scientific reports

Check for updates

OPEN

Interactions between folate intake and genetic predictors of gene expression levels associated with colorectal cancer risk

Cameron B. Haas[1,2]✉, Yu-Ru Su[2], Paneen Petersen[2], Xiaoliang Wang[2], Stephanie A. Bien[2], Yi Lin[2], Demetrius Albanes[3], Stephanie J. Weinstein[3], Mark A. Jenkins[4], Jane C. Figueiredo[5,6], Polly A. Newcomb[2,7], Graham Casey[8], Loic Le Marchand[9], Peter T. Campbell[10], Victor Moreno[11,12,13,14], John D. Potter[2,15], Lori C. Sakoda[2,16], Martha L. Slattery[17], Andrew T. Chan[18,19,20,21], Li Li[22], Graham G. Giles[4,23,24], Roger L. Milne[4,23,24], Stephen B. Gruber[6], Gad Rennert[25,26,27], Michael O. Woods[28], Steven J. Gallinger[29], Sonja Berndt[3], Richard B. Hayes[30], Wen-Yi Huang[3], Alicja Wolk[31], Emily White[1,2], Hongmei Nan[32], Rami Nassir[33], Noralane M. Lindor[34], Juan P. Lewinger[35], Andre E. Kim[6], David Conti[6], W. James Gauderman[6], Daniel D. Buchanan[36,37], Ulrike Peters[1,2] & Li Hsu[2,38]

[1]Department of Epidemiology, University of Washington, Seattle, WA, USA. [2]Public Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA. [3]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [4]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC, Australia. [5]Department of Medicine, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [6]Department of Preventive Medicine and USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [7]School of Public Health, University of Washington, Seattle, WA, USA. [8]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA. [9]University of Hawaii Cancer Center, Honolulu, HI, USA. [10]Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, GA, USA. [11]Oncology Data Analytics Program, Catalan Institute of Oncology-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain. [12]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [13]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona, Spain. [14]ONCOBEL Program, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona, Spain. [15]Center for Public Health Research, Massey University, Wellington, New Zealand. [16]Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA. [17]Department of Internal Medicine, University of Utah, Salt Lake City, UT, USA. [18]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [19]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [20]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [21]Broad Institute of Harvard and MIT, Cambridge, MA, USA. [22]Department of Family Medicine, University of Virginia, Charlottesville, VA, USA. [23]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, VIC, Australia. [24]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC, Australia. [25]Department of Community Medicine and Epidemiology, Lady Davis Carmel Medical Center, Haifa, Israel. [26]Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel. [27]Clalit National Cancer Control Center, Haifa, Israel. [28]Memorial University of Newfoundland, Discipline of Genetics, St. John's, Canada. [29]Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, ON, Canada. [30]Division of Epidemiology, Department of Population Health, New York University School of Medicine, New York, NY, USA. [31]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [32]IU Melvin and Bren Simon Cancer Center, Indiana University, Indianapolis, IN, USA. [33]Department of Pathology, School of Medicine, Umm Al-Qura'a University, Makkah, Saudi Arabia. [34]Department of Health Science Research, Mayo Clinic, Scottsdale, AZ, USA. [35]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [36]Colorectal Oncogenomics Group, Department of Clinical Pathology, The University of Melbourne, Parkville, VIC 3010, Australia. [37]University of Melbourne Centre for Cancer Research, Victorian Comprehensive Cancer Centre, Parkville, VIC 3010, Australia. [38]Department of Biostatistics, University of Washington, Seattle, WA, USA. ✉email: cameron.b.haas@gmail.com

Observational studies have shown higher folate consumption to be associated with lower risk of colorectal cancer (CRC). Understanding whether and how genetic risk factors interact with folate could further elucidate the underlying mechanism. Aggregating functionally relevant genetic variants in set-based variant testing has higher power to detect gene–environment (G × E) interactions and may provide information on the underlying biological pathway. We investigated interactions between folate consumption and predicted gene expression on colorectal cancer risk across the genome. We used variant weights from the PrediXcan models of colon tissue-specific gene expression as a priori variant information for a set-based G × E approach. We harmonized total folate intake (mcg/day) based on dietary intake and supplemental use across cohort and case–control studies and calculated sex and study specific quantiles. Analyses were performed using a mixed effects score tests for interactions between folate and genetically predicted expression of 4839 genes with available genetically predicted expression. We pooled results across 23 studies for a total of 13,498 cases with colorectal tumors and 13,918 controls of European ancestry. We used a false discovery rate of 0.2 to identify genes with suggestive evidence of an interaction. We found suggestive evidence of interaction with folate intake on CRC risk for genes including glutathione S-Transferase Alpha 1 (*GSTA1*; p = 4.3E−4), Tonsuko Like, DNA Repair Protein (*TONSL*; p = 4.3E−4), and Aspartylglucosaminidase (*AGA*: p = 4.5E−4). We identified three genes involved in preventing or repairing DNA damage that may interact with folate consumption to alter CRC risk. Glutathione is an antioxidant, preventing cellular damage and is a downstream metabolite of homocysteine and metabolized by *GSTA1*. *TONSL* is part of a complex that functions in the recovery of double strand breaks and *AGA* plays a role in lysosomal breakdown of glycoprotein.

Folate is a naturally occurring, water-soluble B vitamin that cannot be produced by the human body and plays a key role in DNA formation and is necessary for cellular division and tissue differentiation. It is found abundantly in green leafy vegetables, legumes, fruits, and its more potent form, folic acid, is found in supplements and fortified foods[1]. Supplementary folic acid is routinely prescribed during pregnancy as an evidence-based intervention to prevent neural tube defects in utero[2,3]. Dietary deficiency is typically found in persons subsisting on inadequate diets, as well as chronic alcoholics with diminished absorption[4]. Fortification of grains with folic acid began in the early 1990s to prevent nutritional deficiencies[5,6]. To date, 71 countries have legislative mandates for including folate in the fortification of milled grains[5]. Results pre- and post-fortification and risk of CRC have been somewhat inconsistent[7–13], suggesting that folate might play a more complex role in colorectal carcinogenesis through various interactions[14–16]. Given the complexity of the relationship between CRC and folate, there is a need to elucidate the underlying biological mechanisms and possible differential risk based on individual genetics[15].

Increased folic acid consumption is known to lower circulating levels of homocysteine, a common amino acid that has been associated with numerous diseases[6,17,18]. The absence of folic acid leads to impaired DNA synthesis and disturbances in red blood cell maturation[19]. Due to its role as a carrier of one-carbon groups and in folate-mediated one-carbon metabolism (FOCM), insufficient folate consumption has been implicated as a possible cause of cancer[12,20–23]. Consistent with this hypothesis previous studies have shown evidence that greater folate intake is associated with a reduced risk of colorectal adenomas and cancers (CRC)[11,21,24]. A pooled analysis of 13 prospective studies in 2010 observed a modest effect, estimating a 2% risk reduction for CRC per 100 µg/day increase in total folate consumption[25].

Candidate gene approaches targeting FOCM-related genes have shown associations with CRC risk[24,26,27]. This has raised interest in studying interactions between folate and genetic variants[23,28]. As such, it has been hypothesized that germline mutations to the enzyme 5,10-methylenetetrahydrofolate reductase (MTHFR) would be a driver of the effects on folate on CRC risk[11,29,30]. A common mutation, 677TT in MTHFR has been associated with a greater decreased risk of CRC in high consumers of folate and low alcohol consumption[27,29,31,32] compared to lower folate consumers. However, such analyses have relied on the assumption that FOCM-related genes are the driving genetic force on the pathway from folate consumption to CRC development. A genome-wide approach has the potential to identify novel genes that may modify the folate–CRC association.

To this end, we conducted a novel set-based genome-wide analysis to test interactions between genes and total folate intake on CRC risk. By using a set-based approach we may increase the power to detect associations, which is a common issue in traditional gene–environment interaction studies. We incorporate functional annotation based weights from PrediXcan, a transcriptome prediction tool[33].

Methods

Study participants.

We used epidemiological and genetic data from studies included in three international CRC consortia: the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), the Colorectal Transdisciplinary Study (CORECT) and the Colon Cancer Family Registry (CCFR). Full details have been published previously[34,35], and the demographic characteristics of study participants are summarized in Table 1. We describe the study designs in Supplementary Table 1A and present results for the study design specific effects of total folate on CRC for study designs in Supplementary Table 1B. In case–control study designs, included cases were ascertained using population-based sampling and age-matched controls. In prospective cohorts, cases were identified through linkage to cancer registries. Participants with non-European ancestry were excluded due to

	Cases	Controls	p-value
N	13,498	13,918	
Male (%)	6190 (45.9%)	6014 (43.2%)	< 0.001*
Mean reference age (SD)	65.0 (10.5)	65.0 (9.9)	0.971
BMI in kg/m^2 (%)			
Normal (18.5–24.9)	4415 (34.0%)	5260 (39.0%)	< 0.001*
Overweight (25–30)	5355 (41.2%)	5451 (40.4%)	
Obese (≥ 30)	3215 (24.8%)	2782 (20.6%)	
Mean total energy consumption in kcal/day (SD)	1938.3 (769.1)	1911.3 (718.7)	0.003*
Sex-study specific quantile of total folate consumption in mcg/day (%)			
First quantile	3399 (25.2%)	3176 (22.8%)	< 0.001*
Second quantile	3089 (22.9%)	3235 (23.2%)	
Third quantile	4120 (30.5%)	4162 (29.9%)	
Fourth quantile	2890 (21.4%)	3345 (24.0%)	
Alcohol consumption in g/day (%)			< 0.001*
Nondrinker	7031 (56.2%)	6852 (52.6%)	
1–28 g/day	4464 (35.7%)	5250 (40.3%)	
> 28 g/day	1017 (8.1%)	924 (7.1%)	
Smoking status			0.626
Never smoker	6452 (48.6%)	6736 (49.4%)	
Former smoker	5271 (39.8%)	5328 (39.1%)	
Smoker	1530 (11.5%)	1560 (11.5%)	

Table 1. Characteristics of participants from all studies by colorectal cancer case/control status and Chi-square p-values for statistical differences. Continuous variables were presented as mean and standard deviation (SD) and p-values were calculated using Student's *t* test for difference in means; categorical variables were presented as n (%) and p-values were calculated using Pearson Chi-square test. *Statistically significant difference between cases and controls.

small sample sizes among those with genetic data. Informed consent was given by all participants, and studies were approved by their respective Institutional Review Boards and complies with all relevant ethical regulations.

Genotype data. Details on genotyping and imputation have been reported previously[36]. In brief, DNA was mostly obtained from blood samples, with some from buccal swabs. Several platforms (the Illumina HumanHap 300k, 240k, 550k and OncoArray 610k BeadChip Array system, or Affymetrix platform) were used for genotyping[37,38]. Samples were excluded on the basis of sample call rate ≤ 97%, heterozygosity, unexpected duplicates or relative pairs, gender discrepancy and principal component analysis (PCA) outlier of HapMap2 CEU cluster. SNPs were excluded on the basis of inconsistency across platforms, call rate < 98%, and out of Hardy–Weinberg equilibrium (HWE) in controls ($p < 0.0001$)[37]. SNPs were imputed to the CEU population in Haplotype Reference Consortium (HRC version r1.0) if not directly genotyped[39], and restricted by imputation accuracy ($R^2 > 0.3$).

Genetically predicted gene expression. The sets of genetic variants and weights for predicting gene expression were downloaded from the publicly available PredictDB Repository (https://hakyimlab.org/resource/predixcan/). The weights for the predicted gene expression were obtained by an elastic net penalized regression approach using the genome-wide variant data and transcriptome data from 169 colon tissue samples from the GTEx project (GTEx v6)[40] (Supplementary material). We restricted GTEx data to the transverse colon as it included the entire colonic wall and as such the epithelial layer in the mucosa most relevant to CC development while the GTEx sigmoid colon data only included the muscle layer. Genes for which SNPs explained at least 1% of the variation in CRC risk were selected for interaction analyses. A total of 4839 genes were included.

Exposure assessment. Basic demographics and environmental risk factors were collected using in-person interviews and/or structured questionnaires[35,41–49]. For these data, we carried out a multi-step data harmonization procedure, reconciling each study's unique protocols and data-collection instruments as discussed previously[34]. Folate and folic acid intake were assessed at the reference time using food frequency questionnaires (FFQs). For cohort studies, the reference time was time of enrollment or blood collection. Folate and folic acid intake in each study were determined based on micrograms per day (mcg/day) of folate from foods (i.e., dietary folate) and mcg/day of folic acid from supplements (single or multivitamins) when available. Only two of the 23 studies with dietary folate intake did not capture information regarding supplemental folate. To account for the higher bioavailability of synthetic folic acid vs. natural folate in foods, we calculated total folate intake as dietary folate equivalents (total mcg DFE = mcg of dietary folate + 1.7 × mcg folic acid from supplements)[50]. Because the time of enrollment for some studies overlapped or followed the period of folic acid fortification (1996–

1998), these studies accounted for folic acid fortification when calculating dietary folate intake and entered dietary folate intake as mcg of natural food folate + 1.7 × mcg folic acid from fortified food (see Supplementary Table 1A). Two studies (OFCCR, DALS) entered supplement data as regular user vs. nonuser; for these, we assumed regular use was 400 mcg/day or 400 mcg/tablet (for multivitamins), which corresponds to the generic dose in supplements[25,51]. The primary analysis used sex-study specific quartiles of total folate using controls based on the calculated daily dietary and supplemental intake, if available. By using categorical sex-study specific quartiles we reduce the influence of outliers and skewed distributions and is consistent with the Cancer Cohort Pooling Project[52]. To further explore the differences in bioavailability, secondary analyses we explored sex-study specific quartiles of dietary folate and binary (yes/no) supplemental folate separately.

Statistical analysis. We used the Mixed effects Score Tests for interaction (MiSTi)[53], a mixed effects score test for gene-based interaction test with folate consumption on CRC risk, to conduct a pooled analysis across all studies. MiSTi modeled the gene–environmental interaction effect by two components. The fixed effects component incorporates variant functional information from PrediXcan as weights with our genotype data to calculate the genetically predicted gene expression and then assess its interaction with folate consumption. The random effects component involves residual interaction effects that have not been accounted for by the fixed effects. We used sex- and study-specific quantiles of folate consumption. p-values were calculated separately for fixed and random effects interaction terms, after adjusting for age, sex, study, sex-study specific quartiles of total energy consumption in kcal, and principal components to account for population stratification. We used the MiSTi data-adaptive weighted combination approach to combine the fixed and random effects components.

Genes with p-values less than the Bonferroni correction (0.05/4839 = 1.03E−5) were considered genome-wide statistically significant for an interaction with folate. p-values that reached false discovery rate (FDR) at 20% were considered having suggestive evidence of interaction as it is less stringent than a Bonferroni threshold. We conducted follow-up analyses based on the fixed and random effects p-values. For associations driven by the fixed effects, we investigated the direction and magnitude of these interactions using the generalized linear model, which included all covariates in the original model, folate, standardized predicted gene expression, and an interaction term for folate and predicted gene expression. Genes for which the signal was driven by the random effects component were further investigated to identify individual variants of the gene set as drivers using the same approach with interactions for individual variants and folate while adjusting for all other variants in the gene set. Due to some of the variants having high collinearity, we pruned variants by $R^2 < 0.9$.

All analyses were performed using R version 4.0.1[54].

We performed these additional follow-up analyses for MTHFR, as prior candidate gene studies have shown variants, specifically the C677T mutation, alter the association between folate and CRC[31,32,55–58]. We additionally include the results of the gene–environment interaction between rs1801133 (C677T mutation) per additional effect allele with sex-study specific quantiles of total folate consumption on colorectal cancer.

Results

The final sample included 13,498 cases and 13,918 controls with both folate and energy consumption measures available from 23 studies. We present demographic characteristics of all samples and report on measures for factors associated with CRC risk for study participants by case–control status in Table 1. Cases were more likely to be male, have higher BMI, and report consuming less folate daily and more calories daily compared to controls. Multivariable logistic regression estimated a reduced risk of CRC per-quartile increase in total folate intake, adjusting for sex, age at reference, and total energy intake, and study (OR = 0.91, 95% CI: 0.89, 0.93, p-trend < 0.001, Supplementary Table 1B). Sensitivity analyses included further adjustment for smoking and alcohol consumption, which had little effect on the estimates for total folate and CRC risk.

We found no suggestion of interaction between predicted gene expression for the *MTHFR* gene and sex-study specific folate on risk of CRC in our analysis. Supplementary Table 2A–C present follow-up analyses conducted to test the interaction per standard deviation change in predicted gene expression within sex-study specific quantiles, allowing for a non-linear relationship between folate quantiles, as well as individual variant weights used in the modeling of predicted gene expression to capture the C677T mutation. In the snp-environment interaction analysis for the rs1801133 variant (C677T mutation), no interaction was show between each additional effect allele with sex-study specific quantiles of total folate consumption on risk of CRC (ratio of odds ratio = 1.02; 95% CI = 0.98, 1.06; interaction p-value = 0.235).

The median number of SNPs included in the gene sets was 25 (minimum: 1, inter-quartile range [IQR]: 13–43, maximum: 277). Figure 1 displays the quantile–quantile plot for the G × E test that combined both fixed and random effects using adaptive weight. While there was no G × E interaction that reached the Bonferroni threshold (0.05/4839), three did surpass the false discover rate (FDR) of 0.2.

We present the findings with p-values that surpassed the FDR threshold for gene interactions with total folate consumption and CRC risk in Table 2. We observed suggestive evidence of interactions between total folate intake and 3 independent gene sets on risk of CRC at FDR < 0.2, including Glutathione S-Transferase Alpha 1 (*GSTA1*; p = 4.3E−4), Tonsuko Like, DNA Repair Protein (*TONSL*; p = 4.3E−4), and Aspartylglucosaminidase (*AGA*; p = 4.5E−4). In follow-up analyses for these three genes we observed positive interactions for *GST1A* and *AGA*, showing greater risk for CRC associated with higher gene expression and increasing folate consumption (Table 3). As the signal for *TONSL* primarily came from the random effects, indicating one or a few variants were drivers of the association, we investigated the individual interactions of variants with sex-study specific folate. We see two variants as possible drivers of the signal in our main analysis, 8:144964455_T/C and 8:144965104, as shown in Table 4.

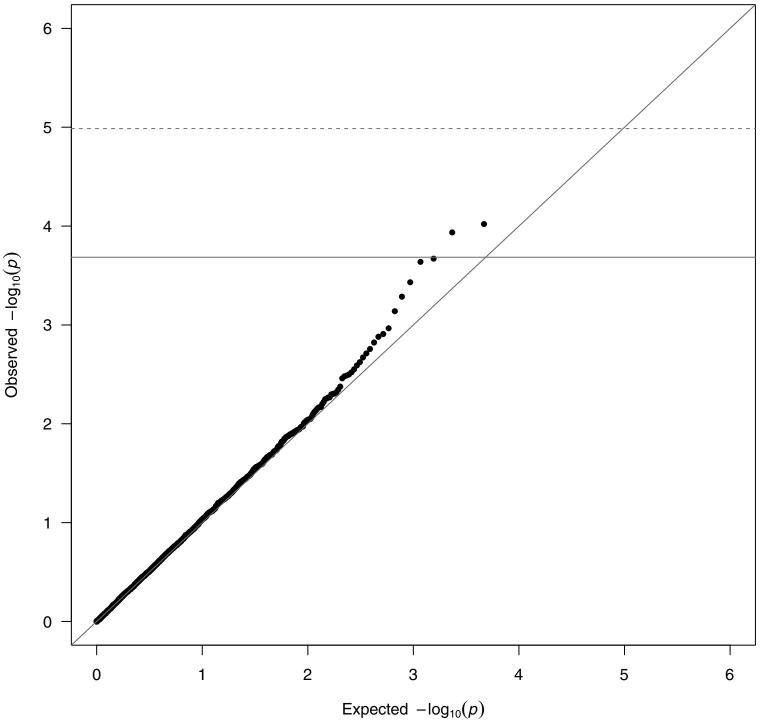

Figure 1. MiSTi results for Adaptive Weight test of predicted gene expression interactions with total folate intake. Dashed line is the Bonferroni corrected threshold, solid is the false discovery rate < 0.2 threshold for p-value significance.

Gene	Chromosome	R^2	Number of SNPs	Fixed effects	Random effects	Adaptive weight P
GSTA1	6	0.17	20	1.6E−4	0.11	4.3E−4
TONSL	8	0.06	28	0.048	1.3E−3	4.3E−4
AGA	4	0.37	39	1.2E−4	0.65	4.5E−4

Table 2. Top genes from MiSTi for gene-based interactions with total folate for colorectal cancer based on Adaptive Weight Test p-values with FDR < 0.2. Models were adjusted for age, sex, study, sex-study specific quartiles of total energy consumption in kcal, and principal components to account for population stratification.

Gene	Sex-study specific quantile of total folate	OR	95% CI	Interaction p-value*
GSTA1	1	1.29	0.86, 1.92	
	2	1.32	0.88, 1.98	0.48
	3	1.40	0.93, 2.10	0.01
	4	1.45	0.97, 2.17	6.5E−4
AGA	1	1.06	0.79, 1.43	
	2	1.16	0.85, 1.58	0.01
	3	1.15	0.85, 1.56	0.02
	4	1.21	0.89, 1.64	3.6E−4

Table 3. Estimated odds ratios (ORs) and 95% confidence intervals (CIs) of colorectal cancer risk per standard deviation change in predicted gene expression stratified by sex-study specific quantiles of total folate consumption. Models were adjusted for age, sex, study, sex-study specific quartiles of total energy consumption in kcal, and principal components to account for population stratification. *p-value tests for difference in quantile specific OR estimate for per standard deviation change in predicted gene expression and the lowest sex-study specific quantile of total folate quantile on colorectal cancer risk.

Variant**	Ratio of odds ratio	Standard error	Interaction p-value
8:144757296_A/G	1.04	0.22	0.86
8:144801243_C/T	1.03	0.02	0.09
8:144801593_C/A	0.93	0.04	0.10
8:144856443_G/T	0.82	0.13	0.13
8:144964455_T/C	0.95	0.02	7.0E−3
8:144965104_G/A	0.95	0.02	5.1E−3
8:145651888_G/A	1.00	0.02	0.78
8:145653006_G/A	1.00	0.02	0.76
8:145668042_G/A	1.01	0.02	0.76
8:146216936_A/G	0.96	0.02	0.03
8:146280802_C/T	1.03	0.02	0.19

Table 4. Variant specific ratio of odds ratios associated with per quantile increase in sex-study specific quantiles of total folate consumption on colorectal cancer risk for variants included in TONSL*. Models were adjusted for age, sex, study, sex-study specific quartiles of total energy consumption in kcal, and principal components to account for population stratification. *Filtering to uncorrelated variants in TONSL ($r^2 < 0.9$). **Chromosomal position and reference/alternative alleles.

Discussion

In this sizable analysis including a large number of studies we harmonized data on folate consumption and genome-wide genetic data to investigate interactions between folate intake and variants in genes on CRC risk. We observed an inverse association between folate intake and CRC risk across 23 studies. Using our novel statistical set-based G × E mixed effects score tests, MiSTi, we identified 3 genes with suggestive interactive effects with total folate consumption on CRC risk: *GSTA1*, *TONSL*, and *AGA*.

We observed a positive interaction between the predicted gene expression of *GSTA1* and folate for CRC risk. *GSTA1* located at 6p12.2 encodes for an enzyme that functions in cellular detoxification of electrophilic compounds through glutathione metabolism. Electrophilic compounds include carcinogens, therapeutic drugs, environmental toxins, and products of oxidative stress. Glutathione is a product of homocysteine metabolism, a key amino acid correlated with folate intake, and is bound to free radicals by *GSTA1*[59]. Our results suggest that folate consumption may increase remethylation of homocysteine to methionine, thus reducing the production of glutathione need for DNA repair. Mutations in *GSTA1* could feasibly alter the binding affinity of glutathione to carcinogenic compounds, leading to variation in cancer susceptibility. Of the 20 SNPs included in our analyses of *GSTA1*, three of the alternative alleles result in missense mutations to the gene[60]. Compromised function of glutathione as an antioxidant due to mutations in *GSTA1* in conjunction with depleted levels of glutathione due to lower homocysteine levels may be a pathway to tumorigenesis[6,22]. Candidate gene studies have shown no association between *GSTA1* and colorectal cancer or adenoma risk[61,62]. However, previous studies have shown interactions between diet, such as cruciferous vegetable consumption, and *GSTA1* genotypes, supporting that associations between this gene and CRC are likely driven by dietary exposures[63–65].

TONSL in the 8q24.3 region codes for a 1378 amino acid protein component of the MMS22L-TONSL complex, which functions in recovery of damaged replication forks[66]. Numerous mutations in *TONSL* are considered pathogenic[60]. Low levels of the MMS22L-TONSL complex result in increased frequency of DNA double-strand breaks and compromised DNA integrity[66]. In combination with increased DNA damage due to deficiencies in folic acid, impaired functionality of the MMS22L-TONSL due to functional mutations may be a pathway to increase tumorigenesis. Follow-up analyses further suggested that possible associations may be primarily driven by a small subset of variants included in our gene set in the main analysis.

We observed increasing risk of CRC per standard deviation increase in predicted gene expression of *AGA* with increasing folate consumption. The *AGA* gene is in the 4q34.3 region and codes for a 346 amino acid protein that functions in pathways related to the innate immune system and asparagine degradation[67]. Once the protein is processed into the mature enzyme it takes part in the catabolism of N-linked oligosaccharides, cleaving asparagine from *N*-acetylglucosamines in one of the final steps in the lysosomal breakdown of glycoproteins. Mutations in the *AGA* gene are known to cause the lysosomal storage disease aspartylglycosaminuria, eventually resulting in neurodegeneration. Previous research has not indicated a link to cancer for this gene.

While we have many strengths in performing the largest investigation of gene–folate interactions to date using a powerful set-based approach that allows to account for functional prediction, some limitations should be considered when interpreting these findings. Approximately half of the studies in our consortium ascertained cases using a cohort study design which may have resulted in earlier and more frequent detection of tumors. Most cohort studies in our consortium used population-based registries for case ascertain. However, one study, The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial, was a randomized trial to determine the effectiveness of screening. While we have adjusted for study in our approach there may be unknown residual effects of this design. Our study population was limited to those of European descent. As gene expression levels may differ across populations of different ancestry, our results may not be generalizable to populations of non-European ancestry. The studies included in our analysis occurred over a range of time and geographic locations. Fortification with folate occurred in different places at different times and we used adjusted dietary equivalents to account for these differences (see Supplementary Table 1A). Study designs also varied. We looked at the effect size of folate on CRC by case/control versus cohort study designs and did not find a substantive difference to justify stratified analyses (see Supplementary Table 1B). Lastly, studies in our consortium generally ascertained folate consumption through standard questionnaires. However, previous work has shown self-reported measures of folate intake to be positively and moderately correlated with plasma levels of folate, particularly when dietary supplement use was included as was generally the case in studies included in our analyses[68].

We utilized colon-specific gene expression data, specifically transverse colon tissue captured by the GTex Project[40]. One limitation of this data is the diversity of cell types aside from epithelial cells of the mucosa of the colon, from which CRC derives given that the entire colonic wall was sampled. The impact of this would cause a dilution of gene expression for the tissue most relevant for CRC. However, we expect this to be an improvement over alternative tissue types including blood or sigmoid colon tissues in GTEx, which were collected from muscle tissues only and would not represent the gene expression profile of interest.

Although MiSTi is a powerful statistical tool, which accounts for both fixed- and random-effects of the gene–folate interaction, none of our findings reached the Bonferroni corrected threshold, which can be overly conservative as many genes are co-expressed. We did not perform independent replication and thus follow-up investigations are warranted, as a FDR of 0.2 should be considered liberal[53]. The previously suggested *MTHFR* gene was not identified in our analysis[27]. However, in using the penalized elastic net to create our predicted gene expression the C677T was not included in the variant weights due to the insignificant contribution to regulation of gene expression. While it was also not seen in the gene–environment interaction analysis either, we believe these results to be representative of an agnostic approach which has not been shown before, as opposed to candidate gene studies.

Our analysis was conducted in the largest pooled analysis of a well characterized and harmonized consortium of CRC with comprehensive genetic data which enabled a hypothesis-free genome-wide investigation of interactions with folate consumption on CRC risk. An extensive number of genes evaluated in prior candidate gene–folate interaction studies, including *MTHFR*, were included among the 4839 genes examined. However, none of those previously hypothesized genes were found to interact with folate consumption in our analysis[31,57,69]. We conducted additional follow-up analysis for *MTHFR* using indicator terms for sex-study-specific folate quantiles and interaction terms for all quantiles with predicted gene expression were null (see Supplementary Table 2A–C). No previous study has agnostically tested for genetic interactions with folate for cancer. Our statistical approach was potentially improved by incorporating functional variant weights and testing gene-sets rather than individual SNPs reducing the penalty for multiple testing. In the end, we found three genes that were suggestive of interacting with folate consumption on risk of CRC, supporting the hypothesis that associations of folate with CRC may be modified by common genetic variation.

The biological functions of our top genes serve to primarily prevent or repair DNA damage. The combined effects of increased DNA damage due to folate deficiencies and compromised functionality of these genes may be an important pathway in CRC tumorigenesis. These findings, particularly for *GSTA1*, warrant follow-up in future studies with comprehensive genetic and data on folate intake in order to confirm the potential role of these genes in interacting with folate on CRC risk.

Data availability

Data will be made available upon request and approval by contacting Dr. Ulrike Peters.

Code availability

Please contact the corresponding author for code.

Received: 19 November 2021; Accepted: 31 October 2022
Published online: 07 November 2022

References

1. Food and Drug Administration. Food standards: amendment of standards of identity for enriched grain products to require addition of folic acid. *Fed Regist*, 8781–8807 (1996).
2. Crider, K. S. *et al.* Population red blood cell folate concentrations for prevention of neural tube defects: Bayesian model. *BMJ* **349**, g4554 (2014).
3. Choumenkovitch, S. F. *et al.* Folic acid intake from fortification in United States exceeds predictions. *J. Nutr.* **132**, 2792–2798 (2002).
4. Finglas, P. M. Dietary reference intakes for thiamin, riboflavin, niacin, vitamin B6, folate, vitamin B12, pantothenic acid, biotin and choline. https://doi.org/10.1016/S0924-2244(01)00010-3 (2000).
5. Initiative FF. FFI—Global Progress. http://ffinetwork.org/global_progress/index.php (Accessed 7 Dec2017).
6. Hoey, L. *et al.* Effect of a voluntary food fortification policy on folate, related B vitamin status, and homocysteine in healthy adults. *Am. J. Clin. Nutr.* **86**, 1405–1413 (2007).
7. Hirsch, S. *et al.* Colon cancer in Chile before and after the start of the flour fortification program with folic acid. *Eur. J. Gastroenterol. Hepatol.* **21**, 436–439 (2009).
8. Lee, J. E. *et al.* Folate intake and risk of colorectal cancer and adenoma: Modification by time. *Am. J. Clin. Nutr.* **93**, 817–825 (2011).
9. Mason, J. B. *et al.* A temporal association between folic acid fortification and an increase in colorectal cancer rates may be illuminating important biological principles: A hypothesis. *Cancer Epidemiol. Biomark. Prev.* **16**, 1325–1329 (2007).
10. Kennedy, D. A. *et al.* Folate intake and the risk of colorectal cancer: A systematic review and meta-analysis. *Cancer Epidemiol.* **35**, 2–10 (2011).
11. Giovannucci, E. Epidemiologic studies of folate and colorectal neoplasia: A review. *J. Nutr.* **132**, 2350S-2355S (2002).
12. Zschäbitz, S. *et al.* B vitamin intakes and incidence of colorectal cancer: Results from the Women's Health Initiative Observational Study cohort. *Am. J. Clin. Nutr.* **97**, 332–343 (2013).
13. Weinstein, S. J. *et al.* One-carbon metabolism biomarkers and risk of colon and rectal cancers. *Cancer Epidemiol. Biomark. Prev.* **17**, 3233–3240 (2008).

14. Kok, D. E. *et al.* Bacterial folate biosynthesis and colorectal cancer risk: More than just a gut feeling. *Crit. Rev. Food Sci. Nutr.* **60**, 244–256 (2020).
15. Ulrich, C. M. & Potter, J. D. Folate and cancer—Timing is everything. *JAMA* **297**, 2408 (2007).
16. Mason, J. B. & Tang, S. Y. Folate status and colorectal cancer risk: A 2016 update. *Mol. Aspects Med.* **53**, 73–79 (2017).
17. Bailey, L., Stover, P., Mcnulty, H., Fenech, M., Gregory, J., Mills, J. *et al.* Biomarkers of nutrition for development-folate review. *J. Nutr.* **145**, 1–5. http://search.proquest.com/docview/1695233723/ (2015).
18. Hao, L. *et al.* Folate status and homocysteine response to folic acid doses and withdrawal among young Chinese women in a large-scale randomized double-blind trial. *Am. J. Clin. Nutr.* **88**, 448–457 (2008).
19. Kruman, I. I. *et al.* Folic acid deficiency and homocysteine impair DNA repair in hippocampal neurons and sensitize them to amyloid toxicity in experimental models of Alzheimer's disease. *J. Neurosci.* **22**, 1752–1762 (2002).
20. Kim, Y. Folate and DNA methylation: A mechanistic link between folate deficiency and colorectal cancer?. *Cancer Epidemiol. Biomark. Prev.* **13**, 511–519 (2004).
21. Kim, Y.-I. Folate and colorectal cancer: An evidence-based critical review. *Mol. Nutr. Food Res.* **51**, 267–292 (2007).
22. Hanley, M. P. & Rosenberg, D. W. One-carbon metabolism and colorectal cancer: Potential mechanisms of chemoprevention. *Curr. Pharmacol. Rep.* **1**, 197–205 (2015).
23. Farias, N. *et al.* The effects of folic acid on global DNA methylation and colonosphere formation in colon cancer cell lines. *J. Nutr. Biochem.* **26**, 818–826 (2015).
24. Figueiredo, J. C., Levine, A. J., Crott, J. W., Baurley, J. & Haile, R. W. Folate-genetics and colorectal neoplasia: What we know and need to know next. *Mol. Nutr. Food Res.* **57**, 607–627 (2013).
25. Kim, D.-H. *et al.* Pooled analyses of 13 prospective cohort studies on folate intake and colon cancer. *Cancer Causes Control* **21**, 1919–1930 (2010).
26. Figueiredo, J. C. *et al.* Genome-wide diet-gene interaction analyses for risk of colorectal cancer. *PLoS Genet.* **10**, e1004228 (2014).
27. Kim, J. W. *et al.* Association between folate metabolism-related polymorphisms and colorectal cancer risk. *Mol. Clin. Oncol.* **3**, 639–648 (2015).
28. Montazeri, Z. *et al.* Systematic meta-analyses, field synopsis and global assessment of the evidence of genetic association studies in colorectal cancer. *Gut* **69**, 1460–1471 (2020).
29. Ma, J. *et al.* Methylenetetrahydrofolate reductase polymorphism, dietary interactions, and risk of colorectal cancer. *Cancer Res.* **57**, 1098–1102 (1997).
30. Nazki, F. H., Sameer, A. S. & Ganaie, B. A. Folate: Metabolism, genes, polymorphisms and the associated diseases. *Gene* **533**, 11–20 (2014).
31. Le Marchand, L., Wilkens, L. R., Kolonel, L. N. & Henderson, B. E. The MTHFR C677T polymorphism and colorectal cancer: The multiethnic cohort study. *Cancer Epidemiol. Biomark. Prev.* **14**, 1198–1203 (2005).
32. Slattery, M. L., Potter, J. D., Samowitz, W., Schaffer, D. & Leppert, M. Methylenetetrahydrofolate reductase, diet, and risk of colon cancer. *Cancer Epidemiol. Biomark. Prev.* **8**, 513–518 (1999).
33. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
34. Hutter, C. M. *et al.* Characterization of gene–environment interactions for colorectal cancer susceptibility loci. *Cancer Res.* **72**, 2036–2044 (2012).
35. Newcomb, P. A. *et al.* Colon Cancer Family Registry: An international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol. Biomark. Prev.* **16**, 2331–2343 (2007).
36. Peters, U. *et al.* Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. *Gastroenterology* **144**, 799-807.e24 (2013).
37. Peters, U. *et al.* Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum. Genet.* **131**, 217–234 (2012).
38. Zanke, B. W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989–994 (2007).
39. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
40. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
41. Slattery, M. L. *et al.* Energy balance and colon cancer–beyond physical activity. *Cancer Res.* **57**, 75–80 (1997).
42. Christen, W. G., Gaziano, J. M. & Hennekens, C. H. Design of Physicians' Health Study II—A randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. *Ann. Epidemiol.* **10**, 125–134 (2000).
43. Prorok, P. C. *et al.* Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Control Clin. Trials* **21**, 273S-309S (2000).
44. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin. Trials* **19**, 61–109 (1998).
45. Hoffmeister, M., Raum, E., Krtschil, A., Chang-Claude, J. & Brenner, H. No evidence for variation in colorectal cancer risk associated with different types of postmenopausal hormone therapy. *Clin. Pharmacol. Ther.* **86**, 416–424 (2009).
46. Brenner, H., Chang-Claude, J., Seiler, C. M., Rickert, A. & Hoffmeister, M. Protection from colorectal cancer after colonoscopy: A population-based, case-control study. *Ann. Intern. Med.* **154**, 22–30 (2011).
47. Küry, S. *et al.* Combinations of cytochrome P450 gene polymorphisms enhancing the risk for sporadic colorectal cancer related to red meat consumption. *Cancer Epidemiol. Biomark. Prev.* **16**, 1460–1467 (2007).
48. Colditz, G. A. & Hankinson, S. E. The Nurses' Health Study: Lifestyle and health among women. *Nat. Rev. Cancer* **5**, 388–396 (2005).
49. Giovannucci, E. *et al.* Aspirin use and the risk for colorectal cancer and adenoma in male health professionals. *Ann. Intern. Med.* **121**, 241–246 (1994).
50. Suitor, C. W. & Bailey, L. B. Dietary folate equivalents: Interpretation and application. *J. Am. Diet Assoc.* **100**, 88–94 (2000).
51. Giovannucci, E. *et al.* Multivitamin use, folate, and colon cancer in women in the Nurses' Health Study. *Ann. Intern. Med.* **129**, 517–524 (1998).
52. Swerdlow, A. J. *et al.* The National Cancer Institute Cohort Consortium: An international pooling collaboration of 58 cohorts from 20 countries. *Cancer Epidemiol. Biomark. Prev.* **27**, 1307–1319 (2018).
53. Su, Y.-R., Di, C.-Z., Hsu, L., Genetics and Epidemiology of Colorectal Cancer Consortium. A unified powerful set-based test for sequencing data analysis of G × E interactions. *Biostatistics* **18**, 119–131 (2017).
54. Team RC. *R: A language and environment for statistical computing*, Vienna. http://www.r-project.org/.
55. Dong, L. M. *et al.* Genetic susceptibility to cancer: The role of polymorphisms in candidate genes. *JAMA* **299**, 2423–2436 (2008).
56. Levine, A. J. *et al.* Genetic variability in the MTHFR gene and colorectal cancer risk using the colorectal cancer family registry. *Cancer Epidemiol. Biomark. Prev.* **19**, 89–100 (2010).
57. Ulrich, C. M. *et al.* Colorectal adenomas and the C677T MTHFR polymorphism: Evidence for gene–environment interaction?. *Cancer Epidemiol. Biomark. Prev.* **8**, 659–668 (1999).
58. Torre, M. L. *et al.* MTHFR C677T polymorphism, folate status and colon cancer risk in acromegalic patients. *Pituitary* **17**, 257–266 (2014).

59. Lushchak, V. I. Glutathione homeostasis and functions: Potential targets for medical interventions. *J. Amino Acids* **2012**, 736837 (2012).
60. Stelzer, G. *et al.* The GeneCards Suite: From gene data mining to disease genome sequence analyses. In *Current Protocols in Bioinformatics* (ed. Vanitha, M.) 1.30.1-1.30.33 (Wiley, 2016).
61. van der Logt, E. M. J. *et al.* Genetic polymorphisms in UDP-glucuronosyltransferases and glutathione S-transferases and colorectal cancer risk. *Carcinogenesis* **25**, 2407–2415 (2004).
62. Economopoulos, K. P. & Sergentanis, T. N. GSTM1, GSTT1, GSTP1, GSTA1 and colorectal cancer risk: A comprehensive meta-analysis. *Eur. J. Cancer* **46**, 1617–1631 (2010).
63. Coles, B. *et al.* The role of human glutathione *S*-transferases (hGSTs) in the detoxification of the food-derived carcinogen metabolite *N*-acetoxy-PhIP, and the effect of a polymorphism in hGSTA1 on colorectal cancer risk. *Mutat. Res.* **482**, 3–10 (2001).
64. Sweeney, C., Coles, B. F., Nowell, S., Lang, N. P. & Kadlubar, F. F. Novel markers of susceptibility to carcinogens in diet: Associations with colorectal cancer. *Toxicology* **181–182**, 83–87 (2002).
65. Tijhuis, M. J. *et al.* GSTP1 and GSTA1 polymorphisms interact with cruciferous vegetable intake in colorectal adenoma risk. *Cancer Epidemiol. Biomark. Prev.* **14**, 2943–2951 (2005).
66. O'Donnell, L. *et al.* The MMS22L-TONSL complex mediates recovery from replication stress and homologous recombination. *Mol. Cell* **40**, 619–631 (2010).
67. GeneCard. Cytochrome P450 Family 2 Subfamily D Member 6. http://www.genecards.org/cgi-bin/carddisp.pl?gene=CYP2D6#aliases_descriptions (Accessed 12 May2018).
68. Park, J. Y. *et al.* Dietary intake and biological measurement of folate: A qualitative review of validation studies. *Mol. Nutr. Food Res.* **57**, 562–581 (2013).
69. Sharp, L. & Little, J. GENOME OF EPIDEMIOLOGY polymorphisms in genes involved in folate metabolism and colorectal. *Neoplasia* **159**, 423–443 (2004).

Acknowledgements

CPS-II: The authors thank the CPS-II participants and Study Management Group for their invaluable contributions to this research. The authors would also like to acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention National Program of Cancer Registries, and cancer registries supported by the National Cancer Institute Surveillance Epidemiology and End Results program. The Colon CFR graciously thanks the generous contributions of their 42,500 study participants, dedication of study staff, and the financial support from the U.S. National Cancer Institute, without which this important registry would not exist. The content of this manuscript does not necessarily reflect the views or policies of the NIH or any of the collaborating centers in the CCFR, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government, any cancer registry, or the CCFR. Harvard cohorts (HPFS, NHS, PHS): The study protocol was approved by the institutional review boards of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health, and those of participating registries as required. We would like to thank the participants and staff of the HPFS, NHS and PHS for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. The authors assume full responsibility for analyses and interpretation of these data. Kentucky: We would like to acknowledge the staff at the Kentucky Cancer Registry. PLCO: The authors thank the PLCO Cancer Screening Trial screening center investigators and the staff from Information Management Services Inc and Westat Inc. Most importantly, we thank the study participants for their contributions that made this study possible. SFCCR: The authors would like to thank the study participants and staff of the Hormones and Colon Cancer and Seattle Cancer Family Registry studies (CORE Studies). WHI: The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf.

Disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer / World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

Author contributions

C.H. wrote the main manuscript text and analyses. All authors reviewed the manuscript.

Funding

This was supported by grant number T32 CA094880, R01-CA201407, and U01-CA164930 from the National Institutes of Health. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NCI, NIH. Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO): National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA164930, U01 CA137088, R01 CA059045, R01201407). Genotyping/Sequencing services were provided by the Center for Inherited Disease Research (CIDR) (X01-HG008596 and X-01-HG007585). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268201200008I. This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA015704. The ATBC Study is supported by the Intramural Research Program of the U.S. National Cancer Institute, National Institutes of Health, Department of Health and Human Services. COLO2&3: National Institutes of Health (R01 CA60987). The Colon Cancer Family Registry (CCFR, http://www.coloncfr.org) is supported in part by funding from the National Cancer Institute (NCI), National Institutes of Health (NIH) (award U01 CA167551). The CCFR Set-1 (Illumina 1M/1M-Duo) and Set-2 (Illumina Omni1-Quad) scans were supported by NIH awards U01 CA122839 and R01 CA143247 (to GC). The CCFR Set-3 (Affymetrix Axiom

CORECT Set array) was supported by NIH award U19 CA148107 and R01 CA81488 (to SBG). The CCFR Set-4 (Illumina OncoArray 600K SNP array) was supported by NIH award U19 CA148107 (to SBG) and by the Center for Inherited Disease Research (CIDR), which is funded by the NIH to the Johns Hopkins University, contract number HHSN268201200008I. The SFCCR Illumina HumanCytoSNP array was supported through NCI award R01 CA076366 (to PAN). Additional funding for the OFCCR/ARCTIC was through award GL201-043 from the Ontario Research Fund (to BWZ), award 112746 from the Canadian Institutes of Health Research (to TJH), through a Cancer Risk Evaluation (CaRE) Program grant from the Canadian Cancer Society (to SG), and through generous support from the Ontario Ministry of Research and Innovation. Colorectal Cancer Transdisciplinary (CORECT) Study: The CORECT Study was supported by the National Cancer Institute, National Institutes of Health (NCI/NIH), U.S. Department of Health and Human Services (grant numbers U19 CA148107, R01 CA81488, P30 CA014089, R01 CA197350, P01 CA196569, R01 CA201407) and National Institutes of Environmental Health Sciences, National Institutes of Health (grant number T32 ES013678). CPS-II: The American Cancer Society funds the creation, maintenance, and updating of the Cancer Prevention Study-II (CPS-II) cohort. This study was conducted with Institutional Review Board approval. DALS: National Institutes of Health (R01 CA48998 to M. L. Slattery). Harvard cohorts (HPFS, NHS, PHS): HPFS is supported by the National Institutes of Health (P01 CA055075, UM1 CA167552, U01 CA167552, R01 CA137178, R01 CA151993, and R35 CA197735), NHS by the National Institutes of Health (R01 CA137178, P01 CA087969, UM1 CA186107, R01 CA151993, and R35 CA197735) and PHS by the National Institutes of Health (R01 CA042182). Kentucky: This work was supported by the following grant support: Clinical Investigator Award from Damon Runyon Cancer Research Foundation (CI-8); NCI R01CA136726. MCCS cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further supported by Australian NHMRC grants 509348, 209057, 251553 and 504711 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry (VCR) and the Australian Institute of Health and Welfare (AIHW), including the National Death Index and the Australian Cancer Database. MEC: National Institutes of Health (R37 CA54281, P01 CA033619, and R01 CA063464). MECC: This work was supported by the National Institutes of Health, U.S. Department of Health and Human Services (R01 CA81488 to SBG and GR). NFCCR: This work was supported by an Interdisciplinary Health Research Team award from the Canadian Institutes of Health Research (CRT 43821); the National Institutes of Health, U.S. Department of Health and Human Serivces (U01 CA74783); and National Cancer Institute of Canada grants (18223 and 18226). The authors wish to acknowledge the contribution of Alexandre Belisle and the genotyping team of the McGill University and Génome Québec Innovation Centre, Montréal, Canada, for genotyping the Sequenom panel in the NFCCR samples. Funding was provided to Michael O. Woods by the Canadian Cancer Society Research Institute. OFCCR/ARCTIC: Additional funding for the OFCCR/ARCTIC was through award GL201-043 from the Ontario Research Fund (to BWZ), award 112746 from the Canadian Institutes of Health Research (to TJH), through a Cancer Risk Evaluation (CaRE) Program grant from the Canadian Cancer Society (to SG), and through generous support from the Ontario Ministry of Research and Innovation. PLCO: Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. Funding was provided by National Institutes of Health (NIH), Genes, Environment and Health Initiative (GEI) Z01 CP 010200, NIH U01 HG004446, and NIH GEI U01 HG 004438. Swedish Mammography Cohort and Cohort of Swedish Men: This work is supported by the Swedish Research Council /Infrastructure grant, the Swedish Cancer Foundation, and the Karolinska Institute´s Distinguished Professor Award to Alicja Wolk. VITAL: National Institutes of Health (K05 CA154337). WHI: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C. SPAIN: Colorectal Cancer Genetics and Genomics, Spanish study was supported by Instituto de Salud Carlos III, co-funded by FEDER funds—a way to build Europe– (grants PI14-613 and PI09-1286), Agency for Management of University and Research Grants (AGAUR) of the Catalan Government (grant 2017SGR723), and Junta de Castilla y León (grant LE22A10-2). Sample collection of this work was supported by the Xarxa de Bancs de Tumors de Catalunya sponsored by Pla Director d'Oncología de Catalunya (XBTC), Plataforma Biobancos PT13/0010/0013 and ICOBIOBANC, sponsored by the Catalan Institute of Oncology. PMH-SCCFR: National Institutes of Health (R01 CA076366 to P. Newcomb and U01 CA074794 to J. Potter).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-23451-y.

Correspondence and requests for materials should be addressed to C.B.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022