From Department of Medical Epidemiology and Biostatistics,
Karolinska Institutet, Stockholm, Sweden

Molecular Epidemiologic Studies on *Helicobacter pylori* Infection and Stomach Cancer Risk

Zongli Zheng

Karolinska Institutet

Stockholm 2011

Published by Karolinska Institutet. Printed by [name of printer]

ISBN 978-91-7457-405-0

To MingTong
In memory of my mother!

ABSTRACT

Helicobacter pylori (H. pylori) infection increases stomach cancer risk. The aim of this thesis was to study genetic susceptibility from the host and to develop molecular methods for future characterization of bacterial virulence factors in longitudinal cohorts.

In Study I, we investigated the association between genetic variation in an O-glycan transferase encoding gene (*a4GnT)* and *H. pylori* infection and gastric cancer risk in a Polish population-based case–control study (273 gastric cancer patients and 377 controls). A haplotype at rs2622694–rs397266 was associated with *H. pylori* infection, with the A-A haplotype associated with a higher risk compared with the most frequent G-G haplotype (odds ratio 2.30; 95% confidence intervals 1.35–3.92). Neither this haplotype nor the tagSNPs were associated with overall gastric cancer risk.

In Study II, we characterized genomic evolution of *H. pylori* over 20 years in the stomach. Whole genome of 21 sequential isolates 20 years apart, from 7 patients, were sequenced using 454 sequencing platform. There were on average 260 point mutations (range 70 to 488) per isolate over 20 years, and 45 recombinations (range 18 to 92). Genes in the cell motility category were overrepresented in point mutations and recombinations. Specifically, mutations often affected genes involved in chemotaxis, vacuolating cytotoxin-like protein, restriction and type IV secretory pathway; and recombinations affected glycosyltransferase involved in lipopolysaccharide biosynthesis. The major form of single nucleotide substitutions was transition (85%) and the minor form was transversion (15%). Mutation was sequence context-dependent.

Clinical samples are often precious and of trace amounts. In Study III, we developed novel methods for DNA shotgun library construction and quantification. As compared with the standard procedure, our double-stranded and Y library construction methods are simpler and more efficient. A highly sensitive Taqman MGB-probe-based quantitative polymerase chain reaction (qPCR) was developed to quantify the amount of effective library. We also demonstrated that the distribution of library molecules on capture beads follows a Poisson distribution. Combining the qPCR and Poisson statistics, the labor intensive and costly titration can be eliminated and trace amounts of starting material is applicable.

Archived formalin-fixed and paraffin-embedded (FFPE) biopsies, coupled with long term follow-up, are valuable resources for molecular epidemiologic studies. Study IV presented a method based on laser capture micro-dissection and modified whole genome sequencing methods to obtain metagenomic profiles of *H. pylori* from 15-year old FFPE biopsy sections.

Keywords: *Helicobacter pylori*, Stomach cancer, Next-generation sequencing, Formalin-fixed and paraffin-embedded biopsy

LIST OF PUBLICATIONS

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I. Zheng Z, Jia Y, Hou L, Persson C, Yeager M, Lissowska J, Chanock SJ, Blaser M, Chow WH, Ye W. **Genetic variation in *a4GnT* in relation to *Helicobacter pylori* serology and gastric cancer risk**. *Helicobacter*. 2009 Oct;14(5):120-5.

II. Zongli Zheng, Kerstin Stake-Nilsson, Anders F. Andersson, Rolf Hultcrantz, Weimin Ye, Lars Engstrand. **Genomic evolution of *Helicobacter pylori* in human stomach over twenty years.** (Manuscript)

III. Zheng Z, Advani A, Melefors Ö, Glavas S, Nordström H, Ye W, Engstrand L, Andersson AF. **Titration-free massively parallel pyrosequencing using trace amounts of starting material.** *Nucleic Acids Res*. 2010 Jul;38(13):e137.

IV. Zongli Zheng, Anders F. Andersson, Weimin Ye, Olof Nyrén, Staffan Normark, Lars Engstrand. **A method for metagenomics of *Helicobacter pylori* from archived formalin-fixed gastric biopsies permitting longitudinal studies of carcinogenic risk.** (Manuscript)

CONTENTS

LIST OF ABBREVIATIONS

H. pylori	*Helicobacter pylori*
FFPE	Formalin-fixed and paraffin-embedded
PCR	Polymerase chain reaction
RAPD-PCR	Random arbitrarily primed DNA-PCR
BLAST	Basic Local Alignment Search Tool
FDR	False discovery rate
OR	Odds ratio
RR	Risk ratio, Relative risk
CI	Confidence interval
LCM	Laser capture microdissection
COG	Clusters of orthologous group
SNP	Single nucleotide polymorphism
GWAS	Genome wide association study
ELISA	Enzyme-linked immunosorbent assay
ICD-O	International classification of disease for oncology
emPCR	Emulsion PCR
qPCR	Quantitative PCR
bp	Base pair

1 INTRODUCTION

Although the age-standardized incidence of stomach cancer has been declining over recent decades, this cancer is still the fourth most common cancer (7.8% of all cancers) and the second leading cause of cancer-related mortality (9.7% of all cancer deaths) worldwide [1]. Nearly one million new stomach cancer cases happened in 2008 and the number is estimated to increase by 76% in 2030, largely due to aging population [1]. *Helicobacter pylori* (*H. pylori*) colonizes about half of the world population [2] and contributes to about 60% to 90% of stomach cancers [3].

H. pylori has been defined as a strong risk factor for gastric cancer, but only a small subset of infected individuals develop gastric cancer [4]. The magnitude of the risk varies in different parts of the world, to get a hint, 4.7% of infected adults (mean age 52) with non-ulcer dyspepsia in Japan developed gastric cancer over 7.8 years [5]. Individuals having duodenal ulcer, which is one of the sequelae of *H. pylori* infection, however are protected from gastric cancer, observed both in Sweden [6] and in Japan [5]. The underlying mechanisms why the infection in the majority has benign course whereas a subset of *H. pylori*-colonized persons develop different outcomes that are mutually exclusive are not known [4,7].

2 BACKGROUND

2.1 DESCRIPTIVE EPIDEMIOLOGY

2.1.1 Stomach cancer

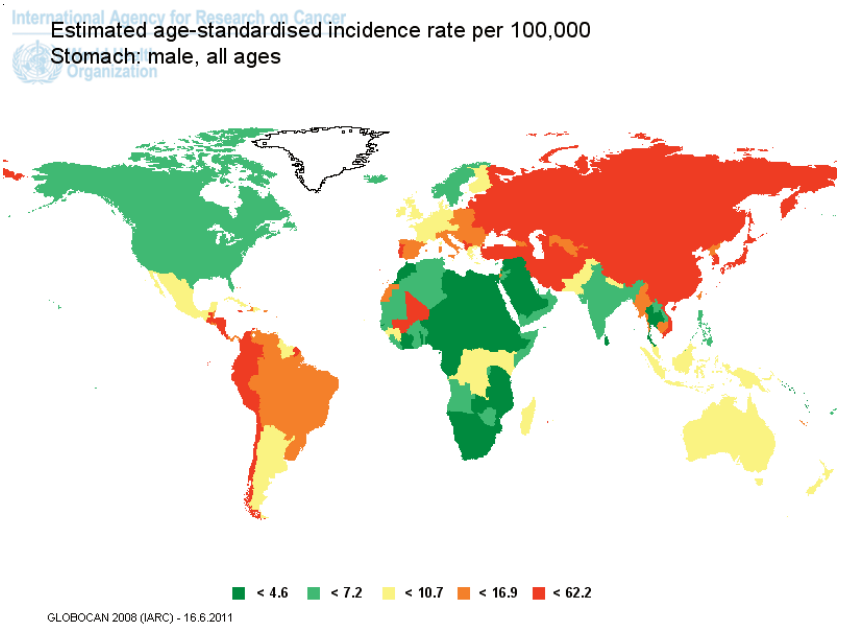


Figure 2.1 Global stomach cancer incidence: estimated age-standardized incidence rate per 100,000, by sex, all ages. Source: GLOBOCAN 2008 (IARC)

Stomach cancer was the most common cancer in the first estimate of global cancer incidence in 1975, and was behind lung, breast and colorectum in 2008. About 50% of stomach cancers occur in Eastern Asia, mainly in China [1]. Age-standardized incidence rates were about twice as high in men as in women, ranging from 3.9 in Northern Africa and 5.8 in Northern America to 42.4 in Eastern Asia for men, and from 2.2 in Southern Africa and 2.8 in Northern America to 18.3 in Eastern Asia for women [1] (Figure 2.1). Overall, the incidence of stomach cancer has declined in the past decades in both high-risk [8,9] and low-risk areas [10].

Stomach cancer has poor survival and is the second leading cause of cancer death in both sexes worldwide, after lung cancer [1]. The highest mortality rates are estimated in Eastern Asia (28.1 per 100,000 in men, 13.0 per 100,000 in women) and a low mortality in Northern America (2.8 per 100,000 in men, 1.5 per 100,000 in women). Five-year relative survival rates approximate 20% in most areas of the world, except in Japan where it approaches 60% owning mostly to a mass screening program, different staging system, and early treatment [8].

2.1.2 Subtypes of stomach cancer

More than 95% of all gastric neoplasms are adenocarcinomas. *H. pylori* infection causes gastric mucosa associated lymphoid tissue (MALT) lymphoma, which could be regressed totally by eradication of *H. pylori* [11]. This thesis focused on gastric adenocarcinoma.

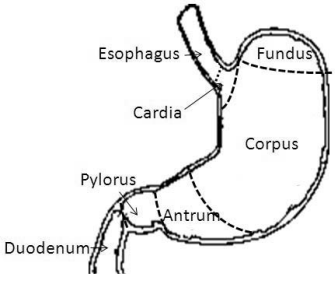


Figure 2.2 Anatomy of the stomach

2.1.2.1 Cardia, non-cardia, and corpus

By anatomy sites, gastric adenocarcinomas can be divided into cardia and non-cardia. Cardia cancer has been defined as an adenocarcinomatous lesion with its center located within 1 cm proximal and 2 cm distal to the esophagogastric junction [12,13] (Figure 2.2). A less precise definition based on ICD-O code (9^{th} version 151.0 and 10^{th} version C16.0) has also been used [14,15,16]. Rising incidence of cardia cancer was first reported in the United State in 1991 [16]. One of the explanations was misclassification

of esophageal adenocarcinoma as cardia cancer [13], as suggested by a study showing stable incidence of cardia cancer identified by meticulous site-specific diagnosis in Sweden [17]. No increase in cardia cancer was noted in Eastern Asia [18,19]. The overall decline in stomach cancer incidence worldwide was due to a decline in non-cardia cancer. The incidence of non-cardia cancer declined among all ethnics and age groups in the United States from 1977 through 2006, except for whites aged 25 to 39 years, for whom it increased [20]. Very recently, among non-cardia cancers, corpus cancer was suggested to be an epidemiologically distinct entity from the rest of non-cardia cancers [21]. In the United States, between 1976 and 2007, corpus cancer incidence rate increased by 1% annually for whites and 3.5% for blacks, and decreased among other ethnics. In a larger but newer database, corpus cancer significantly increased between 1999 and 2007 among younger and middle-aged whites [21].

2.1.2.2 Intestinal and diffuse types

Dr. Laurén subdivided gastric adenocarcinoma into two main histological types, intestinal type and diffuse type, in 1965 [22]. The intestinal type gastric cancer is mainly diagnosed in the elderly and diffuse type in younger age groups. The intestinal type gastric cancer has a long incubation time – about 30 to 50 years and typically evolves from chronic gastritis to atrophic gastritis, intestinal metaplasia, dysplasia and adenocarcinoma – as proposed in the Correa model in 1975 [23]. The intestinal type is more common than diffuse type, and there also exist some tumors showing features of both. The declining gastric cancer incidence during the past decades was attributable to the decreasing incidence of the intestinal type in both high-risk and low-risk areas [24]. A population-based and comprehensive case ascertainment study conducted in Sweden showed that, among non-cardia cancers, the age-standardized incidence of the intestinal type was 7.9, 6.8 and 5.1 per 100,000 person-years in 1989-1990, 1991-1992 and 1993-1994, respectively [17]. The corresponding figures for the diffuse type were 4.5, 4.5 and 3.2 per 100,000 person-years.

2.1.3 Prevalence of *H. pylori* infection

H. pylori colonizes the stomach of about half of the world's population [25,26]. The prevalence of *H. pylori* infection has declined during the last decades, but varies greatly in different countries, age groups and socioeconomic classes [27,28,29]. The prevalence among middle-aged adults ranges from over 80% in the developing countries to 20% - 50% in the industrialized countries [25]. A recent larger study involving 7,465 adults in the US showed that the sero-prevalence of *H. pylori* was 32.5% overall, and was 52.7% among non-Hispanic blacks, 61.6% among Mexican-Americans, and 26.2% among non-Hispanic Caucasians [30]. In Eastern Asia, the sero-prevalence ranges from about 80% to 90% in India and Bangladesh, to 60% in China and Korea, and 40% in Japan [29].

The infection is acquired early in childhood through person-to-person transmission [31]. Culturing for *H. pylori* using the air sampled 0.3 meter away from healthy carriers of *H. pylori* during vomiting could grow positive colonies [32]. By the age of 15 years the sero-prevalence, as determined by the immunoblot assay, was estimated to decrease from 43% in the 1950s, to 36% in the 1960s and 30% in the 1970s in the United Kingdom [28]. In the Dutch population, age 12 -15 years, the sero-prevalence determined by ELISA of anti-*H. pylori* IgG antibodies decreased from 23% in 1978 to 11% in 1993 [27]. The decline suggested a birth-cohort effect, reflecting improvement in hygienic conditions and decreased crowding during childhood for younger generations [27,28].

2.2 STOMACH CANCER RISK FACTORS

Environmental and lifestyle factors may play important roles in the risk for gastric cancer. A recent review concluded that high intake of fresh fruits and vegetables, lycopene and its products, and potentially vitamin C and selenium may be protective [33]. In addition to *H. pylori* infection, suggested risk factors include high intake of nitrosamines, processed meat products, salt and salted foods, and smoking [33,34]. There is little evidence to support the associations between β-carotene, vitamin E, alcohol consumption and the risk for gastric cancer [33]. This thesis focused on host genetic polymorphisms and the genomics of *H. pylori*.

2.2.1 Host genetic susceptibility

A large-scale twin study in northern Europe estimated that heritability accounts for 28% of the variation in susceptibility to gastric cancer [35]. Hereditary diffuse gastric cancer (HDGC) is an autosomal-dominant disease characterized by aggressive and poorly differentiated carcinoma due to germline mutations in the *CDH1* gene which encodes a tumor suppressor intercellular adhesion protein E-cadherin [36].

Other main approaches used to identify the associations between single nucleotide polymorphisms (SNPs) and diseases include the hypothesis-driven candidate gene approach and hypothesis-free genome-wide association study (GWAS).

2.2.1.1 Candidate gene approach

Many studies on genetic susceptibility to gastric cancer have been conducted. A systematic review of meta-analyses was performed focusing on nine genes involved in inflammation (*IL-1β, IL-1RN, IL-8*), detoxification of carcinogens (*GSTs, CYP2E1*), folate metabolism (*MTHFR*), intercellular adhesion (*E-cadherin*) and cell cycle regulation (*p53*) [37]. In the estimations of population attributable risks (PAR, which is

the proportion of gastric cancer cases attributable to the presence of the variant genotype), the most impacting polymorphisms were *E-cadherin* -160A and *IL-1* -511T (which might account for ~20% among Caucasians), *MTHFR* -677T variant and *GSTM1* null (both 10% in Asians and Caucasians), *IL-1RN* *2 (10% in Caucasians) and *IL-8* -251A (10% in Asians) [37].

Relatively few studies of genetic susceptibility have focused on host responses to *H. pylori* within the gastric mucus. The gel layer covering the gastric mucosa is the major reservoir of *H. pylori*. An uneven distribution of *H. pylori* among mucins of varied physiochemical properties, as observed by electron microscopy [38], provides clues to understand how the host environment may influence *H. pylori* survival. *H. pylori* rarely colonize the deeper portions of the normal gastric mucosa [38,39], where the mucins are rich in O-glycans capped with terminal alpha-1,4-linked N-acetylglucosamine (A4GN) [40,41]. Glycans possessing this A4GN residue, but not those without, suppress *H. pylori* growth in vitro [42,43]. The transfer of A4GN to beta-Gal residues with alpha1,4-linkage, forming GlcNAc alpha-1, 4-Gal beta-R structures, is mediated by a transferase encoded by the gene *A4GNT* [40].

The histo-blood group antigen Lewis b (Le^b), which is dominant antigen in the gastric mucosa, was identified to function as a receptor for *H. pylori* [44]. The MUC5AC, a highly glycosylated protein in the mucus, was suggested as the primary receptor for *H. pylori* [45]. In the Portuguese population, which has a relatively high stomach cancer incidence in Europe, the smaller size of *MUC1* variable number tandem repeat (VNTR) alleles were associated with an increased risk for gastric carcinoma [46], as well as chronic atrophic gastritis and incomplete intestinal metaplasia [47]. In addition, the smaller size VNTR alleles of *MUC6* have been associated with *H. pylori* infection [48] and an excess risk of stomach cancer [49]. We previously conducted one study in Poland and the results showed that the *MUC1* haplotype ACTAA at rs4971052-rs4276913-rs4971088-rs4971092-rs4072037 had a nearly doubled risk compared to the common haplotype GTAAG [50]. For *MUC5AC*, the minor allele at rs868903 was associated with an 80% increased risk of stomach cancer [50].

2.2.1.2 GWAS

Although GWAS stands for genome wide association study, it has been typically based on single nucleotide polymorphism (SNP) data, i.e. SNP-based GWAS. Other forms of GWAS are not feasible yet.

To date, the only available stomach cancer GWAS (small chips, two stages: 85,576 and 2,880 SNPs) was from Japan and Korea. The study showed that polymorphisms (rs2976392 and rs2294008) in the prostate stem cell antigen gene (*PSCA*), which

possibly involves in regulating gastric epithelial-cell proliferation, influenced susceptibility to diffuse type gastric cancer (rs2976392: OR = 1.62, $P = 1.11 \times 10^{-9}$; rs2294008: OR = 1.58, $P = 6.3 \times 10^{-9}$) but not to intestinal type gastric cancer [51]. However, an independent validation study in China on these two specific loci showed that the SNP rs2976392 was associated with both intestinal and diffuse types of gastric cancer, and that SNP rs2294008 was not associated with gastric cancer [52].

Following their initial GWAS study [51], the authors continued to study the second most significant locus which was in *MUC1*, and included an additional *MUC1* SNP (rs4072037) [53]. A significant association between rs4072037 and diffuse type gastric cancer was observed and was confirmed with functional study [53]. This was in line with our previous study on this SNP, only that our result pointed to an association for overall gastric cancer (66% intestinal type) in a Polish population [50].

2.2.2 *H. pylori* infection

The International Agency for Research on Cancer (IARC) stated *H. pylori* as a definite (group I) carcinogen to human in 1994 [54]. After evaluation of new evidence, the IARC restated the 1994 statement in 2011 [11]. A comparison in *H. pylori* infection concordance in monozygotic and dizygotic twins estimated that hereditability may explain 57% of infection prevalence [55]. The infection increases risk for non-cardia cancer, but not for cardia cancer in Western countries [11,56,57]. However, one study from China reported an elevated risk for cardia cancer (hazard ratio 1.64, 95% CI 1.26 – 2.14) [58], while another study from Japan showed a non-significant elevated risk [59].

The magnitude of the association between *H. pylori* infection and non-cardia gastric carcinoma risk varied to a great extent in different studies. Cross-sectional studies reported weak association, but prospective studies, where samples were collected many years before stomach cancer diagnosis, generally pointed to strong associations [11], for example, an odds ratio as high as 48.5 was reported when sera were collected at a mean age of 31 (range 16 to 40) and cancer cases diagnosed at a mean age 47 (range 25 to 68) [60]. In studies where *H. pylori* infection was assayed by anti- *H. pylori* IgG ELISA generally an approximately doubled risk of non-cardia cancer was reported [11]. Immunoblot provides higher sensitivity than ELISA and is able to detect low-abundant antibodies that could be missed by ELISA [56]. When immunoblot was used to determine the infection status, the odds ratio increased to 10 - 18 [56,61], and to 68 in one study (antibody against CagA was used) [57]. A meta-analysis of studies shows that CagA-positive strains increase the risk of non-cardia gastric cancer two-fold compared to CagA-negative strains [62].

2.2.3 Antibiotic resistance

When the Scottish bacteriologist Alexander Fleming was awarded the Nobel Prize in 1945, he somewhat prophetically warned that misuse of penicillin would lead to resistance. *H. pylori* eradication trials have demonstrated favourable results for patients with peptic ulcers, pre-cancerous lesions and MALT [63,64,65,66]. After decades of treatment against *H. pylori* infection, antibiotic resistance is becoming more and more common [67]. Success rate of first eradication is often less than 80% in most parts of the world [67]. Opinion leaders have called for a regime with a local eradication rate of > 90% [68]. Development of antibiotic resistance is one of the main concerns why at this stage widespread eradication of all *H. pylori* infections should not be recommended.

2.3 *H. PYLORI* GENOMICS

The carcinogenic action of *H. pylori* infection on gastric cancer is chronic, presumably lasts for decades, during which time profound changes typically occur in the microorganism's niche [69]. In return, *H. pylori* may change its genomic makeup to adapt to the altered micro-environments. Some *H. pylori* pathogenic factors have been characterized such as the *cytotoxin-associated gene pathogenicity island* (*cagPAI*) and *vacuolating cytotoxin A* (*VacA*) [11]. Other unexplored factors in the bacteria genome are likely to be important. The focus of this thesis on *H. pylori* was at the genome level.

H. pylori genome is extremely diverse [70]. Finger printing technique – RAPD-PCR results demonstrated that essentially every two *H. pylori* isolates from different individuals showed different banding patterns [71,72].

Traditional Sanger sequencing was used to sequenced the first *H. pylori* strain 26695 from a patient with gastritis in 1997 [73] and, subsequently, strain J99 from a patient with duodenal ulcer in 1999 [74] and HPAG1 from a patient with chronic atrophic gastritis in 2005 [75]. In the past few years, many *H. pylori* strains (tens, or more than one hundred if including draft genomes) had been sequenced by the means of high-throughput sequencing [76].

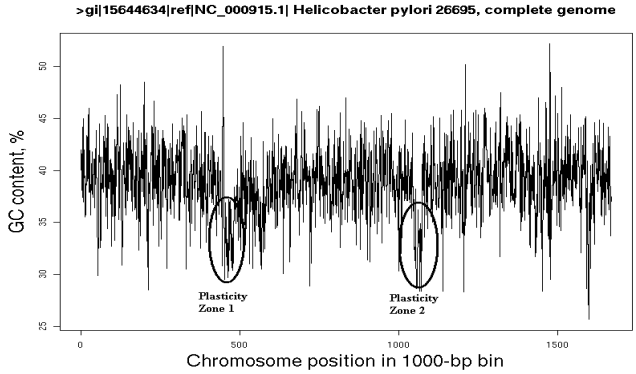


Figure 2.3 GC content of the genome of *H. pylori* 26695, by 1000-bp bin

In the comparison of the genomes of strain 26695 and strain J99, the segments that contain high proportions (46% and 48%) of genes that are unique to one of the strains was named plasticity zones [74]. Plasticity zone segments have low GC content (< 35%, Figure 2.3), as opposite to a percentage of about 39% for the entire genome. A recent study on the plasticity zones of five more strains showed that plasticity zones are complex mosaics of transposable elements (TnPZ) remnants, formed by multiple TnPZ insertions, and spontaneous and transposable element mediated deletions [77]. They were not essential for the viability, but some TnPZ genes affect bacterial phenotypes and fitness [77].

Genomic comparison of *H. pylori* host jump from human to large felines (*Helicobacter acinonychis Sheeba*) revealed profound number of changes resulting in pre-mature stop codons [78]. *H. pylori* appears to lack of mismatch repair function [79]. These findings indicate that *H. pylori* employs efficient strategies to overcome its unusual high genomic mutation rate, which could be deleterious. One previous study showed that about half of the *H. pylori* genome could be exchanged during a course of 41 years [80]. The study was based on only ten housekeeping genes and extrapolated estimates from pairs of sequential isolates at a mean interval of only 1.8 years.

However, it is still largely unknown to what extent and which parts of the genome have been changed in the same *H. pylori* strain after long term colonization in the human stomach. These questions can only be answered with confidence after comparisons of dozens or even hundreds of the genomes with sequential samples collected decades apart. Recent technology advance in high throughput DNA sequencing makes such comparisons possible.

To date, only one study has compared sequential *H. pylori* isolates from the human stomach, at the whole genome level [81]. The authors compared four pairs of sequential

isolates collected three years apart, among them two pairs included additional isolates at 16 years. The four sets of genomes differed by 27 to 232 isolated SNPs and 16 to 441 imported clusters of polymorphisms resulting from recombination [81]. Since the study subjects were from a high risk area and likely harbored mixed strains in their stomachs, the authors further compared evolution of a single vaccine strain from a volunteer and the results showed no evidence of recombination when the stomach is infected with one single strain. However, the time interval of the sampling was only three months.

2.4 DNA SEQUENCING

2.4.1 Shotgun sequencing

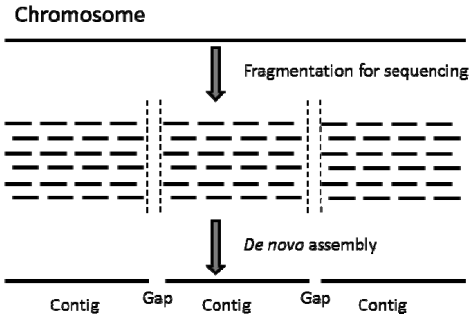


Figure 2.4 Schematic illustration of shotgun DNA sequencing and *De novo* assembly.

Because current DNA sequencing method can only sequence short stretches (50 bp to 1500 bp), long sequences need to be divided into smaller fragments for sequencing, and then pieced together (assembled) into the original long fragments (chromosomes) using the overlapping parts (Figure 2.4). In *de novo* assembly, fragments are pieced together without any prior sequence information. This process requires relative high sequencing depth, such as more than 20 times coverage, to generate good assembly results. Some regions in the chromosome may not be sequenced, either by chance due to an inadequate sequencing depth, repetitive elements or highly difficult structures (homopolymer is difficult for pyro-sequencing, high GC content might be missed by a PCR step in the library preparation for Illunima sequencing), and result in gaps (Figure 2.4).

Mapping refers to alignment of the fragments to a reference genome. The purpose of mapping is usually to look for variants.

2.4.2 High-throughput DNA sequencing technology

Modern DNA sequencing technology has improved dramatically in recent years. One of the main differences in current high-throughput DNA sequencing technology as

compared with traditional Sanger sequencing is that sample template concentration is kept very low to avoid tedious microbial sub-cloning.

Emulsion PCR (emPCR)-based sequencing uses many millions of water-in-oil droplets, each of which serves as a separated amplification compartment [82]. Bridge PCR-based amplification method uses primers covalently mounted on a glass to generate clusters of amplicons [83]. For emPCR, sample library concentration is kept so low that the majority of the droplets contain no library, a small proportion contains single-molecule libraries and an even smaller proportion contains mixed-molecule libraries in a stochastic fashion that follows Poisson distribution. Similarly for bridge-PCR, sample library concentration is kept low so that most of the amplicon clusters are derived from single molecules.

A recent review compared different sequencing platforms in details [84]. The technologies, existing and emerging ones are under rapid developments so that the information available now will be outdated soon, probably in just a few months. Interestingly, the companies often use calendar quarter as the time unit to release new products or upgrades. The videos illustrating the technologies are available from the websites of these platforms (listed below, with Twitter names). Here I summarize very briefly a few key specifications of some platforms.

2.4.2.1 454 sequencing (www.454.com; @454Sequencing; history/upgrades: GS 20, GS FLX standard, GS FLX Titanium, GS Junior Titanium [small version FLX Titanium], GS FLX+)

Roche 454 sequencing was the first commercial massively parallel DNA sequencing platform available in the market after its first publication in 2005 [82]. Among current main sequencing technologies, Roche 454 sequencing is relatively expensive in terms of cost per base. The data per run is about 500 Mbp, 1.2 million reads with modal length 450 bp (~700 bp for FLX+). Its strength lies in its relative long read length than the other common platforms (100 bp in HiSeq2000, 75 bp in SOLiD). Despite its relatively high cost, it is still commonly used for applications such as *de novo* sequencing of microorganisms and amplicon sequencing. The technology (pyro-sequencing) is prone to homopolymer errors (e.g. the amount of lights generated from 6 adenosines [As] in a stretch is smaller than six times the amount of a single A, and may be mistakenly regarded as 5 As) [85].

2.4.2.2 Illumina sequencing (www.illumina.com; @illuminainfo; history/upgrades: Solexa, Genome Analyzer, HiSeq, MiSeq [small version of HiSeq])

The Illumina sequencing is currently the leader in the market share. It has the highest throughput per run (up to 600 Gbp for HiSeq) among all current platforms. The read

length is ~100 bp. The platform is suitable for most applications. The standard Illumina library construction involves a PCR step, which is associated with GC content bias [86].

2.4.2.3 SOLiD (www.appliedbiosystems.com, @LIFECorporation, history/upgrades: SOLiD2, SOLiD3, 5500[xl] SOLiD)

SOLiD sequencing is the third player in the next-generation sequencing market. Its throughput is also high (up to 300 Gbp), and the read length is ~75 bp. This platform has the highest sequencing accuracy (> 99.99%), which is important in the clinical setting.

2.4.2.4 Ion Torrent – Personal Genome Machine (www.iontorrent.com; @iontorrent; upgrades of the chips: 314, 316)

The IonTorrent sequencing was the first post-light sequencing platform. Its strength lies in its fast turnaround time – within one day, as compared with about 2 days for 454 and 1 - 2 weeks for Illumina and SOLiD. The fast speed makes it particularly suitable for pathogen identification. As an example, the microorganism responsible for the recent outbreak of EHEC in Europe was quickly sequenced using the Ion Torrent sequencing by Beijing Genome Institute, followed by a quick crowd-sourcing annotation of the data (https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki). The relative small size machine is also an advantage for small laboratories or hospitals.

2.4.2.5 Pacific Biosciences (www.pacificbiosciences.com; @PacBio)

The PacBio sequencing is a single-molecule real-time sequencing (SMRT) platform and generates longest reads currently available (> 1000 bp). Despite its low sequencing accuracy (currently ~ 91%), the accuracy is constant from the start till the end of a read. The accuracy in all other current available platforms are high (> 99%) at the starts of reads, but drop as they read toward the ends. The long read length is particularly useful for *de novo*, amplicon and copy number variation sequencings. Another very attractive and unique feature is the ability to read epigenetics (methylation) directly – based on the slowdowns when the polymerase encounters methylation sites during real-time reading [87].

There are other platforms such as Helicos (the first single-molecule sequencing platform) which has relatively short reads (35 bp) and has mainly been used for transcriptome sequencing (RNA counts). Complete Genomics has been focused on human genome sequencing and only provides sequencing service (samples in, data out). Emerging technology such as nanopore might provide a totally different way of sequencing. However, each technology has its own strength, and it appears that more

and more studies are using more than one platform, such as 454 plus HiSeq, to complement each other.

3 AIMS

The overall aim of this thesis was towards identification of molecular risk factors for *H. pylori*-associated stomach cancer.

- To study the associations between the genetic variations in one host gene and *H. pylori* infection, and gastric cancer risk.

- To characterize genomic mutations and recombinations of *H. pylori* in the time frame of its carcinogenic actions, presumably decades in the human stomach.

- To develop methods allowing high-throughput DNA sequencing technology using trace amounts of starting material.

- To develop methods for sequencing microbial genomes in decades-preserved formalin-fixed and paraffin-embedded gastrointestinal biopsies.

4 MATERIALS AND METHODS

4.1 CASE-CONTROL STUDY

Paper I used a population-based case-control study on gastric cancer conducted in Warsaw, Poland between 1994 and 1996 [88]. Cases (n = 464) were patients newly diagnosed with histologically confirmed gastric cancer. Controls (n = 480) were randomly selected from the computerized population registry and frequency matched to the cases by sex and age (± 5 years). Blood samples were obtained from 305 (65.7%) cases and 427 (90.0%) controls. Genomic DNA, after uses for previous studies, was available from 273 cases and 377 controls.

Serum levels of IgG antibodies against *H. pylori* whole cell antigens and antibodies against CagA were measured using ELISA, as described [89].

4.2 *H. PYLORI* CULTURE ISOLATES

Paper II used *H. pylori* isolates from patients with dyspepsia who underwent gastroscopies in both 1990 and 2010, at Sandvikens, Gälve, Sweden. Frozen biopsies taken from the corpus were retrieved for culturing of *H. pylori*. Random amplified polymorphic DNA (RAPD) using primer sets 1283 and 1290 as described previously [90] was used to screen for isolates showing consistent patterns between one isolate collected in 1990 and two isolates collected in 2010. A total of 21 isolates from 7 patients were subjected to Roche 454 sequencing.

Paper III used the *H. pylori* strain HPAG1, previously isolated from a patient diagnosed with chronic atrophic gastritis [75].

Paper IV used two frozen biopsies from patients who underwent gastroscopies in 1995 and were diagnosed with non-cardia gastric cancer, from a previous endoscopy clinic-based study [91]. At the time of this study, the frozen biopsies were retrieved and used for *H. pylori* culture.

4.3 FORMALIN-FIXED AND PARAFFIN-EMBEDDED (FFPE) BIOPSY

Paper IV used two FFPE biopsies collected at the same gastroscopy as the frozen biopsies described above. These biopsies had been presevered in ambient temperature since 1995.

4.4 LASER CAPTURE MICRODISSECTION

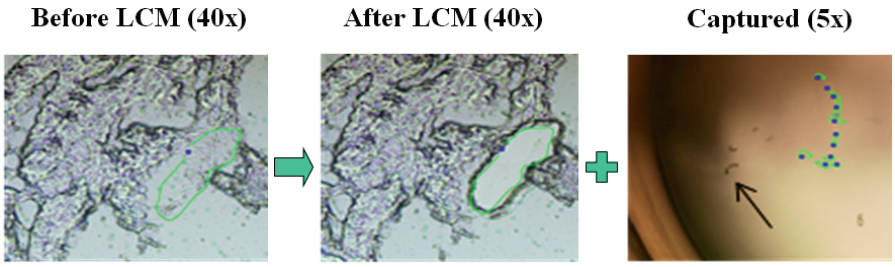


Figure 4.1 Laser capture microdissection (LCM) of *H. pylori*.

From each of the FFPE biopsy blocks, six sections, each 5 µm thick, were cut with disposable and autoclaved microtome blade. The surface section was discarded to reduce contamination. The other 5 sections were placed on autoclaved de-ionized warm (about 45°C) water. To guide the microdissection of *H. pylori* on the adjacent de-paraffinated sections, the second and sixth sections were collected on glass slides for immuno-histochemistry staining of *H. pylori*. The third to fifth sections were collected on molecular grade membrane-mounted PALM slides (Carl Zeiss). These sections were de-paraffinated and followed by laser capture microdissection (Carl Zeiss PALM System, Figure 4.1) of *H. pylori*, which constituted the two metagenomics *H. pylori* samples in Paper IV.

4.5 DNA SEQUENCING

Roche 454 Genome Sequencer FLX Titanium platform was used for DNA sequencing in Papers II, III and IV.

4.6 BIOINFORMATICS ANALYSES

De novo assembly and sequence mapping were performed using Newbler, an algorithm best accommodating pyrosequencing data and is integrated in the Roche 454 gsAssembler and gsMapper.

In the analyses of sequence variants, we analyzed only those mapping results that were deemed to be of high confidence by the software. The gsMapper application uses a combination of flow signal information, quality score information, and difference type information to determine if a difference is of High-Confidence. The general rules are i) there must be at least 3 non-duplicate reads with the difference, unless the expected number (- e option) of reads with the difference is specified, in which case there must be 10% of the expected depth having the difference; ii) there must be both forward and reverse reads showing the difference, unless there are at least 5 reads with quality scores over 20 (< 1% error rate) or 30 (< 0.1% error rate) if the difference involves a 5-mer or higher; iii) if the difference is a single-base overcall or undercall, then the reads with the difference must form the consensus of the sequenced reads (i.e., at that

location, the overall consensus must differ from the reference) and the signal distribution of the differing reads must vary from the matching reads (and the number of bases in that homopolymer of the reference).

Most sequencing platforms generate sequences with accuracy higher than 99%, but are not 100%. Computer algorithms for genome assembly and mapping are not perfect either. Even though cautions have been taken to generate 'High-Confidence' results, in the comparisons of genomes that contain about 1.65 million bases (a very small genome as compared with other organisms), the artifacts stem from sequencing and computer algorithms might be substantial. To overcome this, a novel self-mapping approach was used to reduce errors caused by these artifacts in the analyses of genomic evolutions in Paper II. During self-mapping, raw reads were compared with (aligned to) their *de novo* assembled contigs. We applied this self-mapping approach to a previously published *H. pylori* strain V225d [92] that had been sequenced using Roche 454 sequencing as well. The results showed that there were 22 high-confidence variants when comparing V225d raw reads with V225d assembled contigs. These within strain variants might arise also from, besides the aforementioned technical errors, potential true phenomenon for this highly diverse bacterium. Nevertheless, in whichever case it would confound the strain evolution over time. Thus, these variants identified by self-mapping within the first time isolates in 1990 were subtracted from the variants (overtime comparison) identified in the 2010 isolates to identify true changes over time.

The number of variants for strain V225d identified by self-mapping (n = 22) is consistent with the number of variants identified in the 1990 isolates in our study (range 9 to 41, mean 24), except one isolate (ID p539) showing extraordinary high number of variants (n = 209). This indicates that the isolates cultivated from this patient's 1990 biopsy might contain multiple strains. RAPD patterns for the isolate collected from this patient, one in 1990 and two in 2010, were non-distinguishable, indicating that the same multiple strains co-existed in one stomach over 20 years. In the analyses of genomic changes over time, sequence data from this patient were excluded to ensure valid estimations on *H. pylori* genomic changes over time.

We applied the same cutoff (200 bp) as in a previous study [81] to categorize the variants into two groups, namely isolated single nucleotide change (iSNC) and clustered single nucleotide change (cSNC). When the variants occurred very close (<= 200 bp) to each other, they are more likely to be a result of recombination/import rather than point mutations. To annotate those sequences subjected to point mutation, 60 bases flanking the point mutation were aligned (Standalone NCBI BlastX 2.2.24+ [93]) to the eight annotated *H. pylori* genomes available from GenBank. These strains are 26695, J99, HPAG1, Shi470, G27, P12, B38 and B8. Genomes of two other strains PeCan and

SJM which were also available but lack of annotation at the time of this writing were not included. Both of the sequences in question (60 bases with and without the mutation nucleotides from 2010 and 1990 isolates, respectively) were retrieved for the alignment and, if hit a coding region, the resulting amino acids were compared to identify mutation type (synonymous or non-synonymous). Median and maximum import lengths were calculated for each isolate. To annotate imports, the entire import sequences were aligned (BlastX) to the eight GenBank genomes.

4.7 STATISTICAL METHODS

In **Paper I**, Hardy–Weinberg equilibrium was tested using Pearson's χ^2-test in cancer-free controls. Odds ratios (ORs) with 95% confidence intervals (95% CIs) derived from unconditional logistic regression models were used to assess relative risks. Firth's penalized maximum likelihood estimation was used in case of data separation [94]. Haplotypes were inferred with comparison groups (infected vs uninfected, or cancer cases vs controls) jointly by an Expectation-Maximization algorithm, and analyzed by a "sliding-window" approach with varied window sizes ranging from 2 to 4 tagSNPs [95]. The probabilities, inferred from the EM algorithm, of having certain haplotypes for each individual were used as weights in a logit binomial model [96].

To account for increased type I errors of multiple testing, statistical significance was assessed by empirical *P* values derived from 10,000 permutations (**Paper I**) or adjusted *P* values from the Benjamini and Hochberg method (FDR, **Paper II**).

Generalized estimation equation (GEE) model was applied to account for non-independency among observations – several haplotypes with different probabilities for one individual (**Paper I**) and the mutations and recombinations observed in replicate *H. pylori* isolates from the same patients (**Paper II**).

Poisson distribution was used to predict the distribution of DNA molecules in micro-droplet (**Paper III**). The probability that there are exactly *k* DNA molecules on one library capture bead ($k = 0, 1, 2, ...$) is equal to:

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!},$$

where λ is the input DNA-to-bead ratio. Therefore, the probability of having zero and one DNA molecule on one bead is $e^{-\lambda}$ and $\lambda \bullet e^{-\lambda}$, respectively, and the enrichment fraction is $1 - e^{-\lambda}$.

The precision of a laboratory assay was evaluated by the mean and standard error of the coefficient variation from a serial of triplicate samples (**Paper III**).

The number of genes being sequenced increases with sequencing depth, but levels off at certain point beyond which few new genes are sequenced compared to the amount of additional data. In **Paper IV**, to assess how unique sequences gain with increasing sequencing depth, simulations (n = 1000) with replacement was used. Each simulation contained a specified number of randomly sampled sequences, increased in steps of 100 up to the total number of sequences available. The number of unique sequences for each simulation dataset was calculated, as was the average number derived from 1000 simulation datasets. These average numbers (Y axis) were plotted as a function of the sample numbers (X axis).

Data were analyzed using R (http://www.R-project.org/), SAS 9.2 (SAS Institute, Cary, NC) and Linux shell scripts.

5 RESULTS AND DISCUSSIONS

5.1 STUDY I

We hypothesized that genetic variation in *A4GNT*, which could lead to varied expression or function of this transferase, may be related to the ability of *H. pylori* to establish colonization, and consequent gastric cancer risk. We tested these hypotheses using blood samples from a case-control study conducted in Poland.

All seven tested SNPs were in Hardy-Weinberg equilibrium. SNP rs2622694 heterozygous genotype AG was associated with a 49% reduced risk of *H. pylori* infection compared with the most common genotype AA (OR 0.51, 95% CI 0.26-1.01). A similar magnitude of reduced risk was observed for the homozygous variant genotype GG (OR 0.52, 95% CI 0.21-1.31). For SNP rs397266, the heterozygous genotype AG conferred a 68% higher risk of *H. pylori* infection than the most common genotype GG. All carriers (n=29) of the homozygous variant genotype AA had *H. pylori* infection. Armitage trend test revealed a *P* value of 0.003, which remained significant after correction for multiple tests (0.036). In haplotype analysis, among the fifteen sliding windows, three (rs2622694-rs397266, rs2246945-rs2622694-rs397266 and rs329386- rs2246945-rs2622694-rs397266) had haplotype profiles that differed between the infected and uninfected groups. These three windows all pointed to a significant effect of haplotype at loci rs2622694-rs397266. Compared with the most common (G-G) haplotype, haplotype A-A was associated with a significantly higher risk of *H. pylori* infection (OR 2.30, 95% CI 1.35-3.92, global *P* 0.0019). This association remained significant after correction for multiple tests (empirical global *P* value 0.045).

None of the tested SNPs was associated with gastric cancer risk, in both single-locus and haplotype analyses.

In vitro studies showed that *A4GNT* encodes a transferase that helps forming certain structure of O-glycans [41]. This glycans can inhibit *H. pylori* growth by inhibiting the biosynthesis of cholesteryl-alpha-D-glucopyranoside, a major *H. pylori* cell wall component [42,43,97]. In contrast to an expected beneficial effect of the A4GNT gene product, two small clinical series studies showed up-regulated expression of *A4GNT* mRNA in patients with gastric cancer [98]. One explanation for this discrepancy may be the existence of functionally impotent subclasses of the A4GNT-encoded enzyme. This is supported by the observation that the expression of the A4GNT enzyme, but not of the A4GNalpha1-4Galbeta-R (a mucous gland specific mucin specific glycan that suppresses *H. pylori* growth), was up-regulated amongst patients with *H. pylori* gastritis

and decreased to normal level after *H. pylori* eradication [39]. The up-regulation of A4GNT enzyme accompanied by unchanged amount of A4GNalpha1-4Galbeta-R suggests that there exist subclasses of A4GNT with little or no transferase activity. Hosts with up-regulated expression of non-functional A4GNT subclasses would be less capable of preventing *H. pylori* from establishing colonization and, consequently, could have more severe clinical outcomes, such as gastric cancer [98].

Examining into the amino acid sequence of A4GNT [40] that had been used in the previous *in vitro* studies [42,43], in which the A4GNT, producing a glycan suppressing *H. pylori* growth, had an alanine at position 218 (corresponding to a C allele at rs2246945), consistent with our finding that the C allele rs2246945 was associated with reduced risk for *H. pylori* infection.

In the study, *A4GNT* variation was associated with *H. pylori* seropositivity but was not associated with overall gastric cancer risk. The proportion of the risk haplotype A-A at loci rs2622694-rs397266 was 17% among the uninfected group in this study. Such a relatively low prevalence requires a larger sample-size study to examine for moderate or modest effects of *A4GNT* on gastric cancer risk.

5.2 STUDY II

To characterize genomic mutations and recombinations of *H. pylori* in the time frame of its carcinogenic actions, presumably decades in the human stomach, we sequenced whole genome of 7 pairs of sequential *H. pylori* isolates from 7 patients, with 1 isolate collected in 1990 and 2 replicates collected in 2010 for each patient, totally 21 isolates.

Number of point mutation and import over 20 years

Point mutations and imports scattered across the genomes. The number of point mutations over 20 years was on average 261 (range 70 to 488, Table 1). These mutations affected mostly outer membrane encoding genes (*hopP/sabA, hopZ, hopS/babA, hopK, hopL, hopU/babC*), and genes involving in chemotaxis and flagellar biosynthesis, tRNA synthetase, *cagPAI, vacA*, restriction and modification system and fucosyl- and glyco-transferase. The mean number of imports per isolate was 45 (range 18 to 92). The median length of imports ranged from 49 to 151 bp for all the 12 isolates, with max length of import 2,241 bp observed in one isolate. The imports had been mostly observed in genes having unknown functions, followed by genes encoding outer membrane proteins, *cagPAI*, fucosyl- and glyco-transferase, chemotaxis and flagellar biosynthesis, *vacA*, replicase, restriction and modification system.

Isolate	Number of variants present in 1990	Total number of variants after exclusion of variants present in 1990	Isolated single nucleotide change (iSNC), n				Clustered single nucleotide change (cSNC), n		
			Total	Synonymous	Non-synonymous	Non-coding /intergenic	n	length, median	length, max
p3:1	3	325	203	144	16	43	28	107	783
p3:2	3	327	198	138	16	44	29	92	535
p64:1	14	499	324	235	15	74	36	121	898
p64:2	16	503	328	232	17	79	35	151	898
p70:1	22	799	488	399	25	64	92	87	756
p70:2	28	422	323	254	19	50	31	63	417
p121:1	3	700	235	169	25	41	67	95	1031
p121:2	5	740	241	175	20	46	76	88	1031
p138:1	8	556	290	218	25	47	39	114	624
p138:2	7	184	70	49	3	18	18	121	589
p230:1	5	583	207	149	18	40	43	92	2119
p230:2	6	635	220	153	22	45	51	49	2241
p539:1	96	277	162	116	11	35	35	106	522
p539:2	73	188	80	47	9	24	27	82	546

Table 1. Number of variants in the isolates sampled in 2010 (two isolates :1 and :2) as compared with the isolate sampled in 1990, with mutation types, after exclusion of self-mapping variants already present in 1990

Function category analysis revealed that point mutation and import (recombination) occurred more often in un-annotated genes, 40% and 62%, respectively, as compared with about 29% un-annotated genes in the eight *H. pylori* genomes currently available in GenBank (Figure 5.1). Among annotated genes, point mutations were overrepresented in the function categories N (cell motility), K (transcription), D (cell cycle control, cell division, chromosome partitioning) and P (inorganic ion transport and metabolism), all with statistical significance. Although point mutations occurred in high frequency in category M (cell wall/membrane), its difference in relative proportion with the eight GenBank genomes did not reach statistical significance. Among annotated genes, imports were highly overrepresented in the function category N (cell motility) with statistically significance as compared with the GenBank reference genomes. Imports also occurred in high frequency in the category U (intracellular trafficking), but was not statistically significant.

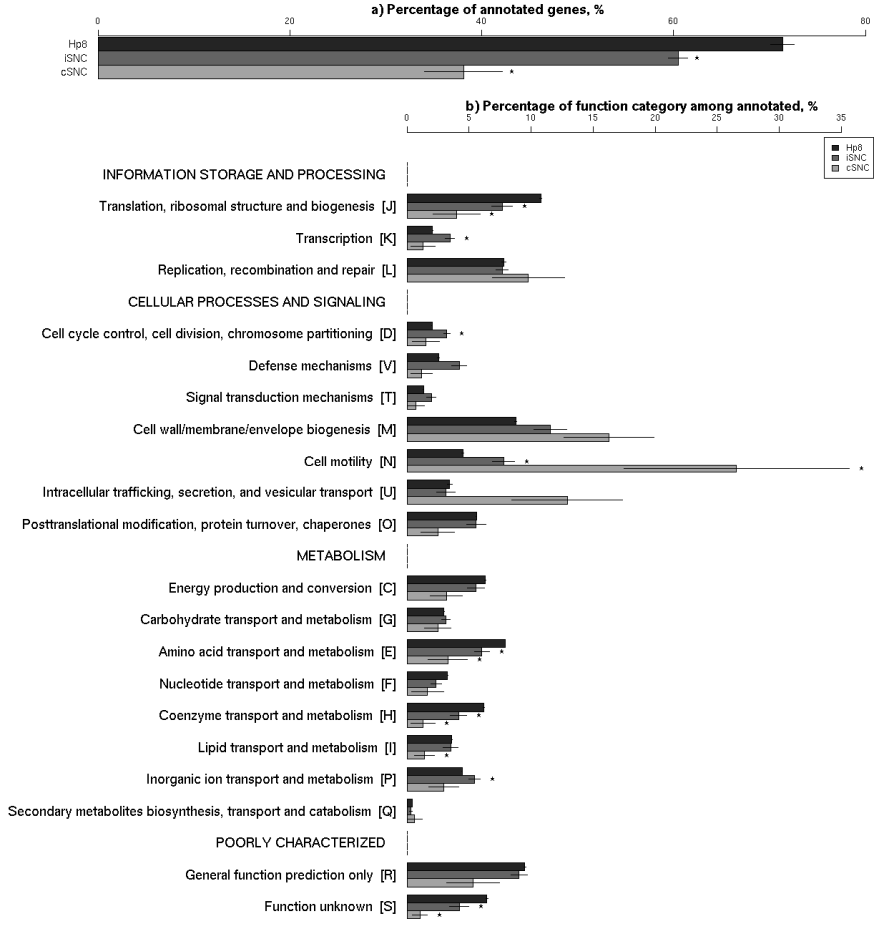


Figure 5.1 Percentage of annotated genes (**a**) and function categories (**b**) among the 8 GenBank *H. pylori* genomes (Hp8), the isolated single nucleotide change (iSNC) and clustered SNC (cSNC) identified in the 12 isolates over 20 years. Horizontal bars indicate standard errors.

The analysis of clusters of orthologous group (COG) showed that point mutations frequently affected COG5651 (vacuolating cytotoxin [*vacA*]-like protein), COG0840 (methyl-accepting chemotaxis protein), COG0610 (type I site-specific restriction-modification system, R [restriction] subunit), COG0086 (DNA-directed RNA polymerase, beta subunit), COG1674 (DNA segregation ATPase FtsK/SpoIIIE) and COG1538 (nickel cobalt outer membrane efflux protein), with their relative proportions statistically significantly higher than those in the eight GenBank reference genomes. For the import group, overrepresented COGs were COG2948 (DNA transformation competence protein ComB10 / VirB10 components / type IV secretion system / *cagY*, $P = 0.02$) and those without statistical significance COG5651 ($P = 0.12$), COG3306

(lipopolysaccharide biosynthesis protein, $P = 0.054$), COG5527 (replication initiation protein A, $P = 0.37$) and COG0840 ($P = 0.22$)

Genomic fusion was a frequent event in *H. pylori* and about 8 fusions per strain had accumulated after 20-year infection in the human stomach. These fusions mostly affected outer membrane genes. About two events of duplications, particularly tandem repeat, had accumulated over 20 years per strain. These duplications occurred mostly in lipopolysaccharide related genes and DNA transformation competence protein ComB10, VirB10 components, and type IV secretion system (*cagY*). Deletion, insertion and substitution were relatively rare events and occurred in sequences with unknown functions.

We observed that 85% of single nucleotide substitutions belonged to transition and 15% to transversion. Mutation sequence context analysis revealed that an A or G nucleotide at the 3' of dinucleotide GC, namely the third nucleotide in GCA or GCG, was subjected to mutation more often than other sequence context (Figure 5.2).

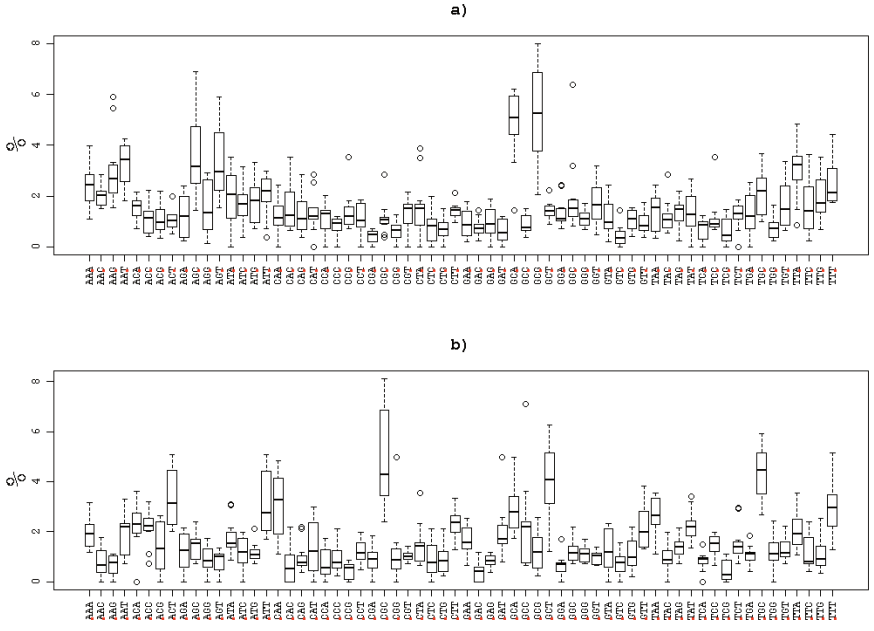


Figure 5.2. Percentages of single nucleotide subjected to mutation (dotted in red), among all isolated single nucleotide changes (iSNCs) occurring in a *H. pylori* isolate after 20-year colonization in the human stomach, in the context of different dinucleotide. The most frequent mutation site was an A or G at the 3' of dinucleotide GC (**a**) or a T or C at the 5' of dinucleotide GC (**b**), or a G the 5' of dinucleotide CT (**b**).

We observed that extensive mutations had occurred in *H. pylori* over 20 years, further supporting previous notion that *H. pylori* is one of the most diverse bacteria. In line with mutations observed in many bacteria, transition, rather than transversion, is the major form of single nucleotide substitution in *H. pylori*. Mutation frequency is associated with specific sequence context that a nucleotide A or G is a hot mutation spot if downstream to the dinucleotide GC.

In the stomachs infected with single *H. pylori* strains, we observed also frequent clustered nucleotide changes which are likely representing recombinations. A previous study comparing a single strain vaccine indicated no recombination [81]. However, it was only 3 months apart for the isolates in comparison. The numbers of recombination over 20 years after single strain colonization we observed (18 - 92) were smaller than the numbers of recombination observed in the previous study (16 - 441). This might be due to colonization of mixed strains in the previous study, which was conducted in a high *H. pylori* prevalent area [81]. However, the sources of homologue recombination for these single strains are unknown. Transient *H. pylori* or homologue genes in other gut microbe might contribute to the recombination. This hypothesis however requires further investigations.

5.3 STUDY III

Current high-throughput DNA sequencing technologies generally require substantial amounts of starting material such as more than 100 nanogram of double-stranded DNA. However, clinical samples are often precious and of low amounts. We designed double-stranded library protocols (simplified AB and Y, Figure 5.3) for the Roche 454 platform to avoid the yield-reducing steps associated with single-stranded library preparation. We designed a highly sensitive Taqman MGB-probe based quantitative PCR method for library quantification (Figure 5.4).

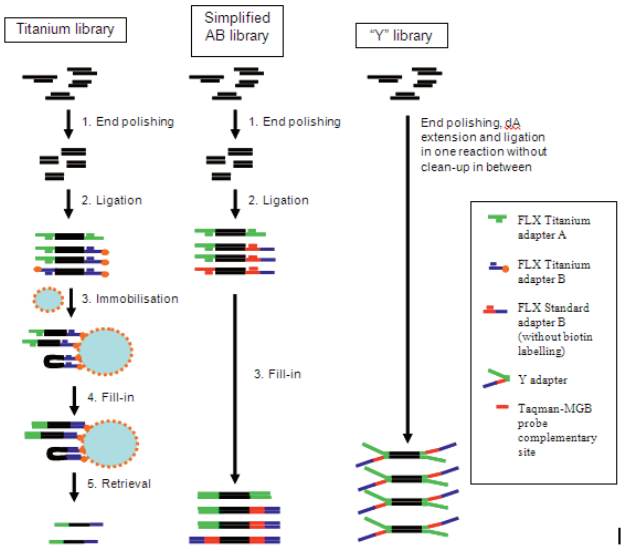


Figure 5.3 Schematic illustrations of three types of library construction.

Our Taqman MGB-probe-based qPCR facilitates quantification of minute amounts of sequencing library, and quantify only those library molecules that are functional. We demonstrated that the distribution of DNA on beads after emulsification follows a Poisson distribution, which enabled us to predict enrichment percentage, a key index for successful sequencing. The ability to avoid labor-intensive and costly titration assay in the standard protocol by combining Poison statistics and our qPCR setup was evaluated retrospectively using previously sequenced FLX Standard libraries and prospectively using newly generated libraries. There was no significant difference between observed and predicted enrichment percentages (paired t-test $P = 0.92$). The precision and reproducibility of this qPCR assay (coefficient variation 9.5%) is at least as good as a digital-qPCR based on a microfluidic system (coefficient variation 11.8%) [99], which is expensive and not generally available in ordinary laboratories.

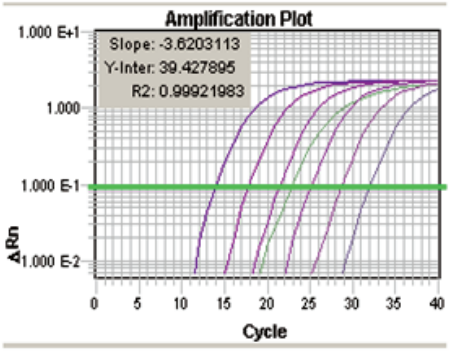


Figure 5.4 Library quantification by qPCR

With the Taqman-MGB probe-based qPCR and Poisson distribution, we were able to avoid the costly and labor-intensive titration assay. Using only one nanogram of fragmented DNA, we prepared enough library (1.15 million amplifiable AB library molecules and 53.6 million Y library molecules) for Roche 454 Titanium sequencing without the need for template pre-amplification by various means of whole genome amplification [100,101]. However, it should be acknowledged that some factors causing sample loss remained, such as fragmentation of genomic DNA, low efficiency of ligation and a limited recovery at the enzymatic reaction clean-up [102]. This was clearly shown by a much higher yield of Y library (sticky-end ligation) than AB library (blunt-end ligation and two additional reaction clean-up steps).

Library quantification by qPCR has been earlier proposed to overcome the lower detection limits of conventional methods [99,103]. We used a Taqman MGB-probe to take advantage of the fact that Taqman MGB-probe provides significantly higher sensitivity, specificity and reproducibility than conventional Taqman probes [104].

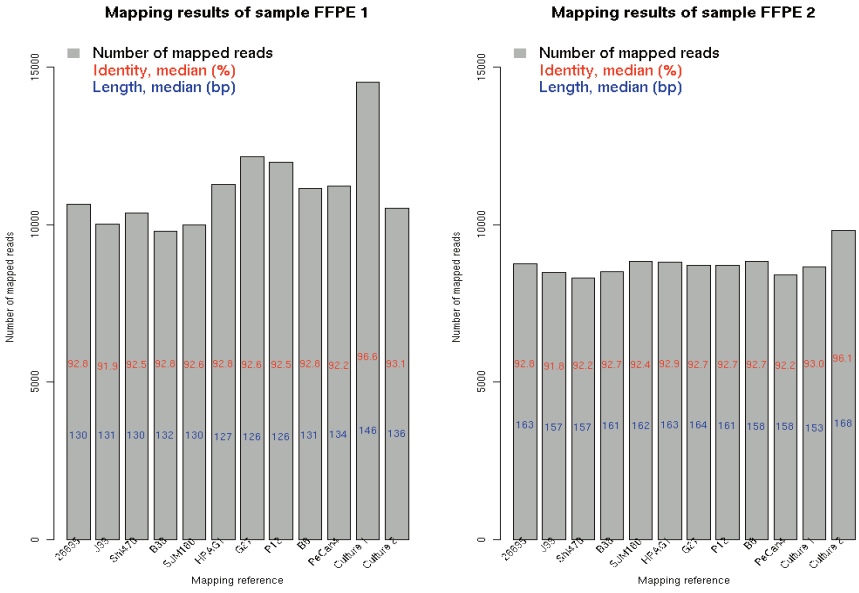
Poisson distribution demonstrates that imprecise library quantifications within two-fold over- or under-estimations will all give satisfactory results when DNA-to-bead ratio is low. In contrast, the same level of inaccuracy will lead to undesired results when DNA-to-bead ratio is high. In a previous study, a linear regression model was used to correlate input DNA-to-bead ratio with enrichment percentage [105]. The linear regression method may suffer from limitations of a positive intercept (meaning that there will be enriched beads even with no input library) and unlimited enrichment percentage (can be higher than 100% when DNA amount increases). However, when the input DNA-to-bead ratio is low, prediction from a linear regression approaches that from a Poisson distribution and is then acceptable.

5.4 STUDY IV

To be able to sequence metagenomic profiles of *H. pylori* from old FFPE samples, we used laser capture microdissection (LCM) for physical separation of *H. pylori* (Figure 4.1), followed by a limited number of cycles of DNA pre-amplification, the improved DNA library construction method for 454 sequencing from Paper III, and customized barcoded Y adapters (Zheng et al, Nature Protocols, in press).

Higher identity was observed when aligning FFPE metagenomic samples with the culture samples from the same hosts (~ 96.4%) than with the reference genomes or with the culture sample of the other patients (~ 92.5%, Figure 5.5). This demonstrates that

the metagenomic profile of *H. pylori* dissected from 15 year old FFPE sections faithfully represents the *H. pylori* populations in its host.



***** Strains HPAG1, G27, P12, B8 and PeCan1 carry plasmids.

Figure 5.5 Alignment identity, length and number of reads of two FFPE samples against each of the ten GenBank *H. pylori* reference genomes and the culture samples of the two patients, respectively

When aligning culture raw sequences with culture assembled contigs (self-mapping) from the same patient, the identity was about 98.7%. However, higher than 99% identity is expected when the same *H. pylori* strain is compared using 454 sequencing and the software gsMapper, as we have shown previously [106]. The lower self-mapping identity suggests the presence of quasispecies in the isolate sweeps. The further lower identity (~ 2.3% lower) when comparing FFPE reads with culture assembled contigs may be due to error prone translesion synthesis during PCR using formalin-fixed DNA [107]. However, given an estimated translesion synthesis error rate of only 0.1% per base [107], other conceivable explanations might be i) a patchy distribution of *H. pylori* in the stomach leading to different strains of *H. pylori* being sampled in the FFPE biopsy and in the culture biopsy; ii) the existence of viable but nonculturable coccoid and/or degenerative forms [108,109] of *H. pylori in vivo* which could only be sequenced with metagenomic methods.

An alternative to microdissection may be a hybridization-based method, which could be used for enrichment-and-sequencing or for direct detection. This approach has been widely used, e.g. for exon sequencing. However, hybridization methods rely on current and thus probably incomplete information on the genomic makeup of the studied

organism, which may be a limitation in the enrichment of highly diverse microbial genomes. Also, hybridization relies on good sequence matching and good hybridization conditions (buffer, temperature, time, etc.) to minimize false hybridizations and missed targets.

Our exploration of sequences common to FFPE and culture in the same patient but not aligning with the comparably large selection of reference genomes demonstrated that even in an imaginary perfect hybridization with probes covering the entire genome of all ten reference strains, some true but highly variant *H. pylori* sequences would still be missed. The omission of true *H. pylori* reads is not surprising as *H. pylori* is one of the most diverse bacteria [70]. One question that might arise concerns how to identify these missed true sequences when, in reality, re-cultures are unavailable in most of the cases. By using our proposed metagenomic approach, functional genes displaying significant difference between cases and controls can be determined using e.g. BLASTx against the "nr" database followed by comparisons of their frequencies in the two groups, without the need for closely matching reference genomes.

The amount of input DNA required in subsequent steps is also critical for the choice of enrichment method. The three commercial hybridization-based methods all require input DNA quantities above one microgram [110], sometimes necessitating pre-amplification. In the previous study using microdissection to analyze microbes from fresh colon biopsies [111], about three nanogram of DNA was dissected from ten 0.8-μm thick sections and was amplified by a multiple displacement amplification method, which could have been unnecessary if the library preparation protocol had been adjusted for nanogram input DNA with the capability to yield tens of millions of reads. Our DNA yield was much less (not detectable by a fluorometry method) due to moderate *H. pylori* density, thinner sections, and degraded DNA and therefore still requires pre-amplification. Whole genome amplification is liable to bias, varying with the amplification method [101], with the multiple displacement amplification method giving less bias than others. However, as multiple displacement amplification requires fragment longer than 500 bp (http://www.qiagen.com/products/repli-gffpekit.aspx#Tabs=t1, last accessed June 11th, 2011), its application on decade-preserved FFPE biopsies is limited. We therefore applied a PCR-based method and amplified for 14 cycles (instead of recommended 25 cycles) to reduce bias. GC content is the most established risk factor for amplification bias in PCR-based methods [101]. We observed similar GC contents in the amplified and sequenced *H. pylori* reads (40% and 41%) as compared with the reference genomes (~39%). Further, the fairly evenly distributed alignment of amplified FFPE *H. pylori* sequences to the reference *H. pylori* genome demonstrated that the amplification bias was minor, although not totally absent.

One important limitation of LCM is that it is a labor intensive task and thus may hamper its use in large scale studies. However, it needs to be balanced, among other considerations, between the amount of work required for LCM and the amount of compromised results acceptable from a quick and prior knowledge-dependent hybridization approach. The balance will shift towards in favor of the latter if the samples are not very precious and are widely available. Ultimately, if one would study different *H. pylori* located in different areas of one biopsy, such as those attached to the epithelium from those able to penetrate deep in the gastric gland and interacting with surrounding stem cells, it may be only possible to study with a physical separation approach.

In all, physical separation through microdissection followed by sequencing provides an appealing alternative to commercial hybridization-based methods when microbial DNA needs to be separated from host, with an additional advantage of being able to capture novel and highly variant genes. Application of this method on FFPE material from the distant past might hopefully unveil markers of potential carcinogens in *H. pylori* strains colonizing the stomach before irreversible damage has occurred, and thus pave the way for targeted chemoprevention.

6 GENERAL DISCUSSIONS

6.1 ON EPIDEMIOLOGICAL ISSUES

Epidemiologic experiments are often limited by ethics and constrains in time, cost and compliance of individuals. For example, a randomized, placebo-controlled, field trial of *H. pylori* eradication among 1630 healthy carriers of *H. pylori* from a high risk area, Fujian Province, China, recruited in July 1994 and followed up until January 2002 had resulted in 7 incident cases of gastric cancer in the eradication arm and 11 in the placebo arm [64]. The size of the study, about 11,000 person-years in total, was too small to adequately assess the effect. Another trial involved 3,365 adults (Linqu, Shandong, China, also a high risk area) with 1,142 individuals randomized to receive *H. pylori* eradication, showed a significant improvement in precancerous gastric lesions after eradication, however there was still lack of statistical power to conclude its effect on gastric cancer [63].

Observational approach has been the main tool for epidemiologists to observe what occur in the natural experiments. In ecologic studies, the units of observation are groups of people and are often hypothesis generating, followed by further investigations. For example, the observations on the association between the prevalence of *H. pylori* infection and the prevalence of gastrointestinal diseases in different areas opened questions for later studies [112,113].

In a cohort study, exposure information of all subjects in a source population is collected first, and then followed up over time to ascertain disease incidence. In a case-control study, conversely, outcome status (case/control) is identified first, followed by assessment of their exposure history. Nested case-control study is an elegant design having the advantages of both cohort study and case-control study, for example, a case-control study on stomach cancer nested within a defined FFPE biopsy cohort. Since stomach cancer is a rare disease even among a group of subjects who undergo gastroscopy, case-control design allows quick recruitment of hundreds of cancer biopsies. Because the biopsies had already been incidentally collected before the outcome disease, a nested case-control study design can control for reversed causality which exists in a conventional case-control design. For example, in the study of carcinogenic factors in related to *H. pylori* infection, the *H. pylori* strains collected at the time of cancer diagnosis may have adapted to the altered stomach environment in the cancer patients, rather than the original ones that responsible for the gastric carcinogenesis. This reversed causality occurs due to two preconditions: 1) *H. pylori* genome has rapid genomic mutation and recombination rates and 2) the incubation time for gastric cancer is long enough to accumulate genomic changes in *H. pylori*.

Selection bias causes invalid results in both case-control and cohort studies. In case-control study, selection bias can occur if the controls are not representative of the population where cases arise. In Paper I, 79% of eligible controls donated blood. The final controls included in the study might have been more likely to be *H. pylori*-infected than those persons (21%) who refused to donate blood, due to probable symptoms caused by *H. pylori* infection. This would lead to an underestimated association between *H. pylori* infection and gastric cancer. For our study on the association between genetic polymorphisms in *A4GnT* and gastric cancer, where *H. pylori* infection is supposed to be a mediator, if the aforementioned selection bias had existed, the associations would have been underestimated. However, this would not affect our study on the association between *A4GnT* polymorphisms and *H. pylori* infection among controls. In a cohort study, selection bias can also occur such as loss of follow-up. For example, *H. pylori* infected individuals may die of cardiovascular diseases (competing risk) on the way of gastric cancer development. If the probability of dying of these diseases among non-infected individuals is lower than that of infected group, namely an association between *H. pylori* infection and cardiovascular diseases, the relative risk of *H. pylori* infection on gastric cancer would be biased towards 0 (although unlikely to across the value 1, from a risk factor to a protective factor, it is theoretically possible if the competing risk is strong enough).

Information bias (misclassification) might exist in Paper I if genotyping and serology assays were incorrect. If misclassification of exposures is differential between cases and controls, the association can either be over- or underestimated. This could have happened if the case and control samples were assayed in different batches. Non-differential misclassification of a dichotomous exposure would bias the association towards null. For an exposure having three or more levels, non-differential misclassification could lead either under- or over-estimation. However, here is a bit of 'unfair' to claim the misclassification as non-differential, as the comparisons were made between two levels (e.g. level 2 vs level 1, and level 3 vs level 1), whereas the non-differential misclassification pertains to all levels (which from the perspective of the two levels involved is indeed differential). In Paper I, if the serology exposures (dichotomous) were misclassified differentially among cases and controls, for example more false-negative of *H. pylori* infection in the cases (due to disease-associated low abundance of anti-*H. pylori* antibodies, which were lower than the detection limit of the ELISA assay) than in controls, a true positive association between *H. pylori* and gastric cancer would be biased towards null. Our study of the association between *A4GnT* polymorphisms and *H. pylori* infection was restricted among controls.

In epidemiological studies, often times, researchers are looking for associations that are causal. Non-causal associations are also useful such as in the search for biomarkers to identify cases. Causal associations are useful to prevent diseases. However, conversions of causal factors might not necessarily reduce disease risk – it might be already too late. One example is that *H. pylori* eradication among individuals having advanced precancerous gastric lesions showed no benefit in terms of preventing the development of gastric cancer, whereas the eradiation among relative healthier (gastritis) individuals helps [64].

Confounding – a confusion of effects [114]. The association between yellow fingers and lung cancer is not causal because of confounding effect by smoking. The theory of directed acyclic graph (DAG) was originally developed in the setting of artificial intelligence [115] and later in epidemiologic setting [116,117]. A modified DAG was used in an attempt to address clinical problems [118]. This approach has helped to adjust for confounding, but relies on comprehensive knowledge about study variables involved. As compared with the complex of human body, not to mention the even more complex social behavior, our knowledge in biomedicine is very limited. It is often unclear whether exposure A is yellow fingers and exposure B is smoking. Thus, DAG might not be a practically useful approach in clinical setting. A non-knowledge based approach, i.e. randomization, will be an effective way to address this issue.

For nested case-control study of rare diseases, there are usually enough eligible controls to sample from. In incidence density sampling, a sampling set (risk-set) contains one case and a set of controls who were free of the diseases at the time of the diagnosis of the index case, with some additional matching factors being the same as those in the index case, such as age, gender and geography area of residence. Randomization is then applied to select the controls for the index case. Occasionally, one control may be eligible for different cases at different time points. It is also possible that one individual may be a control for another individual and later become a case. Some randomization algorithm might specifically avoid these to happen. However, this avoiding actually creates bias; for example, resulting in a group of ‘superman’ among controls, though the magnitude of bias may be small.

Residual confounding might affect the precision of estimates. For example, when age is categorized into < 50, 50 – 60 and > 60, all values in a group is treated as the same. However, we know that 50 and 60 years do differ substantially. For a continuous variable, the difference of every unit is treated equally. In Paper I, the Armetage trend test assigns value 0, 1, 2 to homozygous common allele, heterozygous, and homozygous rare allele genotypes, respectively. As a consequence, the effect of

homozygous rare is assumed to be the square that of heterozygous, which might not be the case and result in imprecise estimate.

6.2 ON MULTIPLE TESTS CORRECTION

Two types of errors can occur in a statistical test, i.e. to reject the null hypothesis when it is true (type I error) or accept the null hypothesis when it is false (type II error). A statistical significance level 0.05 per hypothesis is commonly used as the acceptance level of type I error (false positive). As the number of tests available for one hypothesis increases, the chance of false positive for the hypothesis under study increases. Several common methods exist for controlling false positive including Bonferroni adjustment, permutation, and false discovery rate (FDR).

Bonferroni adjustment uses a significance level 0.05 /n, where 'n' is the number of tests performed. The underlying assumption is that all the tests are completely independent. However, the variables such as SNPs under investigation, for one hypothesis, are often times correlated to some extent. Thus, Bonferroni adjustment is too stringent and resulting in a high false negative.

In contrast, the permutation method preserves the correlations of the tested variables, as it only shuffle the dependent variable in a dataset. For the null hypothesis, the dependent variable (case/control) is not associated with the independent variables (exposures). Thus, the probability of getting a *P* value as extreme as the one observed in the actual data can be empirically calculated in a large number of permutation datasets. For each permutation dataset, a set of multiple tests were performed and the minimum *P* value was recorded. It is in this set-wise manner, the statistical significance is maintained at the hypothesis-wise level. The permutation method is the most accurate method for controlling multiple tests, but suffers from computational constrains when the number of tests is large, and become impossible to perform in the omics era. Figure 6.1 is an example of permutation for the haplotype analysis, which was part of the analyses in Paper I. It took 17 hours (from 20Feb09 to 21Feb09) to run logit binomial model, with sandwich covariance, on 10,000 permutations of a small dataset (273 cases and 377 controls) for 15 haplotype sliding windows using SAS 9.2, run in batch mode, on a desktop PC with duo CPUs, 4 G RAM (the speed was comparable with a Unix server at my department). And, before running the permutations, the efficiency of the program was already improved, for about 10 times faster, from my first draft of the program.

```
start # 1 of 10000 permutations, of sliding _SNP1to2, 20FEB09:17:35:33
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
```

…from the 2nd round, part of log information is turned off (warnings are still on).

```
MPRINT(PERM_GENMOD):   ods listing close;
MPRINT(PERM_GENMOD):   options nonotes nosource nomprint;
start # 2 of 10000 permutations, of sliding _SNP1to2, 20FEB09:17:35:33
start # 5 of 10000 permutations, of sliding _SNP1to2, 20FEB09:17:35:34
start # 10 of 10000 permutations, of sliding _SNP1to2, 20FEB09:17:35:37
start # 20 of 10000 permutations, of sliding _SNP1to2, 20FEB09:17:35:41
start # 50 of 10000 permutations, of sliding _SNP1to2, 20FEB09:17:35:53
start # 100 of 10000 permutations, of sliding _SNP1to2, 20FEB09:17:36:13
start # 200 of 10000 permutations, of sliding _SNP1to2, 20FEB09:17:36:49
start # 500 of 10000 permutations, of sliding _SNP1to2, 20FEB09:17:38:55
start # 1000 of 10000 permutations, of sliding _SNP1to2, 20FEB09:17:42:15
start # 2000 of 10000 permutations, of sliding _SNP1to2, 20FEB09:17:49:06
start # 5000 of 10000 permutations, of sliding _SNP1to2, 20FEB09:18:09:09
start # 10000 of 10000 permutations, of sliding _SNP1to2, 20FEB09:18:35:08

Sliding _SNP2to3: rs405265 rs11928535

start # 1 of 10000 permutations, of sliding _SNP2to3, 20FEB09:18:35:09
```

…the process continues until all the 15 sliding windows were completed.

```
Sliding _SNP4to7: rs329386 rs2246945 rs2622694 rs397266

start # 1 of 10000 permutations, of sliding _SNP4to7, 21FEB09:09:15:53
start # 2 of 10000 permutations, of sliding _SNP4to7, 21FEB09:09:15:54
start # 5 of 10000 permutations, of sliding _SNP4to7, 21FEB09:09:15:56
start # 10 of 10000 permutations, of sliding _SNP4to7, 21FEB09:09:15:58
start # 20 of 10000 permutations, of sliding _SNP4to7, 21FEB09:09:16:03
start # 50 of 10000 permutations, of sliding _SNP4to7, 21FEB09:09:16:16
start # 100 of 10000 permutations, of sliding _SNP4to7, 21FEB09:09:16:38
start # 200 of 10000 permutations, of sliding _SNP4to7, 21FEB09:09:17:25
start # 500 of 10000 permutations, of sliding _SNP4to7, 21FEB09:09:19:45
start # 1000 of 10000 permutations, of sliding _SNP4to7, 21FEB09:09:23:46
start # 2000 of 10000 permutations, of sliding _SNP4to7, 21FEB09:09:31:53
start # 5000 of 10000 permutations, of sliding _SNP4to7, 21FEB09:09:55:42
start # 10000 of 10000 permutations, of sliding _SNP4to7, 21FEB09:10:35:52
```

Figure 6.1 After initial code efficiency improvement, it took 17 hours (from 20Feb09 to 21Feb09) to run permutation for the analysis for a small study.

In a list of declared statistical significant tests (rejected null hypotheses), FDR controls the expected proportion of false positive (incorrectly rejected null hypotheses) [119]. The FDR method deals with only raw *P* values so that computational resource is not a concern. Although not as accurate as permutation, FDR is the approach of choice for large number of tests in the omics era.

6.3 SNPS IN A DIPLOID CELL

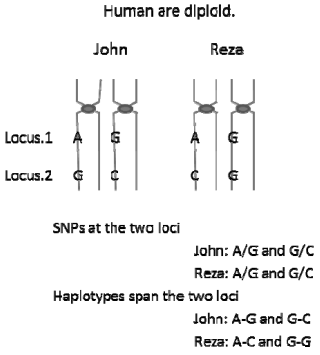


Figure 6.2 Identical SNPs data but different haplotypes in diploid cells.

Single nucleotide polymorphism (SNP) represents genetic information at a single point on a chromosome. For monoploid cell such as most bacteria, genomic variation is wholly represented by SNPs, except some tandem repeats or copy number variations. As the ploidy increases (diploid, triploid, tetraploid…), genomic variations become more and more complex and more difficult to be represented by SNPs. Human are diploid, meaning that a human somatic cell contains two complete haploid sets (sperm and egg gamete). One SNP corresponding to two nucleotides, one inherited from mother and the other from father. Two or more SNPs contain nucleotides in a 'non-connected' form. For example, John's genetic blueprint at locus 1 is A/G and locus 2 G/C (Figure 6.2). Reza has same SNPs data as John at the two loci. However, their genetic makeups at these two loci are clearly different. In molecular biology, genetic information is read through in a stretch (haplotype) by polymerases. The difference in John and Reza's haplotypes at these two loci could result in different amino acids being translated, different levels of mRNA expressions in their downstream region or difference in on/off status of a downstream gene, despite identical SNPs data.

As the number of involving SNPs increases, the possibility of different haplotypes increases rapidly, while the actual number remains two. For example, if a haplotype spans 20 heterozygous loci, over one million of haplotypes (2^20 = 1,048,576) could theoretically be constructed. When high-density oligonucleotides arrays were used to genotype hundreds of thousands of SNPs across the entire genome, genome wide association study (GWAS) became possible. Various statistically methods have been developed to infer (guess) most likely haplotypes, for example, based on maximum-likelihood estimate. The author's computer can only calculate possible combinations of haplotypes up to 1023 heterozygous loci (2^1023 = 9E307; 2^1024 = infinite). Although it might never involve so many loci in a stretch for a certain biology process, 20 loci can already result in 1 million possible guesses of haplotypes and lead to extensive computation in the data analysis for a case-control study

involving hundreds of subjects. Statistical method can be improved for easier computation – a set of SNPs can be partitioned into small units first and recombined subsequently. However, even for haplotypes spans only 20 heterozygous loci, infer 2 true haplotypes (or let's not so picky and allow 20 'representative' haplotypes) from 1 million possibilities, to me, is kind of mission impossible. Thus, statistics is unlikely to be able to faithfully restore a stretch DNA sequence from dozens of SNPs, without losing critical biological information. Although haplotypes at certain significant regions identified by SNPs may be obtained with confidence by re-sequencing, haplotypes that are informative but were not picked up by SNPs in the first place are lost. The loss of critical biological information is likely to be worse for more complex diseases, such as cancers.

6.4 GENOMIC ENRICHMENT

DNA enrichment is useful for hypothesis-driven studies to zoom in certain genomic regions of interest (or genomes of interest such as in Paper IV) and allows for more samples to be sequenced or the same amount of samples with much higher sequencing depths (for identifying rare key mutations in a bulk of majority benign tissue cells).

6.4.1 Sensitivity

Sensitivity – the percentage of target being sequenced over total length of a target. It increases with increasing sequencing depth and levels off at a certain point beyond which little is gained compared to the amount of additional data. The sequencing cost can be reduced by reducing sequencing depth to the level where the sequencing depth versus gene yield curve levels off and the marginal cost per additional gene becomes prohibitive. Thus, more samples, such as in an epidemiological setting, could be sequenced with the same budgetary constraint. The sequencing redundancy in three commercial hybridization-based methods, to achieve a similar level of target (67%, the approximate level-off point for all the three methods), was 57 to 171 times [110]. Foreseeably, sequencing redundancy is rapidly becoming less of a concern as the sequencing cost continues to fall.

6.4.2 Specificity

Specificity – the fraction of sequenced reads that aligned to the region of interest. For the three commercial methods, the number of reads that aligned to the region of interest over total number of reads ranged from 52% to 59% [110]. In Paper IV, about 5% of DNA fragments that we dissected from the FFPE sections could be mapped to *H. pylori*, and the rest was mostly human and unknown. Considering the size of the human genome (~ 3.3 billion base pairs) and that of *H. pylori* (~ 1.65 million base pairs), the 5% being *H. pylori* reflected a contamination of 1 human cell in more than one hundred

H. pylori cells, demonstrating a significant enrichment of *H. pylori*. The human DNA observed in our dissected material might partly derive from tissue cell disruption during long-term storage of the biopsy at the ambient temperature and mechanical trauma during sectioning.

New targets enrichment method such as rolling cycle amplification-based multiple displacement amplification has been shown to achieve high sensitivity and specificity, both are higher than 95% [120]. Further studies are needed to evaluate its performance when being applied on highly diverse microbial genomes, as prior knowledge of the targets is still required to design probes.

7 CONCLUSIONS

- Genetic variants of *A4GNT* are related to *H. pylori* infection, but not to gastric cancer risk.

- In the stomach infected with a single *H. pylori* strain, extensive mutations and recombinations can accumulate in *H. pylori* after 20 years. High recombination rates in genes involving in intracellular trafficking and mutations in the cell defensive and motility systems imply selection of *H. pylori* via horizontal gene transfer for advantageous genes during long term colonization in the stomach.

- Current high-throughput DNA sequencing technology can be applied on trace amount of starting materials by using advanced DNA library construction and quantification method.

- Archived FFPE gastric biopsies can be used to sequence *H. pylori* genome for longitudinal study of *H. pylori* metagenomics and stomach cancer risk.

8 FUTURE PERSPECTIVE

8.1 *H. PYLORI* AND GASTRIC CANCER

The upward trend of gastric corpus cancer observed in the US warrants similar investigations in other parts of the world. After decades of antibiotic action by human, *H. pylori* might start to react. The phenotypes of *H. pylori* currently circulating in the world might be different from those circulating decades ago. Latency period might need to be bear in mind when interpreting the two secular trends in one picture.

A very intriguing question is why some *H. pylori* infections result in gastric cancer, whereas some others result in a condition (duodenal ulcer) that prevents gastric cancer – a phenomenon pertains to both Eastern (high risk area) and Western (low risk area) countries. Whole genome sequencing of the bacteria to compare isolates from duodenal ulcer patients with isolates from pre-cancer patients may provide a clue. A case-control study using deep sequencing to compare targeted host genetic loci involving in *H. pylori* adherence and antigen presentation may also help, as *H. pylori* tend to play critical roles at early stages, where the destiny of the infection might have been shaped.

H. pylori is acquired in childhood and causes both diffuse type and intestinal type of stomach cancers, but at different speed in that diffuse type appears earlier (age ~50) and intestinal type comes in elderly (age 65+). Identification of the determinants of the speed would not only be an interest for cancer etiology, but would also facilitate estimation of stomach cancer burden based on disability-adjusted life year (DALY).

Epigenetic changes in the stomach tissue imprinted by *H. pylori* would be important molecular events in *H. pylori* related carcinogenesis to be revealed.

Availability of non-invasive detection methods of *H. pylori* infection and antibiotic resistances rapid and simple enough for a clinician in the countryside to prescribe the right antibiotics to patients (if eradications are warranted) would be wonderful. Vaccination of *H. pylori* during childhood, if successful after evaluation of efficacy and long-term safety, would be a very effective approach to reduce stomach cancer burden.

8.2 MOLECULAR EPIDEMIOLOGY

Molecular epidemiology would be more attractive if the studies involve small numbers of subjects rather than going towards gigantic studies. Currently, molecular epidemiologic studies appear to be shifting towards larger and larger sample sizes such as thousands or even tens of thousands of cases and controls. This is due to, at least in part, two reasons. First, the resolution for differentiating diseases under study is still low. For example, we have

been able to divide stomach cancers based on anatomical site and histology. However, a defined cancer based on these two criteria still consists of a dynamic bulk of cells having different features, and with different proportions at different time points. Second, we cannot yet measure molecules (the exposures) well enough. Measurements at the single-cell haplotype, single-cell epigenetic haplotype or single-cell transcriptome level would facilitate the unveiling of the fundamental processes of decoding genetic information.

Technology drives biomedicine research. Pathologists use microscope and can differentiate tumors into subtypes, for example, adenocarcinoma from squamous cell carcinoma of the esophagus. With the assistance of immunohistochemistry staining, the resolution can be improved to differentiate, for example, estrogen receptor positive from progesterone receptor positive breast cancers. Oligonucleotides micro-array enables us to look at the whole genome single nucleotides polymorphisms. Next-generation sequencing is a big step towards decoding the fundamental process of molecular biology. Recently, technique for whole-genome single cell haplotyping has been developed [121]. Foreseeably, technique for whole-genome single cell epigenetic haplotyping will be available soon.

Molecular classification of tumors facilitates stratification of cancer patients to receive the 'smart' drug in personalized medicine [122]. Identification of biomarkers in patients with solid tumors has been one of the main barriers in biomarker discovery [123] one of the main reasons is possibly the difficulty in getting access to the tissues. Researchers have now shown that micro-RNA signature circulating in blood can be used for lung cancer diagnosis more than two years — 28 months — before spiral computed tomography (CT) can detect the cancer [124]. Noninvasive prenatal detection of fetal chromosomal aneuploidies is possible by using maternal plasma [125].

With sensitive enough technique, molecular information released into the circulation system would be revealed with surprises, such as the DNA of a biological father, through fetus and placenta, into mother's blood. Biobank, where most of the samples are blood, not tissues, would be an invaluable resource for molecular epidemiologic studies that previously regarded as not possible with blood samples.

9 ACKNOWLEDGEMENTS

I would like to thank all who have helped me in different ways during my time as a PhD student at MEB and SMI. I would especially like to thank:

Weimin Ye, my main supervisor, for your trust in my ability since the first day, for sharing your profound knowledge in epidemiology and biostatistics, for showing your great enthusiasms in science and in discussions of various projects, for your vision in molecular epidemiology and the training programs designed for me.

Olof Nyrén, my co-supervisor, for sharing your profound knowledge in epidemiology, for your always enthusiastic encouragements and supports.

Lars Engstrand, my co-supervisor, for your never-ending supports, for your always positive attitudes and encouragements, for maintaining a fantastic research group, for the wonderful group retreats. The resources and platforms that you provided allowed me to freely test various ideas.

Nancy Pedersen, thank you for accepting me in the beginning and your constant generous helps and encouragements.

Anders Andersson, my supervisor in bioinformatics, for stimulating discussions and insightful comments on the manuscripts, and for sharing your skills in bioinformatics.

The SMI 454 group, Reza Advani, for sharing your lab skills, for your helps in developing the 454 method, for the advices in and out of the lab and I cannot thank you enough; Öjar Melefors, for being always open-minded, helpful discussions and editing the manuscripts; Henrik Nordström, for discussions on every aspect of a PCR/ligation reaction and for the discussions out of the lab; Steve Glavas, for assisting with the lab work. I felt like home here and it was really fun to see Poisson statistics in the wet-lab, after using it for ORs.

Lars Engstrand group, a fantastic group where I benefited a lot from, Wilhelm Paulander, for help with the cloning; Sandra Rodin, Hedvig Jakobsson and Mathilda Lindberg for being the ‘father’ introducing me to SMI and showing me some lab procedures; Lena Eriksson, Kristina Schönmeyr and Marianne Ljungström for excellent helps in the lab, where I easily got everything I need; Sönke Andres and Heather-Marie Schmidt, my fellow students making the time of my PhD study much more enjoyable and memorable; Annika Fahlén, Elin Ludin, Cecilia Svensson, Maria Nygård, Valtteri Wirta, Anna Skoglund, Helene Kling-Bäckhed, Cecilia Jernberg, Britta Björkholm, Karin Wreiber, Katrin Pütsep and Philippe Lehours, for the discussions and nice work

and research atmosphere; Britt-Marie Hoffman, my excellent ski coach, and thanks for all the helps on paperwork in SMI.

Weimin Ye group members and other MEB colleagues, for making MEB a very nice work place.

Zack Yusof, for expedite helps on the Linux system! Alex Ploner, for introduction to R! Chuen Seng Tan, my ultimate statistical consultant and my (Hokkien) brother!

Helena Nordenstedt and Jesper Lagergren, my co-authors, for your generous helps in the first studies after I came to Sweden and your supports in my PhD registration.

Lin Cai, my master thesis mentor, for introducing me to the world of molecular epidemiology and to Sweden; Xu Lin, for helps on my first experiments on DNA extraction.

My father and brother, for supports always!

Wenjing and TongTong, you are everything!

10 REFERENCES

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, et al. (2010) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer.
2. Everhart JE (2000) Recent developments in the epidemiology of Helicobacter pylori. Gastroenterol Clin North Am 29: 559-578.
3. Malfertheiner P, Sipponen P, Naumann M, Moayyedi P, Megraud F, et al. (2005) Helicobacter pylori eradication has the potential to prevent gastric cancer: a state-of-the-art critique. Am J Gastroenterol 100: 2100-2115.
4. Peek RM, Jr., Blaser MJ (2002) Helicobacter pylori and gastrointestinal tract adenocarcinomas. Nat Rev Cancer 2: 28-37.
5. Uemura N, Okamoto S, Yamamoto S, Matsumura N, Yamaguchi S, et al. (2001) Helicobacter pylori infection and the development of gastric cancer. N Engl J Med 345: 784-789.
6. Hansson LE, Nyren O, Hsing AW, Bergstrom R, Josefsson S, et al. (1996) The risk of stomach cancer in patients with gastric or duodenal ulcer disease. N Engl J Med 335: 242-249.
7. Blaser MJ, Atherton JC (2004) Helicobacter pylori persistence: biology and disease. J Clin Invest 113: 321-333.
8. Inoue M (2005) [Epidemiology and current trends in stomach cancer]. Nippon Naika Gakkai Zasshi 94: 3-10.
9. Yang L, Parkin DM, Li L, Chen Y (2003) Time trends in cancer mortality in China: 1987-1999. Int J Cancer 106: 771-783.
10. Plummer M, Franceschi S, Munoz N (2004) Epidemiology of gastric cancer. IARC Sci Publ: 311-326.
11. International Agency for Research on Cancer. (2011) A Review of Human Carcinogens: Biological Agents IARC Monographs on the Evaluation of Carcinogenic Risks to Humans 100: 1-487.
12. Misumi A, Murakami A, Harada K, Baba K, Akagi M (1989) Definition of carcinoma of the gastric cardia. Langenbecks Arch Chir 374: 221-226.
13. Ekstrom AM, Signorello LB, Hansson LE, Bergstrom R, Lindgren A, et al. (1999) Evaluating gastric cancer misclassification: a potential explanation for the rise in cardia cancer incidence. J Natl Cancer Inst 91: 786-790.
14. Percy C, Holten Vv, Muir CS, World Health Organization. (1990) International classification of diseases for oncology. Geneva: World Health Organization. 144 p. + 141 computer disk p.
15. Corley DA, Kubo A (2004) Influence of site classification on cancer incidence rates: an analysis of gastric cardia carcinomas. J Natl Cancer Inst 96: 1383-1387.
16. Blot WJ, Devesa SS, Kneller RW, Fraumeni JF, Jr. (1991) Rising incidence of adenocarcinoma of the esophagus and gastric cardia. Jama 265: 1287-1289.
17. Ekstrom AM, Hansson LE, Signorello LB, Lindgren A, Bergstrom R, et al. (2000) Decreasing incidence of both major histologic subtypes of gastric adenocarcinoma--a population-based study in Sweden. Br J Cancer 83: 391-396.
18. Lee JY, Kim HY, Kim KH, Jang HJ, Kim JB, et al. (2003) No changing trends in incidence of gastric cardia cancer in Korea. J Korean Med Sci 18: 53-57.
19. Inoue M, Tsugane S (2005) Epidemiology of gastric cancer in Japan. Postgrad Med J 81: 419-424.
20. Anderson WF, Camargo MC, Fraumeni JF, Jr., Correa P, Rosenberg PS, et al. (2010) Age-specific trends in incidence of noncardia gastric cancer in US adults. Jama 303: 1723-1728.
21. Camargo MC, Anderson WF, King JB, Correa P, Thomas CC, et al. (2011) Divergent trends for gastric cancer incidence by anatomical subsite in US adults. Gut.
22. Lauren P (1965) 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 64: 31-49.

23. Correa P, Haenszel W, Cuello C, Tannenbaum S, Archer M (1975) A model for gastric cancer epidemiology. Lancet 2: 58-60.
24. Coleman MP, Esteve J, Damiecki P, Arslan A, Renard H (1993) Trends in cancer incidence and mortality. IARC Sci Publ: 1-806.
25. Suerbaum S, Michetti P (2002) Helicobacter pylori infection. N Engl J Med 347: 1175-1186.
26. Parsonnet J (1998) Helicobacter pylori: the size of the problem. Gut 43 Suppl 1: S6-9.
27. Roosendaal R, Kuipers EJ, Buitenwerf J, van Uffelen C, Meuwissen SG, et al. (1997) Helicobacter pylori and the birth cohort effect: evidence of a continuous decrease of infection rates in childhood. Am J Gastroenterol 92: 1480-1482.
28. Banatvala N, Mayo K, Megraud F, Jennings R, Deeks JJ, et al. (1993) The cohort effect and Helicobacter pylori. J Infect Dis 168: 219-221.
29. Fock KM, Ang TL (2010) Epidemiology of Helicobacter pylori infection and gastric cancer in Asia. J Gastroenterol Hepatol 25: 479-486.
30. Everhart JE, Kruszon-Moran D, Perez-Perez GI, Tralka TS, McQuillan G (2000) Seroprevalence and ethnic differences in Helicobacter pylori infection among adults in the United States. J Infect Dis 181: 1359-1363.
31. Kivi M, Tindberg Y, Sorberg M, Casswall TH, Befrits R, et al. (2003) Concordance of Helicobacter pylori strains within families. J Clin Microbiol 41: 5604-5608.
32. Parsonnet J, Shmuely H, Haggerty T (1999) Fecal and oral shedding of Helicobacter pylori from healthy infected adults. Jama 282: 2240-2245.
33. Liu C, Russell RM (2008) Nutrition and gastric cancer risk: an update. Nutr Rev 66: 237-249.
34. Gonzalez CA, Pera G, Agudo A, Palli D, Krogh V, et al. (2003) Smoking and the risk of gastric cancer in the European Prospective Investigation Into Cancer and Nutrition (EPIC). Int J Cancer 107: 629-634.
35. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, et al. (2000) Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 343: 78-85.
36. Guilford P, Hopkins J, Harraway J, McLeod M, McLeod N, et al. (1998) E-cadherin germline mutations in familial gastric cancer. Nature 392: 402-405.
37. Gianfagna F, De Feo E, van Duijn CM, Ricciardi G, Boccia S (2008) A systematic review of meta-analyses on gene polymorphisms and gastric cancer risk. Curr Genomics 9: 361-374.
38. Hidaka E, Ota H, Hidaka H, Hayama M, Matsuzawa K, et al. (2001) Helicobacter pylori and two ultrastructurally distinct layers of gastric mucous cell mucins in the surface mucous gel layer. Gut 49: 474-480.
39. Matsuzwa M, Ota H, Hayama M, Zhang MX, Sano K, et al. (2003) Helicobacter pylori infection up-regulates gland mucous cell-type mucins in gastric pyloric mucosa. Helicobacter 8: 594-600.
40. Nakayama J, Yeh JC, Misra AK, Ito S, Katsuyama T, et al. (1999) Expression cloning of a human alpha1, 4-N-acetylglucosaminyltransferase that forms GlcNAcalpha1-->4Galbeta-->R, a glycan specifically expressed in the gastric gland mucous cell-type mucin. Proc Natl Acad Sci U S A 96: 8991-8996.
41. Zhang MX, Nakayama J, Hidaka E, Kubota S, Yan J, et al. (2001) Immunohistochemical demonstration of alpha1,4-N-acetylglucosaminyltransferase that forms GlcNAcalpha1,4Galbeta residues in human gastrointestinal mucosa. J Histochem Cytochem 49: 587-596.
42. Lee H, Kobayashi M, Wang P, Nakayama J, Seeberger PH, et al. (2006) Expression cloning of cholesterol alpha-glucosyltransferase, a unique enzyme that can be inhibited by natural antibiotic gastric mucin O-glycans, from Helicobacter pylori. Biochem Biophys Res Commun 349: 1235-1241.
43. Kawakubo M, Ito Y, Okimura Y, Kobayashi M, Sakura K, et al. (2004) Natural antibiotic function of a human gastric mucin against Helicobacter pylori infection. Science 305: 1003-1006.
44. Boren T, Falk P, Roth KA, Larson G, Normark S (1993) Attachment of Helicobacter pylori to human gastric epithelium mediated by blood group antigens. Science 262: 1892-1895.

45. Van de Bovenkamp JH, Mahdavi J, Korteland-Van Male AM, Buller HA, Einerhand AW, et al. (2003) The MUC5AC glycoprotein is the primary receptor for Helicobacter pylori in the human stomach. Helicobacter 8: 521-532.
46. Carvalho F, Seruca R, David L, Amorim A, Seixas M, et al. (1997) MUC1 gene polymorphism and gastric cancer--an epidemiological study. Glycoconj J 14: 107-111.
47. Silva F, Carvalho F, Peixoto A, Seixas M, Almeida R, et al. (2001) MUC1 gene polymorphism in the gastric carcinogenesis pathway. Eur J Hum Genet 9: 548-552.
48. Nguyen TV, Janssen M, Jr., Gritters P, te Morsche RH, Drenth JP, et al. (2006) Short mucin 6 alleles are associated with H pylori infection. World J Gastroenterol 12: 6021-6025.
49. Garcia E, Carvalho F, Amorim A, David L (1997) MUC6 gene polymorphism in healthy individuals and in gastric cancer patients from northern Portugal. Cancer Epidemiol Biomarkers Prev 6: 1071-1074.
50. Jia Y, Persson C, Hou L, Zheng Z, Yeager M, et al. (2010) A comprehensive analysis of common genetic variation in MUC1, MUC5AC, MUC6 genes and risk of stomach cancer. Cancer Causes Control 21: 313-321.
51. Sakamoto H, Yoshimura K, Saeki N, Katai H, Shimoda T, et al. (2008) Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. Nat Genet 40: 730-740.
52. Lu Y, Chen J, Ding Y, Jin G, Wu J, et al. (2010) Genetic variation of PSCA gene is associated with the risk of both diffuse- and intestinal-type gastric cancer in a Chinese population. Int J Cancer 127: 2183-2189.
53. Saeki N, Saito A, Choi IJ, Matsuo K, Ohnami S, et al. (2011) A functional single nucleotide polymorphism in mucin 1, at chromosome 1q22, determines susceptibility to diffuse-type gastric cancer. Gastroenterology 140: 892-902.
54. International Agency for Research on Cancer. (1994) Schistosomes, liver flukes and Helicobacter pylori. Lyon: International Agency for Research on Cancer. 270 p. p.
55. Malaty HM, Graham DY, Isaksson I, Engstrand L, Pedersen NL (1998) Co-twin study of the effect of environment and dietary elements on acquisition of Helicobacter pylori infection. American Journal of Epidemiology 148: 793-797.
56. Mitchell H, English DR, Elliott F, Gengos M, Barrett JH, et al. (2008) Immunoblotting using multiple antigens is essential to demonstrate the true risk of Helicobacter pylori infection for gastric cancer. Aliment Pharmacol Ther 28: 903-910.
57. Ekstrom AM, Held M, Hansson LE, Engstrand L, Nyren O (2001) Helicobacter pylori in gastric cancer established by CagA immunoblot as a marker of past infection. Gastroenterology 121: 784-791.
58. Kamangar F, Qiao YL, Blaser MJ, Sun XD, Katki H, et al. (2007) Helicobacter pylori and oesophageal and gastric cancers in a prospective study in China. Br J Cancer 96: 172-176.
59. Sasazuki S, Inoue M, Iwasaki M, Otani T, Yamamoto S, et al. (2006) Effect of Helicobacter pylori infection combined with CagA and pepsinogen status on gastric cancer development among Japanese men and women: a nested case-control study. Cancer Epidemiol Biomarkers Prev 15: 1341-1347.
60. Persson C, Jia Y, Pettersson H, Dillner J, Nyren O, et al. (2011) H. pylori seropositivity before age 40 and subsequent risk of stomach cancer: a glimpse of the true relationship? PLoS One 6: e17404.
61. Siman JH, Forsgren A, Berglund G, Floren CH (1997) Association between Helicobacter pylori and gastric carcinoma in the city of Malmo, Sweden. A prospective study. Scand J Gastroenterol 32: 1215-1221.
62. Huang JQ, Zheng GF, Sumanac K, Irvine EJ, Hunt RH (2003) Meta-analysis of the relationship between cagA seropositivity and gastric cancer. Gastroenterology 125: 1636-1644.
63. You WC, Brown LM, Zhang L, Li JY, Jin ML, et al. (2006) Randomized double-blind factorial trial of three treatments to reduce the prevalence of precancerous gastric lesions. J Natl Cancer Inst 98: 974-983.

64. Wong BC, Lam SK, Wong WM, Chen JS, Zheng TT, et al. (2004) Helicobacter pylori eradication to prevent gastric cancer in a high-risk region of China: a randomized controlled trial. Jama 291: 187-194.
65. Leung WK, Lin SR, Ching JY, To KF, Ng EK, et al. (2004) Factors predicting progression of gastric intestinal metaplasia: results of a randomised trial on Helicobacter pylori eradication. Gut 53: 1244-1249.
66. Mera R, Fontham ET, Bravo LE, Bravo JC, Piazuelo MB, et al. (2005) Long term follow up of patients treated for Helicobacter pylori infection. Gut 54: 1536-1540.
67. Graham DY, Fischbach L (2010) Helicobacter pylori treatment in the era of increasing antibiotic resistance. Gut 59: 1143-1153.
68. Rimbara E, Fischbach LA, Graham DY (2011) Optimal therapy for Helicobacter pylori infections. Nat Rev Gastroenterol Hepatol 8: 79-88.
69. Salama NR, Gonzalez-Valencia G, Deatherage B, Aviles-Jimenez F, Atherton JC, et al. (2007) Genetic analysis of Helicobacter pylori strain populations colonizing the stomach at different times postinfection. J Bacteriol 189: 3834-3845.
70. Salama N, Guillemin K, McDaniel TK, Sherlock G, Tompkins L, et al. (2000) A whole-genome microarray reveals genetic diversity among Helicobacter pylori strains. Proc Natl Acad Sci U S A 97: 14668-14673.
71. Kuipers EJ, Israel DA, Kusters JG, Gerrits MM, Weel J, et al. (2000) Quasispecies development of Helicobacter pylori observed in paired isolates obtained years apart from the same host. J Infect Dis 181: 273-282.
72. Enroth H, Nyren O, Engstrand L (1999) One stomach--one strain: does Helicobacter pylori strain variation influence disease outcome? Dig Dis Sci 44: 102-107.
73. Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, et al. (1997) The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature 388: 539-547.
74. Alm RA, Ling LS, Moir DT, King BL, Brown ED, et al. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. Nature 397: 176-180.
75. Oh JD, Kling-Backhed H, Giannakis M, Xu J, Fulton RS, et al. (2006) The complete genome sequence of a chronic atrophic gastritis Helicobacter pylori strain: evolution during disease progression. Proc Natl Acad Sci U S A 103: 9999-10004.
76. Kawai M, Furuta Y, Yahara K, Tsuru T, Oshima K, et al. (2011) Evolution in an oncogenic bacterial species with extreme genome plasticity: Helicobacter pylori East Asian genomes. BMC Microbiol 11: 104.
77. Kersulyte D, Lee W, Subramaniam D, Anant S, Herrera P, et al. (2009) Helicobacter Pylori's plasticity zones are novel transposable elements. PLoS One 4: e6859.
78. Eppinger M, Baar C, Linz B, Raddatz G, Lanz C, et al. (2006) Who ate whom? Adaptive Helicobacter genomic changes that accompanied a host jump from early humans to large felines. PLoS Genet 2: e120.
79. Bjorkholm B, Sjolund M, Falk PG, Berg OG, Engstrand L, et al. (2001) Mutation frequency and biological cost of antibiotic resistance in Helicobacter pylori. Proc Natl Acad Sci U S A 98: 14607-14612.
80. Falush D, Kraft C, Taylor NS, Correa P, Fox JG, et al. (2001) Recombination and mutation during long-term gastric colonization by Helicobacter pylori: estimates of clock rates, recombination size, and minimal age. Proc Natl Acad Sci U S A 98: 15056-15061.
81. Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, et al. (2011) Helicobacter pylori genome evolution during human infection. Proc Natl Acad Sci U S A 108: 5033-5038.
82. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

83. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456: 53-59.
84. Glenn TC (2011) Field guide to next-generation DNA sequencers. Mol Ecol Resour.
85. Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol 8: R143.
86. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12: R18.
87. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, et al. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 7: 461-465.
88. Chow WH, Swanson CA, Lissowska J, Groves FD, Sobin LH, et al. (1999) Risk of stomach cancer in relation to consumption of cigarettes, alcohol, tea and coffee in Warsaw, Poland. Int J Cancer 81: 871-876.
89. Chow WH, Blaser MJ, Blot WJ, Gammon MD, Vaughan TL, et al. (1998) An inverse relation between cagA(+) strains of Helicobacter pylori infection and risk of esophageal and gastric cardia adenocarcinoma. Cancer Research 58: 588-590.
90. Akopyanz N, Bukanov NO, Westblom TU, Kresovich S, Berg DE (1992) DNA diversity among clinical isolates of Helicobacter pylori detected by PCR-based RAPD fingerprinting. Nucleic Acids Res 20: 5137-5142.
91. Enroth H, Kraaz W, Engstrand L, Nyren O, Rohan T (2000) Helicobacter pylori strain types and risk of gastric cancer: a case-control study. Cancer Epidemiol Biomarkers Prev 9: 981-985.
92. Mane SP, Dominguez-Bello MG, Blaser MJ, Sobral BW, Hontecillas R, et al. (2010) Host-interactive genes in Amerindian Helicobacter pylori diverge from their Old World homologs and mediate inflammatory responses. J Bacteriol 192: 3078-3092.
93. Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. J Comput Biol 7: 203-214.
94. Roder DM (2002) The epidemiology of gastric cancer. Gastric Cancer 5 Suppl 1: 5-11.
95. Cheng R, Ma JZ, Elston RC, Li MD (2005) Fine mapping functional sites or regions from case-control data using haplotypes of multiple linked SNPs. Ann Hum Genet 69: 102-112.
96. French B, Lumley T, Monks SA, Rice KM, Hindorff LA, et al. (2006) Simple estimates of haplotype relative risks in case-control data. Genet Epidemiol 30: 485-494.
97. Lee H, Wang P, Hoshino H, Ito Y, Kobayashi M, et al. (2008) Alpha1,4GlcNAc-capped mucin-type O-glycan inhibits cholesterol alpha-glucosyltransferase from Helicobacter pylori and suppresses H. pylori growth. Glycobiology 18: 549-558.
98. Shimizu F, Nakayama J, Ishizone S, Zhang MX, Kawakubo M, et al. (2003) Usefulness of the real-time reverse transcription-polymerase chain reaction assay targeted to alpha1,4-N-acetylglucosaminyltransferase for the detection of gastric cancer. Lab Invest 83: 187-197.
99. White RA, 3rd, Blainey PC, Fan HC, Quake SR (2009) Digital PCR provides sensitive and absolute calibration for high throughput sequencing. BMC Genomics 10: 116.
100. Blow MJ, Zhang T, Woyke T, Speller CF, Krivoshapkin A, et al. (2008) Identification of ancient remains through genomic sequencing. Genome Res 18: 1347-1353.
101. Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, et al. (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics 7: 216.

102. Maricic T, Paabo S (2009) Optimization of 454 sequencing library preparation from small amounts of DNA permits sequence determination of both DNA strands. Biotechniques 46: 51-52, 54-57.
103. Meyer M, Briggs AW, Maricic T, Hober B, Hoffner B, et al. (2008) From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. Nucleic Acids Res 36: e5.
104. Kutyavin IV, Afonina IA, Mills A, Gorn VV, Lukhtanov EA, et al. (2000) 3'-minor groove binder-DNA probes increase sequence specificity at PCR extension temperatures. Nucleic Acids Res 28: 655-661.
105. Sandberg J, Stahl PL, Ahmadian A, Bjursell MK, Lundeberg J (2009) Flow cytometry for enrichment and titration in massively parallel DNA sequencing. Nucleic Acids Res 37: e63.
106. Zheng Z, Advani A, Melefors O, Glavas S, Nordstrom H, et al. (2010) Titration-free massively parallel pyrosequencing using trace amounts of starting material. Nucleic Acids Res 38: e137.
107. Quach N, Goodman MF, Shibata D (2004) In vitro mutation artifacts after formalin fixation and error prone translesion synthesis during PCR. BMC Clin Pathol 4: 1.
108. Andersen LP, Rasmussen L (2009) Helicobacter pylori-coccoid forms and biofilm formation. FEMS Immunol Med Microbiol 56: 112-115.
109. Enroth H, Wreiber K, Rigo R, Risberg D, Uribe A, et al. (1999) In vitro aging of Helicobacter pylori: Changes in morphology, intracellular composition and surface properties. Helicobacter 4: 7-16.
110. Teer JK, Bonnycastle LL, Chines PS, Hansen NF, Aoyama N, et al. (2010) Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. Genome Res 20: 1420-1431.
111. Wang Y, Antonopoulos DA, Zhu X, Harrell L, Hanan I, et al. (2010) Laser capture microdissection and metagenomic analysis of intact mucosa-associated microbial communities of human colon. Appl Microbiol Biotechnol.
112. Graham DY, Lu H, Yamaoka Y (2009) African, Asian or Indian enigma, the East Asian Helicobacter pylori: facts or medical myths. J Dig Dis 10: 77-84.
113. Holcombe C (1992) Helicobacter pylori: the African enigma. Gut 33: 429-431.
114. Rothman KJ, Greenland S, Lash TL (2008) Modern epidemiology. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins. x, 758 p. p.
115. Pearl J (1988) Probabilistic reasoning in intelligent systems : networks of plausible inference. San Mateo, Calif.: Morgan Kaufmann Publishers. xix, 552 p. p.
116. Greenland S, Pearl J, Robins JM (1999) Causal diagrams for epidemiologic research. Epidemiology 10: 37-48.
117. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA (2002) Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 155: 176-184.
118. Shrier I, Platt RW (2008) Reducing bias through directed acyclic graphs. BMC Med Res Methodol 8: 70.
119. Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological 57: 289-300.
120. Johansson H, Isaksson M, Sorqvist EF, Roos F, Stenberg J, et al. (2011) Targeted resequencing of candidate genes using selector probes. Nucleic Acids Res 39: e8.
121. Fan HC, Wang J, Potanina A, Quake SR (2011) Whole-genome molecular haplotyping of single cells. Nat Biotechnol 29: 51-57.
122. Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, et al. (2010) Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. N Engl J Med 363: 1693-1703.
123. Sawyers CL (2008) The cancer biomarker problem. Nature 452: 548-552.
124. Boeri M, Verri C, Conte D, Roz L, Modena P, et al. (2011) MicroRNA signatures in tissues and plasma predict development and prognosis of computed tomography detected lung cancer. Proc Natl Acad Sci U S A 108: 3713-3718.

125. Lo YM (2009) Noninvasive prenatal detection of fetal chromosomal aneuploidies by maternal plasma nucleic acid analysis: a review of the current state of the art. BJOG 116: 152-157.