs [10 µM]	2 µl
DNA template [2 ng/µl]	1 µl	Expression primer mix	2 µl
Gene-specific primer mix [10µM]	1 µl	Taq polymerase	0.25 µl
ddH_20	10.75 µl	ddH_2O	26.75 µl
		DNA template (1st PCR)	2 µl

Figure 6: Schematic representation of a DNA expression construct
Gene of interest are flanked by elements for in vitro *transcription and translation, as well sequences for terminal tags (6xHis and V5). Adapted from Hufnagel* et al.[91]

Table 12: Thermocycling conditions to generate gene expression constructs

	1st PCR: Q5			2nd PCR: Taq		
Initial denaturation	98°C	2 min		95°C	5 min	
Denaturation	98°C	20 s		95°C	30 s	
Annealing	49°C	20 s	45x	49°C	30 s	35x
Elongation	72°C	1 min		72°C	2 min	
		2 min			4 min	
Final elongation	72°C	10 min		72°C	10 min	
Cooling	10°C			10°C		

PCRs were executed in 96-well plates using a PCR thermocycler (Table 12). An elongation time of 1 min for the 1st PCR and 2 min for the 2nd PCR was chosen for PCR products <2,000 bp, while 2min and 4 min were used for PCR products >2,000 bp, respectively. Subsequently, PCR products were analyzed by gel electrophoresis with the QIAxcel advanced and transferred to 384-well plates without further purification. To prevent evaporation, 5µl betaine (5M) were added to each well.

2.2.5 Spotting of protein microarrays

Microarrays were generated with a Nano-Plotter NP2.1 from GeSIM, a liquid handling device using piezoelectric nano liter pipettes to precisely spot small volumes of liquid onto glass slides. The exact positions of the slides in the spotter, as well as their final layout (number of rows, number of columns, distances, etc.) were defined in a work plate file, which also included the position and depth of the 384-well plates, which contained DNA expression constructs. For each type of microarray, the work plate was adjusted and a transfer file was created according to the manufacturer's instructions. This text file allocated positions on 384-well plates to the corresponding location on the microarrays. It further specified the number of droplets per spot. Up to 33 microarrays were generated at a time, each containing up to 2,001 spots (23 columns, 87 rows).

Before every spotting procedure, the tubing and pipettes of the Nano-Plotter were rinsed for at least 20 min with ddH_2O. Then, epoxysilane coated glass slides were positioned on the cooled slide tray and the exact height of each slide was measured using an integrated Z-sensor. A plate with DNA expression constructs was inserted and the spotting procedure was started by loading the respective transfer file. Eight piezoelectric pipettes aspirated the necessary sample volume (including 1.5 µl dead volume) and spotted two droplets (approximately 0.6 nl) onto the assigned position of each microarray. In between, a stroboscope check was performed to ensure that each pipette tip contained sample. If the aspiration failed, an error was recorded for the

respective DNA expression construct and its spotting was repeat once. Washing steps using 8 µl ddH_20 were included after each spotting step to prevent cross-contamination.

After all DNA expression constructs, including positive and negative controls, were spotted, the S30 T7 High-Yield Protein Expression Kit was prepared by mixing 60 µl S30 Premix Plus, 54 µl T7 S30 Extract and 36 µl nuclease-free H_2O. Eight droplets (approximately 2.4 nl) of this cell-free expression kit were spotted directly on top of the DNA spots. As soon as a microarray was completed, it was transferred to a microarray hybridization cassette, a humidity chamber containing 50 µ of nuclease-free H_2O. Each cassettes was incubated at 37°C for one hour followed by an overnight incubation at 30°C. For long time storage, they were transferred to -20°C.

2.2.6 Determination of *on-chip* expression by staining terminal tags

In order to verify successful on-chip protein expression, N-terminal 6xHis-tags and C-terminal V5 tags were stained. For this purpose, microarrays were transferred from storage at -20°C into 10 ml 5% blocking milk and incubated for 1h at room temperature (RT) on a shaker. Afterwards they were washed 3 times in PBS-T, for 5 min per washing step.

Meanwhile, anti-tag antibodies (anti-6xHis-tag and anti-V5-tag) were diluted 1:1,000 in 5% blocking milk (1 ml per microarray) in a lightproof container.

Subsequently, single chamber incubation frames were mounted on each microarray slide and filled with 1 ml of the antibody solution. Incubation was conducted on a shaker in a lightproof container for 1h. Finally, microarrays were washed three times in PBS-T and rinsed in ddH_20. After air-drying in a lightproof incubator, emitted signals of the N-terminal (532 nm) and C-terminal (635 nm) fluorophores were recorded with a Tecan Power Scanner and saved as tiff files. These scans were subsequently analyzed with the software GenePix Pro 6.0 to obtain raw median fluorescence intensities (MFI) for each spot.

On-chip protein expression was assessed by determining the rate of full-length expressed and partially expressed proteins. For this purpose, two cutoffs were calculated per microarray: one for the N-terminal signal at 532 nm and another for the C-terminal signal at 635 nm. Cutoffs were determined using the median + 5 median absolute deviations (MADs) of the respective negative control signals. Each protein with a C-terminal signal above the respective cutoff was considered full-length expressed, while proteins with only the N-terminal signal above the respective cutoff was considered partially expressed.

2.2.7 Detecting serum antibodies with an immunoassay

Microarrays were transferred from storage at -20°C into 10 ml 5% blocking milk (milk powder in PBS-T) and incubated for 1h at RT on a shaker. Afterwards they were washed 3 times à 5 min in PBS-T. Meanwhile, patient sera were diluted 1:33 in 5% blocking milk, which contained 1 mg/ml *E. coli* BL21 lysate to adsorb antibodies against the bacterial expression kit. Solutions were kept on ice and subsequently incubated for 1h on a shaker.

Single chamber incubation frames were mounted onto the microarrays and the pre-incubated serum was added. After another incubation for 1h at RT on a shaker, the slides were washed three times and 1 ml anti-human antibody (1:133 in 5% blocking milk), which was conjugated to fluorescent dye was added. Microarray were again incubated for another hour at RT in a lightproof container prior to three final washing steps and rinsing in ddH_20.

Signals were generated as described in 2.2.6 at a wavelength of 635 nm.

2.3 Multiplex serology

2.3.1 Preparation of electrocompetent *E. coli* BL21

Five ml LB medium were inoculated with *E. coli* BL21 and incubated overnight at 37°C on a shaker. The next day, the culture was transferred to 250 ml LB medium and incubated further. Once an OD_{500} of 0.6 was reached, the cells were cooled down on ice, pelleted (6,000g, 7min, 4°C) and resuspended in 125 ml precooled 5% glycerol (w/v). This cell suspension was centrifuged again and resuspended in 62.5 ml precooled 5% glycerol (w/v). After a last centrifugation step, the pellet was resuspended in 5 ml precooled 10% glycerol (w/v), aliquoted à 40 µl and quick-frozen in liquid nitrogen. Aliquots were stored at -80°C until usage.

2.3.2 Generation and expression of GST-X-TAG fusion proteins

Antigens for multiplex serology were fused to an N-terminal GST and a C-terminal TAG. Both sequences were encoded on the expression plasmid pGEX-4T3tag. This vector further contained a beta-lactamase for ampicillin resistance, as well as a *lac* operon to enable inducible protein expression with IPTG.
Sequences of the respective antigens were codon optimized for expression in *E. coli* BL21 using EMBOSS Backtranseq.[97] They were ordered as gene syntheses from Eurofins Genomics (Ebersberg), flanked by a BamHI restriction site at the 5' end and a SalI restriction site at the 3' end. Cloning of the synthesized sequences into the pGEX-4T3tag vector was also conducted by Eurofins Genomics (Ebersberg). In case that the cloning process was not successful, DNA sequences were isolated from a standard vector using the BamHI and SalI restriction sites and cloned manually into the expression vector. Restriction digests and respective ligation was performed according to protocols of New England Biolabs (Frankfurt a. M.).

Expression constructs were obtained as lyophilisates and dissolved in ddH_2O to a concentration of 20 ng/µl, of which 1 µl was added to a freshly thawed aliquot of electrocompetent *E. coli* BL21 cells. The mixture was transferred to a pre-cooled electroporation cuvette and electroporated using a Gene Pulser with Pulse Controller with the following settings: voltage: 2.3 kV, resistance: 200 Ω, capacitances: 960 µF (Controller), 25 µF (Gene Pulser).
Then, cells were immediately transferred into 1 ml pre-warmed LB medium and incubated for 1 h at 37°C on a shaker before plating 10 µl on LB_{amp} agar plates. After

overnight incubation at 37°C, one colony was picked and transferred into 5 ml LB_{amp} medium. The culture was grown for 6h at 37°C on a shaker, transferred into 250 ml of fresh LB_{amp} and cultivated further overnight.
On the next day, 50 ml of the transformed *E. coli* culture were separated for plasmid purification (2.3.4), while further 700 µl were mixed with 50% glycerol (w/v) and stored at -80°C to generate a glycerol stock for long-time preservation. The remaining culture was added to 800 ml of fresh LB medium and induced with 500 µl IPTG (stock concentration: 0.5 M), as soon as the OD_{500} reached 0.5-0.6. After agitating the culture for 6h at RT, cells were pelleted (5,000 g, 6 min, 4°C), resuspended in 10 ml PBS and stored at -20°C or directly processed further.
A protease inhibitor tablet was dissolved in 1 ml ddH_2O, of which 0.5 ml were added to the (thawed) protein lysate. Furthermore, 20 µl DTT (stock solution: 1 M) were added per 10 ml cell suspension. Mechanic protein lysis was executed by applying 1,500 - 2,000 bar for 2 min using a pressure homogenizer ('French press'). To remove cell debris, lysates were subsequently centrifuged at 30,000 g for 1h at 4°C. An aliquot of 100 µl was taken before and after the centrifugation step for sodium dodecylsulfate polyacrylamide gel electrophoresis (SDS-PAGE) and Western Blot analysis (2.3.7). Cleared lysates were mixed 1:1 with 100% glycerol and stored at -20°C.

2.3.3 Determination of total protein concentration

To measure protein concentration of cell lysates, one µl lysate was mixed with 800 µl ddH_20 and 200 µl Bradford reagent. After incubating for 5 min, absorption at OD_{595} was determined. A blank control without lysate served as reference. In case the absorption exceed 1.0, the measurement was repeated using a 1:10 dilution. Protein concentrations were calculated applying the following formula, based on a calibration curve with bovine serum albumin: $c_{protein\ lysate}\,[\mu g/\mu l] = OD_{595} * 44$

2.3.4 Analytical DNA digest

Plasmid DNA of 50 ml transformed *E. coli* culture was isolated using the QIAgen plasmid midi kit according to the manufacturer's instructions. For the final elution, 50 µl ddH_2O were used and the DNA concentration was determined with a NanoDrop ND-1000 UV-Vis spectrophotometer. The integrity of each plasmid was determined by three analytical DNA digests: a linear digest to verify the length of the plasmid (L), a symmetric digest to verify the length of the insert (S) and an asymmetric digest to verify the pre-calculated length of two fragments (A).

Each restriction digest was prepared according to the manufacturer's protocol: 5 µl CutSmart Buffer, 1 µg DNA and 1 µl per restriction enzyme were filled up with ddH_2O to a total volume of 50 µl. The respective enzymes are listed in Table 13. Digests were incubated at 37°C for 2h and subsequently analyzed by agarose gel electrophoresis (2.3.6).

Table 13: Analytical DNA digests

Antigen	Digest*	Restriction enzyme	Fragment length [bp]*
HP0003	L	BamHI-HF	5079, 737
	S	BamHI-HF, SalI-HF	4982, 737, 93
	A	AvaI, EcoRV-HF	3841, 1975
HopA	L	BamHI-HF	6431
	S	BamHI-HF, SalI-HF	4982, 1449
	A	PstI-HF	4389, 2042
HP0017N	L	BamHI-HF	6113
	S	BamHI-HF, SalI-HF	4982, 1131
	A	EcoRV-HF	3856, 2257
HP0185	L	SalI-HF	5789
	S	BamHI-HF, SalI-HF	4982, 559, 248
	A	StuI, EcoRV-HF	3614, 2175
HP0385	L	BamHI-HF	5216
	S	BamHI-HF, SalI-HF	4982, 234
	A	BsmI, EcoRV-HF	3242, 1974
HP0477	L	BamHI-HF	6089
	S	BamHI-HF, SalI-HF	4982, 1107
	A	PstI-HF	4296, 1793
HP0527trunc	L	BamHI-HF	5258
	S	BamHI-HF, SalI-HF	4982, 275
	A	PstI-HF	4119, 1139
HP0545	L	BamHI-HF	5609
	S	BamHI-HF, SalI-HF	4982, 627
	A	AvaI, EcoRV-HF	3682, 1927
HP0582	L	BamHI-HF	5960
	S	BamHI-HF, SalI-HF	4982, 978
	A	PstI-HF	4005, 1955
HP0659	L	BamHI-HF	6230
	S	BamHI-HF, SalI-HF	4982, 1248
	A	BstEII-HF	3745, 2485
HP1038	L	BamHI-HF	5486
	S	BamHI-HF, SalI-HF	4982, 504
	A	AvaI, EcoRV-HF	3475, 2001
HP1064	L	BamHI-HF	5267
	S	BamHI-HF, SalI-HF	4982, 285
	A	PstI-HF	4017, 1250

HP1091	L	BamHI-HF	6266
	S	BamHI-HF, SalI-HF	4982, 1284
	A	EcoNI	5232, 1034
HP1355	L	BamHI-HF	5807
	S	BamHI-HF, SalI-HF	4982, 825
	A	BsaBI, EcoRV-HF	3542, 2265
HP1435	L	BamHI-HF	5864
	S	BamHI-HF, SalI-HF	4982, 882
	A	PstI-HF	4554, 1310
HP1570	L	BamHI-HF	5480
	S	BamHI-HF, SalI-HF	4982, 498
	A	AlwNI	3390, 2090
Kor1294N	L	BamHI-HF	6521
	S	BamHI-HF, SalI-HF	4982, 1539
	A	AlwNI	4218, 2303
Lat118	L	BamHI-HF	5175
	S	BamHI-HF, SalI-HF	4982, 192
	A	BtgI-HF	4140, 1034
Lat1540	L	BamHI-HF	5417
	S	BamHI-HF, SalI-HF	4982, 435
	A	PstI-HF	4044, 1373
Lat98	L	BamHI-HF	5732
	S	BamHI-HF, SalI-HF	4982, 750
	A	BssHII	3541, 2191
Mal1434	L	BamHI-HF	5558
	S	BamHI-HF, SalI-HF	4982, 576
	A	MluI-HF	3076, 2482
Mal648	L	BamHI-HF	5264
	S	BamHI-HF, SalI-HF	4982, 282
	A	PsiI, EcoRV-HF	3332, 1932

Enzymes and respective DNA fragment lengths for a linear DNA digest (L), symmetric DNA digest (S), asymmetric DNA digest (A)

2.3.5 Agarose gel electrophoresis

In order to determine the size of DNA fragments, gel electrophoresis was applied. An agarose gels were prepared by dissolving 1.5% agarose (w/v) in heated TAE buffer. PeqGreen was added to the solution (1:10,000) after it cooled down to approximately 60°C. The solidified gel was placed in an electrophoresis chamber and covered with TAE buffer. Each DNA sample was mixed 1:6 with loading dye. Subsequently, 10 µl of each sample were loaded onto the gel alongside 5 µl of DNA size marker to enable size estimation. Electrophoresis took place for 1h applying a voltage of 100V.

Gel electrophoresis of DNA expression constructs for antigen microarrays, was performed in an automated fashion using the QIAxcel advanced according to the manufacturer's instructions.

2.3.6 Anti-TAG ELISA

For a semi-quantitative detection of full-length GST-X-TAG proteins in bacterial lysate, an enzyme-linked immunosorbent assay (ELISA) was performed. First, 96-well polysorb plates were coated with 100 µl ELISA coating buffer per well and incubated overnight at 4°C. The next day, antigen lysates were diluted to a concentration of 2 µg/µl in ELISA blocking buffer to a total volume of 700 µl and divided into two wells of a 96-well microtiter plate. Row-wise, a 1:3 dilution series in ELISA blocking buffer was prepared with buffer only in the last column. Each plate also contained a GST-TAG lysate (without insert) dilution series to serve as an internal control and enable relative quantification.
ELISA coating buffer was discarded from the polysorb plate and each well was filled with 200 µl ELISA blocking buffer. After incubation for 1h at 37°C the buffer was discarded again and 100 µl of each protein lysate dilution were transferred to the polysorb plate. After another incubation for 1h at RT on a shaker, the protein dilutions were discarded and wells were washed three times with 200 µl PBS-T per well. To remove residual liquids, plates were turned over and knocked on paper towels after each washing step. Then, 100 µl mouse anti-TAG antibodies (1:5,000 in ELISA blocking buffer) were added to each well and again incubated for 1h at RT on a shaker, followed by three washing and drying steps. Subsequently, plates were incubated with 100 µl HRP-conjugated goat anti-mouse antibodies (1:10,000 in ELISA blocking buffer) per well for another hour at RT on a shaker. Then, plates were washed three times again and dried thoroughly before adding 100 µl freshly prepared ELISA substrate buffer. After 2-8 min, the reaction was stopped by adding 50 µl 1 M H_2SO_4. The readout was generated by measuring the absorption at 450 nm with a microplate spectrophotometer.

2.3.7 SDS-PAGE

To separate proteins in bacterial lysates by size, SDS-PAGE was performed. For this purpose, acrylamide gels for the Mini-PROTEAN II system were prepared according to the manufacturer's instruction. The ingredients are listed in Table 14. First, the resolving gel was poured into a sealed glass cassette sandwich. After polymerization,

the stacking gel is added on top and a comb is inserted to form sample pockets. Polymerized acrylamide gels were transferred to vertical electrophoresis chambers and covered with SDS running buffer.

Samples from crude and cleared bacterial lysates (2.3.2) were mixed with SDS-PAGE sample buffer and ddH_2O to a final concentration of 1 µg/µl. GST-TAG lysate (without insert) served as positive control and was prepared alongside. Samples were incubated for 5 min at 95°C, before 10 µl of each were loaded onto an acrylamide gel. Alongside, 3 µl pre-stained protein ladder were loaded and a voltage of 200 V was applied for 50 min.

Then, gels were either used to perform Western Blot analysis (2.3.8) or incubated in Coomassie blue solution for approximately 1h to unspecifically dye all proteins. After another incubation over night at 4°C, bands were documented using a GelDoc imager.

Table 14: Reagents for four acrylamide gels

Reagent	Resolving gel (13.5%)	Stacking gel (5%)
ddH_2O	3.2 ml	7.4 ml
30% (w/v) acrylamide	7.5 ml	1.2 ml
1M Tris-HCl (pH 8.8)	9.0 ml	-
1M Tris-HCl (pH 6.8)	-	1.3 ml
10% (w/v) SDS	200.0 µl	100.0 µl
TEMED	10.0 µl	10.0 µl
10% (w/v) APS	100.0 µl	50.0µl

2.3.8 Western Blot

To specifically detect protein tags, Western Blot analysis was performed. For this purpose, protein lysates were separated on an acrylamide gel using SDS-PAGE (2.3.7). Proteins were then blotted to nitrocellulose membranes. For this purpose, the respective acrylamide gel was placed on a nitrocellulose membrane between two Whatman papers and two sponges. The ‘sandwich’ was then soaked in EMBL transfer buffer; 100 V and 350 mA were applied for 1h to transfer the proteins onto the nitrocellulose membrane.

Afterwards, the membrane was incubated in 10% blocking milk for 1h at RT on a shaker to block residual binding sites. Then, the membrane was washed three times in PBS-T for 5 min and a primary antibody, either a rabbit anti-GST (1:10,000 in 5% blocking milk) or a mouse anti-TAG antibody (1:5,000 in 5% blocking milk), was added. After another incubation for 1h at RT on a shaker, the membrane was washed again and incubated with a secondary antibody, either HRP-conjugated goat anti-rabbit antibody (1:10,000 in 5% blocking milk) or HRP-conjugated goat anti-mouse antibody

(1:10,000 in 5% blocking milk). A last washing was performed and the membranes were covered with ECL Western Blotting Detecting reagent according to the manufacturer's instruction. Luminescence signals were visualized with a ChemiDoc imager.

2.3.9 Multiplex serology

Cleared *H. pylori* antigen lysates and a GST-TAG control lysate were diluted with multiplex serology blocking buffer to a protein concentration of 1 mg/ml in a total volume of 1 ml. Glutathione-casein coupled polystyrene beads (2,500 beads per patient sample) of an assigned bead type were added to each antigen. They were prepared by M. Oppenlaender as described.[78] As beads contained fluorophores to enable subsequent identification of the bead type, all incubation steps took place in lightproof containers.
After incubating the beads for 1h at RT on a shaker, they were washed three times. This comprised pelleting the beads by centrifugation (2 min, 17,000g), removing the supernatant and adding 1 ml fresh blocking buffer. Loaded beads were stored in 1 ml multiplex serology storage buffer at 4°C.

The next day, serum or plasma samples were diluted 1:500 in polystyrene flat-bottom plates with multiplex serology pre-incubation buffer to a total volume of 100 µl per sample. Incubation took place for 1h at RT on a shaker and aimed to deplete antibodies which would unspecifically bind to the bead surface or residual bacterial proteins.
Meanwhile, beads were resuspended by alternating sonification and vortexing for approximately 4x30s. If agglomeration was observed, they were pulled multiple times through a syringe to separate the beads. Then, beads were pooled into a single mix and a sufficient volume of blocking buffer was added (total volume: 50 µl per patient sample + extra volume).
Fresh 96-well filter plates were equilibrated with 100 µl ddH_2O per well for approximately 10 min and emptied applying vacuum. Residual water was removed mechanically by knocking with a hammer on the lid of the plate. Then, 50 µl of the bead mix were filled into each well and 50 µl pre-incubated patient samples were added. The final sample dilution was therefore 1:1,000.
After incubation for 1h at RT on a shaker, plates were washed three times by applying vacuum and adding 100 µl fresh blocking buffer per well. Residual buffer was removed mechanically after the last washing. Immediately after, 100 µl biotinylated secondary anti-human antibody (1:1,000 in blocking buffer) were added to each well, followed by

another incubation for 1h at RT on a shaker. In case a well was probed with KT3 antibodies instead of patient sample, a secondary anti-mouse antibody (1:1,000 in blocking buffer) was used instead.
Afterwards, plates were washed again before adding 100 µl Strep PE (1:750 in blocking buffer) to each well. After a last incubation for 30 min at RT on a shaker, plates were washed, filled with 100 µl storage buffer per well and stored at 4°C overnight.

Bound serum antibodies were quantified on the next day with a Luminex 200 analyzer. To retrieve a valid MFI value, at least 100 beads per bead sort needed to be analyzable. Background signals were quantified based on a sample-free well, while patient-specific backgrounds corresponded to the readouts of the GST-TAG control. In order to estimate plate-to-plate variations, each plate contained three control samples. Day-to-day differences were estimated by including the same two plates on each day.

2.4 Study data

2.4.1 Multicase-control study Spain (MCC Spain)

The population-based MCC Spain study was initiated in 2007 to evaluate etiological factors of common tumors (prostate, breast, colorectal, gastro-esophageal and chronic lymphocytic leukaemia) in the Spanish population.[98] GC cases were sampled in 15 hospitals of 9 Spanish provinces (Asturias, Barcelona, Cantabria, Granada, Huelva, León, Madrid, Navarra and Valencia) and were histologically confirmed. Patients were between 20-85 years old, lived for more than 6 months in the respective study area and provided epidemiologically relevant information. Controls were randomly selected from General Practitioners' lists at primary healthcare centers in the proximity of a respective hospitals.

Serum samples from 202 NCGC cases and 2,071 healthy controls were analyzed in a prior study using multiplex serology.[99] For the *H. pylori* microarray experiments, 65 NCGC cases from three regions (Leon, Barcelona, Madrid) were randomly picked, alongside controls matched for sex and age. Fourteen samples were, however, *post hoc* excluded from the statistical evaluation due to an erroneous matching. Baseline characteristics of the samples used for data analysis are summarized in Table 15.

The MCC Spain study was reviewed and approved by the participating institutions. All participants provided written consent to participate. Study sera were kindly provided by Nuria Aragonés.

Table 15: Baseline characteristics of serum samples from the MCC Spain study

	Selected NCGC cases n = 58*	**Matched Controls** n = 58*	**All NCGC cases** n = 202
Sex			
Female	23 (40%)	23 (40%)	76 (38%)
male	35 (60%)	35 (60%)	126 (62%)
Age			
<55	11 (19%)	11 (19%)	36 (18%)
55-64	6 (10%)	6 (10%)	30 (15%)
65-74	17 (29%)	17 (29%)	70 (35%)
≥75	24 (41%)	24 (41%)	66 (33%)
Smoking status			
Never	30 (52%)	27 (47%)	88 (44%)
Former	16 (28%)	22 (38%)	67 (33%)
Current	12 (21%)	8 (14%)	47 (23%)

**Fourteen of the initial 130 sera were excluded due to erroneous matching;*
NCGC = non-cardia gastric cancer; GC = gastric cancer; Modified from Jeske et al.[100]

2.4.2 Shanxi study

This cross-sectional study was conducted between 1997 and 2005 among patients presenting to the Shanxi Cancer Hospital in Taiyuan in the Chinese county Shanxi. It comprised a case-control and a case-only portion. Enrolled cases provided a blood samples and were at least 20 years old and had incident esophageal or gastric cancer (histologically confirmed). For the case-control portion, neighborhood controls were enrolled, matched for sex and age (±5 years). They also provided blood samples.
Serum samples from 214 NCGC cancer cases and 455 controls were available from a previous multiplex serology study (unpublished). Their baseline characteristics is summarized in Table 16. Many of the NCGC cases turned out to be from the case-only portion of the Shanxi study and had no demographic data available.

Table 16: Baseline characteristics of serum samples from the Shanxi study

	All n = 669	**NCGC Cases** n = 214	**Controls** n = 455
Sex			
male	355 (70%)	46 (74%)	309 (69%)
female	155 (30%)	16 (26%)	139 (31%)
missing	159	152	7
Age			
<50	104 (20%)	16 (26%)	88 (20%)
50-60	192 (38%)	23 (37%)	169 (38%)
>60	214 (42%)	23 (37%)	191 (43%)
missing	159		7
mean	57.2	53.9	57.6
Smoking			
Yes	323 (63%)	45 (73%)	278 (62%)
no	187 (37%)	17 (27%)	170 (38%)
missing	159	152	7
Alcohol			
Daily	80 (16%)	12 (19%)	68 (15%)
Weekly	12 (2%)	3 (5%)	92 (2%)
Monthly	159 (31%)	23 (37%)	136 (30%)
Never	259 (51%)	24 (39%)	235 (52%)
missing	159	152	7

NCGC = non-cardia gastric cancer

2.4.3 Linxian General Population Nutrition Intervention Trial (NIT)

The NIT is a population-based cohort initiated in 1985 in the Chinese county Linxian. Details on trial design, objectives and participant characteristics were previously described.[101] Briefly, the cohort comprised 29,584 healthy adults aged 40 – 69. Participants were enrolled for an intervention trial investigating the effects of nutrition supplements on esophageal and GC incidence and mortality in the general population. The intervention period lasted for 5.25 years. Afterwards the study was continued as a population cohort.
Before the start of intervention, each participant provided blood and answered an epidemiological questionnaire. Until the initial end of follow-up (May 31, 2001), 363 NCGC cases were found and histologically confirmed.
In 1999/2000, 80% of all living participants donated blood again and respective plasma samples were stored at -80°C until analyzed. Until December 31, 2006, 118 additional NCGC cases were diagnosed.

For this project, I used serum and plasma samples, which were part of a prior multiplex serology research project on *H. pylori* and GC.[102] It included serum samples from 330 NCGC cases and 330 controls (frequency matched for sex) of the baseline cohort and plasma samples from 118 NCGC cases and 912 controls (age- and sex-stratified subcohort) of the 1999/2000 resurvey. Their characteristics at time of blood drawl is summarized in Table 17. There was no overlap between the study samples from 1985 and 1999/2001.

For the microarray experiments we randomly picked sera from 136 NCGC cases alongside 136 age- and sex-matched controls. After visual examination of the microarray immunoassays, 119 case-control pairs were included into further analyses. For the subsequent multiplex serology measurement, all available serum sample from the baseline and plasma sample from the 1999/2000 collection, were used. MFI readouts were generated for 322 and 113 NCGC cases and 327 and 880 controls, respectively. The remaining samples did not contain enough sample volume.

Table 17: Baseline characteristics of serum and plasma samples from the NIT

	Serum samples			Plasma samples		
	All	NCGC cases	Controls	All	NCGC cases	Controls
	n = 649	n = 322	n = 327	n = 993	n = 113	n = 880
Sex						
Male	435 (67%)	216 (67%)	219 (67%)	490 (49%)	62 (55%)	428 (49%)
Female	214 (33%)	106 (33%)	108 (33%)	503 (51%)	51 (45%)	452 (51%)
Age at blood drawl						
<50	231 (36%)	76 (24%)	155 (47%)	8 (1%)	0 (0%)	8 (1%)
50-60	259 (40%)	152 (47%)	107 (33%)	351 (35%)	35 (31%)	316 (36%)
>60	158 (24%)	94 (29%)	64 (20%)	631 (64%)	78 (69%)	553 (63%)
mean	53.3	55.5	51.1	63.9	64.7	63.8
missing			1			3
Smoking						
Yes	316 (49%)	154 (48%)	162 (50%)	319 (32%)	39 (35%)	280 (32%)
No	332 (51%)	168 (52%)	164 (50%)	671 (68%)	74 (65%)	597 (68%)
missing	1		1	3		3
Alcohol						
yes	177 (27%)	76 (24%)	101 (31%)	258 (26%)	23 (20%)	235 (27%)
no	471 (73%)	246 (76%)	225 (69%)	732 (74%)	90 (80%)	642 (73%)
missing	1		1	3		3
BMI						
mean	21.7	21.5	21.9	21.9	21.3	21.9
Follow-up time						
<2 years		38 (12%)			36 (32%)	
<4 years		83 (26%)			65 (58%)	
>4 years		239 (74%)			48 (42%)	
mean	10.8	7.5	13.9	5.8	3.3	6.1
missing	1					
H. pylori* seroprevalence						
		95 %	84%		97%	94%

**based on previously reported multiplex serology results.*[102]
NCGC = non-cardia gastric cancer; BMI = body mass index

2.5 Statistical analysis

To characterize study populations, categorical variables were described by frequencies and percentages. Continuous values for the variable age were stratified into categories (<50, 50-60, and >60 years) and characterized by means.

Raw MFI readouts from minimized *H. pylori* 26695-microarrays that were probed with patient serum, were normalized to fold-change values by division through a global cutoff. This cutoff was sample-specific and defined as the median + 5 MADs of the 23 negative controls. Antibody responses exhibiting a fold-change value above 1 were considered seropositive. Fold-change values were described group-wise (cases or controls) using means and standard deviations (SDs). They were compared using Mann-Whitney U test.
Raw MFI readouts from *H. pylori* multi-stain microarrays were corrected by subtracting the local background from each individual spot (described in more detail in 3.2.4), before applying a sample-specific cutoff. This cutoff was pre-specified to achieve a 95% specificity, measured by 123 negative controls scattered across each microarray. Each MFI value which exceeded the sample-specific cutoff was considered seropositive.
Difference between NCGC cases and controls were assessed using non-parametric Mann-Whitney U test for continuous data or chi-squared test for dichotomized data.

Multiplex serology readouts were corrected by subtracting the antigen-specific background and the sample-specific GST background. Signals which exceeded the antigen-specific cutoff, which was determined using finite mixture modeling (described in more detail in 3.4.1), were considered seropositive.

Correlations between microarray readouts and multiplex serology results were visualized by scatterplots. They were further analyzed calculating non-parametric rank correlation (Spearman's rho). Furthermore, the performance of *H. pylori* antigens on microarrays were assessed with specificity and sensitivity, using multiplex serology results as a reference.

Seroprevalence in cases and controls were determined by dividing the number of seropositive signals through the overall number of signals. Group-wise differences were analyzed by chi-squared test, or Fisher's exact test in case of small sample sizes. Given the exploratory nature of the study, multiple testing was not accounted for.

Associations between seropositivity to an antigen and NCGC were assessed applying unconditional logistic regression analysis to estimate ORs and 95% CI. Models were adjusted for sex and age as indicated. Including smoking or alcohol as further cofounders did not substantially alter outcomes and where therefore only presented in the supplement.

All analyses were conducted with R 3.3.2. The used packages included ‘glm’ for regression models, ‘epi’ and ‘survival’ for epidemiological analysis, ‘mixsmsn’ and ‘mixtools’ for finite mixture modeling, ‘FactoMineR’ for PCA and ‘randomForest’ to create random forest models. p-values ≤ 0.05 were considered significant.
Figures were either created using the R package ‘ggplot2’ or GraphPad Prism 9.

3 Results

3.1 Identifying NCGC-associated antigens in the MCC Spain study using *H. pylori* 26695-microarrays

In the scope of a Bachelor's and Master's thesis, D. Reininger and B. Turgu started generating microarrays for *H. pylori* under the supervision of Dr. K. Hufnagel based on the common reference strain *H. pylori* 26695. This version of the microarray is referred to as *H. pylori* 26695-microarray. I determined that these microarrays displaying 1,440 antigens reached a protein expression rate of 93% (90% full-length expressed and 3% partially expressed). The design and a representative protein expression staining are visualized in Figure 7A. In order to pre-select immunogenic *H. pylori* proteins, sera from the cross-sectional MCC Spain study were utilized. These samples were characterized by multiplex serology in 2017 using established *H. pylori* antigens.[99] *H. pylori* prevalence among NCGC cases and healthy controls was determined to be 95% and 88%, respectively. This high baseline prevalence was favorable to potentially identify further seroreactive *H. pylori* antigens.

Sera from 65 NCGC cases were randomly selected from 202 available NCGC samples. Alongside, 65 age- and sex-matched healthy controls were picked. Serum pools (n = 26) of either each five cases or five controls were prepared and *H. pylori* 26695-microarrays were probed with these pools as described in the methods. Exemplary immunoassays are visualized in Figure 8. They were analyzed by Dr. Hufnagel to pre-select seroreactive proteins, which exceeded the cutoff (median + 5 MAD of all negative controls), irrespective of case/control status of the serum pool. Subsequently, I supplemented this selection with established *H. pylori* multiplex serology markers to enable a subsequent comparison of microarray results. In total, the selection included 245 *H. pylori* antigens that were subsequently used to generate 'minimized' *H. pylori* 26695-microarrays (Figure 7B, Supplementary Table S2). Three of them, CagA, VacA and HyuA, were split into their N-terminal and a C-terminal parts, in order to enhance protein expression and to resemble the corresponding multiplex serology antigens. I prepared fresh DNA expression constructs for each of these antigens.

Using these DNA expression constructs, I generated one batch (33 slides à 4 microarrays) of minimized *H. pylori* 26695-microarrays and probed them with individual serum samples from the MCC Spain (Figure 9). Case-control pairs were always tested on the same slide.

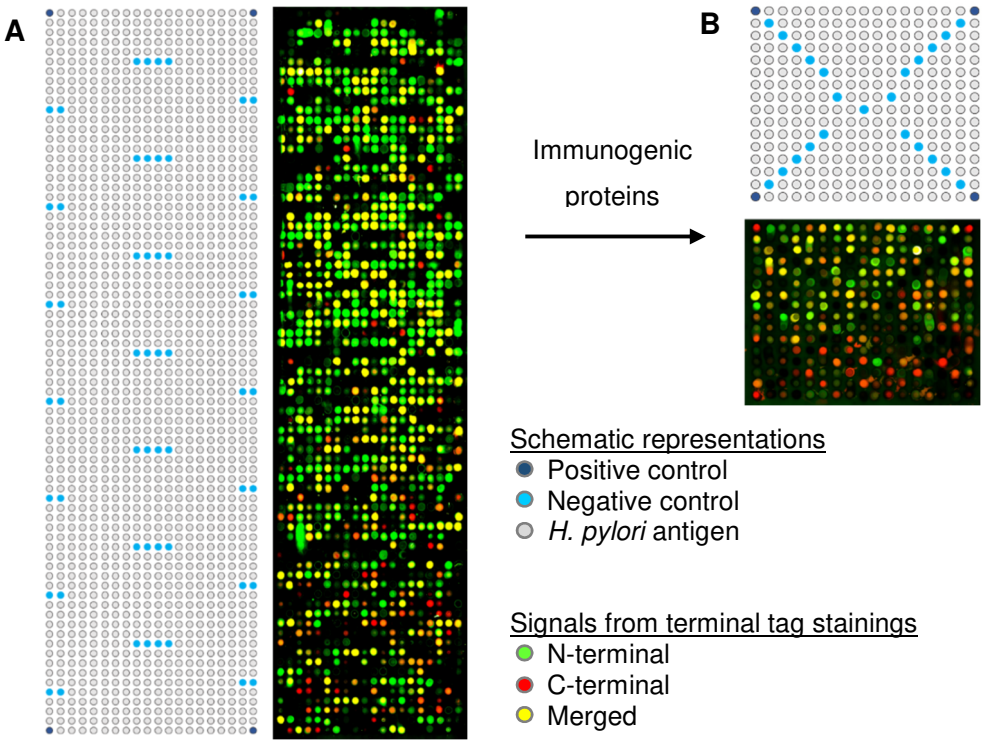


Figure 7:* H. pylori *26695-microarrays and minimized* H. pylori *26695-microarrays
Design and protein expression control staining of an (A) H. pylori *26695-microarray (1,440 proteins) and (B) a minimized* H. pylori *26695-microarray containing 245 pre-selected immunogenic proteins. To optimize visualization, brightness and contrast were adjusted which caused less intense signals to vanish. This did not influence numerical values. Adopted from Jeske* et al.[100]

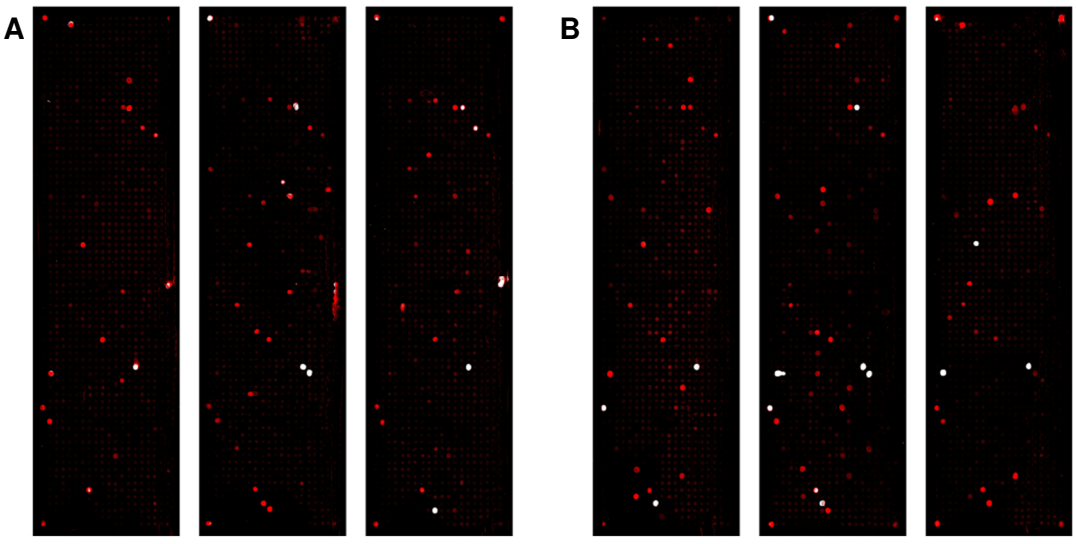


Figure 8: Immunoassays stainings of* H. pylori *26695-microarrays
Exemplary immunoassays of serum pools comprising either five (A) NCGC cases or (B) healthy controls. Bound serum antibodies were visualized with a fluorescent secondary anti-human antibody. Saturated signals appear white. Adapted from Jeske et al.[100]

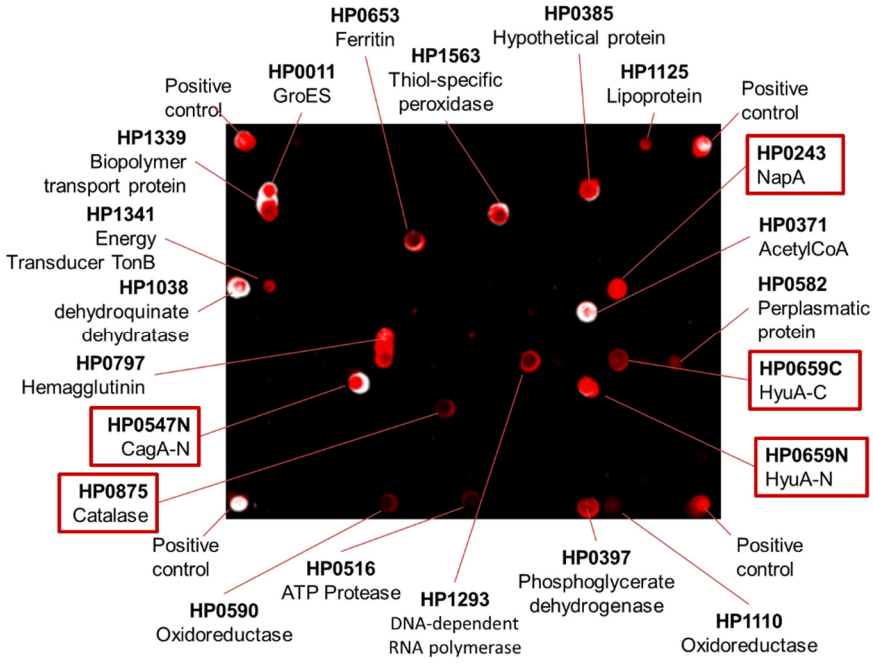


Figure 9: Immunoassays stainings of a minimized* H. pylori *26695-microarrays
Exemplary minimized H. pylori *26695-microarray probed with the serum of a healthy control. Signals were annotated with locus tags and the corresponding proteins. Established* H. pylori *antigens, which are part of the multiplex serology panel, are highlighted by a frame.*

Numeric readouts for each immunoassay were acquired using the software GenePix Pro 6. There was no systematic group-wise difference between NCGC cases and controls, as shown in Table 18. The mean signal measured in negative controls was 4,750 MFI for NCGC cases and 4,276 MFI for healthy controls with SDs of 2,324 MFI (CV: 72%) and 3,269 MFI (CV: 66%), respectively. On average, 17 antigens were found seroreactive in NCGC cases and 20 in controls.
Due to varying patient-specific backgrounds, I normalized MFI signals to fold-change values by dividing raw MFI values through a patient-specific cutoff (median MFI + 5 MAD of all negative controls). The mean fold-change value of seroreactive antigens was 2.68 among NCGC cases and 2.86 among controls.

Eight *H. pylori* antigens showed a significantly different seroprevalence among NCGC cases and controls (Table 19). For two antigens, HP0512 and HP1071, one seroreactive case had to be imputed to calculate an OR. Hence, the specified values for OR and CI can be considered an underestimation. I found that the presence of antibodies against the CagA (HP0547N) and the membrane protein HopA (HP0229) were directly associated with NCGC case status with an OR of 3.64 (95% CI: 1.69 – 7.87) and 4.58 (95% CI: 1.23 – 17.01), respectively.

Table 18: Baseline characteristics for immunoassay of MCC Spain study samples

	NCGC cases (n = 58)	**Controls (n = 58)**	**p-value***
Mean (SD) MFI of negative controls	4,750 (3,296)	4,276 (2,324)	0.59
Mean (SD) number of seroreactive antigens	17 (15)	20 (18)	0.13
Mean (SD) fold-change value among seroreactive antigens	2.68	2.86	0.28

**Mann-Whitney U test*
Signals derived from probing minimized H. pylori 26695-microarrays; SD = standard deviation; MFI = median fluorescence intensity;

Antibodies against the glutamine synthetase HP0512 (OR: 0.08, 95% CI: 0.01 – 0.65), the fumarate hydratase HP1325 (OR: 0.19, 95% CI: 0.05 – 0.71), the phosphatidylserine synthase HP1071 (OR: 0.13, 95% CI: 0.01 – 1.17), the acylamide amidohydrolase HP0294 (OR: 0.19, 95% CI: 0.04 – 0.87), the DNA protection during starvation protein HP0243 (NapA, OR: 0.37, 95% CI: 0.16 – 0.85) and the hypothetical protein HP0385 (OR: 0.35, 95% CI: 0.14 – 0.85) were more frequently detected in sera of controls than cases. Hence, presence of these antibodies was inversely associated with NCGC case status.

I further characterized the numerical fold-change values among seropositive sera for the selected *H. pylori* antigens (data not shown). Only CagA showed elevated fold-change values in seropositive NCGC cases compared to seropositive controls (5.29 and 3.22, respectively). Although the seroprevalence of HopA was increased in cases, the fold-change values were comparable between seropositives of the two groups (1.90 and 1.88, respectively). The same applied to the remaining antigens.
Out of the presented antigens, I selected HopA to be expressed as a GST-X-TAG fusion protein for multiplex serology to validate the potential NCGC risk markers on a high-throughput platform.

Table 19: Associations between seropositivity to individual H. pylori *antigens with NCGC status in the MCC Spain study*

Locus tag	Protein product	Seropositive cases (n = 58)	Seropositive controls (n = 58)	OR (95% CI)*
HP0547N	CagA	38	18	3.64 (1.69 – 7.87)
HP0229	HopA	13	4	4.58 (1.23 – 17.01)
HP0512	Glutamine synthetase	0	10	0.08 (0.01 – 0.65)**
HP1325	Fumarate hydratase	3	13	0.19 (0.05 – 0.71)
HP1071	Phosphatidylserine synthase	0	6	0.13 (0.01 – 1.17)**
HP0294	Acylamide amidohydrolase	2	10	0.19 (0.04 – 0.87)
HP0243	NapA	15	27	0.37 (0.16 – 0.85)
HP0385	Hypothetical protein	14	26	0.35 (0.14 – 0.85)

**ORs derive from unadjusted conditional logistic regression analysis; Controls were matched for sex, age and region*
***One seropositive case was imputed for to enable OR calculation*
OR = Odds ratio; CI = confidence interval;

3.1.1 Minimized *H. pylori* 26695-microarrays vs. multiplex serology

Sera from the MCC Spain study were already characterized by *H. pylori* multiplex serology in an earlier study.[99] This allowed me to compare the results to the microarray readouts using 16 shared antigens. Figure 10 shows scatter plots for each antigen to visualize the signal distributions. Despite seropositive signals in multiplex serology, there were no seropositive microarray readouts for the antigens HP0305, HP1564, HcpC and HpaA and only two for the antigen Cad. Thus, they were excluded from these analyses.

A high number of concordant results between the two methods, is depicted in the upper right and lower left quadrant of the scatter plots. There is also a substantial fraction in the lower right quadrant of each plot, representing sera found seropositive in the *H. pylori* multiplex serology, but seronegative on minimized *H. pylori* 26695-microarrays. This means that the sensitivity of the microarrays is diminished compared to the multiplex serology. The difference ranged from 0.30 for catalase and HP0231 to 0.81 for CagAN. Concordantly, the specificity reaches values between 0.81 for VacAN and 1.00 for CagAN, catalase and HP0231 (Table 20). I also calculated Spearman's rho to evaluate the non-parametric correlation between the two methods. It ranged from 0.26 for VacAN to 0.80 for CagAN and can thus be described as moderate to good. All of the *H. pylori* multiplex serology antigens derive from the strain 26695, except for GroEL and VacAN, which potentially contributed to a diminished correlation.

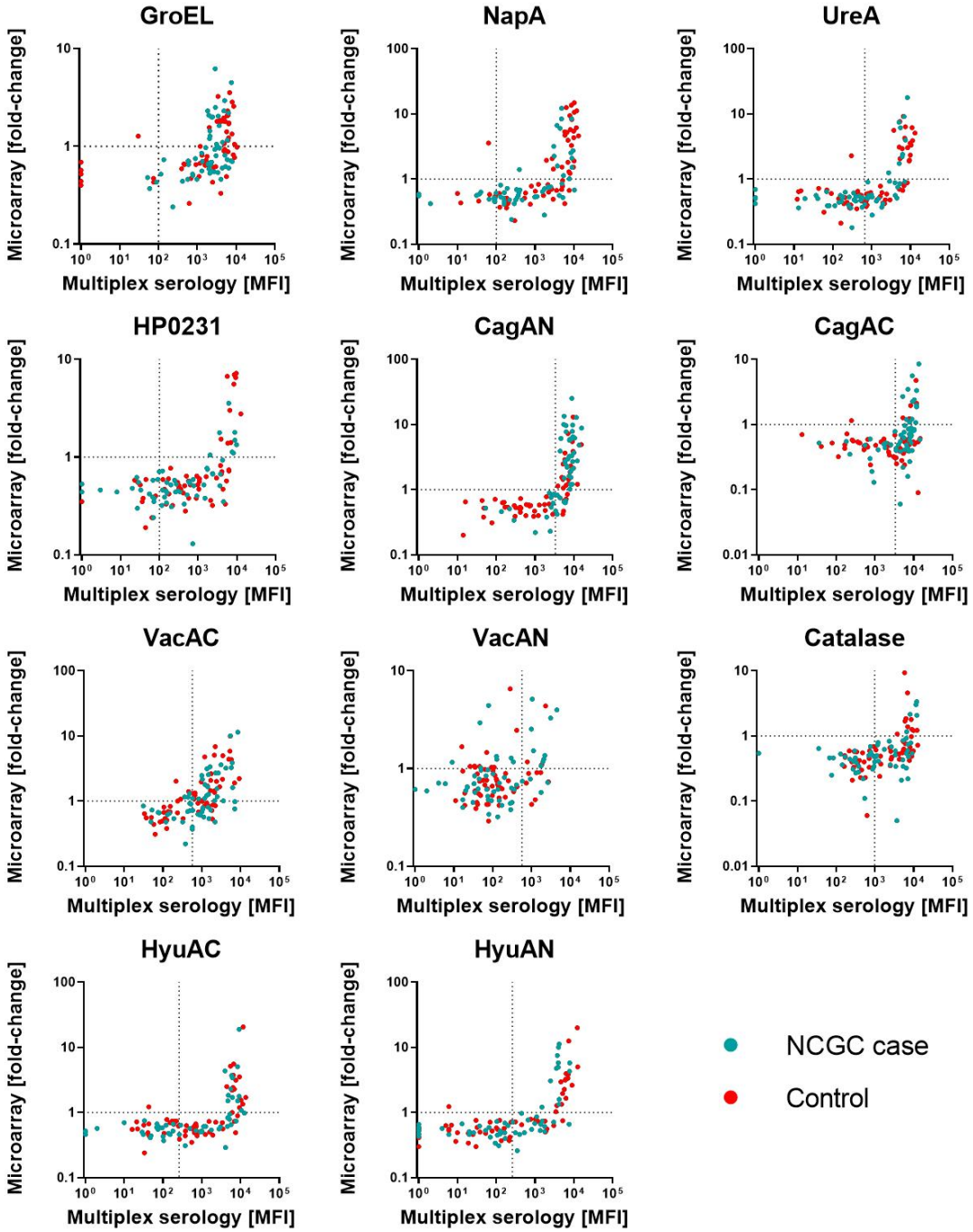


Figure 10: Scatterplots comparing microarrays with multiplex serology
Scatterplots comparing results from minimized H. pylori *26695-microarrays to* H. pylori *multiplex serology based on sera from NCGC cases and matched controls from the MCC Spain study. Dashed lines represent antigen-specific cutoffs.*
MFI = Median fluorescence intensity; NCGC = non-cardia gastric cancer

Table 20: Assay measures of minimized H. pylori *26695-microarrays compared to multiplex serology*

	Specificity [%]	**Sensitivity [%]**	**rho***
GroEL	91	40	0.59
NapA	94	42	0.73
UreA	98	39	0.64
HP0231	100	22	0.57
CagAN	100	81	0.80
CagAC	98	26	0.40
HyuAN	98	51	0.70
HyuAC	98	39	0.60
VacAN	81	52	0.26
VacAC	85	67	0.69
Catalase	100	22	0.54

* *Spearman's rank correlation coefficient*

3.2 Identifying NCGC-associated antigens in two Chinese studies using *H. pylori* multi-strain microarrays

3.2.1 Designing *H. pylori* multi-strain microarrays

The minimized *H. pylori* 26695-microarray showed promise to identify novel cancer-associated serological markers. To advance the platform further and to consider the phylogeographic distribution of *H. pylori*, I pursued a multi-strain approach. For this purpose, I received four strains from the *H. pylori* Genome Project (HpGP), which were isolated in Chile, Korea, Latvia and Malaysia (Table 21). As the HpGP was still in a very early stage, only glycerol stocks and unannotated DNA sequences were available.

For the annotation, I utilized the NCBI Prokaryotic Genome Annotation Pipeline (PGAP).[94] I also re-annotated *H. pylori* 26695, as the original annotations, which were used to generate DNA expression constructs, were from 1997.[103] Meanwhile, a systematic re-annotation project was initiated by the NCBI to update its protein sets, which also included *H. pylori* 26695 (https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/reannotation). I found these revised annotations identical to my results (data not shown), reassuring the validity of the used algorithm and settings. This re-annotation of *H. pylori* 26695 yielded 38 additional protein-coding sequences, while 49 wrongly annotated genes were removed.

Overall, I identified 7,745 *H. pylori* protein-coding genes in the five strains (Table 21). In order to distinguish them, I used temporary identifiers combining the origin (CHI, KOR, LAT, MAL or 26695) and a sequential number (e.g. kor112 was the 112th gene on the KOR-037 genome).

Table 21: H. pylori *strains of the* H. pylori *multi-strain microarray*

***H. pylori* Strain**	**Origin**	**Genome size (kb)**	**Protein-coding genes**
26695	United Kingdom	23.8	1,523
CHI-122*	Chile	28.7	1,599
KOR-037*	Korea	27.4 (+ 7.6 plasmid)	1,514
LAT-022*	Latvia	28.7 (+ 6.6 plasmid)	1,572
MAL-007*	Malaysia	28.3	1,539

**Strains were kindly provided by Dr. M. C. Camargo from the HpGP.*
kb = kilo base pairs

Generating expression constructs for each of the protein-coding genes would require an immense amount of material and lead to redundancies due to sequence homologies. Also, the number of displayable proteins on a microarray is limited.

Therefore, I clustered the 7,745 sequences to identify a non-redundant set of protein-coding genes. Overall, the proteins were assigned to 1,833 clusters (Figure 11A), 1,171 clusters contained genes from every *H. pylori* strain, while 125 clusters contained genes from four strains, 92 clusters contained genes from three strains and 117 clusters contained genes from two strains. Additionally, there were 328 clusters which represented stain-specific antigens.

For each of the clusters I chose a representative to be displayed on the *H. pylori* multi-strain microarray. If possible, I chose a sequence derived from *H. pylori* 26695 as cluster representative, because primer pairs for DNA expression constructs were already available. This was the case for 1,413 clusters. If a cluster did not contain sequences derived from *H. pylori* 26695, I selected the longest sequence as cluster representative: 119 sequences derived from CHI-122, 94 derived from KOR-037, 103 from LAT-022 and 114 from MAL-007 (Figure 11B).

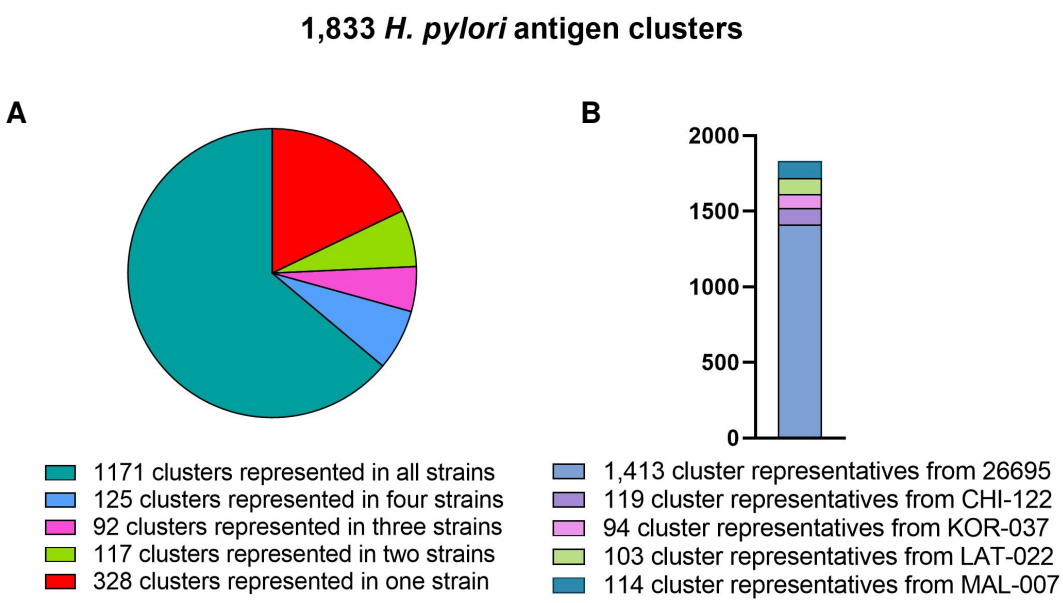


Figure 11: H. pylori antigen cluster

Protein-coding genes (n = 7,745) were assigned to 1,833 antigen clusters. A: Composition of protein clusters. Almost two thirds of the clusters contain proteins of all five utilized strains, while 18% represent strain-specific antigens. B: Strain origin of cluster representatives which are displayed on the H. pylori *multi-strain microarray.*

3.2.2 Generating and probing *H. pylori* multi-strain microarrays with sera from the Shanxi study and NIT

I prepared expression constructs for all of the 1,833 *H. pylori* genes using gene-specific primer pairs and common expression primer pairs in two successive PCRs. Sizes of PCR products were checked by gel electrophoresis (data not shown). If no band was detectable or the size was erroneous, PCR reactions were repeated with adjusted conditions. Eventually, 1,618 expression constructs were successfully generated. For the remaining 215, a successful generation could not be confirmed on DNA level. Nevertheless, I used all of the 1,833 PCR products for the spotting of the *H. pylori* multi-strain microarray, as theoretically only few intact expression constructs are necessary to enable protein expression. Later on, 57 of the 215 expression constructs without detectable bands in gel electrophoresis were found full-length expressed by detecting the C-terminal V5-tag of the *on-chip* translated antigens. Furthermore, 21 were found partially expressed with only the N-terminal 6xHis-tag being detectable.
Alongside the DNA expression constructs for *H. pylori* antigens, I also prepared positive controls and negative controls. The latter were products of template-free PCRs. Positive controls were either expression constructs for the highly prevalent EBV VCA p18 antigen or secondary anti-human antibodies.

I experimentally determined the maximum number of antigens that can be displayed on a microarray to be 2,001, arranged in 23 columns and 87 rows (data not shown). Therefore, I was able to include 20 positive controls and 129 negative controls in addition to the *H. pylori* expression constructs. The design of the final microarray and the distribution of controls are displayed in Figure 13A.

In total, I produced 594 of these *H. pylori* multi-strain microarrays in 18 batches. Ten batches were spotted in two subsequent weeks and the remaining eight batches were generated one month later in another two subsequent weeks. Although I calculated extra volume, some wells ran empty in the course of the microarray spottings. From batch 14 on, I added ddH_2O to these wells diluting the concentration of DNA expression constructs. Because protein expression rates are not correlated to the amount of DNA, and I always measured matched sera together on two subsequent slides in the same batch, the comparison still remained valid.[91]

To ensure that the dilution of DNA concentration did not lead to a decrease in overall protein expression rates, I stained the first/second and last slide of the batches 11-18. An exemplary staining is depicted in Figure 13B. A protein was considered expressed if the signal of the terminal tag exceeded the mean of the negative controls + 3 SDs (positive outliers excluded); partial expression was observed if only the N-terminal signal was detected. The first slide of batch eight was excluded after visual inspection. Overall, the expression rates ranged from 77% to 94% (Figure 12). A systematic decrease across time and batches was not observable.

Sera from the Shanxi (n = 116) and NIT (n = 238) studies were used to probe the microarrays. Case-control pairs were always tested on subsequently spotted microarrays from the same spotting batch.
I performed the immunoassays with help from a S. Blumrich, who I supervised during her practical semester. She used the software GenePix to determine raw MFI signals for each spots on each microarray being blinded to the respective case- control status of the sample. Overall, 3,879 (1.6%) spots from a total of 248,124 spots were manually flanked invalid due to smearing, dust particles or visible spotting failures.

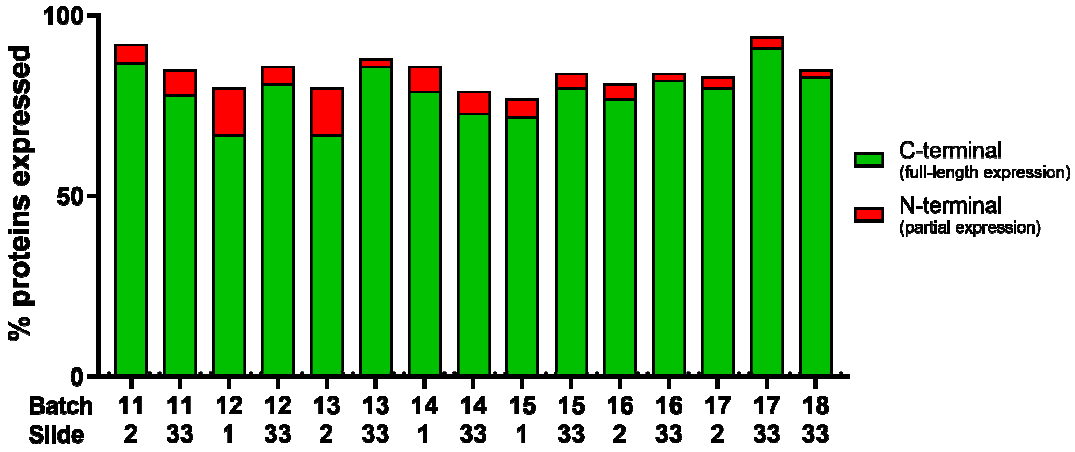


Figure 12: Protein expression control stainings
Terminal tags of the first/second and last slide of batches 11-18 were stained. Full-length expressed proteins are visualized by a green bar while partially expressed proteins are added in red.

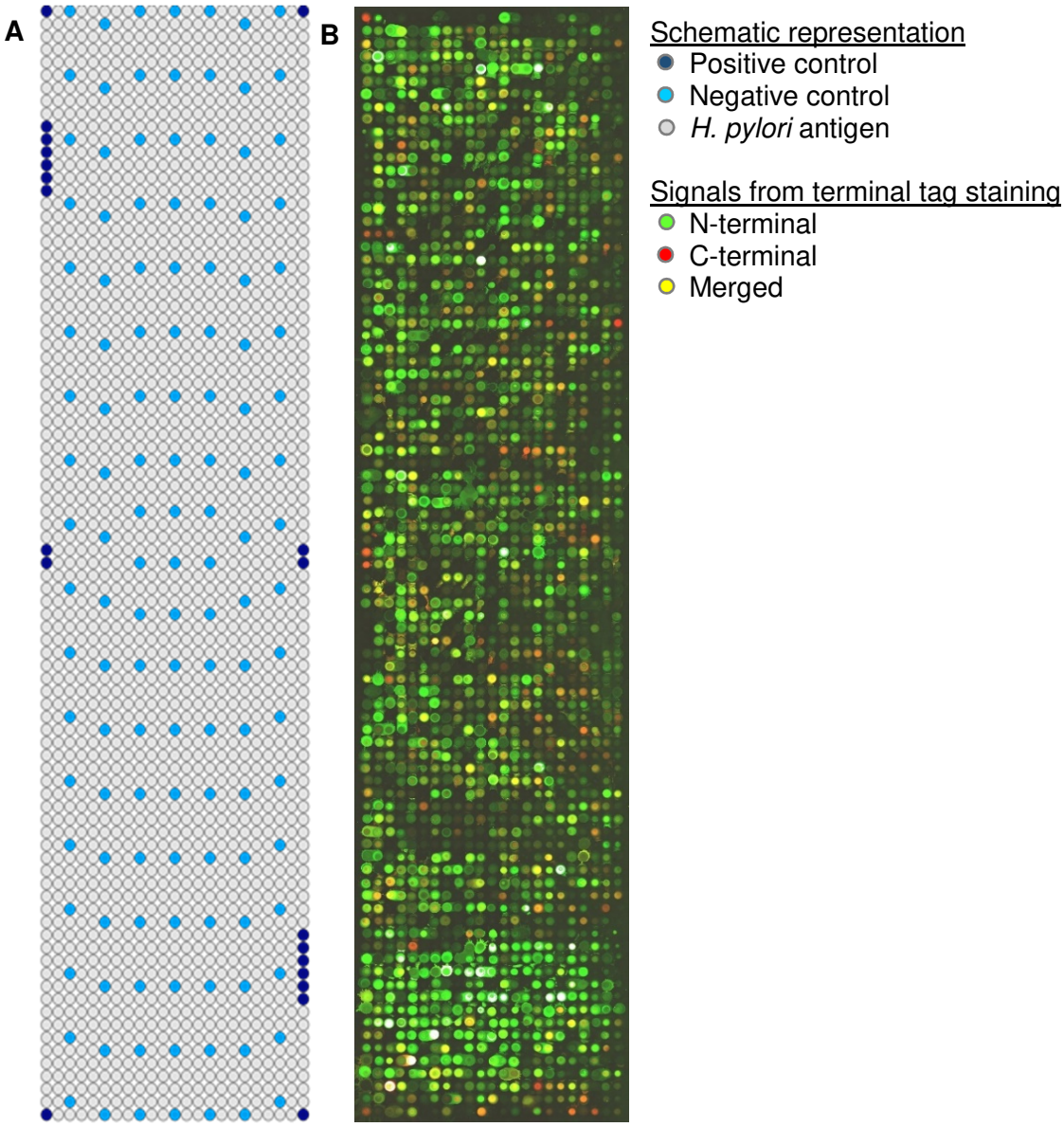


Figure 13: **H. pylori *multi-strain microarray***
A: Setup of the H. pylori *multi-strain microarray. Each slide presents 1,833* H. pylori *antigens (grey points), 129 negative controls (light blue points) and 20 positive controls (dark blue points). Positive controls in groups of six are anti-human antibodies, while others are EBV VCA p18 antigens. B: Expression control staining of on-chip expressed* H. pylori *antigens. Green signals derive from staining of the C-terminal V5-tag, while red signals derive from staining of the N-terminal 6xHis-tag. Merged signals appear yellow, while saturated signals appear white.*

3.2.3 Optimizing signal acquisition for large antigen microarrays

Large protein microarrays often suffer from background effects (gradients, spatial artifacts, smear contaminations or small speckles).[104-106] These effects were negligible on minimized *H. pylori* 26695-microarrays, but very prominent on *H. pylori* multi-strain microarrays. Potential causes included the increased number of spots, a larger surface area and a longer spotting duration per batch. As an example, the protein expression control staining visualized in Figure 13B exhibited a gradient from left to right.

For the dichotomous analyses of expression control stainings, background effects were negligible because monoclonal antibodies (anti-V5 and anti-6xHis) directed against the terminal tags generated a high signal-to-noise ratio. In contrast, serum antibodies showed a broader MFI distribution and in order to distinguish seropositive signals from seronegative signals, I needed to pre-process raw MFI readouts and take background effects into account,

For this purpose, I adapted an analytical approach to calculate local backgrounds for each spot/antigen individually.[104] Thereby, universal background effects across entire arrays can be compensated for. The approach required the majority of antigens to be unreactive, an assumption that held true for large *H. pylori* multi-strain microarrays.

In order to determine the local background for each spot/antigen individually, I averaged the raw MFI values of the respective nearest neighbors (Euclidean distance) assuming that most of these signals corresponded to background noise. To account for the seropositive antigens and positive controls, I excluded positive outliers (raw MFI signals > 2 SDs of all neighbors) from this calculation. The principle is visualized in Figure 14.

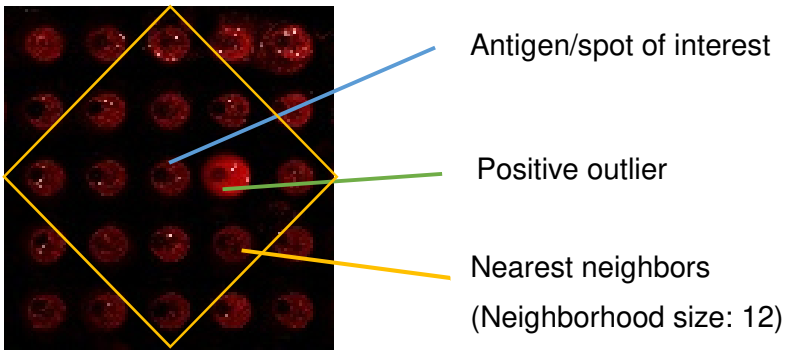


Figure 14: Local background definition for microarray spots
Local backgrounds are determined for each spot of the microarray. Signal readouts from the nearest neighbors (here: 12) surrounding the spot of interest were averaged to define the local background. Positive outliers were excluded from the calculation to account for seropositive signals.

In order to determine the appropriate number of neighbors, I used the SD of negative controls per microarray as a readout. These negative controls were spotted from the same aliquot and therefore represent the technical variation.
For each microarray, I calculated background signals for each spot based on different neighborhood sizes (8, 12, 24, 50 and 100) and subtracted them from the respective raw MFI signals. Using twelve nearest neighbors for the local background calculation, reduced the mean SD of negative controls by 25% (from 2,095 MFI to 1,579 MFI).

After subtracting local backgrounds, I identified seropositive antigens in a subsequent step. For this purpose, I pre-specified a 95% specificity by applying a cutoff which allowed 5% of the negative controls to be seropositive. Signals exceeding this cutoff were considered seropositive. A graphical representation of a background subtraction and identification of seropositive signals is given in Figure 15. A further normalization to account for patient-to-patient variance was not necessary, as these variances were part of the local backgrounds and, hence, already compensated for.

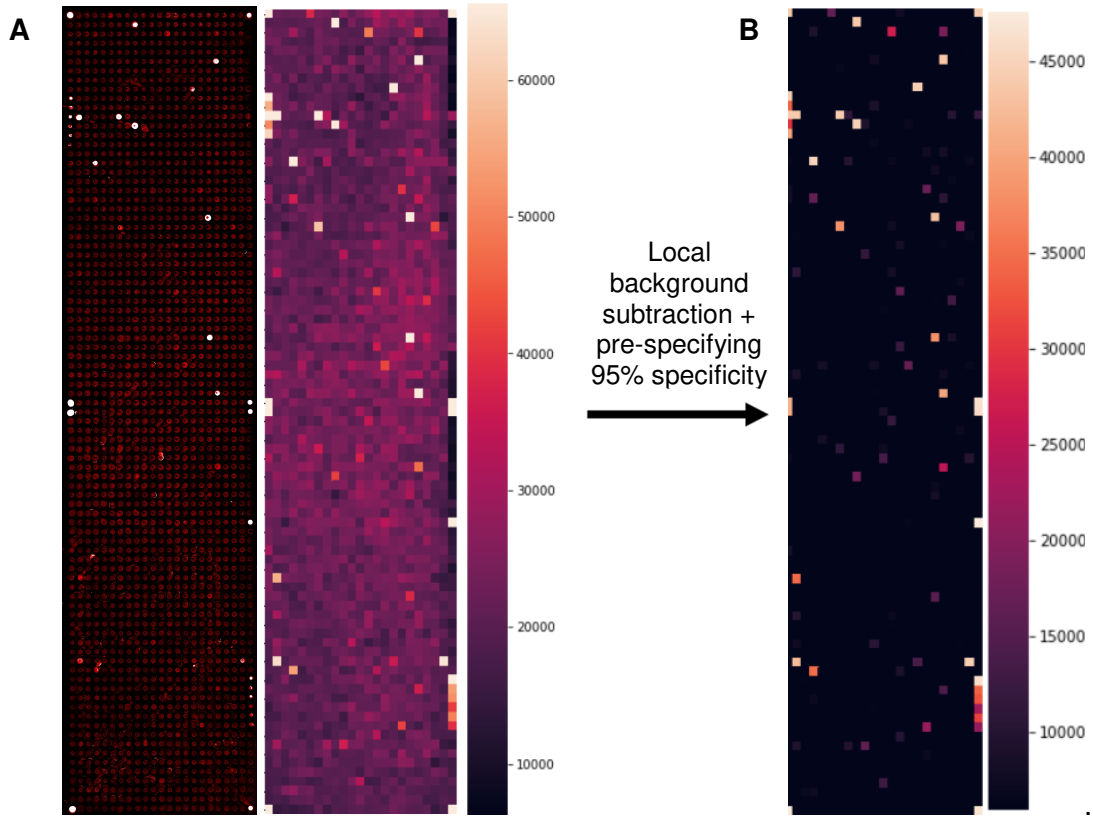


Figure 15: Local background subtraction
A: Exemplary immunoassay and its corresponding visualization as a heat map. B: MFI signals after subtracting local backgrounds and pre-specifying a 95% specificity. MFI values below the cutoff were set to 1.

3.2.4 Optimized signal acquisition vs. multiplex serology

In order to evaluate the new signal acquisition process, I performed a multiplex serology using established *H. pylori* antigens and sera from the Shanxi study, which were used to probe the *H. pylori* multi-strain microarrays.
Non-parametric correlations between raw microarray readouts and multiplex serology were overall poor. The highest correlation was observed for GroEL (rho: 0.56), although the utilized proteins derive from different *H. pylori* strains (multiplex serology: *H. pylori* G27; microarray*: H. pylori* 26695).

Based on these raw signals, I calculated fold-change values normalized to the global patient specific backgrounds (median + 5 MADs of all negative controls), which corresponded to the 'old' signal acquisition process. For minimized *H. pylori* 26695-microarrays, correlations showed a mean rho of 0.59 (Table 22). Applied on large *H. pylori* multi-strain microarrays, global normalization decreased the correlation with multiplex serology compared to the raw microarray results. Only the correlation for HyuAN improved (rho: 0.21 and 0.76, respectively).

In contrast, applying the 'new' signal acquisition process by subtracting local backgrounds from raw microarray signals, improved the correlation with multiplex serology for each antigen. For GroEL and HyuAC, Spearman's rho reached 0.81 and 0.76, respectively. Still, the correlation remained poor for many antigens, especially if baseline seroprevalences according to multiplex serology were low (HP0231, HP1098, Cad). Despite having a high seroprevalence of 71%, the antigen VacAN also showed a low correlation (rho: 0.02).

I also calculated sensitivities and specificities for each antigen to compare the 'old' (global normalization to fold-changes) and 'new' (local background subtraction) signal acquisition processes to multiplex serology. For this purpose, multiplex serology results were dichotomized using cutoffs defined by finite mixture modeling (see 3.4.1). Specificities for globally normalized fold-changes ranged from 0.57 for HP1564 to 0.78 for CagAN and sensitivities ranged from 0.19 for VacAN to 0.88 for NapA.
Background subtracted MFI values showed a higher or equal specificity across all antigens. Sensitivities were increased for some antigens including GroEL, UreA and HP0231 (48% to 78%, 43% to 71% and 40% to 60%, respectively), while it was decreased for others including CagAN, CagAC and VacAN (32% to 25%, 30% to 22%

and 19% to 12%, respectively). The decreases were, however, moderate and compensated by the larger increase in specificities.

Overall, I confirmed the need for background correction by showing that raw signals did not correlate well with multiplex serology. Furthermore, I showed that the new signal acquisition method performed better than a global normalization, which was used for the minimized *H. pylori* 26695-microarrays.

Table 22: Comparison of processed signals with raw microarray readouts

		Raw signals [MFI]	**Global normalization [fold-change]*** **			**Local background subtraction [MFI]*****		
	Prev [%]*	**rho**	**Sp [%]**	**Se [%]**	**rho****	**Sp [%]**	**Se [%]**	**rho****
GroEL	52	0.56	77	48	0.41	88	78	0.81
UreA	11	0.28	74	43	0.24	86	71	0.30
HP0231	4	0.01	74	40	0.01	92	60	0.14
NapA	20	0.42	76	88	0.37	91	88	0.52
HP0305	38	0	73	26	0.05	88	23	0.31
HpaA	6	0.07	77	29	0.02	96	0	0.12
CagAN	60	0.16	78	32	0.10	96	25	0.37
CagAC	44	0.15	73	30	0.05	87	22	0.32
HyuAN	10	0.14	68	46	0.11	96	46	0.25
HyuAC	19	0.21	66	35	0.76	97	30	0.76
VacAN	73	0	67	19	0.01	85	12	0.02
HP1098	21	0.09	64	38	0.05	90	35	0.12
Cad	10	0.13	74	33	0	94	25	0.01
HP1564	59	0.44	57	66	0.35	57	68	0.49

**seroprevalence based on multiplex serology*
***Spearman's rank correlation coefficient*
****signal acquisition method used for the minimized* H. pylori *26695-microarrays*
*****signal acquisition method used for large H. pylori multi-strain microarrays*
Sensitivities and specificities were calculated using multiplex serology as a reference.
Prev = Seroprevalence; Sp = Specificity, Se = Sensitivity; MFI = median fluorescence intensity

3.2.5 Identifying informative *H. pylori* antigens to classify NCGC status using *H. pylori* multi-strain microarray results in the Shanxi study and NIT

After determining net antibody responses for each antigen in each study sample, descriptive and statistical analyses were conducted in order to identify informative antigens to transfer to multiplex serology. I analyzed the two studies separately as biological differences might exist (cross-sectional vs. prospective study).

First, I narrowed down the 1,833 *H. pylori* antigens by removing those that showed seroprevalences below 10% in NCGC cases. Furthermore, I calculated crude ORs and removed inversely associated antigens. Mann-Whitney U tests and chi-squared tests were applied on numerical and dichotomized data sets, respectively, in order to manually evaluate differences between the two groups (Table 23, Supplementary Table S3). Due to the exploratory nature of these analyses, I did not account for multiple testing.

To decide which of the remaining antigens showed promise to be transferred to multiplex serology, I used random forests classification. This supervised machine learning algorithm used numerical microarray readouts to model the respective case/control status based on a tree-like Boolean logic ('if-then-else', Figure 16). In contrast to individual decision trees, it builds numerous trees using bootstrapped subsets of the original data and averages them into a single classification model.
By pre-defining conditions (e.g. maximal tree depths or minimal samples per tree branche), I limited the number of antigens per decision tree and, hence, forced the algorithm to primarily use the most informative antigens to classify sera as NCGC case or control. The more an antigen contributed to a correct classification, the higher was its semi-quantitative importance value. After building numerous trees under various conditions, I selected antigens which reoccurringly showed the highest importance values (Table 23).

Analyzing the Shanxi study, random forests were heterogeneous. Apart from HP1564, also HP1091 and HP1064 were most consistently utilized in the respective models. Prevalences of antibodies against these antigens were 61%, 19% and 60% among NCGC cases and 53%, 34%, and 44% among controls.
For the NIT, HP1038, HP0003 and HopA showed especially high importance values across most of the numerous random forests. Antibodies against these three antigens

were found elevated in NCGC cases (54%, 13% and 39%, respectively) compared to controls (34%, 6% and 29%, respectively).
Apart from these additive decision trees, which assume a dependence between antigens, I also used the classification algorithm in a recursive manner. This means, that I iteratively removed informative antigens while modeling case/control status. This led to the use of antigens, which would otherwise not be included in the models, because they highly correlated with other informative antigens. To potentially maximize the chance of a successful transfer of relevant antigens from microarray to multiplex serology, I decided to include these otherwise redundant antigens as well.

Furthermore, I also included four manually picked *H. pylori* antigens (HP0185, Lat118, HP0527 and HP0385) due to a high baseline prevalence and a notable distribution. Antibodies against HP0185 and Lat118 had a similar prevalence among NCGC cases (52% and 48%, respectively) and controls (50% and 44%, respectively) among the NIT sera, but a decreased prevalence in NCGC cases (45% and 35%, respectively) of the Shanxi study, compared to controls (73% and 56%, respectively). Antibodies against HP0527 were also similar in NCGC cases and controls of the NIT sera (13% and 15%, respectively), but elevated in NCGC cases compared to controls among the Shanxi sera (19% and 10%, respectively). Antibodies against HP0385 were more prevalent in NCGC cases than controls (48% and 35%, respectively) among the NIT sera.

Eventually, I selected 22 promising *H. pylori* antigens to be expressed as GST-X-TAG fusion proteins in a subsequent step (Table 23).

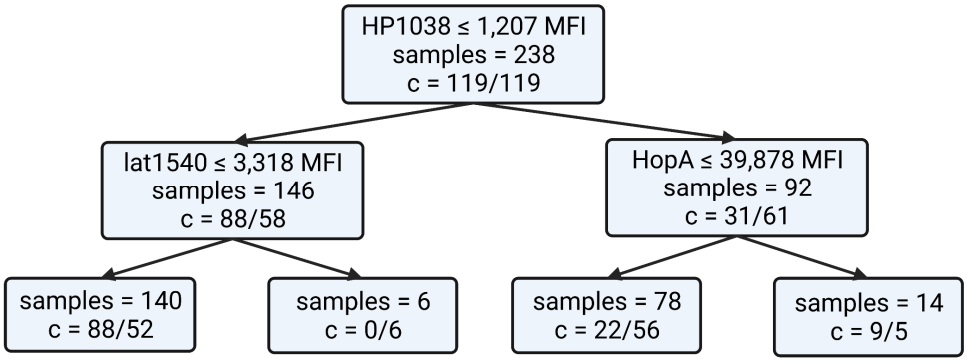


Figure 16: Decision tree classification model
Exemplary decision tree classifying 238 NIT serum samples (119 NCGC cases and 119 sex- and age-matched controls) into two classes (c = NCGC case / control). A maximal tree depth of 3 was specified (3 levels of decision). Random forest algorithms built numerous decision trees and average the results to generate a final output.

T*able 23: Microarray results of* H. pylori *antigens selected for multiplex serology*

		NCGC cases			Controls				
Shanxi		**Prev [%]**	**Median [MFI]***	**IQR [MFI]***	**Prev [%]**	**Median [MFI]***	**IQR [MFI]***	**MWU test**	**χ2 test**
Selected from additive decision trees	HP1564**	61	34,788	37,676	53	21,723	30,208	0.33	0.57
	HP1091	19	668	744	34	1,530	1,089	**0.01**	0.13
	HP0477	60	13,891	25,059	44	17,000	36,240	0.08	0.19
	HP1355	40	2,827	3,098	50	1,764	2,072	0.12	0.45
	HP1064	42	1,642	2,612	44	1,280	1,443	0.13	1.00
	HP0545	32	2,381	4,130	45	2,581	5,181	0.39	0.26
Selected after iterative recursive elimination	HP1091	19	668	744	34	1,530	1,089	**0.01**	0.13
	HP1570	26	1,079	1,139	18	825	825	0.39	0.41
	HP1435	23	2,096	2,336	19	761	594	0.35	0.84
	HP1064	42	1,642	2,612	44	1,280	1,443	0.13	1.00
	HP0017	26	971	1,161	11	1,704	3,542	0.08	0.08
	Lat98	16	3,720	5,437	11	745	1,607	0.33	0.62
	Kor1294	15	1,763	1,168	13	458	501	0.21	1.00
Manually added	HP0185	45	2,617	10,346	73	4,417	10,986	0.44	**0.03**
	Lat118	35	1,113	1,819	56	1,528	2,150	0.06	0.07
	HP0527	19	618	892	10	385	116	0.30	0.22
	HP0385	42	5,493	15,798	48	10,924	25,173	0.43	0.65
NIT									
Selected from additive decision trees	HP1038	54	12,024	24,323	34	14,690	26,870	0.38	**0.01**
	HP0003	13	1,456	2,676	6	502	1,552	0.49	0.08
	HP0659	39	6,568	17,752	29	3,327	23,032	0.42	0.17
	HopA	64	17,261	31,126	55	12,720	32,874	0.46	0.19
	HP0582	91	29,046	26,834	88	30,391	31,325	0.16	0.67
	Mal648	23	3,490	7,117	21	2,325	1,124	**0.04**	0.88
	Mal1434	31	3,065	3,613	24	1,264	2,503	**0.03**	0.31
Selected after iterative recursive elimination	Mal648	23	3,490	7,117	21	2,325	1,124	0.04	0.88
	NapA**	34	6,654	17,713	23	2,490	7,574	0.29	0.06
	HP0003	13	1,456	2,676	6	502	1,552	0.49	0.08
	Mal1434	31	3,065	3,613	24	1,264	2,503	**0.03**	0.31
	HP0659	39	6,568	17,752	29	3,327	23,032	0.42	0.17
	Lat1540	16	2,407	6,676	8	1,186	1,950	0.25	0.11
Manually added	HP0185	52	5,606	19,465	50	8,062	19,126	0.10	0.80
	Lat118	48	1,821	2,990	44	1,667	2,835	0.38	0.60
	HP0527	13	715	1,351	15	550	681	0.26	0.85
	HP0385	48	17,404	32,130	35	21,624	29,180	0.08	0.07

**among seropositive signals only. ** established* H. pylori *multiplex serology markers. Prev. = Serorevalence; IQR = interquartile range; MWU test = non-parametric Mann-Whitney U test; χ2 test = chi-squared test; NCGC = non-cardia gastric cancer; NIT = nutrition intervention trial; MFI = median fluorescence intensity*

3.3 From microarray to multiplex serology: Expressing GST-X-TAG fusion proteins for a high-throughput validation

3.3.1 Cloning and expression of GST-X-TAG fusion protein

Selected antigens were expressed as GST-X-TAG fusion proteins for multiplex serology. The sequences of HP0527, HP0017 and Kor1294 were too long to be full-length expressed in *E.coli* BL21 with 5.8 kb, 2.4 kb and 2.6 kb, respectively. This was confirmed by protein expression stainings on the microarrays, as the C-terminal V5-tag was not detectable. Hence, I decided to truncate these proteins to promote successful protein expression. For the VirB4 homolog HP0017 and the silent information regulator (SIR2) family protein Kor1294, I retained functional N-terminal domains by expressing the first 385 and 511 amino acids, respectively. HP0527, which is part of the *cag* PAI, was already characterized by Soluri *et al.*[107] They identified an immunogenic epitope corresponding to the amino acids 566-655, which I chose to be expressed as a GST-X-TAG fusion protein. In the following, these proteins were referred to as HP0017N, Kor1294N and HP0527trunc.
Final sizes of all selected *H. pylori* antigens, protein descriptions and information about homologues in the fiver further *H. pylori* strains are summarized in Table 24. Sequences are listed in the Supplementary Table S4.

Plasmids for the antigens HP1091 and Kor1294N could not be successfully cloned into the expression plasmid pGEX-4T3tag by Eurofins genomics and were delivered in an alternative vector. Hence, I manually cloned the insert sequences into the pGEX-4T3tag vector using restriction digests and a subsequent ligation.
Transformation of *E. coli* BL 21 cultures and expression of the fusion proteins were performed as described (2.3.2). Final antigen lysate concentrations are listed in Table 26. The expression and subsequent characterization of the plasmids and fusion proteins were performed with the help of N. Kruse, who I supervised in the scope of a practical semester. Furthermore, the expression of recombinant HopA was performed by D. Schulz in the scope of a Bachelor's thesis, also under my supervision.

Table 24: Characterization of new H. pylori *multiplex serology antigens*

Accession	Protein	Accession	Location	Size [bp]	Size [kDa]	Number of *H. pylori* strains encoding respective antigens (max = 5)
HP0229	Outer membrane protein HopA	WP_000751160	membrane	1,449	53	4 (not in MAL-007)
HP1038	3-dehydroquinate dehydratase	WP_000699284	cytoplasm	501	18	5
Mal648	Uncharacterized	WP_053577294	membrane	276	11	1 (only in MAL-007)
HP0003	3-deoxy-8-phosphooctulonate synthase	WP_000858181	membrane	828	30	5
HP0659	SurA_N domain-containing protein	WP_001225999		1,242	48	5
HP0582	TonB_C domain-containing protein	WP_001114818	membrane	1,026	37	5
HP0185	Uncharacterized	WP_001290538	membrane	801	30	5
Lat118	Uncharacterized	WP_164866270		186	7	1 (only in LAT-022)
Mal1434	VirB6 homologue	WP_108562744	membrane	576	17	1 (only in MAL-007)
Lat1540	Uncharacterized	WP_128061807	Membrane	429	20	2 (LAT-022, CHI-122)
HP0527trunc*	Cag pathogenicity island protein cagY	WP_001001434	membrane	267*	10*	4 (not in KOR-037)
HP0385	Uncharacterized	WP_000983762		228	9	5
HP1570	HAD family hydrolase	WP_000593725		492	18	5
HP1435	signal peptide peptidase SppA	WP_000269537	cytoplasm	876	32	5
HP1064	Prokaryotic metallothionein family protein	WP_010875589	membrane	279	11	5
HP0017N*	VirB4 homologue	WP_000890508		1,125*	45*	5
HP0477	Outer membrane protein HopJ	WP_001167977	membrane	1,101	42	3 (only in 26695, KOR-037, CHI-122)
HP1355	Nicotinate-nucleotide pyrophosphorylase	WP_000405980	cytoplasm	819	31	5
HP0545	*Cag* PAI protein cagD	WP_000609477	membrane	621	24	5
Lat98	Uncharacterized	WP_108247824		738	29	1 (only in LAT-022)
Kor1294N*	SIR2 family protein	WP_121292687		1,545*	50*	1 (only in KOR-037)
HP1091	Alpha-ketoglutarate permease	WP_001069529	membrane	1,278	48	5

**Sequences were truncated to enhance successful protein expression*
bp = base pairs; kDa = kilo Dalton;

Table 25: Concentrations of GST-X-TAG protein lysates

Antigen	Lysate concentration [mg/ml]	Antigen	Lysate concentration [mg/ml]
HopA	12.4	HP0527trunc	19.8
HP1038	17.6	HP1570	15.4
Mal648	10.0	HP1435	6.1
HP0003	15.4	HP1064	8.8
HP0659	16.7	HP0017N	8.2
HP0582	6.9	HP0477	7.8
HP0185	15.4	HP1355	20.9
Lat118	15.4	HP0545	6.9
Mal1434	6.1	Lat98	10.5
Lat1540	9.0	Kor1294N	5.5
HP0385	13.2	HP1091	8.0

3.3.2 Verifying DNA integrity

I controlled the integrity of the ordered plasmids at two time points using two different methods. First, I performed multiple analytical DNA digests using isolated plasmid DNA from transformed *E. coli* cultures before induction of protein expression. Each plasmid was examined by three digestion approaches: (L) linear digestion to estimate the length of the linearized plasmid, (S) symmetric digestion separating the insert from the backbone and (A) asymmetric digestion using a restriction site within the insert and another within the plasmid backbone to verify correct fragment sizes. Selected restriction enzymes and the expected fragment sizes are summarized in Table 13. To visualize and evaluate DNA fragments, I used DNA gel electrophoresis (Supplementary Figure S1). Overall, all of the fragments were of the expected sizes. Linear digestion of HP0185 and HP0003 showed three instead of two DNA fragments due to an additional BamHI restriction site within the insert sequence. The asymmetric digestion of HP0185 resulted in three bands, too, due to incomplete digestion. Generally, fragments <300 bp were barely visible on the gels.

Furthermore, I isolating plasmids from the final antigen lysates after induction of protein expression and used these as templates in a subsequent PCR with the primer pair pGEX-T7 and pGEX-T3. The complementary sequence of these primers flanked the inserted antigen. I purified the PCR product, verified the correct size by gel electrophoresis and sent it to Eurofins Genomics (Ebersberg) to be sequenced.
Consensus sequences were created by aligning sequencing results from both strands after reverse complementation of the antisense strand. I confirmed that all of the consensus sequences corresponded to the anticipated antigens.

3.3.3 Characterizing recombinant fusion proteins with anti-TAG ELISA, anti-TAG Western blot and anti-GST Western blot

After checking DNA constructs, I proceeded to detect and characterize the recombinantly expressed GST-X-TAG fusion proteins. For this purpose, I prepared anti-GST and anti-TAG Western blots (Supplementary Figure S2 and Supplementary Figure S3). For each antigen two samples were analyzed, one taken before clearing the lysates by centrifugation and another after. Western blot experiments were complemented by a semi-quantitative anti-TAG ELISA (Figure 17). For both assays, I used a GST-TAG lysate (lacking an inserted antigen) side-by-side as a control or reference for relative quantification.

Recombinant antigens containing Lat98, HP1355, HP0527trunc, HP1570, Lat118, Mal648, HP1038, HP1064, HP0545, HP0003 and HP0385 were verified in both Western blots detecting the N-terminal GST, as well as the C-terminal TAG at the correct sizes. Bands which corresponded to the full-length fusion proteins showed the highest intensity. The respective ELISA results confirmed these results: the sigmoidal curve reached the same plateau as the GST-TAG reference indicating predominantly full-length expressed antigens.
Relative concentrations of full-length fusion proteins were determined setting the GST-X-TAG concentration in relation to the GST-TAG concentration, both at the half-maximal concentration of the GST-TAG ($OD50_{GST\text{-}tag}$) (Table 26). For instance, the concentration of full-length Lat98 was 3.3-fold higher than the full-length GST-TAG in the control lysate, while the concentration of full-length HP0003 only reached the 0.064-fold of the GST-TAG control.

Sigmoidal ELISA curves for the recombinant antigens containing Lat1540, HP1435, HP1091, HP0017N, HP0582, HP0659, HP0017N and HopA showed a lower plateau than the GST-TAG reference. This indicated that a substantial fraction of the fusion proteins was only partially expressed or degraded. This observation was confirmed by Western blot results. While the N-terminal GST was detectable, the C-terminal anti-TAG Western blot only showed very faint or no bands. Consequently, the relative quantity of full-length fusion proteins is decreased, e.g. to the 0.061-fold (Lat1540) or 0.007-fold (HopA) of the GST-TAG lysate.
Recombinant HP0185 did also not reach the same plateau as the GST-TAG reference in the ELISA. The anti-GST Western blots showed a very intense band above the band corresponding to the full-length fusion protein. This bigger protein was also detected

by the anti-TAG antibody and was, hence, possibly caused by protein interactions with *E.coli* host proteins strong enough to resist denaturation during Western blot.
Recombinant antigens containing HP0477, Mal1434 and Kor1294N did not show any signal on the anti-TAG Western blot. Their ELISA results did not reach half-maximum absorbance (OD_{50}) of the GST-TAG control, which obviates a relative quantification. Yet, the measured absorbance was rising with increased antigen lysate concentrations, which indicated that expression of the recombinant protein was possible, but inefficient. This was confirmed by Western Blot results, as very faint bands were seen at the expected sizes of the full-length proteins.

All of the recombinant proteins show additional bands on the Western blots, especially using an anti-GST antibody. This was expected as partial expression and proteolysis are common and part of the natural metabolism of the host bacterium. Purification to enrich full-length protein is possible, however, not required for multiplex serology.
As partially expressed antigens can suffice as targets to subsequently bind serum antibodies, I decided to transfer all of the *H. pylori* fusion proteins to multiplex serology.

Table 26: Relative concentration of full-length GST-X-TAG fusion proteins

Antigen	Concentration of full-length fusion proteins relative to GST-TAG [fold-change]	Antigen	Concentration of full-length fusion proteins relative to GST-TAG [fold-change]
Lat98	3.333	Lat1540	0.061
HP0527trunc	2.000	HP1435	0.056
HP1570	2.000	HP0185	0.047
Lat118	1.667	HP1091	0.023
HP1355	1.667	HP0017N	0.020
Mal648	1.429	HP0582	0.012
HP1038	0.588	HP0659	0.009
HP1064	0.238	HopA	0.007
HP0545	0.164	Mal1434	<0.002*
HP0003	0.083	HP0477	<0.002*
HP0385	0.064	Kor1294N	<0.002*

**OD_{50} did not reach level of OD_{50} (GST-TAG)*

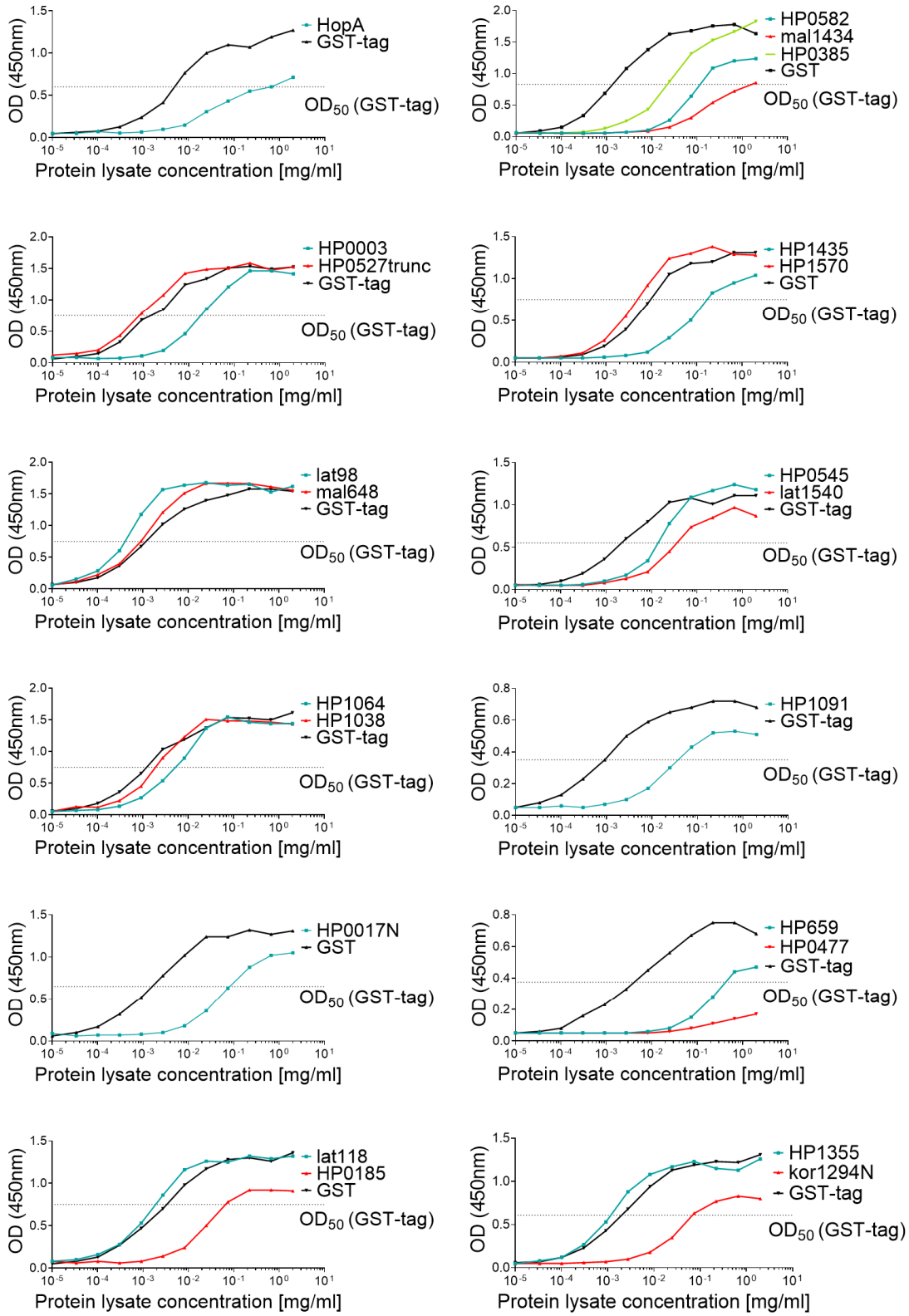


Figure 17: Anti-TAG ELISA

Detection of full-length GST-X-fusion proteins using anti-TAG ELISA. A GST-TAG control, lacking an insert, was used side-by-side to enable relative quantification. Dotted lines show half-maximal absorption (OD_{50}) of the respective GST-TAG control.

3.3.4 Loading recombinant *H. pylori* antigens onto fluorescent beads

For multiplex serology, the 22 new *H. pylori* antigens were loaded onto fluorescent Luminex beads. Additionally, 15 established *H. pylori* antigens (GroEL, UreA, HP0231, NapA, HP0305, HpaA, CagAN, CagAC, HyuAN, HyuAC, Catalase, VacAC, HP1098, Cad, HP1564) and three control antigens from endemic viruses (BK VP1, HPyV6 VP1, EBV VCA p18) were included.

In order to verify successful loading of the GST-X-TAG fusion proteins onto the glutathione-coupled polystyrene beads, the signal of an antibody directed against the C-terminal TAG was measured (Figure 18).
Concordant to the ELISA and Western Blot results, signals from recombinant Mal1434 and HP0477 were very low and barely crossed the lower limit of quantification (50 MFI). The antigens HP0017N, HP0659 and HP1091 also showed reduced signals reaching 8%, 11% and 17% of the GST-TAG control signal, respectively. Their respective anti-TAG Western Blot did also not show any signals and, hence, confirmed a suboptimal expression of full-length GST-X-TAG fusion proteins.

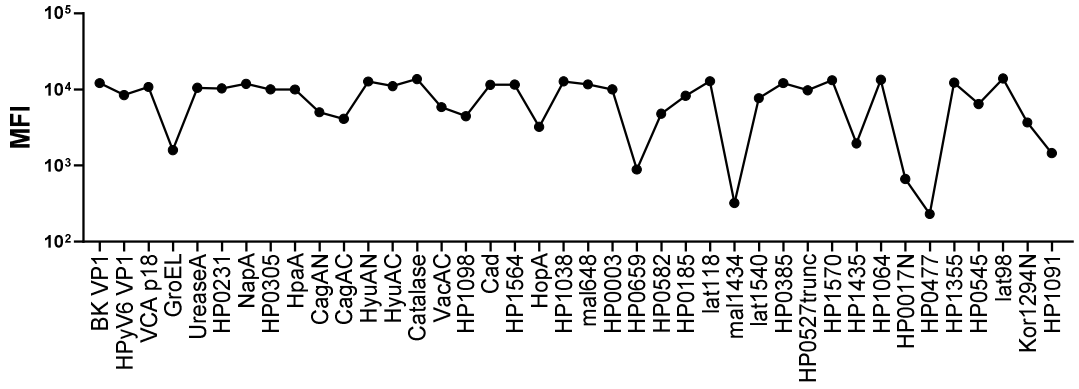


Figure 18: Verification of* H. pylori *antigens on Luminex beads
Detection of GST-X-TAG fusion proteins loaded onto Luminex beads using an antibody against the C-terminal TAG. To enable quantification, a biotinylated secondary anti-mouse antibody was used in combination with the fluorescent reporter conjugate Strep-PE.

3.4 Multiplex serology for new *H. pylori* antigens

3.4.1 Describing distributions, defining cutoffs and determining seroprevalences

Antibodies against 40 antigens were quantified in serum samples from the Shanxi study and NIT by multiplex serology. In contrast to the microarray experiments, all available samples were used (Shanxi: 669; NIT: 649). Additionally, 993 plasma samples from the NIT were available, which were sampled in 1999/2000.

Net MFI readouts against individual antigens were summarized into boxplots, stratified by sample set (Figure 19). Antibody levels were heterogeneous with the highest MFI values deriving from the control antigens VCA p18 (median: 7,790 MFI) and BK VP1 (median: 2,127 MFI), as well as the *H. pylori* antigens CagAN (median: 2,592 MFI), CagAC (median: 1,954 MFI) and HP1564 (median: 1,146 MFI).

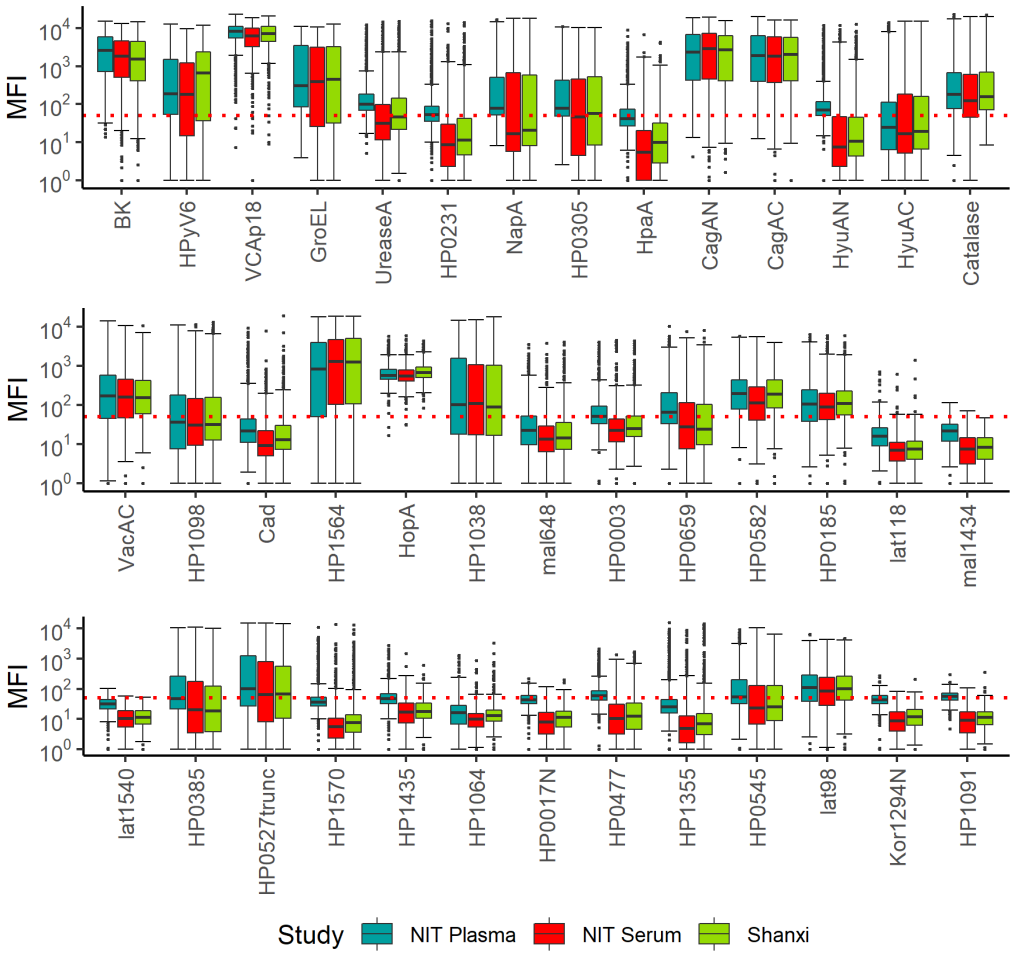


Figure 19: Antibody responses to* H. pylori *and control antigens
Results are stratified by sample set and summarized into boxplots. The dotted red line represents the lower limit of quantification (50 MFI).

Antibody responses against HpaA, HyuAC, HP1098, Cad, Mal648, HP0003, Lat118, Mal1434, Lat1540, HP0385, HP1570, HP1064, HP0017N, HP1355 and Kor1294N were very low with over 50% of the samples falling below the lower limit of quantification (50 MFI).
For these low-titer antigens the difference between sample sets (Shanxi sera, NIT sera or NIT plasma) was very prominent: Over 50% of the antibody measurement against the antigens HP0231, NapA, HyuAN, HP0659, HP0477, HP0545 and HP1091 were below the lower limit of quantification in serum samples, but not in plasma samples. Overall, NIT plasma samples showed consistently elevated MFI values. This systematic trend was observed for most *H. pylori* and control antigens. Hence, I decided to separate the subsequent data processing steps by sample set (Shanxi sera, NIT sera and NIT plasma).

In order to enable statistical analysis, I dichotomized numerical values, which required setting a meaningful cutoff to separate seropositive signals from seronegative signals. Cutoffs, which were used in the previous multiplex serology by Murphy *et al.* were not suitable, as samples had been measured in a lower dilution (1:100, instead of 1:1000).[102] The additional freeze-and-thaw cycle, as well as the long storage time (>10 years), also contributed to decreased MFI signals when comparing the two serology experiments (data not shown).
Hence, I utilized finite mixture modeling to determine cutoffs for antibodies against each *H. pylori* antigen in each sample set. After log transforming the MFI values, I used an expectation maximization (EM) algorithm assuming two mixed skew-normal distributions. Using the example of GroEL, I visualized these distributions by plotting a density histogram and overlaying it with the modeled density mixture function in Figure 20A. The cutoff corresponded to the minimum of this density mixture function.

For some antigens, the EM algorithm could not converge assuming two skew-normal distributions. This was the case if one of the two assumed populations, seropositives or seronegatives, was too small or if the two mixed distributions were overlapping too much. In these cases, I fitted two normal distributions as visualized in Figure 20B using the example of HopA. All cutoffs, assumed underlying distributions and resulting seroprevalences are summarized Table 27.

Seroprevalences of different antigens were highly heterogeneous, as already visualized using numerical MFI values. Low-titer antibodies consequently also showed a low seroprevalence. Antibodies against the *H. pylori* antigens Lat118, Mal1434,

HP1435, HP1064 and Kor1294N were detected in less than 6% of the examined samples. Hence, I excluded them from the subsequent association analyses. Antibodies against Lat1540, HP0017N and HP1091 showed seroprevalences below 2% among serum samples from the Shanxi study and NIT, while seroprevalences in NIT plasma samples reached 18%, 40% and 60%, respectively. As a consequence, I excluded these antigens from the subsequent analyses of the serum samples, but not the NIT plasma samples.

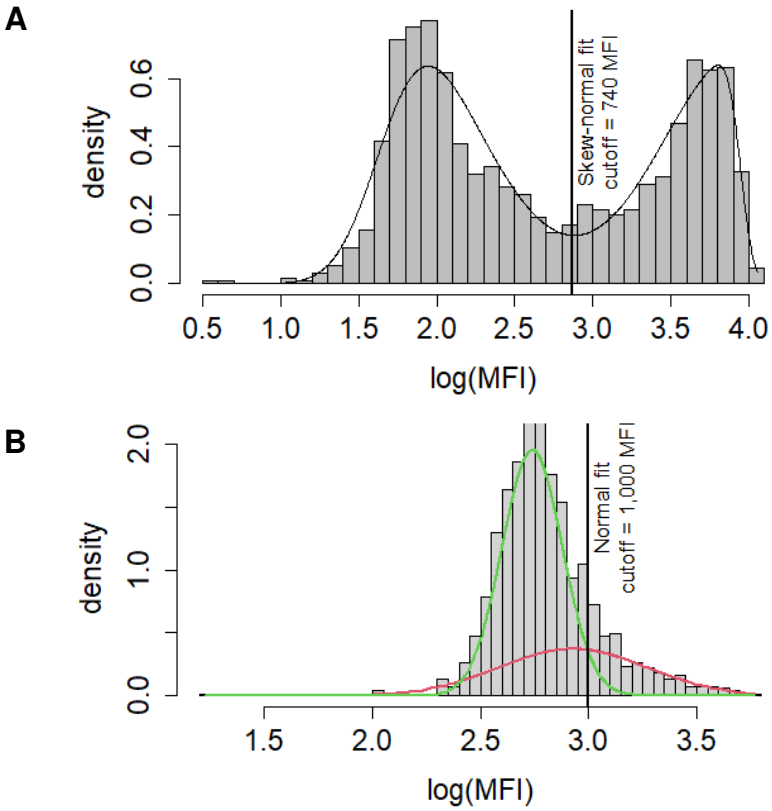


Figure 20: Finite mixture modeling to define cutoffs for seropositivity
A: GroEL antibody responses in NIT plasma. Signal distributions are assumed to be the mix of two skew-normal distributions corresponding to seropositive or seronegative signals. A cutoff (horizontal line) is defined at the minimum of the density mixture function (vertical line). B: HopA antibody responses in NIT plasma. If skew-normal distributions are not applicable, two normal distributions are assumed. In these cases, the cutoff represent the cutpoint of these normal distributions.

For each antigen, I set a minimal cutoff of 50 MFI to account for the quantifiable range of the assay, which subsequently led to decreased seroprevalences of low-titer antibodies. This mainly affected serum samples (Shanxi and NIT sera), due to the elevated MFI signals in NIT plasma.

For the high-titer antibodies against CagAN, CagAC and HP1564, I observed an opposite trend. Although numerical values were in a comparable range, seroprevalences in NIT plasma samples (47%, 46% and 56%, respectively) were decreased compared to Shanxi serum (63%, 65% and 62%, respectively) and NIT serum samples (67%, 65% and 67%, respectively). I made a similar observation for the antigens GroEL with seroprevalences of 47% in Shanxi sera, 49% in NIT sera and 43% in NIT plasma, HP0305 (45%, 50% and 34%, respectively), HP0185 (24%, 25% and 19%, respectively) and HP0545 (31%, 35% and 19%, respectively).
Seroprevalences of antibodies against HyuAN (31%, 35% and 19%, respectively), HyuAC (31%, 35% and 19%, respectively), VacAC (31%, 35% and 19%, respectively) and HP1038 (31%, 35% and 19%, respectively), were elevated in NIT serum samples, while seroprevalences of antibodies against Mal648 (20%, 12% and 26%, respectively), HopA (16%, 11% and 16%, respectively) and HP0003 (20%, 7% and 16%, respectively) were decreased.

Table 27: Seroprevalence, cutoffs and assumed underlying distributions for new H. pylori *multiplex serology antigens*

	Assumed underlying	Shanxi sera		NIT sera		NIT plasma	
Antigen	**distributions**	**cutoff**	**prev [%]**	**cutoff**	**prev [%]**	**cutoff**	**prev [%]**
GroEL	skew-normal	450	47	450	49	740	43
UreaseA	skew-normal	600	14	300**	13	400	12
HP0231	skew-normal	320	7	320	7	320	10
NapA	skew-normal	450	25	180	30	350	27
HP0305	skew-normal	90	45	50*	50	200	34
HpaA	skew-normal	120	11	50*	15	150	13
CagA-N	normal	1150	63	1000	67	2600	47
CagA-C	normal	1050	65	800**	65	2300	46
HyuA-N	skew-normal	180	15	80	20	300	13
HyuA-C	skew-normal	680	18	350	22	900	16
Catalase	skew-normal	1700	19	1700	21	2250	17
VacA-C	skew-normal	150	50	120	57	220	46
HP1098	skew-normal	480	17	900	13	1000	12
Cad	normal	50*	16	95	8	150**	9
HP1564	normal	550	62	340	67	450	56
HopA	normal	1110	16	1150	11	1000	16
HP1038	skew-normal	430	35	250	42	920	29
Mal648	normal	50*	20	85	12	50*	26
HP0003	skew-normal	70	20	130	7	140	16
HP0659	skew-normal	200	17	180	20	400	16
HP0582	normal	1200**	8	1100**	5	1200	7
HP0185	normal	230	24	185	25	350	19
Lat118		50*	1	50*	1	50*	2
Mal1434		50*	0	50*	0	50*	6
Lat1540		50*	0	50*	0	50*	18
HP0385	skew-normal	150	24	170	24	350	22
HP0527trunc	skew-normal	1300	19	1400**	19	1400	24
HP1570	normal	80	7	50*	8	120	7
HP1435	normal	140	1	80	5	180**	4
HP1064	skew-normal	50*	6	50*	3	60	4
HP0017N		50*	1	50*	2	50*	40
HP0477	skew-normal	200	6	400	3	350	6
HP1355	normal	75	11	50*	11	75	13
HP0545	skew-normal	90	31	70	35	350	19
Lat98		50*	72	50*	65	50*	72
Kor1294N		50*	1	50*	1	130**	0
HP1091	skew-normal	50	1	50	1	50	60

**A minimal cutoff of 50 MFI was applied to account for the lower limit of quantification*
***Manually adjusted to meet the visual minimum of the density distribution*
prev = seroprevalence; NIT = nutrition intervention trial;

3.5 Exploring associations between seropositivity to recombinant *H. pylori* antigens and NCGC case status

3.5.1 Association of seropositivity to individual *H. pylori* antigens with NCGC case status in the Shanxi study

The cross-sectional Shanxi study was initiated in 1997 sampling incident NCGC cases, as well as neighborhood controls in the Chinese province Shanxi. A characterization of participants is summarized in Table 16. Out of the 214 available NCGC serum samples, 152 (71%) lacked questionnaire data (age, sex, smoking, etc.). The mean age of the remaining NCGC cases was 53.9 years, while controls were on average 57.6 years. The proportion of smokers was higher in NCGC cases than in controls (73% and 62%, respectively) and the proportion of people never drinking alcohol was smaller (39% and 52%, respectively). Furthermore, more NCGC cases had a higher percentage of males (74%) than controls (69%).

Univariate logistic regression was applied to evaluate the association between seropositivity to individual *H. pylori* antigens and NCGC status based on dichotomized data (Table 28). High seroprevalences were found for antibodies against both CagA fragments (CagAN and CagAC) in sera from NCGC cases (69% and 71%, respectively). On the contrary, only 60% and 61% of healthy controls elucidated these antibodies, respectively, resulting in an OR of 1.42 (95% CI: 1.01 – 2.01) for CagAN and 1.53 (95% CI: 1.06 – 2.19) for CagAC.
A further significant association with NCGC case status was found for antibodies against HP1564 with a seroprevalence of 72% and 58% in NCGC cases and controls, respectively (OR: 1.92; 95% CI: 1.35 - 2.73). Further established *H. pylori* antigens exhibiting a significant association with NCGC case status were GroEL (OR: 1.49, 95% CI: 1.08 - 2.07), NapA (OR: 1.52, 95% CI: 1.06 - 2.19), HP0305 (OR: 1.59, 95% CI: 1.15 - 2.20), HyuAN (OR: 1.81, 95% CI: 1.17 - 2.78) and HyuAC (OR: 1.72, 95% CI: 1.15 - 2.58).

Among the new *H. pylori* antigens, I only found antibodies against HopA to be significantly more prevalent in NCGC cases than in controls (23% and 12%, respectively) with an OR of 2.26 and a 95% CI of 1.48 - 3.46. Presence of antibodies to the new *H. pylori* antigen HP0659 was associated with control status (OR: 0.61; 95% CI: 0.38 – 0.97).

Age and sex are known confounders of NCGC and are usually adjusted for in logistic regression models. However, only 62 (29%) out of the 214 NCGC cases had this information available. Hence, I first repeated univariate logistic regression analysis for these 62 NCGC cases only and adjusted the model for age and sex in a subsequent step (Table 28).

Excluding 71% of the NCGC cases diminished the power of the univariate analyses which was reflected by widened confidence intervals leaving only the association between NCGC cases and antibodies against HP1564 statistically significant (OR: 2.10, 95% CI: 1.15 - 3.82). Point estimates for the other antigens did not change substantially, but lost statistical significance.

Adjusting for age and sex only changed these effect measures marginally with HP1564 still remaining significantly associated with NCGC (OR: 2.10, 95% CI: 1.15 - 3.82).

Table 28: Associations of seropositivity to individual H. pylori *antigens with NCGC status in the Shanxi study*

	Seropositives (%)			**OR (95% CI)**** (sera without questionnaire excluded) n (cases) = 62	**OR (95% CI)***** (sera without questionnaire excluded) n (cases) = 62
Antigens*	**NCGC cases (n = 214)**	**Controls (n = 455)**	**OR (95% CI)****		
GroEL	115 (54)	199 (44)	**1.49 (1.08 - 2.07)**	1.49 (0.87 - 2.54)	1.44 (0.84 - 2.47)
UreA	33 (15)	58 (13)	1.25 (0.79 - 1.98)	1.14 (0.53 - 2.44)	1.16 (0.54 - 2.50)
HP0231	15 (7)	30 (7)	1.07 (0.56 - 2.03)	1.00 (0.34 - 2.94)	0.93 (0.31 - 2.78)
NapA	66 (31)	103 (23)	**1.52 (1.06 - 2.19)**	0.99 (0.52 - 1.87)	1.04 (0.55 - 1.98)
HP0305	113 (53)	188 (41)	**1.59 (1.15 - 2.20)**	1.60 (0.94 - 2.73)	1.63 (0.94 - 2.80)
HpaA	24 (11)	47 (10)	1.10 (0.65 - 1.85)	1.14 (0.49 - 2.65)	1.03 (0.44 - 2.43)
CagAN	147 (69)	276 (61)	**1.42 (1.01 - 2.01)**	1.36 (0.77 - 2.39)	1.24 (0.70 - 2.20)
CagAC	152 (71)	280 (62)	**1.53 (1.08 - 2.18)**	1.42 (0.80 - 2.52)	1.36 (0.76 - 2.42)
HyuAN	44 (21)	57 (13)	**1.81 (1.17 - 2.78)**	1.06 (0.48 - 2.34)	1.10 (0.49 - 2.46)
HyuAC	51 (24)	70 (15)	**1.72 (1.15 - 2.58)**	1.60 (0.84 - 3.06)	1.55 (0.80 - 3.01)
Catalase	43 (20)	84 (18)	1.11 (0.74 - 1.67)	1.18 (0.61 - 2.28)	1.22 (0.62 - 2.38)
VacAC	106 (50)	226 (50)	0.99 (0.72 - 1.38)	0.84 (0.49 - 1.43)	0.88 (0.51 - 1.50)
HP1098	39 (18)	76 (17)	1.11 (0.73 - 1.70)	1.09 (0.54 - 2.19)	1.03 (0.51 - 2.08)
Cad	34 (16)	75 (16)	0.96 (0.61 - 1.49)	0.75 (0.34 - 1.64)	0.81 (0.37 - 1.79)
HP1564	155 (72)	263 (58)	**1.92 (1.35 - 2.73)**	**2.10 (1.15 - 3.82)**	**2.03 (1.11 - 3.72)**
HopA	50 (23)	54 (12)	**2.26 (1.48 - 3.46)**	1.46 (0.70 - 3.06)	1.36 (0.65 - 2.88)
HP1038	81 (38)	150 (33)	1.24 (0.88 - 1.74)	1.05 (0.60 - 1.84)	0.98 (0.56 - 1.73)
Mal648	37 (17)	97 (21)	0.77 (0.51 - 1.17)	0.71 (0.35 - 1.44)	0.72 (0.35 - 1.48)
HP0003	37 (17)	95 (21)	0.79 (0.52 - 1.21)	0.64 (0.30 - 1.34)	0.69 (0.32 - 1.45)
HP0659	27 (13)	87 (19)	**0.61 (0.38 - 0.97)**	0.70 (0.33 - 1.48)	0.47 (0.09 - 2.39)
HP0582	12 (6)	39 (9)	0.63 (0.32 - 1.24)	0.36 (0.08 - 1.53)	0.70 (0.33 - 1.49)
HP0185	43 (20)	119 (26)	0.71 (0.48 - 1.05)	0.49 (0.23 - 1.02)	0.38 (0.09 - 1.62)
HP0385	41 (19)	117 (26)	0.68 (0.46 - 1.02)	0.62 (0.31 - 1.22)	****
HP0527trunc	33 (15)	94 (21)	0.70 (0.45 - 1.08)	0.49 (0.22 - 1.12)	0.63 (0.31 - 1.25)
HP1570	11 (5)	33 (7)	0.69 (0.34 - 1.40)	0.90 (0.31 - 2.63)	0.49 (0.22 - 1.13)
HP0477	11 (5)	30 (7)	0.77 (0.38 - 1.56)	0.71 (0.21 - 2.39)	****
HP1355	23 (11)	50 (11)	0.98 (0.58 - 1.65)	1.01 (0.44 - 2.35)	0.77 (0.22 - 2.62)
HP0545	71 (33)	138 (30)	1.14 (0.81 - 1.61)	0.79 (0.43 - 1.44)	1.03 (0.44 - 2.41)
Lat98	151 (71)	330 (73)	0.91 (0.63 - 1.30)	1.01 (0.56 - 1.84)	0.80 (0.44 - 1.47)

**Excluded due to low seroprevalences: Lat118, Mal1434, Lat1540, HP1064, HP1435, HP0017N, Kor1294N, HP1091 and HP0385*
***unadjusted logistic regression analysis*
****logistic regression analysis adjusted for age and sex*
*****Model did not converge*
Statistically significantly associated are marked in bold; NCGC: non-cardia gastric cancer; OR: odds ratio; CI: confidence interval

3.5.2 Associations between seropositivity to individual *H. pylori* antigens with NCGC case status in the NIT

Study participants of the NIT were enrolled in 1985, each providing a blood sample which was used to isolate serum. In a prior multiplex serology study, samples from 330 study participants who developed NCGC until May 2001 and 330 sex-matched controls were analyzed.[102] For the present study, samples from 322 of the NCGC cases and 327 of the controls were available. Their baseline characteristics are summarized in Table 17. The mean age of participants who developed NCGC cancer was 55.5 years at the time of blood-draw. They were diagnosed on average 7.5 years later. Controls were slightly younger with a mean age of 51.1 years. Both groups comprised 67% males and 37% females. The proportion of smokers was approximately the same between the two groups with 50% current smokers among the controls. The proportion of NCGC cases who drank alcohol was at baseline slightly lower than in controls (24% vs 31%, respectively).

In 1999/2000, all living NIT participants were re-invited for a second survey. From these participants, plasma samples from 113 NCGC cases and 880 controls were available for the current multiplex serology study. Their baseline characteristics at the time of blood-draw are summarized in Table 17. There was no overlap between the serum sample set and the plasma sample set.
Overall, plasma samples derived from older participants than serum samples (63.8 years and 53.3 years, respectively). Consistently, NCGC cases were diagnosed earlier with a mean follow-up time of 3.3 years. Furthermore, they were slightly older than the corresponding control (64.7 years and 63.8 years, respectively).
In this second NIT sample set, NCGC cases and controls were not matched for sex and comprised 55% and 49% males, respectively. The proportion of smokers was again similar between NCGC cases and controls, but overall prevalence of current smokers declined compared to the serum samples which were collected 15 years earlier (32% and 50% among controls, respectively). The proportion of participants drinking alcohol remained approximately the same (31% and 27%, respectively).

Associations between seropositivity against individual *H. pylori* antigens and NCGC case status were analyzed for both NIT sample sets (sera and plasma) applying logistic regression analysis on dichotomized data (Table 29 and Table 30). As serum samples were already matched for sex no adjustment for this variable was necessary. However, since age is an important confounder in the assessment of risk of developing NCGC,

it was included in the model. Plasma samples were not matched on age and sex and therefore I adjusted logistic regression models for these potential confounders, in order to ensure comparability.

For the serum samples, associations between NCGC case status and being seropositive for a respective *H. pylori* antigen were similar to the ones described for the cross-sectional Shanxi study, despite blood samples being drawn 7.5 years before diagnosis.
CagAN, CagAC and HP1564 were found most frequently with a seroprevalence of 73%, 70%, and 72% in NCGC cases and 60%, 61% and 63% in controls, respectively. This corresponded to ORs of 1.89 (95% CI: 1.34 - 2.67), 1.66 (95% CI: 1.18 - 2.33) and 1.65 (95% CI: 1.17 - 2.33), respectively. Furthermore, I also identified an association for antibodies against GroEL (OR: 1.42, 95% CI: 1.03 - 1.96), NapA (OR: 1.85, 95% CI: 1.30 - 2.63), HP0305 (OR: 1.90, 95% CI: 1.37 - 2.62), HyuAC (OR: 1.49, 95% CI: 1.01 - 2.19) and the new antigen HopA (OR: 1.68, 95% CI: 1.01 - 2.81).
Besides HopA, I identified significant associations with antibodies against new *H. pylori* antigens HP0582 and HP0527trunc which were prevalent in 7% and 21% of the NCGC cases, and 3% and 16% of the controls, respectively, exhibiting ORs of 2.32 (95% CI: 1.09 - 4.93) and 1.56 (95% CI: 1.03 - 2.37).

The majority of the NCGC-associated antigens in NIT serum samples were confirmed in the NIT plasma samples, with HP0305 (OR: 1.96, 95% CI: 1.32 - 2.91), CagAC (OR: 1.85, 95% CI: 1.24 - 2.75), HP1564 (OR: 1.95, 95% CI: 1.27 - 2.98) and HopA (OR: 1.79, 95% CI: 1.11 - 2.89) showing statistically significant associations. Additionally, antibodies against VacA were found significant among plasma samples (OR: 1.67, 95% CI: 1.12 - 2.48). Associations of the new *H. pylori* antigens HP0582 (OR: 0.36, 95% CI: 0.11 - 1.16) and HP0527trunc (OR: 1.34, 95% CI: 0.87 - 2.07) with NCGC could not be confirmed in NIT plasma samples.

Logistic regression models adjusted for smoking, alcohol and BMI are summarized in Supplementary Table S5 (serum samples) and Supplementary Table S6 (plasma samples). Point estimates only marginally differed.

Table 29: Associations of seropositivity to individual H. pylori *antigens with NCGC status in the NIT serum samples*

Antigen*	Seropositive NCGC cases (n = 322)	Seropositive controls (n = 327)	OR (95% CI)**
GroEL	169 (52%)	147 (45%)	**1.42 (1.03 - 1.96)**
UreA	43 (13%)	39 (12%)	1.16 (0.72 - 1.87)
HP0231	25 (8%)	23 (7%)	1.08 (0.59 - 1.98)
NapA	117 (36%)	78 (24%)	**1.85 (1.30 - 2.63)**
HP0305	183 (57%)	140 (43%)	**1.90 (1.37 - 2.62)**
HpaA	52 (16%)	43 (13%)	1.24 (0.79 - 1.95)
CagAN	236 (73%)	197 (60%)	**1.89 (1.34 - 2.67)**
CagAC	226 (70%)	198 (61%)	**1.66 (1.18 - 2.33)**
HyuAN	70 (22%)	60 (18%)	1.17 (0.79 - 1.74)
HyuAC	83 (26%)	62 (19%)	**1.49 (1.01 - 2.19)**
Catalase	74 (23%)	64 (20%)	1.20 (0.82 - 1.77)
VacAC	194 (60%)	175 (54%)	1.33 (0.96 - 1.83)
HP1098	43 (13%)	41 (13%)	1.11 (0.69 - 1.79)
Cad	27 (8%)	23 (7%)	0.95 (0.52 - 1.73)
HP1564	233 (72%)	205 (63%)	**1.65 (1.17 - 2.33)**
HopA	45 (14%)	29 (9%)	**1.68 (1.01 - 2.81)**
HP1038	140 (43%)	135 (41%)	1.19 (0.86 - 1.65)
Mal648	39 (12%)	36 (11%)	0.99 (0.60 - 1.63)
HP0003	28 (9%)	20 (6%)	1.44 (0.77 - 2.67)
HP0659	70 (22%)	60 (18%)	1.31 (0.88 - 1.96)
HP0582	23 (7%)	11 (3%)	**2.32 (1.09 - 4.93)**
HP0185	73 (23%)	89 (27%)	0.71 (0.49 - 1.03)
HP0385	75 (23%)	82 (25%)	0.95 (0.66 - 1.38)
HP0527trunc	68 (21%)	53 (16%)	**1.56 (1.03 - 2.37)**
HP1570	29 (9%)	25 (8%)	1.05 (0.59 - 1.87)
HP0477	12 (4%)	7 (2%)	1.69 (0.64 - 4.46)
HP1355	31 (10%)	40 (12%)	0.68 (0.41 - 1.14)
HP0545	116 (36%)	108 (33%)	1.27 (0.90 - 1.77)
Lat98	200 (62%)	219 (67%)	0.79 (0.57 - 1.10)

**Excluded due to low seroprevalences: Lat118, Mal1434, Lat1540, HP1064, HP1091, HP0017N, Kor1294N and HP1435*
***logistic regression analysis adjusted for age*
Statistically significantly associated are marked in bold; NCGC: non-cardia gastric cancer; OR: odds ratio; CI: confidence interval

Table 30: Associations of seropositivity to individual H. pylori *antigens with NCGC status in the NIT plasma samples*

Antigen*	Seropositive NCGC cases (n = 113)	Seropositive controls (n = 880)	OR (95% CI)**
GroEL	53 (47%)	376 (43%)	1.21 (0.81 - 1.79)
UreA	15 (13%)	107 (12%)	1.10 (0.61 - 1.97)
HP0231	7 (6%)	89 (10%)	0.59 (0.26 - 1.30)
NapA	37 (33%)	227 (26%)	1.41 (0.92 - 2.15)
HP0305	54 (48%)	280 (32%)	**1.96 (1.32 - 2.91)**
HpaA	10 (9%)	120 (14%)	0.62 (0.31 - 1.22)
CagAN	57 (50%)	412 (47%)	1.21 (0.81 - 1.80)
CagAC	66 (58%)	389 (44%)	**1.85 (1.24 - 2.75)**
HyuAN	20 (18%)	112 (13%)	1.44 (0.85 - 2.43)
HyuAC	23 (20%)	132 (15%)	1.44 (0.88 - 2.37)
Catalase	20 (18%)	146 (17%)	1.04 (0.62 - 1.74)
VacAC	64 (57%)	392 (45%)	**1.67 (1.12 - 2.48)**
HP1098	7 (6%)	113 (13%)	0.46 (0.21 - 1.01)
Cad	9 (8%)	84 (10%)	0.82 (0.40 - 1.68)
HP1564	79 (70%)	479 (54%)	**1.95 (1.27 - 2.98)**
HopA	27 (24%)	136 (15%)	**1.79 (1.11 - 2.89)**
HP1038	37 (33%)	255 (29%)	1.22 (0.80 - 1.86)
Mal648	29 (26%)	227 (26%)	1.00 (0.64 - 1.57)
HP0003	17 (15%)	146 (17%)	0.92 (0.53 - 1.59)
HP0659	18 (16%)	137 (16%)	1.07 (0.62 - 1.83)
HP0582	3 (3%)	64 (7%)	0.36 (0.11 - 1.16)
HP0185	18 (16%)	167 (19%)	0.85 (0.50 - 1.45)
Lat1540	21 (19%)	160 (18%)	1.05 (0.63 - 1.73)
HP0385	20 (18%)	194 (22%)	0.78 (0.47 - 1.30)
HP0527trunc	33 (29%)	209 (24%)	1.34 (0.87 - 2.07)
HP1570	8 (7%)	63 (7%)	0.97 (0.45 - 2.08)
HP0017N	41 (36%)	356 (40%)	0.84 (0.56 - 1.26)
HP0477	6 (5%)	49 (6%)	1.00 (0.42 - 2.40)
HP1355	16 (14%)	112 (13%)	1.14 (0.65 - 2.02)
HP0545	24 (21%)	169 (19%)	1.18 (0.73 - 1.91)
Lat98	86 (76%)	633 (72%)	1.24 (0.79 - 1.96)
HP1091	73 (65%)	518 (59%)	1.24 (0.82 - 1.86)

**Excluded due to low seroprevalences: Lat118, Mal1434 Kor1294N, HP1435 and HP1064*

***logistic regression analysis adjusted for age and sex*

Statistically significantly associated are marked in bold; NCGC: non-cardia gastric cancer; OR: odds ratio; CI: confidence interval

To address potential reverse causality, I excluded all NCGC cases, which were diagnosed within 2 years after enrollment in a further analysis. The NIT serum sample set was reduced by 38 (12%) cases, while the plasma sample set was reduced by 36 (32%) NCGC cases. This increased the mean follow-up times to 8.4 and 4.4 years, respectively. I repeated the calculations adjusting the serum samples for age and the plasma samples for sex and age. Results for antigens, which showed a significant association in any of the prior analyses are summarized in Figure 21.

ORs in NIT serum samples were overall similar to the previous analysis. However, GroEL and HopA, which were borderline significant using all NCGC samples, lost significance excluding the cases with a follow-up time <2 years, due to wider CIs (OR: 1.38, 95% C I: 0.99 – 1.91 and OR: 1.57, 95% CI: 0.92 – 2.68, respectively).
On the contrary, OR for seropositivity against HP0582 increased and the 95% CI moved further away from the 1, despite widening (OR: 2.55, 95% CI: 1.18 – 5.47). Re-analyzing plasma samples, excluding NCGC cases with a follow-up time <2 years led to CagAN gaining significance (OR: 1.76, 95% CI: 1.09 – 2.85). Other associations between NCGC case status and seropositivity to other antigens, which were found significant before, remained.

Point estimates determined in the plasma samples consistently trended in the same direction as point estimates of the serum samples, except for seropositivity to HP0582. For serum samples, I observed a direct association to NCGC case status (OR: 2.55, 95% CI: 1.18 - 1.46), while seropositivity was inversely associated to NCGC in the plasma samples (OR: 0.35, 95% CI: 0.08 – 1.46).
Seropositivity to HyuAN was significantly associated with NCGC case status in the cross-sectional Shanxi study (OR: 1.81, 95% CI: 1.17 - 2.78). The significance could not be confirmed in the NIT study.

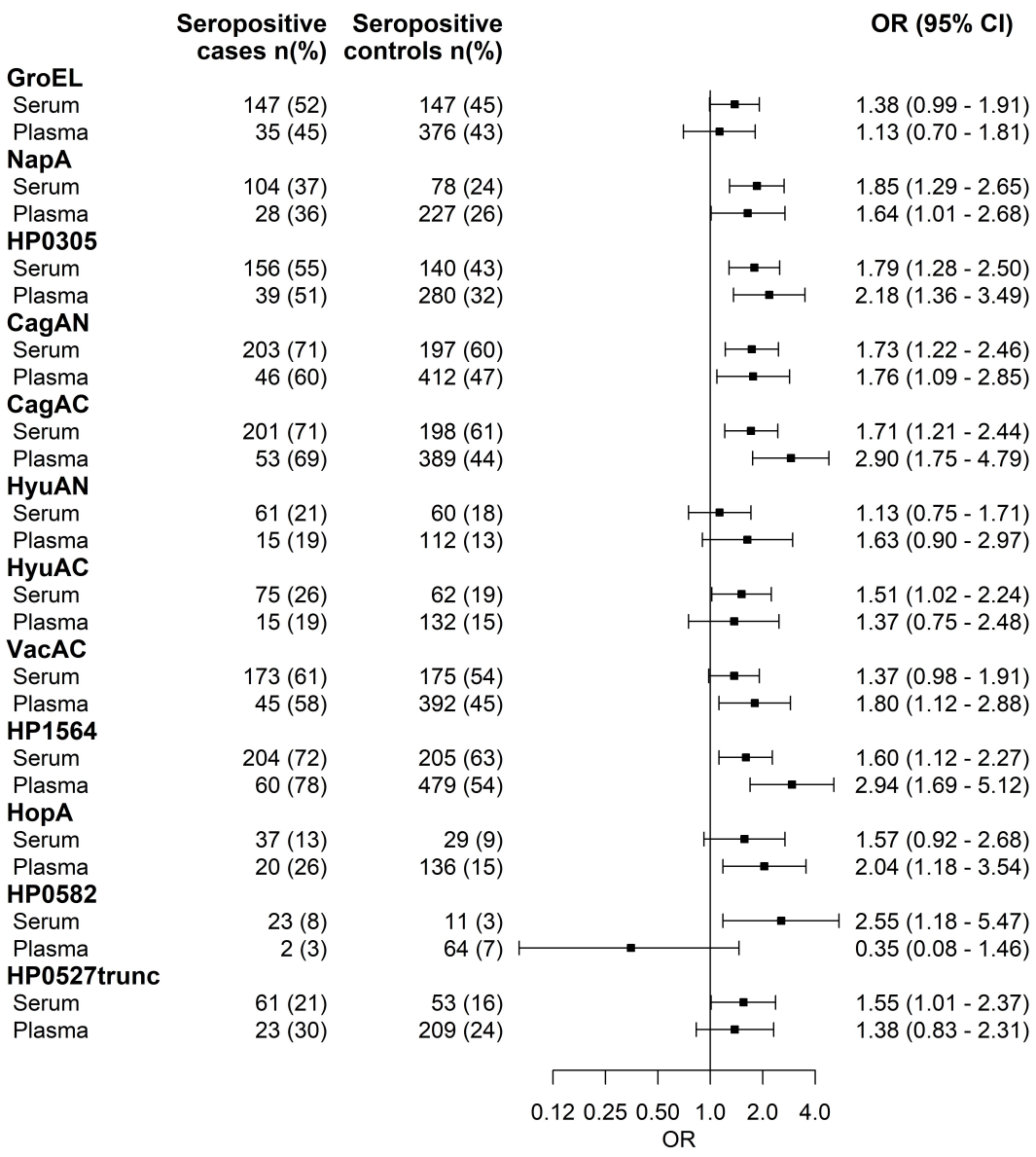


Figure 21: Associations of seropositivity to H. pylori antigens with NCGC status in the NIT excluding cases with a follow-up time >2 years

ORs and 95% CI were determined by logistic regression analyses for selected H. pylori *antigens after excluding NCGC cases with a follow-up time <2 years in NIT serum samples (n = 284 NCGC cases and 327 controls) and NIT plasma samples (77 NCGC cases and 880 controls). Models were sex-adjusted for the NIT serum samples; and age- and sex-adjusted for NIT plasma samples.*

OR = odds ratio, CI = confidence interval

3.5.3 Associations between seropositivity to multiple *H. pylori* antigens with NCGC case status

After exploring the associations between seropositivity to individual *H. pylori* antigens with NCGC case status in the cross-sectional Shanxi study and the two prospective sample sets of the NIT, I evaluated if seropositivity to multiple *H. pylori* outperformed individual antigens.
For this purpose, I used different combinations of antigens which were found significantly associated in the previous analyses (Table 31 and Table 32). First, I used the combination of seropositivity to HP0305 and HP1564 which was proposed by Epplein *et al.*[89] I found antibodies against these two antigens significantly associated with NCGC case status in all three sample sets.

In the Shanxi study, sera which were seropositive for one of the two antigens were more likely to be a NCGC case with an OR of 2.43 (95% CI: 1.55 – 3.79) compared to being seronegative for both antigens. This 'and/or' algorithm generated slightly higher ORs than seropositivity to HP0305 (OR: 1.59, 95% CI: 1.15 - 2.20) or HP1564 (OR: 1.92, 95% CI: 1.35 - 2.73) individually. Being simultaneously seropositive for both antigens did not increase the OR further (OR: 2.20, 95% CI: 1.47 – 3.29). Furhtermore, all of the 95% CI overlap.
Among NIT serum samples, the OR for seropositivity to one of the two antigens (OR: 1.60, 95% CI: 1. 02 - 2.49) was higher than being seronegative for both antigens. Nevertheless, the point estimate was lower than for individual seropositivity to HP0305 (OR: 1.90, 95% CI: 1.37 - 2.62) or HP1564 (OR: 1.65, 95%: 1.17 - 2.33).
Observations among the NIT plasma samples were accordingly, but being seropositive to both antigens showed an increased OR compared to the individual associations in both sample sets (OR: 2.14, 95% CI: 1.46 – 3.15 and OR: 2.44, 95% CI: 1.51 – 3.94, respectively). However, again all of the 95% CI are overlapping.

In a next step, I added the two CagA fragments (CagAN and CagAC) to the combination of HP0305 and HP1564. CagA is the most prominent *H. pylori* virulence factor and has been discussed as a potential antigen candidate for GC screenings.
I found the point estimates of associations very similar compared to using only HP0305 and HP1564: in the Shanxi study, the highest category (being positive for three or four of the antigens) showed only marginally higher point estimates for NCGC case status than the highest category using HP1564 and HP0305 only (OR: 2.53, 95% CI:

1.59 - 4.01 and OR: 2.20, 95% CI: 1.47 – 3.29, respectively). Observations for the NIT serum and plasma samples were similar.

Stronger associations were observed considering all *H. pylori* antigens which were found to be individually associated with NCGC. For the Shanxi samples these were the antigens GroEL, NapA, HP0305, CagAC, CagAN, HyuAN, HyuAC, HP1564 and HopA. Again, I created three categories with seropositivity against up to one antigen serving as the reference category. Being seropositive against two to four of these antigens increased the strength of the association to NCGC status to an OR of 3.16 (95% CI: 1.83 – 5.46). Seropositivity against more than four antigens increased it even further to an OR of 3.46 (95% CI: 2.01 – 5.97). These effect measures were higher than for individual antigens, but 95% CI overlapped again.

Results were similar for the NIT serum samples and plasma samples using the respective significantly associated antigens. The category which was based on being seropositive for most of the antigens (six to ten or four to five, respectively) consistently showed the strongest association to NCGC case status (OR: 3.39, 95% CI: 2.03 – 5.68 and OR: 2.64, 95% CI: 1.62 – 4.30, respectively).

In a last approach, I calculated OR using the only four *H. pylori* antigens which were consistently found associated in all three sample sets: HP1564, HP0305, CagAC and HopA. Logistic regression models based on these four antigens reached similar point estimates as using HP0305 and HP1564 only or using the combination of HP0305, HP1564, CagAN and CagAC.

Table 31: Associations of simultaneous seropositivity to multiple H. pylori *antigens with NCGC status among Shanxi serum samples and NIT serum samples*

Shanxi serum samples

	Seropositive antigens (n)	NCGC cases (n = 214)	Controls (n = 455)	OR (95% CI)*
HP0305, HP1564	0	47 (22%)	178 (39%)	REF
	1	66 (31%)	103 (23%)	2.43 (1.55 – 3.79)
	2	101 (47%)	174 (38%)	2.20 (1.47 – 3.29)
HP0305, HP1564, CagAN, CagAC	0	28 (13%)	121 (27%)	REF
	1-2	51 (24%)	103 (23%)	2.14 (1.26 – 3.64)
	3-4	135 (63%)	231 (51%)	2.53 (1.59 – 4.01)
All antigens found significantly associated with NCGC***	0-1	19 (9%)	111 (24%)	REF
	2-4	93 (43%)	172 (38%)	3.16 (1.83 – 5.46)
	5-9	102 (48%)	172 (38%)	3.46 (2.01 – 5.97)
HP0305, CagAC, HP1564, HopA	0	30 (14%)	131 (29%)	REF
	1-2	79 (37%)	148 (33%)	2.33 (1.44 – 3.77)
	3-4	105 (49%)	176 (39%)	2.61 (1.64 – 4.15)

NIT serum samples

	Seropositive antigens (n)	NCGC cases (n = 322)	Controls (n = 327)	OR (95% CI)**
HP0305, HP1564	0	74 (23%)	116 (35%)	REF
	1	80 (25%)	77 (24%)	1.60 (1. 02 - 2.49)
	2	168 (52%)	134 (41%)	2.14 (1.46 – 3.15)
HP0305, HP1564, CagAN, CagAC	0	45 (14%)	84 (26%)	REF
	1-2	64 (20%)	70 (21%)	1.69 (1.01 – 2.82)
	3-4	213 (66%)	172 (53%)	2.52 (1.64 – 3.87)
All antigens found significantly associated with NCGC****	0-1	33 (10%)	74 (23%)	REF
	2-5	172 (53%)	166 (51%)	2.49 (1.54 – 4.02)
	6-10	117 (36%)	86 (26%)	3.39 (2.03 – 5.68)
HP0305, CagAC, HP1564, HopA	0	50 (16%)	88 (27%)	REF
	1-2	114 (35%)	114 (35%)	1.79 (1.14 – 2.80)
	3-4	158 (49%)	124 (38%)	2.46 (1.59 – 3.80)

**unadjusted logistic regression model*
***logistic regression models were adjusted for age*
****GroEL, NapA, HP0305, CagAC, CagAN, HyuAN, HyuAC, HP156 and HopA;*
*****GroEL, NapA, HP0305, CagAC, CagAN, HyuAN, HP1564, HopA, HP0582, HP0527trunc*
NCGC = non-cardia gastric cancer; OR = odds ratio; CI = confidence interval; NIT = nutrition intervention trial; REF = reference

Table 32: Associations of simultaneous seropositivity to multiple H. pylori *antigens with NCGC status among NIT plasma samples*

NIT plasma samples

	Seropositive antigens (n)	NCGC cases (n = 113)	Controls (n = 880)	OR (95% CI)*
HP0305, HP1564	0	30 (27%)	383 (44%)	REF
	1	33 (29%)	235 (27%)	1.81 (1.07 – 3.05)
	2	50 (44%)	262 (30%)	2.44 (1.51 – 3.94)
HP0305, HP1564, CagAN, CagAC	0	23 (20%)	312 (35%)	REF
	1-2	30 (27%)	213 (24%)	1.90 (1.07 – 3.36)
	3-4	60 (53%)	352 (40%)	2.39 (1.44 – 3.96)
All antigens found significantly associated with NCGC**	0-1	28 (25%)	388 (44%)	REF
	2-3	34 (30%)	211 (24%)	2.21 (1.30 – 3.75)
	4-5	51 (45%)	276 (31%)	2.64 (1.62 – 4.30)
HP0305, CagAC, HP1564, HopA	0	17 (15%)	245 (28%)	REF
	1-2	37 (33%)	313 (36%)	1.67 (0.91 – 3.05)
	3-4	59 (52%)	317 (36%)	2.72 (1.54 – 4.79)

**logistic regression models were adjusted for age and sex*
***HP0305, CagAC, VacAC, HP1564 and HopA*
NCGC = non-cardia gastric cancer; OR = odds ratio; CI = confidence interval; NIT = nutrition intervention trial; REF = reference

3.6 Exploring numerical *H. pylori* multiplex serology results to predict NCGC

So far, association analyses were based on knowledge-driven approaches, especially combining seropositivity to multiple antigens. In a last step, I explored the numerical *H. pylori* multiplex serology readouts in an unbiased fashion to investigate potential underlying patterns and to better understand interdependencies between antigens. In order to prevent statistical noise at the lower level, I changed all values below the lower limit of quantification (50 MFI) to 50 MFI.

In the following, I explored the numerical readouts for the NIT serum samples. First, I performed a principle component analysis (PCA) to the scaled data to reduce complexity. This technique is used to remove redundancies from large data sets while keeping variation, summarized into multiple uncorrelated principle components.
The first two principle components accounted for 15.8% and 6.1% of the total variation of the dataset. Color-coding NCGC cases and controls, I observed that the two groups were not distinguishable in any of the dimensions, except for the first one. However, the corresponding 95% CI ellipses substantially overlapped (Figure 22). Antigens which correlate to this first dimension are visualized in Figure 23. The strongest correlations were found for the *H. pylori* antigens CagAC, HP1564, CagAN and GroEL.

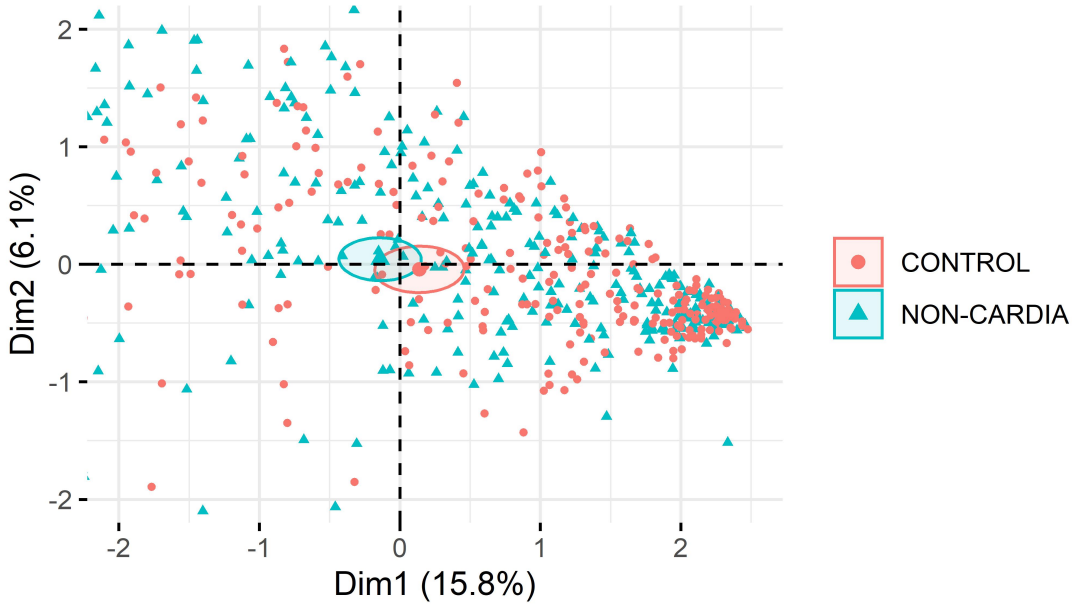


Figure 22: PCA for NIT serum samples
First and second dimension of PCA using NIT serum samples. The figure shows only an excerpt in order to visualize 95% confidence ellipses.

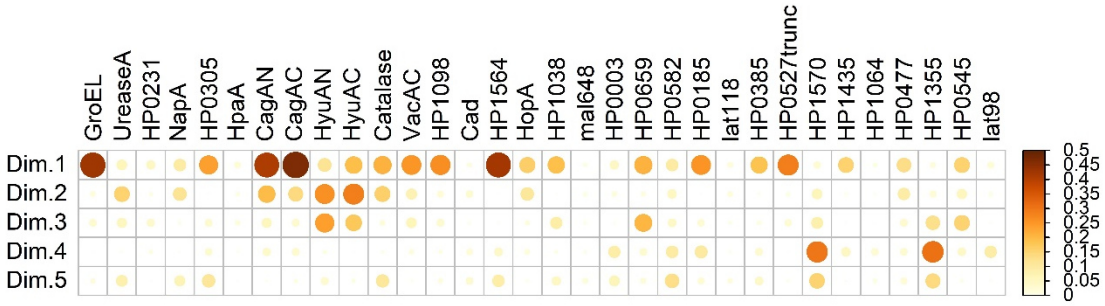


Figure 23: Correlation between antibody responses and principle components

An algorithm could be trained based on the numerical data to potentially distinguish NCGC cases from controls. To explore this further, I split the multiplex serology results of the NIT serum samples into a training and a test set (80% and 20% of the individuals, respectively) to prevent information leakage.

Using the training set, I built a random forest classification model supplying the numerical data and the corresponding class (NCGC case or control). Antigens with the highest relative contributions to this classification model are visualized in Figure 24. While CagAC was the 'important' antigen for this model, CagAN was only on 10th position. HopA was also not among the most important values (12th position). Instead, the antigens HP0305, NapA, as well as the new antigen HP0003, were considered particularly important.

I repeated analyzing seropositivity to multiple antigens using the four 'most important' antigens from the random forest model (Table 33). Exhibiting antibodies against three or four of these antigens showed an over 3-fold increase in odds to develop NCGC (OR: 3.21, 95% CI: 1.98 – 5.21). This combination generated higher OR compared to individual antigens and to other combinations which were created knowledge-driven.

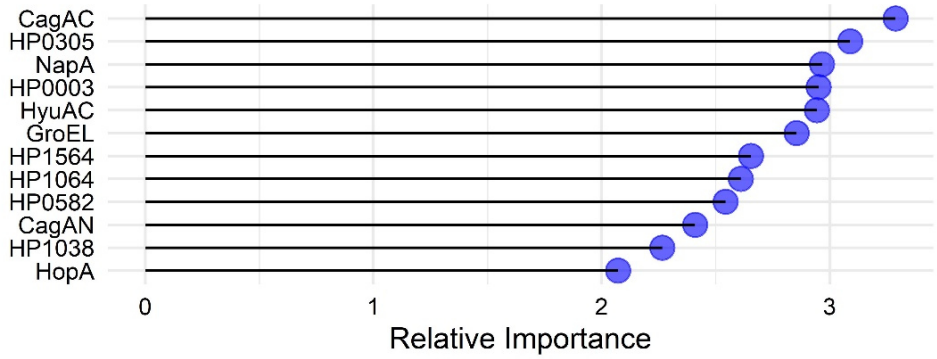


Figure 24: Relative importances of* H. pylori *antigens in a random forest model
A random forest classification model was trained with 80% of the NIT serum samples; Relative importance values of the top twelve antigens are shown.

Table 33: Associations of simultaneous seropositivity to selected H. pylori *antigens among NIT serum samples*

	Seropositive antigens (n)	NCGC cases (n = 322)	Controls (n = 327)	OR (95% CI)*
CagAC, HP0305, NapA and HP0003	0	24	350	REF
	1-2	66	381	2.35 (1.57 – 3.51)
	3-4	33	146	3.21 (1.98 – 5.21)

** adjusted for sex and age*
NCGC = non-cardia gastric cancer; OR = odds ratio; CI = confidence interval; REF = reference

Only the combination based on the ten individually associated antigens showed slightly higher ORs when being seropositive against at least six of them (OR: 3.39, 95% CI: 2.03 - 5.68). However, all of the respective 95% CI are again overlapping.
Using the same combination (CagAC, HP0305, NapA and HP0003) to analyze combined seropositivity the Shanxi samples and in the NIT samples showed an OR of 2.21 (95% CI: 1.28 – 3.80) and 2.30 (95% CI: 1.26 – 4.20), respectively, when being seropositive for at least three of the antigens.

As I initially only used 80% of the NIT serum samples to train the random forest model, I was able to determine a sensitivity of 59% and a specificity of 52% among the remaining 20% (Table 34). The same model predicted case-control status for the Shanxi serum samples and the NIT plasma samples with sensitivities of 59% and 64%, respectively, and specificities of 54% and 56%, respectively.
I also performed the explorative analyses for the Shanxi sera and the NIT plasma samples (data not shown). Overall, the capacity to distinguish NCGC cases from controls showed similar results. The four 'most important' antigens for the Shanxi sera were HopA, CagAC, HP1564 and HP0185, while VacAC, HopA, HyuAN and NapA were determined for NIT plasma samples.

Table 34: Sensitivities and specificities of a random forest classification model

Training data	NIT sera test set*	Shanxi serum samples	NIT plasma samples
NIT sera train set (80% of the samples)	52% spec 59% sens	54% spec 59% sens	56% spec 64% sens

** 20% of the NIT sera, which was not used to train the random forest model*
spec = specificity; sens = sensitivity; NIT = nutrition intervention trial

4 Discussion

Infection with the bacterium *H. pylori* is the major cause for GC and particularly NCGC. Although global infection rates have been declining for decades, the bacterium is still highly prevalent in some parts of the world. Overaging and population growth are reasons why absolute numbers of GC are rising. To date, GC is still the fifth most common cancer globally, while ranking fourth for cancer-related deaths.[1, 108]

To enable or facilitate secondary prevention of GC, pre-defining a high-risk population eligible for endoscopy has been proposed. To achieve this pre-selection, serology is a favored method, because it is only minimally-invasive and can be conducted cost-efficiently in a high-throughput manner.

A promising approach is using antibodies against *H. pylori,* the major cause of GC. In countries with high *H. pylori* prevalence rates, however, a mere detection of the bacterium itself is not enough to reduce the number of people eligible for endoscopy. A further risk stratification is desirable, e.g. by addressing additional *H. pylori* antigens, which are associated with GC development.

During my thesis I conducted an unbiased *de novo* approach to identify NCGC-associated antibodies against specific *H. pylori* antigens. For this purpose, I adapted a technique to generate *H. pylori* microarrays presenting a multitude of antigens in parallel. A pilot study was conducted probing minimized *H. pylori* 26695-microarrays displaying 245 antigens with sera from NCGC patient and healthy controls from the MCC Spain study. Thereby, I confirmed the well-known association between NCGC and antibodies against CagA. Furthermore, I identified a new association with antibodies against the outer membrane protein HopA.

In a next step, I expanded the antigen repertoire of *H. pylori* microarrays to a total of 1,833 antigens by including further *H. pylori* strains of different origins to account for the genomic variety and phylogeographic pattern of the bacterium. *H. pylori* multi-strain microarrays were probed with sera (NCGC cases and healthy controls) from the cross-sectional Shanxi study and the prospective NIT, two Chinese studies from areas with high GC incidence rates.

After optimizing signal acquisition, I selected 22 antigens which were found most informative to classify sera as a NCGC case or healthy control. These antigens were expressed as GST-X-TAG fusion proteins and transferred to multiplex serology for further validation in a low-density, but high-throughput manner. After determining antibody responses to the 22 new and 15 established *H. pylori* antigens in three

sample sets (Shanxi serum samples, NIT serum samples and NIT plasma samples), I evaluated their associations (individual and combined) with NCGC status.
Seropositivity to HopA was confirmed to be associated with NCGC and showed comparable point estimates as established serological markers, including CagA. Combining seropositivity to multiple antigens did not substantially increase the association to NCGC in a Chinese population with an overall high *H. pylori* prevalence. Despite clear trends, numerical analysis indicated that NCGC cases cannot be sufficiently discriminated from controls using multiplex serology readouts only. Additional studies are required to evaluate the performance of *H. pylori* antigens in combination with other clinical biomarkers, e.g. PGI and PGII.

4.1 *On-chip* expression of *H. pylori* antigens

4.1.1 Advantages and limitations

The generation of *H. pylori* antigen microarrays was based on a double spotting technique, which was originally developed for the seroepidemiological research on *C. trachomatis*.[91] Briefly, DNA expression constructs encoding the antigen of interest, were spotted individually on epoxysilane coated slides and subsequently overlaid with spots containing cell-free protein expression kit. During a following incubation period, DNA expression constructs were transcribed and translated into the respective protein directly on the chip. Each of these antigens was flanked by an N-terminal V5- and a C-terminal 6xHis-tag to enable verification of successful protein expression.[91]
This *on-chip* expression technique obviates the need for conventional protein expression, purification and subsequent spotting saving time and resources. Within the frame of this thesis, these microarrays enabled a unique high-density representation of 1,833 *H. pylori* antigens in parallel allowing to screen for novel serological biomarkers in a virtually unbiased fashion.

These advantages are, however, accompanied by some limitations. A major disadvantage is that *on-chip* expressed proteins cannot be characterized apart from detecting their terminal tags and verifying full-length or N-terminal expression. These signals can only be interpreted in a dichotomous fashion, as they do not necessarily correlate with the amount of translated antigens, because a high expression can also cause a suboptimal accessibility of the terminal tags (oral communication with Dr. Hufnagel).

Another major drawback is that a correct folding of the proteins cannot be verified. Although this is considered a general problem of recombinant protein expression, it particularly affects cell-free expression systems.[109] Physiologically, proteins have a designated secondary and tertiary structure, but sometimes require chaperones, other co-factors or the milieu of a specific cell compartment (e.g. periplasm) to fold correctly.[110] Especially integral membrane proteins are prone to deviate from their natural conformation when expressed under non-physiological conditions. There are approaches to account for this, e.g. adding membrane-mimicking structures.[109] This is, however, not feasible in a high-density approach. Therefore, most likely a certain proportion of the antigens presented on the microarray were not exhibiting their natural conformation. However, this does not imply that they were not immunogenic. Regarding membrane proteins, which were among the immunogenic proteins in this project, epitopes are physiologically located at exposed domains and as long as these domains are still accessible for antibodies, a misfolded hydrophobic domain would not interfere with detecting an antibody response.

An additional limitation is that *on-chip* expressed antigens were not purified prior to being probed with patient serum. I used slides which were coated with epoxysilane to generate the microarrays. Epoxysilane is a substrate that unspecifically binds numerous kinds of molecules, including DNA expression constructs and components of the cell-free *E. coli* expression kit. This means, that *on-chip* expressed *H. pylori* antigens were not presented as an individual target, but as part of a complex mixture. To account for off-targets, sera were pre-incubated with *E. coli* BL21 lysate to adsorb antibodies, which would otherwise bind to components of the expression kit.
However, it is possible that the *E. coli* BL21 did not include all potential off-targets causing some antigens to bind to components of the expression kit. Presumably, this majorly contributes to the patient-to-patient variation of the background and optimization of the serum pre-incubation buffer should be addressed for future array developments.

4.1.2 Reproducibility and upscaling

Developing the *on-chip* expression technique for *C. trachomatis*, Dr. Hufnagel reported a Pearson correlation coefficient of 0.77 staining terminal tags of *on-chip* expressed antigens on two slides of the same production batch.[91] I confirmed this on the minimized *H. pylori* 26695-microarrays, determining correlation coefficients between

0.76 and 0.98. The correlation for *H. pylori* multi-strain microarrays was similar with a mean r of 0.83 within a batch and a mean of 0.77 between batches.

Despite this good correlation, some limitations became apparent when increasing the size of minimized *H. pylori* microarrays to a >4-fold larger area and upscaling the number of tested sera from 116 (one spotting batch) to 354 (18 spotting batches). During the development, I spotted test batches using a single randomly chosen DNA expression construct (HP0395) and negative controls for all 2,001 spots. I found that 5% of the signals were randomly missing without an error report from the microarray spotter. I assume that this also affected *H. pylori* microarrays introducing a bias, as it was not possible to check for these missing spots for each individual slide. An additional 4% of the spots were missing on average per production batch, which was documented by the spotter. These spots were equally missing on all microarrays of a batch and hence introduces a bias towards the null, as NCGC cases and respective controls were always tested on subsequently spotted microarrays.

In the same experiment I also assessed intra-slide variability. Quantification of terminal tags revealed a high coefficient of variation of 56% for raw MFI readouts. High variability is a general problem of protein microarrays, which is driven by differences in spotting processes, uneven coating or uneven probing.[106] Using *on-chip* protein expression added another component to the total variation.
As an example, I observed a spotting gradient which was presumably caused by accumulation of expression kit at the print head pin of the microarray spotter.[111] The gradually changed concentration led to a gradually decreased protein expression towards the end of the printing process. For the minimized *H. pylori* microarrays, I did not observe this effect, as the number of spots was much smaller (272 vs. 2,001, respectively). An experimental solution for this problem would be rinsing the print-head within the spotting of a single microarray, which would introduce another type of variation. Instead, I corrected the gradient on an analytical level by subtracting the local background, as it has been proposed for other protein microarrays which encountered similar problems.[104, 106]
On average, this background correction reduced the SD of negative controls by 25%. Most importantly, local background corrected values showed a better correlation with multiplex serology readouts than fold-change values derived from global normalization. Although I still consider the variance on microarrays rather high, this background correction showed promise to enable selecting antigens for multiplex serology.

4.2 Identifying informative *H. pylori* antigens

4.2.1 Addressing the phylogeographic distribution of *H. pylori*

Numerous studies have been conducted to identify the cause for varying GC rates in different parts of the world. To some extent, this variation can be attributed to the virulence of the locally predominant *H. pylori* strains.[46, 112] In particular, CagA and VacA variants were found to determine the virulence of a strain.[113, 114] However, there is no published assay which is able to distinguish genetic variants serologically. Hence, additional antigens might be informative to characterize the virulence of an *H. pylori* strain using serology.

For example, seropositivity to the hypothetical protein HP0305 and the lipoprotein HP1564 were found to be associated with increased risk for GC in an Asian population. Double-seropositivity to these antigens showed an OR of 3.34 (95% CI: 2.27–4.91) compared to being seronegative for both.[88, 115] The OR was even higher for precancerous lesions (OR: 7.43, 95% CI: 5.59–9.88).[89]

It is well possible that further undiscovered risk markers exist, especially with about 23% of the *H. pylori* proteins still not being functionally characterized.[107]

The plasmid-encoded protein pGP3, for example, is an infection marker for *C. trachomatis*, and was only recently discovered and validated. It outperforms the previously used assays, which were based on major outer membrane proteins.[116] Likewise, antibodies against the EBV protein LF2 were only recently discovered to be strongly associated with an increased risk to develop nasopharyngeal cancer, because the LF2 open reading frame is lacking from the EBV reference strain B95-8.[117, 118]

As *H. pylori* exhibits a particularly high intra-species genetic diversity, potentially informative antigens might not be encoded on the common reference strain 26695, especially considering that it was originally isolated from a European gastritis patient without any signs of GC.[103, 119]

Therefore, I addressed the genetic diversity of *H. pylori* by complementing the original 26695-microarray with antigens from other strains. Regarding *H. pylori,* genetic distance corresponds to geographic distance. Therefore, I aimed to access strains from different locations, ideally one representative for each *H. pylori* clade.[120, 121] For this purpose, I was kindly supported by Dr. M.C. Camargo who provided four strains from GC patients isolated in countries with high GC prevalence rates (Latvia, Malaysia,

Chile and Korea), which potentially increased the chance to identify GC-associated antibodies.

Despite the high genetic variance of *H. pylori*, these strains shared many antigens due to homology. It would have been redundant to present all of them on the *H. pylori* multi-strain microarrays, because serological assays are generally prone to cross-reactivity. This can be explained by the vast majority of antigens exhibiting multiple epitopes. If a certain degree of sequence similarity exists, at least some epitopes are shared, causing cross-reactive antibody responses. Therefore, I clustered all of the encoded proteins with an amino acid identity above 70%, according to observations between cross-reactive allergens.[122]
Eventually, all antigens were assigned to 1,833 clusters. Antigens from *H. pylori* 26695 were present in 1,387 of them, while 446 clusters did not have a homologue encoded by *H. pylori* 26695. Out of the 1,833, 328 clusters represented strain-specific antigens, which could potentially be specific for the region they were isolated from.
Regarding the *H. pylori* strains included on the multi-strain microarray, Chinese isolates would presumably be most closely related to the Korean and Malaysian strains. Antibody responses against the respective strain-specific antigens from KOR-037 and MAL-007 was, however, not particularly high.

The most recent publication comparing *H. pylori* genomes was published in 2017 and was based on 346 genomes. Van Vliet reported 2,252 *H. pylori* clusters using a cutoff of 60% (results for 70% deviated only marginally).[123] Until now, the number of *H. pylori* genomes found in the RefSeq database increased to 2,362, each encoding approximately 1,400 proteins. This presumably means that that the number of *H. pylori* clusters also increased due to strain-specific antigens.[124, 125]
In summary, using five strains to generate *H. pylori* microarrays is likely an underrepresentation given the increasing number of sequenced genomes in the last years. Nevertheless, the *H. pylori* multi-strain microarray enables parallel analysis of the highest number of non-redundant antigens so far. Adding further antigens would most probably barely increase the chances to find a sensitive marker, as these require a certain overall prevalence in multiple rather than a single cancer-associated strain.

4.2.2 Selecting informative antigens from minimized microarrays

In order to identify *H. pylori* antigens which could be used as risk markers for NCGC, I performed two major experiments: Probing minimized *H. pylori* 26695-microarrays with sera from the MCC Spain study and subsequently probing *H. pylori* multi-strain microarrays with Chinese sera from the Shanxi study and the NIT.

I started with minimized *H. pylori* 26695-microarrays to show that the technique is generally compatible with *H. pylori* by replicating known GC-markers. With an OR of 3.64 (95% CI: 1.69 - 7.87), I found anti-CagAN to be significantly associated with NCGC based on 58 cases and 58 controls from the MCC Spain study. This confirmed results from a previous multiplex serology experiment reporting an OR of 3.65 (95% CI: 2.44 - 5.46) for 200 NCGC cases and 2,052 controls from the MCC Spain study.[99] Despite a wider 95% CI, point estimates were very similar.
Seroprevalences, however, differed between the two approaches. While Fernández de Larrea-Baz *et al*. reported a seroprevalence of 79% in NCGC cases and 40% in controls, for anti-CagA antibodies, I determined a seroprevalence of 66% and 31%, respectively.[99] This discrepancy was most probably caused by using the median + five MAD of all negative controls as a cutoff. Transferred to a normal distribution, setting a cutoff at the mean plus 5 SDs would correspond to an one-in-a-million chance that a finding is random. This indicates that the applied cutoffs were extremely stringent, virtually eliminating type 1 errors (false positives). Thus, low- and medium-titer antibodies were filtered out leaving only high-titer antibodies to be analyzed.
The advantage of high-titer antibodies is their transferability to clinical settings, as serological assays do not need to be particularly sensitive to detect them.

Further analyzing microarray results for the MCC Spain study, another *H. pylori* antigen which was significantly more prevalent in NCGC cases than controls (22% and 7%, respectively) was HopA. With an OR of 4.58 (95% CI: 1.23 – 17.01), the point estimate was even higher than for CagA antibodies (OR: 3.64, 95% CI: 1.69 - 7.87). However, the wider 95% CI also indicated a lower precision.
HopA is a monomeric protein that belongs to the Hop family. It is membrane-embedded and forms a porine to enable passive diffusion.[56] Prior to this study, HopA has not been described in the context of cancer.[100] However, other members of the Hop family are involved in NCGC development, e.g. BabA, OipA, SabA and HopQ.[56-59, 126]
The expression of HopA was found to correlate with an increasing pH, which is characteristic during the development of GC and caused by a reduced rate of basal

gastric acid secretion.[127, 128] This could suggest that antibodies against HopA are more likely to be found in individuals who already have NCGC or are in the process of developing it. Selecting HopA to be expressed as a recombinant antigen for multiplex serology enabled measuring antibody responses in a larger number of samples and further investigate this hypothesis by comparing results for cross-sectional and prospective studies.

4.2.3 Selecting informative antigens from multi-strain microarrays

No further *H. pylori* antigen besides CagA and HopA was found significantly associated with NCGC status in the MCC Spain study using microarrays. Comparing results for shared antigens with multiplex serology, microarrays exhibited a high specificity, but only a moderate sensitivity. Therefore, I decided to use a lower cutoff for the subsequent *H. pylori* multi-strain microarray experiments, to also consider potential low- and medium-titer antibodies. By using a pre-determined 95% specificity, I allowed a type I error of 5%. This means, that chance findings per microarray (and thereby patient serum) were expected. Post-hoc comparisons with multiplex serology confirmed an increase in sensitivity, but also a moderate decrease in specificity.

For the selection of NCGC-associated antigens from *H. pylori* multi-strain microarray experiments, I evaluated results of Mann-Whitney U tests for numerical data, and chi-squared tests for dichotomized data.
I found that most antigens, which turned out significantly different in any of the statistical tests, showed low seroprevalences among NCGC cases or were inversely associated with NCGC status. My previous findings, however, suggested that antibodies with a low seroprevalence were only poorly correlated with multiplex serology readouts, both on the minimized 26695-microarrays and on the multi-strain microarrays. This particularly affected antigens with a seroprevalence <20%. Furthermore, antigens with low seroprevalences can per definition only reach suboptimal sensitivity, which is not desirable for a serological marker. However, this could potentially be compensated using multiple antigens in combination.

Altogether, the results did not provide clear lead antigens discriminating NCGC cases and controls. This confirmed two other studies which were meanwhile published. Song *et al*. also used antigen microarrays, while Soluri *et al.* performed phage display to identify new serological *H. pylori* markers. Except for antibodies against HP0527 (CagD) found by phage display, only inverse associations with NCGC were reported

as statistically significant in both studies.[107, 129] As both groups used cross-sectional case-control studies, this decreased seroprevalence in GC cases could potentially be caused by the hypothesized 'hit-and-run' mechanism of *H. pylori* leading to a loss of infection.[19, 20]

To account for this, my experiments included sera from a prospective cohort study. Furthermore, I investigated antibody responses in an East Asian population with a generally higher *H. pylori* prevalence and high GC rates, while the previously mentioned studies both used European samples. Most importantly, an association between GC and anti-CagA antibodies was not reported in either of the two publications, which indicated that the here generated microarrays exhibit methodological advantages.

As already described, I could not identify clear lead antigens to distinguish NCGC cases from controls. Therefore, I proceeded by selecting 'informative' antigens, which are not necessarily able to distinguish NCGC cases from controls individually, but in combination. As the goal was to transfer *H. pylori* antigens to multiplex serology, using combined seropositivity was a reasonable strategy.

For this purpose, I modeled random forest classifiers and decision trees.[130] By supplying input data (microarray readouts) and the corresponding class (NCGC case or control), a model was built classifying as many samples correctly as possible. In contrast to simple decision trees, random forest algorithms vary the input data by bootstrapping, which prevents over-fitting.[131] The advantage of decision-tree based algorithms, including random forests, is that they are easy to interpret and robust to unscaled and missing data. By altering different parameters, I was able to force the algorithm to e.g. restrict the selection to a certain number of antigens or only use antigens which classify a minimum number of samples. Each of these random forests included relative importance values as an output; the higher this value was, the higher was the contribution of an antigen to separate NCGC cases from controls.

4.2.4 Characterizing selected antigens from multi-strain microarrays

Among the numerous random forests I modeled, in order to find informative *H. pylori* antigens, some antigens were reoccurring. Random forests built using the Shanxi microarray readouts, most often used antibody responses against the antigen HP1564, which is an established multiplex serology antigen and a verified NCGC risk marker in East Asian populations.[88, 89] Finding HP1564 among the most informative antigens was

reassuring, especially considering that a recent *in vitro* study found it involved in the translocation of CagA into host cells.[88, 89, 115]

A similar function was also reported for HP0545 (CagD), which was also among the most important antigens in the Shanxi study. HP0545 is encoded on the *cag PAI* and secreted by *H. pylori*.[132] This means, that it is actively involved in pathogenesis and also accessible to host antibodies.
Another *H. pylori* antigen which was found informative to classify Shanxi samples, was the VirB4 homolog HP0017. This protein is part of a specialized type IV secretion system essential for the natural competence of *H. pylori* enabling horizontal gene transfer.[124, 133, 134] Strains carrying this protein have a survival advantage and could be potentially associated with a more persistent infection.
Further reoccurring antigens were HP0477 and HP1091. HP0477 is a membrane protein that belongs to the same family as HopA. As already discussed, different members of this family were already described to be associated with NCGC status.[56-59, 135] HP1091 is a membrane-embedded alpha-ketoglutarate permease, a transport system which was found essential for colonization of the gastric mucosa.[136, 137]
HP0527 is encoded on the *cag PAI* and antibodies against it were found enriched in the serum of GC patients in an exploratory study using phage display.[107] I manually selected this antigen, as it showed a decreased seroprevalence among NCGC cases from the cross-sectional Shanxi study. Replicating this observation with multiplex serology, would support the 'hit-and-run' hypothesis.

Building classification models for the prospectively collected NIT outputs, I found HopA among the most informative antigens. This contradicted my prior hypothesis of HopA being a late antigen. However, NIT serum samples also included NCGC cases with a short follow-up time (<2 years). These cases might have driven the selection process. Analyzing anti-HopA antibody responses from the complete NIT study might deliver insights into this observation.
By far, the highest recurrence among the NIT samples, was observed for antibodies against the dehydroquinate dehydratase HP1038. This enzyme was reported to be essential for *H. pylori* as part of the shikimic acid pathway, but a potential association with cancer remains unclear.[138]
Furthermore, I also found the established multiplex serology antigen NapA among the most informative antigens in the NIT study. This protein protects *H. pylori* DNA from oxidative stress damage and counteracts a high acid environment.[139, 140] Serum antibodies against NapA have been found associated with GC in Asian populations.

However, in Western populations an inverse association was reported.[99, 129 88, 141] The discrepancy could potentially derive from environmental factors or differences in hosts' genetics.

An antigen that was found informative and not encoded by *H. pylori* 26695 was Mal1434, a type IV secretion system protein VirB6. These secretion systems are commonly found among Gram-negative bacteria with *H. pylori* encoding multiple variants.[124, 142] Members of other type IV secretion systems, especially the *cag PAI*, have been described as virulence factors.[120]

Among the selected antigens, I also identified proteins of unknown function, e.g. Mal648 and Lat1540. By far most functional studies among *H. pylori* proteins were conducted using the reference strain 26695. Hence, antigens that are not encoded on this reference strain are more likely to be not yet functionally characterized. Nevertheless, both antigens were predicted to be membrane-embedded, making them potentially accessible for host antibodies.

The same applies to the two membrane proteins HP0582 and HP0659. Although their exact function is unknown, they contain domains of other immunogenic proteins. HP0659 includes parts of the membrane chaperone SurA, while HP0582 incorporates the C-terminal domain of the protein TonB.[143, 144] On both *H. pylori* microarrays (minimized and multi-strain), TonB was by far the most immunogenic antigen, with antibodies found in >97% of the sera. Further studies are needed to evaluated, if this was unspecific or if TonB could serve as an *H. pylori* infection marker.

None of the antigens selected based on the Shanxi samples, overlapped with the selection from the NIT samples. This could potentially reflect biological differences (cross-sectional vs. prospective) or mean that there were no clear 'lead antigens' able to distinguish between NCGC cases and controls. In latter case, selected antigens would correspond chance findings.

In order to investigate this, all of the selected *H. pylori* antigens were transferred to multiplex serology and antibody responses were measured in all three sample sets. This included 669 serum samples from the Shanxi study, 649 NIT serum samples and 993 NIT plasma samples.

4.3 From microarray to multiplex serology

4.3.1 Methodological similarities and differences

Multiplex serology is a high-throughput assay enabling simultaneous quantification of antibodies against up to 100 antigens in up to 2,000 samples per day. It was originally developed for HPV serology and since supplemented with antigens from numerous other viruses and bacteria.[78, 83, 135, 145, 146]

The technique offers several desirable advantages, including a high-throughput capacity and high reproducibility. Comparing multiplex results I generated in the scope of this thesis with published results for the NIT by Murphy *et al*. showed correlation coefficients above 0.8 for virtually all of the shared antigens.[102] Hence, multiplex serology is a powerful tool to establish and validate new serological markers. However, discovering new informative antigens in an unbiased fashion is barely feasible, as only up to 100 antigens can be included at once. Combining high-density microarrays with high-throughput multiplex serology is a desirable approach.

However, these two serological methods have substantial differences. While antigens on the microarrays are flanked by terminal V5- and 6xHis-tags, multiplex serology requires the expression of GST-X-TAG fusion proteins. GST enables loading the recombinant antigens onto glutathione-derivatized beads and increases the solubility of the proteins, which enables expressing membrane proteins and keeping them in solution to a certain degree.[147] However, the size of GST can sterically hinder the antigen to fold correctly or make parts of it inaccessible, which is usually not caused by a small V5 or 6xHis-tag.

Altogether, both methods are prone to alter protein folding and the recombinant antigens might hence not be comparable. Also, the methods are based on different protein expression techniques: while antigens for multiplex serology are transcribed and translated in *E. coli* cells, microarrays proteins are expressed in a cell-free milieu, which could potentially cause an alteration of the final protein product.

Furthermore, the presentation of antigens is different. While microarray antigens are covalently bound and presented on a planar surface, multiplex serology antigens are expressed and presented in solution, which could potentially influence binding of antibody from serum.

4.3.2 Expression of GST-X-TAG fusion proteins

In order to make as few changes to the selected antigens as possible, I decided to keep full-length sequences, without excluding domains which encode signal peptides or transmembrane domains. Only three antigens (HP0017, HP0527 and Kor1294) needed to be shortened to increase the chance for a successful expression in *E. coli*. For the same reason, I also optimized the codon usage for all of the sequences before ordering gene synthesis and cloning into the expression plasmid pGEX-4T3tag from Eurofins genomics (Ebersberg). The integrity of all plasmids was checked by PCR, sequencing and multiple restriction digests and eventually, I confirmed that all of the 22 DNA constructs were correct.

On the protein expression level, however, certain antigens exhibited low expression yield. Anti-TAG ELISA readouts for more than half of the antigen lysates (e.g. Lat1540, HP1435, HP1091 and Kor1294N) indicated a low relative concentration of full-length expressed protein. This was supported by a lack of signal for the C-terminal anti-TAG Western blot while signals of the N-terminal anti-GST Western blots appeared very intense. This indicates that a substantial fraction of the antigens lacked the C terminus. Multiple reasons can lead to partial protein expression. On a DNA level, unfavorable base pair compositions can cause the RNA to form hairpins or other secondary structures and sterically hinder RNA polymerases to finish transcription. This might e.g. be the case for HP1091 and Kor1294N which already exhibited difficulties during cloning. A similar effect could also affect protein translation. Secondary structures or composition of the RNA or the nascent polypeptide chain might sterically hinder ribosomes and cause a premature termination.
Nevertheless, even partially expressed proteins can be good serological targets. As already described, antigens comprise multiple epitopes and there is a high chance that some of these epitopes are presented even without full-length expression. Hence, I used all GST-X-TAG fusion proteins for multiplex serology.
Analyzing loading controls and antibody levels determined in the multiplex serology confirmed that almost all of the antigens were successfully presented on the beads. Only the presentation of HP0477, HP0017N and Mal1434 on the beads could not be verified as the anti-TAG antibody showed barely a signal. Respective antibody responses were later found to fall below the lower limit of quantification when measured in sera. Antibody levels measured in plasma samples were slightly higher, but most measurements were below 50 MFI, too.

4.3.3 Determining cutoffs for seropositivity

In order to evaluate if antibodies against a specific *H. pylori* antigen are associated with NCGC, antibody levels were classified into “seropositive” or “seronegative” by setting a cutoff.
Classically, these cutoffs were calculated using the mean plus three SDs of a seronegative reference population. For the validation of the established *H. pylori* multiplex serology antigens, a German reference population was used.[83]
This approach has, however, several limitations. Fixed cutoffs are inflexible and potentially not transferable to other populations, e.g. *H. pylori* ELISAs which were validated in a Western population showed a suboptimal performance in Chinese patients.[148] Furthermore, it assumes that signals derived from a seronegative population derive from a single Gaussian distribution, which in reality is often not the case.[149]

Cutoffs were adapted in different *H. pylori* multiplex serology studies conducted over the years using a visual inflection point method.[88, 99, 102, 150] Nevertheless, this was not entirely objective and required expertise. It can therefore not be applied to new antigens, which have not been characterized yet. I decided to use an automated and systematic approach based on finite mixture modeling to determine suitable cutoffs for the new and also for the established *H. pylori* antigens.[150] This approach is solely based on data distributions and does not require a designated seronegative population. Hence, cutoffs are versatile and can be adapted to a target population in an objective manner.
I based my work on the assumption, that antibody responses against a specific antigen derived from one of two populations – seropositives or seronegatives – and that the antibody levels in these two populations were reasonably distinct.

To model the two populations, I used skew-lognormal distributions as proposed by Domingues *et al.*[149] They analyzed serological responses to different herpes viruses and found the high modeling flexibility of skewed distribution especially appropriate to reflect serological data. I confirmed this observations for the antibody responses against most of the *H. pylori* antigens.
Even though cutoffs sometimes varied by sample set, e.g. cutoffs for plasma samples being generally higher, resulting seroprevalences were comparable between the three sample sets. This was expected as the NIT sample sets derive from the same cohort

and the Shanxi samples were sampled from the same birth cohort in a neighboring county.

However, the algorithm did not converge if one of the two populations was too small. In these cases, the algorithm interpreted one population as the skew of the other population. I bypassed this problem by using two lognormal distributions, in order to determine a cutoff. In the case of *H. pylori*, low prevalence antigens were always accompanied by low antibody titers, which is not desirable for a serological marker. Indeed, many of these antigens (e.g. Mal1434, Lat1540 or HP1355) were excluded in the course of the data analyses, hence the influence on the results was marginal.
The algorithm also had problems to converge, if the seropositive and seronegative population were not distinct enough, like the antibody responses against VacAC and HopA. I also chose to assume two lognormal distributions instead, in order to set a cutoff. Further studies could test shorter fragments of these antigens, as merged populations could be a result of cross-reactivity.

Furthermore, I also fitted normal distributions to the antibody responses to CagAN and CagAC. Visually, antibodies responses did not form a typical bivalent distribution, but rather resembled a mixture of at least three populations. Other studies have shown, it can be statistically appropriate to assume more than two populations for modeling, e.g. to account for resolving infections.[151, 152] However, spontaneous eradication of *H. pylori* infections are not typical and especially anti-CagA antibodies are considered to be stable over time.[153] Further studies are necessary to investigate, if these different CagA populations correlate with clinical features or outcomes (e.g. diffuse or intestinal-type GC). This information was not available for any of the used study samples. It also needs to be investigated if the different antibody responses correlate to CagA subtypes.
Altogether, I decided to use a systemic ‘one-for-all’ approach for this project and apply a bimodal distribution to both CagA fragments. This resulted in high cutoffs for the NIT plasma samples and, hence, a reduced seroprevalence. Nevertheless, as the cutoffs affects NCGC cases and controls to the same degree, they are expected to only marginally affect point estimates for the association with case status.

4.4 Validation of NCGC-associated *H. pylori* antigens

4.4.1 Association of seropositivity to individual *H. pylori* antigens in three sample sets

I calculated associations between *H. pylori* antigens and NCGC case status for each antigen applying logistic regression analysis. As described, I was not able to adjust point estimates of the Shanxi study for age and sex due to missing questionnaire data for most of the NCGC cases. Hence, I here compare unadjusted ORs to the results from the NIT sample sets, which were age- and sex-adjusted.

Overall, associations between *H. pylori* antigens and NCGC were very similar in Shanxi serum samples and NIT serum samples. The two studies were initiated in neighboring counties in Northern China. Patients were enrolled into the NIT in 1985 with NCGC cases being diagnosed on average 7.5 years later. NCGC cases for the Shanxi study were sampled between 1997 and 2005. This means that the NIT covered NCGC cases which occurred slightly earlier than the cases of the Shanxi study. Despite an overall declining *H. pylori* infection rate, I did not expect substantial differences as the time interval was too small to encounter strong birth cohort effects.
In contrast, results from the NIT plasma samples showed more discrepancies. This difference was already indicated by higher MFI values in the multiplex serology and consequently higher cutoffs to distinguish between seropositive and seronegative samples. The NIT plasma samples derive from the same cohort as the serum samples, but were collected at a later time point in 1999/2000. Although there was no direct overlap between individual serum and plasma samples, meaning that all samples derived from different individuals, I expected similar prerequisites and outcomes. However, I found fewer *H. pylori* antigens significantly associated with NCGC using NIT plasma samples, in accordance with a previous multiplex serology study using NIT serum and plasma samples.[102]
Potentially, these systematic differences are attributable to different sample types. Plasma samples usually exhibit higher background values in multiplex, which could impair the signal-to-noise ratio serology (oral communication with Dr. Waterboer). Other potential influence factors were the preparation of samples and storage conditions.

Nevertheless, some observations were consistent between the three sample sets, e.g. the association between anti-CagA antibodies and NCGC. In the Shanxi study, the

associations were only borderline significant with an OR of 1.42 (95% CI: 1.01 - 2.01) for CagAN and 1.53 (95% CI: 1.08 - 2.18) for CagAC, while point estimates determined in the NIT serum samples were slightly higher with ORs of 1.89 (95%CI: 1.34 – 2.67) and 1.66 (95%CI: 1.18 – 2.33), respectively. Overall these effect measures were consistent with published multiplex serology results for anti-CagA antibodies and NCGC in prospective East Asian cohort studies.[88, 102]

For Western populations, a stronger association has been determined, due to overall lower *H. pylori* prevalences and occurrence of CagA negative *H. pylori* strains. In a Swedish population, the OR was as high as 9.20 (95% CI: 5.47 – 15.48).[154] The OR of 3.64 (95% CI: 1.69 – 7.87), which I determined probing minimized 26695-microrrays with sera from the MCC Spain study, fitted well into this range.

Analyzing NIT plasma samples, CagAC, but not CagAN, showed a significant association with NCGC. This was unexpected, as usually results for both CagA fragments are highly correlated.

As discussed earlier, comparatively high cutoffs were set for CagAN and CagAC in the plasma samples, causing seroprevalences to fall <50%, while seroprevalences determined in serum samples were approximately 65%. However, cutoffs were applied to NCGC cases and controls and should therefore barely influence the risk estimate.

When excluding NCGC cases with a follow-up time <2 years from the plasma samples, the association between CagAN and NCGC cases became significant again. This could indicate that NCGC cases loose CagAN antibodies prior to diagnosis, but is not supported by the results from the other studies presented in this thesis, especially considering the Shanxi study and the MCC Spain study for which NCGC cases were sampled at diagnosis.[100] This inconclusive observation was supported by results from the earlier multiplex serology study published by Murphy *et al.*, who did not find a significant association between NCGC and anti-CagA antibodies (fragments combined, OR: 1.66, 95% CI of 0.95 - 2.89).[102]

Altogether, the results suggest that apart from a technical difference (plasma vs. sera), possibly also a biological difference separates the NIT plasma samples from the NIT serum samples and the Shanxi serum samples.

Antibodies against two further established *H. pylori* antigens, HP1564 and HP0305, were found to be associated with NCGC cases in all three sample sets. This was in accordance to the results from Murphy *et al.* and other East Asian cohorts.[88, 89, 102]

In Western populations, results have been more ambiguous. Anti-HP1564 antibodies were associated with 2.50-fold higher odds for NCGC in Sweden, while the association

was found insignificant in Germany and Spain. In contrast, anti-HP0305 antibodies were associated with NCGC case status in the German, but not the Spanish or Swedish study.[99, 154, 155] Additional studies are required to investigate the role of antibodies against HP1564 and HP0305 in more detail.

Another antigen which was consistently associated with NCGC in all three sample sets was the new *H. pylori* antigen HopA. As already described, HopA was identified as a potential NCGC-marker probing minimized *H. pylori* microarrays with serum samples from the MCC Spain study. Furthermore, it was among the selected antigens probing the *H. pylori* multi-strain microarrays with samples from the NIT.
Analyzing the MCC Spain experiments, I hypothesized that HopA could be a 'late antigen', meaning that antibodies occurred with GC development or shortly before. Verifying a statistically significant association with NCGC in the cross-sectional Shanxi study supported this hypothesis, but anti-HopA antibodies were also found significantly associated in the prospective NIT study. Among serum samples with a mean follow-up time of 7.5 years, the association was, however, only borderline significant (OR: 1.68, 95% CI: 1.01 - 2.81).
NIT plasma samples had a shorter follow-up time of only 3.5 years on average, corresponding to a stronger association between NCGC cases and anti-HopA antibodies. This supported my initial hypothesis, especially as the association in NIT serum samples became insignificant excluding cases with a follow-up time <2 years. The development of gastric cancer is a continuous process that may take many years and it could be very well possible that specific antigens, e.g. HopA, are expressed by *H. pylori* to adapt to the changing environment.
Lastly, HopA was found associated with NCGC in both Chinese studies, as well as the MCC Spain study, which could potentially make it an *H. pylori* antigen that is universally associated with NCGC. Additional studies using different populations are needed to investigate this observation further.

This raises the question why HopA has not been described in the context of GC, although it is encoded by the common reference strain 26695. A potential explanation could be the length of the here used HopA antigens. Neither the minimized *H. pylori* 26695-microarrays, nor the large *H. pylori* multi-strain microarrays contained full-length expressed HopA. The C-terminal signal was missing in the protein expression control stainings. For multiplex serology, HopA was expressed as GST-HopA-TAG fusion protein, but again partially expressed antigen was dominant.

Alm *et al*. compared different members of the Hop family and found that they share a common C-terminal domain, as well as a short N-terminal signal peptide.[156, 157] These shared domains could potentially have caused cross-reactivities and mask NCGC-associated epitopes of HopA in other studies. Further experiments and the expression of a truncated HopA antigen could confirm or reject this hypothesis.
The other new *H. pylori* antigens showed only suboptimal performances when analyzed individually. Antibodies against HP0582 showed an association with NCGC case status in the NIT serum samples and the OR increased when excluding NCGC cases with a follow-up time <2 years. Further studies covering NCGC-cases with long follow-up times could potentially investigate if HP0582 is an early risk marker.
Furthermore, antibodies against HP0527trunc were found significantly associated with NCGC among NIT serum samples. This antigen was selected manually from the microarrays results due to an increased seroprevalence among NCGC cases of the cross-sectional Shanxi study. This discrepancy was potentially caused by the truncation of HP0527 for multiplex serology leading to the presentation of different epitopes. Further studies are required to evaluate if HP0527trunc could be used as a serological NCGC risk marker.

To summarize, I confirmed the association between NCGC and well-known *H. pylori* antigens including CagAN, HP1564 and HP0305. Furthermore, HopA showed promise to be a late marker for NCGC development. The other new *H. pylori* antigens were not able to individually discriminate NCGC cases from controls, which generally confirmed results from the microarray experiments.

4.4.2 Association of seropositivity to multiple *H. pylori* antigens in three sample sets

After demonstrating that seropositivity against certain individual *H. pylori* antigens was associated with an increased odds for NCGC, I evaluated if seropositivity to multiple antigens corresponded to an even stronger risk. Generally, serological markers can be translated into clinical application by developing an ELISA or a lateral flow test. This can theoretically be done for multiple antigens, but additional work and expenses need to be justified by enhanced assay performances.

Based on multiple East Asian cohorts, Epplein *et al.* determined an OR of 3.34 (95% CI: 2.27–4.91) for being double-seropositive to the antigens HP1564 and HP0305 compared to being seronegative for both of them.[88, 115] My observations were

highly consistent in all three Chinese sample sets despite lower point estimates among the Shanxi sera (OR: 2.20, 95% CI: 1.47 – 3.29), the NIT sera (OR: 2.14, 95% CI: 1.46 - 3.15) and the NIT plasma (OR: 2.44, 95% CI: 1.51 – 3.94), as 95% CI overlapped.

I further tried different other knowledge-driven combinations, e.g. adding the two CagA fragments (CagAN and CagAC), adding all *H. pylori* antigens found significantly associated with NCGC in the respective sample set, or using antigens which were commonly found significantly associated in all three sample sets.
Overall, all of the OR were in the same range with the highest ones being achieved when using all significant antigens per sample set: for the Shanxi samples the OR was 3.46 (95% CI: 2.01 – 5.97) for being seropositive for at least five out of nine antigens, for NIT serum samples, the OR reached 3.39 (95% CI: 2.03 – 5.68) for being seropositive for at least six out of ten antigens and for NIT plasma samples an OR of 2.64 (95% CI: 1.62 – 4.30) was determine for being seropositive for at least four out of five antigens. Summing up, all of the 95% CI strongly overlapped and the slight increase of point estimates would most probably not justify the additional cost of including up to ten antigens in a potential screening assay.

These rather steady point estimates indicated a high number of redundancy and correlations between the different *H. pylori* antigens, which was not surprising as they were all part of the humoral response to the same infection.
They were either caused by ‘paired’ antigens or involved multiple antigens. For a better characterization, I explored the numerical multiplex serology readouts by performing PCA. This method is used to break down the large dataset into uncorrelated principle components, which corresponded to the eigenvectors of the correlation matrix.
Visualizing the distribution of NCGC cases and controls in the different principle components, I observed that the respective centers corresponding to the NCGC cases or healthy controls could only be discriminated in the first principle component (= first dimension). However, 95% confidence ellipses strongly overlapped, which indicated that NCGC cases could not be sufficiently distinguished from controls using *H. pylori* multiplex readouts only. Unsurprisingly, this first dimension was strongly correlated with antibody responses against those *H. pylori* antigens which I already identified associated with NCGC case status.

In order to estimate which antibody response could actually contribute independently to distinguish NCGC cases from controls, I trained a random forest classification model with 80% of the multiplex serology readouts using the NIT serum samples as an example.
Based on relative importance values of each of the antibody responses, I was able to evaluate which *H. pylori* antigen actually added information to the model. Among NIT serum samples, CagAC was found most important to discriminate NCGC cases from controls in the NIT serum sample set. Although CagAN showed an equal correlation to the first principle component, it was only on 10th position when it came to importance. This confirmed that antibody responses against CagAN and CagAC were redundant and using both did not add information to the classification model. Instead, the algorithm used antibody responses against HP0305, NapA, HP0003, etc. to supplement CagAC readouts.

HP0003 is a new *H. pylori* antigen that did not stand out in the individual analyses. It was selected from *H. pylori* multi-strain microarrays probed with NIT serum samples, being among the most important antigens to model the case-control status. Here, I confirmed the new *H. pylori* antigens have the potential to add independent value to classification models.
To estimate the overall performance of the random forest classifier, I used the remaining 20% of the NIT samples as a test set and determined a specificity of 52% and a sensitivity of 59% to detect NCGC. I also tested the eligibility of the model to classify Shanxi serum samples and NIT plasma samples which yielded specificities of 54% and 56%, while sensitivities reached only 59% and 64%, respectively.

Clearly, the model did not perform well enough substantially narrow down individuals at risk for NCGC and tuning the classification algorithms did also barely improve the outcome (data not shown). It was also not possible to include other risk factors as sex and age into the model, due to matching for these factors.
However, investigating interdependencies between antibody responses indicated that current approaches might still profit from the new *H. pylori* antigens. Recently, Murphy *et al*. presented a prediction model which amongst others combined patient characteristics with pepsinogen measurements and antibody responses against the *H. pylori* antigens HP0305, HP1564 and UreA.[90] Pre-specifying a specificity of 75%, this prediction model achieved a higher sensitivity (59%) than risk stratification by the ABC method (53%). The new *H. pylori* antigens might be able to add independent value to these kind of prediction models.

4.5 Conclusion and outlook

By combining high-density *H. pylori* multi-strain microarrays with multiplex serology, I enabled an unbiased *de novo* identification of NCGC-associated *H. pylori* antigens and subsequent validation in a high-throughput fashion.
I described a new *H. pylori* marker, HopA, which was significantly associated with increased risk for NCGC in two independent Chinese studies. The observed association was as strong as with seropositivity to the established *H. pylori* virulence factor CagA.

The expression of HopA was described to be pH-dependent and could therefore be potentially enhanced during GC formation, which entails drastic pH changes in the gastric lumen and mucosa.[127, 128] This could even mean that expression of HopA by *H. pylori*, and the corresponding antibody response by its human host, correlate with time to diagnosis. My results supported this hypothesis, as antibodies against HopA showed a stronger association in the cross-sectional MCC Spain and Shanxi study, compared to the prospective NIT. However, 95% CI overlapped, meaning that additional studies are required to further investigate this hypothesis. For this purpose, a cohort comprising serial samples would be ideal.
Additionally, HopA fragments that lack common N-terminal and C-terminal sequences will be expressed. Potentially, this could decrease cross-reactivities and improve the signal-to-noise ratio.

In case antibody responses against HopA increase closer to diagnosis, numerical values would need to be evaluated in order to fully exploit the clinical value. However, my results showed that numerical antibody responses as measured in multiplex serology were not generalizable between studies, e.g. by requiring different cutoffs for seropositivity or by exhibiting overall different MFI ranges.
Potentially, this could be overcome by using self-contained measures or normalization, e.g. ratios between antigen responses, similar to the PGI/PGII biomarker. Respective analyses are pending.

Cutoffs for seropositivity were defined according to finite mixture modeling results. Using this approach, I also re-analyzed multiplex serology results which were generated for the initial validation of the *H. pylori* multiplex serology panel.[83] Currently, this initial panel comprises 15 *H. pylori* antigens and being seropositive for at least four of them showed the highest concordance to a commercial *H. pylori* ELISA detecting

infection.[83] Using modeled cutoffs for seropositivity, I was able to achieve an increased concordance while requiring less *H. pylori* antigens. This would allow to decrease the current *H. pylori* multiplex serology panel, in order to include new *H. pylori* antigens or antigens from other pathogens, e.g. EBV. Recently, serological risk markers for EBV-driven cancers were validated.[117, 118] As EBV can also play a causative role in the development of GC, it is of interest to investigate the role of co-infections more closely.

Alternatively, other new *H. pylori* antigens could be included into the panel. Despite being not found significantly associated in the here presented studies, I showed that they can still add independent value to a potential risk stratification model.
Overall, ORs determined throughout my analyses, were of only moderate strength, which was in accordance to existing observations for a Chinese population with a high baseline prevalence for *H. pylori* infection. This means that a risk stratification for GC cannot be solely achieved measuring serum antibodies against specific *H. pylori* virulence factors in this population. Nevertheless, a combined approach which considers further variables, e.g., demographic factors or PGI/PGII results could enable a risk stratification as demonstrated by Murphy *et al*.[90]
This requires further studies with different study designs, as the here described populations were matched for relevant confounders, e.g. sex or age, or were missing this information.
Moreover, associations between NCGC and the new *H. pylori* antigens need to be re-analyzed in other populations to assess the generalizability of the results. Generally, serological *H. pylori* assays have been exhibiting differences across populations. E.g., antibodies against HP1564 and HP0305 enabled risk stratification in East Asian populations, but exhibited ambiguous results in European population.[99, 154, 155] *H. pylori* antigens or a risk stratification signature based on multiple antigens that could be universally used, is highly desirable, but could be population-dependent.

Altogether, I showed that antigen microarrays can be a useful tool to pre-select serological risk markers in a high-density fashion for further application in multiplex serology. Associations of selected *H. pylori* antigens with NCGC were, however, of only moderate strength and need to be combined with other GC-associated biomarkers.
The respective applicability of the identified markers in GC risk stratification needs to be elucidated with regard to clinical settings, in order to potentially contribute to the prevention of GC.

5 References

1. Sung, H., J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal*, et al.*, *Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.* CA Cancer J Clin, 2021. **71**(3): p. 209-249.
2. National Health Commission Of The People's Republic Of, C., *Chinese guidelines for diagnosis and treatment of gastric cancer 2018 (English version).* Chin J Cancer Res, 2019. **31**(5): p. 707-737.
3. Howson, C.P., T. Hiyama, and E.L. Wynder, *The decline in gastric cancer: epidemiology of an unplanned triumph.* Epidemiol Rev, 1986. **8**: p. 1-27.
4. Arnold, M., J.Y. Park, M.C. Camargo, N. Lunet, D. Forman, and I. Soerjomataram, *Is gastric cancer becoming a rare disease? A global assessment of predicted incidence trends to 2035.* Gut, 2020. **69**(5): p. 823-829.
5. Zeng, H., R. Zheng, Y. Guo, S. Zhang, X. Zou, N. Wang*, et al.*, *Cancer survival in China, 2003-2005: a population-based study.* Int J Cancer, 2015. **136**(8): p. 1921-30.
6. Rawla, P. and A. Barsouk, *Epidemiology of gastric cancer: global trends, risk factors and prevention.* Prz Gastroenterol, 2019. **14**(1): p. 26-38.
7. Hamashima, C., K. Yoshimura, and A. Fukao, *A study protocol for expanding the screening interval of endoscopic screening for gastric cancer based on individual risks: prospective cohort study of gastric cancer screening.* Ann Transl Med, 2020. **8**(23): p. 1604.
8. Chang, Y., B. Cho, K.Y. Son, D.W. Shin, H. Shin, H.K. Yang*, et al.*, *Determinants of gastric cancer screening attendance in Korea: a multi-level analysis.* BMC Cancer, 2015. **15**: p. 336.
9. Allemani, C., T. Matsuda, V. Di Carlo, R. Harewood, M. Matz, M. Niksic*, et al.*, *Global surveillance of trends in cancer survival 2000-14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries.* Lancet, 2018. **391**(10125): p. 1023-1075.
10. Karimi, P., F. Islami, S. Anandasabapathy, N.D. Freedman, and F. Kamangar, *Gastric cancer: descriptive epidemiology, risk factors, screening, and prevention.* Cancer Epidemiol Biomarkers Prev, 2014. **23**(5): p. 700-13.
11. Conteduca, V., D. Sansonno, G. Lauletta, S. Russi, G. Ingravallo, and F. Dammacco, *H. pylori infection and gastric cancer: state of the art (review).* Int J Oncol, 2013. **42**(1): p. 5-18.
12. Colquhoun, A., M. Arnold, J. Ferlay, K.J. Goodman, D. Forman, and I. Soerjomataram, *Global patterns of cardia and non-cardia gastric cancer incidence in 2012.* Gut, 2015. **64**(12): p. 1881-8.
13. Liou, J.M., P. Malfertheiner, Y.C. Lee, B.S. Sheu, K. Sugano, H.C. Cheng*, et al.*, *Screening and eradication of Helicobacter pylori for gastric cancer prevention: the Taipei global consensus.* Gut, 2020. **69**(12): p. 2093-2112.
14. Lauren, P., *The Two Histological Main Types of Gastric Carcinoma: Diffuse and So-Called Intestinal-Type Carcinoma. An Attempt at a Histo-Clinical Classification.* Acta Pathol Microbiol Scand, 1965. **64**: p. 31-49.
15. Hu, B., N. El Hajj, S. Sittler, N. Lammert, R. Barnes, and A. Meloni-Ehrig, *Gastric cancer: Classification, histology and application of molecular pathology.* J Gastrointest Oncol, 2012. **3**(3): p. 251-61.
16. Correa, P. and M.B. Piazuelo, *The gastric precancerous cascade.* J Dig Dis, 2012. **13**(1): p. 2-9.
17. Uno, Y., *Prevention of gastric cancer by Helicobacter pylori eradication: A review from Japan.* Cancer Med, 2019. **8**(8): p. 3992-4000.

18. Vogelstein, B., E.R. Fearon, S.R. Hamilton, S.E. Kern, A.C. Preisinger, M. Leppert*, et al.*, *Genetic alterations during colorectal-tumor development.* N Engl J Med, 1988. **319**(9): p. 525-32.
19. Hatakeyama, M., *Structure and function of Helicobacter pylori CagA, the first-identified bacterial protein involved in human cancer.* Proc Jpn Acad Ser B Phys Biol Sci, 2017. **93**(4): p. 196-219.
20. Chen, S., L. Ying, M. Kong, Y. Zhang, and Y. Li, *The Prevalence of Helicobacter pylori Infection Decreases with Older Age in Atrophic Gastritis.* Gastroenterol Res Pract, 2013. **2013**: p. 494783.
21. Kusters, J.G., A.H. van Vliet, and E.J. Kuipers, *Pathogenesis of Helicobacter pylori infection.* Clin Microbiol Rev, 2006. **19**(3): p. 449-90.
22. Lin, Y., Z. Wu, W. Guo, and J. Li, *Gene mutations in gastric cancer: a review of recent next-generation sequencing studies.* Tumour Biol, 2015. **36**(10): p. 7385-94.
23. Lee, J.Y., E.J. Gong, E.J. Chung, H.W. Park, S.E. Bae, E.H. Kim*, et al.*, *The Characteristics and Prognosis of Diffuse-Type Early Gastric Cancer Diagnosed during Health Check-Ups.* Gut Liver, 2017. **11**(6): p. 807-812.
24. Ansari, S., B. Gantuya, V.P. Tuan, and Y. Yamaoka, *Diffuse Gastric Cancer: A Summary of Analogous Contributing Factors for Its Molecular Pathogenicity.* Int J Mol Sci, 2018. **19**(8).
25. La Vecchia, C., E. Negri, S. Franceschi, and A. Gentile, *Family history and the risk of stomach and colorectal cancer.* Cancer, 1992. **70**(1): p. 50-5.
26. Skierucha, M., A.N. Milne, G.J. Offerhaus, W.P. Polkowski, R. Maciejewski, and R. Sitarz, *Molecular alterations in gastric cancer with special reference to the early-onset subtype.* World J Gastroenterol, 2016. **22**(8): p. 2460-74.
27. Humans, I.W.G.o.t.E.o.C.R.t., *IARC monographs on the evaluation of carcinogenic risks to humans. Ingested nitrate and nitrite, and cyanobacterial peptide toxins.* IARC Monogr Eval Carcinog Risks Hum, 2010. **94**: p. v-vii, 1-412.
28. Ladeiras-Lopes, R., A.K. Pereira, A. Nogueira, T. Pinheiro-Torres, I. Pinto, R. Santos-Pereira*, et al.*, *Smoking and gastric cancer: systematic review and meta-analysis of cohort studies.* Cancer Causes Control, 2008. **19**(7): p. 689-701.
29. Cavaleiro-Pinto, M., B. Peleteiro, N. Lunet, and H. Barros, *Helicobacter pylori infection and gastric cardia cancer: systematic review and meta-analysis.* Cancer Causes Control, 2011. **22**(3): p. 375-87.
30. Helicobacter and G. Cancer Collaborative, *Gastric cancer and Helicobacter pylori: a combined analysis of 12 case control studies nested within prospective cohorts.* Gut, 2001. **49**(3): p. 347-53.
31. Mukaisho, K., T. Nakayama, T. Hagiwara, T. Hattori, and H. Sugihara, *Two distinct etiologies of gastric cardia adenocarcinoma: interactions among pH, Helicobacter pylori, and bile acids.* Front Microbiol, 2015. **6**: p. 412.
32. Burke, A.P., T.S. Yen, K.M. Shekitka, and L.H. Sobin, *Lymphoepithelial carcinoma of the stomach with Epstein-Barr virus demonstrated by polymerase chain reaction.* Mod Pathol, 1990. **3**(3): p. 377-80.
33. Shinozaki-Ushiku, A., A. Kunita, and M. Fukayama, *Update on Epstein-Barr virus and gastric cancer (review).* Int J Oncol, 2015. **46**(4): p. 1421-34.
34. Shibata, D., M. Tokunaga, Y. Uemura, E. Sato, S. Tanaka, and L.M. Weiss, *Association of Epstein-Barr virus with undifferentiated gastric carcinomas with intense lymphoid infiltration. Lymphoepithelioma-like carcinoma.* Am J Pathol, 1991. **139**(3): p. 469-74.
35. Lee, J.H., S.H. Kim, S.H. Han, J.S. An, E.S. Lee, and Y.S. Kim, *Clinicopathological and molecular characteristics of Epstein-Barr virus-associated gastric carcinoma: a meta-analysis.* J Gastroenterol Hepatol, 2009. **24**(3): p. 354-65.

36. Singh, S. and H.C. Jha, *Status of Epstein-Barr Virus Coinfection with Helicobacter pylori in Gastric Cancer.* J Oncol, 2017. **2017**: p. 3456264.
37. Davila-Collado, R., O. Jarquin-Duran, L.T. Dong, and J.L. Espinoza, *Epstein-Barr Virus and Helicobacter Pylori Co-Infection in Non-Malignant Gastroduodenal Disorders.* Pathogens, 2020. **9**(2).
38. El-Omar, E.M., M. Carrington, W.H. Chow, K.E. McColl, J.H. Bream, H.A. Young*, et al.*, *Interleukin-1 polymorphisms associated with increased risk of gastric cancer.* Nature, 2000. **404**(6776): p. 398-402.
39. Persson, C., P. Canedo, J.C. Machado, E.M. El-Omar, and D. Forman, *Polymorphisms in inflammatory response genes and their association with gastric cancer: A HuGE systematic review and meta-analyses.* Am J Epidemiol, 2011. **173**(3): p. 259-70.
40. Warren, J.R. and B. Marshall, *Unidentified curved bacilli on gastric epithelium in active chronic gastritis.* Lancet, 1983. **1**(8336): p. 1273-5.
41. Marshall, B.J. and J.R. Warren, *Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration.* Lancet, 1984. **1**(8390): p. 1311-5.
42. IARC Working Group, *IARC monograph on the evaluation of carcinogenic risks to humans: Schistosomes, liver flukes and Helicobacter pylori.* Lyon, France: International Agency for Research on Cancer, 1994.
43. Hooi, J.K.Y., W.Y. Lai, W.K. Ng, M.M.Y. Suen, F.E. Underwood, D. Tanyingoh*, et al.*, *Global Prevalence of Helicobacter pylori Infection: Systematic Review and Meta-Analysis.* Gastroenterology, 2017. **153**(2): p. 420-429.
44. Falush, D., T. Wirth, B. Linz, J.K. Pritchard, M. Stephens, M. Kidd*, et al.*, *Traces of human migrations in Helicobacter pylori populations.* Science, 2003. **299**(5612): p. 1582-5.
45. Linz, B., F. Balloux, Y. Moodley, A. Manica, H. Liu, P. Roumagnac*, et al.*, *An African origin for the intimate association between humans and Helicobacter pylori.* Nature, 2007. **445**(7130): p. 915-918.
46. Yamaoka, Y., M. Kato, and M. Asaka, *Geographic differences in gastric cancer incidence can be explained by differences between Helicobacter pylori strains.* Intern Med, 2008. **47**(12): p. 1077-83.
47. Cover, T.L., *Helicobacter pylori Diversity and Gastric Cancer Risk.* mBio, 2016. **7**(1): p. e01869-15.
48. Censini, S., C. Lange, Z. Xiang, J.E. Crabtree, P. Ghiara, M. Borodovsky*, et al.*, *cag, a pathogenicity island of Helicobacter pylori, encodes type I-specific and disease-associated virulence factors.* Proc Natl Acad Sci U S A, 1996. **93**(25): p. 14648-53.
49. Hatakeyama, M., *Oncogenic mechanisms of the Helicobacter pylori CagA protein.* Nat Rev Cancer, 2004. **4**(9): p. 688-94.
50. Song, X., N. Xin, W. Wang, and C. Zhao, *Wnt/beta-catenin, an oncogenic pathway targeted by H. pylori in gastric carcinogenesis.* Oncotarget, 2015. **6**(34): p. 35579-88.
51. Li, Q., J. Liu, Y. Gong, and Y. Yuan, *Association of CagA EPIYA-D or EPIYA-C phosphorylation sites with peptic ulcer and gastric cancer risks: A meta-analysis.* Medicine (Baltimore), 2017. **96**(17): p. e6620.
52. Basso, D., C.F. Zambon, D.P. Letley, A. Stranges, A. Marchet, J.L. Rhead*, et al.*, *Clinical relevance of Helicobacter pylori cagA and vacA gene polymorphisms.* Gastroenterology, 2008. **135**(1): p. 91-9.
53. Cover, T.L. and S.R. Blanke, *Helicobacter pylori VacA, a paradigm for toxin multifunctionality.* Nat Rev Microbiol, 2005. **3**(4): p. 320-32.
54. Xu, C., D.M. Soyfoo, Y. Wu, and S. Xu, *Virulence of Helicobacter pylori outer membrane proteins: an updated review.* Eur J Clin Microbiol Infect Dis, 2020. **39**(10): p. 1821-1830.

55. Zhang, J.Y., J.J. Qian, X.L. Zhang, and Q.M. Zou, *Outer membrane inflammatory protein A, a new virulence factor involved in the pathogenesis of Helicobacter pylori.* Molecular Biology Reports, 2014. **41**(12): p. 7807-7814.
56. Exner, M.M., P. Doig, T.J. Trust, and R.E. Hancock, *Isolation and characterization of a family of porin proteins from Helicobacter pylori.* Infect Immun, 1995. **63**(4): p. 1567-72.
57. Papamichael, K. and G.J. Mantzaris, *Pathogenesis of Helicobacter pylori infection: colonization, virulence factors of the bacterium and immune and non-immune host response.* Hospital Chronicles, 2012. **7**(1): p. 32-37.
58. Rad, R., M. Gerhard, R. Lang, M. Schoniger, T. Rosch, W. Schepp*, et al.*, *The Helicobacter pylori blood group antigen-binding adhesin facilitates bacterial colonization and augments a nonspecific immune response.* J Immunol, 2002. **168**(6): p. 3033-41.
59. Yamaoka, Y., O. Ojo, S. Fujimoto, S. Odenbreit, R. Haas, O. Gutierrez*, et al.*, *Helicobacter pylori outer membrane proteins and gastroduodenal disease.* Gut, 2006. **55**(6): p. 775-81.
60. Ohno, T., M. Sugimoto, A. Nagashima, H. Ogiwara, R.K. Vilaichone, V. Mahachai*, et al.*, *Relationship between Helicobacter pylori hopQ genotype and clinical outcome in Asian and Western populations.* J Gastroenterol Hepatol, 2009. **24**(3): p. 462-8.
61. Shiota, S., R. Suzuki, and Y. Yamaoka, *The significance of virulence factors in Helicobacter pylori.* J Dig Dis, 2013. **14**(7): p. 341-9.
62. Hocker, M. and P. Hohenberger, *Helicobacter pylori virulence factors--one part of a big picture.* Lancet, 2003. **362**(9391): p. 1231-3.
63. Jun, J.K., K.S. Choi, H.Y. Lee, M. Suh, B. Park, S.H. Song*, et al.*, *Effectiveness of the Korean National Cancer Screening Program in Reducing Gastric Cancer Mortality.* Gastroenterology, 2017. **152**(6): p. 1319-1328 e7.
64. Sekiguchi, M., I. Oda, T. Matsuda, and Y. Saito, *Epidemiological Trends and Future Perspectives of Gastric Cancer in Eastern Asia.* Digestion, 2022. **103**(1): p. 22-28.
65. Hamashima, C., G. Systematic Review, and G. Guideline Development Group for Gastric Cancer Screening, *Update version of the Japanese Guidelines for Gastric Cancer Screening.* Jpn J Clin Oncol, 2018. **48**(7): p. 673-683.
66. Cui, D., *Background of Gastric Cancer prewarning and Early Diagnosis System*, in *Gastric Cancer Prewarning and Early Diagnosis System* D. Cui, Editor. 2017, Springer Science+Business Media B.V. and Shanghai Jiao Tong University Press: Shanghai.
67. Zagari, R.M., S. Rabitti, D.C. Greenwood, L.H. Eusebi, A. Vestito, and F. Bazzoli, *Systematic review with meta-analysis: diagnostic performance of the combination of pepsinogen, gastrin-17 and anti-Helicobacter pylori antibodies serum assays for the diagnosis of atrophic gastritis.* Aliment Pharmacol Ther, 2017. **46**(7): p. 657-667.
68. Yamaguchi, Y., Y. Nagata, R. Hiratsuka, Y. Kawase, T. Tominaga, S. Takeuchi*, et al.*, *Gastric Cancer Screening by Combined Assay for Serum Anti-Helicobacter pylori IgG Antibody and Serum Pepsinogen Levels--The ABC Method.* Digestion, 2016. **93**(1): p. 13-8.
69. Miki, K., M. Ichinose, A. Shimizu, S.C. Huang, H. Oka, C. Furihata*, et al.*, *Serum pepsinogens as a screening test of extensive chronic gastritis.* Gastroenterol Jpn, 1987. **22**(2): p. 133-41.
70. Miki, K., *Gastric cancer screening by combined assay for serum anti-Helicobacter pylori IgG antibody and serum pepsinogen levels - "ABC method".* Proc Jpn Acad Ser B Phys Biol Sci, 2011. **87**(7): p. 405-14.
71. Kishino, T., T. Oyama, A. Tomori, A. Takahashi, and T. Shinohara, *Usefulness and Limitations of a Serum Screening System to Predict the Risk of Gastric Cancer.* Intern Med, 2020. **59**(12): p. 1473-1480.

72. Cai, Q., C. Zhu, Y. Yuan, Q. Feng, Y. Feng, Y. Hao*, et al.*, *Development and validation of a prediction rule for estimating gastric cancer risk in the Chinese high-risk population: a nationwide multicentre study.* Gut, 2019. **68**(9): p. 1576-1587.
73. Pan, K.F., L. Formichella, L. Zhang, Y. Zhang, J.L. Ma, Z.X. Li*, et al.*, *Helicobacter pylori antibody responses and evolution of precancerous gastric lesions in a Chinese population.* Int J Cancer, 2014. **134**(9): p. 2118-25.
74. Fan, X., X. Qin, Y. Zhang, Z. Li, T. Zhou, J. Zhang*, et al.*, *Screening for gastric cancer in China: Advances, challenges and visions.* Chin J Cancer Res, 2021. **33**(2): p. 168-180.
75. IARC, I.A.f.R.o.C.-. *Helicobacter pylori Eradication as a Strategy for Preventing Gastric Cancer.* IARC Working Group Reports, 2013. **8**.
76. Park, J.Y., D. Forman, E.R. Greenberg, and R. Herrero, *Helicobacter pylori eradication in the prevention of gastric cancer: are more trials needed?* Curr Oncol Rep, 2013. **15**(6): p. 517-25.
77. Yang, L., C. Kartsonaki, P. Yao, C. de Martel, M. Plummer, D. Chapman*, et al.*, *The relative and attributable risks of cardia and non-cardia gastric cancer associated with Helicobacter pylori infection in China: a case-cohort study.* Lancet Public Health, 2021. **6**(12): p. e888-e896.
78. Waterboer, T., P. Sehr, K.M. Michael, S. Franceschi, J.D. Nieland, T.O. Joos*, et al.*, *Multiplex human papillomavirus serology based on in situ-purified glutathione s-transferase fusion proteins.* Clin Chem, 2005. **51**(10): p. 1845-53.
79. Sehr, P., M. Muller, R. Hopfl, A. Widschwendter, and M. Pawlita, *HPV antibody detection by ELISA with capsid protein L1 fused to glutathione S-transferase.* J Virol Methods, 2002. **106**(1): p. 61-70.
80. Mentzer, A.J., N. Brenner, N. Allen, T.J. Littlejohns, A.Y. Chong, A. Cortes*, et al.*, *Identification of host-pathogen-disease relationships using a scalable multiplex serology platform in UK Biobank.* Nat Commun, 2022. **13**(1): p. 1818.
81. Shakeri, R., R. Malekzadeh, D. Nasrollahzadeh, M. Pawilta, G. Murphy, F. Islami*, et al.*, *Multiplex H. pylori Serology and Risk of Gastric Cardia and Noncardia Adenocarcinomas.* Cancer Research, 2015. **75**(22): p. 4876-4883.
82. Chen, H., S. Werner, J. Butt, I. Zornig, P. Knebel, A. Michel*, et al.*, *Prospective evaluation of 64 serum autoantibodies as biomarkers for early detection of colorectal cancer in a true screening setting.* Oncotarget, 2016. **7**(13): p. 16420-32.
83. Michel, A., T. Waterboer, M. Kist, and M. Pawlita, *Helicobacter pylori multiplex serology.* Helicobacter, 2009. **14**(6): p. 525-35.
84. Krah, A., S. Miehlke, K.P. Pleissner, U. Zimny-Arndt, C. Kirsch, N. Lehn*, et al.*, *Identification of candidate antigens for serologic detection of Helicobacter pylori-infected patients with gastric carcinoma.* Int J Cancer, 2004. **108**(3): p. 456-63.
85. Haas, G., G. Karaali, K. Ebermayer, W.G. Metzger, S. Lamer, U. Zimny-Arndt*, et al.*, *Immunoproteomics of Helicobacter pylori infection and relation to gastric disease.* Proteomics, 2002. **2**(3): p. 313-24.
86. Bumann, D., P.R. Jungblut, and T.F. Meyer, *Helicobacter pylori vaccine development based on combined subproteome analysis.* Proteomics, 2004. **4**(10): p. 2843-8.
87. Butt, J., M.G. Varga, W.J. Blot, L. Teras, K. Visvanathan, L. Le Marchand*, et al.*, *Serologic Response to Helicobacter pylori Proteins Associated With Risk of Colorectal Cancer Among Diverse Populations in the United States.* Gastroenterology, 2019. **156**(1): p. 175-186 e2.
88. Cai, H., F. Ye, A. Michel, G. Murphy, S. Sasazuki, P.R. Taylor*, et al.*, *Helicobacter pylori blood biomarker for gastric cancer risk in East Asia.* Int J Epidemiol, 2016. **45**(3): p. 774-81.

89. Epplein, M., J. Butt, Y. Zhang, L.H. Hendrix, C.C. Abnet, G. Murphy*, et al.*, *Validation of a Blood Biomarker for Identification of Individuals at High Risk for Gastric Cancer.* Cancer Epidemiol Biomarkers Prev, 2018. **27**(12): p. 1472-1479.
90. Murphy, J.D., A.F. Olshan, F.C. Lin, M.A. Troester, H.B. Nichols, J. Butt*, et al.*, *A Predictive Model of Noncardia Gastric Adenocarcinoma Risk Using Antibody Response to Helicobacter pylori Proteins and Pepsinogen.* Cancer Epidemiol Biomarkers Prev, 2022. **31**(4): p. 811-820.
91. Hufnagel, K., S. Lueong, M. Willhauck-Fleckenstein, A. Hotz-Wagenblatt, B. Miao, A. Bauer*, et al.*, *Immunoprofiling of Chlamydia trachomatis using whole-proteome microarrays generated by on-chip in situ expression.* Sci Rep, 2018. **8**(1): p. 7503.
92. Sehr, P., K. Zumbach, and M. Pawlita, *A generic capture ELISA for recombinant proteins fused to glutathione S-transferase: validation for HPV serology.* J Immunol Methods, 2001. **253**(1-2): p. 153-62.
93. MacArthur, H. and G. Walter, *Monoclonal antibodies specific for the carboxy terminus of simian virus 40 large T antigen.* J Virol, 1984. **52**(2): p. 483-91.
94. Tatusova, T., M. DiCuccio, A. Badretdin, V. Chetvernin, E.P. Nawrocki, L. Zaslavsky*, et al.*, *NCBI prokaryotic genome annotation pipeline.* Nucleic Acids Res, 2016. **44**(14): p. 6614-24.
95. Steinegger, M. and J. Soding, *MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.* Nat Biotechnol, 2017. **35**(11): p. 1026-1028.
96. Cock, P.J., T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke*, et al.*, *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-3.
97. Madeira, F., Y.M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan*, et al.*, *The EMBL-EBI search and sequence analysis tools APIs in 2019.* Nucleic Acids Res, 2019. **47**(W1): p. W636-W641.
98. Castano-Vinyals, G., N. Aragones, B. Perez-Gomez, V. Martin, J. Llorca, V. Moreno*, et al.*, *Population-based multicase-control study in common tumors in Spain (MCC-Spain): rationale and study design.* Gac Sanit, 2015. **29**(4): p. 308-15.
99. Fernandez de Larrea-Baz, N., B. Perez-Gomez, A. Michel, B. Romero, V. Lope, M. Pawlita*, et al.*, *Helicobacter pylori serological biomarkers of gastric cancer risk in the MCC-Spain case-control Study.* Cancer Epidemiol, 2017. **50**(Pt A): p. 76-84.
100. Jeske, R., D. Reininger, B. Turgu, A. Brauer, C. Harmel, N. Fernandez de Larrea-Baz*, et al.*, *Development of Helicobacter pylori Whole-Proteome Arrays and Identification of Serologic Biomarkers for Noncardia Gastric Cancer in the MCC-Spain Study.* Cancer Epidemiol Biomarkers Prev, 2020. **29**(11): p. 2235-2242.
101. Li, B., P.R. Taylor, J.Y. Li, S.M. Dawsey, W. Wang, J.A. Tangrea*, et al.*, *Linxian nutrition intervention trials. Design, methods, participant characteristics, and compliance.* Ann Epidemiol, 1993. **3**(6): p. 577-85.
102. Murphy, G., N.D. Freedman, A. Michel, J.H. Fan, P.R. Taylor, M. Pawlita*, et al.*, *Prospective study of Helicobacter pylori antigens and gastric noncardia cancer risk in the nutrition intervention trial cohort.* Int J Cancer, 2015. **137**(8): p. 1938-46.
103. Tomb, J.F., O. White, A.R. Kerlavage, R.A. Clayton, G.G. Sutton, R.D. Fleischmann*, et al.*, *The complete genome sequence of the gastric pathogen Helicobacter pylori.* Nature, 1997. **388**(6642): p. 539-47.
104. Zhu, X., M. Gerstein, and M. Snyder, *ProCAT: a data analysis approach for protein microarrays.* Genome Biol, 2006. **7**(11): p. R110.

105. Schaeferling, M., S. Schiller, H. Paul, M. Kruschina, P. Pavlickova, M. Meerkamp*, et al.*, *Application of self-assembly techniques in the design of biocompatible protein microarray surfaces.* Electrophoresis, 2002. **23**(18): p. 3097-105.
106. Sboner, A., A. Karpikov, G. Chen, M. Smith, D. Mattoon, L. Freeman-Cook*, et al.*, *Robust-linear-model normalization to reduce technical variability in functional protein microarrays.* J Proteome Res, 2009. **8**(12): p. 5451-64.
107. Soluri, M.F., S. Puccio, G. Caredda, P. Edomi, M.M. D'Elios, F. Cianchi*, et al.*, *Defining the Helicobacter pylori Disease-Specific Antigenic Repertoire.* Front Microbiol, 2020. **11**: p. 1551.
108. de Martel, C., D. Georges, F. Bray, J. Ferlay, and G.M. Clifford, *Global burden of cancer attributable to infections in 2018: a worldwide incidence analysis.* Lancet Glob Health, 2020. **8**(2): p. e180-e190.
109. Zemella, A., L. Thoring, C. Hoffmeister, and S. Kubick, *Cell-Free Protein Synthesis: Pros and Cons of Prokaryotic and Eukaryotic Systems.* Chembiochem, 2015. **16**(17): p. 2420-31.
110. Wittung-Stafshede, P., *Role of cofactors in protein folding.* Acc Chem Res, 2002. **35**(4): p. 201-8.
111. Hartmann, M., J. Sjodahl, M. Stjernstrom, J. Redeby, T. Joos, and J. Roeraade, *Non-contact protein microarray fabrication using a procedure based on liquid bridge formation.* Anal Bioanal Chem, 2009. **393**(2): p. 591-8.
112. Park, J.Y., D. Forman, L.A. Waskito, Y. Yamaoka, and J.E. Crabtree, *Epidemiology of Helicobacter pylori and CagA-Positive Infections and Global Variations in Gastric Cancer.* Toxins (Basel), 2018. **10**(4).
113. Nejati, S., A. Karkhah, H. Darvish, M. Validi, S. Ebrahimpour, and H.R. Nouri, *Influence of Helicobacter pylori virulence factors CagA and VacA on pathogenesis of gastrointestinal disorders.* Microbial Pathogenesis, 2018. **117**: p. 43-48.
114. Batista, S.A., G.A. Rocha, A.M. Rocha, I.E. Saraiva, M.M. Cabral, R.C. Oliveira*, et al.*, *Higher number of Helicobacter pylori CagA EPIYA C phosphorylation sites increases the risk of gastric cancer, but not duodenal ulcer.* BMC Microbiol, 2011. **11**: p. 61.
115. Varga, M.G., C.R. Wood, J. Butt, M.E. Ryan, W.C. You, K. Pan*, et al.*, *Immunostimulatory membrane proteins potentiate H. pylori-induced carcinogenesis by enabling CagA translocation.* Gut Microbes, 2021. **13**(1): p. 1-13.
116. Wills, G.S., P.J. Horner, R. Reynolds, A.M. Johnson, D.A. Muir, D.W. Brown*, et al.*, *Pgp3 antibody enzyme-linked immunosorbent assay, a sensitive and specific assay for seroepidemiological analysis of Chlamydia trachomatis infection.* Clin Vaccine Immunol, 2009. **16**(6): p. 835-43.
117. Simon, J., Z. Liu, N. Brenner, K.J. Yu, W.L. Hsu, C.P. Wang*, et al.*, *Validation of an Epstein-Barr Virus Antibody Risk Stratification Signature for Nasopharyngeal Carcinoma by Use of Multiplex Serology.* J Clin Microbiol, 2020. **58**(5).
118. Song, L., M. Song, M.C. Camargo, J. Van Duine, S. Williams, Y. Chung*, et al.*, *Identification of anti-Epstein-Barr virus (EBV) antibody signature in EBV-associated gastric carcinoma.* Gastric Cancer, 2021. **24**(4): p. 858-867.
119. Linz, B. and S.C. Schuster, *Genomic diversity in Helicobacter and related organisms.* Res Microbiol, 2007. **158**(10): p. 737-44.
120. Olbermann, P., C. Josenhans, Y. Moodley, M. Uhr, C. Stamer, M. Vauterin*, et al.*, *A global overview of the genetic and functional diversity in the Helicobacter pylori cag pathogenicity island.* PLoS Genet, 2010. **6**(8): p. e1001069.
121. Montano, V., X. Didelot, M. Foll, B. Linz, R. Reinhardt, S. Suerbaum*, et al.*, *Worldwide Population Structure, Long-Term Demography, and Local Adaptation of Helicobacter pylori.* Genetics, 2015. **200**(3): p. 947-63.

122. Aalberse, R.C., *Structural biology of allergens.* J Allergy Clin Immunol, 2000. **106**(2): p. 228-38.
123. van Vliet, A.H., *Use of pan-genome analysis for the identification of lineage-specific genes of Helicobacter pylori.* FEMS Microbiol Lett, 2017. **364**(2).
124. Fischer, W., L. Windhager, S. Rohrer, M. Zeiller, A. Karnholz, R. Hoffmann*, et al.*, *Strain-specific genes of Helicobacter pylori: genome evolution driven by a novel type IV secretion system and genomic island transfer.* Nucleic Acids Res, 2010. **38**(18): p. 6089-101.
125. Tettelin, H., D. Riley, C. Cattuto, and D. Medini, *Comparative genomics: the bacterial pan-genome.* Curr Opin Microbiol, 2008. **11**(5): p. 472-7.
126. Su, Y.L., H.L. Huang, B.S. Huang, P.C. Chen, C.S. Chen, H.L. Wang*, et al.*, *Combination of OipA, BabA, and SabA as candidate biomarkers for predicting Helicobacter pylori-related gastric cancer.* Sci Rep, 2016. **6**: p. 36442.
127. Konturek, S.J., T. Starzynska, P.C. Konturek, E. Karczewska, K. Marlicz, M. Lawniczak*, et al.*, *Helicobacter pylori and CagA status, serum gastrin, interleukin-8 and gastric acid secretion in gastric cancer.* Scand J Gastroenterol, 2002. **37**(8): p. 891-8.
128. Allan, E., C.L. Clayton, A. McLaren, D.M. Wallace, and B.W. Wren, *Characterization of the low-pH responses of Helicobacter pylori using genomic DNA arrays.* Microbiology-Sgm, 2001. **147**: p. 2285-2292.
129. Song, L., M. Song, C.S. Rabkin, S. Williams, Y. Chung, J. Van Duine*, et al.*, *Helicobacter pylori Immunoproteomic Profiles in Gastric Cancer.* J Proteome Res, 2021. **20**(1): p. 409-419.
130. Ho, T.K., *Random decision forests.* Proceedings of 3rd International Conference on Document Analysis and Recognition, 1995. **1**: p. 278-282.
131. Cutler, A. and J.R. Stevens, *Random forests for microarrays.* Methods Enzymol, 2006. **411**: p. 422-32.
132. Cendron, L., M. Couturier, A. Angelini, N. Barison, M. Stein, and G. Zanotti, *The Helicobacter pylori CagD (HP0545, Cag24) protein is essential for CagA translocation and maximal induction of interleukin-8 secretion.* J Mol Biol, 2009. **386**(1): p. 204-17.
133. Chang, K.C., Y.C. Yeh, T.L. Lin, and J.T. Wang, *Identification of genes associated with natural competence in Helicobacter pylori by transposon shuttle random mutagenesis.* Biochem Biophys Res Commun, 2001. **288**(4): p. 961-8.
134. Hofreuter, D., S. Odenbreit, and R. Haas, *Natural transformation competence in Helicobacter pylori is mediated by the basic components of a type IV secretion system.* Mol Microbiol, 2001. **41**(2): p. 379-91.
135. Jeske, R., L. Dangel, L. Sauerbrey, D. Frangoulidis, L.R. Teras, S.F. Fischer*, et al.*, *Development of High-Throughput Multiplex Serology to Detect Serum Antibodies against Coxiella burnetii.* Microorganisms, 2021. **9**(11).
136. Kavermann, H., B.P. Burns, K. Angermuller, S. Odenbreit, W. Fischer, K. Melchers*, et al.*, *Identification and characterization of Helicobacter pylori genes essential for gastric colonization.* J Exp Med, 2003. **197**(7): p. 813-22.
137. Baldwin, D.N., B. Shepherd, P. Kraemer, M.K. Hall, L.K. Sycuro, D.M. Pinto-Santini*, et al.*, *Identification of Helicobacter pylori genes that contribute to stomach colonization.* Infect Immun, 2007. **75**(2): p. 1005-16.
138. Chalker, A.F., H.W. Minehart, N.J. Hughes, K.K. Koretke, M.A. Lonetto, K.K. Brinkman*, et al.*, *Systematic identification of selective essential genes in Helicobacter pylori by genome prioritization and allelic replacement mutagenesis.* J Bacteriol, 2001. **183**(4): p. 1259-68.
139. De Re, V., O. Repetto, S. Zanussi, M. Casarotto, L. Caggiari, V. Canzonieri*, et al.*, *Protein signature characterizing Helicobacter pylori strains of patients with autoimmune atrophic gastritis, duodenal ulcer and gastric cancer.* Infect Agent Cancer, 2017. **12**: p. 22.

140. Cooksley, C., P.J. Jenks, A. Green, A. Cockayne, R.P.H. Logan, and K.R. Hardie, *NapA protects Helicobacter pylori from oxidative stress damage, and its production is influenced by the ferric uptake regulator.* J Med Microbiol, 2003. **52**(Pt 6): p. 461-469.
141. Liu, J.J., H.M. Liu, T.T. Zhang, X.Y. Ren, C. Nadolny, X.Q. Dong*, et al.*, *Serum Helicobacter pylori NapA antibody as a potential biomarker for gastric cancer.* Scientific Reports, 2014. **4**.
142. Fronzes, R., E. Schafer, L. Wang, H.R. Saibil, E.V. Orlova, and G. Waksman, *Structure of a type IV secretion system core complex.* Science, 2009. **323**(5911): p. 266-8.
143. Utt, M., I. Nilsson, A. Ljungh, and T. Wadstrom, *Identification of novel immunogenic proteins of Helicobacter pylori by proteome technology.* J Immunol Methods, 2002. **259**(1-2): p. 1-10.
144. Meinke, A., M. Storm, T. Henics, D. Gelbmann, S. Prustomersky, Z. Kovacs*, et al.*, *Composition of the ANTIGENome of Helicobacter pylori defined by human serum antibodies.* Vaccine, 2009. **27**(25-26): p. 3251-9.
145. Brenner, N., A.J. Mentzer, J. Butt, A. Michel, K. Prager, J. Brozy*, et al.*, *Validation of Multiplex Serology detecting human herpesviruses 1-5.* PLoS One, 2018. **13**(12): p. e0209379.
146. Brenner, N., A.J. Mentzer, J. Butt, K.L. Braband, A. Michel, K. Jeffery*, et al.*, *Validation of Multiplex Serology for human hepatitis viruses B and C, human T-lymphotropic virus 1 and Toxoplasma gondii.* PLoS One, 2019. **14**(1): p. e0210407.
147. Schafer, F., N. Seip, B. Maertens, H. Block, and J. Kubicek, *Purification of GST-Tagged Proteins.* Methods Enzymol, 2015. **559**: p. 127-39.
148. Leung, W.K., E.K. Ng, F.K. Chan, S.C. Chung, and J.J. Sung, *Evaluation of three commercial enzyme-linked immunosorbent assay kits for diagnosis of Helicobacter pylori in Chinese patients.* Diagn Microbiol Infect Dis, 1999. **34**(1): p. 13-7.
149. Domingues, T., H. Mouriño, and N. Sepúlveda, *Analysis of Antibody Data Using Skew-normal and Skew-T Mixture Models* REVSTAT-Statistical Journal, 2022.
150. Migchelsen, S.J., D.L. Martin, K. Southisombath, P. Turyaguma, A. Heggen, P.P. Rubangakene*, et al.*, *Defining Seropositivity Thresholds for Use in Trachoma Elimination Studies.* PLoS Negl Trop Dis, 2017. **11**(1): p. e0005230.
151. Parker, R.A., D.D. Erdman, and L.J. Anderson, *Use of mixture models in determining laboratory criterion for identification of seropositive individuals: application to parvovirus B19 serology.* J Virol Methods, 1990. **27**(2): p. 135-44.
152. Nhat, N.T.D., S. Todd, E. de Bruin, T.T.N. Thao, N.H.T. Vy, T.M. Quan*, et al.*, *Structure of general-population antibody titer distributions to influenza A virus.* Sci Rep, 2017. **7**(1): p. 6060.
153. Formichella, L., L. Romberg, H. Meyer, C. Bolz, M. Vieth, M. Geppert*, et al.*, *Validation of a Novel Immunoline Assay for Patient Stratification according to Virulence of the Infecting Helicobacter pylori Strain and Eradication Status.* J Immunol Res, 2017. **2017**: p. 8394593.
154. Song, H., A. Michel, O. Nyren, A.M. Ekstrom, M. Pawlita, and W. Ye, *A CagA-independent cluster of antigens related to the risk of noncardia gastric cancer: associations between Helicobacter pylori antibodies and gastric adenocarcinoma explored by multiplex serology.* Int J Cancer, 2014. **134**(12): p. 2942-50.
155. Gao, L., A. Michel, M.N. Weck, V. Arndt, M. Pawlita, and H. Brenner, *Helicobacter pylori infection and gastric cancer risk: evaluation of 15 H. pylori proteins determined by novel multiplex serology.* Cancer Res, 2009. **69**(15): p. 6164-70.

156. Coppens, F., G. Castaldo, A. Debraekeleer, S. Subedi, K. Moonens, A. Lo*, et al.*, *Hop-family Helicobacter outer membrane adhesins form a novel class of Type 5-like secretion proteins with an interrupted beta-barrel domain.* Mol Microbiol, 2018. **110**(1): p. 33-46.
157. Alm, R.A., J. Bina, B.M. Andrews, P. Doig, R.E. Hancock, and T.J. Trust, *Comparative genomics of Helicobacter pylori: analysis of the outer membrane protein families.* Infect Immun, 2000. **68**(7): p. 4155-68.

6 Supplement

6.1 Abbreviations

6xHis	Polyhistidine
°C	Degrees Celsius
amp	Ampicillin
APS	Ammoniumperoxodisulfate
BMI	Body mass index
bp	Base pairs
Cad	Cinnamyl alcohol dehydrogenase
cag	*Cytotoxin-associated gene*
CagA	Cytotoxin-associated gene A
cag PAI	*cag* pathogenicity island
CBS-K	Superchemiblock
CDH1	*Cadherin-1*
CGC	Cardia gastric cancer
CI	Confidence interval
cm	Centimeter
C-terminus	Carboxy-terminus
ddH_2O	Double-distilled water
DNA	Deoxyribonucleic acid
DTT	1,4-Dithiothreitol
EBER	Epstein–Barr virus-encoded small RNA
EBV	Epstein-Barr virus
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetate
e.g.	Exempli gratia
EM	expectation maximization
ELISA	Enzyme-linked immunosorbent assay
et al.	*Et alii*
g	g-force
GC	Gastric cancer
GST	Glutathione-S-transferase
h	Hour
HcpC	*Helicobacter* cysteine-rich protein C
His	Histidine
Hor	Hop-related proteins
Hop	*Helicobacter* outer membrane porin
HpaA	*H. pylori* adhesin A
HpGP	Helicobacter pylori genome project
H. pylori	*Helicobacter pylori*
HRP	Horse radish peroxidase
HUSAR	Heidelberg Unix Sequence Analysis Resources
HyuA	Hydantoin utilization protein A
IARC	International Agency for Research on Cancer
ICD-10	International Classification of Diseases 10th Revision
IPTG	Isopropyl β-D-1-thiogalactopyranoside

Kat	Catalase
kb	kilo base pairs
kDa	kilo Dalton
kV	kilovolt
LB	Lysogenic broth
µF	microfarad
µl	microliter
M	Molar
MAD	Median absolute deviation
MALT	mucosa-associated lymphoid tissue
MCC Spain	Multicase-control study Spain
MEGA	Molecular Evolutionary Genetics Analysis
MFI	Median fluorescence intensity
min	minute
ml	Milliliter
mm	Millimeter
mM	millimolar
n	Number
NapA	Neutrophil-activating protein A
NCBI	National Center for Biotechnology Information
NCGC	Non-cardia gastric cancer
NIT	Nutrition intervention trial
N-terminus	Amino-terminus
OD	Optical density
OMP	Outer membrane protein
OR	Odds ratio
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PBS-T	PBS-Tween
PCA	Principle component analysis
PCR	Polymerase chain reaction
PGI	Pepsinogen I
PGII	Pepsinogen II
PGAP	Prokaryotic Genome Annotation Pipeline
PVA	Polyvinylalcohol
PVP	Polyvinylpyrrolidone
PVX	Synonym for PVA plus PVP
r	Correlation coefficient
RAM	random access memory
REF	reference
RNA	Ribonucleic acid
RT	Room temperature
s	second
SD	Standard deviation
SDS	Sodium dodecyl sulfate
SIR2	silent information regulator 2
Strep-PE	Streptavidin-R-Phycoerythrin
SV40	Simian virus 40
T4SS	type IV secretion system

TAG	C-terminal peptide comprising the last seven amino acids of the SV40 large T antigen
TEMED	N,N,N',N'-Tetramethylethylendiamin
tiff	Tagged Image File Format
TMB	Tetramethylbencidine
UK	United Kingdom
UreA	Urease subunit A
USA	United States of America
V5	V protein of simian virus 5
VacA	Vacuolating cytotxin A
VCA p18	Viral Capsid Antigen p18
VCPUs	virtual centralized processing unit

6.2 List of Figures

6.3 List of Tables

6.4 Supplementary tables

Supplementary Tables S1 – S4 can be found on the attached CD-ROM.

Supplementary Table S1: Primer sequences to for all H. pylori expression constructs
Supplementary Table S2: H. pylori antigens on minimized H. pylori 26695-microarray
Supplementary Table S3: Statistical analyses of H. pylori multi-strain microarray probed with sera from the Shanxi study and the NIT
Supplementary Table S4: Sequences of new H. pylori multiplex serology antigens

Supplementary Table S5: Associations, adjusted for all potential confounders, of seropositivity to individual H. pylori *antigens with NCGC status in the NIT serum samples*

Antigen*	Seropositive NCGC cases (n = 322)	Seropositive controls (n = 327)	OR (95% CI)**
GroEL	169 (52%)	147 (45%)	**1.44 (1.05 - 1.99)**
UreA	43 (13%)	39 (12%)	1.18 (0.73 - 1.91)
HP0231	25 (8%)	23 (7%)	1.07 (0.58 - 1.97)
NapA	117 (36%)	78 (24%)	**1.92 (1.34 - 2.73)**
HP0305	183 (57%)	140 (43%)	**1.93 (1.39 - 2.68)**
HpaA	52 (16%)	43 (13%)	1.23 (0.78 - 1.94)
CagAN	236 (73%)	197 (60%)	**1.87 (1.33 - 2.65)**
CagAC	226 (70%)	198 (61%)	**1.65 (1.18 - 2.33)**
HyuAN	70 (22%)	60 (18%)	1.22 (0.81 - 1.81)
HyuAC	83 (26%)	62 (19%)	**1.55 (1.05 - 2.28)**
Catalase	74 (23%)	64 (20%)	1.23 (0.83 - 1.81)
VacAC	194 (60%)	175 (54%)	1.30 (0.94 - 1.81)
HP1098	43 (13%)	41 (13%)	1.10 (0.68 - 1.76)
Cad	27 (8%)	23 (7%)	0.93 (0.51 - 1.69)
HP1564	233 (72%)	205 (63%)	**1.66 (1.17 - 2.34)**
HopA	45 (14%)	29 (9%)	1.63 (0.97 - 2.72)
HP1038	140 (43%)	135 (41%)	1.20 (0.86 - 1.66)
Mal648	39 (12%)	36 (11%)	0.99 (0.60 - 1.64)
HP0003	28 (9%)	20 (6%)	1.41 (0.75 - 2.63)
HP0659	70 (22%)	60 (18%)	1.29 (0.86 - 1.93)
HP0582	23 (7%)	11 (3%)	**2.27 (1.06 - 4.90)**
HP0185	73 (23%)	89 (27%)	0.69 (0.47 - 1.01)
HP0385	75 (23%)	82 (25%)	0.94 (0.65 - 1.37)
HP0527trunc	68 (21%)	53 (16%)	**1.58 (1.03 - 2.40)**
HP1570	29 (9%)	25 (8%)	1.00 (0.56 - 1.79)
HP0477	12 (4%)	7 (2%)	1.65 (0.62 - 4.41)
HP1355	31 (10%)	40 (12%)	0.68 (0.41 - 1.14)
HP0545	116 (36%)	108 (33%)	1.25 (0.89 - 1.76)
Lat98	200 (62%)	219 (67%)	0.79 (0.57 - 1.11)

**Excluded due to low seroprevalences: Lat118, Mal1434, Lat1540, HP1064, HP1091, HP0017N, Kor1294N and HP1435*
***logistic regression analysis adjusted for age, sex, alcohol, smoking and BMI*
Statistically significantly associated are marked in bold; NCGC: non-cardia gastric cancer; OR: odds ratio; CI: confidence interval

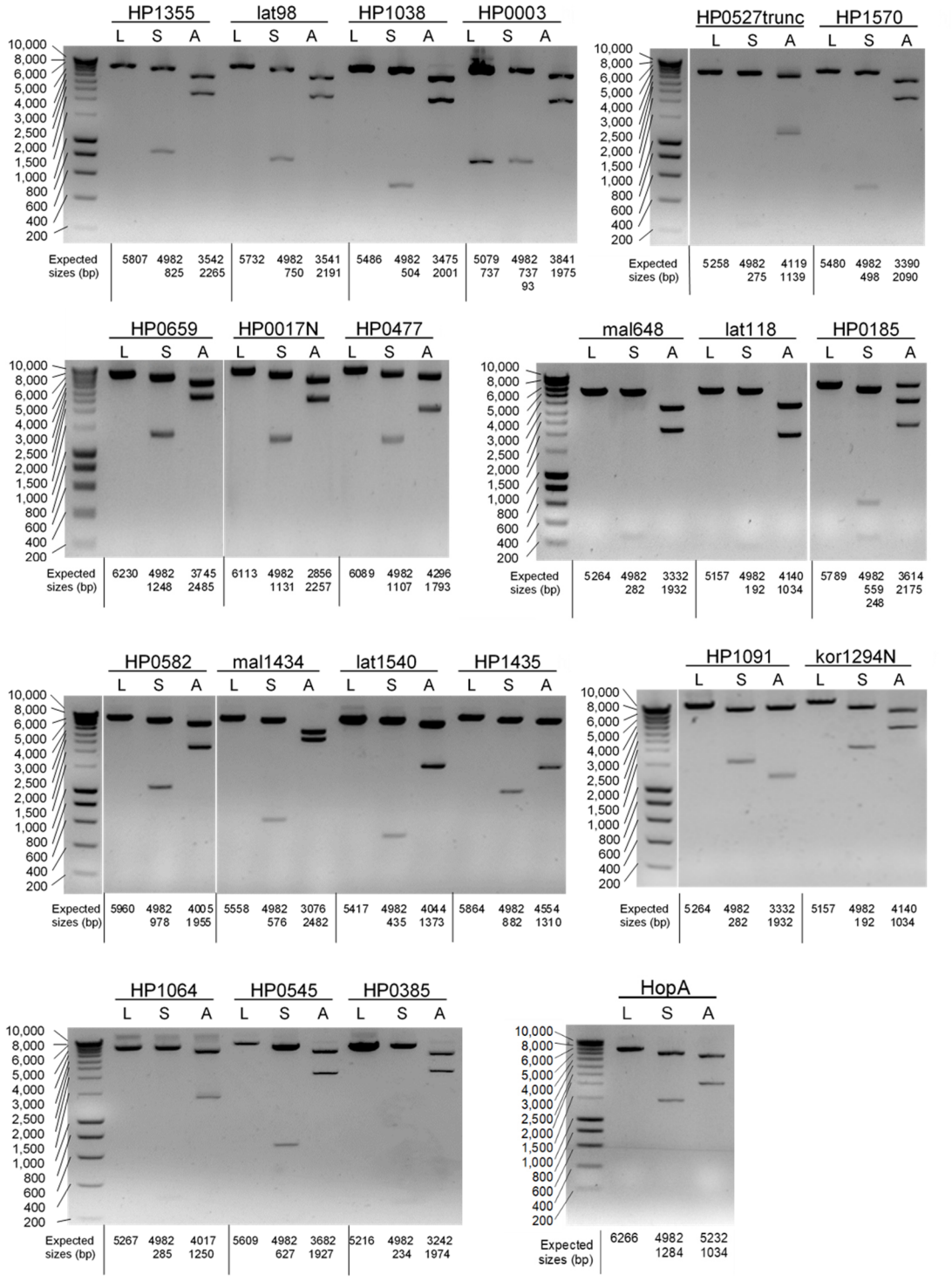
Supplementary Table S6: Associations, adjusted for all potential confounders, of seropositivity to individual *H. pylori* antigens with NCGC status in the NIT plasma samples

Antigen*	Seropositive NCGC cases (n = 113)	Seropositive controls (n = 880)	OR (95% CI)**
GroEL	53 (47%)	376 (43%)	1.22 (0.82 - 1.82)
UreA	15 (13%)	107 (12%)	1.09 (0.61 - 1.96)
HP0231	7 (6%)	89 (10%)	0.57 (0.26 - 1.26)
NapA	37 (33%)	227 (26%)	1.38 (0.90 - 2.10)
HP0305	54 (48%)	280 (32%)	**1.92 (1.29 - 2.86)**
HpaA	10 (9%)	120 (14%)	0.61 (0.31 - 1.21)
CagAN	57 (50%)	412 (47%)	1.21 (0.81 - 1.80)
CagAC	66 (58%)	389 (44%)	**1.88 (1.25 - 2.82)**
HyuAN	20 (18%)	112 (13%)	1.43 (0.84 - 2.43)
HyuAC	23 (20%)	132 (15%)	1.42 (0.87 - 2.35)
Catalase	20 (18%)	146 (17%)	1.06 (0.63 - 1.78)
VacAC	64 (57%)	392 (45%)	**1.67 (1.12 - 2.50)**
HP1098	7 (6%)	113 (13%)	0.47 (0.21 - 1.05)
Cad	9 (8%)	84 (10%)	0.84 (0.41 - 1.73)
HP1564	79 (70%)	479 (54%)	**1.96 (1.28 – 3.00)**
HopA	27 (24%)	136 (15%)	**1.82 (1.12 - 2.94)**
HP1038	37 (33%)	255 (29%)	1.25 (0.82 - 1.91)
Mal648	29 (26%)	227 (26%)	1.01 (0.64 - 1.59)
HP0003	17 (15%)	146 (17%)	0.93 (0.54 - 1.62)
HP0659	18 (16%)	137 (16%)	1.05 (0.61 - 1.81)
HP0582	3 (3%)	64 (7%)	0.36 (0.11 - 1.16)
HP0185	18 (16%)	167 (19%)	0.84 (0.49 - 1.44)
Lat1540	21 (19%)	160 (18%)	1.06 (0.64 - 1.77)
HP0385	20 (18%)	194 (22%)	0.78 (0.46 - 1.30)
HP0527trunc	33 (29%)	209 (24%)	1.33 (0.86 - 2.06)
HP1570	8 (7%)	63 (7%)	0.93 (0.43 - 2.02)
HP0017N	41 (36%)	356 (40%)	0.83 (0.55 - 1.26)
HP0477	6 (5%)	49 (6%)	1.00 (0.41 - 2.41)
HP1355	16 (14%)	112 (13%)	1.10 (0.62 - 1.94)
HP0545	24 (21%)	169 (19%)	1.13 (0.69 - 1.84)
Lat98	86 (76%)	633 (72%)	1.22 (0.77 - 1.93)
HP1091	73 (65%)	518 (59%)	1.24 (0.82 - 1.88)

**Excluded due to low seroprevalences: Lat118, Mal1434 Kor1294N, HP1435 and HP1064*

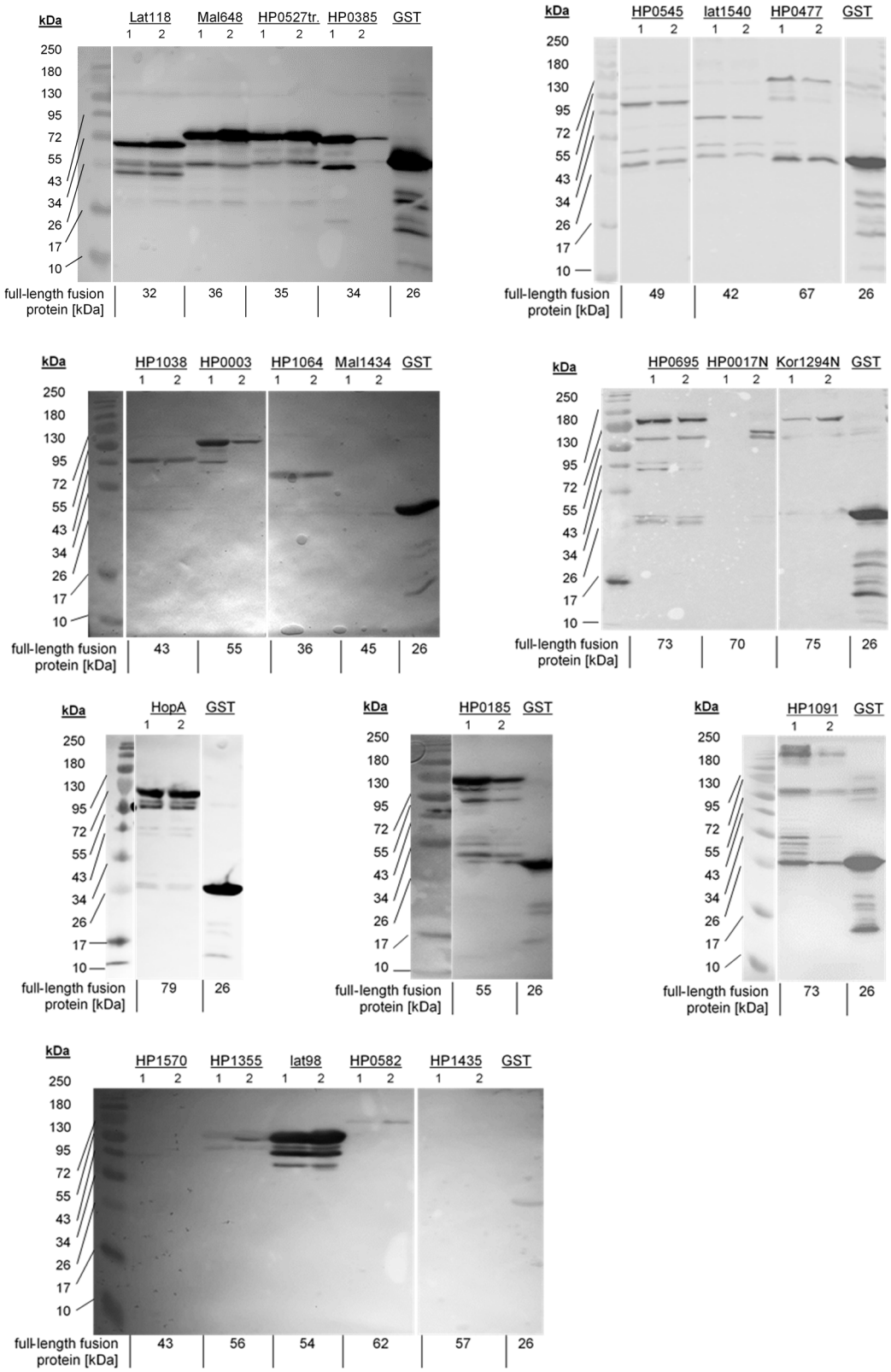
***logistic regression analysis adjusted for age, sex, alcohol, smoking and BMI*
Statistically significantly associated are marked in bold; NCGC: non-cardia gastric cancer; OR: odds ratio; CI: confidence interval

6.5 Supplementary figures



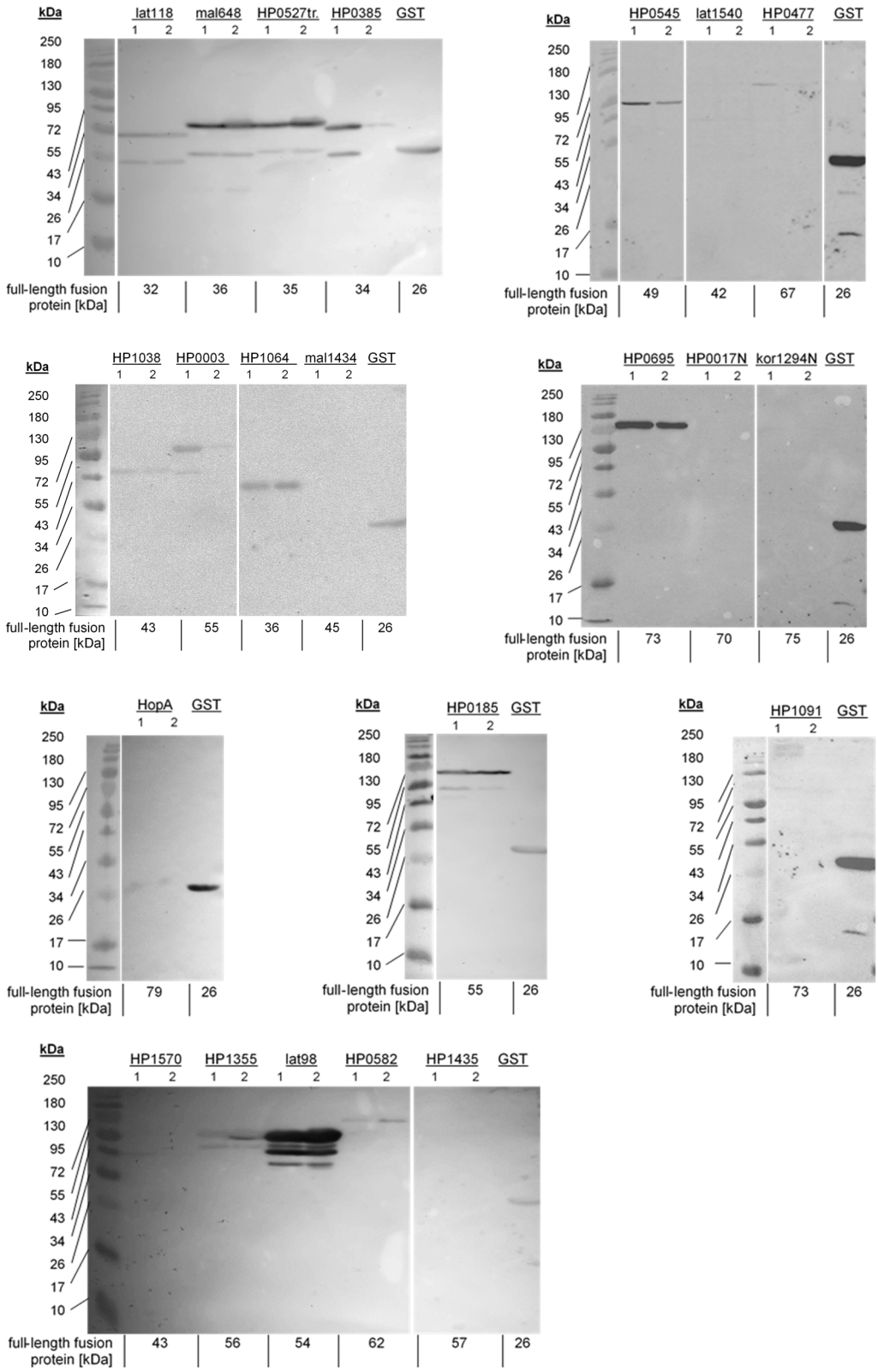
Supplementary Figure S1: Analytical DNA digests

Protein expression plasmids were analyzed by three analytical DNA digests: a linear digest to verify the length of the plasmid (L), a symmetric digest to verify the length of the insert (S) and an asymmetric digest to verify the pre-calculated length of two fragments (A). Respective antigens are stated above the gels and expected fragment sizes below. The first band of each gel shows a DNA ladder and is annotated with the respective sizes [bp].



Supplementary Figure S2: anti-GST Western blots

Respective antigens are stated above the gels and expected protein sizes below. The first band of each gel shows a protein ladder and is annotated with the respective sizes [kDa]. Two samples per protein lysate were analyzed, one taken before clearing the lysate by centrifugation (1) and another after (2)



Supplementary Figure S2: anti-GST Western blots

Respective antigens are stated above the gels and expected protein sizes below. The first band of each gel shows a protein ladder and is annotated with the respective sizes [kDa]. Two samples per protein lysate were analyzed, one taken before clearing the lysate by centrifugation (1) and another after (2)

6.6 Publications

Jeske R, Merle U, Müller B, Waterboer T & Butt J (2022). Performance of Dried Blood Spot Samples in SARS-CoV-2 Serolomics. *Submitted*

Yao, P., Millwood, I., Kartsonaki, C., Mentzer, A. J., Allen, N., **Jeske, R**., ... & Chen, Z. (2022). Sero-prevalence of 19 infectious pathogens and associated factors among middle-aged and elderly Chinese adults: a cross-sectional study. *BMJ open*, *12*(5), e058353-e058353.

Artola-Borán, M., Fallegger, A., Priola, M., **Jeske, R**., Waterboer, T., Dohlman, A. B., ... & Müller, A. (2022). Mycobacterial infection aggravates Helicobacter pylori-induced gastric preneoplastic pathology by redirection of de novo induced Treg cells. *Cell reports*, *38*(6), 110359.

Yang, L., Kartsonaki, C., Yao, P., Chapman, D., …, **Jeske R.**, ... & Chen, Z. (2021). The relative and attributable risks of cardia and non-cardia gastric cancer associated with H. pylori infection in China: a case-cohort study. *Lancet Public Health*, *6*(12).

Jeske, R., Dangel, L., Sauerbrey, L., Frangoulidis, D., Teras, L. R., Fischer, S. F., & Waterboer, T. (2021). Development of High-Throughput Multiplex Serology to Detect Serum Antibodies against Coxiella burnetii. *Microorganisms*, *9*(11), 2373.

Jeske, R., Reininger, D., Turgu, B., Brauer, A., Harmel, C., de Larrea-Baz, N. F., ... & Hufnagel, K. (2020). Development of Helicobacter pylori Whole-Proteome Arrays and Identification of Serologic Biomarkers for Noncardia Gastric Cancer in the MCC-Spain Study. *Cancer Epidemiology and Prevention Biomarkers*, *29*(11), 2235-2242.

Alberts, C. J., **Jeske, R.**, de Martel, C., den Hollander, W. J., Michel, A., Prins, M., ... & Waterboer, T. (2020). Helicobacter pylori seroprevalence in six different ethnic groups living in Amsterdam: The HELIUS study. *Helicobacter*, *25*(3), e12687.

Aistleitner, K., Sieper, T., Stürz, I., **Jeske, R.**, Tritscheller, S., Mantel, S., ... & Wölfel, R. (2020). NOTIFy (non-toxic lyophilized field)-FISH for the identification of biological agents by Fluorescence in situ Hybridization. *PloS one*, *15*(3), e0230057.

Aistleitner, K., **Jeske, R**., Wölfel, R., Wießner, A., Kikhney, J., Moter, A., & Stoecker, K. (2018). Detection of Coxiella burnetii in heart valve sections by fluorescence in situ hybridization. *Journal of Medical Microbiology*, *67*(4), 537-542.