Integrating genetics and transcriptomics to decipher susceptibility genes for risk stratification of gastric cancer and effect modification of *Helicobacter pylori* treatment

Zhou-Yi Yin,[a] Heng-Min Xu,[a] Meng-Yuan Wang,[a] Xin-Ling Wang,[a] Zong-Chao Liu,[a] Yu Jin,[a] Yang Zhang,[b] Jing-Ying Zhang,[b] Tong Zhou,[b] Wei-Cheng You,[b] Kai-Feng Pan,[a,**] and Wen-Qing Li[a,*]

[a]State Key Laboratory of Holistic Integrative Management of Gastrointestinal Cancers, Department of Cancer Epidemiology, Peking University Cancer Hospital & Institute, Beijing, 100142, China
[b]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Cancer Epidemiology, Peking University Cancer Hospital & Institute, Beijing, 100142, China

eBioMedicine
2025;116: 105767
Published Online xxx
https://doi.org/10.1016/j.ebiom.2025.105767

Summary

Background In a transcriptome-wide association study, we deciphered susceptibility genes that may predict gastric cancer (GC) risk and modify the effects of *Helicobacter pylori (H. pylori)* treatment.

Methods Genetically predicted expression models of 4518 genes were developed based on the GTEx and applied to a nested case–control study (935 GCs and 1869 controls) of the Mass Intervention Trial in Linqu, Shandong Province (MITS), with genes associated with GC risk further validated in BioBank Japan (7921 GCs and 159,201 controls). Transcriptome risk scores (TRSs) integrating key genes were constructed, utilizing imputed transcriptomes from the Shandong Intervention Trial (SIT) and UK Biobank, and observed transcriptomes from the National Upper Gastrointestinal Cancer Early Detection (UGCED) program. We also examined whether TRS may modify the association of *H. pylori* infection and anti-*H. pylori* treatment with GC risk.

Findings Integrating 11 independent GC-associated genes identified based on the MITS (FDR-q < 0.05) and BioBank Japan ($P < 0.05$), the TRS demonstrated a dose-dependent association with an elevated risk of incident GC in both the SIT (P-trend = 0.003) and UK Biobank (P-trend = 0.008), and exhibited an upward trend as gastric lesions progressed based on the UGCED program (P-trend = 5.01×10^{-4}). In the SIT, the increased risk of GC associated with *H. pylori* infection (P-interaction = 0.03) and beneficial effect of successful *H. pylori* eradication (P-interaction = 0.05) were significant for individuals with high TRSs.

Interpretation We identified a gene panel which may predict GC risk across populations of multiple ancestries, which offers important insights into GC risk stratification and presents a precision approach to primary prevention.

Funding Funders are listed in the Acknowledgement.

Keywords: Transcriptome-wide association study; Transcriptome risk score; Gastric cancer; Risk stratification; Primary prevention; *Helicobacter pylori*

Introduction

Gastric cancer (GC) ranks fifth globally in terms of both cancer-related incidence and mortality, with more than half of cases with GC worldwide occurring in East Asia, predominantly in China.[1] *Helicobacter pylori (H. pylori)* is the most well-established risk factor inducing the progression of gastric lesions and development of GC, which infects more than 40% of the world's population.[2,3] The Shandong Intervention Trial (SIT) conducted by our team has provided early evidence supporting *H. pylori* eradication as a means to eliminate GC,[4–6] while the subsequent Mass Intervention Trial in Linqu, Shandong Province (MITS) further reinforces the effectiveness and overall feasibility of

*Corresponding author. State Key Laboratory of Holistic Integrative Management of Gastrointestinal Cancers, Department of Cancer Epidemiology, Peking University Cancer Hospital & Institute, 52 Fu-cheng Road, Haidian District, Beijing, 100142, China.
**Corresponding author. State Key Laboratory of Holistic Integrative Management of Gastrointestinal Cancers, Department of Cancer Epidemiology, Peking University Cancer Hospital & Institute, 52 Fu-cheng Road, Haidian District, Beijing, 100142, China.
E-mail addresses: wenqing_li@bjmu.edu.cn (W.-Q. Li), pan-kf@263.net (K.-F. Pan).

Research in context

Evidence before this study

Gastric cancer remains a major public health threat worldwide. Our capacity to discern potential high-risk populations and those who stand to gain the most from preventive interventions is yet limited. Developing a generalizable tool which may predict gastric cancer risk across populations of multiple ancestries is of great public health significance.

Added value of this study

Our two-stage transcriptome-wide association study in East Asia identified a panel of genes associated with the risk of developing gastric cancer. Integrating 11 independent genes, the derived transcriptome risk score can predict gastric cancer risk in independent, cross-ancestry prospective cohorts. The associations of *Helicobacter pylori* infection and successful *H. pylori* eradication with gastric cancer risk were prominent for individuals with a high transcriptome risk score (top quintile).

Implications of all the available evidence

This study provides insights into the genetic architecture of gastric cancer and represents an important attempt to construct a risk score for assessing gastric cancer risk by integrating susceptibility genes identified through transcriptome-wide association study. The transcriptome risk score defined in this study offers a generalizable tool for gastric cancer risk stratification and may potentially serve as a new avenue for optimized primary prevention of gastric cancer.

H. pylori treatment.[7] Alongside primary prevention, endoscopic screening-based secondary prevention can significantly lower the risks of GC incidence and mortality.[8] Nevertheless, our comprehension of the pivotal molecular mechanisms underlying the progression of gastric lesions and the onset of GC remains constrained. Additionally, our capacity to discern potential high-risk populations and those who stand to gain the most from preventive interventions is limited, leading to GC prevention strategies that fall short of the standards set by the precision public health paradigm.[9]

The occurrence of GC is intricately linked to genetic susceptibility.[10] Previous genome-wide association studies (GWASs) have unveiled a number of genetic loci associated with GC,[11–15] and the integrated polygenic risk scores (PRSs) have exhibited promising potential in distinguishing high-risk populations and predicting GC risk,[12,13] offering insights into the formulation of tailored prevention strategies that cater to individual risk and benefit profiles.[13] However, most of the identified genetic loci are in non-coding regions of the genome, providing limited genetic explanatory power for GC risk; advanced biological knowledge of the underlying mechanisms of GC and the cross-ancestry portability of PRS remain challenging.[16–18]

Drawn upon the concept that a multitude of genetic variants exert influence on complex traits by modulating gene expression,[19–21] transcriptome-wide association studies (TWASs) assess the association between genetically imputed gene expression within specific tissues and complex traits,[22,23] thereby facilitating the identification of susceptibility genes and providing deeper understanding into the underlying molecular mechanisms.[24] The aggregation of TWAS-identified gene expression into a transcriptional risk score (TRS) has exhibited notable advantages in risk stratification across diverse populations, either constructed utilizing genetically predicted[25–28] or observed gene expression data.[28–30] However, few TWASs is available on GC,[15] and none of them have yet established a TRS for assessing GC risk or evaluating the effectiveness of prevention strategies for individuals with varying TRS levels.

In our study, we conducted a two-stage TWAS to identify GC susceptibility genes in East Asia and then utilized independent genes to construct TRS in gastric tissues to investigate their association with GC risk, using imputed transcriptomes or observed transcriptomes from our team and publicly accessible databases of cross-ancestry. To further enhance the public health implications, we examined how TRS may modify the association of *H. pylori* infection and anti-*H. pylori* treatment on risk of GC.

Methods

Overall study design and participants

The overall study design is illustrated in Fig. 1, which involves three in-house cohorts, including the MITS, SIT, and National Upper Gastrointestinal Cancer Early Detection (UGCED) program enrolled in Linqu county, Shandong Province, a recognized high-risk area for GC.[6,7,31] All participants are of Chinese ethnicity. There were no overlapping participants across three cohorts. Additionally, the study utilizes three publicly accessible databases, including the Genotype-Tissue Expression (GTEx, v8), the UK Biobank (UKB), as well as the BioBank Japan (BBJ) project.

GTEx

The GTEx project (https://gtexportal.org/) was launched in 2010, aiming to characterize the relationship between genetic variations and gene expression across different human tissues. Details of the project have been described

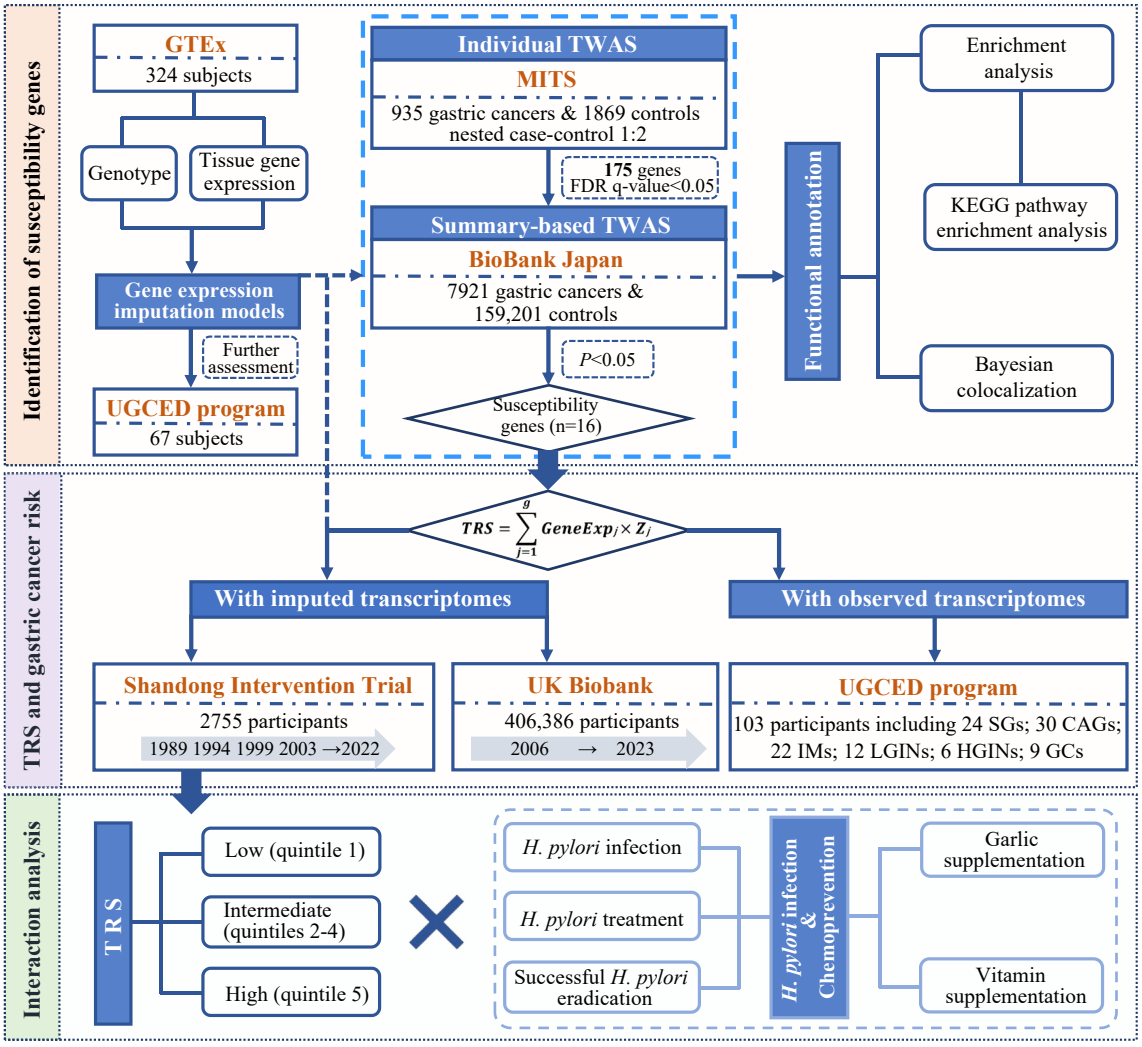


***Fig. 1:* Study outline of work flow**. Abbreviations: CAG, chronic atrophic gastritis; FDR, false discovery rate; GTEx, Genotype-Tissue Expression project; HGIN, high-grade intraepithelial neoplasia; *H. pylori*, *Helicobacter pylori*; IM, intestinal metaplasia; KEGG, Kyoto Encyclopedia of Genes and Genomes; LGIN, low-grade intraepithelial neoplasia; MITS, Mass Intervention Trial in Linqu, Shandong Province; SG, superficial gastritis; TRS, transcriptional risk score; TWAS, transcriptome-wide association study; UGCED, Upper Gastrointestinal Cancer Early Detection.

previously.[21,32] The gene expression imputation model used in this study was constructed based on 324 individuals with both whole-blood genotype and RNA sequencing (RNA-seq) data of gastric tissues. Study subjects were free of GC or other tumours. Over 80% were of Whites ancestry. The generation of the expression model was performed by the Mancuso laboratory (https://www.mancusolab.com/) and is stored on the FUSION website (http://gusevlab.org/projects/fusion/).

National UGCED program in Linqu, Shandong Province of China

Within the framework of the UGCED program, approximately 3500 residents in Linqu County aged 40–69 years receiving endoscopic examinations for free per year. Details have been described previously.[8,33,34] Information on age and sex were confirmed based on the national identity card. Biopsy samples taken from seven standard sites and reviewed following the criteria proposed by the Chinese Association of Gastric Cancer and Updated Sydney System.[35] From April 21, 2012 to May 07, 2019, a total of 103 subjects aged 40–69 years were randomly selected for the assessment of TRS based on observed transcriptomes, diagnosed with superficial gastritis (SG, n = 24), chronic atrophic gastritis (CAG, n = 30), intestinal metaplasia (IM, n = 22), low-grade intraepithelial neoplasia (LGIN, n = 12), high-grade intraepithelial neoplasia (HGIN, n = 6), or GC (n = 9). Among them, 67 subjects who were free of GC and had both genotype and RNA-seq data were also used to assess the predictive stability of imputation weights from GTEx.

MITS

The MITS is a community-based, cluster-randomized, controlled, superiority intervention trial (no. ChiCTR-TRC-10000979) to test the GC prevention effect of anti-*H. pylori* therapy launched in March 2011. A total of 180,284 eligible individuals in 980 villages were enrolled. Age and sex were extracted from the national identity card. Information on smoking, alcohol intake and dietary habits was collected through standardized questionnaires at cohort enrolment. Blood samples were collected for all individuals. Details of MITS were described previously.[7] A total of 1035 incident cases with GC were identified over 11.8 years of follow-up (updated to December 31, 2022). The current study employed a nested case–control design, wherein newly diagnosed patients with GC identified during the follow-up period served as the case group, while controls were randomly selected in a 1:2 ratio. Finally, a total of 935 cases with GC and 1869 healthy controls, which had qualified genotype data, were included.

BBJ with GWAS summary data

The BBJ is a nation-wide hospital-based prospective project launched in 2003, which collaboratively gathered serum samples along with clinical data of approximately 200,000 participants from 66 hospitals affiliated with 12 medical institutes. The project collected DNA samples from all participants at baseline and were genotyped with the Illumina HumanOmniExpressExome and HumanExome beadchip.[14] The GWASs were performed for each phenotype using the SAIGE software (v.0.37). We downloaded the summary statistics of their GWAS on GC and gastric adenocarcinoma (accessible at https://pheweb.jp/), which include data from 7921 GCs and 159,201 controls.

SIT

The SIT is a recognized trial for GC prevention on *H. pylori* treatment, garlic supplementation, and vitamin supplementation.[4–6] In 1989, we initiated a 5-year prospective follow-up study involving scheduled endoscopic examinations among 3386 residents in Linqu County, Shandong Province.[31] In 1994, eligible individuals and additional endoscopically screened residents aged 35–64 were invited, with a total of 3365 individuals free of GC or other cancers included in a double-blinded, randomized, placebo-controlled trial conducted in 1995 (ClinicalTrials.gov, NCT00339768).[4–6] Of them, 2258 *H. pylori*-seropositive participants were randomly assigned to receive 3 interventions, including 2-week *H. pylori* treatment with amoxicillin and omeprazole and/or 7.3-year garlic supplementation and/or 7.3-year vitamin supplementation or placebo in a 2 × 2 × 2 factorial design. Meanwhile, 1107 *H. pylori*-seronegative participants were randomly assigned to receive garlic supplementation and/or vitamin supplementation or placebos in a 2 × 2 factorial design. Under this study design, each *H. pylori*-seropositive participant may receive up to 3 active interventions, while each *H. pylori*-seronegative participant may receive up to 2 active interventions, with each intervention being allocated evenly across the participants. 382 participants who had *H. pylori* treatment failure received another 2-week *H. pylori* retreatment. GC incidence was ascertained mainly from scheduled gastroscopies (in 1989, 1994, 1999 and 2003 for all participants) and cancer registries, with follow-up extended until August 31, 2022. For all participants, age and sex were extracted from the national identity card. Information on environmental and lifestyle factors was collected via standardized questionnaires at baseline. For the current study, a total of 2755 participants with qualified genotype data were included, of whom 2548 attended the intervention trial. Given that the official designation of SIT did not occur until the inception of the actual trial in 1995, we refer to participants of the whole study period (1989–2022) as the 'SIT cohort' for convenience.

UKB

Details of the UKB (https://www.ukbiobank.ac.uk/, Application No. 90999) including study design, procedures, genotyping and imputation were previously described.[36,37] Briefly, approximately 500,000 participants aged 40–69 years in the United Kingdom were recruited from 2006 to 2010. We excluded related individuals and those who had withdrawn their consent to participate, had no available genetic data, had a diagnosis of GC at baseline, reported a gender mismatch with genetic data, or had abnormal heterozygosity or high missingness rates. As a result, a total of 406,386 participants were included in the analysis. Information on age and sex was derived from National Health Service (NHS) records, while data on race/ethnicity, smoking, alcohol intake, and dietary variables were collected through touchscreen questionnaires. Outcomes of GC events in our study were ascertained through both the cancer registry data and the hospital inpatient records, with complete follow-up updated to September 1, 2023.

Overall study design

Our study utilized the pre-computed weights of single-nucleotide polymorphisms (SNPs) for the prediction of gene expression in gastric tissues based on the GTEx (v8, n = 324) project, the validity of which was further assessed using in-house data from the UGCED program in Linqu, Shandong Province (n = 67). A two-stage TWAS was then conducted to identify GC susceptibility genes based on a nested case–control study (n = 2804, including 935 GCs and 1869 controls) of the MITS, along with summary statistics of SNPs generated for 7921 GCs and 159,201 controls in BBJ. By integrating independent GC susceptibility genes identified in TWAS, a TRS was generated and then applied to

evaluate GC risk in the SIT (n = 2755) and UKB (n = 406,386) for cross-ancestry validation. We further examined the performance of the TRS in predicting GC risk using the RNA-seq data of 103 individuals with different gastric lesions or GC from the UGCED program. In addition, we examined whether the TRS would interact with *H. pylori* infection, anti-*H. pylori* treatment, and nutrition supplementation among participants of the SIT.

The diagnosis of GC was coded using the World Health Organization's International Classification of Diseases, Tenth Revision (ICD-10), classified under C16. The location of GC was classified based on the ICD-10, which included cardia gastric cancer (C16.0), non-cardia gastric cancer (C16.1–16.8), and unspecified gastric cancer (C16.9), while the pathologic type of gastric cancer was defined based on the ICD for Oncology, Third Edition (ICD-O-3).

RNA-seq and genome-wide screening

RNA-seq for the observed transcriptomic profile of 103 subjects was performed using PE100 strategy on DNBSEQ platform. Peripheral blood leucocyte DNA samples were genotyped using the Global Screening Array beadchip for UGCED and SIT participants, and the Asian Screening Array beadchip for MITS participants, followed by the standard quality control and imputation procedures described previously.[13] Genotyped variants were excluded if they had a call rate <95%, a *P*-value for Hardy–Weinberg Equilibrium $<1.0 \times 10^{-6}$ or a minor allele frequency (MAF)≤0.5% or were duplicated variants. Samples were removed if they had abnormal missing rate of variants (>5%) or heterozygosity rate (>±3 standard deviation from the mean), sex discrepancy, or were duplicated samples. For SNP imputation, we phased genotypes using SHAPEIT (v2) and performed imputation with IMPUTE2 (v2.3.1). Imputed variants were excluded if they had an imputation quality score INFO≤0.3, a MAF ≤ 0.5%, or a missing rate >10%.

For the UKB, participants were genotyped on the UK BiLEVE Axiom array and UK Biobank Axiom array.[37] Further quality control and adjustment and harmonization of forward and reverse strand and risk genotypes were performed for all study subjects. We then adopted a post-imputation quality control process consistent with the aforementioned.

Statistics

Gene expression imputation models for TWAS analysis

We utilized the SNP-weights for gastric gene expression downloaded from FUSION[23] (http://gusevlab.org/projects/fusion/), which were precomputed using genotype and gastric gene expression data of 324 healthy subjects in GTEx (v8). Briefly, the heritability of genes explained by *cis*-SNPs (SNPs located within ± 500 kb of the gene boundary) was calculated based solely on HapMap3 variants, and the further analysis was restricted to 6127 genes with a heritability *P*-value < 0.01. Then, the SNP-weights for the expression of each gene were estimated using four predictive linear models, and a five-fold cross-validation process was followed to determine the best-performing imputation model for each gene, yielding its cross-validation R^2. To assess the predictive stability of these imputation models across different populations, we calculated the R^2 value for prediction performance by comparing the imputed gene expression with the actual RNA-seq data of the 67 non-GCs included from the UGCED program. The final set of genes selected for the TWAS adhered to the criteria that were previously reported.[38] For genes that were detected by RNA-seq and had SNP data passing the quality control standards for the generation of imputed gene expression in UGCED participants, a prediction $R^2 > 0.01$ (corresponding to >10% correlation) in both GTEx and UGCED projects was considered eligible. Otherwise, for genes that could not be evaluated in UGCED (due to missing RNA-seq data or incomplete genotype data after quality control), we required a prediction $R^2 > 0.04$ (corresponding to >20% correlation) in the GTEx project. For the included genes, a Pearson correlation analysis was conducted to calculate the correlation coefficient between the prediction R^2 in the UGCED program and the prediction R^2 values obtained from GTEx.

Two-stage TWAS for GC

Using the predictive model developed for each gene based on GTEx, we employed the individual genotype data to impute predicted transcriptomes for the subjects in the nested case–control study of the MITS. Standardized effect sizes (Z_{TWAS}) and hazard ratios (HRs) with 95% confidence intervals (CIs) were then calculated for the associations between the imputed gene expression levels and incident GC risk. Instead of conventional Cox regression analysis, inverse probability weighted Cox regression models as appropriate for the nested case–control study,[39] were employed for the analysis of the MITS subjects, which accounts for selection bias to enhance the representativeness of the sample and provides an unbiased estimate of the association effect. The analyses were conducted adjusting for age, sex, *H. pylori* infection and treatment status, and five principal components (PCs). Transcriptome-wide significant hits that passed the false discovery rate (FDR)-corrected significance threshold (FDR q-value < 0.05) in the MITS were subsequently examined in the validation stage based on GC GWAS summary statistics of BBJ, leveraging linkage disequilibrium (LD) references derived from the East Asian population within the 1000 Genomes Project. The association between predicted expression and GC was estimated with ImpG-Summary[40] as a linear combination of elements of Z, a vector of standardized effect sizes (z-scores) of SNPs associated with GC at a given *cis*

locus, with the precomputed weights w, calculated as $Z_{TWAS} = w'Z/(w'Dw)^{1/2}$, where D denotes the LD matrix.[23] Genes with FDR q-value <0.05 in MITS and $P < 0.05$ in BBJ were defined as key GC susceptibility genes in this study. The Manhattan plot was generated using the R package "ggplot2".

Functional annotation of TWAS significant associations

Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were conducted to elucidate the biologically relevant pathways associated with key GC susceptibility genes. For two GC susceptibility genes at newly identified loci exceeding Bonferroni-corrected significance threshold ($P < 1.11 \times 10^{-5}$, 0.05/4518 genes), Bayesian colocalization analysis was performed to assess whether the GC GWAS signals and expression quantitative trait loci (eQTL) signals in gastric tissues share common causal variants. Summary statistics from the BBJ were integrated with gastric tissue eQTL data from GTEx (v8), using the "coloc" R package. Five posterior probabilities (PP0-PP4) were computed, where PP4 > 75% was considered strong evidence for colocalization, indicating a high probability (>75%) that both association signals originate from the same underlying causal variant. Enrichment analyses were conducted using the "clusterProfiler" R package. LocusZoom (http://locuszoom.org/) was employed to visualize SNP associations with gene expression levels and GC risk.

Construction and assessment of TRS with imputed transcriptomes based on prospective studies

By integrating the independent key genes associated with GC susceptibility, the TRS of an individual was calculated as the sum of gene expression ($GeneExp_j$) weighted by the effect size (Z_j), computed over g genes. The formula is as follows:

$$TRS = \sum_{j=1}^{g} GeneExp_j \times Z_j \quad (1)$$

The effect size (Z_j) for each gene was determined by incorporating all GC susceptibility genes, identified through the two-stage TWAS, into a single model of the inverse probability weighted Cox regression analysis, based on the nested case–control study within the MITS. When multiple susceptibility genes were identified at a single genomic locus (defined by an intergenic distance of <500 kb), the gene with the lowest *P*-value in the initial TWAS for MITS was retained. We further assessed associations between TRS and GC risk based on the prospective follow-up of the MITS, SIT, and UKB, using imputed gene expression levels (min–max standardized) for study participants. Restricted cubic spline analysis was conducted to evaluate potential nonlinear relationships. The Cox proportional hazards regression models (specifically, inverse probability weighted Cox regression models for MITS) were utilized to calculate the HRs and 95% CIs for continuous TRS and its quintiles associated with GC, and the *P*-value for trend (*P*-trend), adjusting for age, sex, *H. pylori* infection and treatment status (for MITS and SIT only), and five PCs. In addition to GC overall, analyses were also conducted for cardia and non-cardia GC respectively based on the MITS and UKB respectively. Sensitivity analyses were performed by additionally adjusting for smoking status, alcohol consumption, and dietary factors (total vegetable intake, total fruit intake, and preference for salty food) as covariates. Stratified analyses were conducted by these lifestyle and dietary factors.

To evaluate the performance of the TRS developed in this study compared to the conventional PRS model, we utilized a PRS model constructed from the largest existing GC GWAS[12] (10,254 Asian GCs and 10,914 controls) containing 112 SNPs, and examined the associations of the PRS with the risk of GC in the MITS, SIT, and UKB populations respectively. For each cohort, we employed regression models identical to those used in the TRS evaluation.

Construction and assessment of TRS with observed transcriptomes in the UGCED program

For the calculation of TRS for UGCED subjects, we leveraged the observed gene expression levels derived from RNA-seq instead of imputed transcriptomes, and examined the dynamic changes in TRS across the multistage progression of gastric lesions, from SG to CAG, IM, LGIN, HGIN, and ultimately GC. A rank-based ANOVA (Kruskal–Wallis test), followed by Dunn's post-hoc test with Benjamini-Hochberg correction for multiple comparisons, was used for the differences of TRS across subjects in different categories. Unconditional logistic regression models were employed to calculate the odds ratios (ORs) and 95% CIs for the associations of TRS with the risk of GC, as well as gastric neoplasia (including GC, HGIN, and LGIN) adjusting for age and sex. Considering the modest sample size and a high proportion of missing data regarding *H. pylori* infection for UGCED subjects (Table S1), the analysis did not adjust for *H. pylori* infection as a covariate.

Interactions of TRS with H. pylori *infection, anti-*H. pylori *treatment and nutrition supplementation*

Leveraging the prospective follow-up of SIT participants, we examined the risk of incident GC associated with *H. pylori* infection (1989–2022) and anti-*H. pylori* treatment (1995–2022) for individuals with a low (quintile 1), intermediate (quintiles 2–4), and high TRS (quintile 5) respectively. Among the 2755 participants in the SIT, the subjects receiving active *H. pylori* treatment in 1995

were excluded for the analysis on *H. pylori* infection, leaving 1878 subjects for this analysis. A total of 2548 attended the 2 × 2 × 2 factorial-designed intervention trial in 1995. The analysis on anti-*H. pylori* treatment was restricted to 1811 *H. pylori* positive individuals receiving either active treatment (one capsule with amoxicillin and omeprazole twice daily for 2 weeks) or its placebo (a look alike capsule) in 1995, following the intention-to-treat approach. We further examined the association of GC with successful eradication compared with those receiving placebo for *H. pylori* treatment. Additionally, the associations of garlic supplementation and vitamin supplementation with GC risk across different TRS groups were evaluated for 2548 trial participants. Kaplan–Meier curves, along with log-rank tests were employed to compare the cumulative incidence based on *H. pylori* infection or intervention status among individuals with different TRSs. The Cox proportional hazards regression models were employed to calculate the HRs (95% CIs) for the associations. *P*-values for interaction (*P*-interaction) were calculated by adding the interaction term in the regression models. For each intervention, absolute risk reduction (ARR) in GC incidence and the number of participants needed to treat (NNT) to prevent one case of GC over the 27.1-year follow-up period were calculated.

FUSION,[23] R (version 4.3.3) and PLINK (version 1.9) were utilized for analysis.

Ethics

This study was approved by the Institutional Review Board of Peking University Cancer Hospital (approval No. 2023KT12). All participants of the UGCED, MITS, and SIT provided written informed consent.

Role of funders

The funders had no role in study design, data collection, analysis and interpretation, writing of the report, or decision to publish.

Results

Two-stage TWAS for GC risk in MITS and BBJ

The SNP-weights for genetically predicted expression in gastric tissue of 6127 genes had been precomputed based on GTEx (v8, n = 324) and further assessed based on 67 participants of the UGCED program. Of them, 3341 genes met the prediction $R^2 > 0.01$ criterion in both GTEx and UGCED, while an additional 1177 genes, though unevaluable in UGCED, had a prediction $R^2 > 0.04$ in GTEx. Ultimately, these 4518 genes were included in our TWAS analysis. A significant correlation between the R^2 of prediction performance in UGCED and the prediction R^2 in GTEx (r = 0.37, Pearson correlation, $P = 5.85 \times 10^{-109}$, Fig. S1), supporting the applicability of *cis*-SNP weights from GTEx and applicability of the gene expression imputation models across populations.

Applying these imputation models to the nested case–control study within the MITS (935 incident GCs: mean age 48.2, standard deviation [SD] 5.8 years; and 1869 controls: mean age 42.9, SD 7.5 years; Table S1), the predicted expression of 175 genes was significantly associated with incident GC risk (inverse probability weighted Cox regression, FDR q-value <0.05, Fig. 2 and Table S2), enriched in pathways involving Endocytosis, extracellular matrix–receptor interaction, and PI3K-Akt signalling pathway (Table S3). We tested the proportional hazards assumption and did not find violations (Table S4). Among them, 16 genes located in 11 independent genomic loci were successfully replicated in another East Asian cohort, BBJ, at a nominal threshold of $P < 0.05$ (ImpG-Summary, Fig. 2 and Table 1). 1q22 (*GBAP1* and *THBS3*), 4q28.1 (*ANKRD50*), 5p13.1 (*PRKAA1*), and 6p21.33 (*HLA-C*) have been previously reported as GC susceptibility loci, but *GBA* and *RUSC1-AS1* in 1q22 have not been shown as the targets of GWAS-identified risk SNPs. We also paid special attention to the located genes of previously reported GC risk variants listed in the GWAS Catalogue, for which the association of the predicted expression with GC risk are shown in Table S5.

A total of 7 loci were newly associated with GC risk (Fig. 2 and Table 1). The predicted expression of *AP1AR* (4q25) and *RCCD1* (15q26.1) exceeded Bonferroni-corrected significance threshold (inverse probability weighted Cox regression, $P = 6.22 \times 10^{-7}$ and 1.44×10^{-6} respectively) in the MITS. Bayesian colocalization suggested that GWAS and gastric tissue eQTL signals share the same variant at the *AP1AR* locus (PP4 = 0.92), with rs6836717 identified as the SNP most significantly associated with both GC risk and *AP1AR* expression ($P\text{-}_{GWAS} = 5.09 \times 10^{-5}$; $P\text{-}_{eQTL} = 1.66 \times 10^{-9}$, Fig. S2). The PP4 for *RCCD1* was 0.31, indicating relatively low confidence in the colocalization of the GWAS and eQTL signals at this locus.

TRS based on imputed transcriptomes and incident GC risk

We integrated 11 independent lead genes on key genomic loci to derive a TRS, based on imputed transcriptomes. TRS = (−1.84 × *GBA*) − (3.99 × *HIBCH*) + (4.19 × *CFAP44*) − (2.99 × *AP1AR*) + (3.32 × *ANKRD50*) − (2.45 × *PRKAA1*) − (4.30 × *HLA-C*) − (4.84 × *HRCT1*) − (4.45 × *SMUG1*) − (2.85 × *RNASEH2B-AS1*) − (4.71 × *RCCD1*).

The associations between TRS and GC risk were then examined based on the nested case–control study of MITS, and SIT, as well as the cross-ancestry UKB cohort. While the MITS and SIT only comprises ethnically Chinese participants, the majority of UKB participants are Whites (93.7%) (Table S1). The mean (SD) age at enrolment was 46.9 (9.1) years for the 2755 participants in SIT, and 56.5 (8.1) years for the 406,386 participants in UKB. Women accounted for 50.9% and

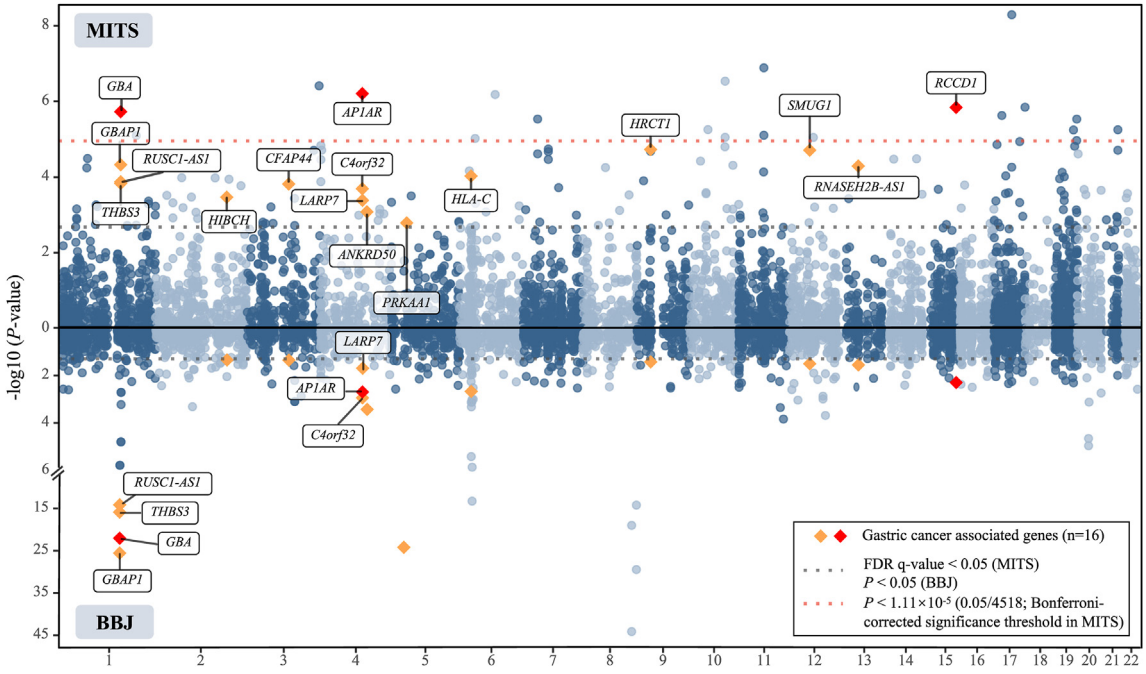


***Fig. 2:* Manhattan plot for the two-stage transcriptome-wide association study**. Each point on the graph represents a gene. Gastric gene expression imputation models precomputed based on the Genotype-Tissue Expression project (v8, n = 324) were downloaded from FUSION and utilized. The predicted expression of 175 genes were associated with gastric cancer risk (FDR q-value <0.05, dashed grey line) in MITS, and 16 of them were further successfully replicated in the BBJ at a nominal threshold of $P < 0.05$ (dashed grey line). Among them, *GBA*, *AP1AR*, and *RCCD1* passed the Bonferroni-corrected significance threshold ($P < 1.11 \times 10^{-5}$, dashed red line) in the MITS. Abbreviations: BBJ, BioBank Japan; FDR, false discovery rate; MITS, Mass Intervention Trial in Linqu, Shandong Province.

53.9% of the participants in these datasets, respectively (Table S1). During the follow-up period, a total of 146 incident cases with GC were recorded in the SIT (1989–2022) and 1144 cases in the UKB (2006–2023).

From the lowest to highest quintile of TRS, the risk of GC increased in a dose-dependent manner in the MITS (HR = 2.83, 95% CI: 2.31–3.48 for the highest quintile; HR = 1.51, 95% CI: 1.42–1.62 per SD increase in TRS, inverse probability weighted Cox regression, *P*-trend = 2.93×10^{-34}). Such similar associations were found for the SIT and UKB, with HRs of 1.27 (95% CI: 1.09–1.50, Cox regression, *P*-trend = 0.003) and 1.08 (95% CI: 1.02–1.15, Cox regression, *P*-trend = 0.008) per one SD increase in TRS respectively (Table 2 and Fig. S3). We tested the proportional hazards assumption and found no violations (Table S4). The restricted cubic spline curve also demonstrates a gradual increase in GC risk with rising TRS levels (Fig. S4). Sensitivity analyses additionally adjusting for smoking status, alcohol consumption, and dietary patterns yielded no material change in findings (Table S6).

As shown in Table S7, the majority of GCs with specified locations occurred at non-cardia sites in the MITS (89.8%, 666/742) and SIT (83.3%, 60/72), while a higher proportion of cardia GC compared to non-cardia GC was observed in the UKB. Regarding histological types, the vast majority were adenocarcinomas in MITS (96.9%, 850/877 with confirmed information) and UKB (80.2%, 438/546), but such information was unavailable for the SIT. Given the more complete data on clinical characteristics for GC in MITS and UKB, we examined the association for the risk of non-cardia and cardia GC respectively (Table S8). The results demonstrated significant associations of TRS with both non-cardia and cardia GC in two cohorts. For per SD increase in TRS, inverse probability weighted Cox regression analysis yield an HR of 1.53 (95% CI: 1.41–1.66, *P*-trend = 1.60×10^{-25}) for non-cardia GC and 1.51 (95% CI: 1.20–1.90, *P*-trend = 3.85×10^{-4}) for cardia GC in MITS. In UKB, the HR was 1.15 (95% CI: 1.03–1.30, Cox regression, *P*-trend = 0.02) for non-cardia cancer and 1.10 (95% CI: 1.01–1.19, Cox regression, *P*-trend = 0.03) for cardia cancer.

Comparison of associations between TRS and PRS with GC risk

Based on the PRS derived from the largest previously conducted GWAS on GC, we examined the risk of GC for the MITS, SIT, and UKB cohorts. The HR per SD PRS increase was 1.16 (95% CI: 1.09–1.24, inverse probability weighted Cox regression, *P*-trend = 2.85×10^{-6}) for MITS, 1.17 (95% CI: 0.99–1.38,

Region	Gene	Type	Heritability[b]	Prediction R^2		MITS				BioBank Japan	
				GTEx	UGCED	HR (95% CI)	Z_{TWAS}	*P*	FDR q-value	Z_{TWAS}	*P*
Previously reported loci											
1q22	*GBA*	Protein-coding	0.13	0.02	0.03	0.36 (0.24–0.55)	−4.77	1.87×10^{-6}	8.76×10^{-4}	−9.82	8.87×10^{-23}
	GBAP1	Pseudogene	0.30	0.23	0.13	0.57 (0.43–0.75)	−4.07	4.70×10^{-5}	0.005	−10.61	2.62×10^{-26}
	RUSC1-AS1	LncRNA	0.05	0.03	0.02	1.61 (1.26–2.06)	3.82	1.34×10^{-4}	0.01	7.79	6.97×10^{-15}
	THBS3	Protein-coding	0.27	0.20	0.02	2.27 (1.49–3.46)	3.80	1.46×10^{-4}	0.01	8.27	1.37×10^{-16}
4q28.1	*ANKRD50*	Protein-coding	0.21	0.10	0.09	2.02 (1.34–3.04)	3.34	8.30×10^{-4}	0.03	3.56	3.72×10^{-4}
5p13.1	*PRKAA1*	Protein-coding	0.10	0.02	0.04	0.67 (0.52–0.86)	−3.15	0.002	0.04	−10.30	6.25×10^{-25}
6p21.33	*HLA-C*	Protein-coding	0.60	0.26	0.20	0.43 (0.28–0.66)	−3.91	9.38×10^{-5}	0.01	−3.07	0.002
Newly identified loci[c]											
2q32.2	*HIBCH*	Protein-coding	0.43	0.40	0.01	0.56 (0.40–0.77)	−3.58	3.39×10^{-4}	0.02	−2.01	0.04
3q13.2	*CFAP44*	Protein-coding	0.18	0.15	0.01	1.84 (1.34–2.51)	3.79	1.52×10^{-4}	0.01	2.01	0.04
4q25	*AP1AR*	Protein-coding	0.13	0.08	0.04	0.61 (0.50–0.74)	−4.98	6.22×10^{-7}	4.61×10^{-4}	−3.09	0.002
	C4orf32	Protein-coding	0.08	0.03	0.03	1.50 (1.21–1.86)	3.72	2.03×10^{-4}	0.01	3.25	0.001
	LARP7	Protein-coding	0.11	0.02	0.03	2.74 (1.57–4.79)	3.53	4.11×10^{-4}	0.02	2.32	0.02
9p13.3	*HRCT1*	Protein-coding	0.23	0.09	0.06	0.37 (0.23–0.58)	−4.28	1.87×10^{-5}	0.003	−2.08	0.04
12q13.13	*SMUG1*	Protein-coding	0.35	0.10	0.04	0.44 (0.31–0.64)	−4.27	1.94×10^{-5}	0.003	−2.16	0.03
13q14.3	*RNASEH2B-AS1*	LncRNA	0.34	0.17	0.02	0.23 (0.11–0.46)	−4.05	5.12×10^{-5}	0.005	−2.20	0.03
15q26.1	*RCCD1*	Protein-coding	0.34	0.28	0.03	0.62 (0.51–0.75)	−4.82	1.44×10^{-6}	7.59×10^{-4}	−2.81	0.01

Abbreviations: CI, confidence interval; FDR, false discovery rate; GTEx, Genotype-Tissue Expression; GWAS, genome-wide association study; HR, hazard ratio; lncRNA, long non-coding RNA; MITS, Mass Intervention Trial in Linqu, Shandong Province; TWAS, transcriptome-wide association study; UGCED, Upper Gastrointestinal Cancer Early Detection. [a]Gastric gene expression imputation models previously developed based on GTEx (v8, n = 324) were downloaded from FUSION and utilized. The associations between the imputed gene expression levels (mean standardized) in the MITS and the incident gastric cancer risk were assessed using inverse probability weighted Cox regression models, adjusting for age, sex, *H. pylori* infection and treatment status, and five principal components. Genes with FDR q-value <0.05 (corrected for 4518 genes) were subsequently examined based on GWAS summary statistics of BioBank Japan, and those that were replicated at a nominal *P*-value <0.05 were identified as susceptibility genes for gastric cancer and are presented here. [b]Heritability of genes had been precalculated based on GTEx. [c]Newly identified loci are defined as genomic regions more than 500 kb from any gastric cancer risk variants identified in GWAS (as listed in the GWAS Catalogue up to March 28, 2024).

Table 1: **16 susceptibility genes identified by two-stage transcriptome-wide association study for gastric cancer.[a]**

Cox regression, *P*-trend = 0.06) for SIT, and 0.96 (95% CI: 0.91–1.02, Cox regression, *P*-trend = 0.19) for UKB (Table S9). In each cohort, the TRS exhibited a stronger association with the risk of GC compared to the PRS, demonstrating more robust correlations with GC risk even in the UKB. This supports the superior generalizability of the newly developed TRS across diverse populations (Fig. S5).

TRS based on observed transcriptomes in the UGCED program

We further evaluated the performance of the TRS integrating 11 genes based on the observed transcriptomes of 103 participants from the UGCED program (Table S10). TRS exhibited an upward trend with the progression of gastric lesions (linear regression, *P*-trend = 5.01×10^{-4}), with GCs showing significantly

TRS	MITS			SIT			UK Biobank		
	No. of cases (person-years)	HR (95% CI)	*P*	No. of cases (person-years)	HR (95% CI)	*P*	No. of cases (person-years)	HR (95% CI)	*P*
Quintile 1	147 (5893)	1.00 (Reference)		16 (15,784)	1.00 (Reference)		200 (1,139,990)	1.00 (Reference)	
Quintile 2	154 (5792)	1.29 (1.03–1.62)	0.03	30 (15,648)	1.76 (0.96–3.24)	0.07	207 (1,139,269)	1.02 (0.84–1.24)	0.83
Quintile 3	166 (5655)	1.41 (1.13–1.76)	0.002	31 (15,630)	1.96 (1.07–3.59)	0.03	239 (1,140,996)	1.18 (0.98–1.42)	0.09
Quintile 4	220 (5488)	2.11 (1.71–2.60)	2.67×10^{-12}	33 (15,418)	2.15 (1.18–3.90)	0.01	245 (1,140,665)	1.20 (1.00–1.45)	0.05
Quintile 5	248 (5320)	2.83 (2.31–3.48)	2.43×10^{-23}	36 (15,356)	2.38 (1.32–4.30)	0.004	253 (1,140,650)	1.23 (1.02–1.49)	0.03
Per one SD score increase	935 (28,149)	1.51 (1.42–1.62)	2.93×10^{-34}	146 (77,837)	1.27 (1.09–1.50)	0.003	1144 (5,701,570)	1.08 (1.02–1.15)	0.008

Abbreviations: CI, confidence interval; HR, hazard ratio; MITS, Mass Intervention Trial in Linqu, Shandong Province; NA, not applicable; SD, standard deviation; SIT, Shandong Intervention Trial; TRS, transcriptional risk score. [a]Participants were divided into quintiles according to their TRSs. The HR (95% CI) of gastric cancer for each quintile was calculated using Cox proportional hazards regression models (specifically, inverse probability weighted Cox regression models for MITS) with quintile 1 as the reference, adjusting for age, sex, *H. pylori* infection and treatment status (for MITS and SIT only), and five principal components. The HR (95% CI) per one SD increase in TRS and the *P*-value for trend were also calculated.

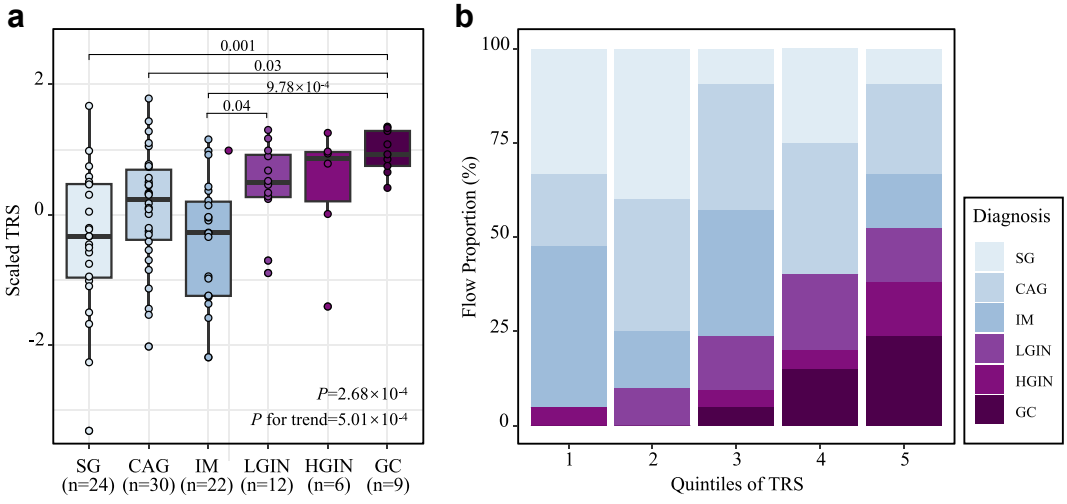
Table 2: **Transcriptional risk score based on imputed transcriptomes and incident gastric cancer risk in the nested case–control study of MITS, SIT, and UK Biobank.[a]**

higher TRS compared to those with SG, CAG and IM (Fig. 3a). Analysis by quintiles of TRS showed that GC was not observed in the lowest two TRS quintiles, with more than half of the cases occurring in the highest quintile (Fig. 3b). Consistent with the analyses based on imputed expression in other cohorts, a strong association between TRS and GC was observed (OR = 9.23 per SD increase in TRS, 95% CI: 1.76–48.46, logistic regression, *P*-trend = 0.009). Constrained by the modest number of GC (n = 9), we further considered gastric neoplasia (GC, HGIN, or LGIN) as a secondary outcome and found that individuals with higher TRS had a significantly increased risk of gastric neoplasia (OR = 3.48 per SD increase in TRS, 95% CI: 1.70–7.12, logistic regression, *P*-trend = 6.25×10^{-4}, Fig. 3c).

Interaction between TRS and chemoprevention on GC risk in the SIT

Stratifying SIT participants by TRS levels, we found that the association of *H. pylori* infection on the development of GC was most pronounced for individuals with high TRS (HR = 5.76, 95% CI: 1.56–21.29) during the entire follow-up period (1989–2022), demonstrating a significant interaction between *H. pylori* infection and TRS levels on the risk of GC (Cox regression, *P*-interaction = 0.03, Fig. 4).

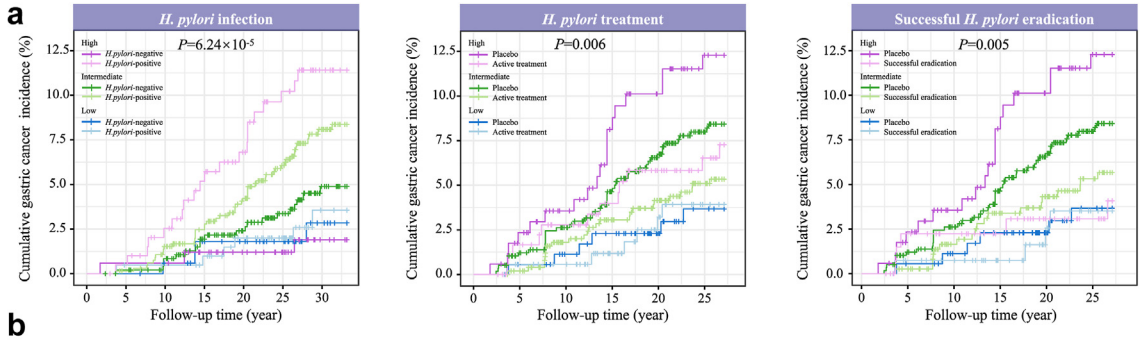
Among 2548 trial participants in 1995, 138 cases with GC were documented during the subsequent 27.1-year follow-up (1995–2022), including 113 cases for baseline *H. pylori* positive individuals (n = 1811) and 25 cases for *H. pylori* negative ones (n = 737) (Table S11). The beneficial effect of *H. pylori* treatment on reducing



c

TRS	No. of gastric neoplasia/non-gastric neoplasia	OR (95% CI)	*P*
Quintile 1	1/20	1.00 (Reference)	NA
Quintile 2	2/18	1.47 (0.12-18.64)	0.77
Quintile 3	5/16	4.78 (0.48-47.98)	0.18
Quintile 4	8/12	13.42 (1.40-128.54)	0.02
Quintile 5	11/10	23.63 (2.51-222.45)	0.006
Per one SD score increase	27/76	3.48 (1.70-7.12)	6.25×10^{-4}

***Fig. 3:* Transcriptional risk score based on observed transcriptomes in the UGCED program**. a. Boxplot of TRS according to gastric lesions and GC. A rank-based ANOVA (Kruskal–Wallis test) with Dunn's post-hoc adjusted by Benjamini-Hochberg method was used to compare the TRS in different categories. The *P*-value for trend was calculated using a linear regression model. As the Kruskal–Wallis test involved 15 pairwise comparisons, only significant *P* values are presented for brevity. b. Proportion of different gastric lesions and GC according to quintiles of TRS. c. TRS and risk of gastric neoplasia. The OR (95% CI) of gastric neoplasia (including GC, HGIN, and LGIN) for each quintile was calculated using unconditional logistic regression models with quintile 1 as the reference, adjusting for age and sex. The OR (95% CI) per one SD increase in TRS and the *P*-value for trend were also calculated. Abbreviations: CAG, chronic atrophic gastritis; CI, confidence interval; GC, gastric cancer; HGIN, high-grade intraepithelial neoplasia; IM, intestinal metaplasia; LGIN, low-grade intraepithelial neoplasia; NA, not applicable; OR, odds ratio; SD, standard deviation; SG, superficial gastritis; TRS, transcriptional risk score; UGCED, Upper Gastrointestinal Cancer Early Detection.



b

TRS categories	*H. pylori* infection: No. of cases (person-years) Negative	Positive	HR (95% CI)	*P*-int	*H. pylori* treatment: No. of cases (person-years) Placebo	Treatment	HR (95% CI)	ARR (%)	NNT	*P*-int	Successful *H. pylori* eradication: No. of cases (person-years) Placebo	Eradication	HR (95% CI)	ARR (%)	NNT	*P*-int
Overall	27 (19,578)	76 (25,472)	1.91 (1.22-3.00)	0.03	70 (22,126)	43 (20,916)	0.66 (0.45-0.97)	2.79	35.87	0.21	70 (22,126)	28 (15,321)	0.61 (0.39-0.95)	3.37	29.64	0.05
Low	4 (4,928)	6 (5,905)	0.96 (0.22-4.13)		6 (4,389)	6 (4,206)	1.32 (0.39-4.49)	NA	NA		6 (4,389)	4 (3,172)	1.53 (0.43-5.45)	NA	NA	
Intermediate	20 (13,431)	49 (18,723)	1.62 (0.95-2.78)		45 (13,854)	25 (12,406)	0.62 (0.38-1.02)	3.11	32.11		45 (13,854)	19 (8,920)	0.70 (0.41-1.20)	2.81	35.54	
High	3 (4,838)	21 (5,433)	5.76 (1.56-21.29)		19 (3,883)	12 (4,304)	0.42 (0.19-0.93)	5.08	19.69		19 (3,883)	5 (3,229)	0.24 (0.08-0.71)	8.14	12.28	

***Fig. 4:* Association of *H. pylori* infection, anti-*H. pylori* treatment, and successful eradication with the risk of incident gastric cancer by different TRS levels in the Shandong Intervention Trial**. Participants were categorized by low (quintile 1), intermediate (quintiles 2–4), and high (quintile 5) levels of TRS constructed with imputed transcriptomes. Analysis of *H. pylori* infection was conducted based on the entire study period (1989–2022), excluding *H. pylori*-seropositive individuals receiving active treatment. Analysis for interventions was conducted based on the 27.1-year follow-up period after trial randomization (1995–2022). a. Cumulative gastric cancer incidence for individuals with different *H. pylori* infection or treatment status according to TRS categories. Log-rank tests were employed to compare the cumulative incidence. b. Stratified analysis on the association of *H. pylori* infection, anti-*H. pylori* treatment, and successful eradication with the risk of incident gastric cancer by TRS levels. The Cox proportional hazards regression models were conducted to estimate HRs (95% CIs), and *P*-values for interaction were calculated by adding the interaction term in the regression models. ARR and NNT for *H. pylori* treatment and successful eradication were calculated over the 27.1-year follow-up period. NNT = 1/ARR. Abbreviations: ARR, absolute risk reduction; CI, confidence interval; *H. pylori*, *Helicobacter pylori*; HR, hazard ratio; NA, not applicable; NNT, number needed to treat; *P*-int, *P*-value for interaction; TRS, transcriptional risk score.

GC risk was only observed for individuals with the high TRS (HR = 0.42, 95% CI: 0.19–0.93, ARR = 5.08%, NNT = 19.69), and appeared even stronger for high TRS individuals with successful eradication (HR = 0.24, 95% CI: 0.08–0.71, ARR = 8.14%, NNT = 12.28). Such significant effect was not found for those with intermediate or low TRS, demonstrating an effect modification of the TRS level on successful *H. pylori* eradication (Cox regression, *P*-interaction = 0.05) (Fig. 4). We also examined the effect of vitamin supplementation and garlic supplementation but did not find significant interactions by TRS (Cox regression, *P*-interaction = 0.59 for garlic supplementation and *P*-interaction = 0.99 for vitamin supplementation) (Fig. S6).

Additionally, we conducted stratified analyses by key lifestyle factors-including smoking status, alcohol consumption, and dietary habits (total vegetable intake, total fruit intake, and preference for salty foods) to assess their potential effect modification on the TRS-GC association. No statistically significant interactions were observed (Cox regression [inverse probability weighted Cox regression for MITS], *P*-interaction > 0.05 for all strata, Table S12).

Discussion

Based on a two-stage TWAS, we identified a panel of genes associated with the risk of developing GC, which may provide new insights into the genetic architecture and aetiology of GC. The derived TRS may predict GC risk in independent, cross-ancestry prospective cohorts and potentially serve as a generalizable tool for risk stratification. The TRS also modified the associations of *H. pylori* infection and successful *H. pylori* eradication with GC risk, possibly opening up new avenues for optimized primary prevention of GC.

In our study, the two stages of TWAS were conducted based on Chinese (MITS) and Japanese (BBJ) populations respectively. Based on the BBJ GWAS summary statistics, 16 out of 175 genes were significantly replicated for GC risk. In addition, MITS participants were enrolled from an established high-risk area for GC while the BBJ included hospital-based cases. The recognized genetic heterogeneity even among East Asian populations,[41] coupled with potentially distinct genetic susceptibilities for individuals from high-risk areas, may partly account for the relatively limited number of genes that were successfully replicated. Even

so, our findings demonstrate the value of TWAS in both validating previously identified GC susceptibility loci and discovering further loci, contributing to the further understanding of the genetic susceptibility of GC.

Prior GWASs have revealed multiple genetic loci associated with GC.[11–15] Our TWAS confirmed 4 previously identified GC susceptibility loci providing evidence for the impact of dysregulated gene expression on GC risk. Upregulation of *ANKRD50* (4q28.1) in GC has also been reported in a previous TWAS study based on meta-analysis of ten European GWASs involving 5815 patients and 10,999 controls.[15] In addition, we newly identified 7 genomic loci associated with GC on chromosome 2q32.2, 3q13.2, 4q25, 9p13.3, 12q13.13, 13q14.3, and 15q26.1 respectively, among which the expression of *AP1AR* on chromosome 4q25 (encoding the Adaptor Related Protein Complex 1 Associated Regulatory Protein) met the Bonferroni correction for multiple comparisons. While a study has reported the role of AP1AR and its binding partner PICALM as regulators of pro-inflammatory signalling,[42] there is still a lack of research investigating their contributions to the initiation and progression of tumorigenesis. Bayesian colocalization analysis indicates rs6836717 as the common causal variant that underlies both GWAS signals and gastric tissue eQTL signals at the *AP1AR* locus. It potentially mediates GC susceptibility through regulation of *AP1AR* expression, thereby providing a target for future research. Another newly identified candidate gene that passed the Bonferroni-corrected significance threshold was *RCCD1* (15q26.1), which encodes the RCC1 Domain-Containing Protein 1, may be important for cell-cycle-regulated transcriptional repression in centromeric regions and accurate mitotic division spindle organization.[43] *RCCD1* has been reported to be associated with breast cancer,[44] ovarian cancer,[45] and pancreatic cancer.[46] Here, we presented evidence for its possible association with the development of GC but further studies are warranted to clarify the potential underlying mechanisms.

Compared with PRS, TRS exhibits several potential complementary advantages. Previous studies have indicated that PRS solely based on genetic level data frequently encounter the challenge of cross-ancestry generalizability.[16–18] In our study, the comparison between PRS and TRS reveals that, although the PRS retains its potential predictive value within populations of the same ancestry, it demonstrates limited generalizability across different ancestries. In contrast, the derived transcriptome-based TRS exhibited a consistent association with GC risk across diverse cohorts, underscoring its portability and applicability for risk stratification across different ancestry groups. Existing studies have suggested that TRS might offer advantages by harnessing eQTL variants, which tie it to biological mechanisms potentially shared among various ancestry groups.[25,26] In addition, it was proposed that the portability of TRS may also stem from the methods and ancestry composition used to construct the models.[21,26] Beyond cross-ancestry robustness, another key advantage of TRS lies in its utilization of trait-associated gene panels that focus on functionally relevant tissues, thereby enhancing its histological specificity and biological interpretability.[21,24] Moreover, gene-level analysis provides technical advantages over SNP-based approaches, including reduced model optimization complexity and significantly improved computational efficiency.[20] Therefore, while there is ongoing debate regarding whether TRS demonstrates superiority in risk prediction over PRS,[25–28] the fact that TRS can be applied across ethnic populations through genome interrogation establishes a foundation for its translation in identifying high-risk populations and early detection of GC. This further underscores the potential of an integrative approach combining TRS with PRS to enhance prediction accuracy and cross-ancestry generalizability in future studies.[25,26,47]

While TRS may be translationally relevant without the need for actual transcriptome data,[23,27] our study also yielded robust findings when actual transcriptome data is available for the UGCED participants, revealing that the TRS signature was associated with the cascade progression of gastric lesions and the development of GC. Previous studies have indicated that TRSs based on observed gene expression data capture a richer array of information than *cis*-eQTL genetic variants.[29] This includes trans-acting genetic effects or environmental influences, thereby potentially reflecting a more dynamic aspect of biological regulation and expanding the phenotypic variance that can be accounted for beyond PRSs.[29] Even so, studies measuring gene expression have frequently encountered obstacles due to limitations in specimen availability and cost. Further large-scale studies are necessary to investigate whether TRS, which incorporates actual transcriptome data, would demonstrate enhanced predictive accuracy for risk assessment compared to approaches that rely solely on genomic interrogation.

In our study, we succeeded in validating the TRS for GC based on cross-ancestry cohorts. The stronger effect magnitude observed for MITS compared to UKB may still be partly attributed to differences in race and ethnicities. Heterogeneity in epidemiologic and etiologic factors for GC has been well known across different populations. For example, *H. pylori* infection rate varies between East Asian and Western Population. However, although *H. pylori* infection status may act as a confounder or interact with genetic susceptibility in influencing GC risk, TWAS relies on germline genetic variants for imputing gene expression. Indeed, *H. pylori* infection is not typically regarded as a determinant of inherited genetic susceptibility in this context. To further enhance the robustness of the constructed TRS for GC risk, we observed consistent associations for

both cardia and non-cardia GC in the Chinese (MITS) and Western (UKB) populations, respectively.

Previous studies have elaborated on the pivotal role of host genetic factors in the *H. pylori* infection-related gastric lesion progression and GC development.[48–50] Specific genes may interact with host molecules involved in the response to *H. pylori* infection or its resultant effects, thereby potentially influencing the pathogenesis of GC. Our study reinforces the potential impact of *H. pylori* infection and its eradication particularly in the context of TRS, with the beneficial effect of anti-*H. pylori* treatment and successful eradication on preventing GC particularly noteworthy for those with the top quintile of TRS. In line with our previous study,[13] our findings corroborate the benefits of *H. pylori* treatment in individuals with a high genetic risk of GC and suggests the integration of genetic susceptibility with transcription profiling to enhance the implementation of effective primary prevention strategies for GC, ultimately facilitating more efficient GC primary prevention.

The pathological variants of hereditary cancer genes, typically very rare, are well-recognized to have profound impacts and offer significant clinical value for the diagnosis and treatment of hereditary GCs.[50] However, their extremely low MAFs may limit their utility in risk assessment and the development of prevention strategies for sporadic cancers in population-based settings, such as conducting endoscopic screening and implementing primary prevention measures like *H. pylori* treatment. Indeed, in standard GWAS and TWAS quality control pipelines, rare variants are commonly excluded to guarantee robust association signals.[51] For example, although GTEx employed comprehensive genotyping techniques that encompasses whole-genome sequencing, whole-exome sequencing, and SNP arrays covering both common and rare variants, the official TWAS protocol for constructing gene expression imputation models recommends utilizing genetic variants with MAF>1% and restricting to Hapmap3 SNPs for analysis.[23] This approach has been empirically proven to efficiently capture common variants and exhibit robust imputation accuracy across different genotype arrays. Due to the inherent constraints of the standard TWAS pipeline, we were unable to investigate hereditary cancer genes based on pathologic variants. We recognize this as a limitation of our study.

Our study has several other limitations. First, not all genes exhibit a significant hereditary component in their expression regulation, which limited the scope of our investigation into these genes. For instance, we were not able to examine *MUC1* and *PLCE1,* for which previous studies have consistently reported the genetic variants associated with GC. The inability to detect heritable levels of gene expression can be partly attributed to the relatively small number of samples in the GTEx. It is imperative to develop gene expression prediction models by leveraging expanded RNA-seq resources. Second, the GTEx project, mainly comprising Whites, was utilized for the construction of gene expression imputation models. Despite the validation of the model performance based on the UGCED program in our study, utilizing reference populations with diverse genetic backgrounds and performing gene expression profiling would be ideal for constructing gene expression imputation models, as this would greatly enhance their generalizability and reliability across different ancestral groups. During the TWAS analysis, although the robustness of our findings was strengthened by independent validation in the BBJ and UKB, caution is needed when generalizing our results to other Chinese populations, as the in-house cohorts of UGCED, MITS, and SIT were all recruited from a high-risk area for gastric cancer. Third, a matched nested case–control study would be preferable for the discovery stage of TWAS analysis. However, during the study design, great emphasis was placed on ensuring that the sampled controls were representative of the entire MITS population. Consequently, the controls were randomly selected without matching to GCs on major confounders like age, sex, and *H. pylori* infection. Fourth, due to the absence of information on *H. pylori* infection for the majority of UKB participants, we were unable to adjust for *H. pylori* infection in the analysis or conduct stratified analysis or interaction analysis between TRS and *H. pylori* infection on GC risk within the UKB cohort. Further study would also be warranted to examine the TRS associated with risk of GC according to *H. pylori* serotypes by virulence factors in different populations. Fifth, although we attempted to leverage actual transcriptome data, we had a modest sample size of subjects from UGCED program with observed transcriptomes. Sixth, while TWASs can propose potential relationships between genes and GC that are more biologically interpretable compared to SNP associations, they cannot answer the mechanisms by which these genes influence cancer risk. Additional research into the functional and biological mechanisms is necessary to clarify their specific roles in the development of GC.

In conclusion, our TWAS unveiled a gene panel that predicts GC risk in cross-ancestry populations, which may possibly serve as a generalizable risk stratification tool for GC. The TRS may modify the effect of *H. pylori* infection, suggests the potential of TRS in precision prevention of GC, indicating that chemoprevention strategies could be customized based on TRS for effective GC prevention.

Contributors

Dr. Li has had full access to all the data in this study and takes responsibility for the integrity of the data and the accuracy of the data analysis, with overall supervision of the study. The concept and design of the study were conceived by ZY Yin and WQ Li. Data acquisition, analysis, and interpretation were carried out by ZY Yin, HM Xu, MY Wang, XL Wang, ZC Liu, Y Jin, and WQ Li. ZY Yin performed the

statistical analysis. ZY Yin, HM Xu and WQ Li verified the underlying data. The manuscript was draughted by ZY Yin, with KF Pan and WQ Li contributing to its critical revision. Y Zhang, JY Zhang, T Zhou, WC You, KF Pan, and WQ Li provided administrative, technical, or material support. All authors have read and approved the final version of the manuscript.

Data sharing statement

The RNA sequencing data of the subjects from the Upper Gastrointestinal Cancer Early Detection Program are deposited to the Genome Sequence Archive in National Genomics Data Center, China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA-Human: HRA011102) that are publicly accessible after journal's acceptance of the manuscript at https://ngdc.cncb.ac.cn/gsa-human. The SNP-weights for gastric gene expression are freely available for download from FUSION (http://gusevlab.org/projects/fusion/), and summary statistics of the genome-wide association study on gastric cancer in BioBank Japan are publicly available from https://pheweb.jp/. The UK Biobank resource (https://www.ukbiobank.ac.uk/) was accessed under Application No. 90999.

Declaration of interests

The authors declare no potential conflicts of interest.

Acknowledgements

This study was funded by the National Natural Science Foundation of China (No. 82273704), Noncommunicable Chronic Diseases-National Science and Technology Major Project (No. 2023ZD0501400-2023ZD0501402), Beijing Hospitals Authority's Ascent Plan (DFL20241102), Beijing Hospitals Authority Clinical Medicine Development of Special Funding Support (No. ZLRK202325), China Postdoctoral Science Foundation (2024M760152), Peking University Medicine Fund for World's Leading Discipline or Discipline Cluster Development (No. BMU2022XKQ004), Science Foundation of Peking University Cancer Hospital (No. BJCH2024BJ02, XKFZ2410, BJCH2025CZ04, and 2022–27). We thank all individuals who participated in this study and donated samples. The UK Biobank resource was used for this research under Application No. 90999. We acknowledge the contributions of participants and investigators of both the UK Biobank and BioBank Japan.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.ebiom.2025.105767.

References

1 Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2024;74(3):229–263.
2 Chen Y-C, Malfertheiner P, Yu H-T, et al. Global prevalence of Helicobacter pylori infection and incidence of gastric cancer between 1980 and 2022. *Gastroenterology*. 2024;166(4):605–619.
3 Liu Z, Xu H, You W, Pan K, Li W. Helicobacter pylori eradication for primary prevention of gastric cancer: progresses and challenges. *J Natl Cancer Cent*. 2024;4:299.
4 You WC, Brown LM, Zhang L, et al. Randomized double-blind factorial trial of three treatments to reduce the prevalence of precancerous gastric lesions. *J Natl Cancer Inst*. 2006;98(14):974–983.
5 Ma JL, Zhang L, Brown LM, et al. Fifteen-year effects of Helicobacter pylori, garlic, and vitamin treatments on gastric cancer incidence and mortality. *J Natl Cancer Inst*. 2012;104(6):488–492.
6 Li WQ, Zhang JY, Ma JL, et al. Effects of Helicobacter pylori treatment and vitamin and garlic supplementation on gastric cancer incidence and mortality: follow-up of a randomized intervention trial. *BMJ*. 2019;366:l5016.
7 Pan K-F, Li W-Q, Zhang L, et al. Gastric cancer prevention by community eradication of Helicobacter pylori: a cluster-randomized controlled trial. *Nat Med*. 2024;30:3250.
8 Li WQ, Qin XX, Li ZX, et al. Beneficial effects of endoscopic screening on gastric cancer and optimal screening interval: a population-based study. *Endoscopy*. 2022;54(9):848–858.
9 Roberts MC, Holt KE, Del Fiol G, Baccarelli AA, Allen CG. Precision public health in the era of genomics and big data. *Nat Med*. 2024;30(7):1865–1873.
10 Mucci LA, Hjelmborg JB, Harris JR, et al. Familial risk and heritability of cancer among twins in nordic countries. *JAMA*. 2016;315(1):68–76.
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(4):641–651.
12 Jin G, Lv J, Yang M, et al. Genetic risk, incident gastric cancer, and healthy lifestyle: a meta-analysis of genome-wide association studies and prospective cohort study. *Lancet Oncol*. 2020;21(10):1378–1386.
13 Xu H-M, Han Y, Liu Z-C, et al. Helicobacter pylori treatment and gastric cancer risk among individuals with high genetic risk for gastric cancer. *JAMA Netw Open*. 2024;7(5):e2413708.
14 Sakaue S, Kanai M, Tanigawa Y, et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet*. 2021;53(10):1415–1424.
15 Hess T, Maj C, Gehlen J, et al. Dissecting the genetic heterogeneity of gastric cancer. *eBioMedicine*. 2023;92:104616.
16 Abdellaoui A, Yengo L, Verweij KJH, Visscher PM. 15 years of GWAS discovery: realizing the promise. *Am J Hum Genet*. 2023;110(2):179–194.
17 Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51(4):584–591.
18 Gu Y, Yan C, Wang T, Hu B, Zhu M, Jin G. Construction and evaluation of the functional polygenic risk score for gastric cancer in a prospective cohort of the European population. *Chin Med J (Engl)*. 2023;136(14):1671–1679.
19 Nicolae DL, Gamazon E, Zhang W, Duan S, Dolan ME, Cox NJ. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet*. 2010;6(4):e1000888.
20 Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet*. 2017;18(2):117–127.
21 The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*. 2020;369(6509):1318–1330.
22 Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47(9):1091–1098.
23 Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 2016;48(3):245–252.
24 Wainberg M, Sinnott-Armstrong N, Mancuso N, et al. Opportunities and challenges for transcriptome-wide association studies. *Nat Genet*. 2019;51(4):592–599.
25 Liang Y, Pividori M, Manichaikul A, et al. Polygenic transcriptome risk scores (PTRS) can improve portability of polygenic risk scores across ancestries. *Genome Biol*. 2022;23(1):23.
26 Hu X, Qiao D, Kim W, et al. Polygenic transcriptome risk scores for COPD and lung function improve cross-ethnic portability of prediction in the NHLBI TOPMed program. *Am J Hum Genet*. 2022;109(5):857–870.
27 Pain O, Glanville KP, Hagenaars S, et al. Imputed gene expression risk scores: a functionally informed component of polygenic risk. *Hum Mol Genet*. 2021;30(8):727–738.
28 Mo A, Nagpal S, Gettler K, et al. Stratification of risk of progression to colectomy in ulcerative colitis via measured and predicted gene expression. *Am J Hum Genet*. 2021;108(9):1765–1779.
29 Cabana-Domínguez J, Llonga N, Arribas L, et al. Transcriptomic risk scores for attention deficit/hyperactivity disorder. *Mol Psychiatry*. 2023;28(8):3493–3502.
30 Marigorta UM, Denson LA, Hyams JS, et al. Transcriptional risk scores link GWAS to eQTLs and predict complications in Crohn's disease. *Nat Genet*. 2017;49(10):1517–1521.
31 You WC, Blot WJ, Li JY, et al. Precancerous gastric lesions in a population at high risk of stomach cancer. *Cancer Res*. 1993;53(6):1317–1321.
32 The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580–585.
33 Huang S, Guo Y, Li ZW, et al. Identification and validation of plasma metabolomic signatures in precancerous gastric lesions that progress to cancer. *JAMA Netw Open*. 2021;4(6):e2114186.
34 Li X, Zheng NR, Wang LH, et al. Proteomic profiling identifies signatures associated with progression of precancerous gastric lesions and risk of early gastric cancer. *eBioMedicine*. 2021;74:103714.

35 Dixon MF, Genta RM, Yardley JH, Correa P. Classification and grading of gastritis. The updated Sydney system. International workshop on the histopathology of gastritis, Houston 1994. *Am J Surg Pathol*. 1996;20(10):1161–1181.
36 Elliott P, Peakman TC. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. *Int J Epidemiol*. 2008;37(2):234–244.
37 Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203–209.
38 Wu L, Wang J, Cai Q, et al. Identification of novel susceptibility loci and genes for prostate cancer risk: a transcriptome-wide association study in over 140,000 European descendants. *Cancer Res*. 2019;79(13):3192–3204.
39 Støer NC, Samuelsen SO. Inverse probability weighting in nested case-control studies with additional matching–a simulation study. *Stat Med*. 2013;32(30):5328–5339.
40 Pasaniuc B, Zaitlen N, Shi H, et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*. 2014;30(20):2906–2914.
41 Yuan Y, Yang L, Shi M, et al. Identification of well-differentiated gene expressions between Han Chinese and Japanese using genome-wide microarray data analysis. *J Med Genet*. 2013;50(8):534–542.
42 Mertins P, Przybylski D, Yosef N, et al. An integrative framework reveals signaling-to-transcription events in toll-like receptor signaling. *Cell Rep*. 2017;19(13):2853–2866.
43 Marcon E, Ni Z, Pu S, et al. Human-chromatin-related protein interactions identify a demethylase complex required for chromosome segregation. *Cell Rep*. 2014;8(1):297–310.
44 Cai Q, Zhang B, Sung H, et al. Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nat Genet*. 2014;46(8):886–890.
45 Kar SP, Beesley J, Amin Al Olama A, et al. Genome-wide meta-analyses of breast, ovarian, and prostate cancer association studies identify multiple new susceptibility loci shared by at least two cancer types. *Cancer Discov*. 2016;6(9):1052–1067.
46 Zhong J, Jermusyk A, Wu L, et al. A transcriptome-wide association study identifies novel candidate susceptibility genes for pancreatic cancer. *J Natl Cancer Inst*. 2020;112(10):1003–1012.
47 Cai X, Li H, Cao X, et al. Integrating transcriptomic and polygenic risk scores to enhance predictive accuracy for ischemic stroke subtypes. *Hum Genet*. 2025;144(1):43–54.
48 Malfertheiner P, Camargo MC, El-Omar E, et al. Helicobacter pylori infection. *Nat Rev Dis Primers*. 2023;9(1):19.
49 Zhang X, Liu F, Bao H, et al. Distinct genomic profile in h. pylori-associated gastric cancer. *Cancer Med*. 2021;10(7):2461–2469.
50 Usui Y, Taniyama Y, Endo M, et al. Helicobacter pylori, homologous-recombination genes, and gastric cancer. *N Engl J Med*. 2023;388(13):1181–1190.
51 Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc*. 2010;5(9):1564–1573.