Identification of Genetic Risk Variants for the Development of Gastric Adenocarcinoma

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch-Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Timo Heß

aus

Troisdorf

Bonn 2022

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Prof. Dr. Markus M. Nöthen

2. Gutachter: Prof. Dr. Walter Witke

Tag der Promotion: 31.01.2023

Erscheinungsjahr: 2023

Table of Content

Abbreviations

A	Adenine
app.	Approximately
bp	Base pair
BMI	Body Mass Index
BO	Barrett's oesophagus
°C	Degree Celsius
C	Cytosine
cDNA	complementary deoxyribonucleic acid
CEU	Central European
Chr	Chromosome
CI	Confidence interval
CR	Call rate
DE	Differential expression
DNA	Deoxyribonucleic acid
EAS	East Asian
EDTA	Ethylenediamine tetraacetate
ENCODE	Encyclopedia of DNA elements
eQTL	Expression quantitative trait locus
EtOH	Ethanol
FC	Fold change
G	Guanine
GC	Gastric adenocarcinoma
gDNA	Genomic deoxyribonucleic acid
GERD	Gastroesophageal reflux disease
GI	Gastrointestinal
GOJ	Gastroesophageal junction
GSA	Global Screening Array
GTEx	Genotype-Tissue Expression
GWAS	Genome-wide association study
HCl	Hydrochloric acid
HDGC	Hereditary diffuse gastric cancer
hg	Human genome
HP	*Helicobacter pylori*

HWE	Hardy-Weinberg equilibrium
LD	Linkage disequilibrium
kb	Kilo base
MAF	Minor allele frequency
Mb	Mega base
mg	Milligram
mRNA	Messenger ribonucleic acid
µl	Microlitre
ng	Nanogram
NGS	Next Generation Sequencing
nl	Nanolitre
nm	Nanometer
OAC	Oesophageal adenocarcinoma
OR	Odds ratio
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
PEER	Probabilistic estimation of expression residuals
QC	Quality control
QQ	Quantile-quantile
REML	Restricted maximum-likelihood
RNA	Ribonucleic acid
rRNA	Ribosomal ribonucleic acid
RR	Relative risk
SNP	Single nucleotide polymorphism
SD	Standard deviation
SNV	Single nucleotide variant
T	Thymin
TPM	Transcripts per million
TSS	Transcription start site
TWAS	Transcriptome-wide association analysis
UTR	Untranslated region
UKBB	UK Biobank

List of Figures

List of Tables

1 Introduction

An effective utilization of nutrients is of central importance for the metabolism and survival of organisms. For this purpose, all mammalian species developed a gastrointestinal (GI) tract, which disintegrates and absorbs all essential components of ingested food. At the same time the GI builds a barrier against the intrusion of potentially harmful agents. In humans, the GI tract comprises all organs along the passageway of food from the mouth to the anus, including the oesophagus, the stomach and the intestines [1].

The human stomach plays a central role in the digestive process, having developed during evolution from a simple muscular tube into a complex exocrine organ. It is mainly involved in the pre-digestion and proportioning of food, as well as the protection against pathogenic microorganisms [2]. For this purpose, it harbours a comparable extreme environment, producing highly acidic gastric juices enriched with proteolytic enzymes. To avoid infections on the one hand and self-digestion processes on the other, a complex physiological system needs to be maintained [2]. Imbalances in this system can easily lead to adverse effects, such as chronic inflammatory reactions in the gastric mucosa. Over time, such conditions can result in a number of pathogenic conditions, including the development of gastric cancer (GC) [3]. In the past, the identification of environmental risk factors promoting these adverse conditions, such as specific dietary habits or the infection with certain microorganisms, as well as intended or unintended interventions have led to significant reduction in the incidence of GC worldwide [3]. However, up till today, GC is the fourth most common tumour entity worldwide with a poor prognosis. Many processes having an influence on GC development are not fully understood.

Beside exogenous noxes, also genetic risk factors are of relevance for GC development. However, only a small fraction of GC develops on the background of monogenetic tumour syndromes, in which mutations with a high penetrance contribute to the development of the phenotype. The majority of GC cases occur sporadically and are of multifactorial genesis. As such, an unfavourable combination of environmental and genetic risk factors, each with a comparable small effect size, trigger the development of cancer in an individual [4–13].

The identification of such genetic risk factors is of importance to understand the underlying pathomechanisms, for unravelling correlations to other diseases, and to predict individual risks for the development of a specific phenotype. However, their detection and functional interpretation is laborious, as large patient and control

samples are required. Comprehensive genotyping and comparisons between the two groups may reveal genetic variants associated to the phenotype under examination, utilizing a method called genome-wide association study (GWAS) [14].

For the functional interpretations of the identified variants further follow-up analyses and the systematic intersection with other omic-studies are required. As such, transcriptome data from tissues of relevance have been proven to be a powerful resource by unravelling gene regulatory mechanisms [15].

In case of GC, multiple GWAS could identify a number of genetic risk loci, but have been mainly performed in samples of Asian ethnicity. Studies in the European population replicating the Asian findings are scarce [4–13]. In addition, functional interpretation of associated variants and the interpretation of the genetic risk background beyond the single marker level is lacking.

In this thesis, we present the results of the largest European case-control GC GWAS sample and the follow-up of the results using the largest and most comprehensive transcriptome dataset on gastric mucosa.

In the following chapters the physiology of the human stomach, the current knowledge on GC related risk factors and classification systems, as well as the theoretical background for the identification and interpretation of genetic risk variants will be introduced.

1.1 The Human Stomach and Gastric Cancer

1.1.1 The Human Stomach

The human stomach is an important part of the human GI tract and is responsible for various aspects of the digestion process, including a major part in the disintegration and portioning of all kinds of food. Three layers of smooth muscle tissue encapsulate a highly specialized mucosa composed of a variety of specialized cell types and glands, which take over different parts in food pre-processing [1,2].

Anatomically and histologically, the stomach can be subdivided into different regions (Figure 1). The oesophagus connects to the stomach at a region called the cardia, whereby the transition area is called gastro-oesophageal junction (GOJ). Superior to the cardia, the fundus is located. The cardia and fundus connect to the main part of stomach, the corpus. The corpus goes over into the pyloric antrum, which funnels down to the pyloric sphincter, forming the connection with the duodenum [2].

The surrounding muscle layers enable the stomach to expand and shrink, maintaining a volume of about 4 litres at its maximal capacity. Furthermore, they enable the mixing and the mechanical breakdown of the contained food to chyme [2].

The mucosa is highly populated with different types of gastric glands, which produce gastric juices of different cellular compositions according to the stomach region. The fundus and corpus are the main site of chemical digestion, which is driven by a variety of secretory cells, including the parietal-, chief-, mucous neck- and foveolar cells [1]. Parietal cells primarily produce hydrochloric acid (HCl) and intrinsic factor. The secretion of HCl causes the extremely acidic milieu in the stomach ranging from a pH from 1.5 to 2.0. The low pH denatures proteins, activates enzymes, such as pepsin, and destroys most of the microorganisms ingested together with food. Intrinsic factor is essential for the absorption of vitamin B_{12} later on in the small intestine [1]. Chief cells produce the zymogen pepsinogen, which gets converted at low pH to pepsin, playing a major part in protein digestion. The foveolar and mucous neck cells produce alkaline mucous that protect the gastric epithelium from acidic erosion. The glands of the antrum produce mucous and contain G cells producing the hormone gastrin. Gastrin secretion is induced upon the presence of peptides and promotes HCl secretion as well as gastric emptying [16].

Apart from these functions and cell types, a variety of pathways may have an influence the stomach glandular secretions, thereby tightly regulating the production of HCl, ensuring stomach emptying and digestion enzyme production. Factors adversely influencing this balanced system, can easily harm the integrity of the mucosa,

exposing the epithelium to the extreme gastric environment thereby leading to cellular damage and inflammation, commonly described as gastritis. If the cellular stress prevails, the tissue may gradually change and accumulate genomic alterations, which may ultimately culminate in a malignant tumour [3].

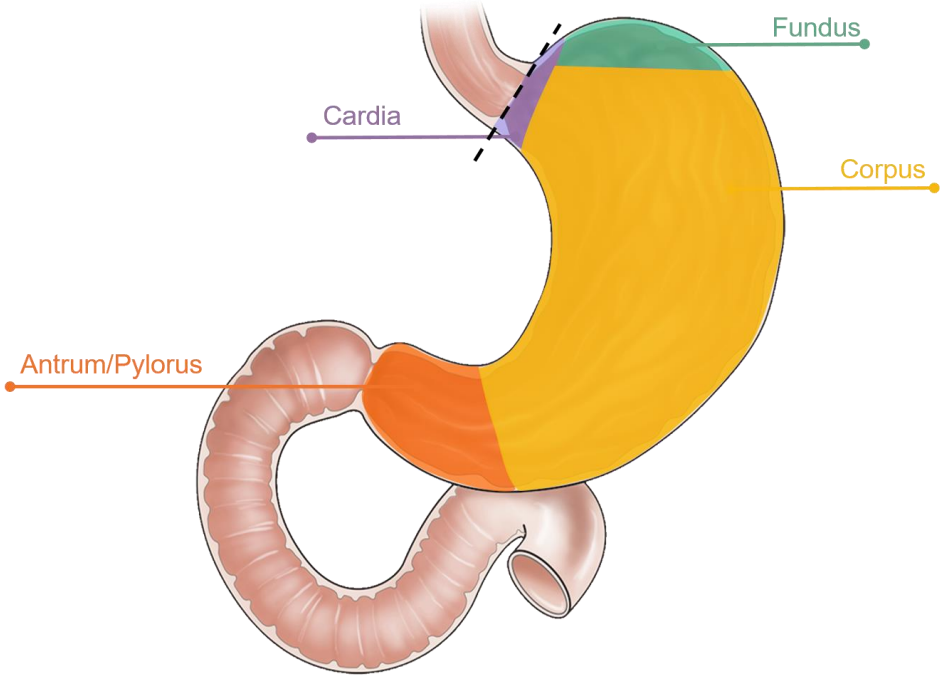


Figure 1: Schematic representation of the anatomical regions of the human stomach (modified from Smyth et al. 2020 [3]).

1.1.2 Epidemiology, Prognosis and Treatment of Gastric Cancer

The vast majority of gastric tumours are adenocarcinomas (80-90%), while the remaining entities are mainly attributable to neuroendocrine tumours, lymphomas and gastrointestinal stromal tumours. In the following the focus remains on gastric adenocarcinomas referred to as GC [3].

With more than one million cases and over 750.000 deaths annually, GC is the fourth most common and third deadliest malignant neoplasm worldwide [17]. Representing the most common cancer entity at the beginning of the 20th century, there has been a dramatic and steady decline in incidence over the past decades, especially in industrialised nations. However, the relative proportion of GC in malignant tumour cases remains at a high level in large parts of East Asia, Eastern Europe and Latin America [18].

Due to a late onset of symptoms, GC is often diagnosed at an advanced stage, which is one of the main reasons for a rather poor prognosis with a 5-year survival rate below 25% worldwide [19]. A noteworthy exception is South Korea and Japan with survival rates beyond 60%, mainly due to sophisticated screening programs enabling detection and treatment at an early stage of the disease [20].

The only potentially curative treatment for localized GC is surgical resection. For non-metastatic advanced GC total or subtotal gastrectomy is indicated, often in combination with neo-adjuvant and/or postoperative chemoradiotherapy. However, there is no general standard for adjuvant therapies [21].

1.1.3 Classification of Gastric Adenocarcinoma

Adenocarcinoma of the stomach summarize a rather diverse groups of tumours with significant differences in their aetiology. For a better characterisation a number of different GC classification systems have been implemented, which group GC tumours either according to their anatomical location, their histological or molecular genetic characteristics [22–26].

The most widely used system to differentiate GC tumour types histologically is the classification according to Lauren, discriminating between a diffuse, intestinal and a mixed tumour type [27]. This classification is most often used in combination with an anatomical differentiation between carcinomas affecting the proximal (cardia) and distal (non-cardia) part of the stomach (Figure 2).

The intestinal tumour type is characterized by well differentiated cells predominantly forming a glandular like epithelium, with cells resembling intestinal columnar cells [28]. Tumours of the intestinal type are usually the endpoint of an inflammatory cascade, starting with a chronic active gastritis, which can be of multifactorial genesis. During the progress of inflammation, a gradual loss of glandular mucosa is observed (multifocal atrophy), further progressing to the replacement of the gastric mucosa with an epithelium resembling that of the intestine (intestinal metaplasia), which may transform to a malignant neoplasm [28].

By contrast, the diffuse type is characterized by poorly differentiated and poorly cohesive cells with diffuse infiltrative margins. Tumours showing a mixture of both types are accounted to the intermediate or mixed type. Carcinomas of the diffuse type appear in the absence of any premalignant condition. They seem to occur at an earlier age and are reported to appear more frequently in the cardia, as compared to tumours of the intestinal type, which are more frequent in the corpus and antrum [28].

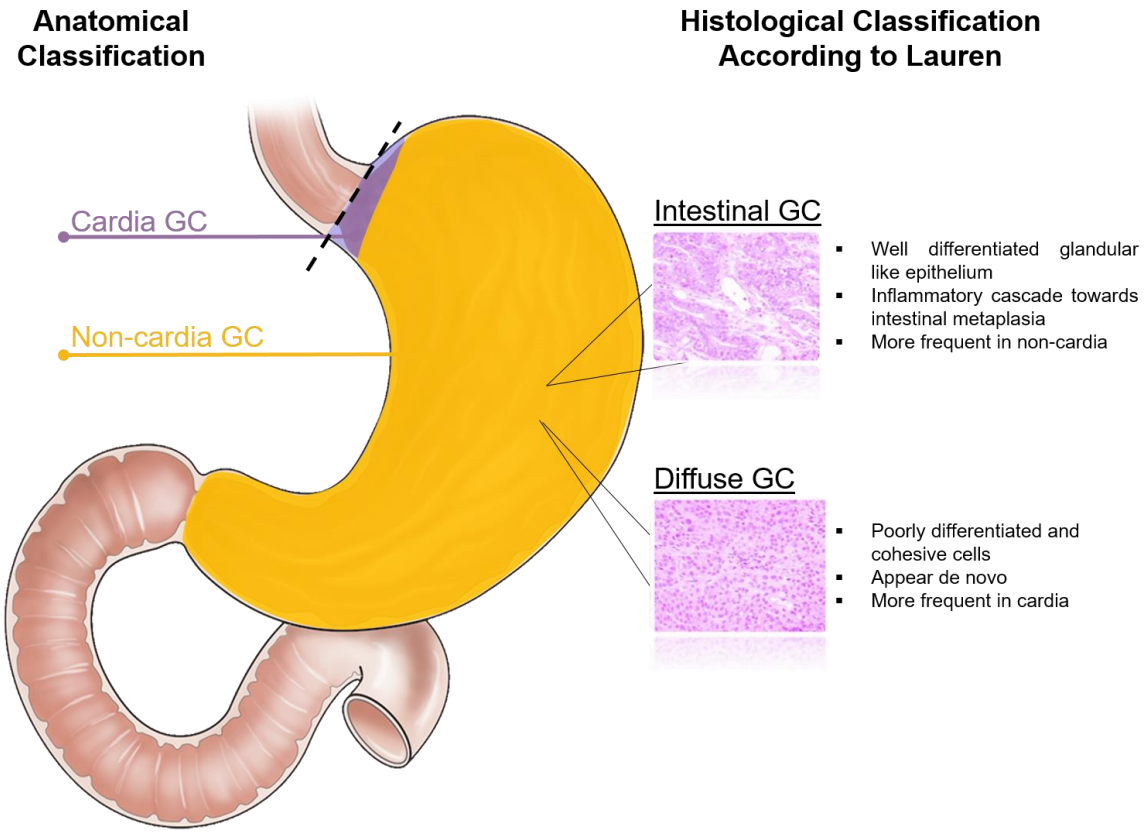


Figure 2: Classification of gastric adenocarcinoma according to their anatomical location into cardia or non-cardia and according to their histological appearance according to Lauren into diffuse or intestinal type (modified from Smyth et al. 2020 [3]).

1.1.4 Environmental Risk Factors

The dramatic decline of GC cases over the past century and the unequal geospatial distribution indicate that GC development is closely linked to modifiable risk factors. Indeed, this development can be primarily accounted for intentional and unintentional reduction of environmental risk exposures and their unequal distributions across the globe [29]. Widely adopted changes in food preservation, like refrigeration, initiated a decrease in GC incidences in industrialized countries. These were merely driven by a shift of food supply from a primarily salt and smoke based preservation and hygienically often insufficient standard, towards a diet that included more fresh fruits, vegetables and unprocessed meat. Consequently, a high salt diet, the intake of smoked and processed meat contribute to GC cancer risk, primarily by inducing stress to the gastric mucosa, leading to gastritis [29]. On the other hand, a Mediterranean diet, which is rich in vegetables, unrefined cereals, fruits and fish, and low in non-fish meat products has been shown to reduce GC risk [30].

Another main risk factor for GC development is infection with *Helicobacter pylori* (HP), a highly specialized gram-negative bacterium. It is able to overcome the harsh gastric

environment to colonize the gastric mucosa, thereby inducing a type B gastritis. In addition, several strains harbour virulence factors, such as cytotoxin-associated gene A (cagA) and vacuolating cytotoxin A (vacA) that can further contribute to the progression towards GC [31]. A better monitoring and eradication of HP infections led to a further significant reduction of GC, primarily in industrialized countries. Although, being of primary significance for the intestinal tumour type, HP infection also contributes to the risk of diffuse GC. Due to this correlation, HP is so far the only bacterium currently classified as class I carcinogen by the World Health Organization [31]. Apart from HP, other infectious agents such as the Epstein-Barr-Virus could be identified to confer to GC development [31]. Other environmental risk factors include smoking, independent of the tumour type and location, as well as above high level alcohol consumption [32]. On the other hand, a high socioeconomic status is associated with a reduced risk for the development of GC [33].

Gender represents another GC risk factor. In general, males are about twice more likely to be diagnosed with GC as compared to females [34]. However, there are differences in the ratio regarding the subtypes. For the intestinal type, the male to female ratio is reported to be 2.3 and for the diffuse type 1.5 [28]. The reason for this gender disparities are not fully elucidated, but are partly attributed to a higher prevalence of risk factors such as smoking and high alcohol consumption in males [18].

Also for the GC location, disparities in risk factors could be identified. As such, obesity and gastroesophageal reflux disease (GERD) are only associated to an increased cardia GC risk, but not with non-cardia GC [30].

1.1.5 Cardia Gastric Cancer and Oesophageal Adenocarcinoma

Although there is a general decrease in the incidence of GC in Western countries, there are discrepancies when examining the tumour location. Compared to non-cardia GC, cardia GC is rarer, but the overall incidence increased over the past decades, representing one of the fastest growing cancer entities overall [35]. The same development is observed for oesophageal adenocarcinoma of the lower oesophagus (OAC). Interestingly, both cancer entities share a couple of risk factors, such as GERD and obesity, which are not associated to non-cardia GC [35]. This raises the question, whether those entities might be related. Indeed, both cancer types arise close or across the gastroesophageal junction (GOJ) and discrimination of OAC and cardia GC has proven to be problematic [38]. The most prominent system to discriminate carcinoma at the GOJ is the Siewert classification, whereby Siewert type I cancers belong to OAC and are located between 1 and 5 cm above the GOJ [36]. These

carcinomas usually arise on the background of a Barrett's Oesophagus (BO), representing the endpoint of a metaplasia-dysplasia-neoplasia cascade often induced by a frequent reflux of gastric juices damaging the oesophageal squamous epithelium [36]. By contrast, Siewert type II cancers are located between 1 cm above and 2 cm below the GOJ and are considered to gastric cardia tumours. Siewert type III cancers are located between 2 and 5 cm below the GOJ, with invasion of the oesophagus, and are considered as subcardiac gastric cancers [37]. There is an ongoing discussion whether cardia GC should be accounted to the gastric cancer entity, or if it is closer related to OAC [38].

1.1.6 Genetic Risk Factors for Gastric Cancer

The majority of GC cases have a multifactorial aetiology and several environmental risk factors contributing to GC development could be identified. However, multifactorial phenotypes always have a genetic component. Based on a twin study, the contribution of genetic factors to GC has been estimated to be 28% [39]. As such, there is a large interest in the identification of these risk factors, which is done primarily by utilizing genome-wide association studies (GWAS). The theoretical concept of GWAS will be introduced in chapter 1.2.2. In the following, the results of the so far conducted GC GWAS will be presented.

A total of ten GC GWAS has been performed to date, identifying a total of twelve risk loci. As the incidence of GC is high in East Asia, all but one study has been conducted in the Chinese and Japanese population. Four of them focused on non-cardia GC [8–11] and three also included cardia GC samples [4–6]. No information on histopathological types were available in most studies except two [12,13], because the Lauren classification is not widely used in Asia. All GWAS are listed in Table 1 along with the association findings. In summary, GC risk loci were identified on chromosome 1q22 (*MUC1*), 3q11 (*NSUN3*), 3q13 (*ZBTB20*), 4q28 (*ANKRD50*), 5p13 (*PRKAA1/PTGER4*), 5q14 (*Inc-POLR3G-4*), 6p21 (*LRFN2*), 8q24 (*PSCA*), 9q34 (*ABO*), 10q23 (*PLCE1/NOC3L*), 12q24 (*CUX2*), and 20q11 (*DEFB16*) [4–13]. Only one GC GWAS has been carried out in Europeans so far. This study used an Icelandic sample and showed GC associations within the gene *ATM* on chromosome 11q22. In addition, the risk East Asian loci on chromosome 1q22 (*MUC1*), 5p13 (*PRKAA1/PTGER4*) and 8q24 (*PSCA*) were replicated [40]. However, because this GWAS only used only 400 cases – along with 2,100 first-/second-degree relatives of patients that were counted as cases – only limited information on GC location and Lauren type was available.

Table 1: Overview of previous GC GWAS in the East Asian and European population. The GC type (Lauren, location) that was used in each GWAS is shown. Lead and genome-wide significant associated GC risk SNPs ($p < 5 \times 10^{-08}$) and their chromosomal location and position are indicated. In addition, genes are listed that were implicated as GC candidate genes in each study. Replication and/or downstream samples that were used are not listed.

GWAS (publication)	GC type	GWAS sample (country)	Lead-SNP per locus	SNP location (in bp (hg38))	Implicated gene(s)
GC GWAS in the East Asian population					
Sakamoto et al. (2008) Nat. Genet. [13]	Lauren diffuse	188 cases, 752 controls (Japan)	rs2294008	8q24 (142,680,513)	PSCA
Abnet et al. (2010) Nat. Genet. [4]	cardia, non-cardia	1,625 cases, 2,100 controls (China)	rs3781264 (f)	10q23 (94,310,618)	PLCE1, NOC3L
Shi et al. (2011) Nat. Genet. [8]	non-cardia	1,006 cases, 2,273 controls (China)	rs9841504	3q13 (11,4643,917)	ZBTB20
			rs13361707	5p13 (40,791,782)	PRKAA1, PTGER4
Jin et al. (2012) Am. J. Hum. Genet. [9]	non-cardia	1,006 cases, 4,016 controls (China) (a)	rs2494938	6p21 (40,568,389)	LRFN2
Hu et al. (2015) Gut [5]	cardia, non-cardia	2,350 cases, 2,708 controls (China) (b)	rs10074991 (g)	5p13 (40,790,449)	PRKAA1, PTGER4
			rs2294693 (h)	6p21 (41,037,763)	UNC5CL
Wang et al. (2017) Gut [10]	non-cardia	2,031 cases, 4,970 controls (China) (c)	rs4072037	1q22 (155,192,276)	MUC1
			rs80142782 (i)	1q22 (155,515,486)	MUC1, ASH1L
			rs7712641	5q14 (89,607,397)	lnc-POLR3G
			rs2294008	8q24 (142,680,513)	PSCA
Tanikawa et al. (2018) Cancer Sci. [12]	Lauren diffuse, intestinal	6,171 cases, 27,178 controls (Japan)	rs1057941	1q22 (155,216,951)	MUC1
			rs13361707	5p13 (40,791,782)	PRKAA1, PTGER4
			rs2294008	8q24 (142,680,513)	PSCA
			rs7849280	9q34 (133,251,249)	ABO
			rs6490061	12q24 (111,335,541)	CUX2
			rs2376549	20q11 (31,411,284)	DEFB
Yan et al. (2019) Gut [11]	non-cardia	3,771 cases, 5,426 controls (China) (d)	rs760077	1q22 (155,208,991)	MUC1
			rs6897169	5p13 (40,726,036)	PRKAA1, PTGER4
			rs10509671	10q23 (94,309,297)	PLCE1, NOC3L
			rs7624041	3q11 (94,389,819)	NSUN3
			rs10029005	4q28 (124,530,209)	ANKRD50
Ishigaki et al. (2020) Nat. Genet. [6]	cardia, non-cardia	6,563 cases, 195,745 controls (Japan)	rs760077	1q22 (155,208,991)	MTX1, THBS3
			rs3805495	5p13 (40,755,466)	TTC33
			rs2978977	8q24 (142,674,302)	JRK, PSCA
			rs11167159	20q11 (31,321,457)	DEFB16
GC GWAS in the European population					
Helgason et al. (2015) Nat. Genet. [40]	cardia, non-cardia	2,500 cases, 205,652 controls (Iceland) (e)	rs760077	1q22 (155,209,241)	MUC1
	Lauren diffuse, intestinal		p.Gln852*, p.Ser644* (j)	11q22 (108,222,832-108,369,099)	ATM

(a) samples used in Shi et al. [8]
(b) including samples used in Abnet et al. [4]
(c) including samples used in Abnet et al. [4] and Shi et al. [8]
(d) including samples used in Abnet et al. [4], Shi et al. [8] and Wang et al. [10]
(e) 400 GC patients and 2,100 first- or second-degree relatives from GC patients used as cases
(f) association was restricted to cardia GC and not present in non-cardia GC
(g) association to both cardia and no-cardia GC
(h) association was restricted to non-cardia GC and not present in cardia GC
(i) independent risk locus on chromosome 1q22, still genome-wide significant after conditioning on rs4072037
(j) identified using whole-genome sequencing data and GWAS significance threshold of $p < 3.1 \times 10^{-06}$

1.1.7 Monogenic Gastric Cancer Syndromes

The majority GC cases are sporadic and arise on the background of environmental and genetic risk factors, each contributing only with small effects to the overall individual risk. However, in about 10% of the cases a familial aggregation of GC is observed. Repeated cases in close relatives are an indication for highly penetrant genetic variants contributing with large effect sizes to GC development. And indeed, for 1-3% of the total GC cases an underlying genetic mutation can be identified [41].

By accounting for 1% of cases, hereditary diffuse gastric cancer (HDGC) is the most prominent monogenic cause of GC. This highly penetrant tumour syndrome is caused by heterozygous mutations in the gene *CDH1*, leading to a lifetime risk of 67% males and 83% in females for the development GC of the diffuse type or lobular breast cancer. In addition to *CDH1,* there is increasing evidence that also mutations *CTNNA1* are causal for HDGC [42].

Apart from HDGC there are several other tumour syndromes increasing the risk for the development of GC, although most of them being primarily associated with other tumour entities. In general, an increased risk for GC could be observed in patients diagnosed with Lynch syndrome, hereditary breast and ovarian carcinoma (HBOC), familial adenomatous polyposis (FAP), Li-Fraumeni syndrome, Peutz-Jeghers syndrome and juvenile polyposis [43]. Besides, a number of studies implicated further genes being of relevance for the development of GC that still need to be confirmed [44–46].

1.2 The Variability of the Human Genome

1.2.1 Common Variability of the Human Genome

In general, the human genome describes the entire heritable information contained in cells, whereby this work will focus entirely on the nuclear genome. The genetic information is encoded by specific sequences of the four bases adenine (A), cytosine (C), guanine (G) and thymine (T) in long chains of deoxyribonucleic acid (DNA). It comprises about 3.2 billion base pairs, which are distributed over 23 chromosomes and are diploid in most somatic cells, each set originating from one parent respectively [47].

Although the central dogma of molecular biology states that the most important function of the DNA is to encode for ribonucleic acid (RNA) transcripts, which are further translated into proteins, less than 2% of the genome sequence code for about

21,000 proteins [48]. The majority of the remaining genomic sequence is non-coding and assigning specific functions to these sequences is challenging [49].

In general, the genomic DNA sequence is more than 99% identical when comparing two unrelated individuals and yet it differs in millions of bases. These genetic differences can be subclassified according to the number of contiguous bases affected. In general, the frequency of variant types found in an individual decrease, the more contiguous bases are involved.

Aberrancies of chromosomal segments or whole chromosomes, represent the largest and rarest form of genetic variation [47]. Due to the large number of genes involved in such aberrations, it will usually result in non-viable pregnancies or severe phenotypes, leading to large selectional pressure. On the other hand, submicroscopic variations affecting more than 1 kb are more frequent and found in every individual. On average, more than 2000 of such copy number variants (CNVs) are present in every genome, deviating from the normal diploid status.

Deletions or insertions of smaller nucleotide segments (smaller than 1 kb), are referred to as indels. Each human genome harbours approximately 500,000 indels [47].

The smallest and thereby most common genetic variation is the exchange of single nucleotides. They are called single nucleotide variants (SNV) or single nucleotide polymorphisms (SNP). It is estimated that each individual carries approximately 4 million SNPs that differ from the reference sequence [50]. In total, over 700 million different SNPs have been described in the human genome so far, of which over 24 million are common and occur at a minor allele frequency (MAF) ≥ 1% in the respective population (Human build 154; [51,52]). For this thesis, the small variants including small indels and SNPs are of primary interest and will be introduced in more detail.

The large number of variants present in each individual indicates, that genomic variation is not the exception, but the rule in the human genome. Obviously, these variants most often do not have adverse or highly penetrant phenotypic effects for individual carriers as they are commonly seen across different populations. Most often, those variants are located in non-coding regions of the genome and/or do not have a direct effect on the function of encoded proteins, which is why they are also referred to as polymorphisms [47]. As they are under no, or only small selectional pressure, common variants are often quite old and got distributed across the globe. Due to their large number and their equal distribution across the genome, SNPs were soon utilized as genetic markers, enabling the examination of human genomes at a high resolution, providing many insights into the history of human development and events such as migrations or genetic bottlenecks.

However, from a technical point of view, the direct genotyping of several million SNPs across large samples of individuals was not feasible for a long time. This limitation was overcome when realizing that SNPs are not inherited independently, but that alleles are linked over larger segments in a block like structure across the genome [53]. This effect arises from intrachromosomal recombinations that occur during meiosis, in which segments of homologous chromosomes are exchanged (crossing-over) [54]. The breaking points of these blocks are not randomly distributed, but often affect specific genomic regions. Within these regions, the redistribution of alleles occurs less frequently, forming so called haplotypes. The allelic patterns within those haplotypes are linked, making it possible to deduce allelic information from one variant to another [55]. For this reason, this phenomenon is referred to as linkage disequilibrium (LD) and it can be quantified by using genotype information from a reference set of larger genome studies. Usually, the degree of linkage between variants is expressed by the measure r^2 and D'. Both measures can have values between 0 and 1. A value of 0 between two SNPs indicates a complete linkage equilibrium and an independent inheritance, whereas a value of 1 describes a perfect linkage in which the alleles of both SNPs are nearly always inherited together. This phenomenon allows to predict the genotype of SNPs in a same haplotype by inferring the genotypic information of some known SNPs in LD [56]. This method is called imputation and is mainly used for genotyping of large patients and control samples. Thus, the experimental determination of a few thousand carefully selected variants in an individual, for example by using SNP arrays, allows the statistical prediction of alleles of millions of other SNPs with a certain likelihood, given that there are reliable reference datasets available [57]. With the technical progress of genome sequencing, nowadays such reference datasets include thousands of individuals and allow an increasingly accurate imputation of genetic variants. Today, SNPs with a MAF > 0.5% can be imputed with sufficient precision in most populations.

1.2.2 Identification of Common Disease-Associated Variants using GWAS

Although common variants are referred to as polymorphisms, they play an important role in the development of multifactorial diseases. Multifactorial diseases are usually not confined to a single causal factor, but are the result of an unfavourable combination of environmental and genetic risk factors, whereby each risk factor usually contributes a relatively low increase to the overall disease risk of the individual carrier [14]. Due to the small effect sizes, there is no large selectional pressure on the risk conferring variants and they can be common in the population. Due to these characteristics, SNPs are usually used to identify such risk conferring genetic loci with a method called genome-wide association analysis. The basic principle is to compare genotype frequencies between case and control samples. If a specific allelic expression of a SNP occurs significantly more often in affected patients than in healthy or population-based controls, it is assumed that this variant is associated with the disease and, for example, increases the risk for the disease under investigation [58].

As described above, today several million variants can be genotyped and analysed at once. Due to the high number of variants under investigation, a correction for multiple testing is required. Here, it has been established that variants with a significance threshold of $P \leq 5 \times 10^{-8}$ are considered to be genome-wide significantly associated with the investigated phenotype [59].

At the beginning of the so-called "GWAS era", it was thought that only a few genetic risk loci would explain the entire genetic contribution to a specific multifactorial phenotype. However, it soon became clear that individual risk loci usually only account for a small proportion of the heritability (h^2), a measure of the inheritance of traits, and the risk of developing the phenotype under investigation [58]. Since then, patient and control samples have been increased in size to gain statistical power, making it possible to identify risk variants with smaller effect sizes. So far, over 200,000 associations to hundreds of phenotypes could be identified [60]. Nevertheless, for no multifactorial phenotype the genetic background could be fully elucidated. Whether this is due to the nature of multifactorial diseases or due to the contribution of uninvestigated rare variants is a matter of an ongoing scientific investigations and discussions [61,62].

Another drawback is that elucidating the biological mechanisms mediated by the identified risk loci has turned out to be difficult. In most cases, the identified variants are not causal, but are only in linkage to the actual risk conferring variant [63]. In addition, a large proportion of the identified risk variants are located in non-coding regions and, thus, do not directly affect the structure and function of individual protein,

but for example, indirectly regulate their expression being located in regulatory regions [49].

Still, the hypothesis free approach of this type of analysis, provided a lot of insights into pathways involved in disease development and showed up new overlaps and connections between phenotypes, which was not possible until the application of GWAS.

1.2.3 Assigning Common Genetic Variants to Function

As common genetic variants often reside in non-coding regions and only confer to disease susceptibility with small effects, elucidating their functional role is often difficult.

Large-scale projects, such as ENCODE, have shown, that large parts of the non-coding sequences are involved in various biochemical processes, playing an important role in regulating gene expression [49]. Due to this orchestration, non-coding sequences fundamentally represent the operating system of the genome, which can determine tissue differentiation and control responses to external and internal stimuli [49]. This highly complex interplay of transcription factor binding, DNA methylation, histone modifications, and other processes leads to highly individualized compositions of transcribed RNAs contained in each tissue and cell type at a given time, the so-called transcriptome. Ultimately, these transcriptome profiles mediate a large part of phenotypic characteristics of an organism. Consequently, the determination and quantification of the transcriptomes has become an important tool when studying signal pathways providing an indirect, but comprehensive insight into the outcome of gene regulatory processes described above. For this purpose, various gene expression array and RNA sequencing methods have been established, enabling transcriptome-wide examination of gene expression levels in any tissue of interest.

For human tissues, the Genotype-Tissue Expression project (GTEx) is the most comprehensive data resource of human transcriptome data, comprising 52 different tissues from over 800 individuals [64]. The high number of individuals included in this project allowed to conduct so called expression quantitative trait locus (eQTL) analyses. The aim of these eQTL studies is to identify genetic variants that have a quantitative influence on the expression of transcripts. For this purpose, the common variants of each individual are genotyped and correlated with the individual gene specific expression data. Comparing genotype and related gene expression levels across the whole sample may unravel changes in gene expression depending on the examined genotypes. Via this hypothesis free approach gene regulatory variants can be identified [64]. Cross-referencing such regulatory variants with disease associated

variants identified in GWAS, can provide strong hints to an underlying causal mechanism.

Depending on the distance between the variant and the regulated transcript, a distinction is made between *cis*- and *trans*-regulatory variants. Usually, variants with a distance <1 Mb from the effector gene are defined as *cis* and beyond as *trans*. *Cis*-regulatory effects are believed to be primarily driven by a direct variant transcript relationship. For example, a variant may alter the binding site of a transcription factor, thereby directly altering the expression of the nearby target gene. By contrast, *trans* -regulatory effects usually occur indirectly and independently of the genomic position [64]. For example, a regulatory variant may alter the expression of a transcription factors in *cis*, and the change in transcription factor expression then in turn may influence the expression of target genes downstream, which can be located on entirely different chromosomes. Such *cis-trans* effects also provide insights into potential signalling pathways. The above described GTEx project could identify *cis*-eQTL variants for over 18,000 protein-coding genes.

Aside from the annotation of single variants identified in GWAS, the genotype-transcript relationship can also be utilized for association testing of multifactorial disorders on a gene-based level. For this purpose, a tissue specific expression model is trained with transcriptome and genotype data, allowing to correlate genetic profiles with gene expression. This dataset can then be used to predict gene expression profiles solely based on genotype data. When imputing the gene expression for case and control data, for example for a GWAS dataset, a so-called transcriptome-wide association study (TWAS) can be performed. Here, the inferred gene expression levels are compared between cases and control groups, testing for an association of gene expression levels with a phenotype. The advantage of this method, amongst others, is a significant reduction of the multiple testing burden and a direct association of a phenotype with the expression status of a specific gene [15].

1.2.4 Genetic Correlation of Multifactorial Phenotypes

Unravelling the architecture of common variation in the human genome, its haplotype block structure and the availability of large datasets facilitated the development of methods to examine the phenotypes under investigation beyond the single marker level.

First, for cross-referencing significant loci across multiple phenotypes, a phenome-wide association study (PheWAS) can be performed to uncover pleiotropic effects of the investigated risk loci [65].

Another, approach to unravel genetic correlations between different phenotypes is LDscore regression (LDSC). This technique is based on a regression analysis examining the relationship between a test statistic of a GWAS and linkage disequilibrium data, thereby quantifying polygenic effects from confounding factors. From that, for a trait under investigation, SNP-based heritability estimates can be deducted and correlations to other phenotypes can be estimated. However, this type of analysis relies on rather large GWAS sample sizes (> 5000 cases) and heritability estimates need to be considered with some caution [66,67].

Another method used to estimate the genetic overlap between phenotypes is polygenic risk score analysis (PRS). Here, risk alleles identified in a GWAS are summed and weighted by their effect sizes. This analysis is not restricted to genome-wide significant variants, but can also include variants beyond the threshold of genome-wide significance. These scores are usually generated in a base dataset and then can be applied on a target dataset, e.g. using summary statistics of a trait of interest. Correlating these scores can give a strong hint of a causal relationship between phenotypes [68].

2 Aim of the Thesis

This thesis aims at the identification and interpretation of the genetic basis of GC, one of the most common malignant tumour entities affecting the human GI tract with a multifactorial aetiology. For this purpose, we collected the worldwide largest sample of European GC patients analysed so far. Of note, only a few germline genetic studies on GC in the European population have been carried out at the time of writing.

The first aim was to identify common genetic risk variants for GC development through a GWAS meta-analysis. Besides analysing the entire sample, we also stratified according to tumour location and histologic characteristics in order to identify risk variants that show subtype-specific disease association. We then carried out a TWAS and eQTL study using our GWAS data and expression data from gastric corpus and antrum mucosa in order to identify plausible GC risk genes and mechanisms among GWAS loci.

Another aim of this study was to determine the genetic correlation between GC and reported risk factors using LD score regression. Here, we used GWAS data from Europeans for various obesity-, reflux-, smoking-, alcohol- and education-/employment-related phenotypes that are publicly available. Finally, we carried out a PRS analysis in order to test whether cardia GC and OAC/BO, which are all located at the GOJ, share genetic aetiology. For this analysis we used a large European in-house OAC/BO GWAS sample.

In summary, this work should provide new scientific insights into disease-causing mechanisms of GC by the integration of different omics data as well as accounting for different GC subtypes on the phenotypic level.

2 Zielsetzung

Diese Arbeit befasst sich mit der Identifizierung und Interpretation molekulargenetischer Ursachen des Magenkarzinoms, eine der häufigsten malignen Tumorerkrankungen des Gastrointestinaltraktes mit multifaktorieller Ätiologie. Zu diesem Zweck wurde das bisher größte Kollektiv von Magenkarzinompatienten europäischer Herkunft gesammelt und genotypisiert.

Das erste Ziel war die Identifizierung häufiger genetischer Risikovarianten für die Entwicklung von Magenkarzinomen durch eine GWAS-Metaanalyse. Neben der Analyse des gesamten Kollektivs, wurde auch nach der Lokalisation und der histologischen Klassifikation der Tumore stratifiziert, um Risikovarianten zu identifizieren, welche eine subtypspezifische Krankheitsassoziation aufweisen. Anschließend wurde eine TWAS- und eine eQTL-Studie durchgeführt, bei welcher die GWAS-Daten und intern generierte Transkriptomdatensätze aus verschiedenen Bereichen der Magenschleimhaut verwendet wurden, um kausale Gene und Mechanismen an den Risikoloci aufzudecken.

Ein weiteres Ziel dieser Studie war es, die genetische Korrelation zwischen Magenkarzinomen und bekannten Risikofaktoren mithilfe von LD score regression zu bestimmen. Hierfür wurden öffentlich zugängliche GWAS-Datensätze mit übergewicht-, reflux-, rauch-, alkohol- und bildungs-/arbeitsplatzbezogenen Phänotypen untersucht. Zusätzlich wurde eine PRS-Analyse durchgeführt, um festzustellen ob Karzinome der Kardia und Adenokarzinome des Ösophagus, welche beide am gastroösophagealen Übergang lokalisiert sind, eine gemeinsame genetische Ätiologie haben. Für diese Analyse wurde ein interner GWAS-Datensatz zum Ösophaguskarzinom und Barrett-Ösophagus verwendet.

Zusammenfassend sollte diese Arbeit neue wissenschaftliche Erkenntnisse über krankheitsverursachende Mechanismen für die Entstehung von Magenkarzinomen identifizieren. Hierzu wurden omics-Daten verschiedener Ebenen integriert, sowie verschiedene Subtypen des Magenkarzinoms berücksichtigt.

3 Materials and Methods

Supplement A gives an overview of the

- devices
- chemicals, buffers, solutions
- commercial systems (kits)
- software and databases

used to conduct this study.

3.1 Patient and Control Samples

The samples described below build the basis for the genetic data generated in this work. All studies were conducted in accordance to the Declaration of Helsinki and were approved by locally appointed ethic committees at the respective study centres. Written informed consent has been obtained from all participants.

3.1.1 Gastric Cancer GWAS Sample

For the GC GWAS a total of ten samples were collected across Europe, subdivided according to the main nationalities and ethnicities to account for population stratification (Figure 3).

The cases were either collected specifically for this study, were already available in different biobanks, or had been collected in the context of other studies. Only cases with a histopathologically confirmed gastric adenocarcinoma were included. In addition, data on the age of onset, tumour location, classification according to Lauren, and the HP status were collected whenever possible. Details on the recruitment periods and the different study centres can be found in Supplementary Table 1. For the cases retrieved from the Estonian Biobank [69] and the UK Biobank [70] samples with a gastric adenocarcinoma according to the ICD-10 code C16 “malignant neoplasm of the stomach” were included.

Population based controls mainly originated from other studies, or were collected and genotyped for this study. Controls were selected to match with the respective sample of cases with regard to the population and the array technology used for genotyping. Details on the control samples regarding the recruitment period and information on the original studies are provided in Supplementary Table 2.

A total of 5,815 GC patients and 10,999 controls could be included in the GWAS. For the subtype-specific analyses 1,291 cardia, 3,183 non-cardia, 1,308 diffuse, and 1,696 intestinal GC patients were included. A detailed description of all samples is given in Table 2.

The overall sample composition matched roughly to epidemiologic data reported for GC, with a total male to female ratio of 1.86 [71]. Also the ratio between non-cardia and cardia gastric cancers of 2.47 lies within the margin of what is expected from previous reports, while four times more males than females were affected by cardia GC. There was only a marginal difference in the incidence between the intestinal and diffuse tumour type (1.3). Male to female ratios regarding tumour type were roughly the same for the diffuse, but about twice as high for the intestinal subtype.

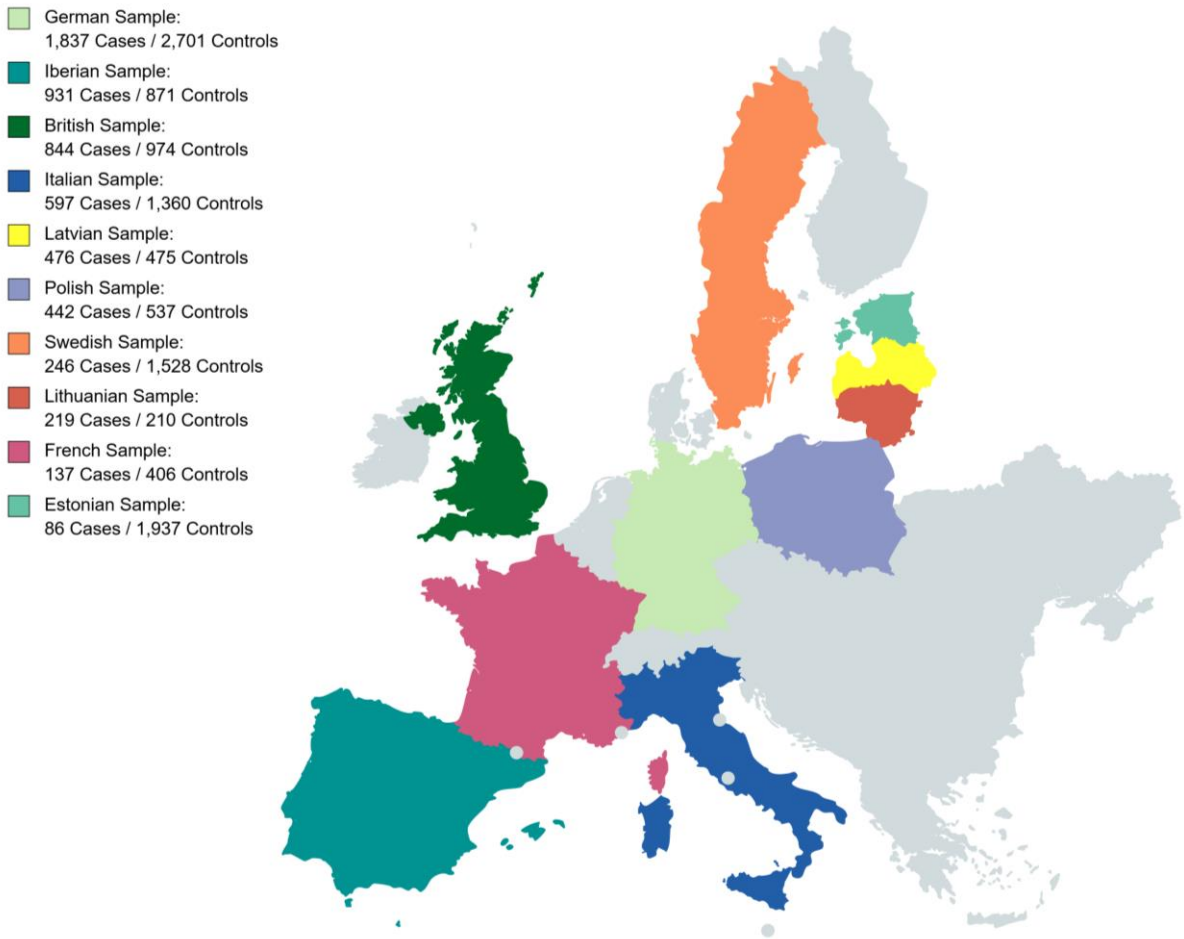


Figure 3: European map summarising the different samples collected for the GC GWAS meta-analysis.

Table 2: GC-GWAS sample overview. The number of GC patients are listed according to their origin, sex (female, male), location type (cardia, non-cardia) and Lauren type (diffuse, intestinal). In addition, the number of controls is listed according to their origin and sex (female, male).

Samples	**Entire GC cases**		**GC location type**				**GC Lauren type**				**Controls**	
	all	**female/male**	**cardia**	**female/male**	**non-cardia**	**female/male**	**diffuse**	**female/male**	**intestinal**	**female/male**	**all**	**female/male**
German sample	1,837	619/1,218	549	107/442	1,110	444/666	543	253/290	651	158/493	2,701	1,352/1,349
Iberian sample	931	322/609	175	25/150	733	291/442	281	133/148	408	137/271	871	429/442
British sample	844	244/600	351	65/286	88	34/54	-	-/-	-	-/-	974	532/442
Italian sample	597	231/366	75	20/55	401	158/243	100	44/56	271	100/171	1,360	506/854
Latvian sample	476	188/288	-	-/-	83	37/46	127	48/79	100	47/53	475	244/231
Polish sample	442	172/270	29	6/23	351	141/210	164	82/82	165	46/119	537	268/269
Swedish sample	246	95/151	27	12/15	135	51/84	-	-/-	1	-/1	1,528	842/686
Lithuanian sample	219	83/136	16	9/7	150	58/92	84	25/59	96	42/54	210	85/125
French sample	137	37/100	64	9/55	73	28/45	9	5/4	4	1/3	406	122/284
Estonian sample	86	45/41	5	2/3	59	32/27	-	-/-	-	-/-	1,937	1,101/836
Total	**5,815**	**2,036/3,779**	**1,291 (a)**	255/1,036	**3,183 (b)**	1,274/1,909	**1,308**	590/718	**1,696**	537/1,159	**10,999**	5,481/5,518

(a) from 513 cardia GC patients, Lauren information on diffuse GC type (N = 142 (27.68%)) or intestinal GC type (N = 371 (72.32%)) was available

(b) from 2,156 non-cardia GC patients, Lauren information on diffuse GC type (N = 1,002 (46.47%)) or intestinal GC type (N = 1,154 (53.53%)) was available

3.1.2 Gastric Tissue Gene Expression Sample

In order to determine the transcriptomic differences between stomach regions and to conduct eQTL and TWAS analyses, we collected tissue samples from healthy individuals undergoing a routine gastroscopy. From all individuals, biopsies from the corpus and the antrum were taken. For a subset of participants, additional biopsies from the cardia, fundus, and angulus were obtained. To control for environmental confounders that may influence the uniformity of the sample, all biopsies were collected during the morning and the probands fasted for at least eight hours prior to the procedure. Each region was sampled twice in direct proximity. One biopsy was preserved for the isolation of nucleic acids and one was taken for histological examination.

Only samples with a regular mucosa, which means no signs of gastritis or low-grade gastritis and no infection with HP were included. Histological examination was performed at the Institute of Pathology, Clinic of Bayreuth, Bayreuth, Germany.

A total of 422 participants were included in the study. Details on the recruitment centres and number of samples collected are given in Supplementary Table 3.

3.1.3 Cross-Trait GWAS Sample

To examine the genetic overlap between the different GC subtypes, OAC, and BO, data from a GWAS previously performed at the Institute of Human Genetics, University of Bonn, Bonn, Germany was used [72]. Raw genotype data required for several analyses were only available for the German in house sample comprising 2,646 cases and 2,732 controls. For the overall meta-analysis the entire sample of 10,279 OAC/BO cases and 27,326 controls was used. Details on the OAC/BO dataset are given in (Supplementary Table 4)

3.2 Extraction and Quality Control of Nucleic Acids

3.2.1 DNA Extraction from EDTA Blood

In-house array genotyping of samples used in the GC GWAS was performed with genomic DNA extracted from EDTA Blood. EDTA blood was collected at different study centres, stored at -20°C and sent on dry ice to the Institute of Human Genetics, University of Bonn, Bonn, Germany. For DNA extraction, the samples were thawed and subjected to the extraction process using the chemagic MSM I Instrument and the chemagic DNA Blood 10k Kit (Perkin Elmer, USA) according to the protocol provided by the manufacturer.

For the extraction process, a lysis buffer is added to the blood sample, releasing the genomic DNA into solution. The DNA binds to magnetic beads, which are transferred through a series of different washing buffers, which clear the sample of proteins and other unwanted contaminants. In the last step, the purified DNA is eluted into buffer TE and stored until further use at 4°C for short- and at -80°C for long-term storage.

3.2.2 Nucleic Acid Extraction from Gastric Tissue Samples

Genotyping and expression analysis for the gastric transcriptome study was performed with DNA and RNA extracted from tissue biopsies obtained during routine gastroscopies as described above. To prevent RNA degradation, the biopsies were directly transferred into vessels containing RNALater Solution (ThermoFisher, USA), stored over night at 4°C to immerse the sample with the stabilizing solution and subsequently stored at -20°C until shipment and at -80°C until extraction. To prevent batch effects, the samples were picked randomly for extraction.

For DNA and RNA extraction, the tissue biopsies (app. 5-10 mg) had to be homogenized first. For this purpose, the samples were transferred into a tube containing ceramic beads and a lysis solution. The tubes were loaded into a Precellys tissue homogenizer (Bertin Instruments, France), shaking the samples with the beads vigorously, thereby grinding the tissue due to bead collisions. The resulting solution was subsequently subjected to extraction using the Allprep DNA/RNA Mini Kit (Qiagen, Germany) according to the protocol provided by the manufacturer. During the silica-column based extraction method, DNA is bound to a silica membrane whereas the RNA is collected in the flow-through. After the addition of ethanol to the flow-through, it is applied to a second silica-column, binding the RNA. After additional washing steps contaminants are removed and the DNA and RNA samples can be eluted, subjected to quality control (QC) and stored at -80°C.

3.2.3 Quality Control and Quantification of Nucleic Acids

To determine the quality and quantity of the DNA and RNA samples, all samples were measured spectrophotometrically using a NanoDrop ND-1000 or ND-8000 photometer (ThermoFisher, USA). The concentration of the samples can be determined by absorbance measurements at different wavelengths. This method also allows to make assumptions on the presence of protein and salt contaminants, which may impair downstream applications.

For samples that were subjected to RNA-Seq, a more sensitive concentration measurement was required. For this purpose, the Qubit RNA BR Assay and the Varioscan Platereader (ThermoFisher, USA) were used according to the protocols given by the manufacturer. Instead of measuring the absorbance of the sample directly, in this assay an intercalating dye is added to the sample. By intercalating with the analyte of interest, the dye becomes fluorescent. The fluorescence intensity is directly proportional to the amount of analyte present in the sample. Measuring the fluorescence intensity with a photometer and by referring it to a previously determined standard of known concentrations, allows a precise determination of the concentration of the analyte of interest.

As RNA samples degrade quickly due to contamination with RNAses, the integrity of RNA samples had to be determined to prevent a possible confounding effect during the expression analysis. For this purpose, the RNA samples were subjected to a high-resolution gel electrophoresis using the RNA ScreenTape Kit and the TapeStation 4200 Device (Agilent, USA) as recommended by the manufacturer. For this analysis, the RNA in the sample is labelled with an intercalating dye and subjected to a gel electrophoresis. A marker added to the sample serves as a start- and end-point window during which fluorescent signals are measured. By comparison to a standard ladder, the fluorescent intensities can be used to precisely determine the size and concentration of RNA molecules present in the sample. To determine the integrity of the sample, the abundance of the 28S and the 18S ribosomal RNA (rRNA) subunits are set into relation. Briefly, a lower abundance of 28S rRNA in relation to 18S rRNA and a low concentration of the rRNA subunits in general indicate RNA degradation. From this value, the RNA Integrity Number (RIN) is deduced, which can range from 10 (no degradation) to 0 (complete degradation) [73]. For this study only samples with RIN > 5 were included in transcriptome analysis.

3.3 Genotyping and Gene Expression Analysis

3.3.1 Genome-Wide Genotyping via Illumina Bead Arrays

For the GC GWAS and the gastric eQTL dataset, genotyping was done using bead arrays from Illumina (USA), with exception of genotype data received from the UK Biobank (Affymetrix, USA). The array types used for each sample are given in Table 3.

All genotypes used for the gastric eQTL dataset and the majority of cases for the GC GWAS were generated at the Institute of Human Genetics, University of Bonn, Bonn, Germany. The remaining datasets were either obtained from cooperation partners, or downloaded from the respective database repositories.

For in-house genotyping, the different array types were processed semiautomatically according to the protocol provided by the manufacturer (Illumina, USA). The prepared arrays were scanned by an iScan System (Illumina, USA) to generate genotype raw data for downstream analysis. The Infinium array technology relies on silica beads to which specific DNA oligos are bound. The oligos contain individual address sequences enabling to identify each bead type, and a target sequence, which is complementary to a specific human genomic DNA sequence, targeting a single variant of interest. The different bead types are mixed, corresponding to the specific SNP content of the final array type. The beads are then fixated on the surface of silicon waver, within tiny, regular spaced wells. The wells are designed to bind a single bead each. Due to the individual address sequences contained in the oligos, the positions of the individual beads can be deduced (mapping). For the actual genotyping, the genomic DNA sample is prepared and loaded onto the array. Thereby, the respective sequences containing the variants of interest hybridize to their complementary oligos attached to the beads and get fixated. In a subsequent single base extension step, the target oligo is extended with a single fluorescently labelled nucleotide, corresponding to the complementary genotype of the SNP under investigation. Afterwards, the array is scanned and the fluorescent signals of each position are measured. From the fluorescent signal, the intensity and the bead type, the specific genotype of the SNP under investigation can be deduced.

Different array types are designed for different purposes, like examination of specific phenotypes or rare variants. Typically, the SNP content ranges from 500.000 up to several million. The present study used arrays of the Omni and Global Screening Array (GSA) family (see Table 3). These arrays are specifically designed for an efficient imputation of common variants.

To call the genotypes, the raw data were clustered against a reference dataset using the software Genomestudio v2 and the module Genotyping v2 (Illumina, USA) and exported in Plink format.

Table 3: Overview of GWAS genotyping arrays used in each sample. With exception for the data from the UK Biobank (Affymetrix, USA), all genotypes were determined using an Illumina (USA) microarray platform.

GC GWAS Sample	Array Types Cases	Array Types Controls
British Sample	UK Biobank Axiom Array	UK Biobank Axiom Array
Estonian Sample	OmniExpress v1.3	OmniExpress v1.3
French Sample	OmniExpressExome v1.3, Omni2.5Exome v1.3	Omni5M
German Sample	OmniExpressExome v1.3, Omni2.5Exome v1.3	OmniExpress v1.0, OmniExpress v1.1, Omni1Quad v1.0
Iberian sample	OmniExpressExome v1.3, Omni2.5Exome v1.3	OmniExpress v1.3, OmniEURHD
Italian sample	OmniExpressExome v1.3, Omni2.5Exome v1.3	Human1M-Duo
Latvian sample	OmniExpressExome v1.3, Omni2.5Exome v1.3	OmniExpress v1.3, OmniEURHD
Lithuanian sample	OmniExpressExome v1.3, Omni2.5Exome v1.3	OmniExpress v1.3, OmniEURHD
Polish sample	OmniExpressExome v1.3, Omni2.5Exome v1.3	OmniExpress v1.3
Swedish sample	OmniExpressExome v1.3, Omni2.5Exome v1.3	OmniExpress v1.0
Gastric eQTL Dataset	GSA v2.0	

3.3.2 Gastric Gene Expression Analysis via Illumina Bead Arrays

To examine the transcriptomic landscape across different parts of the human stomach, gene expression profiles were analysed in 47 RNA samples. The samples were derived as described above from tissue biopsies taken from the cardia, fundus, antrum and angulus from 9-11 individuals each.

For expression analysis, the HumanHT-12v4 Expression BeadChip (Illumina, USA) was used according to the protocol provided by the manufacturer. Same as for the genotyping array, the gene expression array relies on beads bound to specific oligos that can be mapped on the surface of a silica waver. Instead of targeting SNPs, the target sequences are complementary to coding regions of specific transcripts. To determine the expression levels of the respective transcripts, the RNA samples are reversely transcribed into complementary DNA (cDNA) and fluorescently labelled in a downstream process. The labelled cDNAs are hybridized to the complementary oligos present on the array. Subsequently, the array gets scanned, and the determined signal intensities are traced back to each bead and transcript. The fluorescent signal intensities thereby correspond to the abundance of the respective transcript in the sample and can be set into relation to other samples from different individuals or tissue types.

More than 47.000 transcripts can be analysed utilizing the array type described above. The raw data were assigned to the respective transcript and background signal intensities were subtracted using the software Genomestudio and the module Gene Expression (Illumina, USA). The resulting datasets were then exported for further QC

and data normalization. Only probes with a $P_{detection}$ < 0.01 in more than 5% of the samples were included for analysis. Furthermore, all probes were filtered for unique alignment and perfect or good quality as reported in the R package illuminaHumanv4.db. The resulting data were then subjected to explorative and differential gene expression analyses.

3.3.3 Gastric Gene Expression Analysis via RNA-Seq

Due to the discontinued production of the used expression array, the method of expression analysis for the larger eQTL and TWAS dataset was changed to 3' mRNA sequencing.

791 samples from the corpus and antrum were subjected to RNA-Seq using the QuantSeq 3′ mRNA-Seq Library Prep Kit FWD for Illumina (Lexogen, Austria). Libraries were sequenced on a HiSeq 2500 v4 (Illumina, USA) with a read length of 1x50bp, targeting for a coverage of 10 million reads per sample. To avoid batch effects, the samples were randomly picked for the library preparation.

In contrast to the array technology described above, during RNA-Seq transcripts are directly quantified by sequencing. In a first PCR step, oligodT primer containing Illumina read linker sequences align to the 3' poly(A) tail of the mRNA present in the sample and synthesize a complementary cDNA strand towards the 5' end of the transcript. Afterwards, the RNA is degraded, and in a second PCR random primers, containing an additional linker sequence, are incorporated in the opposite direction. This results in a double stranded cDNA library specifically containing the 3' portion of the original transcript. In a subsequent library amplification step, index sequences, enabling the identification of a specific sample, as well as adapter sequences are added to the cDNA strands. After purification, these samples can be pooled and subjected to sequencing.

The sequencing takes place in a highly parallelized manner. For this, the prepared DNA fragments bind randomly to the surface of a glass chamber, called flow cell, by hybridizing to printed counterparts of the incorporated adapter sequences. In a process called bridge amplification, the bound fragments are amplified. Because the fragments are bound to the flow cell surface, the produced copies stay in close proximity and form clusters. The actual nucleic acid sequence of these clusters is determined in a process called sequencing by synthesis. Thereby, the bound clusters are used as template strands in a reversible single base extension PCR with base specific fluorescently labelled nucleotides, followed by a scan for fluorescent intensity. After each base elongation, the labels are enzymatically cut off and the cycle is repeated. The position, the type and the intensity of the fluorescent signals are

registered. The resulting changes in fluorescence for the different clusters represent the actual nucleotide sequence of the initially bound nucleic acid strand. The incorporated index sequences allow the identification of the associated sample, for which all sequences are collected and stored in a single fastq file.

As compared to full transcript sequencing, the 3´-mRNA sequencing approach results in a single sequence read per transcript. The advantage is that a comparable low read coverage leads to robust data for quantification of gene expression given in transcripts per million (TPM).

To determine the expression levels of single transcripts and to make them usable for eQTL and TWAS analyses, the data needs to be controlled for quality issues and the expression values need to be normalized between samples. First, the adapter sequences need to be removed and the sequences need to be aligned to a reference, assigning them to a specific transcript. The quality control (QC) was performed with FastQC v0.11.7. For adaptor trimming, bbduk from the BBMap v 37.44 was used. The read alignment against GRCh38 was performed with STAR Aligner 2.5.2b. FeatureCounts v1.5.1 was used for quantification of gene expression using the Ensembl annotation GRCh38.89 as reference [74,75]. All parameters were used as recommended by the Lexogen‘s QuantSeq 3’ mRNA-Seq Kit and an integrated data analysis platform. All samples covered with at least three million reads were included in the downstream analyses. Uninformative transcripts were removed from analysis keeping only those with an expression above six counts per million in at least 20% of samples. The count values between samples were normalized using edgeR trimmed mean of M value normalization and expression for each gene was normalized by inverse normal transformation [76].

3.4 Quality Control, Statistics and Downstream Analyses

QC and statistical analyses of the retrieved data were performed in close collaboration with the Institute of Medical Biometry, Informatics and Epidemiology (IMBIE) and the Institute for Genomic Statistics and Bioinformatics (IGSB), University of Bonn, Bonn, Germany. Especially, many analyses were done in close cooperation with PD Dr. Michael Knapp, B.Sc. Jan Gehlen, M.Sc. Vitalia Schueller, Dr. Carlo Maj and Dr. Oleg Borisov.

3.4.1 Preimputation Quality Control of Genotype Data

Prior to imputation, the genotype datasets were checked for problems concerning data quality issues due to mismatched patient-phenotype information or technical reasons during processing. Due to the different array platforms used, the QC criteria differed slightly between the GWAS and the TWAS/eQTL dataset.

First, data sets with a call rate < 97% were excluded from further analysis. Insufficient call rates may be due to problems during the genotyping procedure or due to insufficient DNA quality.

Second, to identify genotyping issues for specific variants across the datasets, SNPs were removed from each control and case sample of the GWAS dataset, if the MAF was < 1%, if the genotyping rate was < 95%, or if the *P*-value for Hardy-Weinberg equilibrium (HWE) was < 10^{-04} in controls and < 10^{-06} in cases. For the TWAS/eQTL the same criteria were applied, except that the limit for the genotyping rate was < 90% and the *P*-value for the HWE was < 10^{-06}.

In a third step, the reported and the genetic gender were checked for mismatches to identify possible sample swaps. For this purpose, the rate of X-chromosomal heterozygous genotypes was determined. A rate of heterozygous SNPs < 2% was supposed for males and >10% for female individuals. Mismatches with the reported gender were followed up to identify the source of given sample swaps and corrected. If the swap could not be dissolved, the samples were excluded from further analysis.

Furthermore, closely related individuals or duplicates were removed from the analysis using PLINK version 1.9 [77] and KING [78]. From each pair of individuals with an estimated identity by descend probability > 0.2 or kinship coefficient > 0.0884, the individual with higher rate of missing genotypes was discarded.

To account for population stratification, each sample was checked for individuals being outliers in a multidimensional scaling analysis indicating an origin deviating from the Central European (CEU) population investigated in this study. Respective outliers were excluded from the further analysis.

3.4.2 Imputation and Postimputation Quality Control of Genotype Data

With the exception of the already imputed British and Estonian case-control samples, the genotype datasets were subjected to imputation to deduce the genotypes of common variants that were not directly determined by array genotyping.

For the GWAS samples, the TOPMed Imputation Server using the TOPMed Reference panel was used [79]. For the TWAS/eQTL dataset, Impute2 was used [80], utilizing the 1000 Genomes Phase 3 as reference [50].

In the post-imputation QC, variants with an $r^2 < 0.3$ (for the GWAS data) and an information score < 0.8 (for the TWAS/eQTL dataset), a HWE deviation of $P < 10^{-6}$ in patients or $P < 10^{-4}$ in non-patients, a minor allele frequency (MAF) < 1% or a SNP-missing rate > 5% for best-guessed genotypes at posterior probability > 0.9 were excluded.

3.4.3 Genome-Wide Association Study and Meta-Analysis

Significant differences in genotype allele frequencies between cases and controls indicate an association of a variant to the phenotype under investigation. Thus, we used the imputed genotype data to perform a GWAS in the entire GC sample and the GC subtypes.

We performed an association test considering an additive genetic model adjusting for five principal components (PCs) using PLINK2 for each sample [77]. The association analysis was performed for all national samples separately. Afterwards, a meta-analysis considering the fixed-effects inverse variance-weighting approach implemented in METAL was performed [81]. The same analysis was repeated after filtering the cases for the subtypes considering tumour location (cardia and non-cardia), tumour type according to Lauren (diffuse and intestinal) and gender. As subtype information was not present for all samples, only national collectives with $N_{cases} > 10$ were included in the respective subtype meta-analyses. By convention, variants with a *P*-value $< 5 \times 10^{-8}$ were considered being genome-wide significant [82].

In order to estimate whether associated loci are subtype specific, the cases of the cardia and non-cardia subsamples, and of the intestinal and diffuse cases were compared in a case against case comparison as described above.

To identify independent association signals, which means that an association signal prevails after stratification for the top variant at the locus, a conditional analysis was performed. For this purpose, a stepwise selection procedure as implemented in GCTA-COJO (GCTA version 1.93.0beta) was applied [83].

As the study revealed associated variants at the *ABO* locus, we inferred the ABO blood types of the individual samples according to Groot *et. al.* 2020 [84]. By combining the allelic expression of two single variants (rs8176746 and rs8176719), which taq the blood groups A/B and O/non-O respectively, the blood types can be determined. The

inferred blood groups were then tested for GC association using logistic regression analysis.

3.4.4 Phenome-Wide Association Study

In order to check whether the identified loci in the GWAS overlap with already identified disease associations, the lead variants were tested for significant association in already published datasets comprising the GWAS Catalog [60], UKBiobank [85] and FinnGen [86]. For this purpose, the database Open Targets Genetics Portal was used [87], considering a $P < 0.005$ as significant.

3.4.5 Differential Gene Expression across Five Gastric Regions

To characterize the transcriptomic landscape across the human stomach and in order to prioritize the gastric regions of interest for the larger eQTL dataset, the dataset with expression array data from five different stomach locations was subjected to explorative and differential gene expression (DE) analyses. For explorative analysis, the intensity data was quantile normalized using the R package Limma [88]. By principle component analysis (PCA) and unsupervised clustering, outliers and uniformity of transcriptomes between gastric locations was examined. DE analysis was done by determining significant differences in expression intensities ($P < 0.05$), with a delta in expression intensities > 80, and a fold change (FC) ≤ -2 or ≥ 2 across the examined groups. For pathway enrichment analysis, the online tool Enrichr was used [89].

3.4.6 Expression Quantitative Trait Locus Analysis

To identify variants with an effect on gene expression, it was examined whether the identified GC associated variants were listed in the Genotype-Tissue Expression project (GTEx) database [64] or were present in the corpus and antrum eQTL datasets generated within this study.

For the eQTL analysis, the quality controlled and normalized RNA-seq and genotype datasets were used as described above. The analysis was conducted for the corpus and antrum tissue samples separately. eQTLs were called using QTLtools [90] according to parameters used by GTEx project [64]. Briefly, sex, three genotype-based PCs and a set of probabilistic estimations of expression residuals (PEER) factors derived from the normalized expression data were used for adjustment. A window of 1 Mb around the transcription start site (TSS) of each gene was defined for *cis*-eQTL detection. SNP-gene pairs were considered significant with a nominal *P*-

value below a genome-wide empirical *P*-value threshold (P_t) determined for each gene by extrapolation from a Beta distribution fitted to adaptive permutations.

3.4.7 Transcriptome-Wide Association Analysis

Using the expression and genotype data from antrum and corpus, a TWAS was performed in order to prioritize GC risk genes at GWAS loci. The analysis was done for both tissues in all GC subsamples. We created expression prediction models for all genes in the transcriptome data with FUSION [91] using local SNPs present in the HapMap3 and our GC GWAS data (500 kb up- and downstream of the annotated gene TSS). Due to computational reasons, we omitted the use of the bslmm model. For each gene, the predicted gene expression was correlated with the GWAS data using LD data from prediction models in order to identify significant expression disease associations after Bonferroni-correction for the number of tested genes (P_{corpus} = 0.05 / 3,269 = 1.5 x 10^{-5}, P_{antrum} = 0.05 / 4,182 = 1.1 x 10^{-5}). Conditioning the data for the expression of identified genes was done to ensure that the association at the respective locus was sufficiently explained by the identified expression patterns.

3.4.8 LD Score Regression Analysis

To estimate the SNP-based heritability of GC and to identify genetic correlation to other phenotypes, a LD score regression analysis was performed using LDSC (v1.0.1) without changing the default parameters [92]. Thereby, the focus was set on phenotypes related to known GC risk factors (obesity, smoking, alcohol intake and socio economic status), which were available in the LD Hub database [93] based on data from the UKBiobank [85] and other publicly available GWAS summary statistics. A minimum number of app. 5.000 cases is required to perform a LD score regression. For this reason, only the entire GC GWAS sample was analysed, omitting the GC subtypes. Considering the total number of investigated traits, we defined an experiment-wide significance threshold using Bonferroni-correction (P = 0.05 / 20 = 2.5 x 10^{-3}).

3.4.9 Polygenic Risk Score Analysis

As LD score regression was not suitable to determine genetic overlap between GC subtypes due to small sample sizes, we utilized PRS analysis to examine whether a shared polygenic risk architecture between cardia GC, non-cardia and OAC exists. First, individual level PRS were computed by considering in-house GWAS data from 2,646 German OAC/BO patients and 2,732 German controls as base associations. The dataset was part of a GWAS meta-analysis published previously [94] and for which individual genotype data were available (Supplementary Table 4). To prevent an overlap of controls, all German controls were excluded in the OAC/BO sample, that were used in the GC-GWAS before. PRS were calculated after clumping (250 kb regions, clump-p = 1, clump-r2 = 0.1) by testing different *P*-value thresholds (from genome-wide significant ($P = 5 \times 10^{-8}$) to the full model ($P = 1$)) using PRSice tool [95] . As for the GWAS, the analysis was performed independently in each national sample considering the overall GC status. Logistic regression models between PRS and the phenotypic status were computed. The single-sample PRS regression coefficients (beta and standard error) were then combined into meta-analysis using the restricted maximum-likelihood (REML) estimator as implemented in the R package metaphor.

3.4.10 Cross-Trait GWAS Meta-Analysis

As the PRS analysis indicated a significant polygenic correlation between cardia GC as well as OAC and OAC/BO (see chapter 4.3.2), we conducted a cross-trait GWAS meta-analysis using the cardia GC and the German OAC/BO samples to identify additional risk loci. In addition, we used GWAS summary statistics from the other OAC/BO datasets that were published previously [94]. Supplementary Table 4 lists all samples that were included in the cross-trait GWAS meta-analysis, which was performed considering the fixed-effects inverse variance-weighting approach implemented in METAL [81].

4 Results

4.1 Gastric Cancer GWAS

4.1.1 Genome-Wide Significant Gastric Cancer Risk Loci

A total of six loci reached genome-wide significance in the GC meta-analysis across all subtypes examined: *MUC1* (1q22), *NEGR1* (1p31), *ALK* (2q23), *PSCA* (8q24), *HNF1B* (17q12), and *KLF6* (10p15). One independent signal was revealed by conditional analyses on 1q22. In addition, three previously described risk loci in the Asian population were replicated to confer to the risk of developing GC in the European population: *ANKRD50* (4q28), *PTGER4* (5p13), and *ABO* (9q34). For most of the identified loci, the eQTL and TWAS analysis revealed functional evidence by implicating transcriptional changes.

Details on the lead variants regarding the association signals in the different subtypes and their frequency across populations are given in Table 4. An overview of the genomic inflation quantile-quantile (QQ) plots and Manhattan-plots are given in Supplementary Figure 1 and Supplementary Figure 2. The effect sizes and directions across the single European samples included in the meta-analysis are presented in Supplementary Figure 3.

In the following, each locus will be presented in detail including a presentation of the results of the eQTL and TWAS analyses performed. The eQTL and TWAS are presented in more detail in the chapters 4.2.2 and 4.2.3.

Table 4: Lead associations of genome-wide significant and replicated GC risk loci. The associations are shown for the risk alleles (effect alleles) in the entire GC sample as well as in the location-specific (cardia, non-cardia) and Lauren-specific (diffuse, intestinal) GC samples. *P*-values, odds ratios (ORs) and the corresponding 95% confidence intervals (CIs) are shown. Allele frequencies for the associated SNPs among patients and controls are not given, as the GWAS samples were meta-analyzed. Instead the frequency of effect alleles in the European population are shown according to gnomAD [52]. In addition, on chromosome 1q22, 4q28, 5q13, 8q24 and 9q34 the LD between the lead GC SNPs in the present and East Asian GWAS are shown. At all five loci, the same alleles contribute to GC risk across populations.

SNP	Chromosome (position in bp (hg38))	Effect / other allele (a)	Entire GC sample			Cardia GC sample			Non-cardia GC sample			Diffuse GC sample			Intestinal GC sample		
			P-value	OR	95% CI	*P*-value	OR	95% CI	*P*-value	OR	95% CI	*P*-value	OR	95% CI	*P*-value	OR	95% CI
rs760077 (b)	1q22 (155.208.991)	T/A	5.23E-21	1.27	1.21-1.34	4.45E-02	1.09	1.00-1.19	5.12E-17	1.31	1.23-1.40	7.41E-11	1.35	1.23-1.48	1.78E-07	1.24	1.14-1.35
rs67579710 (c)	1q22 (155.203.736)	G/A	5.89E-13	1.21	1.11-1.31	-	-	-	4.62E-12	1.27	1.15-1.42	-	-	-	-	-	-
rs11677924	2q23 (29.500.326)	G/C	2.37E-07	1.19	1.11-1.27	5.91E-03	1.17	1.04-1.32	1.29E-05	1.20	1.10-1.30	1.90E-03	1.20	1.07-1.35	2.04E-08	1.34	1.21-1.49
rs10029005 (b)	4q28 (124.530.209)	A/G	4.69E-04	1.09	1.04-1.15	9.31E-01	1.00	0.91-1.09	4.37E-04	1.12	1.05-1.19	2.24E-02	1.11	1.01-1.21	7.72E-02	1.08	0.99-1.16
rs6897169 (b)	5q13 (40.726.036)	C/T	2.96E-04	1.13	1.06-1.21	2.65E-01	1.07	0.95-1.19	9.40E-05	1.18	1.09-1.28	4.20E-02	1.13	1.00-1.23	1.18E-02	1.15	1.03-1.27
rs2920293 (b)	8q24 (142.683.996)	G/C	2.84E-32	1.39	1.31-1.47	1.14E-01	1.08	0.98-1.19	1.80E-30	1.46	1.36-1.55	8.10E-17	1.46	1.33-1.60	4.05E-09	1.27	1.17-1.37
rs532436 (b)	9q34 (133,274,414)	A/G	7.51E-05	1.22	1.07-1.22	4.95E-01	1.04	0.93-1.17	5.82E-06	1.19	1.10-1.28	8.52E-07	1.29	1.17-1.44	3.02E-03	1.15	1.05-1.26
rs17138478	17q12 (37.713.312)	C/A	4.30E-06	1.19	1.10-1.28	3.96E-02	1.15	1.00-1.31	6.98E-04	1.17	1.07-1.29	8.83E-01	1.00	0.88-1.14	1.83E-08	1.44	1.27-1.64

(a) Frequency of effect alleles in the European (non-Finnish) population according to gnomAD [52]. rs760077 allele T 59%, rs67579710 allele G 90%, rs11677924 allele G 15%, rs10029005 allele A 42%, rs6897169 allele C 17%, rs2920293 allele G 47%, rs17138478 allele C 86%.

(b) rs760077 corresponds to the lead GC SNP on chromosome 1q22 in the East Asian population [6, 11]. r^2 of 0.63 between rs10029005, the lead GC SNP on chromosome 4q28 in the East Asian population [11], and rs7667950, the lead GC SNP at this locus in Europeans ($P = 9.42 \times 10^{-05}$ (OR = 1.12) in non-cardia GC cases). r^2 of 0.70 between rs6897169, the lead GC SNP on chromosome 5q13 in the East Asian population [11], and rs16870224, the lead GC SNP at this locus in Europeans ($P = 3.55 \times 10^{-05}$ (OR = 1.23) in non-cardia GC cases. r^2 of 0.88 between rs2920293 and rs2978977, the lead GC SNP on chromosome 8q24 in the East Asian population. rs532436 corresponds to the lead SNP rs7849280 in the East Asian population [12] and is in LD (r^2=0.76), however, no LD is observed in the European population (r^2=0.01) [6].

(c) rs67579710 was the only variant that showed additional association in the entire study after conditioning on lead SNPs using COJO [83]. The independent association signal on chromosome 1q22 appeared in the entire and non-cardia GC sample and, thus, the associations are not shown for the other location- or Lauren-specific GC samples

4.1.1.1 Genome-Wide Significant Risk Locus on Chromosome 1q22

On chromosome 1q22 a strong association signal was observed in the entire GC sample with rs760077 being the lead variant (odds ratio [OR] = 1.27; 95% confidence interval [CI] = 1.21-1.34; $P = 5.23 \times 10^{-21}$). Analysis of the subtypes revealed a strong association in the non-cardia (OR = 1.31; 95% CI = 1.23-1.40; $P = 5.12 \times 10^{-17}$) and diffuse GC subsamples (OR = 1.35; 95% CI = 1.23-1.48; $P = 7.41 \times 10^{-11}$), whereas the signal was considerably lower in the intestinal (OR = 1.24; 95% CI = 1.14-1.35; $P = 1.78 \times 10^{-07}$) and almost absent in the cardia GC subsample (OR = 1.09; 95% CI = 1.00-1.19; $P = 0.45$) (Table 4 and Figure 4).

rs760077 is a missense variant located within the coding region of exon 1 of *MTX1* and about 16 kb upstream of *MUC1*. In the in-house eQTL datasets, significant *cis*-eQTLs were identified, showing an upregulation of *MUC1* being associated to the GC risk conferring genotype in the corpus and antrum, and an upregulation for *THBS3* in the antrum (Table 9 and Figure 5).

These results were further supported by the results of the TWAS analyses, showing a transcriptome-wide significant upregulation of *MUC1* and *THSB3* in the antrum and an upregulation of *MUC1* in the corpus expression dataset for the non-cardia and intestinal GC subsamples respectively (Table 10). To assess the amount of residual association after accounting for the predicted *MUC1*-expression, a conditional analysis was performed. This revealed that the *MUC1*-expression explains most of the GWAS signals in corpus mucosa (Supplementary Figure 10). The upregulated *MUC1*-expression was considerably less significant GC-associated in antrum mucosa (non-cardia GC: $P = 5.26 \times 10^{-06}$, diffuse GC: $P = 1.01 \times 10^{-05}$) than in corpus mucosa (Table 10).

The same variant was already reported be associated to GC in the Asian population (Table 4). Moreover, the variant showed a number of significant associations to other phenotypes including weight and red blood cell count (Supplementary Table 6).

After conditioning for rs760077, an independent signal in the entire GC (OR = 1.21; 95% CI = 1.11-1.31; $P = 5.89 \times 10^{-13}$) and non-cardia GC sample (OR = 1.27; 95% CI = 1.15-1.42; $P = 4.62 \times 10^{-12}$) remained for rs67579710, which is located app. 5 kb upstream rs760077 within intron 4 of *THSB3*.

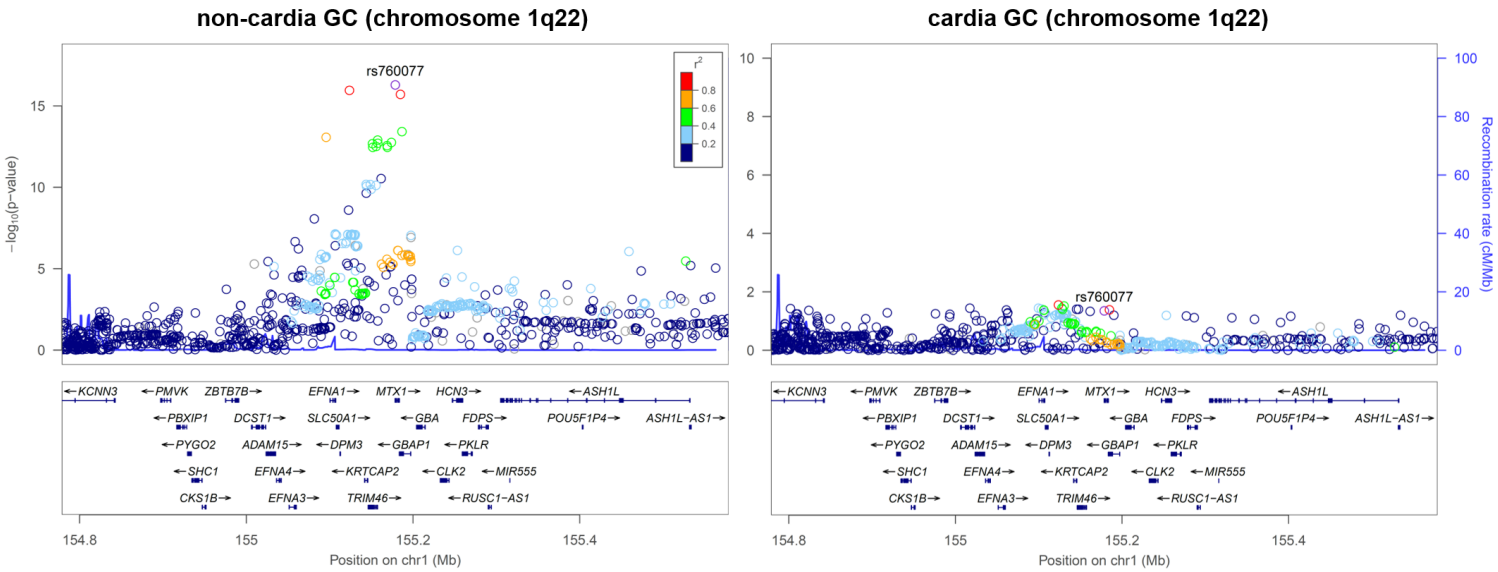


Figure 4: Regional association plots of GC risk locus 1q22 for non-cardia and cardia GC. Associations (-log10(*P*-values)) are shown for SNPs flanking 400 kb on either side of the lead associated SNP (position in hg19). The lead variant is shown in purple. Other markers at each locus are displayed by different colours, indicating different levels of LD (r^2) to the lead SNP. Furthermore, annotated genes within each region are shown with arrows indicating their transcription direction.

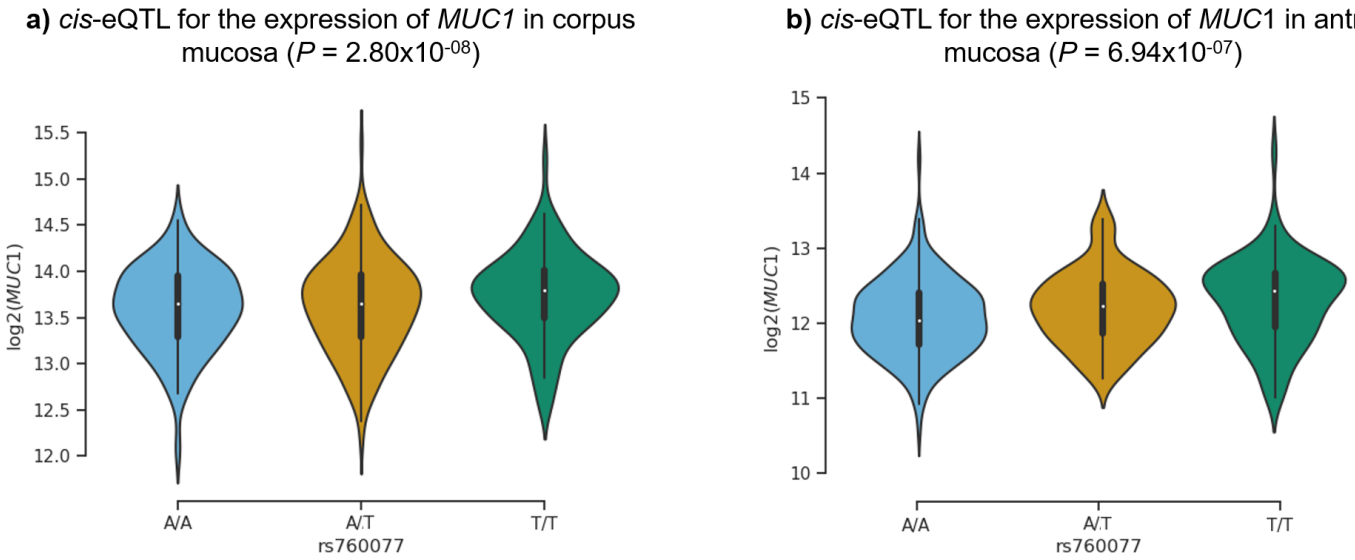


Figure 5: eQTL effects for the expression of *MUC1* on chromosome 1q22 in (a) corpus and (b) antrum. Log2 gene expression, error bars for median log2 expression and standard deviation are shown as box plots (y axis) sorted by SNP genotypes (x axis) with the GC risk allele on the left.

4.1.1.2 Genome-Wide Significant Risk Locus on Chromosome 2q23

In the intestinal GC sample, rs11677924 on chromosome 2q23 showed genome-wide significant association (OR = 1.34; 95% CI = 1.21-1.49; $P = 2.04 \times 10^{-8}$). The signal was also present in the entire GC sample, although not being genome-wide significant (OR = 1.19; 95% CI = 1.11-1.27; $P = 2.37 \times 10^{-7}$). By contrast, in the diffuse GC sample the variant showed only weak association (OR = 1.20; 95% CI = 1.07-1.35; $P = 0.002$) (Table 4 and Figure 6).

rs11677924 is located in intron 4 of *ALK*. The locus has not been described in other GC GWAS before. In addition, no significant eQTL effects, TWAS or PheWAS associations could be identified for the indicated variant nor the respective locus.

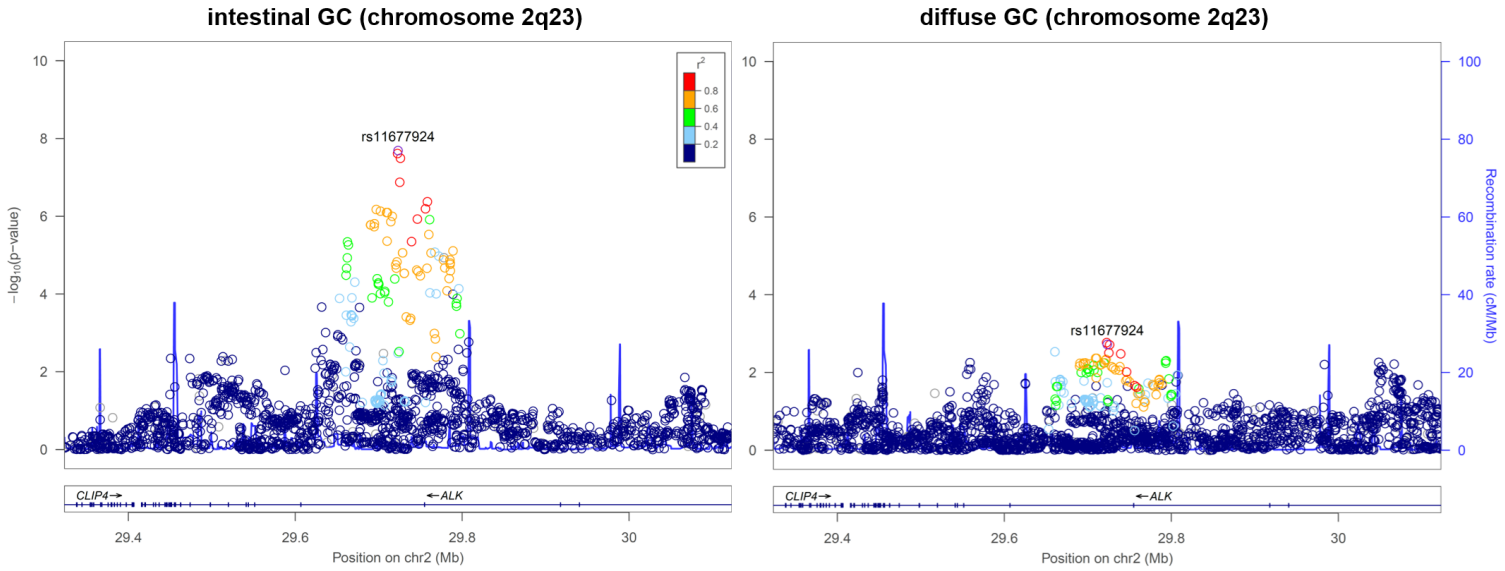


Figure 6: Regional association plots of GC risk locus 2q23 for intestinal and diffuse GC. Associations (-log10(*P*-values)) are shown for SNPs flanking 400 kb on either side of the lead associated SNP (position in hg19). The lead variant is shown in purple. Other markers at each locus are displayed by different colours, indicating different levels of LD (r^2) to the lead SNP. Furthermore, annotated genes within each region are shown with arrows indicating their transcription direction.

4.1.1.3 Genome-Wide Significant Risk Locus on Chromosome 8q24

The strongest GC association signal in the entire sample was observed for rs2920293 (OR = 1.39; 95% CI = 1.31-1.47; $P = 2.84 \times 10^{-32}$). The variant was also found to be highly significant in the non-cardia GC (OR = 1.46; 95% CI = 1.36-1.55; $P = 1.80 \times 10^{-30}$) and the diffuse GC sample (OR = 1.46; 95% CI = 1.33-1.60; $P = 8.10 \times 10^{-17}$). By contrast, the signal was of considerably lower significance and effect size in the intestinal sample (OR = 1.27; 95% CI = 1.17-1.37; $P = 4.05 \times 10^{-09}$), and absent in cardia GC (OR = 0.92; 95% CI = 0.83-1.02; $P = 0.114$) (Figure 7). The direct case to case comparison revealed a significant signal for the non-cardia versus cardia cases (OR = 1.29.; 95% CI = 1.16-1.45; $P = 4.00 \times 10^{-06}$) (Supplementary Table 5). rs2920293 is not biallelic and as such not present in publicly available and the in-house eQTL datasets. For this reason, we included rs2920292 in the analysis, being located 285 bp upstream and showing similar effect sizes in the entire GC sample (OR = 1.34; 95% CI = 1.28-1.41; $P = 5.53 \times 10^{-31}$), and the strongest signal in the non-cardia sample (OR = 1.46; 95% CI = 1.37-1.55; $P = 3.48 \times 10^{-31}$).

The variant rs2920292 is located in a non-coding region app. 1 kb upstream the TSS of *PSCA*.

In the in-house eQTL datasets, rs2920292 was found to account for one of the strongest eQTL effects observed in the entire sample, whereby the risk allele is associated with an increase in *PSCA* transcription in corpus as well as antrum tissue. To a lesser extent, *THEM6* is upregulated also in both tissues. Furthermore, in the corpus tissue, *LY6K* is up- and *LYNX1* is downregulated (Table 9).

Comparable effects for the locus were observed in the TWAS analysis, with *PSCA*, *LY6K*, *THEM6* and *LYNX1* showing transcriptome-wide significance (Table 10). In addition, conditional analyses in the TWAS revealed *PSCA* as the only gene whose expression explained most of the GWAS signals (Supplementary Figure 9). As observed in the GWAS, the TWAS signals were not present in the cardia subsample and more pronounced in the diffuse type as compared to the intestinal type.

The 8q24 locus has been described to be associated to GC in several GWAS in the Asian population before (Table 4). This finding was replicated in the PheWAS along with an increased risk for malignant neoplasms of the bladder and a protective effect for the development of gastric and duodenal ulcers (Supplementary Table 7).

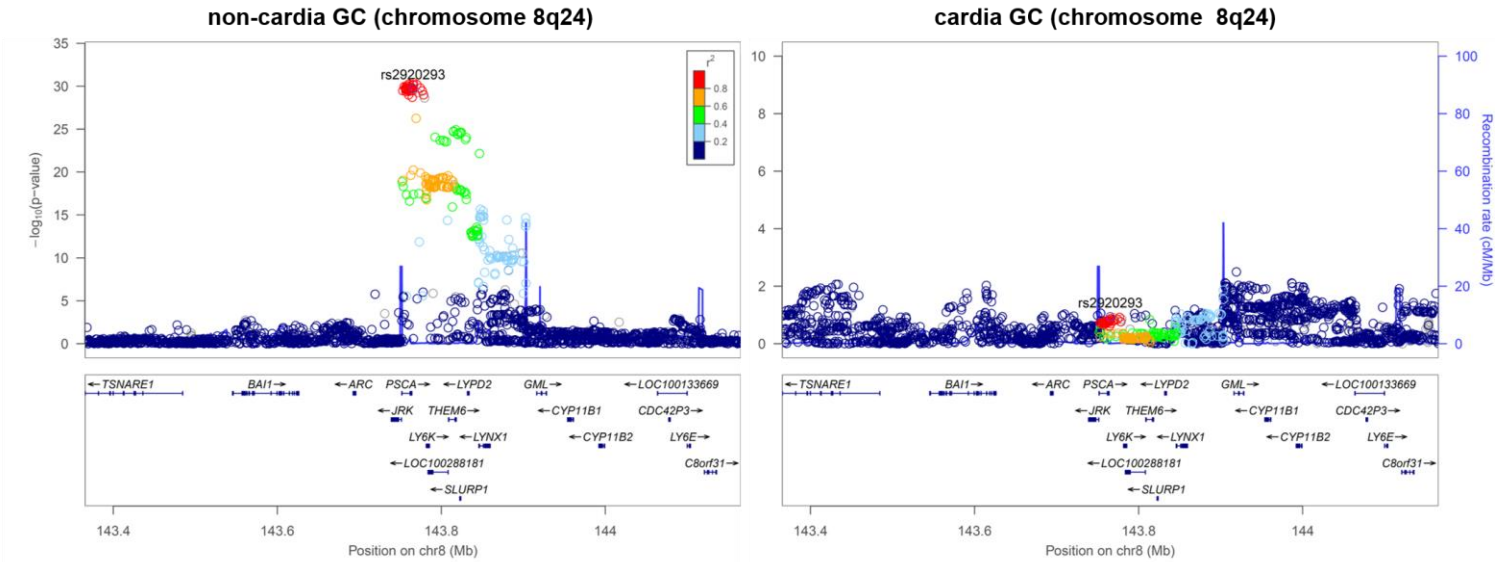


Figure 7: Regional association plots of GC risk locus 8q24 for non-cardia and cardia GC. Associations (-log10(p-values)) are shown for SNPs flanking 400 kb on either side of the lead associated SNP (position in hg19). The lead variant is shown in purple. Other markers at each locus are displayed by different colours, indicating different levels of LD (r^2) to the lead SNP. Furthermore, annotated genes within each region are shown with arrows indicating their transcription direction.

a) *cis*-eQTL for the expression of *PSCA* in corpus mucosa ($P = 3.64x10^{-114}$)

b) *cis*-eQTL for the expression of *PSCA* in antrum mucosa ($P = 1.28x10^{-104}$)

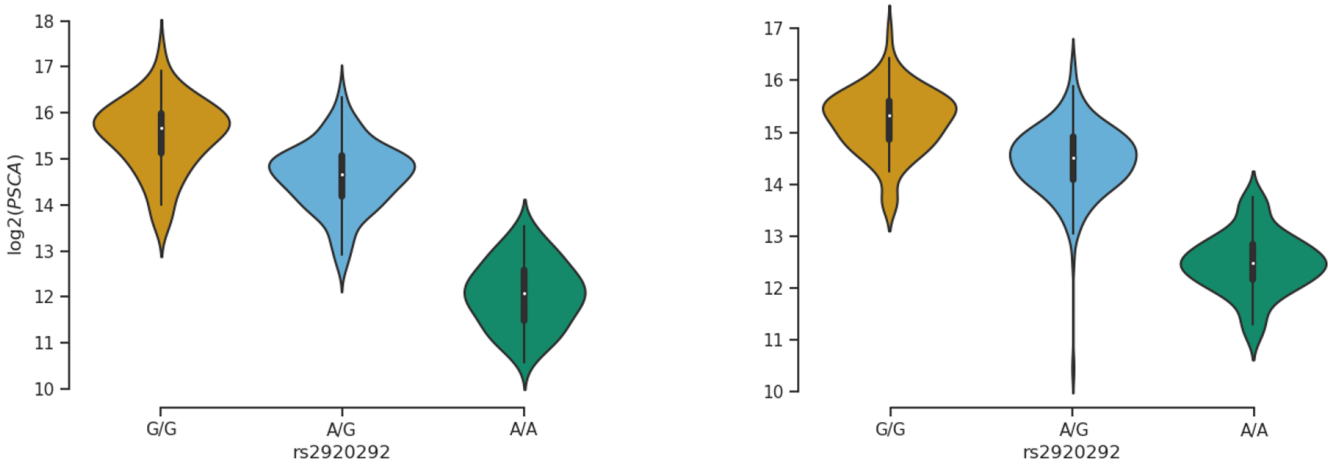


Figure 8: eQTL effects for the expression of *PSCA* on chromosome 8q24 in (a) corpus and (b) antrum. Log2 gene expression, error bars for median log2 expression and standard deviation are shown as box plots (y axis) sorted by SNP genotypes (x axis) with the GC risk allele on the left.

4.1.1.4 Genome-Wide Significant Risk Locus on Chromosome 17q12

Another genome-wide significant signal was observed in the GC intestinal sample for the variant rs17138478 on chromosome 17q12 (OR = 1.44; 95% CI = 1.27-1.64; *P* = 1.83 x 10^{-8}) (Figure 9). This variant showed a suggestive association in the entire GC (OR = 1.19; 95% CI = 1.10-1.28; *P* = 4.30 x 10^{-6}), but no association in the diffuse GC sample (*P* = 0.883) (Table 4). The direct case to case comparison between intestinal and diffuse GC, indicates a specific association for the intestinal tumour type (OR = 1.47; 95% CI = 1.25-1.73; *P* = 4.38 x 10^{-6}).

The variant rs17138478 is located in the intron 4 of *HNF1B*. No significant eQTLs or TWAS signals were detected at this locus.

The variant has not been described in the context of GC before. However, in the PheWAS the variant could be associated, among others, with an elevated risk for prostate cancer, elevated liver enzyme transaminase levels and cholelithiasis (Supplementary Table 8).

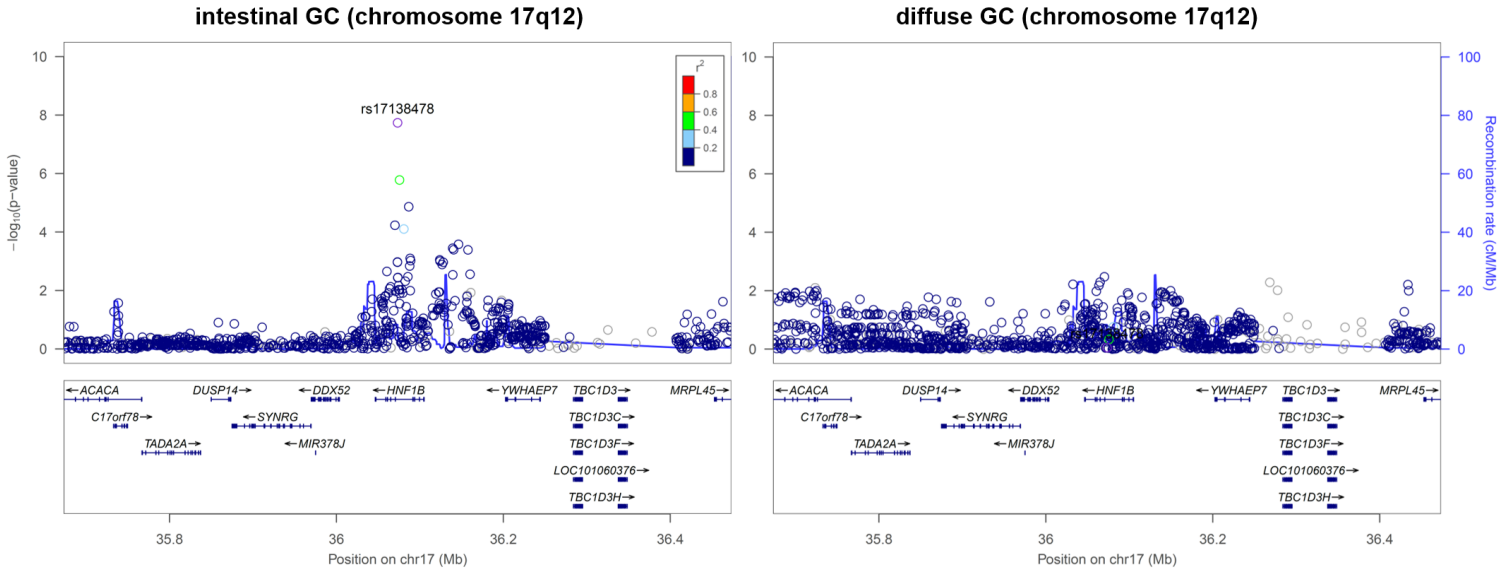


Figure 9: Regional association plots of GC risk locus 17q12 for intestinal and diffuse GC. Associations (-log10(*P* -values)) are shown for SNPs flanking 400 kb on either side of the lead associated SNP (position in hg19). The lead variant is shown in purple. Other markers at each locus are displayed by different colours, indicating different levels of LD (r^2) to the lead SNP. Furthermore, annotated genes within each region are shown with arrows indicating their transcription direction.

4.1.2 Sex-Specific Gastric Cancer Risk Loci

As incidences and risk factors for developing GC are deviating between males and females [34], we also performed GWAS stratifying for gender. For the genome-wide significant loci described above, no sex specific effects were observed (data not shown). However, two additional genome-wide significant loci could be identified at chromosome 1p31 and 10p15, specifically being associated in the female cardia and female non-cardia subsamples respectively (Table 5).

Table 5: Lead associations of genome-wide significant and sex specific GC risk loci. The associations are shown for the risk alleles (effect alleles) in the entire GC sample as well as in the location-specific (cardia, non-cardia) GC samples stratified according to gender. *P*-values, odds ratios (ORs) and the corresponding 95% confidence intervals (CIs) are shown. Allele frequencies for the associated SNPs among patients and controls are not given, as the GWAS samples were meta-analyzed. Instead, the frequency of effect alleles in the European population are shown according to gnomAD [52].

SNP	Chromosome (position in bp (hg38))	Effect / other allele (a)	Entire GC			Cardia GC			Non-cardia GC		
			***P*-value**	**OR**	**95% CI**	***P*-value**	**OR**	**95% CI**	***P*-value**	**OR**	**95% CI**
Female											
rs2590943	1p31 (72,408,773)	A/G	9.02E-03	1.15	1.04-1.28	1.21E-09	1.93	1.56-2.38	3.30E-01	1.07	0.94-1.22
rs1547179	10p15 (4,379,326)	G/T	1.70E-07	1.27	1.16-1.39	2.06E-01	1.13	0.93-1.37	2.18E-08	1.38	1.23-1.54
Male											
rs2590943	1p31 (72,408,773)	A/G	8.76E-01	1.01	0.92-1.10	8.32E-01	1.01	0.90-1.13	8.32E-01	1.01	0.90-1.13
rs1547179	10p15 (4,379,326)	G/T	1.85E-01	0.95	0.88-1.02	2.20E-01	0.94	0.86-1.04	2.20E-01	0.94	0.86-1.04

(a) Frequency of effect alleles in the European (non-Finnish) population according to gnomAD [52]. rs2590943 allele A 18%, rs1547179 allele G 70%.

4.1.2.1 Female-Specific Risk Locus on Chromosome 1p31

One sex-specific locus is located on chromosome 1p31 with rs2590943 showing the strongest association signal in the female cardia sample (OR = 1.93; 95% CI = 1.56-2.38, $P = 1.21 \times 10^{-9}$). The signal was absent in the female non-cardia sample ($P = 0.330$), as well as absent in the male cardia sample ($P = 0.675$).

rs2590943 is a non-coding variant, app. 125 kb upstream the nearest gene *NEGR1*. No antrum or corpus specific eQTLs could be identified in this region.

The variant has not been described in the context of GC before. In the PheWAS analysis, a strong association for an elevated BMI and related phenotypes such as GERD were observed (Supplementary Table 9). We further checked whether for these phenotypes sex-specific effects can be observed in publicly available datasets [96], however, no clear effect was seen (Table 6).

Table 6: Associations of rs2590943 for BMI and GERD in the UKBB [70] stratified according to gender are shown. *P*-values, odds ratios (ORs) and the corresponding 95% confidence intervals (CIs) are shown.

SNP	Chromosome (position in bp (hg38))	Effect / other allele	BMI			GERD		
			P-value	OR	95% CI	*P*-value	OR	95% CI
			Female					
rs2590943	1p31 (72,408,773)	A/G	3.08E-07	1.02	1.01-1.03	5.01E-04	1.00	1.00-1.00
			Male					
rs2590943	1p31 (72,408,773)	A/G	3.02E-18	1.04	1.03-1.05	9.38E-02	1.00	1.00-1.00

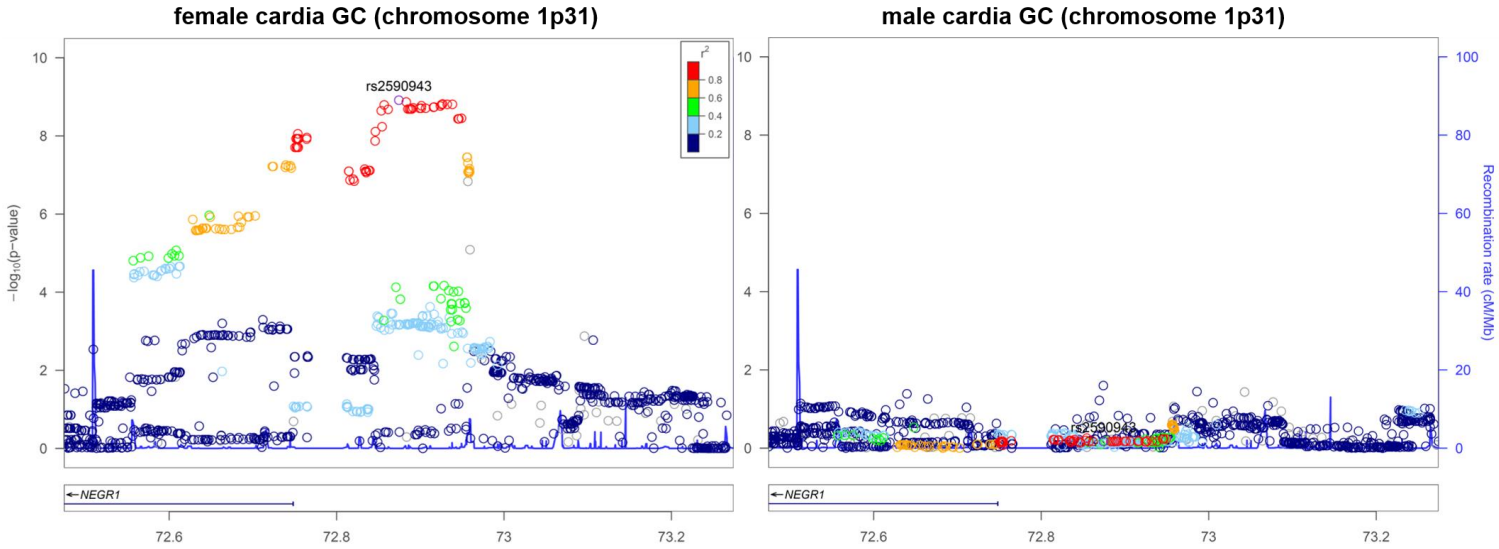


Figure 10: Regional association plots of GC risk locus 1p31 for female and male cardia GC. Associations (-log10(*P*-values)) are shown for SNPs flanking 400 kb on either side of the lead associated SNP (position in hg19). The lead variant is shown in purple. Other markers at each locus are displayed by different colours, indicating different levels of LD (r^2) to the lead SNP. Furthermore, annotated genes within each region are shown with arrows indicating their transcription direction.

4.1.2.2 Female-Specific Risk Locus on Chromosome 10p15

The second sex-specific genome-wide significant signal was found on chromosome 10p15. The variant rs1547179 showed the strongest signal in the female non-cardia sample (OR = 1.38; 95% CI = 1.23-1.54, *P* = 2.18 x 10^{-8}) and was absent both in the female cardia sample (*P* = 0.206) and cardia male sample (*P* = 0.345).

The variant is located in a non-coding region with *AKR1E2* being the closest coding gene located app. 450 kb downstream. No antrum or corpus specific eQTLs could be identified. Also in the PheWAS, no significant associations could be observed.

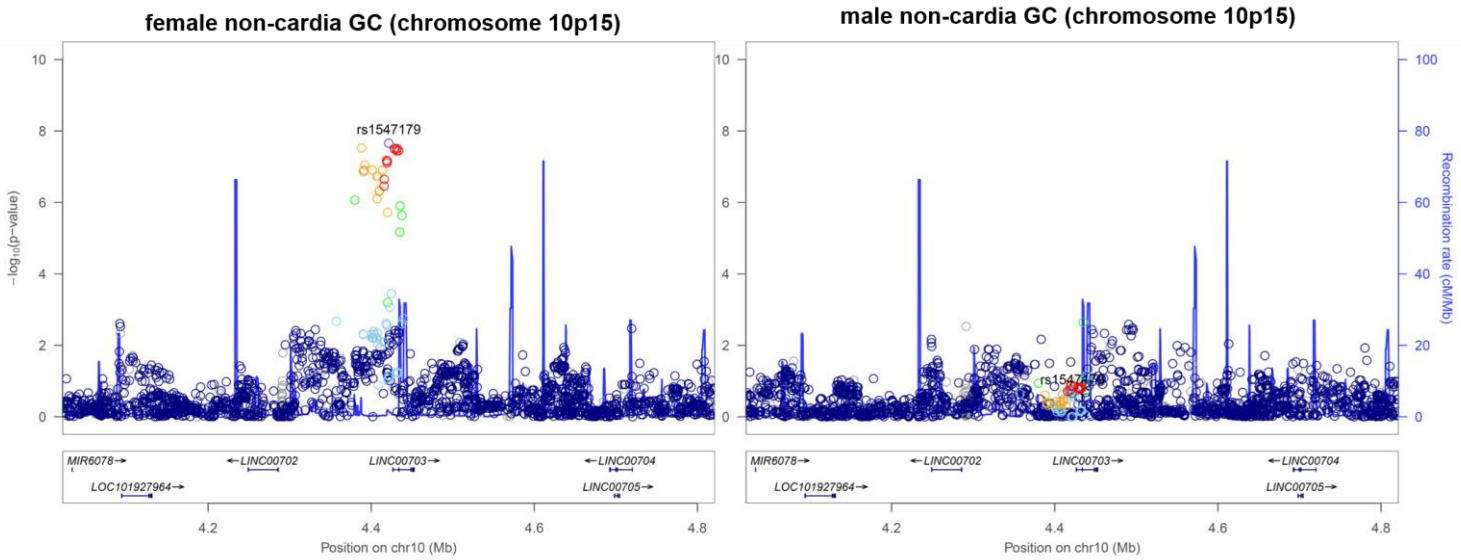


Figure 11: Regional association plots of GC risk locus 10p15 for female and male non-cardia GC. Associations (-log10(*P*-values)) are shown for SNPs flanking 400 kb on either side of the lead associated SNP (position in hg19). The lead variant is shown in purple. Other markers at each locus are displayed by different colours, indicating different levels of LD (r^2) to the lead SNP. Furthermore, annotated genes within each region are shown with arrows indicating their transcription direction.

4.1.3 Replication of Asian Gastric Cancer Risk Loci

As described above, the risk loci on chromosome 1q22 and 8q24 were already known to be associated to GC in the Asian population and showed genome-wide significance in the present European analysis (Table 4). In addition, we checked the remaining eight risk loci reported in the Asian GWAS for replication in our sample (Table 7). We confirmed three loci to be associated in the European population, while for the remaining five loci no association was found.

Table 7: GC associations of East Asian risk loci in Europeans. All associations are shown for the risk alleles (effect alleles) obtained in the East Asian GWAS studies. *P*-values are shown for the entire GC sample in the present study. *P*-values, odds ratios (ORs) and the corresponding 95% confidence intervals (CIs) are shown for the entire GC sample in the present study.

GWAS SNP	SNP location (in bp (hg38))	GWAS (publication)	Implicated gene(s)	Effect allele / other allele	*P*-value	OR	95% CI
rs4072037	1q22 (155,192,276)	Wang et al. (2017) Gut [10]	*MUC1*	T/C	7.88E-07	1.13	1.07 to 1.19
rs760077	1q22 (155,208,991)	Yan et al. (2019) Gut [11] ; Ishigaki et al. (2020) Nat. Genet. [6]	*MTX1, THBS3*	T/A	5.23E-21	1.27	1.21 to 1.34
rs80142782 (a)	1q22 (155,515,486)	Wang et al. (2017) Gut [10]	*MUC1, ASH1L*	NA	NA	NA	NA
rs1057941	1q22 (155,216,951)	Tanikawa et al. (2018) Cancer Sci. [12]	*MUC1*	G/A	1.9E-14	1.22	1.16 to 1.28
rs117950304	1p31 (80,854,393)	Tanikawa et al. (2018) Cancer Sci. [12]	*RPL7P10*	A/G	NA	NA	NA
rs78390645	2q24 (159,394,138)	Tanikawa et al. (2018) Cancer Sci. [12]	*BAZ2B*	C/A	NA	NA	NA
rs7624041	3q11 (94,389,819)	Yan et al. (2019) Gut [11]	*NSUN3*	G/A	0.251	1.05	0.96 to 1.16
rs9841504	3q13 (11,4643,917)	Shi et al. (2011) Nat Genet [8]	*ZBTB20*	G/C	0.768	1.01	0.92 to 1.11
rs10029005	4q28 (124,530,209)	Yan et al. (2019) Gut [11]	*ANKRD50*	A/G	4.69E-04	1.09	1.04 to 1.05
rs6897169	5p13 (40,726,036)	Yan et al. (2019) Gut [11]	*PRKAA1, PTGER4*	C/T	2.69E-04	1.13	1.05 to 1.20
rs3805495	5p13 (40,755,466)	Ishigaki et al. (2020) Nat. Genet. [6]	*TTC33*	C/T	3.04E-04	1.11	1.04 to 1.17
rs10074991	5p13 (40,790,449)	Hu et al. (2015) Gut [5]	*PRKAA1, PTGER4*	G/A	7.01E-04	1.10	1.04 to 1.16
rs13361707	5p13 (40,791,782)	Shi et al. (2011) Nat Genet [8], Tanikawa et al. (2018) Cancer Sci. [12]	*PRKAA1, PTGER4*	C/T	6.56E-04	1.10	1.04 to 1.16
rs7712641	5q14 (89,607,397)	Wang et al. (2017) Gut [10]	*lnc-POLR3G*	T/C	0.422	1.02	0.96 to 1.07
rs2494938	6p21 (40,568,389)	Jin et al. (2012) Am J Hum Genet [9]	*LRFN2*	G/A	0.499	0.98	0.93 to 1.03
rs2294693	6p21 (41,037,763)	Hu et al. (2015) Gut [5]	*UNC5CL*	C/T	0.791	0.99	0.93 to 1.05
rs2978977	8q24 (142,674,302)	Ishigaki et al. (2020) Nat. Genet. [6], Tanikawa et al. (2018) Cancer Sci. [12]	*JRK, PSCA*	A/C	0.122	1.10	0.97 to 1.26
rs2294008	8q24 (142,680,513)	Sakamoto et al. (2008) Nat Genet [6], Wang et al. (2017) Gut [10], Tanikawa et al. (2018) Cancer Sci. [12]	*PSCA*	C/T	7.51E-30	0.74	0.70 to 0.78
rs7849280	9q34 (133,251,249)	Tanikawa et al. (2018) Cancer Sci. [12]	*ABO*	A/G	0.049	0.99	0.90 to 1.09
rs532436	9q34 (133,274,414)	Replication in European population (b)	*ABO*	A/G	7.51E-05	1.22	1.07 to 1.22
rs10509671	10q23 (94,309,297)	Yan et al. (2019) Gut [11]	*PLCE1, NOC3L*	G/T	0.992	1.00	0.94 to 1.05
rs3781264	10q23 (94,310,618)	Abnet et al. (2010) Nat Genet [4]	*PLCE1, NOC3L*	G/A	0.936	1.00	0.94 to 1.05
rs6490061	12q24 (111,335,541)	Tanikawa et al. (2018) Cancer Sci. [12]	*CUX2*	T/C	0.700	0.99	0.93 to 1.05
rs11167159	20q11 (31,321,457)	Ishigaki et al. (2020) Nat. Genet. [6]	*DEFB16*	G/GT	0.521	1.04	0.91 to 1.19
rs2376549	20q11 (31,411,284)	Tanikawa et al. (2018) Cancer Sci. [12]	*DEFB121, DEFB119*	C/T	0.837	1.01	0.96 to 1.06

(a) rs80142782 is monoallelic in the European population, therefore no data are given for this variant.

(b) rs7849280 is in high LD to rs532436 in the East Asian population (r^2=0.76), but not in the European population (r^2=0.01), as an association for rs532436 was seen it was included to examine replication

4.1.3.1 Replication of Susceptibility Locus 4q28

On chromosome 4q28 the variant rs10029005 revealed the strongest association in the European GC non-cardia sample (OR = 1.12; 95% CI = 1.05-1.19; P = 4.37 x 10^{-4}). By contrast, no association was found in the European GC cardia sample (P = 0.931) (Table 4 and Figure 12).

rs10029005 is a non-coding variant located app. 130 kb downstream the coding region of *ANKRD50*. A significant upregulation of *ANKRD50* in the corpus in risk allele carriers was identified in the in-house eQTL dataset (Table 9 and Figure 13).

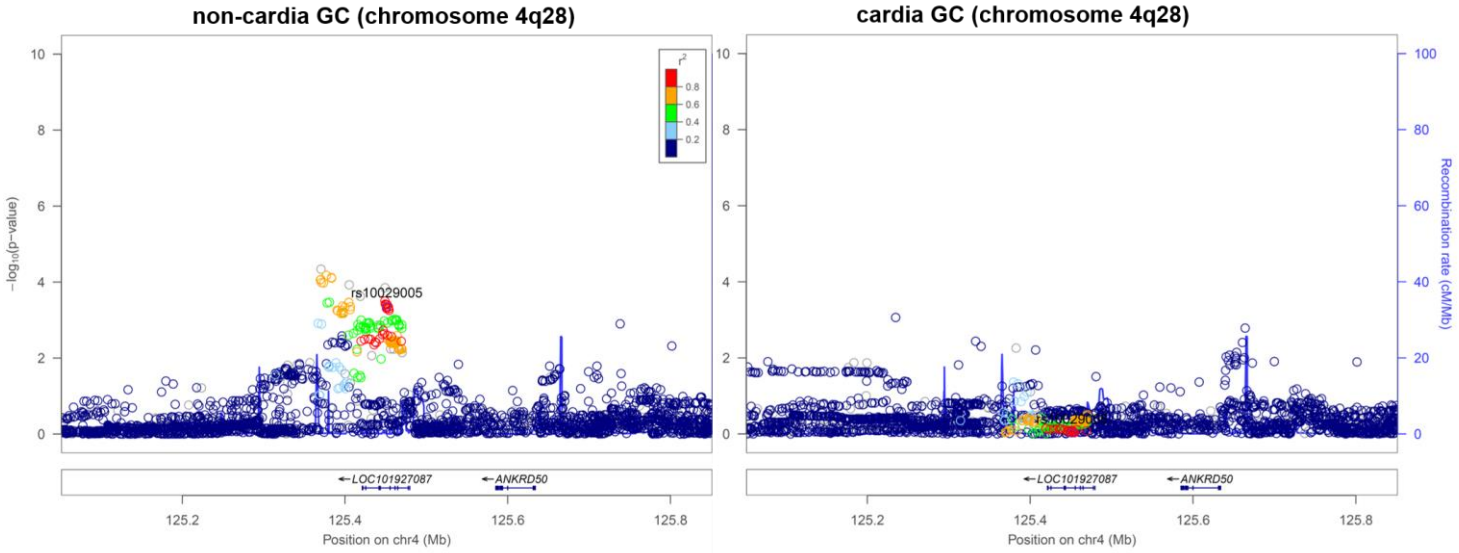


Figure 12: Regional association plots of GC risk locus 4q28 for non-cardia and cardia GC. Associations (-log10(*P*-values)) are shown for SNPs flanking 400 kb on either side of the lead associated SNP (position in hg19). The lead variant is shown in purple. Other markers at each locus are displayed by different colours, indicating different levels of LD (r2) to the lead SNP. Furthermore, annotated genes within each region are shown with arrows indicating their transcription direction.

cis-eQTL for the expression of *ANKRD50* in corpus mucosa (P = 4.10x10^{-17})

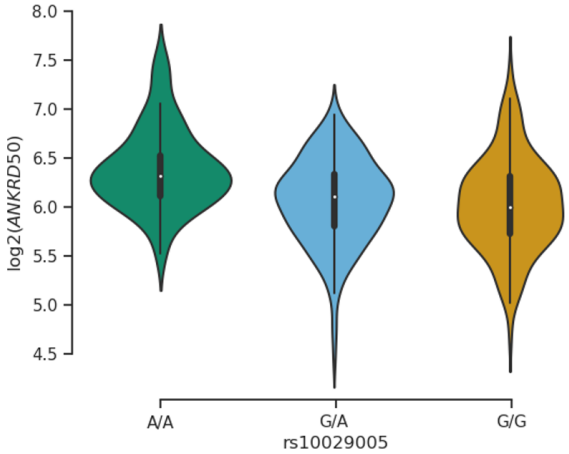


Figure 13: eQTL effects for the expression of *ANKRD50* on chromosome 4q28 in corpus mucosa. Log2 gene expression, error bars for median log2 expression and standard deviation are shown as box plots (y axis) sorted by SNP genotypes (x axis) with the GC risk allele on the left.

4.1.3.2 Replication of Susceptibility Locus 5p13

The second locus that was replicated in the European population is located on chromosome 5q13 with rs6897169 showing the strongest association. Again, the most significant signal was observed in the non-cardia GC sample (OR = 1.18; 95% CI = 1.09-1.28; $P = 9.40 \times 10^{-5}$), while there was no significant association in the cardia GC sample ($P = 0.265$) (Table 4 and Figure 14).

The variant rs6897169 is located in intron 4 of *TTCC3* and app. 45 kb downstream of *PTGER4*. A significant eQTL effect could be identified in the corpus as well as the antrum mucosa datasets, associating risk allele carriers with an increased expression of *PTGER4* (Table 9).

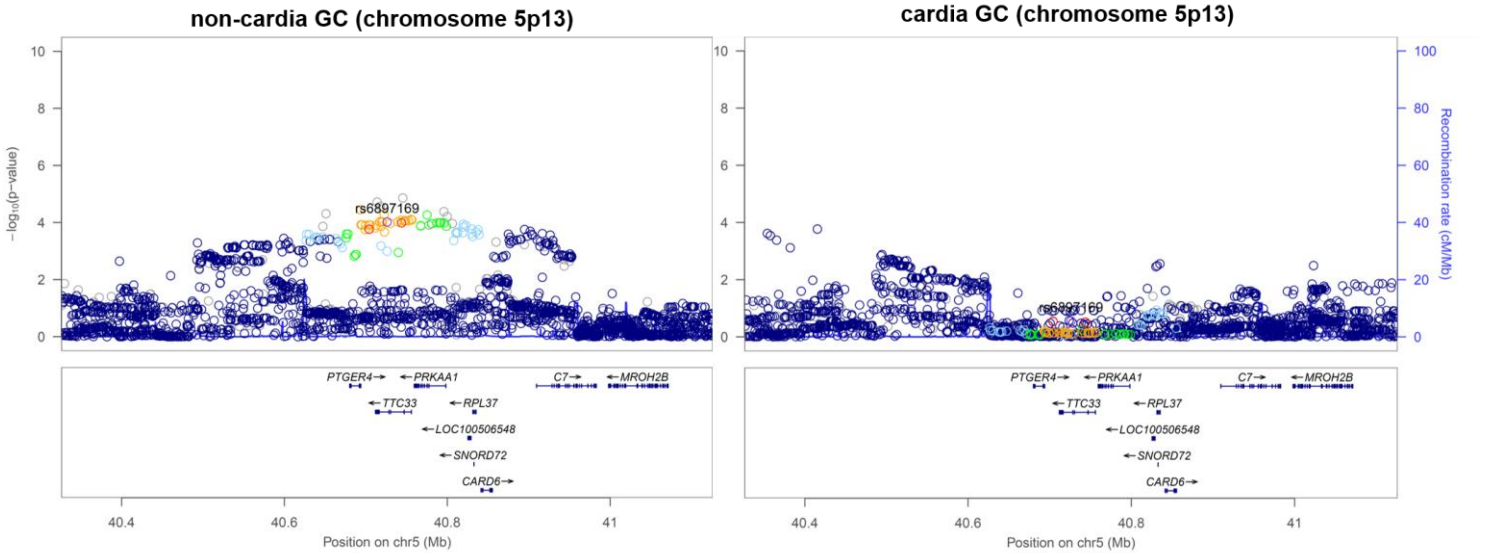


Figure 14: Regional association plots of GC risk locus 5p13 for non-cardia and cardia GC. Associations (-log10(*P*-values)) are shown for SNPs flanking 400 kb on either side of the lead associated SNP (position in hg19). The lead variant is shown in purple. Other markers at each locus are displayed by different colours, indicating different levels of LD (r^2) to the lead SNP. Furthermore, annotated genes within each region are shown with arrows indicating their transcription direction.

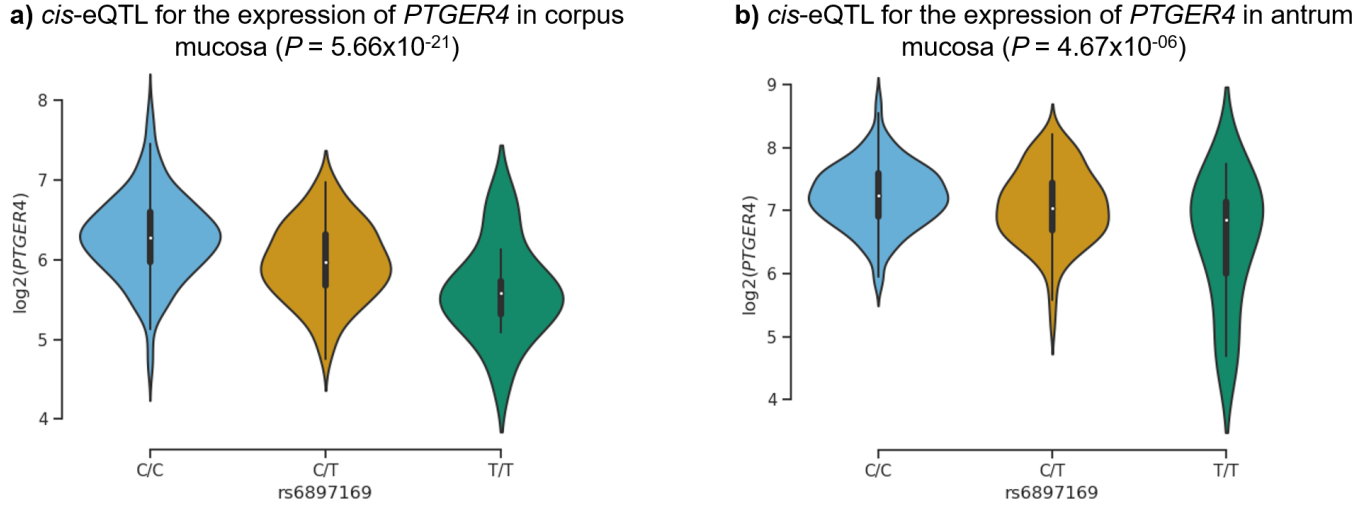


Figure 15: eQTL effects for the expression of *PTGER4* on chromosome 5p13 in (a) corpus and (b) antrum. Log2 gene expression, error bars for median log2 expression and standard deviation are shown as box plots (y axis) sorted by SNP genotypes (x axis) with the GC risk allele on the left.

4.1.3.3 Replication of Susceptibility Locus 9q34

No significant association was identified for the reported Asian lead variant rs7849280 (OR = 0.99; 95% CI = 0.90-1.09; *P* = 0.810). However, there was a strong association for a nearby locus in the European dataset with rs532436 showing the strongest association signal in the entire GC sample (OR = 1.22; 95% CI = 1.07-1.22; *P* = 7.51 x 10^{-5}). rs7849280 is in high LD to rs532436 (r^2=0.76) in the East Asian population, but not in the European population (r^2=0.01).

Concerning subsamples, the signal was prominent in the non-cardia GC (OR = 1.19; 95% CI = 1.10-1.28; *P* = 5.82 x 10^{-6}) and absent in the cardia GC sample (*P* = 0.495) (Figure 16). The strongest signal was seen in the diffuse GC (OR = 1.29; 95% CI = 1.17-1.44; *P* = 8.52 x 10^{-7}) and it was less pronounced in the intestinal GC sample (OR = 1.15; 95% CI = 1.05-1.26; *P* = 3.02 x 10^{-3}) (Table 4).

eQTL effects were observed for the variant in the corpus and antrum mucosa, indicating an upregulation of *ABO* in risk allele carriers (Table 9).

The PheWAS analysis revealed a high association of this locus to a number of phenotypes including haemoglobin concentration and haematocrit percentage. The lead variant was found in a number of other GWAS and could be associated with von Willebrand factor, serum alkaline phosphatase and E-selectin levels (Supplementary Table 10).

As different alleles of *ABO* determine the phenotypic expression of the ABO blood groups, we inferred the blood types from the genetic data according to Groot *et. al.* 2020 [84], and tested the blood groups for GC association (Table 8). This revealed that blood-group O is protective against non-cardia (OR = 0.85; 95% CI = 0.81-0.89; *P* = 2.1 x 10^{-04}) and diffuse GC (OR = 0.76; 95% CI = 0.69-0.82; *P* = 3.0 x 10^{-05}), while blood-group A increases the risk for both GC subtypes (non-cardia GC: OR = 1.28; 95% CI = 1.23-1.32; *P* = 9.3 x 10^{-10}), diffuse GC: OR = 1.31; 95% CI = 1.25-1.37; *P* = 8.0 x 10^{-06}).

Table 8: Association of the ABO blood groups with GC. *P*-values, odds ratios (ORs) and the corresponding 95% confidence intervals (CIs) are shown for the comparison between blood types O versus non-O, as well as A versus non-A.

GC Sample	O versus non-O			A versus non-A		
	***P*-value**	**OR**	**95% CI**	***P*-value**	**OR**	**95% CI**
Entire	0.16	0.95	0.91-0.98	2.63E-05	1.14	1.11-1.17
Cardia	0.07	1.11	1.05-1.17	0.85	1.01	0.95-1.07
Non-Cardia	2.1E-04	0.85	0.81-0.89	9.27E-10	1.28	1.23-1.32
Intestinal	0.01	0.86	0.80-0.91	0.03	1.12	1.06-1.17
Diffuse	3.05E-05	0.76	0.69-0.82	8.01E-06	1.31	1.25-1.37

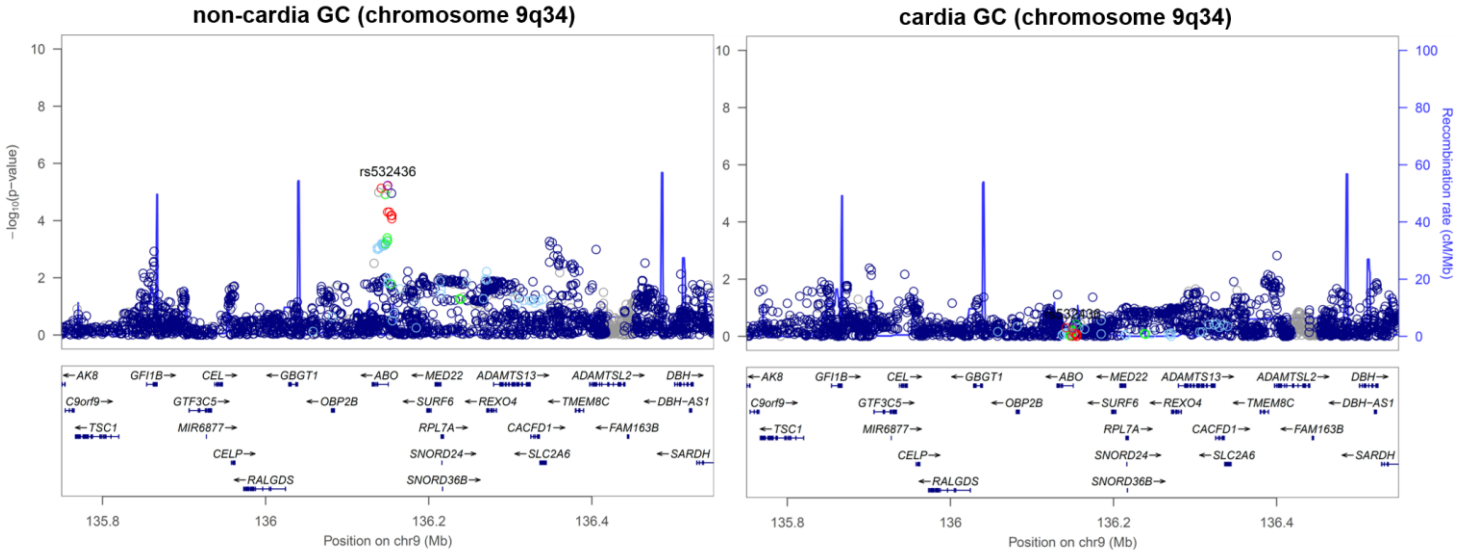


Figure 16: Regional association plots of GC risk locus 9q34 for non-cardia and cardia GC. Associations (-log10(*P*-values)) are shown for SNPs flanking 400 kb on either side of the lead associated SNP (position in hg19). The lead variant is shown in purple. Other markers at each locus are displayed by different colours, indicating different levels of LD (r^2) to the lead SNP. Furthermore, annotated genes within each region are shown with arrows indicating their transcription direction.

a) *cis*-eQTL for the expression of *ABO* in corpus mucosa ($P = 2.94x10^{-17}$)

b) *cis*-eQTL for the expression of *ABO* in antrum mucosa ($P = 4.67x10^{-06}$)

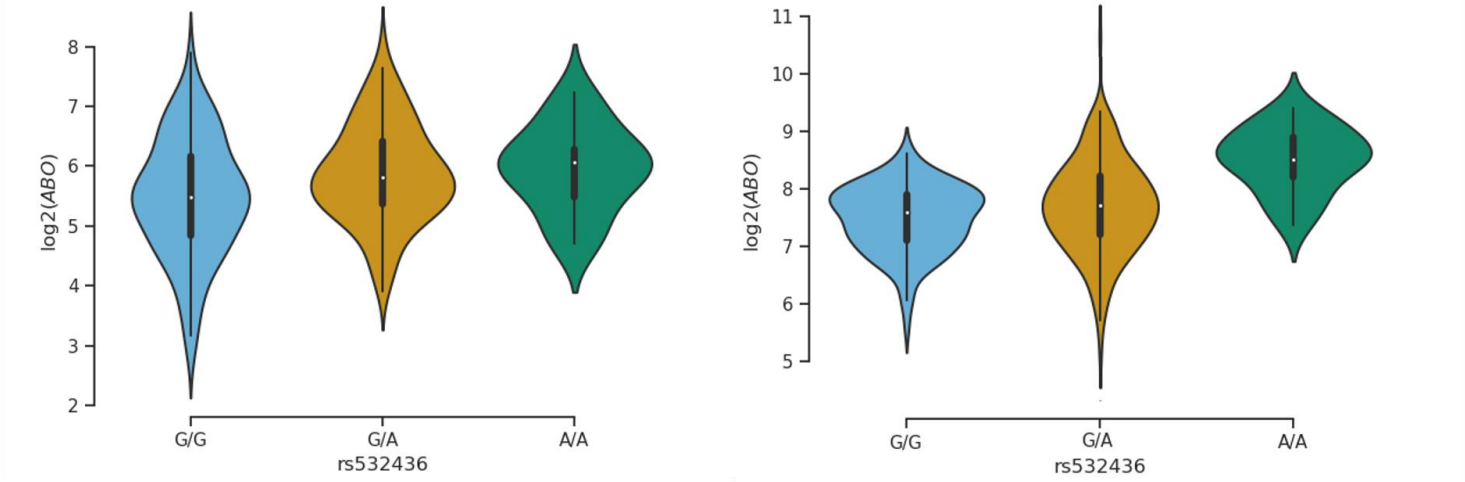


Figure 17: eQTL effects for the expression of *ABO* on chromosome 9q34 in (a) corpus and (b) antrum. Log2 gene expression, error bars for median log2 expression and standard deviation are shown as box plots (y axis) sorted by SNP genotypes (x axis) with the GC risk allele on the right.

4.2 Functional Characterization of Gastric Cancer Risk Variants Using Transcriptome Data

4.2.1 Differential Gene Expression between Stomach Regions

The expression profiles of five different regions of the stomach (cardia, corpus, fundus, antrum, and angulus) were analysed in order to prioritize the tissue of interest for the eQTL and TWAS analysis (Figure 1). The primary aim was to identify uniform expression patterns within the regions to enable a powerful eQTL analysis, as well as to identify the most important regions representing the gastric transcriptome. For this purpose, RNA extracted from biopsies from 11 individuals was analysed using the HumanHT-12v4 Expression BeadChip (Illumina, USA) and the data was analysed as described in chapter 3.3.2. A total of 47 samples were successfully included in the analysis after QC, comprising 9 samples of the cardia, corpus and, antrum, and 10 samples of the fundus and angulus, respectively.

Explorative analysis revealed a clear distinction between the corpus/fundus and the antrum/angulus datasets, however only few differences within the respective datasets were observed. By contrast, the cardia expression profiles did not show a uniform expression signature between individuals, but resembled those from either the corpus/fundus or antrum/angulus signature (see Supplementary Figure 4). This explorative result was confirmed by examination of the differential gene expression (DE) between tissues. A large number of DE genes was found in comparisons between the corpus/fundus and antrum/angulus samples, but not within those groups. Comparisons with cardia showed a considerable overlap with the corpus/fundus, as well as antrum/angulus groups (Supplementary Figure 5).

4.2.2 eQTL Analysis in Corpus and Antrum

The expression profile analysis of all five stomach regions revealed that the corpus and antrum areas cover most of the transcriptomic variation across the human stomach (see above, chapter 4.2.1 and Supplementary Figure 5). Thus, we selected samples from a total of 434 individuals of which RNA-seq data for corpus (N=410) or antrum (N=381) samples from the antrum were generated. Explorative analysis revealed two well separated expression profiles according to stomach location (Supplementary Figure 6). After QC, 362 samples from the corpus and 342 samples from the antrum were included in the eQTL analysis.

A total of 4,229 genes with at least one significant *cis*-eQTL were identified in the corpus sample as compared to 5,706 genes in the antrum samples. Of those, 3,179 overlapped between both samples, while 2,526 were specific in the antrum and 1,049 in the corpus dataset. For five risk loci identified in the GC GWAS, an overlap with *cis*-eQTLs was observed (Table 9).

Table 9: Overview of significant *cis*-eQTLs identified in the in-house antrum and corpus datasets, overlapping with the lead genome-wide or replicated risk variants associated with GC.

GWAS locus	rsID	Tissue	Gene Symbol	*P*-value	Effect (beta)
1q22	rs760077	Antrum	*THBS3*	9.24E-06	0.20
	rs760077	Antrum	*MUC1*	6.94E-07	0.11
	rs760077	Corpus	*MUC1*	2.80E-08	0.13
8q24	rs2920292	Antrum	*THEM6*	7.36E-12	0.22
	rs2920292	Antrum	*PSCA*	1.28E-104	0.92
	rs2920292	Corpus	*LY6K*	4.80E-07	0.33
	rs2920292	Corpus	*PSCA*	3.64E-114	0.97
	rs2920292	Corpus	*LYNX1*	4.77E-21	-0.51
	rs2920292	Corpus	*THEM6*	1.66E-09	0.23
4q28	rs10029005	Corpus	*ANKRD50*	4.10E-17	0.40
5q13	rs6897169	Corpus	*PTGER4*	5.66E-21	-0.53
	rs6897169	Antrum	*PTGER4*	4.67E-06	-0.28
9q34	rs532436	Antrum	*ABO*	2.94E-17	0.51
	rs532436	Antrum	*SURF1*	8.90E-05	-0.16
	rs532436	Corpus	*ABO*	4.03E-07	0.34

4.2.3 Transcriptome-wide Association Analysis (TWAS)

The TWAS using the mucosal corpus and antrum expression dataset revealed transcriptome-wide associations at three loci (Table 10). The Manhattan plots for the TWAS in the GC subsamples are presented in Supplementary Figure 7 and Supplementary Figure 8. The loci 1q22 and 8q24 were already identified to be genome-wide significant associated on single marker or eQTL level (see chapter 4.1.1). In addition to these two loci, one additional locus on 6q24 in the non-cardia GC subsample for the corpus expression dataset was identified. However, the association level of the locus was not substantially affected by conditioning on the identified gene (Supplementary Figure 11).

Table 10: Genes with expression models in corpus and antrum that showed transcriptome-wide significant GC associations. In total, three risk loci (1q22, 6p24, 8q24) and GC types (non-cardia, diffuse, intestinal) as well as seven genes were implicated. The number SNPs with non-zero weights included in the model (NWGT SNPs) in the best predicting expression model is shown. In addition, TWAS Z scores indicating the effect of GC expression association (downregulated/upregulated) and corresponding TWAS *P*-values are shown.

Chromosomal region	Tissue	GC type	Gene	NWGT SNPs	TWAS Z score	TWAS *P*-value
1q22	Corpus	non-cardia	*MUC1*	2	7.26	3.83E-13
		diffuse	*MUC1*	2	6.33	2.40E-10
	Antrum	non-cardia	*MUC1*	1	4.55	5.26E-06
			THBS3	401	6.52	7.06E-11
		diffuse	*MUC1*	1	4.41	1.01E-05
			THBS3	401	6.22	4.97E-10
6p24	Corpus	non-cardia	*TMEM14C*	50	4.54	5.52E-06
8q24	Corpus	non-cardia	*PSCA*	30	11.46	2.14E-30
			LY6K	1	8.73	2.46E-18
			THEM6	1	9.31	1.23E-20
			LYNX1	26	-7.23	4.85E-13
		diffuse	*PSCA*	30	8.14	4.11E-16
			LY6K	1	6.31	2.84E-10
			THEM6	1	6.34	2.28E-10
			LYNX1	26	-4.65	3.35E-06
			PSCA	30	5.96	2.60E-09
		intestinal	*LY6K*	1	5.11	3.19E-07
			THEM6	1	5.24	1.62E-07
	Antrum	non-cardia	*PSCA*	46	11.34	8.42E-30
			LY6K	26	6.75	1.46E-11
			THEM6	1	9.20	3.48E-20
			LYNX1	1	-6.57	5.17E-11
		diffuse	*PSCA*	46	8.04	8.84E-16
			LY6K	26	5.90	3.69E-09
			THEM6	1	6.26	3.87E-10
			LYNX1	1	-4.29	1.81E-05
		intestinal	*PSCA*	46	5.81	6.32E-09
			THEM6	1	5.17	2.39E-07

4.3 Gastric Cancer Heritability and Correlation with other Traits

4.3.1 LD Score Regression Analysis

LD score regression was performed to estimate the SNP-based heritability of GC and the genetic correlation between GC and related phenotypes.

The SNP-based heritability of GC was estimated to be 8.48 ± 3.12% standard deviation (SD).

In the genetic correlation analysis, we examined traits belonging to five phenotype-categories that represent known risk factors for GC development (smoking, reflux, obesity, education, alcohol intake) [30,32,33]. An experiment-wide significant genetic correlation with GC was identified for three obesity-related traits, one smoking- and one alcohol intake-related trait. The correlations are presented in Figure 18 and are listed in Supplementary Table 12. Body mass index (r_g = 0.303, P = 6.0 x 10^{-4}), hip circumference (r_g = 0.269, P = 2.3 x 10^{-3}) and body weight (r_g = 0.262, P = 2.4 x 10^{-3}) showed a Bonferroni-corrected positive GC correlation along with pack years of adult smoking (r_g = 0.352, P = 2.0 x 10^{-3}) and alcohol intake 10 years previously (r_g = 0.361, P = 2.0 x 10^{-3}). Furthermore, all education-/employment-related traits serving as proxy for socio-economic status were nominal significant and negatively GC-correlated.

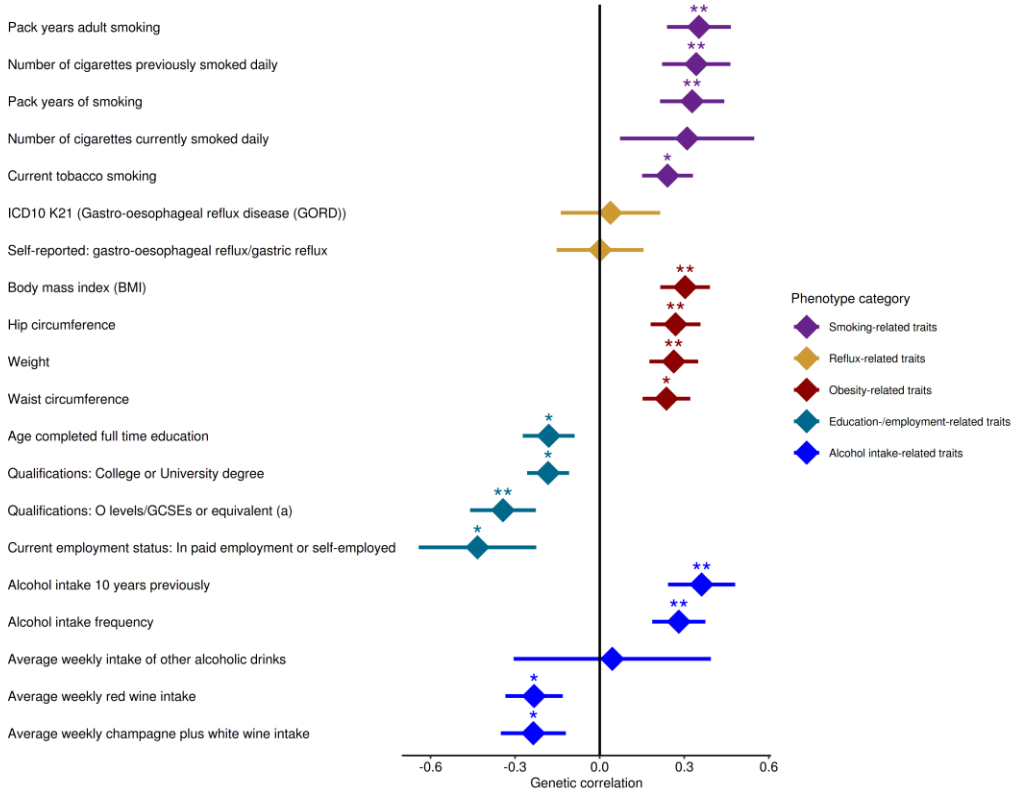


Figure 18: Genetic correlations determined with LD score regression between GC and 20 traits belonging to five phenotype-categories that represent risk factors for GC development. For each trait, the genetic correlation (dot) and the standard deviation (line) are shown. The significance levels of the genetic correlation are indicated by asterisks (* P < 0.05, ** P < 0.0025).

4.3.2 Polygenic Risk Score Analysis

Polygenic risk score (PRS) analysis was carried out to examine the genetic relation between cardia GC and OAC. PRS was calculated in the two in-house discovery GWAS datasets with OAC as well as OAC and its precursor lesion BO [94]. The cardia GC GWAS was used as target dataset to examine the genetic relation to OAC and OAC/BO. In addition, the GWAS of the entire GC, non-cardia GC and cardia-versus-non-cardia GC was used as target sets to determine whether associations between cardia GC and OAC as well as OAC/BO are specific. Supplementary Table 13 and Supplementary Table 14 show the *P*-value thresholds, the number of SNPs included in each PRS and the observed associations. Figure 19 gives a graphical overview of the associations identified using the OAC as a discovery dataset. Highly significant associations were identified between cardia GC and PRS derived from OAC ($P_{threshold}$ = 0.001, $P_{association}$ = 2.37 x 10^{-08}) and OAC/BO ($P_{threshold}$ = 0.2, $P_{association}$ = 2.79 x 10^{-17}). In contrast, no associations were present between other GC case-control datasets and PRS derived from OAC or OAC/BO. Accordingly, the case to case comparison (cardia-versus-non-cardia GC) revealed significant associations to PRS derived from OAC ($P_{threshold}$ = 0.5, $P_{association}$ = 2.18 x 10^{-09}) and OAC/BO ($P_{threshold}$ = 0.2, $P_{association}$ = 4.33 x 10^{-07}).

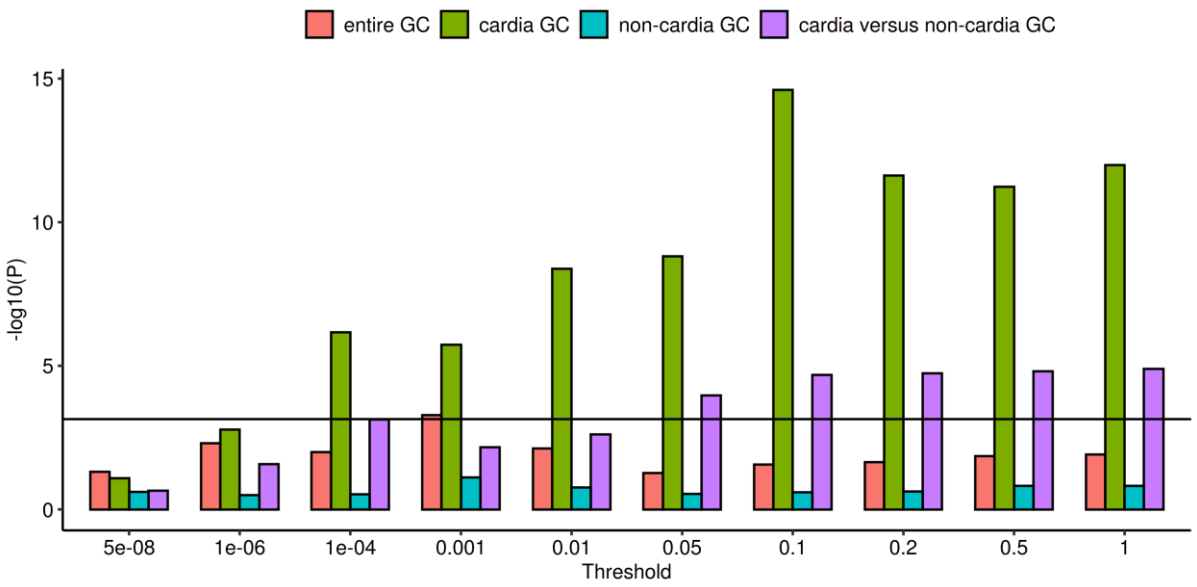


Figure 19: Polygenic risk score associations for OAC in the target GC subtypes. The association values in dependence of the different significance thresholds for PRS variant selection are given. The horizontal black line represents the Bonferroni correction threshold.

4.3.3 Cross Trait Meta-Analysis

Based on the shared polygenic risk architecture of cardia GC and OAC/BO, a meta-analysis combining the GWAS datasets (1,291 cardia GC cases, 10,279 OAC/BO cases, 27,326 controls) was performed. This cross trait meta-analysis revealed 17 genome-wide significant associated loci for oesophago-gastric adenocarcinoma. Two of these loci have not been described before [94]. Figure 20 shows the corresponding Manhattan plot an Supplementary Table 15 lists all significant associated loci.

Of the newly identified risk loci, rs1817002 near *HNF4G* on chromosome 8q21 showed an association with $P = 4.10 \times 10^{-08}$ (OR = 1.11; 95% CI = 1.06-1.14). The second new locus is located on chromosome 15q26 near *SPATA8* and *NR2F2*. Here, rs234506 showed disease association with $P = 1.56 \times 10^{-09}$ (OR = 1.12; 95% CI = 1.07-1.16).

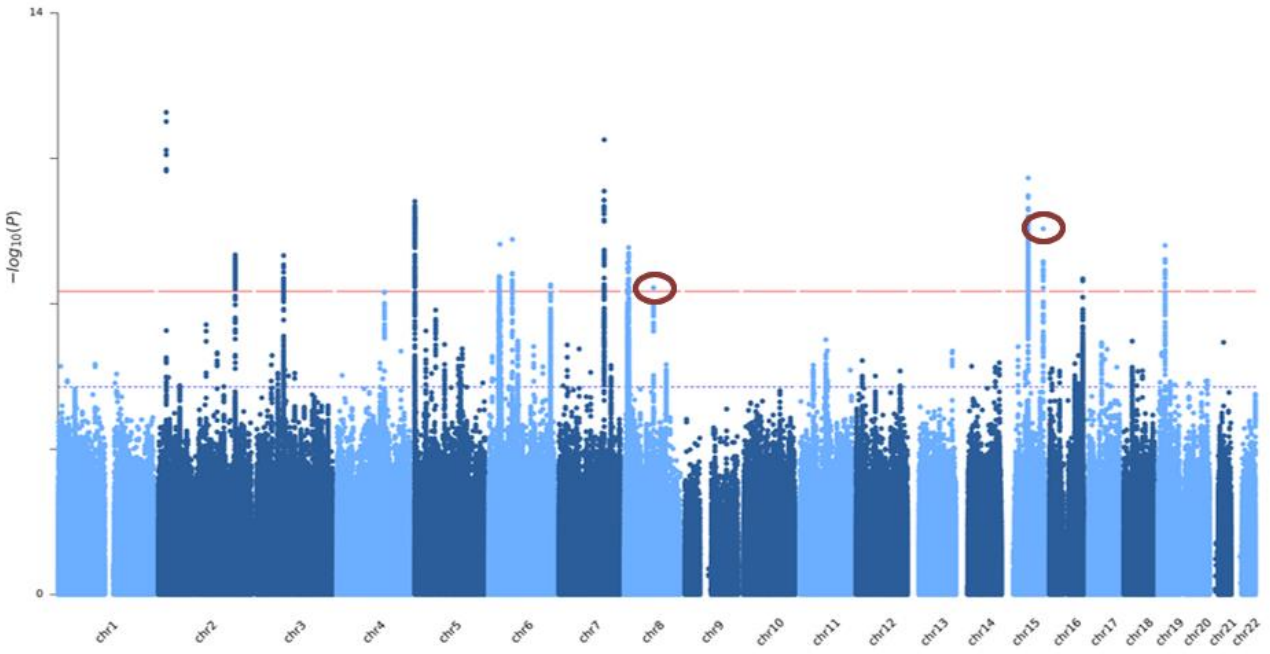


Figure 20: GWAS Manhattan plot from the GWAS of the combined cardia GC/OAC/BE samples are. All SNPs have been plotted against their chromosomal positions (X axis) and the observed -log10(*P*-values) in the GWAS (Y axis). The horizontal red line indicates the threshold of genome-wide significant association ($P < 5 \times 10^{-8}$) while the dashed blue line represents the suggestive associations ($P < 5 \times 10^{-5}$). The red circles indicate the new genome-wide significant signals, not found in the OAC/BE GWAS before.

5 Discussion

Gastric adenocarcinoma summarises a group of neoplasms with a complex aetiology. So far, the focus for the identification of genetic risk factors has been in the Asian population, mainly examining tumours from the non-cardia region without major consideration of the tumour histopathology. In the context of this doctoral thesis, we collected the to-date largest European sample of GC cases, building the basis for the identification of common genetic risk factors contributing GC development by carrying out a GWAS. Furthermore, we collected the to-date largest gastric gene expression dataset including samples from different stomach gastric regions, which allowed in-depth functional characterisation of the identified GWAS risk loci.

In the course of the discussion, the transcriptome dataset will be discussed first, followed by the functional interpretation of the identified GWAS loci. Then, all GWAS analyses beyond the single marker level and the cross phenotype analysis of cardia GC and OAC will be presented.

5.1 The Transcriptomic Landscape of the Human Stomach

In general, most GWAS risk loci do not confer to the risk of a disease by directly altering the coding sequence and thereby affecting the function of a protein, but by having a gene regulatory impact [97]. Although the underlying functional mechanisms can be diverse, the phenotypic effect is oftentimes revealed by a change in gene expression. For this reason, TWAS and eQTL studies proved to be a powerful tool to assign functional relevance to non-coding risk loci. However, as transcriptomes are highly tissue and context specific, unravelling the risk conferring effect can be challenging [97]. In the context of GC, the tumours arise from the epithelial tissue lining the stomach, thereby representing the primary tissue of interest. However, due to the different anatomical parts of the stomach and the acidic milieu, gastric transcriptome analysis provides specific challenges. So far, the only large eQTL dataset from human stomach tissue was published by the GTEx consortium comprising samples from 324 individuals [64]. However, the GTEx dataset has two major limitations. First, the GTEx consortium collects tissue samples *post mortem*. Due to the highly acidic milieu, autolysis progresses rapidly as soon the metabolic processes break down, which ensure the maintenance the mucosa, such as the production of alkaline mucus. Such autolytic processes may adversely affect RNA quality and confound transcriptome profiles [98]. In case of GTEx, a substantial number of samples was collected after 10 hours or more ischemic time [64]. Thus, possible adverse effects cannot be ruled out and are difficult to quantify. A second limitation of the GTEx data is the fact that tissue samples were collected from only one gastric region, namely corpus [64]. As there are

other gastric regions with substantial histological differences compared to corpus, data from this location does not reflect the entire stomach transcriptional landscape. To account for these limitations, we generated an in-house dataset derived from fresh tissue samples from healthy individuals to examine the transcriptional landscape of the human stomach and to perform a TWAS and eQTL analysis in the identified GWAS regions.

For the prioritization of stomach regions, we collected samples from the cardia, fundus, corpus, antrum, and angulus and examined the differential gene expression in each region exploratively and qualitatively. Aside from expected gender specific differences, most variability in gene expression was explained by the different gastric regions, showing a distinct grouping in PCA and unsupervised clustering (Supplementary Figure 6). In line with the cellular composition of the glandular epithelium, which mainly differentiates between oxyntic glands (corpus/fundus) and pyloric glands (angulus/antrum) [16], region specific changes in gene expression were detected (Supplementary Figure 5). An enrichment of pathways related to protein digestion, gastric acid secretion and vitamin B12 absorption was observed in the corpus, representing important physiological functions of this gastric region [16]. In the gastric antrum, primarily immune response related pathways were enriched, which is in line with its function in prolonged food storage leading to mucosal irritation [99] (Supplementary Table 11). Transcriptome profiles of the cardia, however, showed inter-individual divergent expression patterns. Some samples were indistinguishable from the corpus/fundus and others from the antrum/angulus region (Supplementary Figure 5). One reason for this observation can be the selection of study participants, whose indication for gastroscopy often was related to GERD related symptoms. Gastric acid induced irritation of the cardia mucosa due to GERD may explain the antrum like inflammatory signature. In addition, identification and taking biopsies from the cardia can be difficult during gastroscopy, which may lead to miss-sampling of the adjacent oesophagus, corpus or fundus region. Due to the inconsistend expression pattern and the small number of DE genes as compared to the other regions (Supplementary Figure 5), the cardia region was not taken into consideration for the main project. Instead, the corpus and antrum region were prioritized, showing region specific and homogenous expression patterns, which represent the expression profile of the entire stomach.

In the following, the worldwide largest and most comprehensive stomach expression dataset was collected, building the basis for the TWAS and eQTL analysis. These datasets enabled to characterise GC GWAS loci, which will be discussed in the following chapters.

5.2 Gastric Cancer GWAS

For the GC-GWAS meta-analysis, 5,815 GC cases and 10,999 ethnically matched controls across 10 European subsamples were included.

On the single marker level, six loci reached genome-wide significance and additional three risk loci, previously described in the Asian population, could be replicated. Furthermore, their role in specific GC subtypes concerning tumour location and histopathological type could be elucidated, demonstrating that the known risk loci on chromosome 1q22, 4q28, 5p13, 8q24, and 9q34 [5,6,8,10,12,13] contribute almost exclusively to non-cardia GC risk (Table 4). In addition, the data show that the risk loci on chromosome 8q24 and 9q34 confer to substantially higher risk for diffuse GC than intestinal GC. In the following, each locus will be discussed in detail.

5.2.1 Genome-Wide Significant Gastric Cancer Risk Loci

5.2.1.1 Genome-Wide Significant Risk Locus on Chromosome 1q22

The identified locus on chromosome 1q22 (rs760077) has previously been shown to confer to GC risk in the European population [40]. Although, an association to cardia GC has been described in the Chinese population [4], we identified a tumour location specific signal only for non-cardia GC (OR = 1.31; 95% CI = 1.23-1.40; $P = 5.12 \times 10^{-17}$) versus cardia GC ($P = 0.45$). Moreover, the signal was more pronounced for the diffuse GC tumour type (OR = 1.35; 95% CI = 1.23-1.48; $P = 7.41 \times 10^{-11}$) versus intestinal GC ($P = 1.78 \times 10^{-07}$).

The locus on chromosome 1q22 is located near *MUC1,* which has been attributed concurringly as the risk conferring gene at the locus [10,12]. Functional studies suggest that associated GC risk variants lead to a change in tandem repeats in exon 2 [40,100] and a change in a splice acceptor site, having an impact on the promotor activity and the expressed isoform of *MUC1*, which leads to an increased expression of *MUC1* on the risk background [101]. This observation could be partly confirmed by our TWAS and eQTL study indicating an upregulation of *MUC1* for risk allele carriers (Figure 5).

Mucins comprise a family of O-glycolsylated proteins that play an essential role in building protective mucous barriers on epithelial surfaces. *MUC1*, which is highly expressed in the human stomach, encodes for Mucin-1. It belongs to the membrane bound mucins, which is cleaved into an alpha and a beta subunit, forming a heterodimeric complex [102]. Although primarily forming a protective mucous layer, Mucin-1 is considered to confer to the development of cancer via a variety of

oncogenic effects during malignant transformation, exerting oncogenic effects via a variety of roles in cell signalling and cell adhesion processes. In line, overexpression and aberrant glycosylation are described for many cancer entities on the somatic level [103]. Interestingly, it has also been shown that Mucin-1 can repress the tumour suppressor function of Cadherin-1, which is encoded by *CDH1*, the causal gene for hereditary diffuse gastric cancer (HDGC), representing the most common monogenic GC form [104]. This provides a possible link to the fact, that we and others see a pronounced effect for the risk of the diffuse GC subtype [101].

Beside the upregulation of *MUC1*, also *THSB3* was found to be upregulated in the gastric mucosa on the background of the GWAS risk genotype (Table 10). *THSB3* encodes an adhesive glycoprotein that mediates cell-to-cell interactions and cell-to-matrix interactions. It has been reported to play a role in skeletal maturation, but otherwise little is known about its function [105]. A risk conferring role in the context of GC cannot be excluded.

Beside for GC, the risk locus on chromosome 1q22 has also been shown to have pleiotropic effects on a variety of cardiometabolic, renal and haematological traits, consistently prioritizing *MUC1* as the causal gene (Supplementary Table 6). Thus, apart from a direct oncogenic effect, there may be other mechanisms conferring to GC development, which are only sparsely discussed and lack a direct functional link so far [106].

5.2.1.2 Genome-Wide Significant Risk Locus on Chromosome 2q23

The second genome-wide significant locus was identified on chromosome 2q23 (rs11677924), showing a specific signal in the intestinal GC subsample (OR = 1.34; 95% CI = 1.21-1.49; $P = 2.04 \times 10^{-8}$). The locus has never been described in the context of GC or other phenotypes before and it is the first locus described so far, which specifically confers to the risk to intestinal GC.

The lead variant is located in intron 4 of the gene *ALK*. *ALK* encodes a receptor tyrosine kinase. Activating mutations in the germline are causal for hereditary neuroblastoma [107]. In addition, somatic rearrangements and fusion genes involving *ALK* were identified to drive a variety of cancer entities, such as breast, colorectal, thyroid and non-small cell lung cancer [108,109] and is a valuable therapy target [110]. In GC, *ALK* fusions are reported, but occur rarely [111]. The underlying pathomechanisms conferring to the GC risk at the identified locus are unclear. A possible explanation is a change of enzymatic activity of the tyrosine kinase or a priming for gene fusions. However, this needs to be evaluated in future studies.

5.2.1.3 Genome-Wide Significant Risk Locus on Chromosome 8q24

The risk locus at chromosome 8q24 (rs2920293) was the most significantly associated risk locus in the whole GC GWAS and already has been reported for GC before [6,10,12,13]. The signal was specific for non-cardia GC (OR = 1.46; 95% CI = 1.36-1.55; $P = 1.80 \times 10^{-30}$) versus cardia GC ($P = 0.114$) and more pronounced for the diffuse GC type (OR = 1.46; 95% CI = 1.33-1.60; $P = 8.10 \times 10^{-17}$) versus the intestinal type (OR = 1.27; 95% CI = 1.17-1.37; $P = 4.05 \times 10^{-09}$), which is in line with previous studies [13].

The lead variants lie in close proximity to the gene *PSCA*, which has been prioritized as risk conferring gene in previous studies [6,10,12,13]. Our gastric mucosa transcriptome data supports this presumption. The most significant findings in the eQTL and TWAS analyses indicate a strong upregulation of *PSCA* in normal gastric mucosa for GC risk allele carriers (Figure 7 and Table 10). This is in line with studies, that suggests that a risk allele in high LD (rs2294008, T allele) creates a novel translation start site, thereby extending the encoded protein by 9 amino acids, and at the same time leads to an upregulation of the transcriptional activity of the gene in stomach mucosa and a change in subcellular localization of the encoded protein [13,112].

PSCA encodes Prostate stem cell antigen, a small 123-amino-acid glycosylphosphatidylinositol-anchored cell surface protein, which was originally identified and isolated as a tumour antigen over-expressed in prostate cancer [113]. In normal tissues, *PSCA* is prominently expressed in the epithelial cells of the prostate, urinary bladder, kidney, skin, oesophagus, stomach and placenta. The physiological functions of *PSCA* are not fully elucidated and discussed controversially. Overall, an involvement of the protein in several cell signalling pathways promoting cell renewal, proliferation as well as triggering tumour specific immunity are suspected. This mediates oncogenic as well as tumour suppressor effects depending on the cellular context and the tissue type investigated [113,114]. Interestingly, the identified risk locus has an inverse effect for the risk of the development of duodenal ulcers (Supplementary Table 7) [114]. It has also been shown, that the risk locus and *PSCA* expression are associated to the progression from mild to severe atrophic gastritis [115,116]. This led to the hypothesis, that a lower expression of *PSCA* accounts for a suppressed epithelial growth upon tissue damage, which would prime for duodenal ulcer formation. *Vice versa*, an increased *PSCA* expression, as shown in risk allele carriers, may promote epithelial proliferation, priming for progressive gastritis and GC development [116]. However, so far none of the proposed functional hypothesis have been proved sufficiently.

Aside from *PSCA*, also *LY6K*, *THEM6* and *LYNX1* showed transcriptome-wide significant association to GC (Table 10). However, after conditioning for *PSCA* expression, no significant associations remained (Supplementary Figure 9). Still, a causal influence of these genes cannot be ruled out, but *PSCA* is the most promising candidate gene.

5.2.1.4 Genome-Wide Significant Risk Locus on Chromosome 17q12

On chromosome 17q12 (rs17138478) a genome-wide significant locus could be identified specifically conferring to the risk of intestinal GC (OR = 1.44; 95% CI = 1.27-1.64; $P = 1.83 \times 10^{-8}$), which has not been described in the context of GC before.

The lead risk variant is located in intron 4 of *HNF1B,* which represents an interesting candidate gene. It encodes for the hepatocyte nuclear factor 1-beta, which is involved in the embryonic and fetal development of the liver, kidney, pancreas and biliary system and is also highly expressed in tubule forming epithelia in adult tissue [117]. As such, the locus has been reported, to confer to the risk of developing cholelithiasis and cholecystitis [118,119].

Cholelithiasis and gallbladder stones have previously been identified as a risk factor for GC, postulating a duodenogastric bile reflux induced gastritis as a possible causal relationship [120,121]. Our data support this hypothesis, as the intestinal tumour type usually arises on the background of a gastritis [28], for which the identified risk locus specifically confers to.

5.2.2 Sex-Specific Gastric Cancer Risk Loci

For the first time, the present GWAS found two sex specific GC risk loci, each conferring to the risk of GC development in females.

5.2.2.1 Female Specific Risk Locus on Chromosome 1q31

At the locus on chromosome 1p31, rs2590943 showed the strongest association in the female cardia sample (OR = 1.93; 95% CI = 1.56-2.38, $P = 1.21 \times 10^{-9}$), while being absent in the female non-cardia sample ($P = 0.330$), as well as in the male cardia sample ($P = 0.675$).

The locus has been previously described to contribute to the risk of increased BMI and obesity [122], GERD [123] and psychiatric disorders, such as depression [124]. As obesity and GERD are known risk factors cardia GC [35], there may be an overlap in the causal mechanism for this variant between the different phenotypes.

Most association studies indicate *NEGR1*, encoding for neuronal growth regulator 1, as the risk conferring gene at the locus. It encodes a neural cell adhesion and growth protein, which has been shown to be significantly upregulated in the hypothalamus and blood of depression patients [125]. The biological mechanisms contributing to the specific risk of disease development are largely unknown. However, in several mouse models a relationship between loss of *Negr1* expression and priming for lipid synthesis and accumulation could be shown [122]. Stratification for sex showed a higher tolerance to a high fat diet without excessive weight gain or development of glucose intolerance in male mice in comparison to females [122]. In humans, the described risk locus was shown to contribute substantially more to the risk of obesity and related traits in females than in males [126]. However, when stratifying for sex in large GWAS datasets on BMI and GERD, only minor effect sizes and no gender specific effects were seen (Table 6).

In summary, it is inconclusive whether there is a female specific risk effect at this locus, or if these results are biased, for example due to an overrepresentation of obese females in the cardia GC sample. In general, such biases have recently been discussed to contribute to sex specific findings in GWAS [127].

5.2.2.2 Female Specific Risk Locus on Chromosome 10p15

The other sex-specific genome-wide significant signal was found on chromosome 10p15 with rs1547179 showing a female non-cardia specific association signal (OR = 1.38; 95% CI = 1.23-1.54, $P = 2.18 \times 10^{-8}$).

As there were no further PheWAS or TWAS findings, functional interpretation of the risk locus is lacking. The most promising gene at the locus is *KLF6*, encoding Kruppel-like factor 6, a zinc finger transcription factor with tumour suppressor functions, which has been described to be mutated in several cancer types including GC [128]. However, *KLF6* is located app. 600 kb upstream the risk variant and no data elucidating a sex-specific influence of the locus or *KLF6* are available.

5.2.3 Replication of Asian Gastric Cancer Risk Loci

Apart from the above discussed novel genome-wide significant loci, the present thesis examined whether already identified genome-wide significant GC risk loci in the Asian population could be replicated in the European sample. In addition to the already described loci on chromosomes 1q22 and 8q24, three further loci were successfully replicated (Table 7).

5.2.3.1 Replication of Susceptibility Locus 4q28

At the GC risk locus on chromosome 4q28 the variant rs10029005 revealed a non-cardia specific effect in the European non-cardia sample (OR = 1.12; 95% CI = 1.05-1.19; $P = 4.37 \times 10^{-4}$).

The variant was identified in non-cardia GC cases in the Chinese population, however, no assumptions on the functional impact of the locus were reported [11]. We observed an eQTL effect with an upregulation of the gene *ANKRD50* in risk allele carriers in our gastric mucosa transcriptome dataset (Table 9). *ANKRD50* encodes for the Ankyrin Repeat Domain 50, which is involved in endosome-to-plasma membrane trafficking and recycling of cargo protein [129]. *ANKRD50* was described in the context of anorexia nervosa [130]. However, functional descriptions in the context of GC are scarce and require further follow-up.

5.2.3.2 Replication of Susceptibility Locus 5p13

The second locus that was replicated in the European population is located on chromosome 5q13 with rs6897169 showing the strongest association. Again, the most significant signal was observed in the non-cardia sample (OR = 1.18; 95% CI = 1.09-1.28; $P = 9.40 \times 10^{-5}$), while there was no significant association in the cardia sample (P = 0.265) (Table 4).

Based on findings in East Asians it has been hypothesized that rs59133000 is the causal GC risk SNP at this locus by conferring to a downregulation of *PRKAA1,* which encodes a AMP-activated protein kinase [11]. However, rs6897169, which is the lead associated GC variant in the Asian population, is in nearly perfect LD to rs59133000 in East Asians ($r^2 = 0.96$), shows only moderate LD to rs59133000 in Europeans ($r^2 = 0.60$). For rs59133000, no significant eQTL effects could be observed, however, for rs6897169, an increased *PTGER4*-expression in risk allele carriers was observed, representing another plausible GC pathomechanism at this locus. *PTGER4* encodes the prostaglandin E2 (PGE2) receptor 4, which mediates cellular responses to PGE2. It has been previously shown that PGE2 is important for the inflammatory microenvironment in tumours and maintenance of gastric stemness [131].

5.2.3.3 Replication of Susceptibility Locus 9q34

A third locus was replicated on chromosome 9q34 showing a non-cardia specific effect (rs532436; OR = 1.19; 95% CI = 1.07-1.22; $P = 7.51 \times 10^{-6}$) being more pronounced in the diffuse subtype (OR = 1.29; 95% CI = 1.17-1.44; $P = 8.52 \times 10^{-7}$).

The lead variant is located in an intronic region of the gene *ABO*, for which a significant upregulation in the gastric mucosa transcriptome of risk allele carriers was identified (Figure 17), a finding also reported in the Asian population [12].

ABO encodes a glycosyltransferase whose exact physiological role still remains to be clarified. However, by catalysing the transfer of carbohydrates to the so-called substance H or H antigen, a polysaccharide presented on red blood cells, it determines the ABO blood type. The ABO blood type was the first, and up to today, most important blood classification system discovered in humans, differentiating between the blood types A, B, AB and O. It is utilized, amongst others, to predict the compatibility for blood transfusions. On a genetic level, the different blood groups are determined essentially by three different alleles altering the activity of the encoded glycosyltransferase. The alleles determining type A and type B differ in four amino acid substitutions, which alter the protein function, phenotypically resulting in different modifications of H antigen. By contrast, blood type O is genetically determined by a truncating variant, which leads to an inactivation of the encoded protein, leaving the H antigen unmodified. The combination of the described alleles thereby determines the phenotypic appearance of the H antigen and the type of corresponding antibodies present in an individual. Confronted with incompatible blood types, immune complexes may form and cause agglutination, which may lead to adverse reactions for example after blood transfusions [132].

Apart from GC, variants at the *ABO* locus have been shown to be associated with a wide spectrum of phenotypes, including duodenal ulcer [114], pancreatic cancer [133] and COVID-19 [134]. These studies also described associations of the respective phenotype to the actual ABO blood type. Also for GC the ABO blood types are described to influence the risk of tumour development, with non-A blood types being protective versus non-O types increasing the risk for GC, especially for the non-cardia and diffuse tumour type [135,136]. Based on these reports, we inferred the ABO blood types from the genotype data and confirmed non-A blood types to decrease and non-O blood types to increase risk for GC development, especially for non-cardia and diffuse type GC (Table 8).

The observed upregulation of *ABO* in the gastric mucosa in risk allele carriers may also confirm an involvement of blood types as modulators of GC risk. As the O-allele encodes a truncated protein, the respective transcript may be subject of nonsense-mediated decay and thereby appear downregulated in comparison to type A and B transcripts. However, no functional studies are available that support this hypothesis.

Still, it remains unclear which underlying mechanism may confer to the increased GC risk and several hypothesis have been proposed [135]. An unfavourable inflammatory response in the context of HP infection is one plausible theory. It could be shown that individuals with blood type A were more susceptible to HP infection and the development of preneoplastic lesions [135,137]. Another study showed an increased acute inflammatory response to HP infection in persons of blood type O, due to an enhanced bacterial binding to the gastric mucosa surface. This may prime for the development of ulcers [138]. *Vice versa*, blood group type A may prime for a chronic inflammatory response, favouring cancer development through the promotion of a preneoplastic cascade [12,138]. In line with this assumption, the risk for the development of duodenal ulcers is increased in individuals with non-A blood types [114]. However, according to this argumentation an increased risk for intestinal GC should be expected for the non-O blood types, which contradicts the pronounced risk for the development of diffuse GC observed in our sample.

To summarise, the present data clearly indicate *ABO* as risk conferring gene. However, to elucidate the underlying mechanism further follow-up studies are required.

5.3 Gastric Cancer Heritability and Correlation with other Traits

In addition to the comprehensive genetic analysis and functional interpretation on the single marker level, the overall genetic architecture concerning SNP based heritability and correlations to other traits were analysed.

Heritability estimates on GC are scarce. Only one twin study is available, estimating the h^2 to be 28% [39]. With 8.48 ± 3.12% SD, our SNP-based heritability estimate thereby explains around one third of the twin-based estimate. Twin- and SNP based heritability estimates commonly deviate in that order magnitude for many phenotypes, a problem commonly referred to as missing heritability. Several causes for these discrepancies are discussed, including the omitted influence of rare variants and the lack of power due to limited sample size in GWAS [139]. The present study reveals that many signals are subtype specific. An increased sample size and focus on specific subtypes probably will lead to the identification of additional loci.

For the first time, the present dataset revealed genetic correlations between GC and risk traits, which have only been described in observational studies before. The LDSC analysis confirmed a positive correlation of GC risk with weight, alcohol intake and smoking related traits and a negative correlation with traits related to a higher education (Figure 18). Observed protective effects of weekly red and white wine intake

are probably confounded by their correlation to a higher education and thereby awareness of nutrition, lifestyle and disease prevention [140]. Reflux related traits showed no correlation to GC. As GERD represents a cardia GC specific risk factor [35], it indicates the need for subtype specific LDSC analyses, for which additional samples would be required.

5.4 Genetic Correlation of Gastric Cancer to Oesophageal Carcinoma

Except for the female cardia specific locus on 1p31, none of the genome-wide significant loci were associated in the cardia GC subsample (Table 5). This observation raised the question, whether cardia GC might be considered as a separate cancer entity, being more closely related to carcinomas affecting the nearby GOJ and the oesophagus. In general, there is an ongoing debate in the scientific field whether both cancer types represent a single disease entity, as they share common epidemiological and clinical characteristics, such as a rise in incidence worldwide and common risk factors like obesity and GERD [38]. For this reason, we were interested in estimating the germline genetic overlap of the cardia GC sample with BO and OAC.

As the sample size of the respective subsample were not sufficient for a robust LDSC analysis, the genetic overlap was examined utilizing PRS. For this purpose, we used a subset of a previously published GWAS on BO/OAC as a basis [72] and the cardia and non-cardia GC samples as target datasets. Highly significant associations of cardia GC with OAC and BO/OAC were observed. By contrast, no significant associations to non-cardia GC were identified (Figure 19). In conclusion, the PRS analysis enabled to discriminate between cardia and non-cardia GC, giving a strong hint on a shared genetic aetiology between cardia GC and OAC as well as OAC/BO as opposed to non-cardia GC.

To examine whether the genetic overlap may even help to identify shared genetic risk loci, we combined the OAC/BO GWAS with the cardia GC subsample in a meta-analysis. This resulted in the identification of two additional risk loci, which were not significant in the OAC/BO GWAS published previously (Supplementary Table 15). *HNF4G*, the nearest gene to the risk variant on chromosome 8q21, and *NR2F2*, the second nearest gene to the risk variant on chromosome 15q26, are interesting candidate genes on the functional level. NR2F2 is a known co-regulator of HNF4G and both genes play a prominent role in the development of intestinal metaplasia in gastric cell lineages [141]. It has been shown, that an upregulated *HNF4G*-expression

together with a downregulated *NR2F2*-expression lead to intestinal-like cell transformations [141].

In summary, our analysis revealed a strong genetic overlap between cardia GC and OAC/BO, and enabled the discovery of two additional risk loci in a meta-analysis of both datasets. These findings contribute to the notion that cardia GC should be accounted as separate tumour entity when examining GC.

5.5 Limitations and Outlook

The present study has several limitations that need to be taken into consideration when interpreting the results. In addition, follow-up analyses are indicated needed to further dissect the genetic architecture of GC.

Even though the largest European GC sample with the worldwide most detailed phenotypic information on GC subtypes was collected, the power of the study is still limited. The GWAS revealed strong subtype specific effects at the individual genome-wide significant loci. However, the sample sizes within those subtypes was considerably smaller, reducing the power to detect risk loci with moderate to small effects (Table 2). An increase in sample size within the subtypes would probably lead to the identification of additional risk loci, as were seen for non-cardia GC in the Asian population [12]. The same issue also limited the utilization of methods for analysing genetic correlation between GC subtypes and other traits via LDSC or their causal relationships via Mendelian randomization [142], which would be interesting to examine in the future.

Another aspect that should be taken into consideration is the GC classification system. The stratification for tumour location and Lauren type represents a rather old GC classification system with limited clinical utility [26]. Accordingly, utilizing more state-of-the-art classification systems, like somatic mutational signatures, as proposed by Bass *et al.* [26], might be beneficial to obtain further insides into tumour development and specific risk factors on the germline genetic level. However, decentralized collection of samples, missing preservation of matching tumour material and missing standards for data acquisition and tumour classification provide high hurdles for the collection of relevant sample sizes.

From a statistical point of view, stratification for various subtypes, as done in the present GWAS analysis, would require further correction for multiple testing. Still, we reported all variants reaching the genome-wide significant threshold of $P < 5 \times 10^{-8}$. This significance threshold is based on the outdated assumption of the presence of

one million independent SNPs in the human genome, but is nowadays common sense and applied by convention [14]. Thus, most multi-trait and subtype specific analyses are based on this conventional *P*-value threshold, irrespectively of the number of analysed traits [143,144]. It is assumed that the applied significance threshold still reduces the risk of false positive detection to an acceptable extent. Another way of confirming the validity of GWAS associations is the replication of associations in independent datasets [14]. As we only had access to one GWAS discovery sample, replication of the findings in an independent European sample was not possible. However, some loci were already reported in the Asian population and all identified loci showed uniform effect directions when dissecting the discovery dataset to national samples (Supplementary Figure 3). This provides evidence, that the findings of the present GWAS study represent true positives. In addition, the replication of a total of five loci from the Asian population indicates that it would be worth to conduct a trans-ancestry meta-analysis on GC in the near future (Table 7).

For all identified risk loci this study aimed to provide functional hypothesis in the context of GC development, mostly based on the collected mucosa transcriptome data. GWAS, eQTL and TWAS analyses, for the detection of single genetic risk variants and regulatory variants respectively, are *per se* hypothesis free approaches considering functional causality. As such, an overlap between risk variants for a phenotype and regulatory variants for a tissue type do not equate to a causal relationship with the phenotype under investigation. Co-regulation of multiple genes by single variants, cell type heterogeneity between samples and pleiotropy across tissues are some examples for mechanisms, which could result in false positive assumptions [145]. Although some of the discussed findings were supported *in vitro* by previously published studies, other loci and deducted hypotheses would need further functional follow-up for confirmation. Such functional experiments are usually very laborious and often only carried out for individual and carefully selected loci [146] and were beyond the scope of this study. However, our study provides valuable insights for the design of follow-up *in vitro* experiments.

For the present sample, it would be of great value to expand the TWAS and eQTL analyses to obtain further functional insights. As such, for the corpus and antrum transcriptome datasets, a *trans*-eQTL analysis may unravel gene interactions and pathways of relevance for some risk loci. As an example, we showed in a previous study for the risk locus on chromosome 8q24, an involvement of *trans*-regulatory effects, indicating a downregulation of *MBOAT7* encoded on chromosome 19 in risk allele carriers [147]. However, in the present datasets technical limitations have to be solved upfront. For example, the utilized 3’ mRNA-sequencing technology leads to many false positive findings due to an overlap with pseudogenic sequences,

mimicking trans-effects. On the other hand, due to a high burden of multiple testing, a larger dataset with more statistical power would be beneficial.

Another important aspect concerning eQTL and TWAS is the tissue and the context of transcriptome datasets used for analysis. While this study focused entirely on the transcriptomes of healthy gastric mucosa, some risk variants may exert their effect only in a disease relevant cellular context. We were able to show the relevance of such disease specific contexts in a previous study, when examining eQTLs in the context of an activated innate immune response and their overlap to phenotypes involving the immune system, not seen in the naïve state [148]. In the context of GC, an interesting setting would be the examination of eQTLs in HP infected mucosa samples, which we are currently investigating.

Finally, we focused in our study on the analysis of common genetic variants conferring to GC risk. However, it is widely assumed, that rare variants also contribute to the genetic risk in complex genetic phenotypes, which cannot be identified with array-based genotyping methods applied in this study [139]. This notion is supported by a wide spectrum of genes causing tumour syndromes that also lead to a strong increase of the risk for GC development. Several additional candidate genes could be identified in the past years [149]. However, some of these candidate genes still need to be confirmed and others probably remain to be identified. For this purpose, we are currently preparing a large whole exome sequencing dataset, including over 500 GC cases with an early age-at-onset (diagnosis < 50 years). The results should give further hints and insights to the contribution of rare variants in the development of GC and thus will complement the findings of the current study.

6 Summary

GC is a clinically heterogeneous and one of the most common malignant tumour entities affecting the human GI tract with a multifactorial aetiology. Beside environmental risk factors, also genetic risk variants contribute to the development of GC. Their identification and functional interpretation using GWAS already resulted in valuable insights into the pathophysiology and genetic architecture of the disease. However, up to today, these studies were almost exclusively conducted in the Asian population and relevant sub-classifications like the tumour location (cardia and non-cardia) and the histological tumour types (diffuse and intestinal) were often not considered.

The aim of this study was to characterize the genetic risk architecture of GC in the European population according to subtypes and to estimate their genetic overlap with known risk factors and other tumour entities. To this end, we collected the largest European sample of GC cases. In addition, we collected gastric corpus and antrum mucosa biopsies from healthy donors to conduct eQTL and TWAS analyses. These samples built the basis for the, to date, largest European GWAS meta-analysis, comprising 5.816 patients and 10.999 controls, and the largest gastric mucosa transcriptome datasets, including 361 samples from the corpus and 342 samples from the antrum.

The GWAS led to the identification and replication of nine GC risk loci. Stratification for the tumour location revealed that all but one locus contributed specifically to the risk of non-cardia GC, showing no association to cardia GC. Furthermore, two loci specifically contributed to the risk of developing tumours of the intestinal type and three loci showed a pronounced effect for the risk of developing diffuse type GC. This finding highlights the heterogeneous pathophysiology of GC and exemplifies the need for the identification and stratification of clinical relevant subtypes.

The gastric mucosa based TWAS and eQTL analysis provided evidence for the prioritization of the gene *MUC1* at chromosome 1q22, *ANKRD50* at chromosome 4q28, *PTGER4* at chromosome 5p13, *PSCA* at chromosome 8q24 and *ABO* at chromosome 9q34 as risk conferring genes. In line with the prioritization of *ABO*, we found that the blood group O exerts protective effects for non-cardia and diffuse GC, while blood group A increases risk for both GC subtypes.

Furthermore, *NEGR1* at chromosome 1p31, *ALK* at chromosome 2q23, *KLF6* at chromosome 10p15 and *HNF1B* at chromosome 17q12 are promising candidate genes at the respective risk loci.

On the polygenic level, LDSC revealed a positive correlation of GC with obesity, smoking and alcohol consumption related traits, whereas a higher education and socioeconomic status have protective effects. For the first time, these findings confirm the results of observational studies on a genetic level.

As cardia GC is suspected to represent a separate tumour entity, being more closely related to OAC, the genetic correlation between both entities was examined utilizing PRS analysis and a large European in-house OAC/BO GWAS. The analysis revealed that cardia GC and OAC are genetically homogenous at the polygenic level and can be discriminated from non-cardia GC. This finding was further supported by the identification of additional shared risk loci, after meta-analysing cardia GC subtype and OAC/BO.

All in all, the presented GWAS meta-analysis and follow-up studies provided new insights into the pathophysiology of GC at the single variant and polygenic level. The results indicate that GC is genetically heterogeneous in respect to location and histopathology. Moreover, the findings point to common molecular mechanisms underlying cardia GC and OAC/BO.

7 References

1 Ogobuiro I, Gonzales J, Tuma F. Physiology, Gastrointestinal. *StatPearls* Published Online First: 21 April 2022.https://www.ncbi.nlm.nih.gov/books/NBK537103/ (accessed 26 Jul 2022).

2 Chaudhry SR, Liman MNP, Peterson DC. Anatomy, Abdomen and Pelvis, Stomach. *StatPearls* Published Online First: 14 October 2021.https://www.ncbi.nlm.nih.gov/books/NBK482334/ (accessed 7 Aug 2022).

3 Smyth EC, Nilsson M, Grabsch HI, *et al.* Gastric cancer. *Lancet* 2020;**396**:635–48. doi:10.1016/S0140-6736(20)31288-5

4 Abnet CC, Freedman ND, Hu N, *et al.* A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nat Genet* 2010;**42**:764–7. doi:10.1038/ng.649

5 Hu N, Wang Z, Song X, *et al.* Genome-wide association study of gastric adenocarcinoma in Asia: a comparison of associations between cardia and non-cardia tumours. *Gut* 2016;**65**:1611–8. doi:10.1136/GUTJNL-2015-309340

6 Ishigaki K, Akiyama M, Kanai M, *et al.* Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat Genet* 2020;**52**:669–79. doi:10.1038/S41588-020-0640-3

7 Sampson JN, Wheeler WA, Yeager M, *et al.* Analysis of heritability and shared heritability based on genome-wide association studies for thirteen cancer types. *J Natl Cancer Inst* 2016;**107**:1–11. doi:10.1093/jnci/djv279

8 Shi Y, Hu Z, Wu C, *et al.* A genome-wide association study identifies new susceptibility loci for non-cardia gastric cancer at 3q13.31 and 5p13.1. *Nat Genet* 2011;**43**:1215–8. doi:10.1038/ng.978

9 Jin G, Ma H, Wu C, *et al.* Genetic variants at 6p21.1 and 7p15.3 are associated with risk of multiple cancers in Han Chinese. *Am J Hum Genet* 2012;**91**:928–34. doi:10.1016/J.AJHG.2012.09.009

10 Wang Z, Dai J, Hu N, *et al.* Identification of new susceptibility loci for gastric non-cardia adenocarcinoma: pooled results from two Chinese genome-wide association studies. *Gut* 2017;**66**:581–7. doi:10.1136/GUTJNL-2015-310612

11 Yan C, Zhu M, Ding Y, *et al.* Meta-analysis of genome-wide association studies and functional assays decipher susceptibility genes for gastric cancer in Chinese populations. *Gut* 2020;**69**:641–51. doi:10.1136/GUTJNL-2019-318760

12 Tanikawa C, Kamatani Y, Toyoshima O, *et al.* Genome-wide association study identifies gastric cancer susceptibility loci at 12q24.11-12 and 20q11.21. *Cancer Sci* 2018;**109**:4015–24. doi:10.1111/cas.13815

13 Sakamoto H, Yoshimura K, Saeki N, *et al.* Genetic variation in PSCA is associated with susceptibility to diffuse-type gastric cancer. *Nat Genet* 2008;**40**:730–40. doi:10.1038/ng.152

14 Uffelmann E, Huang QQ, Munung NS, *et al.* Genome-wide association studies. *Nat Rev Methods Prim 2021 11* 2021;**1**:1–21. doi:10.1038/s43586-021-00056-9

15 Gamazon ER, Wheeler HE, Shah KP, *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 2015;**47**:1091–8. doi:10.1038/ng.3367

16 Flemström G, Isenberg JI. Gastroduodenal mucosal alkaline secretion and mucosal protection. *News Physiol Sci* 2001;**16**:23–8. doi:10.1152/PHYSIOLOGYONLINE.2001.16.1.23

17 Sung H, Ferlay J, Siegel RL, *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* 2021;**71**:209–49. doi:10.3322/CAAC.21660

18 Yang X, Zhang T, Zhang H, *et al.* Temporal trend of gastric cancer burden along with its risk factors in China from 1990 to 2019, and projections until 2030: comparison with Japan, South Korea, and Mongolia. *Biomark Res* 2021;**9**. doi:10.1186/S40364-021-00340-6

19 Crew KD, Neugut AI. Epidemiology of gastric cancer. *World J Gastroenterol* 2006;**12**:354–62. doi:10.3748/WJG.V12.I3.354

20 Sekiguchi M, Oda I, Matsuda T, *et al.* Epidemiological Trends and Future Perspectives of Gastric Cancer in Eastern Asia. *Digestion* 2022;**103**:22–8. doi:10.1159/000518483

21 Cuzzuol BR, Santos Vieira E, Rocha G, *et al.* Gastric Cancer: A Brief Review, from Risk Factors to Treatment. *Arch Gastroenterol Res* 2020;**1**:34–9. doi:10.33696/GASTROENTEROLOGY.1.008

22 Röcken C. Molecular classification of gastric cancer. *Expert Rev Mol Diagn* 2017;**17**:293–301. doi:10.1080/14737159.2017.1286985

23 Sano T, Kodera Y. Japanese classification of gastric carcinoma: 3rd English edition. *Gastric Cancer* 2011;**14**:101–12. doi:10.1007/S10120-011-0041-5

24 Machado JC, Nogueira AMMF, Carneiro F, *et al.* Gastric carcinoma exhibits distinct types of cell differentiation: an immunohistochemical study of trefoil peptides (TFF1 and TFF2) and mucins (MUC1, MUC2, MUC5AC, and MUC6). *J Pathol* 2000;**190**:437–43. doi:10.1002/(sici)1096-9896(200003)190:4

25 Goseki N, Takizawa T, Koike M. Differences in the mode of the extension of gastric cancer classified by histological type: new histological classification of gastric carcinoma. *Gut* 1992;**33**:606–12. doi:10.1136/GUT.33.5.606

26 Bass AJ, Thorsson V, Shmulevich I, *et al.* Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* 2014;**513**:202–9. doi:10.1038/nature13480

27 Palli D, Bianchi S, Cipriani F, *et al.* Reproducibility of histologic classification of gastric cancer. *Br J Cancer* 1991;**63**:765–8. doi:10.1038/BJC.1991.171

28 Grabsch HI, Tan P. Gastric cancer pathology and underlying molecular mechanisms. *Dig Surg* 2013;**30**:150–8. doi:10.1159/000350876

29 Larsson SC, Orsini N, Wolk A. Processed meat consumption and stomach cancer risk: a meta-analysis. *J Natl Cancer Inst* 2006;**98**:1078–87. doi:10.1093/JNCI/DJJ301

30 Chen Y, Liu L, Wang X, *et al.* Body mass index and risk of gastric cancer: a meta-analysis of a population with more than ten million from 24 prospective studies. *Cancer Epidemiol Biomarkers Prev* 2013;**22**:1395–408. doi:10.1158/1055-9965.EPI-13-0042

31 Sexton RE, Al Hallak MN, Diab M, *et al.* Gastric cancer: a comprehensive review of current and future treatment strategies. *Cancer Metastasis Rev* 2020;**39**:1179–203. doi:10.1007/S10555-020-09925-3

32 Buckland G, Travier N, Huerta JM, *et al.* Healthy lifestyle index and risk of gastric adenocarcinoma in the EPIC cohort study. *Int J cancer* 2015;**137**:598–606. doi:10.1002/IJC.29411

33 Nagel G, Linseisen J, Boshuizen HC, *et al.* Socioeconomic position and the risk of gastric and oesophageal cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC-EURGAST). *Int J Epidemiol* 2007;**36**:66–76. doi:10.1093/IJE/DYL275

34 Bray F, Ferlay J, Soerjomataram I, *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;**68**:394–424. doi:10.3322/CAAC.21492

35 Tytgat GNJ, Bartelink H, Bernards R, *et al.* Cancer of the esophagus and gastric cardia: recent advances. *Dis esophagus Off J Int Soc Dis Esophagus* 2004;**17**:10–26. doi:10.1111/J.1442-2050.2004.00371.X

36 Oo AM, Ahmed S. Overview of gastroesophageal junction cancers. *Mini-invasive Surg* 2019;**3**:13. doi:10.20517/2574-1225.2019.02

37 Siewert JR, Stein HJ. Classification of adenocarcinoma of the oesophagogastric junction. *Br J Surg* 1998;**85**:1457–9. doi:10.1046/J.1365-2168.1998.00940.X

38 Hayakawa Y, Sethi N, Sepulveda AR, *et al.* Oesophageal adenocarcinoma and gastric cancer: should we mind the gap? *Nat Rev Cancer* 2016;**16**:305–18. doi:10.1038/NRC.2016.24

39 Lichtenstein P, Holm N V., Verkasalo PK, *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med* 2000;**343**:78–85. doi:10.1056/NEJM200007133430201

40 Helgason H, Rafnar T, Olafsdottir HS, *et al.* Loss-of-function variants in ATM confer risk of gastric cancer. *Nat Genet* 2015;**47**:906–10. doi:10.1038/ng.3342

41 N S, JW C, DG D, *et al.* Familial Gastric Cancers. *Oncologist* 2015;**20**:1523–6. doi:10.1634/THEONCOLOGIST.2015-0205

42 Clark DF, Michalski ST, Tondon R, *et al.* Loss-of-function variants in CTNNA1 detected on multigene panel testing in individuals with gastric or breast cancer. *Genet Med* 2020;**22**:840–6. doi:10.1038/S41436-020-0753-1

43 Oliveira C, Pinheiro H, Figueiredo J, *et al.* Familial gastric cancer: genetic susceptibility, pathology, and implications for management. *Lancet Oncol* 2015;**16**:e60–70. doi:10.1016/S1470-2045(14)71016-2

44 Lu C, Xie M, Wendl MC, *et al.* Patterns and functional implications of rare germline variants across 12 cancer types. *Nat Commun* 2015;**6**:10086. doi:10.1038/ncomms10086

45 Hansford S, Kaurah P, Li-Chang H, *et al.* Hereditary Diffuse Gastric Cancer Syndrome: CDH1 Mutations and Beyond. *JAMA Oncol* 2015;**1**:23–32. doi:10.1001/JAMAONCOL.2014.168

46 Sahasrabudhe R, Lott P, Bohorquez M, *et al.* Germline Mutations in PALB2, BRCA1, and RAD51C, Which Regulate DNA Recombination Repair, in Patients With Gastric Cancer. *Gastroenterology* 2017;**152**:983-986.e6. doi:10.1053/J.GASTRO.2016.12.010

47 Frazer KA, Murray SS, Schork NJ, *et al.* Human genetic variation and its contribution to complex traits. *Nat Rev Genet* 2009;**10**:241–51. doi:10.1038/NRG2554

48 Pertea M, Shumate A, Pertea G, *et al.* CHESS: A new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. *Genome Biol* 2018;**19**:208. doi:10.1186/s13059-018-1590-2

49 Dunham I, Kundaje A, Aldred SF, *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74. doi:10.1038/nature11247

50 Auton A, Abecasis GR, Altshuler DM, *et al.* A global reference for human genetic variation. Nature. 2015;**526**:68–74. doi:10.1038/nature15393

51 Sherry ST, Ward MH, Kholodov M, *et al.* DbSNP: The NCBI database of genetic variation. *Nucleic Acids Res* 2001;**29**:308–11. doi:10.1093/nar/29.1.308

52 Karczewski KJ, Francioli LC, Tiao G, *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;**581**:434–43. doi:10.1038/S41586-020-2308-7

53 Daly MJ, Rioux JD, Schaffner SF, *et al.* High-resolution haplotype structure in the human genome. *Nat Genet* 2001;**29**:229–32. doi:10.1038/ng1001-229

54 Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 2001;**29**:217–22. doi:10.1038/ng1001-217

55 Thorisson GA, Smith A V., Krishnan L, *et al.* The International HapMap Project Web site. *Genome Res* 2005;**15**:1592–3. doi:10.1101/gr.4413105

56 Johnson GCL, Esposito L, Barratt BJ, *et al.* Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001;**29**:233–7. doi:10.1038/ng1001-233

57 Yun L, Willer C, Sanna S, *et al.* Genotype imputation. Annu. Rev. Genomics Hum. Genet. 2009;**10**:387–406. doi:10.1146/annurev.genom.9.081307.164242

58 Visscher PM, Wray NR, Zhang Q, *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 2017;**101**:5–22. doi:10.1016/J.AJHG.2017.06.005

59 Pe'er I, Yelensky R, Altshuler D, *et al.* Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* 2008;**32**:381–5. doi:10.1002/gepi.20303

60 Buniello A, Macarthur JAL, Cerezo M, *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* 2019;**47**:D1005–12. doi:10.1093/nar/gky1120

61 Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017;**169**:1177–86. doi:10.1016/j.cell.2017.05.038

62 Young AI. Solving the missing heritability problem. *PLOS Genet* 2019;**15**:e1008222. doi:10.1371/journal.pgen.1008222

63 Visscher PM, Wray NR, Zhang Q, *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. Am. J. Hum. Genet. 2017;**101**:5–22. doi:10.1016/j.ajhg.2017.06.005

64 Aguet F, Barbeira AN, Bonazzola R, *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* 2020;**369**:1318–30. doi:10.1126/SCIENCE.AAZ1776

65 Robinson JR, Denny JC, Roden DM, *et al.* Genome-wide and Phenome-wide Approaches to Understand Variable Drug Actions in Electronic Health Records. *Clin Transl Sci* 2018;**11**:112–22. doi:10.1111/CTS.12522

66 Lee JJ, McGue M, Iacono WG, *et al.* The accuracy of LD Score regression as an estimator of confounding and genetic correlations in genome-wide association studies. *Genet Epidemiol* 2018;**42**:783–95. doi:10.1002/GEPI.22161

67 Bulik-Sullivan B, Finucane HK, Anttila V, *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015;**47**:1236–41. doi:10.1038/NG.3406

68 Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* 2020;**15**:2759–72. doi:10.1038/s41596-020-0353-1

69 Leitsalu L, Haller T, Esko T, *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int J Epidemiol* 2015;**44**:1137–47. doi:10.1093/IJE/DYT268

70 Sudlow C, Gallacher J, Allen N, *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 2015;**12**. doi:10.1371/JOURNAL.PMED.1001779

71 Colquhoun A, Arnold M, Ferlay J, *et al.* Global patterns of cardia and non-cardia gastric cancer incidence in 2012. *Gut* 2015;**64**:1881–8. doi:10.1136/GUTJNL-2014-308915

72 Gharahkhani P, Fitzgerald RC, Vaughan TL, *et al.* Genome-wide association studies in oesophageal adenocarcinoma and Barrett's oesophagus: a large-scale meta-analysis. *Lancet Oncol* 2016;**17**:1363–73. doi:10.1016/S1470-2045(16)30240-6

73 Schroeder A, Mueller O, Stocker S, *et al.* The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol* 2006;**7**. doi:10.1186/1471-2199-7-3

74 Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30. doi:10.1093/BIOINFORMATICS/BTT656

75 Dobin A, Davis CA, Schlesinger F, *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21. doi:10.1093/BIOINFORMATICS/BTS635

76 Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40. doi:10.1093/BIOINFORMATICS/BTP616

77 Purcell S, Neale B, Todd-Brown K, *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007;**81**:559–75. doi:10.1086/519795

78 Manichaikul A, Mychaleckyj JC, Rich SS, *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* 2010;**26**:2867–73. doi:10.1093/BIOINFORMATICS/BTQ559

79 Taliun D, Harris DN, Kessler MD, *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021;**590**:290–9. doi:10.1038/s41586-021-03205-y

80 Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009;**5**. doi:10.1371/JOURNAL.PGEN.1000529

81 Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide

association scans. *Bioinformatics* 2010;**26**:2190–1. doi:10.1093/BIOINFORMATICS/BTQ340

82 Panagiotou OA, Willer CJ, Hirschhorn JN, *et al.* The Power of Meta-Analysis in Genome Wide Association Studies. *Annu Rev Genomics Hum Genet* 2013;**14**:441. doi:10.1146/ANNUREV-GENOM-091212-153520

83 Yang J, Ferreira T, Morris AP, *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 2012;**44**:369–75. doi:10.1038/ng.2213

84 Groot HE, Sierra LEV, Said MA, *et al.* Genetically determined ABO blood group and its associations with health and disease. *Arterioscler Thromb Vasc Biol* 2020;**40**:830–8. doi:10.1161/ATVBAHA.119.313658

85 Sudlow C, Gallacher J, Allen N, *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med* 2015;**12**:e1001779. doi:10.1371/JOURNAL.PMED.1001779

86 Kurki MI, Karjalainen J, Palta P, *et al.* FinnGen: Unique genetic insights from combining isolated population and national health register data. *medRxiv* 2022;:2022.03.03.22271360. doi:10.1101/2022.03.03.22271360

87 Ghoussaini M, Mountjoy E, Carmona M, *et al.* Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res* 2021;**49**:D1311–20. doi:10.1093/NAR/GKAA840

88 Smyth G. Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. New York: : Springer 2005. 397–420.http://master.bioconductor.org/help/course-materials/2005/BioC2005/labs/lab01/printing/6.appendix.pdf (accessed 15 Aug 2014).

89 Kuleshov M V., Jones MR, Rouillard AD, *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**:W90–7. doi:10.1093/NAR/GKW377

90 Fort A, Panousis NI, Garieri M, *et al.* MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. *Bioinformatics* 2017;**33**:1895–7. doi:10.1093/BIOINFORMATICS/BTX074

91 Gusev A, Ko A, Shi H, *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* 2016;**48**:245–52. doi:10.1038/ng.3506

92 Bulik-Sullivan B, Finucane HK, Anttila V, *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* 2015;**47**:1236–41. doi:10.1038/NG.3406

93 Zheng J, Erzurumluoglu AM, Elsworth BL, *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* 2017;**33**:272. doi:10.1093/BIOINFORMATICS/BTW613

94 Gharahkhani P, Fitzgerald RC, Vaughan TL, *et al.* Genome-wide association studies in oesophageal adenocarcinoma and Barrett's oesophagus: a large-scale meta-analysis. *Lancet Oncol* 2016;**17**:1363–73. doi:10.1016/S1470-2045(16)30240-6

95 Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience* 2019;**8**. doi:10.1093/GIGASCIENCE/GIZ082

96 Khera A V., Chaffin M, Wade KH, *et al.* Polygenic Prediction of Weight and Obesity Trajectories from Birth to Adulthood. *Cell* 2019;**177**:587-596.e9. doi:10.1016/j.cell.2019.03.028

97 Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. *Front Genet* 2020;**11**:1–21. doi:10.3389/fgene.2020.00424

98 Sacco MA, Abenavoli L, Juan C, *et al.* Biological Mechanisms behind Wischnewsky Spots Finding on Gastric Mucosa: Autopsy Cases and Literature Review. *Int J Environ Res Public Health* 2022;**19**. doi:10.3390/IJERPH19063601

99 Nookaew I, Thorell K, Worah K, *et al.* Transcriptome signatures in Helicobacter pylori-infected

mucosa identifies acidic mammalian chitinase loss as a corpus atrophy marker. *BMC Med Genomics* 2013;**6**. doi:10.1186/1755-8794-6-41

100 Silva F, Carvalho F, Peixoto A, *et al.* MUC1 gene polymorphism in the gastric carcinogenesis pathway. *Eur J Hum Genet* 2001;**9**:548–52. doi:10.1038/SJ.EJHG.5200677

101 Saeki N, Saito A, Choi IJ, *et al.* A functional single nucleotide polymorphism in mucin 1, at chromosome 1q22, determines susceptibility to diffuse-type gastric cancer. *Gastroenterology* 2011;**140**:892–902. doi:10.1053/J.GASTRO.2010.10.058

102 Carson DD. The cytoplasmic tail of MUC1: A very busy place. *Sci Signal* 2008;**1**:1–5. doi:10.1126/scisignal.127pe35

103 Supruniuk K, Radziejewska I. MUC1 is an oncoprotein with a significant role in apoptosis (Review). *Int J Oncol* 2021;**59**:1–11. doi:10.3892/ijo.2021.5248

104 Rajabi H, Hiraki M, Tagde A, *et al.* MUC1-C activates EZH2 expression and function in human cancer cells. *Sci Rep* 2017;**7**:1–13. doi:10.1038/s41598-017-07850-0

105 Hankenson KD, Hormuzdi SG, Meganck JA, *et al.* Mice with a Disruption of the Thrombospondin 3 Gene Differ in Geometric and Biomechanical Properties of Bone and Have Accelerated Development of the Femoral Head. *Mol Cell Biol* 2005;**25**:5599. doi:10.1128/MCB.25.13.5599-5606.2005

106 Yang S, He X, Liu Y, *et al.* Prognostic Significance of Serum Uric Acid and Gamma-Glutamyltransferase in Patients with Advanced Gastric Cancer. *Dis Markers* 2019;**2019**. doi:10.1155/2019/1415421

107 Mossé YP, Laudenslager M, Longo L, *et al.* Identification of ALK as a major familial neuroblastoma predisposition gene. *Nature* 2008;**455**:930–5. doi:10.1038/NATURE07261

108 Soda M, Choi YL, Enomoto M, *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* 2007;**448**:561–6. doi:10.1038/nature05945

109 Lin E, Li L, Guan Y, *et al.* Exon array profiling detects EML4-ALK fusion in breast, colorectal, and non-small cell lung cancers. *Mol Cancer Res* 2009;**7**:1466–76. doi:10.1158/1541-7786.MCR-08-0522

110 Ross JS, Ali SM, Fasan O, *et al.* ALK Fusions in a Wide Variety of Tumor Types Respond to Anti-ALK Targeted Therapy. *Oncologist* 2017;**22**:1444–50. doi:10.1634/THEONCOLOGIST.2016-0488

111 Kim H, Ahn S, Kang WK, *et al.* Two Gastric Cancers With Uncommon ALK Fusion Diagnosed With Comprehensive Panel Sequencing and Confirmed With Companion Diagnostic Assay. *AJSP Rev Reports* 2022;**27**:9–12. doi:10.1097/PCR.0000000000000480

112 Rizzato C, Kato I, Plummer M, *et al.* Genetic Variation in PSCA and Risk of Gastric Advanced Preneoplastic Lesions and Cancer in Relation to Helicobacter pylori Infection. *PLoS One* 2013;**8**:73100. doi:10.1371/JOURNAL.PONE.0073100

113 Saeki N, Gu J, Yoshida T, *et al.* Prostate stem cell antigen: A Jekyll and Hyde molecule? *Clin Cancer Res* 2010;**16**:3533–8. doi:10.1158/1078-0432.CCR-09-3169

114 Tanikawa C, Urabe Y, Matsuo K, *et al.* A genome-wide association study identifies two susceptibility loci for duodenal ulcer in the Japanese population. *Nat Genet* 2012;**44**:430–4. doi:10.1038/ng.1109

115 Toyoshima O, Tanikawa C, Yamamoto R, *et al.* Decrease in PSCA expression caused by Helicobacter pylor infection may promote progression to severe gastritis. *Oncotarget* 2018;**9**:3936–45. doi:10.18632/oncotarget.23278

116 Oze I, Masahiro T, Yasumasa N, *et al.* GWAS analysis reveals a significant contribution of PSCA to the risk of Heliobacter pylori-induced gastric atrophy. 2019.

117 Hojny J, Bartu M, Krkavcova E, *et al.* Identification of novel HNF1B mRNA splicing variants and their qualitative and semi-quantitative profile in selected healthy and tumour tissues. *Sci Rep* 2020;**10**. doi:10.1038/S41598-020-63733-X

118 Sakaue S, Kanai M, Tanigawa Y, *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* 2021;**53**:1415–24. doi:10.1038/S41588-021-00931-X

119 Fairfield CJ, Drake TM, Pius R, *et al.* Genome-wide analysis identifies gallstone-susceptibility loci including genes regulating gastrointestinal motility. *Hepatology* 2022;**75**:1081–94. doi:10.1002/HEP.32199

120 Kang SH, Kim YH, Roh YH, *et al.* Gallstone, cholecystectomy and risk of gastric cancer. *Ann hepato-biliary-pancreatic Surg* 2017;**21**:131. doi:10.14701/AHBPS.2017.21.3.131

121 Nogueira L, Freedman ND, Engels EA, *et al.* Gallstones, Cholecystectomy, and Risk of Digestive System Cancers. *Am J Epidemiol* 2014;**179**:731–9. doi:10.1093/AJE/KWT322

122 Kaare M, Mikheim K, Lilleväli K, *et al.* High-Fat Diet Induces Pre-Diabetes and Distinct Sex-Specific Metabolic Alterations in Negr1-Deficient Mice. *Biomedicines* 2021;**9**. doi:10.3390/BIOMEDICINES9091148

123 Ong JS, An J, Han X, *et al.* Multitrait genetic association analysis identifies 50 new risk loci for gastro-oesophageal reflux, seven new loci for Barrett's oesophagus and provides insights into clinical heterogeneity in reflux diagnosis. *Gut* 2022;**71**. doi:10.1136/GUTJNL-2020-323906

124 Lee PH, Anttila V, Won H, *et al.* Genomic Relationships, Novel Loci, and Pleiotropic Mechanisms across Eight Psychiatric Disorders. *Cell* 2019;**179**:1469-1482.e11. doi:10.1016/J.CELL.2019.11.020

125 Levey DF, Stein MB, Wendt FR, *et al.* Bi-ancestral depression GWAS in the Million Veteran Program and meta-analysis in >1.2 million individuals highlight new therapeutic directions. *Nat Neurosci* 2021;**24**. doi:10.1038/S41593-021-00860-2

126 Rana S, Mobin M. Association of the NEGR1 rs2815752 with obesity and related traits in Pakistani females. *Ups J Med Sci* 2020;**125**:226–34. doi:10.1080/03009734.2020.1756996

127 Pirastu N, Cordioli M, Nandakumar P, *et al.* Genetic analyses identify widespread sex-differential participation bias. *Nat Genet* 2021;**53**:663–71. doi:10.1038/S41588-021-00846-7

128 Cho YG, Kim CJ, Park CH, *et al.* Genetic alterations of the KLF6 gene in gastric cancer. *Oncogene* 2005;**24**:4588–90. doi:10.1038/SJ.ONC.1208670

129 Himmerich H, Bentley J, Kan C, *et al.* Genetic risk factors for eating disorders: an update and insights into pathophysiology. *Ther Adv Psychopharmacol* 2019;**9**:204512531881473. doi:10.1177/2045125318814734

130 Duriez P, Ramoz N, Gorwood P, *et al.* A Metabolic Perspective on Reward Abnormalities in Anorexia Nervosa. *Trends Endocrinol Metab* 2019;**30**:915–28. doi:10.1016/J.TEM.2019.08.004

131 Echizen K, Hirose O, Maeda Y, *et al.* Inflammation in gastric cancer: Interplay of the COX-2/prostaglandin E2 and Toll-like receptor/MyD88 pathways. *Cancer Sci* 2016;**107**:391–7. doi:10.1111/CAS.12901

132 Yamamoto F. Molecular genetics and genomics of the ABO blood group system. *Ann Blood* 2021;**6**. doi:10.21037/AOB-20-71

133 Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, *et al.* Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet* 2009;**41**:986–90. doi:10.1038/NG.429

134 Shelton JF, Shastri AJ, Ye C, *et al.* Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. *Nat Genet* 2021;**53**:801–8. doi:10.1038/S41588-021-00854-7

135 Wang Z, Liu L, Ji J, *et al.* ABO Blood Group System and Gastric Cancer: A Case-Control Study and Meta-Analysis. *Int J Mol Sci* 2012;**13**:13308. doi:10.3390/IJMS131013308

136 Duell EJ, Bonet C, Muñoz X, *et al.* Variation at ABO histo-blood group and FUT loci and diffuse and intestinal gastric cancer risk in a European population. *Int J cancer* 2015;**136**:880–93. doi:10.1002/IJC.29034

137 Nakao M, Matsuo K, Ito H, *et al.* ABO genotype and the risk of gastric cancer, atrophic gastritis, and Helicobacter pylori infection. *Cancer Epidemiol Biomarkers Prev* 2011;**20**:1665–72. doi:10.1158/1055-9965.EPI-11-0213

138 Alkout AM, Blackwell CC, Weir DM. Increased inflammatory responses of persons of blood group O to Helicobacter pylori. *J Infect Dis* 2000;**181**:1364–9. doi:10.1086/315375

139 Young AI. Solving the missing heritability problem. *PLoS Genet* 2019;**15**. doi:10.1371/JOURNAL.PGEN.1008222

140 Rosoff DB, Clarke TK, Adams MJ, *et al.* Educational attainment impacts drinking behaviors and risk for alcohol dependence: results from a two-sample Mendelian randomization study with ~780,000 participants. *Mol Psychiatry* 2021;**26**:1119. doi:10.1038/S41380-019-0535-9

141 Sousa JF, Nam KT, Petersen CP, *et al.* miR-30-HNF4γ and miR-194-NR2F2 regulatory networks contribute to the upregulation of metaplasia markers in the stomach. *Gut* 2016;**65**:914–24. doi:10.1136/GUTJNL-2014-308759

142 Sanderson E, Glymour MM, Holmes M V., *et al.* Mendelian randomization. *Nat Rev Methods Prim 2022 21* 2022;**2**:1–21. doi:10.1038/s43586-021-00092-5

143 Hautakangas H, Winsvold BS, Ruotsalainen SE, *et al.* Genome-wide analysis of 102,084 migraine cases identifies 123 risk loci and subtype-specific risk alleles. *Nat Genet 2022 542* 2022;**54**:152–60. doi:10.1038/s41588-021-00990-0

144 Campbell PJ, Getz G, Korbel JO, *et al.* Pan-cancer analysis of whole genomes. *Nat 2020 5787793* 2020;**578**:82–93. doi:10.1038/s41586-020-1969-6

145 Wainberg M, Sinnott-Armstrong N, Mancuso N, *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* 2019;**51**:592. doi:10.1038/S41588-019-0385-Z

146 Claussnitzer M, Dankel SN, Kim K-H, *et al.* FTO Obesity Variant Circuitry and Adipocyte Browning in Humans. *N Engl J Med* 2015;:150819140043007. doi:10.1056/NEJMoa1502214

147 Heinrichs SKM, Hess T, Becker J, *et al.* Evidence for PTGER4,PSCA, and MBOAT7 as risk genes for gastric cancer on the genome and transcriptome level. *Cancer Med* 2018;**7**:5057. doi:10.1002/CAM4.1719

148 Kim S, Becker J, Bechheim M, *et al.* Characterizing the genetic basis of innate immune response in TLR4-activated human monocytes. *Nat Commun* 2014;**5**:5236. doi:10.1038/ncomms6236

149 Cosma L-S, Schlosser S, Tews HC, *et al.* Hereditary Diffuse Gastric Cancer: Molecular Genetics, Biological Mechanisms and Current Therapeutic Approaches. *Int J Mol Sci 2022, Vol 23, Page 7821* 2022;**23**:7821. doi:10.3390/IJMS23147821

Supplement A: Materials and Methods

Supplement A1: List of used devices, reagents, chemicals, kits, software and databases used to conduct this study.

Devices

Autoclave
- Systec D-150, Systec GmbH
- Varioklav® 135 S Dampfsterilisator, H+P Labortechnik GmbH

Automated liquid handling
- Biomek® NX MC Laboratory Automation Workstation, Beckman Coulter GmbH
- Biomek® NX S8G Laboratory Automation Workstation, Beckman Coulter GmbH

DNA-Biobanking
- SmartScan Solo™ 2D Barcode Reader, Thermo Fisher Scientific GmbH
- SmartScan 96 2D Barcode Reader, Thermo Fisher Scientific GmbH

Genotyping System
- iScan System, Illumina Inc.

Lab water purification system
- Milli-Q A10 Synthesis, Merck KGaA

Nucleic acid extraction
- Homogenisator, Precellys® 24, VWR International GmbH
- Magnetic Separation Module I, Perkin Elmer Chemagen Technologie GmbH
- QIAcube, Qiagen GmbH

Nucleic acid quantification and quality control
- NanoDrop® 1000 Spectrophotometer, Peqlab Biotechnology GmbH
- NanoDrop® 8000 Spectrophotometer, Peqlab Biotechnology GmbH
- Qubit 3 Fluorometer, Thermo Fisher Scientific GmbH
- Varioskan™ LUX multimode microplate reader, Thermo Fisher Scientific GmbH
- 4200 TapeStation System, Agilent Technologies Deutschland GmbH

Pipettes
- accu-jet® pro, BRAND GmbH & Co. KG
- Eppendorf Research® (variabel), different sizes, Eppendorf GmbH
- Multipette® plus, Eppendorf AG
- Transferpette®, BRAND GmbH & Co. KG
- Transferpette® S-8, BRAND GmbH & Co. KG

- Transferpette®-8/-12 electronic, BRAND GmbH & Co. KG

Shakers and heating devices

- Overheadshaker REAX 2, Heidolph Instruments GmbH & Co. KG
- Platformshaker UNIMAX 1010, Heidolph Instruments GmbH & Co. KG
- Thermomixer comfort, Eppendorf AG
- Platformshaker TITRAMAX 101, Heidolph Instruments GmbH & Co. KG.
- Vortex Genie 2, Scientific Industries Inc.
- Vortex Mixer IKA MS2-S8, Agilent Technologies Deutschland GmbH

Sequencer

- cBot System, Illumina Inc.
- HiSeq 2500 v4, Illumina Inc.

Thermocycler

- Eppendorf Mastercycler X50, Eppendorf AG

Centrifuges

- Megafuge 1.0 R, Heraeus GmbH
- Biofuge stratos, Heraeus GmbH
- Biofuge fresco, Heraeus GmbH
- Biofuge pico, Heraeus GmbH
- neoLab Mini-Zentrifuge Spectrafuge®, neoLab
- Concentrator Plus System, Eppendorf AG

Chemicals and buffers

- Ethanol absolut (C_2H_5OH) (EtOH) (100%), AppliChem GmbH
- RNAfree water, Qiagen GmbH
- RNAlater™ Stabilysing solution, Thermo Fisher Scientific GmbH
- 10X TBE Puffer, Life Technologies GmbH
- Tris-EDTA (TE^{-4}; 0,1 mM EDTA, 10 mM Tris-HCl pH 8,0)
- Water, HPLC grade, Merck KGaA

Commercial Systems (Kits)

- AllPrep DNA/RNA Mini Kit, Qiagen GmbH
- Chemagic DNA Blood 10k, PerkinElmer Chemagen Technologie GmbH
- D1000 ScreenTape & Reagents, Agilent Technologies Deutschland GmbH
- HiSeq SBS Kit V4 50 cycle kit, Illumina Inc.
- Infinium Omni2.5Exome-8 Kit, Illumina Inc.
- Infinium Global Diversity Array-8 Kit, Illumina Inc.
- QuantSeq 3' mRNA-Seq Library Prep Kit FWD for Illumina, Lexogen Inc.
- Qubit dsDNA HS Assay Kit, Thermo Fisher Scientific GmbH
- Qubit RNA BR Assay Kit, Thermo Fisher Scientific GmbH

- RNA ScreenTape & Reagents, Agilent Technologies Deutschland GmbH

Software and Databases

- Biomek® Software 3.2, Beckman-Coulter GmbH
- dbSNP database (http://www.ncbi.nlm.nih.gov/projects/SNP/)
- edgeR (https://bioconductor.org/packages/release/bioc/html/edgeR.html)
- ENCODE database (https://www.encodeproject.org/)
- Enrichr (https://maayanlab.cloud/Enrichr/)
- ENSEMBL genome browser (http://www.ensembl.org/index.html)
- FastQC v0.11.7 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- FeatureCounts v1.5.1 (http://subread.sourceforge.net/)
- FinnGen (https://www.finngen.fi/en)
- FUSION (http://gusevlab.org/projects/fusion/)
- GCTA-COJO (GCTA version 1.93.0beta)
- Genomestudio 2.0, Illumina Inc.
- Genotype-Tissue Expression project (GTEx) database (https://gtexportal.org/)
- GWAS Catalog (https://www.ebi.ac.uk/gwas/)
- Impute2 (https://mathgen.stats.ox.ac.uk/impute/impute.html)
- KING (https://www.kingrelatedness.com/)
- LD Hub (https://ldsc.broadinstitute.org/)
- LDlink (https://ldlink.nci.nih.gov/)
- LDSC (v1.0.1) (https://github.com/bulik/ldsc)
- Limma (https://www.bioconductor.org/packages/release/bioc/html/limma.html)
- Locus Zoom (https://my.locuszoom.org/)
- MatrixEQTL (http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/)
- METAL (https://github.com/statgen/METAL)
- NanoDrop® ND-1000 v3.3.0, Peqlab Biotechnology GmbH
- NanoDrop® ND-8000 v2.2.1, Peqlab Biotechnology GmbH
- Open Targets Genetics Portal (https://genetics.opentargets.org/)
- PLINK v1.9 & 2.0 (http://pngu.mgh.harvard.edu/~purcell/plink/)
- PRSice tool (https://www.prsice.info/)
- Pubmed (https://www.ncbi.nlm.nih.gov/pubmed/)
- QTLtools (https://qtltools.github.io/qtltools/)
- R (https://cran.r-project.org/bin/windows/base/)
- SAS Software (SAS Institute 2008)
- STAR Aligner 2.5.2b (https://github.com/alexdobin/STAR)
- TOPMed Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov/#!)
- UCSC Genome Browser (https://genome.ucsc.edu/)
- UKBiobank (https://www.ukbiobank.ac.uk)

Supplementary Table 1: Description of GC GWAS samples. All patients were of European descent and were recruited across nine different countries. Only patients were included with a primary and histopathologically confirmed diagnosis of gastric adenocarcinoma. Informed consent was obtained from all cases and approval was obtained from ethic boards at each participating site. Patients from Spain and Portugal were merged as Iberian cases. In addition, patients of European descent from the from the UK Biobank [70] and The Estonian Biobank (EstBB) [69] with a gastric adenocarcinoma according to ICD-10 code C16 were included (British and Estonian samples, not listed).

Country	Recruiting site	Recruitment period	Cases (female/male)	Involved clinical institutions
Sweden	Stockholm	2014-2017	246 (95/151)	Patients were recruited through the Department of Upper GI Diseases, Karolinska University Hospital, Stockholm, through the National Registry for Esophageal- and Gastric Cancer (NREV)
Latvia	Riga	2007-2017	476 (188/288)	Patients were recruited at the Institute of Clinical and Preventive Medicine at the Latvia Oncology Centre, Riga East University Hospital, University of Latvia, Riga
Lithuania	Kaunas	2010-2017	219 (83/136)	Patients were recruited at the Gastroenterology Department and Institute for Digestive Research, Lithuanian University of Health Sciences, Kaunas
Poland	Szczecin	1990-2007	388 (158/230)	Patients were recruited at the Department of Gastroenterology, Pomeranian Medical University, Szczecin
	Lublin	2016-2017	37 (6/31)	Patients were recruited at 2nd Department of General Surgery, Medical University of Lublin, Lublin
Germany	Magdeburg	2013-2017	688 (427/261)	Patients were recruited through the Department of Gastroenterology, Hepatology and Infectious Diseases, Otto-von-Guericke University Hospital, Magdeburg, in a nation-wide study involving university or community hospitals in Magdeburg, Bonn, Erlangen, Solingen, Mainz and Leipzig
	Bonn	2013-2017	75 (27/48)	Patients were recruited through the Institute of Human Genetics, University Hospital Bonn, Bonn, at oncological centres in Dortmund and Troisdorf and the community hospital Koblenz
	Frankfurt	2010-2015	384 (102/282)	Patients were recruited through the Krankenhaus Nordwest, University Cancer Center, Frankfurt, as part of a therapy study at 38 German trial sites
	Hamburg	2005-2014	77 (23/54)	Patients were recruited at the Department of General and Abdominal Surgery, University Hospital Hamburg-Eppendorf, Hamburg
	Heidelberg (DKFZ)	1996-2003	92 (32/60)	Patients were recruited through the Division of Clinical Epidemiology and Aging Research at the German Cancer Research Center (DKFZ) in Heidelberg at community hospitals in the state Saarland (Germany)
	Heidelberg (University)	1999-2017	261 (89/172)	Patients were recruited at the Department of General, Visceral and Transplantation Surgery, Heidelberg University Hospital, Heidelberg
	Cologne	1998-2016	70 (44/26)	Patients were recruited at the Department of General, Visceral, Cancer and Transplant Surgery, University Hospital of Cologne, Cologne
	Berlin	1999-2006	136 (51/85)	Patients were recruited at the Department of Surgery and Surgical Oncology, Robert-Rössle Klinik, Charité, Berlin
	München	1999-2005	66 (13/53)	Patients were recruited at the Department of Surgery, Klinikum rechts der Isar, Technical University of Munich, München
France	Lille	2015-2017	147 (42/105)	Patients were recruited through the clinicobiological database French EsoGastric Tumours (FREGAT Database), CHU de Lille, CHU de Lyon, CHU de Rennes, Centre Oscar Lambret (Lille), and CHU de Bordeaux
Spain	Zaragoza	2002-2012	652 (226/426)	Patients were recruited through the Instituto de Investigación Sanitaria Aragón, Zaragoza, in a nation-wide study involving 16 community hospitals belonging to seven regions of Spain (Aragon, Catalonia, Basque Country, Madrid, Castile and Leon, Andalusia, Canary Islands)
	Madrid	2008-2013	223 (66/157)	Patients were recruited through collaborating hospitals from 10 regions of Spain (Asturias, Barcelona, Cantabria, Granada, Huelva, León, Madrid, Murcia, Navarra and Valencia)
Portugal	Porto	2001-2008	55 (30/25)	Patients were recruited at the Digestive Clinic at the Portuguese Institute of Oncology of Porto, Porto
Italy	San Giovanni Rotondo	2004-2016	185 (74/111)	Patients were recruited at the Division of Gastroenterology of IRCCS 'Casa Sollievo della Sofferenza', San Giovanni Rotondo
	Rome	2002-2012	168 (74/94)	Patients were recruited at the A. Gemelli'' teaching hospital, Rome
	Aviano	2012-2014	91 (31/60)	Patients were recruited at the Unit of Oncological Gastroenterology, Centro di Riferimento Oncologico, National Cancer Institute, IRCCS Aviano, Aviano
	Cremona	2016-2017	146 (50/96)	Patients were recruited at the Medical Oncology Unit, ASST of Cremona, Cremona

Supplementary Table 2: Description of control GWAS samples. All controls were of European descent and were recruited across eight different countries. The Spanish controls were used for the association analysis with GC patients from Spain and Portugal (Iberian sample, see supplementary table 3). As for cases, informed consent was obtained from all controls and approval was obtained from ethic boards at each participating site. In addition, healthy controls of European descent without history of oncological diseases from the UK Biobank [70] and The Estonian Biobank (EstBB) [69] were used at a ratio of 1:1 (British sample) and 1:20 (Estonian sample).

Country	Recruiting site	Recruitment period	Subjects (female/male)	Involved clinical institutions
Sweden	Umea	1988	1,528 (824/686)	Controls were part of the Betula study, a population-based longitudinal cohort with the objectives to study how memory and health develop across the adult lifespan, and to determine factors underlying the heterogeneity in aging
Latvia	Riga	2013-2017	475 (244/231)	Controls were part of the GISTAR study (general middle-aged population), a randomized study of H. pylori eradication and pepsinogen testing for prevention of GC mortality
Lithuania	Kaunas	2008-2016	210 (85/125)	Controls were blood donors recruited at the Lithuanian University of Health Sciences in Kaunas, Lithuania
Poland	Poznan	2013	537 (268/269)	Controls were blood donors attending the Regional Centre of Blood Donation and Treatment in Poznan, Poland
Germany	Essen	2000-2003	2,701 (1,352/1,349)	Controls were part of the Heinz Nixdorf Recall Study (HNR), a population-based cohort investigating cardiovascular diseases
France	Paris	2011-2012	406 (122/284)	Controls were recruited through a systematic urologic screening program by the CeRePP network and included subjects without history or symptoms of urological cancers
Spain	Santiago de Compostela	2006-2008	871 (429/442)	Controls were recruited through the Spanish meningococcal disease (MD) research network ESIGEM involving 43 university or community hospitals. A healthy and geographically matched control was selected for each MD patient
Italy	Milan	1993	1,360 (506/854)	Controls were part of the HYPERGENES project and were collected in continental Italy or Sardinia. All subjects were healthy and were clinically followed for more than 15 years up until at least 55 years of age (hyper-controls)

Supplementary Table 3: Overview and description of study samples with expression data for the TWAS and eQTL analyses. All individuals were of German descent and were recruited according to standardised procedures at five sites between 2016 and 2017. In all participants, a gastroscopy was performed because of unclear upper abdominal symptoms. In addition to diagnostic biopsies, mucosa tissue samples from five defined parts of the stomach were collected from each individual, namely cardia, corpus, antrum, angulus and fundus. Only participants who showed a regular gastric mucosa without HP infection in the histopathologic examination at the Institute of Pathology in Bayreuth, Germany, were included in the study. In addition, only subjects were included in whom other medical causes for the unclear upper abdominal symptoms could be excluded. Among others, this included a normal complete blood count. Informed consent was obtained from all participants and a study approval was obtained from the ethic board at the University of Bonn, Germany.

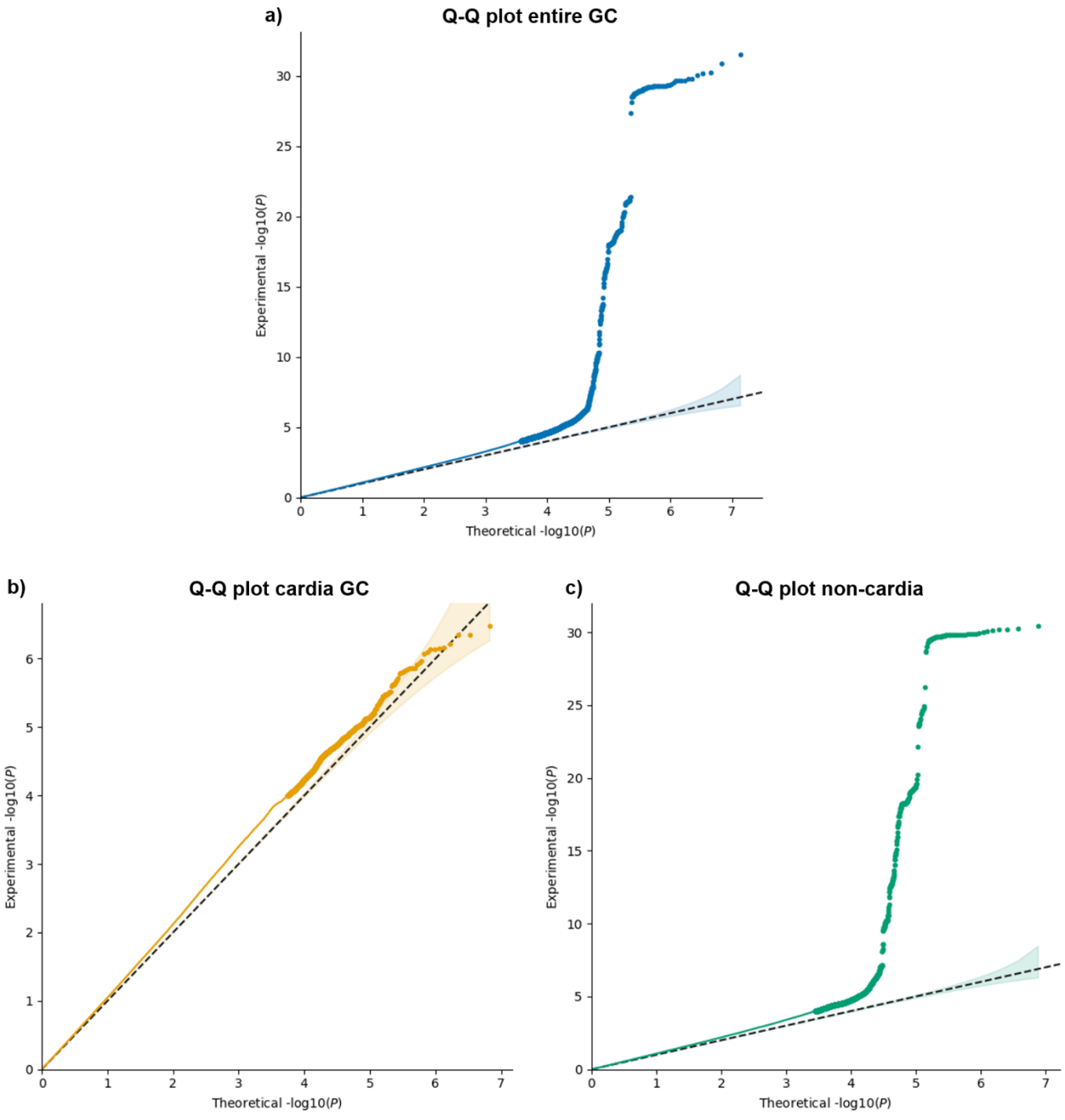
Recruiting site	Participants (female/male)	Involved clinical institutions
Ahrweiler	27 (15/12)	Department of Gastroenterology, Marienhaus Hospital Ahrweiler, Ahrweiler
Bayreuth	123 (77/46)	Gastroenterological Centre, Dr. Geppert, Bayreuth
Bonn	133 (57/76)	Gastroenterological Centre, Dr. Plaßmann, Bonn
	19 (12/7)	Department of Gastroenterology, St. Elisabeth Hospital Bonn, Bonn
Cologne	25 (17/8)	Gastroenterological Centre, Dr. Hofer, Cologne
	2 (2/-)	Department of Gastroenterology, Cologne-Holweide and Merheim Medical Centre, Cologne
Koblenz	67 (46/21)	Gastroenterological Centre, Dr. Benner, Dr. Dommermuth, Koblenz
Magdeburg	10 (10/-)	Department of Gastroenterology, Otto-von-Guericke University Hospital, Magdeburg
Siegburg	16 (16/-)	Department of Gastroenterology, Helios Hospital Siegburg, Siegburg

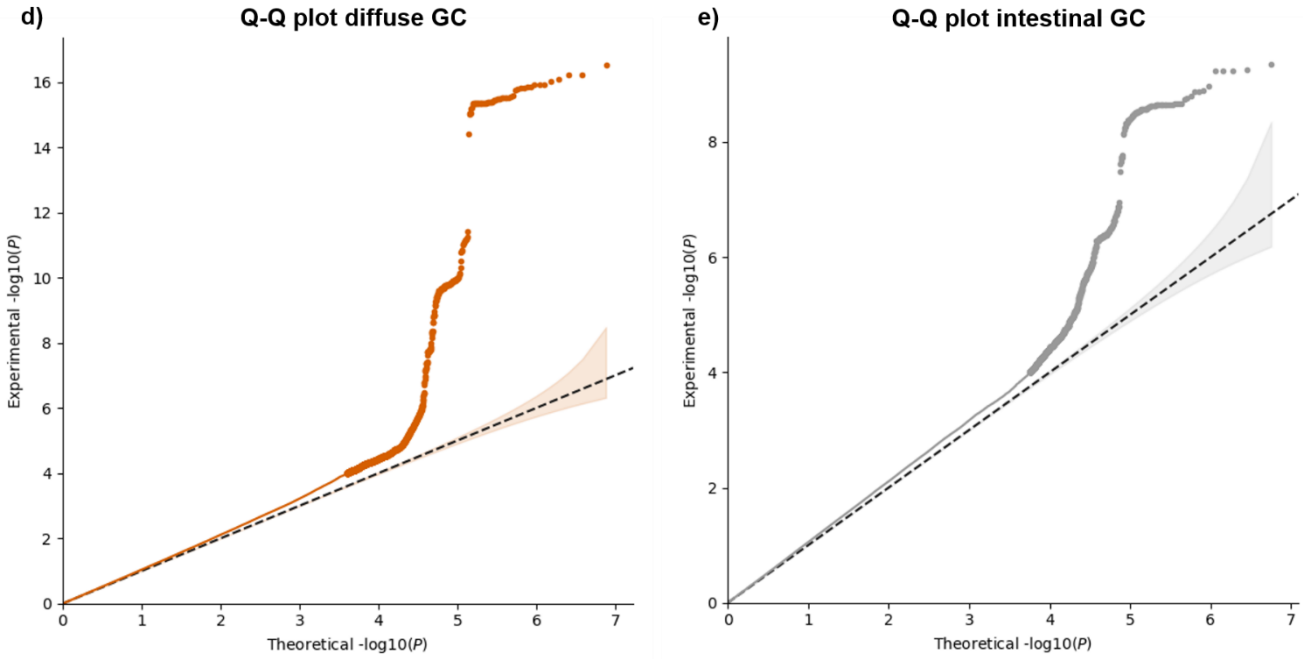
Supplementary Table 4: Description of the OAC/BO PRS and OAC/BO cross-trait GWAS samples. The OAC/BO GWAS datasets that have been published previously [94] consist of a German case-control sample (German/Bonn sample) and of three case-control samples from UK and US/Australia. Only for the German OAC/BO sample individual GWAS data were available that were used in the OAC/BO PRS study (2,646 patients, 2,732 controls). Prior to this analysis we excluded all German controls in the OAC/BO sample that were also used in the GC GWAS study to ensure that no individual was subject to both datasets. In the cross-trait GWAS meta-analysis, the German and all remaining OAC/BO samples were used together with the cardia GC sample.

OAC/BO samples	**Cases**		**Controls**	
	all	**female/male**	**all**	**female/male**
GWAS individual in-house data				
– German/Bonn sample	2,646	548/2,098	2,732	1,344/1,388
GWAS summary statistic data				
– BEACON (US/Australia)	3,914	754/3,160	6,718	2,704/4,014
– Cambridge sample	1,868	368/1,500	3,408	1,711/1,697
– Oxford sample	1,851	364/1,487	3,469	1,769/1,700
Total	**10,279**	2,034/8,245	**16,327**	7,528/8,799

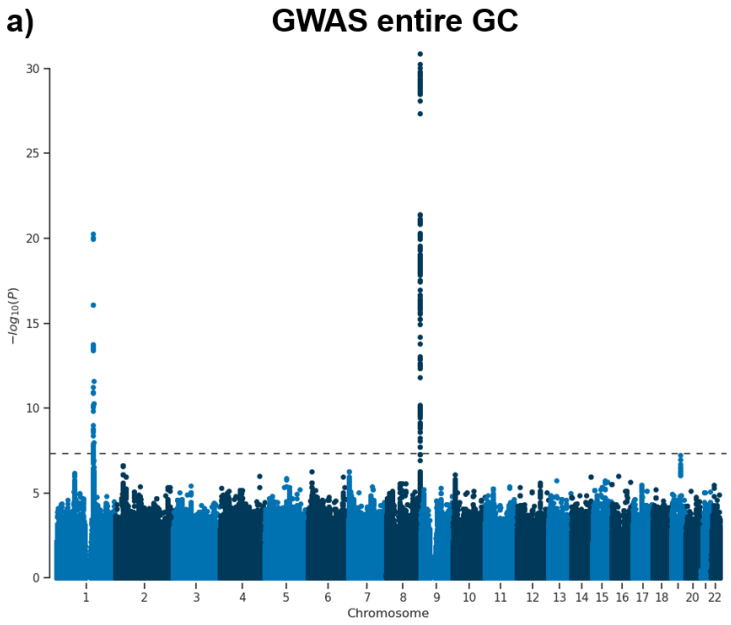
Supplement B: Results

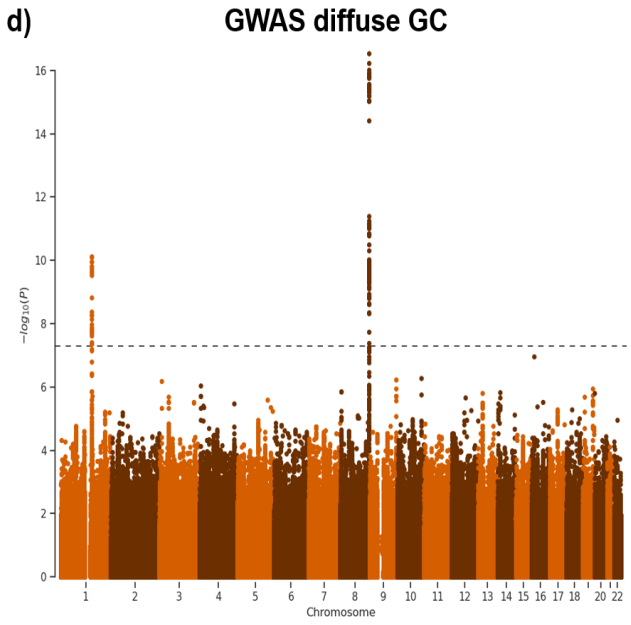
Supplementary Figure 1: GC GWAS Quantile-Quantile (Q-Q) plots. Q-Q plots for association *P*-values obtained from the GWAS of the entire (a), cardia (b), non-cardia (c), diffuse (d) and intestinal (e) GC samples are shown. The X axis shows the expected distribution of -log10(*P*-values) under the null hypothesis of no association. The Y axis shows the distribution of the observed -log10(*P*-values) in each GWAS. The dashed indicator lines show where X=Y. The genomic inflation factor lambda was 1.11, 1.07, 1.11, 1.06 and 1.07 for entire, cardia, non-cardia, diffuse and intestinal GC GWAS, respectively.

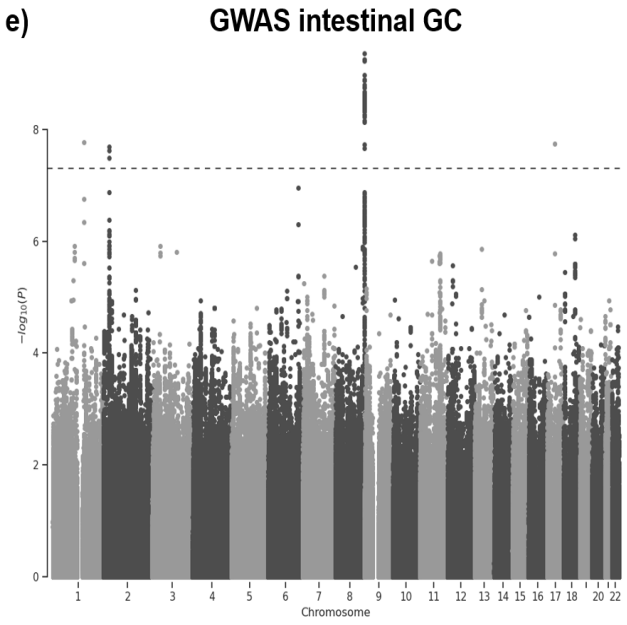


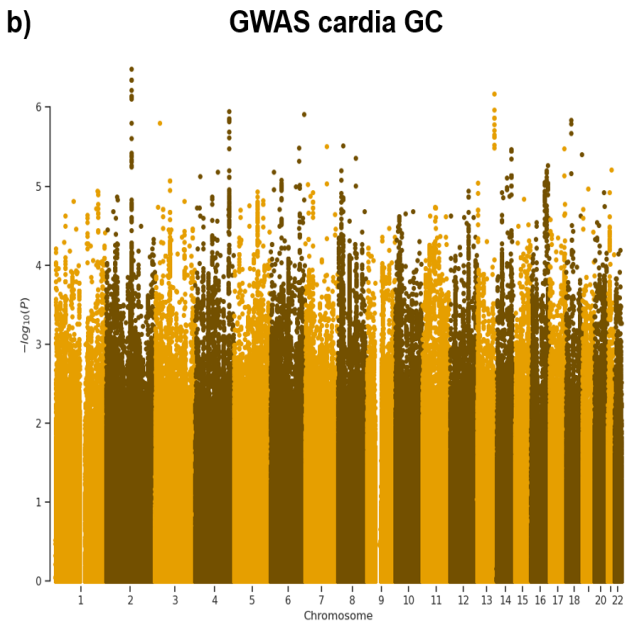


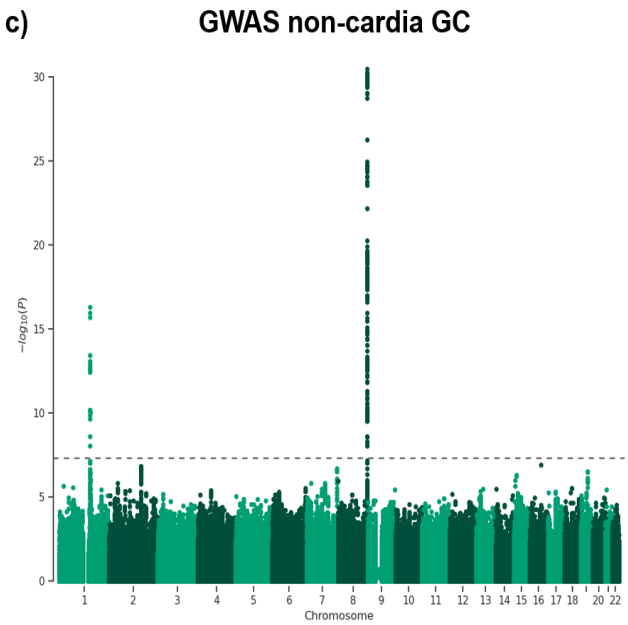
Supplementary Figure 2: GC GWAS Manhattan plots. The associations obtained from the GWAS of the entire (a), cardia (b), non-cardia (c) diffuse (d) and intestinal (e) GC samples are shown. All SNPs have been plotted against their chromosomal positions (X axis) and the observed -log10(*P*-values) in the GWAS (Y axis). All SNPs on each chromosome are shown in the same colour but a distinct colour from that of the adjacent chromosome. The horizontal lines indicate the threshold of genome-wide significant association.



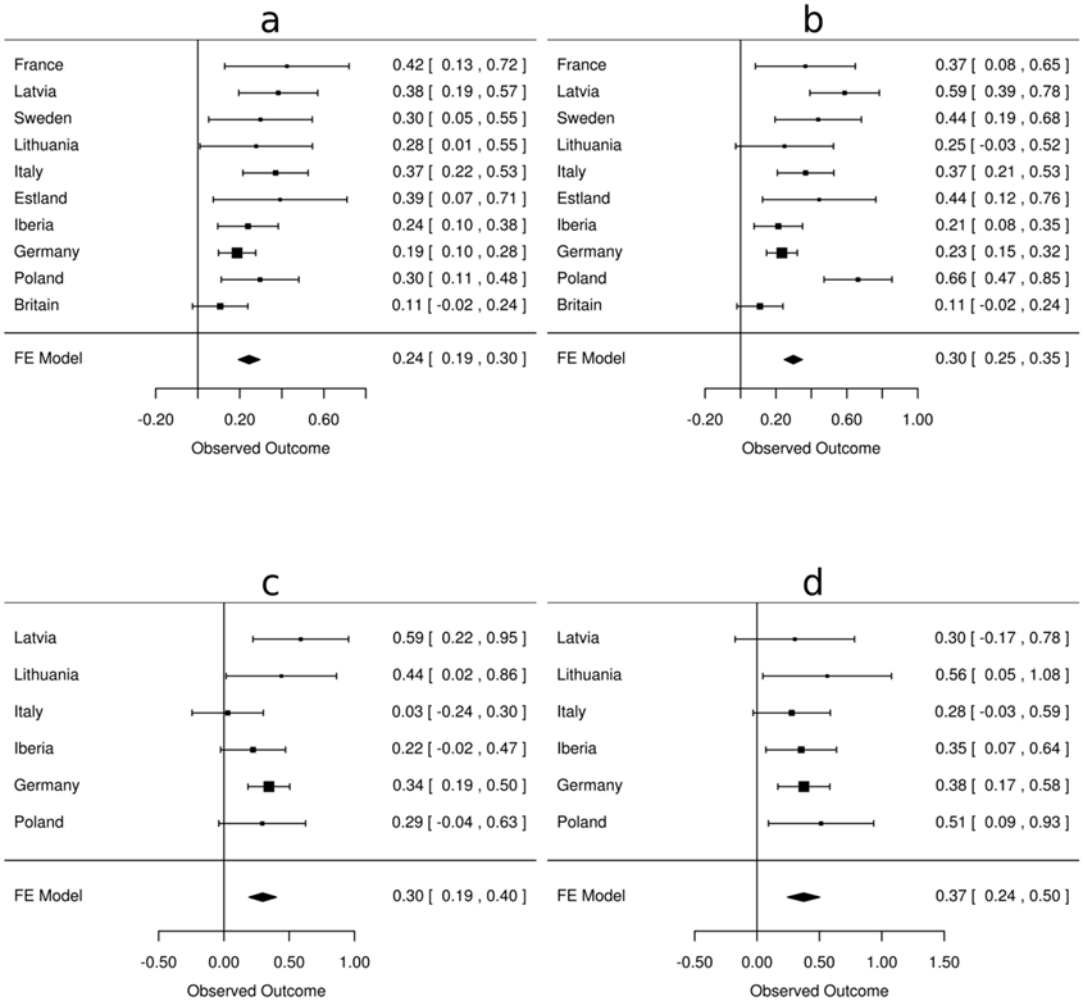

d)
GWAS diffuse GC
−log10(P)
Chromosome


e)
GWAS intestinal GC
−log10(P)
Chromosome

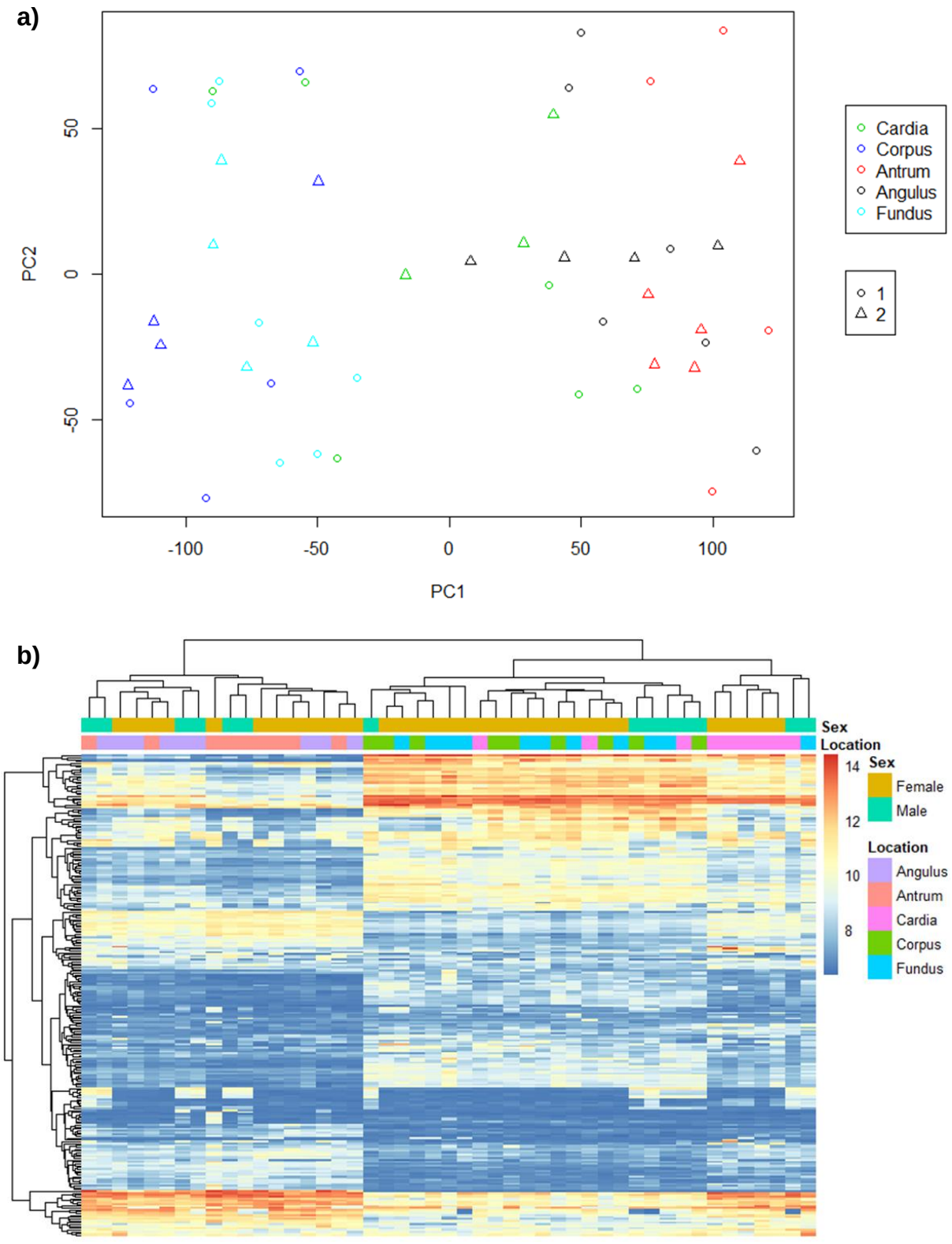

b)
GWAS cardia GC
−log10(P)
Chromosome


c)
GWAS non-cardia GC
−log10(P)
Chromosome

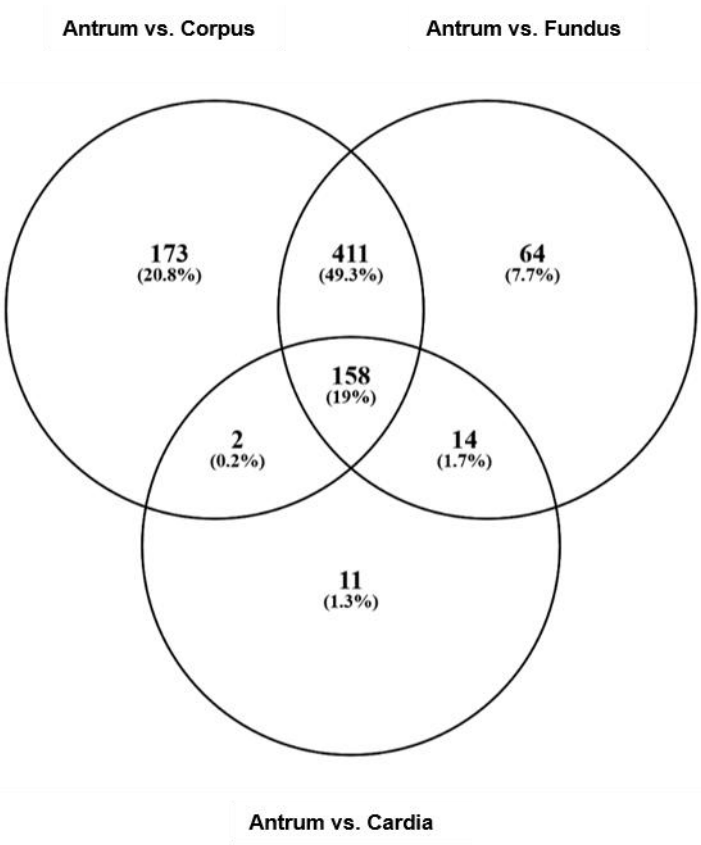
Supplementary Figure 3: Forest plots of genome-wide associated GC SNPs across all included European samples. Single country and meta-analysis association for the entire GC: (a) leading variant rs760077 at the 1q22 locus, (b) leading rs2920293 at the 8q24 locus; and for intestinal GC: (c) leading variant rs11677924 at the 2q23 locus, (d) leading variant rs17138478 at the 17q12 locus. In the figure are reported the beta (and 95% CI) from the single GWAS and from the fixed effect meta-analysis (due to limited information on GC type for some subsamples only samples with at least 50 cases are reported in panel c and d).

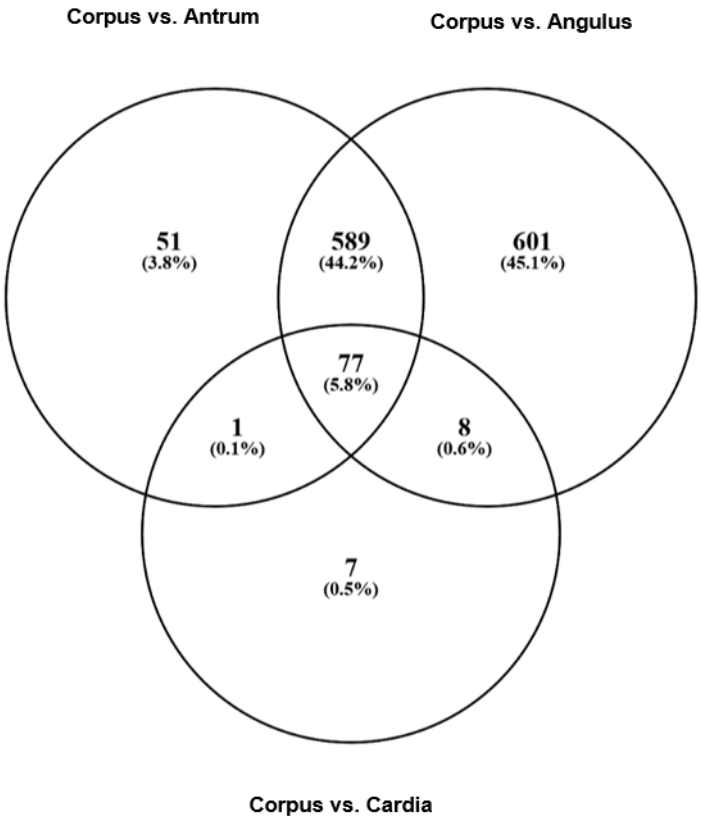


Supplementary Figure 4: Explorative analysis of expression array data of tissue biopsies taken from the cardia (N=9), corpus (N=9), fundus (N=9), antrum (N=10) and angulus (N=10). **a)** Plot of the first two principle components of the top 250 most variable expressed genes. Stomach locations are indicated by colour and batches of array processing by shape of plot symbols **b)** Unsupervised hierarchical clustering of the top 250 most variable expressed genes. Both plots show a clear differentiation between samples from the corpus and fundus compared with samples from antrum or angulus. Cardia expression data showed no uniform expression pattern among individuals, resembling either the corpus/fundus or the antrum/angulus profiles.



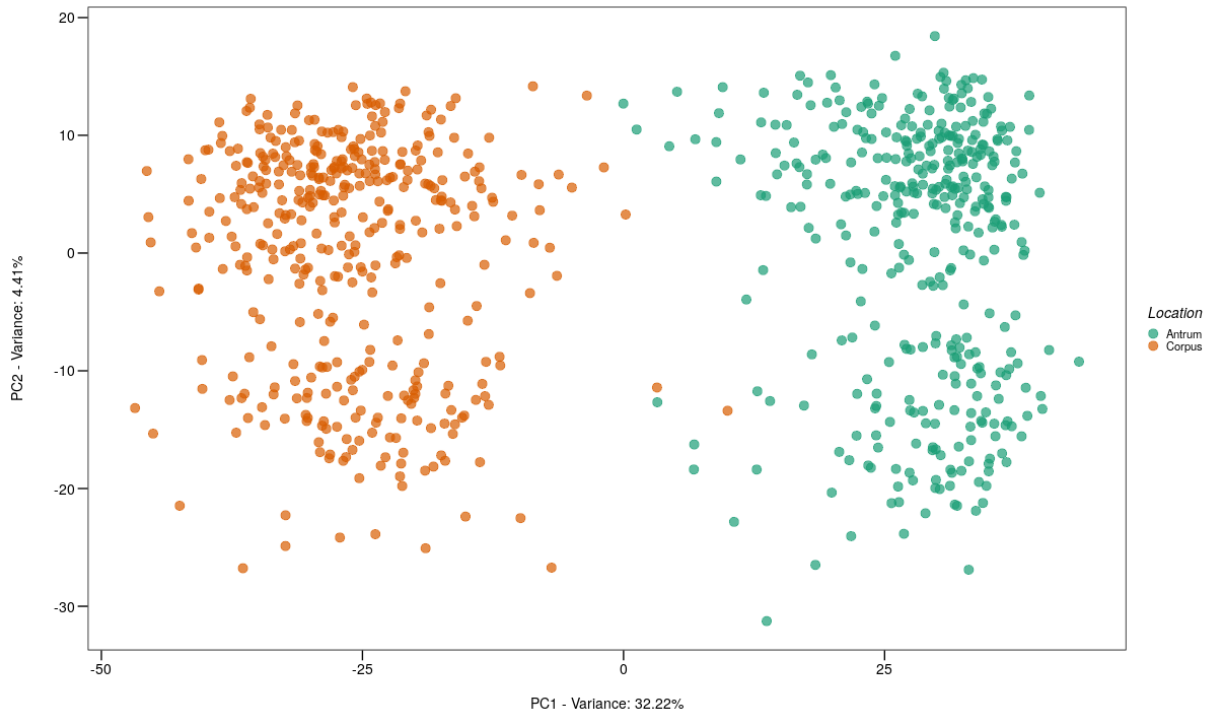
Supplementary Figure 5: Venn diagram showing the overlap of DE genes (P < 0.05; FC ≤ -2; FC ≥ -2) comparing transcriptome data from different parts of the stomach.



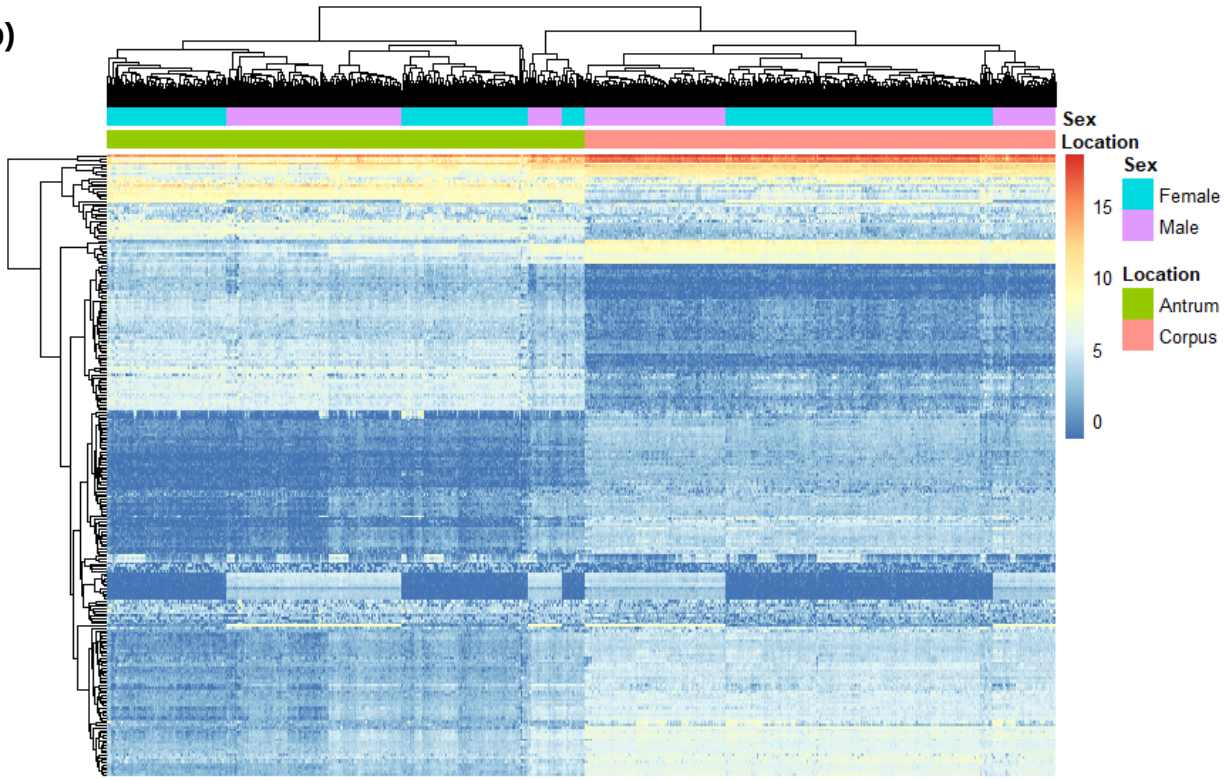


Supplementary Figure 6: Explorative analysis of RNA-Seq data from tissue biopsies of the corpus (N=362) and antrum (N=342). **a)** Plot of the first two principle components of the top 250 most variable expressed genes. Stomach locations are indicated by the colour of plot symbols. **b)** Unsupervised hierarchical clustering of the top 250 most variable expressed genes derived from RNA-Seq data from tissue biopsies of the corpus (N=362) and antrum (N=342).

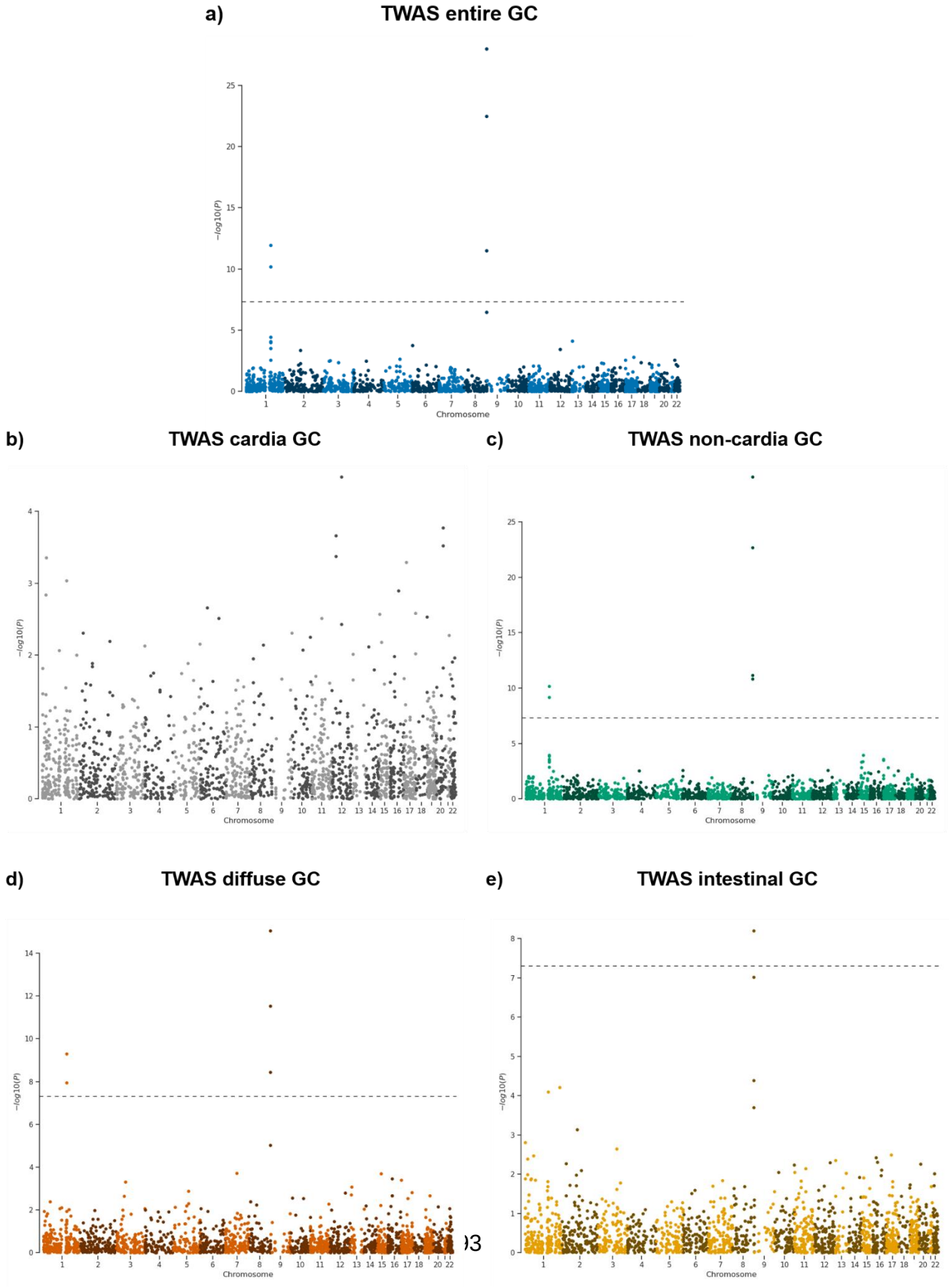
a)



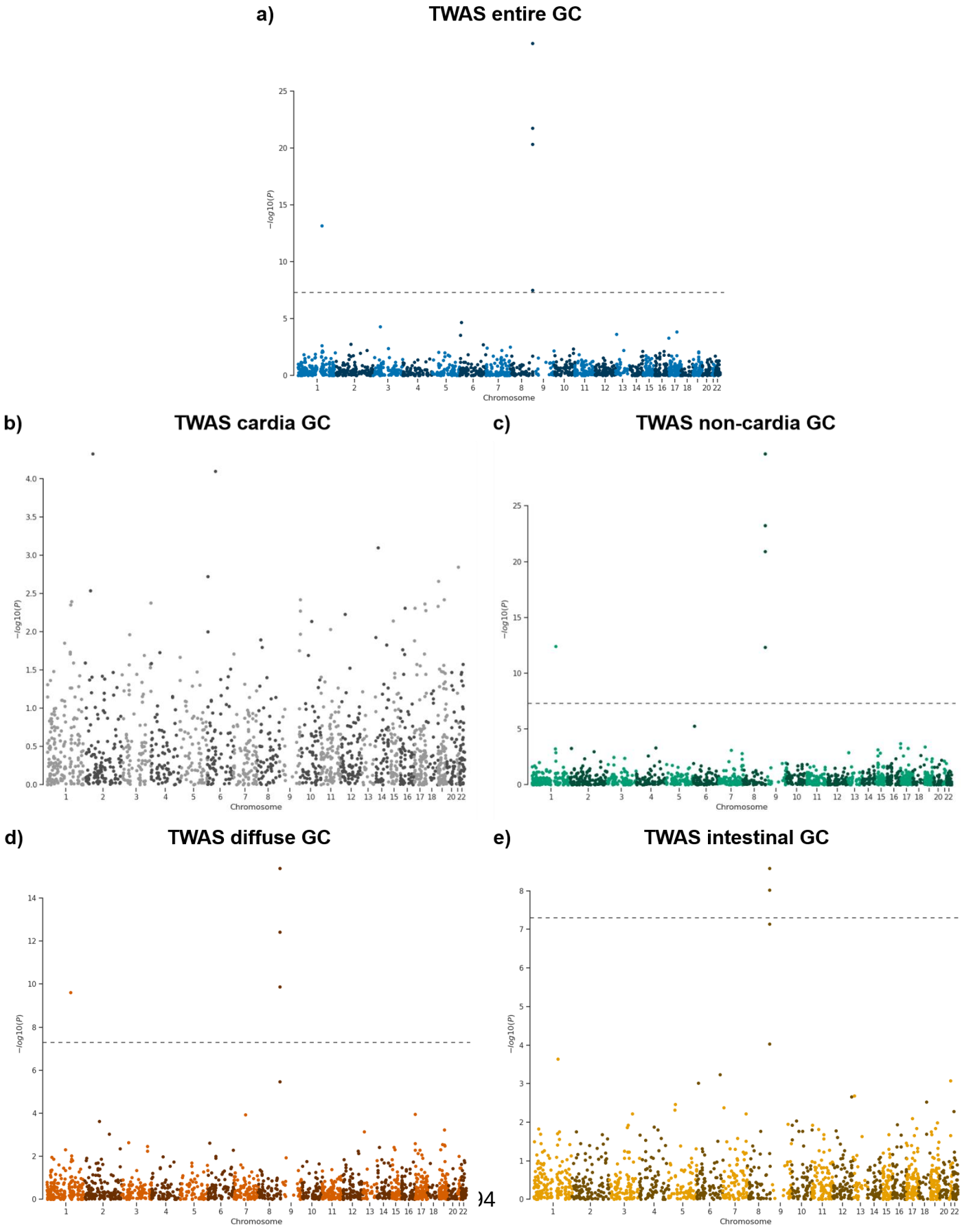
b)



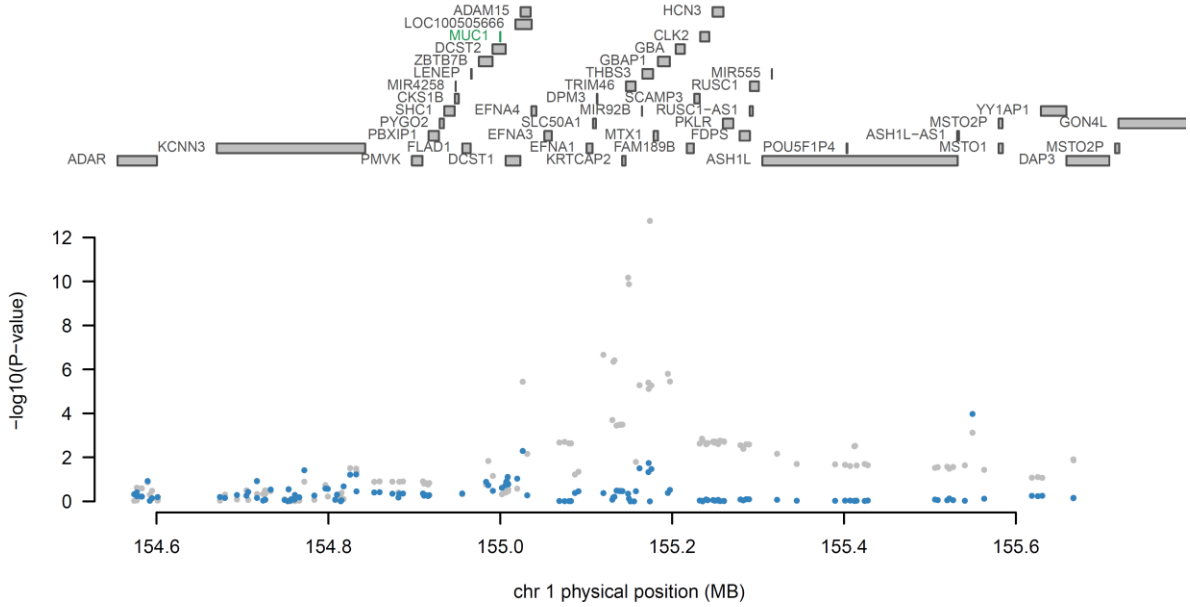
Supplementary Figure 7: Association between antrum gene expression and GC. Manhattan plots of TWAS-identified genes with predicted expression models that are associated with entire GC **(a)** as well as with GC types according to location (cardia **(b)**, non-cardia **(c)**) and Lauren type (diffuse **(d)**, intestinal **(e)**). Each point represents a single gene with physical position plotted on the x-axis and *P*-values of GC association plotted on the y-axis. The threshold for transcriptome-wide significant association is highlighted as dashed line.



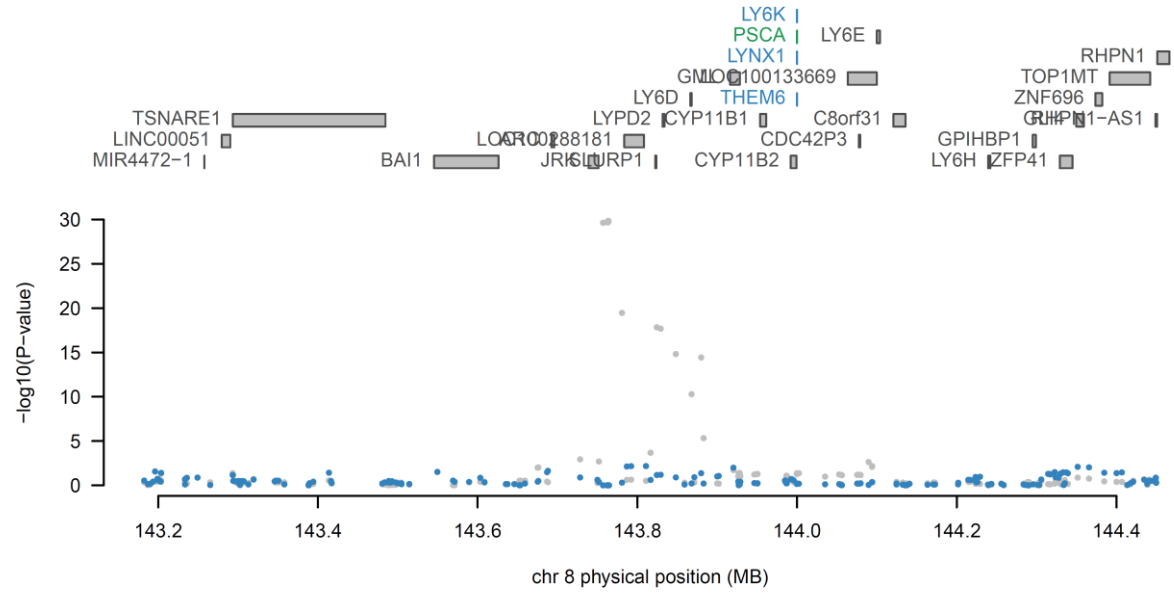
Supplementary Figure 8: Association between corpus gene expression and GC. Manhattan plots of TWAS-identified genes with predicted expression models that are associated with entire GC **(a)** as well as with GC types according to location (cardia **(b)**, non-cardia **(c)**) and Lauren type (diffuse **(d)**, intestinal **(e)**). Each point represents a single gene with physical position plotted on the x-axis and *P*-values of GC association plotted on the y-axis. The threshold for transcriptome-wide significant association is highlighted as dashed line.



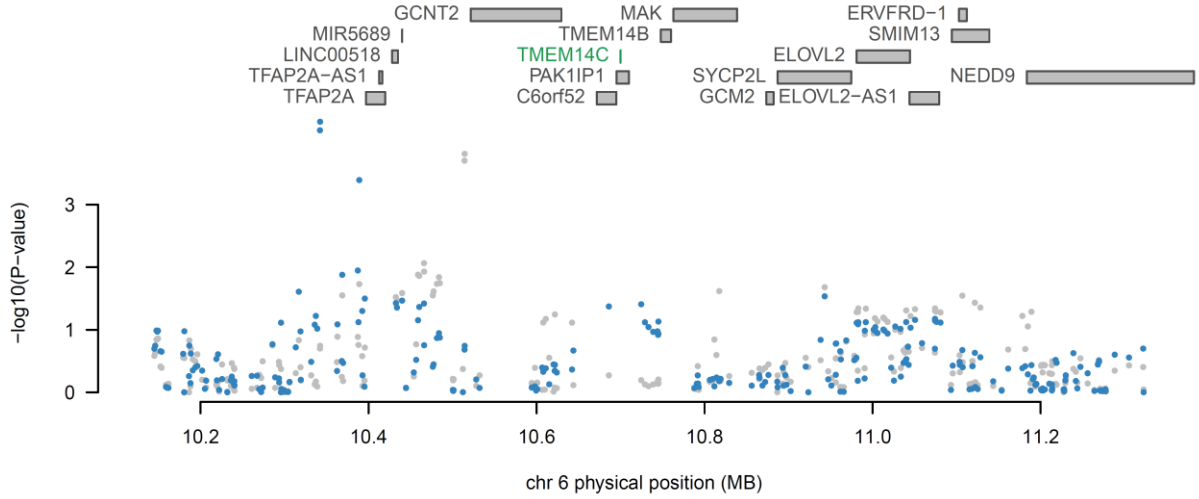
Supplementary Figure 10: Regional TWAS association plot on chromosome 1q22 in corpus mucosa. The associations of MUC1 and THBS3 expression in non-cardia are shown. The top panel shows all genes at this locus. TWAS genes associated with non-cardia GC are highlighted in green. The bottom panels show the TWAS associations before (grey) and after (blue) conditioning on the green genes.



Supplementary Figure 9 Regional TWAS association plot on chromosome 8q24 in antrum mucosa. The associations of PSCA, LY6K, THEM6 and LYNX1 expression in non-cardia GC are shown. The top panel shows all genes at this locus. TWAS genes associated with non-cardia GC are highlighted in green. The bottom panels show the TWAS associations before (grey) and after (blue) conditioning on the green genes.



Supplementary Figure 11: Regional TWAS association plot on chromosome 6p24 in corpus mucosa. The association of TMEM14C expression in non-cardia GC is shown. The top panel shows all genes at this locus. TWAS genes associated with non-cardia GC are highlighted in green. The bottom panels show the TWAS associations before (grey) and after (blue) conditioning on the green genes.



Supplementary Table 5: GC case-case comparison according to location and Lauren type. Associations of genome-wide significant GC SNPs between non-cardia and cardia GC patients as well as diffuse and intestinal GC patients are shown. All associations are shown for the risk alleles (effect alleles) in the entire GC sample (see table 1 in the main manuscript). *P*-values, odds ratios (ORs) and the corresponding 95% confidence intervals (CIs) are shown.

SNP	Chromosome (position in bp (hg38))	Effect allele / other allele	Non-Cardia versus Cardia GC sample			Intestinal versus diffuse GC sample		
			P-value	OR	95% CI	*P*-value	OR	95% CI
rs760077	1q22 (155.208.991)	T/A	0.045	1.11	1.00 to 1.23	0.090	1.09	0.98 to 1.22
rs11677924	2q23 (29.500.326)	G/C	0.269	1.08	0.94 to 1.24	0.040	0.86	0.75 to 0.99
rs2920293	8q24 (142.683.996)	G/C	4.00E-06	1.30	1.16 to 1.45	0.012	1.15	1.03 to 1.28
rs17138478	17q12 (37.713.312)	C/A	0.921	0.99	0.84 to 1.16	4.38E-06	0.68	0.58 to 0.80

Supplementary Table 6: Top 10 PheWAS results **(a)** and overlap with lead variants of previously published GWAS studies **(b)** for rs760077, being the lead variant of locus 1q22 in the GC-GWAS.

a) PheWAS

Trait	*P*-value	Effect (beta)	Author (Year)	Number of Cases	Number overall
Haematocrit percentage	5.81E-45	0.099	UKB Neale v2	NA	350475
Haemoglobin concentration	6.91E-35	0.029	UKB Neale v2	NA	350474
Red blood cell (erythrocyte) count	2.79E-27	0.009	UKB Neale v2	NA	350475
Benign neoplasm of other parts of digestive system	1.13E-17	0.173	UKB SAIGE	5280	400581
Impedance of leg (left)	7.17E-14	0.584	UKB Neale v2	NA	354811
Impedance of leg (right)	4.27E-13	0.566	UKB Neale v2	NA	354817
Leg fat-free mass (right)	4.96E-09	-0.017	UKB Neale v2	NA	354798
Leg predicted mass (right)	6.23E-09	-0.016	UKB Neale v2	NA	354798
None of the above | mouth/teeth dental problems	1.05E-08	-0.028	UKB Neale v2	141495	359841
Dentures | mouth/teeth dental problems	1.45E-08	0.035	UKB Neale v2	60977	359841

b) Overlap GWAS

Trait	Lead Variant	*P*-value	LD (r^2) to variant examined	Author (Year)	Study PMID
Serum cancer antigen 15.3 levels	rs760077	1.00E-300	1.00	Olafsson S	PMID:31666285
Blood urea nitrogen levels	rs760077	6.00E-126	1.00	Sakaue S	PMID:34594039
Hematocrit	rs760077	1.00E-82	1.00	Chen MH	PMID:32888493
Hematocrit	rs6676150	8.80E-77	0.88	Chen MH	PMID:32888493
Blood urea nitrogen levels	rs760077	2.00E-67	1.00	Wuttke M	PMID:31152163
Hematocrit	rs6676150	3.40E-61	0.88	Vuckovic D	PMID:32888494
Hemoglobin concentration	rs760077	6.00E-59	1.00	Chen MH	PMID:32888493
Hemoglobin concentration	rs6676150	1.80E-54	0.88	Chen MH	PMID:32888493
Hematocrit	rs760077	2.00E-53	1.00	Sakaue S	PMID:34594039
Red blood cell count	rs760077	3.00E-50	1.00	Chen MH	PMID:32888493

Supplementary Table 7: Top 10 PheWAS results **(a)** and overlap with lead variants of previously published GWAS studies **(b)** for rs2920293, being the lead variant of locus 8q24 in the GC-GWAS.

a) PheWAS

Trait	*P*-value	Effect (beta)	Author (Year)	Number of Cases	Number overall
Cancer of bladder	2.25E-11	0.20	UKB SAIGE	2427	407223
Duodenitis	3.59E-11	-0.11	UKB SAIGE	7655	385779
Malignant neoplasm of bladder	6.07E-11	0.20	UKB SAIGE	2146	406942
Ankle spacing width	7.81E-11	-0.08	UKB Neale v2	NA	206589
Duodenal ulcer	1.46E-10	-0.17	UKB SAIGE	3002	404527
Gastroduodenal ulcer	5.30E-10	-0.14	FINNGEN_R5	4510	194205
Duodenal ulcer	9.20E-09	-0.20	FINNGEN_R5	1691	191386
Malignant neoplasm of stomach (all cancers excluded)	1.65E-08	0.32	FINNGEN_R5	633	174639
Trunk fat mass	2.05E-08	-0.07	UKB Neale v2	NA	354597
Malignant neoplasm of stomach	3.17E-08	0.31	FINNGEN_R5	633	218792

b) Overlap GWAS

Trait	Lead Variant	*P*-value	LD (r^2) to variant examined	Author (Year)	Study PMID
Gastric cancer	rs2920280	4.00E-49	0.88	Sakaue S	PMID:34594039
Gastric cancer	rs2294008	1.00E-44	1.00	Tanikawa C	PMID:30281874
Gastric cancer	rs2978977	3.00E-43	0.57	Ishigaki K	PMID:32514122
Duodenal ulcer [Additive]	rs2294008	2.00E-33	1.00	Tanikawa C	PMID:22387998
Pepsinogen I/II ratio	rs2920283	4.00E-27	0.99	Hishida A	PMID:30753327
Gastric ulcer	rs35464379	3.00E-25	0.88	Sakaue S	PMID:34594039
Gastric ulcer	rs2976397	6.00E-24	0.56	Sakaue S	PMID:34594039
Peptic ulcer disease	rs2976388	2.00E-14	0.92	Wu Y	PMID:33608531
Severe gastric atrophy	rs2920283	2.00E-13	0.99	Hishida A	PMID:30753327
Gastric cancer	rs2976394	2.00E-13	1.00	Park B	PMID:30189721

Supplementary Table 8: Top 10 PheWAS results **(a)** and overlap with lead variants of previously published GWAS studies **(b)** for rs17138478, being the lead variant of locus 17q12 in the GC-GWAS.

a) PheWAS

Trait	*P*-value	Effect (beta)	Author (Year)	Number of Cases	Number overall
Prostate cancer	5.92E-24	0.12	Schumacher FR	79148	140254
Liver enzyme levels (alanine transaminase)	4.90E-23	-0.01	Pazoki R	NA	437267
Sex hormone-binding globulin levels adjusted for BMI	3.30E-14	0.01	Ruth KS	NA	368929
Sex hormone-binding globulin levels	4.90E-11	0.02	Barton AR	NA	397043
Sex hormone-binding globulin levels	8.20E-09	0.01	Ruth KS	NA	370125
Cholelithiasis and cholecystitis	3.21E-08	0.10	UKB SAIGE	16225	407532
Cholelithiasis	6.50E-08	0.10	UKB SAIGE	13777	405084
Sex hormone-binding globulin levels adjusted for BMI	2.20E-07	0.01	Ruth KS	NA	188908
Sex hormone-binding globulin levels adjusted for BMI	2.50E-07	0.01	Ruth KS	NA	180094
Cholelithiasis	5.04E-07	0.08	FINNGEN_R5	19023	214167

b) Overlap GWAS

Trait	Lead Variant	*P*-value	LD (r^2) to variant examined	Author (Year)	Study PMID
C-reactive protein	rs17138478	2.00E-25	1.00	Sakaue S	PMID:34594039
Liver enzyme levels (alanine transaminase)	rs17138478	4.90E-23	1.00	Pazoki R	PMID:33972514
Alanine aminotransferase levels	rs17138478	2.00E-22	1.00	Ward LD	PMID:34315874
C-reactive protein levels	rs17138478	3.00E-20	1.00	Sinnott-Armstrong N	PMID:33462484
Alanine aminotransferase levels	rs17138478	5.00E-18	1.00	Sinnott-Armstrong N	PMID:33462484
Alanine aminotransferase levels	rs17138478	2.00E-15	1.00	Chen VL	PMID:33547301
Aspartate aminotransferase to alanine aminotransferase ratio	rs17138478	4.00E-15	1.00	Sinnott-Armstrong N	PMID:33462484
Alanine aminotransferase levels	rs17138478	2.00E-14	1.00	Sakaue S	PMID:34594039
Sex hormone-binding globulin levels adjusted for BMI	rs17138478	3.30E-14	1.00	Ruth KS	PMID:32042192
Cholelithiasis	rs17138478	1.00E-11	1.00	Sakaue S	PMID:34594039

Supplementary Table 9: Top 10 PheWAS results **(a)** and overlap with lead variants of previously published GWAS studies **(b)** for rs2590943, being the lead variant of locus 1p31 in the GC-GWAS.

a) PheWAS

Trait	*P*-value	Effect (beta)	Author (Year)	Number of Cases	Number overall
Comparative body size at age 10	2.79E-24	0.02	UKB Neale v2	NA	354996
Leg fat percentage (left)	1.31E-21	0.15	UKB Neale v2	NA	354791
Body mass index (bmi)	3.47E-21	0.13	UKB Neale v2	NA	359983
Body mass index (bmi)	7.06E-21	0.13	UKB Neale v2	NA	354831
Leg fat mass (left)	2.92E-20	0.04	UKB Neale v2	NA	354788
Leg fat mass (right)	2.90E-19	0.04	UKB Neale v2	NA	354807
Weight	8.13E-19	0.37	UKB Neale v2	NA	360116
Leg fat percentage (right)	9.95E-19	0.14	UKB Neale v2	NA	354811
Weight	2.32E-18	0.36	UKB Neale v2	NA	354838
Whole body fat mass	7.86E-18	0.23	UKB Neale v2	NA	354244

b) Overlap GWAS

Trait	Lead Variant	*P*-value	LD (r^2) to variant examined	Author (Year)	Study PMID
Body mass index	rs1993709	1.00E-52	0.84	Pulit SL	PMID:30239722
Body mass index	rs2613498	4.00E-40	0.85	Kichaev G	PMID:30595370
Body mass index	rs61765651	2.00E-38	0.86	Zhu Z	PMID:31669095
Leg fat percentage (left)	rs34361149	1.25E-27	0.85	UKB Neale v2	NA
Body mass index (bmi)	rs34361149	3.34E-27	0.85	UKB Neale v2	NA
Body mass index (bmi)	rs1460940	6.95E-27	0.82	UKB Neale v2	NA
Adult body size	rs2613499	1.00E-26	0.85	Richardson TG	PMID:32376654
Body mass index	rs34361149	2.00E-26	0.80	Sakaue S	PMID:34594039
Leg fat mass (left)	rs34361149	2.66E-26	0.85	UKB Neale v2	NA
Leg fat mass (right)	rs34361149	2.39E-25	0.85	UKB Neale v2	PMID:30239722

Supplementary Table 10: Top 10 PheWAS results **(a)** and overlap with lead variants of previously published GWAS studies **(b)** for rs532436, being the lead variant of locus 9q34 in the GC-GWAS.

a) PheWAS

Trait	*P*-value	Effect (beta)	Author (Year)	Number of Cases	Number overall
Haemoglobin concentration	1.94E-141	-0.0761949	UKB Neale v2		350474
Red blood cell (erythrocyte) count	2.85E-121	-0.0252011	UKB Neale v2		350475
Haematocrit percentage	5.73E-121	-0.207538	UKB Neale v2		350475
Blood clot in the leg	3.22E-64	0.36287686	UKB Neale v2	7386	360527
Deep venous thrombosis	5.23E-64	0.36588362	UKB Neale v2	7237	361141
Phlebitis and thrombophlebitis of lower extremities	4.94E-47	0.465	UKB SAIGE	3587	373179
Phlebitis and thrombophlebitis	2.43E-45	0.437	UKB SAIGE	3900	373492
Pulmonary embolism	9.94E-44	0.46440584	UKB Neale v2	2999	361141
Blood clot in the lung	1.70E-42	0.45870702	UKB Neale v2	2984	360527
Pulmonary heart disease	5.08E-36	0.367	UKB SAIGE	4257	406632

b) Overlap GWAS

Trait	Lead Variant	*P*-value	LD (r^2) to variant examined	Author (Year)	Study PMID
vWF levels	rs8176685	4.49E-324	0.98	Sabater-Lleal M	PMID:30586737
Serum alkaline phosphatase levels	rs2519093	4.49E-324	1.00	Kanai M	PMID:29403010
Serum alkaline phosphatase levels	rs507666	4.49E-324	1.00	Sinnott-Armstrong N	PMID:33462484
E-selectin levels	rs8176643	2.20E-308	0.99	Folkersen L	PMID:33067605
Serum 25-Hydroxyvitamin D levels	rs115478735	2.20E-308	0.99	Manousaki D	PMID:32059762
Sulfhydryl oxidase 2 measurement	rs115478735	2.20E-308	0.99	Pietzner M	PMID:33328453
E-selectin levels	rs11244061	2.20E-308	0.55	Folkersen L	PMID:33067605
Soluble E-selectin levels	rs2519093	4.00E-305	0.99	Sliz E	PMID:31217265
Platelet endothelial cell adhesion molecule levels	rs8176643	7.89E-302	0.99	Folkersen L	PMID:33067605
Red blood cell count	rs550057	1.00E-296	0.97	Chen MH	PMID:32888493

Supplementary Table 11: Top 10 results of a pathway analyses using the Enrichr tool and the pathway database BioPlanet 2019 as source [89]. DE genes comparing transcriptome data of tissue samples taken from the **a)** antrum (378 genes) and the **b)** corpus (357 genes) were used as input.

a) Top 10 pathways upregulated in antrum.

Pathway	***P*-value**	**Adjusted *P*-value**	**Odds Ratio**	**Combined score**
Interferon signaling	7.4E-10	6.0E-07	6.73	141.42
Immune system signaling by interferons, interleukins, prolactin, and growth hormones	6.8E-09	2.8E-06	4.75	89.28
Interferon alpha/beta signaling	3.8E-08	1.0E-05	10.77	184.08
TGF-beta regulation of extracellular matrix	6.9E-08	1.4E-05	3.22	53.06
Oncostatin M	2.0E-07	3.3E-05	4.02	61.9
Drug metabolism: cytochrome P450	4.6E-06	6.3E-04	7.08	86.96
Metapathway biotransformation	8.0E-06	9.3E-04	4.55	53.42
Type I hemidesmosome assembly	1.6E-05	1.6E-03	40.85	450.4
Biological oxidations	1.7E-05	1.6E-03	4.9	53.64
TAp63 pathway	8.9E-05	7.3E-03	7.49	69.83

b) Top 10 pathways upregulated in corpus.

Pathway	***P*-value**	**Adjusted *P*-value**	**Odds Ratio**	**Combined score**
Branched-chain amino acid catabolism	9.1E-09	4.8E-06	39.27	727.1
Amino acid metabolism	8.5E-08	2.2E-05	5.47	89.02
Valine, leucine and isoleucine degradation	9.6E-07	1.7E-04	12.48	172.94
Ghrelin-mediated regulation of food intake and energy homeostasis	2.0E-06	2.7E-04	34.86	457.21
Creatine metabolism	3.3E-06	3.5E-04	74.18	935.22
Metabolism	2.3E-05	2.1E-03	1.97	21.02
Response to elevated platelet cytosolic calcium	1.2E-04	8.8E-03	5.98	54.13
Amino acid biosynthesis and interconversion (transamination)	1.5E-04	1.0E-02	18.54	162.82
Insulin-like growth factor (IGF) activity regulation by insulin-like growth factor binding proteins (IGFBPs)	2.0E-04	1.2E-02	17.11	145.94
Gastric acid secretion	3.5E-04	1.9E-02	5.84	46.48

Supplementary Table 12: Genetic correlations using LD Score regression between GC and 20 traits belonging to five phenotype-categories that represent risk factors for GC development. For each trait the genetic GC correlation (rg) along with the standard error (SE) and the corresponding significance level (*P*-value) are shown.

Phenotype category	Trait	r_g	SE	Z	*P*-value
Obesity-related traits	Body mass index (BMI)	0.3031	0.0881	3.4424	6.0E-04
	Hip circumference	0.2692	0.0884	3.0458	2.3E-03
	Weight	0.2628	0.0865	3.04	2.4E-03
	Waist circumference	0.2368	0.0846	2.7978	5.1E-03
Reflux-related traits	ICD10 K21 (Gastro-esophageal reflux disease (GERD))	0.0383	0.1764	0.217	0.828
	Self-reported: gastro-esophageal reflux/gastric reflux	0.0016	0.1541	0.0101	0.991
Smoking-related traits	Pack years adult smoking	0.3521	0.1133	3.1064	2.0E-03
	Pack years of smoking	0.3281	0.1139	2.8799	4.0E-03
	Number of cigarettes previously smoked daily	0.3425	0.1212	2.8251	4.7E-03
	Current tobacco smoking	0.2406	0.0902	2.6685	7.6E-03
	Number of cigarettes currently smoked daily	0.3102	0.2381	1.3028	0.193
Alcohol intake-related traits	Alcohol intake 10 years previously	0.3615	0.1191	3.0339	2.0E-03
	Alcohol intake frequency	0.2807	0.0943	2.9772	3.0E-03
	Average weekly red wine intake	-0.2327	0.1018	-2.2853	0.022
	Average weekly champagne plus white wine intake	-0.2353	0.1153	-2.0413	0.041
	Average weekly intake of other alcoholic drinks	0.0447	0.3498	0.1279	0.898
Education-/employment-related traits	Qualifications: O levels/GCSEs or equivalent (a)	-0.343	0.1167	-2.9405	3.3E-03
	Qualifications: College or University degree	-0.183	0.0745	-2.4563	0.014
	Current employment status: In paid employment or self-employed	-0.4335	0.2086	-2.0783	0.036
	Age completed full time education	-0.1807	0.0921	-1.9631	0.049

(a) GCSE refers to General Certificate of Secondary Education

Supplementary Table 13: Genetic relation between cardia GC and OAC. PRS for OAC were calculated using different *P*-value thresholds and different numbers of SNPs, respectively. The proportion of variance explained by each PRS was then tested in four GC target samples. *P*-values indicate, whether PRS derived from OAC are associated to the target samples.

P-value threshold	Number of SNPs	Target sample	*P*-value
5.00E-08	10	cardia GC	0.109
		entire GC	0.165
		non-cardia GC	0.761
		cardia versus non-cardia GC	0.229
1.00E-06	28	cardia GC	0.001
		entire GC	0.021
		non-cardia GC	0.960
		cardia versus non-cardia GC	0.020
1.00E-04	287	cardia GC	3.75E-07
		entire GC	0.01
		non-cardia GC	0.4
		cardia versus non-cardia GC	4.30E-04
1.00E-03	1,602	cardia GC	2.37E-08
		entire GC	0.003
		non-cardia GC	0.151
		cardia versus non-cardia GC	0.008
0.01	9,878	cardia GC	2.65E-06
		entire GC	0.009
		non-cardia GC	0.187
		cardia versus non-cardia GC	0.005
0.05	34,794	cardia GC	4.40E-04
		entire GC	0.047
		non-cardia GC	0.277
		cardia versus non-cardia GC	1.42E-04
0.1	58,522	cardia GC	0.683
		entire GC	0.012
		non-cardia GC	0.168
		cardia versus non-cardia GC	3.30E-08
0.2	96,776	cardia GC	0.069
		entire GC	0.009
		non-cardia GC	0.147
		cardia versus non-cardia GC	3.49E-09
0.5	176,876	cardia GC	0.095
		entire GC	0.005
		non-cardia GC	0.085
		cardia versus non-cardia GC	2.18E-09
1	244,252	cardia GC	0.021
		entire GC	0.005
		non-cardia GC	0.089
		cardia versus non-cardia GC	1.85E-08

Supplementary Table 14: Genetic relation between cardia GC and OAC/BO. PRS for OAC/BO were calculated using different *P*-value thresholds and different numbers of SNPs, respectively. The proportion of variance explained by each PRS was then tested in four GC target samples. *P*-values indicate, whether PRS derived from OAC/BO are associated to the target samples.

P-value threshold	Number of SNPs	Target sample	*P*-value
5.00E-08	17	cardia GC	2.88E-06
		entire GC	0.045
		non-cardia GC	0.743
		cardia versus non-cardia GC	1.47E-03
1.00E-06	46	cardia GC	8.84E-04
		entire GC	0.096
		non-cardia GC	0.920
		cardia versus non-cardia GC	0.030
1.00E-04	354	cardia GC	0.135
		entire GC	0.202
		non-cardia GC	0.792
		cardia versus non-cardia GC	0.064
1.00E-03	1,812	cardia GC	0.018
		entire GC	0.312
		non-cardia GC	0.864
		cardia versus non-cardia GC	0.012
0.01	10,431	cardia GC	5.20E-08
		entire GC	0.077
		non-cardia GC	0.701
		cardia versus non-cardia GC	2.30E-03
0.05	35,81	cardia GC	4.02E-05
		entire GC	0.205
		non-cardia GC	0.678
		cardia versus non-cardia GC	3.51E-05
0.1	60,117	cardia GC	9.05E-16
		entire GC	0.140
		non-cardia GC	0.778
		cardia versus non-cardia GC	7.20E-07
0.2	98,276	cardia GC	2.79E-17
		entire GC	0.082
		non-cardia GC	0.627
		cardia versus non-cardia GC	4.33E-07
0.5	177,268	cardia GC	1.13E-16
		entire GC	0.062
		non-cardia GC	0.469
		cardia versus non-cardia GC	3.50E-06
1	244,383	cardia GC	4.52E-17
		entire GC	0.067
		non-cardia GC	0.535
		cardia versus non-cardia GC	1.99E-06

Supplementary Table 15: Lead associations of genome-wide significant loci for oesophago-gastric adenocarcinoma. The associations are shown for the risk alleles (effect alleles) in the combined cardia GC/OAC/BO sample. *P*-values, odds ratios (ORs) and the corresponding 95% confidence intervals (CIs) are shown. The heterogeneity *P*-values (Het *P*-value) indicate whether the observed associations are equally attributable to cardia GC and OAC/BO. Allele frequencies for the associated SNPs among patients and controls are not given, as the GWAS samples were meta-analysed. Instead the frequency of effect alleles in the European population are shown according to gnomAD [52].

SNP	Chromosome (position in bp (hg38))	Effect allele / other allele (a)	*P*-value	OR	95% CI	Het *P*-value
rs7255	2p24 (20.679.060)	T/C	2.46E-12	1.13	1.09 to 1.17	0.910
rs896350	2q33 (199.163.981)	G/A	6.66E-09	1.11	1.07 to 1.14	0.590
rs2687202	3p13 (70.880.832)	T/C	6.91E-09	1.11	1.07 to 1.15	0.110
rs3749615	5p15 (601.370)	T/C	3.43E-10	1.14	1.09 to 1.19	0.864
rs9257809	6p22 (29.388.554)	A/G	3.68E-09	1.21	1.13 to 1.28	0.267
rs62423175	6q11 (61.485.463)	A/G	2.81E-09	1.16	1.11 to 1.22	1.000
rs73014164	6q25 (160.555.588)	C/T	3.49E-08	1.13	1.08 to 1.18	0.932
rs17451754	7q31 (117.616.658)	G/A	1.13E-11	1.18	1.12 to 1.24	0.988
rs4382480	8p23 (8.863.963)	A/G	1.04E-08	1.10	1.06 to 1.14	0.684
rs28630503	8p23 (10.151.506)	T/C	1.36E-08	1.11	1.07 to 1.15	0.202
rs1478892	8p23 (11.591.020)	T/G	6.07E-09	1.11	1.07 to 1.14	0.427
rs1817002	8q21 (75.653.577)	G/A	4.10E-08	1.11	1.06 to 1.14	0.907
rs7852462	9q22 (97.548.219)	C/T	2.39E-08	1.11	1.06 to 1.14	0.072
rs2464469	15q21 (58.069.827)	G/A	9.39E-11	1.12	1.08 to 1.15	0.214
rs234506	15q26 (97.035.863)	A/G	1.56E-09	1.12	1.07 to 1.16	0.815
rs2353694	16q24 (86.431.413)	G/C	2.80E-08	1.12	1.07 to 1.14	0.584
rs10404726	19p13 (18.723.704)	C/T	3.94E-09	1.11	1.07 to 1.14	0.198

(a) Frequency of effect alleles in the European (non-Finnish) population according to gnomAD [52]: rs7255 allele T 49%, rs896350 allele G 60%, rs2687202 allele T 32%, rs3749615 allele T 17%, rs9257809 allele A 91%, rs62423175 allele A 17%, rs73014164 allele C 83%, rs17451754 allele G 84%, rs4382480 allele A 44%, rs28630503 allele T 29%, rs1478892 allele T 34%, rs1817002 allele G 64%, rs7852462 allele C 60%, rs2464469 allele G 42%, rs234506 allele A 28%, rs2353694 allele G 25%, rs10404726 allele C 54%.

Acknowledgement

My doctoral thesis was one of many projects I was involved in during my time at the Institute of Human Genetics at the University Hospital Bonn and the Centre for Human Genetics at the University Hospital Marburg. The one constant that connects all the projects, people, and places I was privileged to encounter over the years, is Prof. Dr. Johannes Schumacher. He provided guidance and support during scientifically interesting, but also during professionally and personally challenging times. He showed me that there is always a solution, however unconventional it may be. I deeply respect him as a mentor, person and friend.

I would like to thank Prof. Dr. Markus M. Nöthen, who kindly agreed to take on the responsibility as first supervisor and who made this study possible by providing an excellent infrastructure, supporting and promoting my professional development, and by always sparing a few arrays.

I would like to thank Prof. Dr. Walter Witke for taking on the responsibility as second supervisor. I would like to thank Prof. Dr. Michael Pankratz and Prof. Dr. Frank Kurth for being part of the doctoral committee and for their straightforward commitment.

A lot of cooperation partners were involved in the collection and processing of samples and data. I would like to thank all people who spent time and effort, especially in the context of the staR project and the gastric transcriptome project.

At the Institutes of Human Genetics in Bonn and Marburg, I would like to thank all colleagues helping to extract, prepare, genotype and sequence the samples, especially Dr. Jessica Trautmann, Nadine Fricker, Sandra Barth, Bärbel Lippke, Peter Teßmann, Annegret Reinscheid, Sandra Mürb, Laura Köbbe and Elaine Gurich-Hahn. In addition, I would like to thank my fellow sufferers and PhD students, who became good friends and with who I share many good memories.

Furthermore, I would like to thank PD. Dr. Michael Knapp, Vitalia Schüller, Jan Gehlen, Dr. Carlo Maj and Dr. Oleg Borisov for their great support in the statistical and bioinformatic analyses.

I would like to express my special gratitude and respect to all people who consented to take part in this study. When handling piles of vials, in order to generate gigabytes of anonymised data, one can easily forget that behind each sample is an individual life and destiny. No participant could expect a personal benefit, and yet the motivation was high to contribute. I hope that the generated datasets and the upcoming projects will make the trust and the efforts worthwhile.

I want to sincerely thank my family (especially Marion, Sven, Dietmar, Helga and Silke) and my friends for their support and understanding. I apologise for all the times when they came second when they should have come first and I am deeply grateful for having so many people in my life that I can fully trust and rely on.

Last but not least, I want to thank my wife Anne who had to endure the most. She always supported me and I owe her a lot more, than I can possibly give back. I love you, I am grateful that we found each other, and I am looking forward to a future with many projects and adventures that finally will be unrelated to this thesis.

– Dedicated to my Father –