

Genome-wide interaction study of dietary intake of fibre, fruits, and vegetables with risk of colorectal cancer



Nikos Papadimitriou,^a Andre Kim,^b Eric S. Kawaguchi,^b John Morrison,^b Virginia Diez-Obrero,^{c,d,e,f} Demetrius Albanes,^g Sonja I. Berndt,^g Stéphane Béziau,^h Stephanie A. Bien,ⁱ D Timothy Bishop,^j Emmanouil Bouras,^k Hermann Brenner,^{l,m,n} Daniel D. Buchanan,^{o,p,q} Peter T. Campbell,^r Robert Carreras-Torres,^{d,s} Andrew T. Chan,^{t,u,v,w,x,y} Jenny Chang-Claude,^{z,aa} David V. Conti,^b Matthew A. Devall,^{ab,ac} Niki Dimou,^a David A. Drew,^v Stephen B. Gruber,^{ad} Tabitha A. Harrison,ⁱ Michael Hoffmeister,ⁱ Jeroen R. Huyghe,ⁱ Amit D. Joshi,^x Temitope O. Keku,^{ae} Anshul Kundaje,^{af,ag} Sébastien Küry,^h Loïc Le Marchand,^{ah} Juan Pablo Lewinger,^b Li Li,^{ai} Brigid M. Lynch,^{aj,ak} Victor Moreno,^{c,d,e,f} Christina C. Newton,^{al} Mireia Obón-Santacana,^{c,d,e} Jennifer Ose,^{am,an} Andrew J. Pellatt,^{ao} Anita R. Peoples,^{am,an} Elizabeth A. Platz,^{ap} Conghui Qu,ⁱ Gad Rennert,^{aq,ar,as} Edward Ruiz-Narvaez,^{at} Anna Shcherbina,^{af} Mariana C. Stern,^b Yu-Ru Su,ⁱ Duncan C. Thomas,^b Claire E. Thomas,ⁱ Yu Tian,^{z,au} Konstantinos K. Tsilidis,^{k,av} Cornelia M. Ulrich,^{am,an} Caroline Y. Um,^{al} Kala Visvanathan,^{ap} Jun Wang,^{b,aw} Emily White,^{i,ax} Michael O. Woods,^{ay} Stephanie L. Schmitz,^{az,ba} Finlay Macrae,^{bb} John D. Potter,ⁱ John L. Hopper,^{aj,bc} Ulrike Peters,^{i,ax,be,*} Neil Murphy,^{a,be} Li Hsu,^{i,bd,be,**} Marc J. Gunter,^{a,av,be,***} and W. James Gauderman^{b,be,****}

^aNutrition and Metabolism Branch, International Agency for Research on Cancer, Lyon, France

^bDivision of Biostatistics, Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

^cUnit of Biomarkers and Susceptibility, Oncology Data Analytics Program, Catalan Institute of Oncology, Barcelona, 08908, Spain

^dColorectal Cancer Group, ONCOBELL Program, Bellvitge Biomedical Research Institute, Barcelona, 08908, Spain

^eConsortium for Biomedical Research in Epidemiology and Public Health, Barcelona, 08908, Spain

^fDepartment of Clinical Sciences, Faculty of Medicine and Health Sciences and Universitat de Barcelona Institute of Complex Systems (UBICS), University of Barcelona (UB), L'Hospitalet de Llobregat, 08908, Barcelona, Spain

^gDivision of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

^hService de Génétique Médicale, Centre Hospitalier Universitaire (CHU) Nantes, Nantes, France

ⁱPublic Health Sciences Division, Fred Hutchinson Cancer Center, Seattle, WA, USA

^jLeeds Institute of Cancer and Pathology, University of Leeds, Leeds, UK

^kDepartment of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina, Greece

^lDivision of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany

^mDivision of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumour Diseases (NCT), Heidelberg, Germany

ⁿGerman Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

^oColorectal Oncogenomics Group, Department of Clinical Pathology, Melbourne Medical School, The University of Melbourne, Parkville, Australia

^pUniversity of Melbourne Centre for Cancer Research, The University of Melbourne, Parkville, Australia

^qGenomic Medicine and Family Cancer Clinic, The Royal Melbourne Hospital, Parkville, Victoria, Australia

^rDepartment of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA

^sDigestive Diseases and Microbiota Group, Girona Biomedical Research Institute (IDIBGI), 17190 Salt, Girona, Spain

^tDivision of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

^uChanning Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

^vClinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

^wBroad Institute of Harvard and MIT, Cambridge, MA, USA

^xDepartment of Epidemiology, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA

^yDepartment of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA

^zDivision of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

^{aa}University Medical Centre Hamburg-Eppendorf, University Cancer Centre Hamburg (UCC), Hamburg, Germany

^{ab}Center for Public Health Genomics, Department of Public Health Sciences, University of Virginia, Charlottesville, VA, USA

^{ac}Department of Public Health Sciences, Center for Public Health Genomics, Charlottesville, VA, USA

*Corresponding author. Fred Hutchinson Cancer Center, 1100 Fairview Avenue N, M4-B402, Seattle, WA, 98109-1024, USA.

**Corresponding author. Fred Hutchinson Cancer Center, 1100 Fairview Avenue N, M4-B402, Seattle, WA, 98109-1024, USA.

***Corresponding authors. Department of Epidemiology and Biostatistics, Imperial College London, School of Public Health, St.Mary's Campus, Norfolk Place, London, W2 1PG, UK.

****Corresponding author. Department of Population and Public Health Sciences, Keck School of Medicine, University of Southern California, 1975, Zonal Ave, Los Angeles, CA, 90033, USA.

E-mail addresses: upeters@fredhutch.org (U. Peters), lih@fredhutch.org (L. Hsu), m.gunter@imperial.ac.uk (M.J. Gunter), jimg@usc.edu (W.J. Gauderman).

^{be}These authors contributed equally as senior authors.

- ^{ad}Department of Medical Oncology & Therapeutics Research and Center for Precision Medicine, City of Hope National Medical Center, Duarte, CA, USA
- ^{ae}Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, NC, USA
- ^{af}Department of Genetics, Stanford University, Stanford, CA, USA
- ^{ag}Department of Computer Science, Stanford University, Stanford, CA, USA
- ^{ah}University of Hawaii Cancer Center, Honolulu, HI, USA
- ^{ai}Department of Family Medicine, University of Virginia, Charlottesville, VA, USA
- ^{aj}Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Victoria, Australia
- ^{ak}Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia
- ^{al}Department of Population Science, American Cancer Society, Atlanta, GA, USA
- ^{am}Huntsman Cancer Institute, Salt Lake City, UT, USA
- ^{an}Department of Population Health Sciences, University of Utah, Salt Lake City, UT, USA
- ^{ao}Department of Cancer Medicine, University of Texas MD Anderson Cancer Center, Houston, TX, USA
- ^{ap}Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
- ^{aq}Department of Community Medicine and Epidemiology, Lady Davis Carmel Medical Center, Haifa, Israel
- ^{ar}Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa, Israel
- ^{as}Clalit National Cancer Control Center, Haifa, Israel
- ^{at}Department of Nutritional Sciences, University of Michigan School of Public Health, Ann Arbor, MI, USA
- ^{au}School of Public Health, Capital Medical University, Beijing, China
- ^{av}Department of Epidemiology and Biostatistics, Imperial College London, School of Public Health, London, UK
- ^{aw}Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA
- ^{ax}Department of Epidemiology, University of Washington School of Public Health, Seattle, WA, USA
- ^{ay}Memorial University of Newfoundland, Discipline of Genetics, St. John's, Canada
- ^{az}Genomic Medicine Institute, Cleveland Clinic, Cleveland, OH, USA
- ^{ba}Population and Cancer Prevention Program, Case Comprehensive Cancer Center, Cleveland, OH, USA
- ^{bb}The Royal Melbourne Hospital, Melbourne, Victoria, Australia
- ^{bc}Department of Epidemiology, School of Public Health and Institute of Health and Environment, Seoul National University, Seoul, South Korea
- ^{bd}Department of Biostatistics, University of Washington, Seattle, WA, USA

eBioMedicine

2024;104: 105146

Published Online 14 May 2024

<https://doi.org/10.1016/j.ebiom.2024.105146>

1016/j.ebiom.2024.105146

Summary

Background Consumption of fibre, fruits and vegetables have been linked with lower colorectal cancer (CRC) risk. A genome-wide gene-environment (G × E) analysis was performed to test whether genetic variants modify these associations.

Methods A pooled sample of 45 studies including up to 69,734 participants (cases: 29,896; controls: 39,838) of European ancestry were included. To identify G × E interactions, we used the traditional 1-degree-of-freedom (DF) G × E test and to improve power a 2-step procedure and a 3DF joint test that investigates the association between a genetic variant and dietary exposure, CRC risk and G × E interaction simultaneously.

Findings The 3-DF joint test revealed two significant loci with p-value < 5 × 10⁻⁸. Rs4730274 close to the *SLC26A3* gene showed an association with fibre (p-value: 2.4 × 10⁻³) and G × fibre interaction with CRC (OR per quartile of fibre increase = 0.87, 0.80, and 0.75 for CC, TC, and TT genotype, respectively; G × E p-value: 1.8 × 10⁻⁷). Rs1620977 in the *NEGR1* gene showed an association with fruit intake (p-value: 1.0 × 10⁻⁸) and G × fruit interaction with CRC (OR per quartile of fruit increase = 0.75, 0.65, and 0.56 for AA, AG, and GG genotype, respectively; G × E p-value: 0.029).

Interpretation We identified 2 loci associated with fibre and fruit intake that also modify the association of these dietary factors with CRC risk. Potential mechanisms include chronic inflammatory intestinal disorders, and gut function. However, further studies are needed for mechanistic validation and replication of findings.

Funding National Institutes of Health, National Cancer Institute. Full funding details for the individual consortia are provided in acknowledgments.

Copyright © 2024 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND IGO license (<http://creativecommons.org/licenses/by-nc-nd/3.0/igo/>).

Keywords: Diet; Fibre; Gene-environment interaction; Colorectal cancer; GWAS

Research in context

Evidence before this study

It is unknown whether genetic polymorphisms modify the associations of fruits, vegetables and dietary fibre with colorectal cancer risk as current evidence is limited. Previous genome-wide diet-gene interaction studies were underpowered due to insufficient sample size. Additionally, new statistical techniques have recently been developed that aim to further increase the statistical power of the interaction analyses.

Added value of this study

In this large-scale genome-wide interaction analysis, we found two $G \times E$ interactions for fibre, fruits, and colorectal cancer risk. The most significant finding was rs4730274 close to the

SLC26A3 gene, which provides supportive evidence for an interaction between fibre consumption, chronic inflammatory intestinal disorders, overall gut function, and colorectal cancer risk development. A second signal involved rs1620977 in the *NEGR1* gene and this has been linked with obesity and food preference.

Implications of all the available evidence

Our study identified a genetic polymorphism that could modify the protective effect of dietary fibre on colorectal cancer risk through various mechanisms. Additional studies are needed to understand functional implications and to replicate these findings both in European ancestry and other racial/ethnic populations.

Introduction

Colorectal cancer (CRC) is one of the most common cancer types at a global level, responsible for almost 2 million new cancer cases and over 900,000 related deaths in 2020.¹ Current evidence suggests that high consumption of fruits, vegetables, wholegrains and foods containing dietary fibre are associated with a lower CRC risk.² The World Cancer Research Fund classified the evidence for relationship between wholegrains and foods containing dietary fibre with CRC as strong while the evidence for fruits and vegetables and CRC associations was classified as limited.²

Recent genome-wide association studies (GWAS) have identified over 200 independent loci associated with CRC risk explaining up to 35% of total heritability according to twin studies.^{3–8} It has been suggested that gene-environment interactions ($G \times E$) might be able to explain some of this missing heritability.⁹ However, previous genome-wide diet-gene interaction analyses have reported only one significant interaction between rs4143094 (10p14/near *GATA3*) and intake of processed meat.^{10,11} Other, earlier studies focusing solely on CRC susceptibility loci did not find evidence of strong gene-diet interactions, except for an interaction for vegetable consumption and rs16892766, on chromosome 8q23.3, near the genes *EIF3H* and *UTP23*.^{12,13} A recent umbrella review of gene environment evidence and CRC assigned this finding a weak plausibility score due to the strong genetic effect and the weak environmental effects of vegetable intake on CRC risk.¹⁴ Currently, we are still at the early stages of exploring gene-diet interactions and prior studies were limited by the relatively small sample sizes.

These prior studies which conducted traditional interaction analyses were probably underpowered due to

insufficient sample size (16,739 to 18,509 participants). Additional new statistical techniques have recently been developed that aim to increase the statistical power of the interaction analyses including joint tests, which consider both main and interaction effects, and two-step methods, which apply a filtering step before the interaction testing thus prioritizing single nucleotide polymorphisms (SNPs) and decreasing the burden of multiple testing.^{15–17}

Accordingly, the aim of our study was to identify new gene-environment interactions for the consumption of fruits, vegetables, and fibre with risk of CRC cancer by applying both traditional and modern techniques to a large-scale dataset of up to 69,734 participants from three genetic consortia with information on the consumption of fruits, vegetables, and dietary fibre.

Methods

Study population

Up to 45 studies of individuals of European ancestry from three CRC genetic consortia: the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), the Colorectal Cancer Transdisciplinary Study (CORECT) and the Colon Cancer Family Registry (CCFR) were included in this analysis (Supplemental Table S1).³ For cohort studies, nested case-control sets were assembled via risk-set sampling, whereas cancer-free controls were used for case-control studies. Controls were mostly matched on age at enrolment or diagnosis, sex, race, and enrolment date/trial group, when applicable. Cases were defined as colorectal adenocarcinoma and were confirmed by medical records, pathology reports, or death certificate information. In total, 69,599 (cases: 29,820; controls: 39,779), 69,734 (cases: 29,896;

controls: 39,838), and 44,890 (cases: 20,749; controls: 24,141) participants were included in the analyses of fruits, vegetables, and fibre respectively. Analyses were limited to individuals of European ancestry, based on self-reported race and clustering of principal components with the 1000 Genomes EUR population.

Exposure definition

Information on dietary intake of fruits, vegetables, and total fibre was ascertained through food frequency questionnaires and diet histories while data on demographics and additional factors were collected using in-person interviews and/or structured questionnaires within a period ranging from three months to two years prior to diagnosis for case-control studies and at enrolment for cohort studies. For both fruits and vegetables consumption, within each study multiple questions were used to derive the total consumption of each dietary variable. Most studies reported fruits and vegetables consumption as servings per day. For total fibre, grams/day was the most common measure across contributing studies and it was derived by multiplying intakes of all fibre containing foods in the dietary questionnaires by the fibre content data in nutrient databases. All data were centrally harmonised at Fred Hutchinson Cancer Center.¹² In summary, a multi-step approach was implemented for the data harmonization using a priori defined common data elements (CDEs) to which all questionnaires and data dictionaries were mapped to. Definitions, permissible values, and standardized coding were implemented into a single database via SAS and T-SQL and the overall data were checked for potential outlying values and other errors. The harmonized consumption data were expressed either as servings per day (fruits, vegetables) or grams per day (fibre) and entered into the analysis as sex- and study-specific quartiles, where the quartile groups were coded with the median value of the quartile within each study and sex. In the case-control studies, the study and sex-specific quartiles were based on the control distribution as this best represents the underlying study populations.

Genotyping and imputation

Details on quality control and genotyping were previously published and the genotyping arrays used are summarized in [Supplemental Table S1](#).^{3,18} In brief, SNPs were excluded for: missing call rate more than 2–5%, departure from Hardy–Weinberg equilibrium (HWE) ($p < 1 \times 10^{-4}$), differences between self-reported and genotypic sex, and discordant genotype calls within duplicate samples. All individual studies genotyped using the same build (GRCh37). Autosomal SNPs were imputed to the Haplotype Reference Consortium (HRC) r1.1 panel through the University of Michigan Imputation Server and converted into a binary format for data

management and analyses using the BinaryDosage R package (<https://cran.r-project.org/web/packages/BinaryDosage>).^{19,20} The imputed SNPs were filtered based on a pooled minor allele frequency (MAF) of less than 1% and imputation accuracy R^2 over 0.8. Overall, 7,250,911 SNPs that passed the quality control and were present in all participating studies were included in our analysis. Principal component analysis (PCA) for population stratification assessment was performed using PLINK1.9 on 30,000 randomly sampled imputed SNPs with MAF >5% and $R^2 > 0.99$.

Statistics

Main effects

Logistic regression models adjusted for age (continuous), sex, total energy consumption (kcal/day, continuous divided by 1000), and three population-stratification principal components were used to estimate the association between fibre, fruits, and vegetables intakes and CRC cancer in each of the participating studies. The study-specific results were then meta-analyzed using random-effects models (Hartung-Knapp) to calculate the overall summary odds ratios (ORs) and 95% confidence intervals (CIs).²¹ Presence of heterogeneity was examined and quantified using the Cochran's Q and I^2 statistic respectively.²² Additional stratified analyses were conducted to assess the relationship between the three exposures with CRC risk by sex, tumour site (proximal colon, distal colon, or rectum), and study design (case-control or cohort study). All meta-analyses were performed using the R package Meta.

Interaction effects

Genome-wide interaction scans were performed to identify interaction effects using the R package $G \times E$ ScanR (<https://cran.r-project.org/web/packages/GxEScanR>), which implements a series of different tests such as the traditional logistic regression models, case-only analyses and joint tests of G-E association. Imputed SNP dosages were modelled as continuous variables and a p-value threshold of 5×10^{-8} was chosen to denote the significant associations.²³ In the context of the current study the following notation was used to describe the methods used. E corresponds to the exposures of interest (fibre, fruits, or vegetables), G corresponds to SNPs, D corresponds to CRC status, and C corresponds to the adjustment covariates in the models. Logistic regression models were implemented to test for multiplicative scale interaction: $\text{logit}(\text{Pr}(D = 1|G)) = \beta_0 + \beta_G G + \beta_E E + \beta_{G \times E} G \times E + \beta_C C$. The adjustment covariates included age at the reference time, sex, total energy intake (kcal/day), three population stratification principal components, and study as a fixed effect which adjusts for the potential for the individual studies to act as potential confounders. In this case we tested the hypothesis $H_0 : \beta_{G \times E} = 0$, which corresponds to the multiplicative interaction. Apart from this traditional

approach we also applied two more advanced $G \times E$ techniques as our primary tests. The 3-degrees of freedom (DF) test further expands the main hypothesis and considers both the primary D|G the G|E associations in the combined case/control population: $H_0: \beta_G = \beta_{G \times E} = \delta_G = 0$, where δ_G represents the association between G and E in a combined case-control sample, in an effort to increase the overall statistical power.²⁴ This test examines a fundamentally different hypothesis than a $G \times E$ analysis since significant results can occur due to strong D|G effect, $G \times E$ effect, and/or G|E correlation. However, as it has been shown that an underlying $G \times E$ interaction can induce both D|G and G|E associations in a case-control sample, the 3-DF test increases the chance that important interacting loci will be identified and can be further investigated in additional studies.²⁴ We also implemented a two-step approach that gives priority to potential interactions after weighting $G \times E$ interaction tests (step 2) based on ranks of an independent filtering statistic (step 1).^{25,26} Given the independence between steps 1 and 2, this approach can reduce the burden of multiple comparisons and increase the statistical power to detect significant interactions. In the weighted testing framework proposed by Ioanita-Laza et al.,²⁶ SNPs are partitioned based on their Step-1 p-value into exponentially larger bins, each assigned with an increasingly more stringent threshold for interaction test significance. Bins of size $k_1 = 5$, $k_2 = 2k_1$, $k_3 = 2k_2$ etc, were recommended with each bin having step 2 significance threshold $\frac{\alpha}{\binom{2}{5} \binom{4}{10}}$, $\frac{\alpha}{\binom{8}{20}}$ etc., where $\alpha = 0.05$. Therefore, SNPs with higher step 1 ranks were prioritized and interaction testing of these SNPs conducted at more liberal thresholds. The step 1 filtering statistic that was applied was a combination of the marginal D|G associations and the G|E association in a combined case-control population (EDGE).²⁵ Since many SNPs are in linkage disequilibrium, it is likely that the top bins are populated by correlated markers from the same locus. Therefore, we implemented a modified two-step approach accommodating for the correlated SNPs while properly controlling for type I error.²⁷ Specifically, SNPs were partitioned into bins based on the step 1 Pvalue thresholds, which were calculated based on the original predetermined bin sizes, assuming uniform distribution of 1 million independent tests. For step 2 GxE testing, we accounted for the influx of correlated markers into each bin by correcting for the effective number of tests, estimated using principal component analysis performed on genotype correlated matrices for each bin.²⁸ For any significant findings, the models were further adjusted for body mass index (BMI), diabetes, alcohol, red, and processed meat consumption to examine the robustness of the results.

We examined the extent of genomic inflation using quantile-quantile (Q-Q) plots and by calculating the genomic inflation factor (λ). Because λ

scales with sample size, we also calculated λ_{1000} , which corresponds to the genomic inflation factor for a study of 1000 cases and 1000 controls.^{29,30}

Regional plots were generated for significant hits which present the strength of association, the extent of association signal and linkage disequilibrium (LD) with other SNPs, as well as the position of loci in relation to genes in the region. The software LocusZoom v1.3 was used to generate these plots.³¹ Measures of LD were estimated using European populations of the 1000 Genomes Project study population. Potential expression quantitative trait loci (eQTL) relationships were explored using the Genotype-Tissue Expression (GTEx V8) and the University of Barcelona and University of Virginia genotyping and RNA sequencing project (BarcUVa-Seq)^{32,33} using the available online tool at <https://barcuvaseq.org/cotrex/> (accessed on January 2023). Based on the results from eQTL-gene associations, we then tested the predicted expression of the eQTL-associated gene of interest for an interaction with fibre, fruits, and vegetables consumption in data from the three consortia involved in this study.

Functional annotation plots were also created to visualize chromatin accessibility across the functional datasets and to plot $-\log_{10}$ (p-value) signal tracks. We used ATAC-seq, DNASE-seq, H3K27ac histone ChIP-seq, and H3K4me1 histone ChIP-seq datasets (Supplemental Table S2) of primary tissue from healthy colon and tumour primary tissue samples from Scacheri et al.,³⁴ as well as from three colorectal cancer cell lines (SW480, HCT116, COLO205). These datasets were processed through ENCODE ATAC-seq/DNASE-seq³⁵ and histone ChIP-seq pipelines³⁶ to perform alignment and peak calling. $-\log_{10}$ (p-value) tracks were extracted from the MACS2 step of the pipeline for visualization in genome browsers. Irreproducible Discovery Rate (IDR)³⁷ peak calls for ATAC-seq and DNASE-seq datasets, as well as naive overlap peak calls for histone ChIP-seq datasets, were determined from the ENCODE pipelines. The pyGenomeTracks³⁸ software package was used to visualize chromatin accessibility across the functional datasets and to plot $-\log_{10}$ (p-value) signal tracks. Peaks across samples from the same assay were concatenated across datasets, cropped to within 200 bp centered on the peak summit, and merged using bedtools³⁹ merge.

Interaction analyses for rare variants

As a separate secondary analysis, we further conducted $G \times E$ testing for rare variants. Interaction tests of the three exposures of interest and aggregated rare variant sets at the gene and enhancer level were performed using the Mixed effects Score Tests for interactions (MiSTi) method.⁴⁰ This regression framework includes the interaction between E and the burden component of rare variants as the fixed effect and heterogeneous $G \times E$ effects as random effects. A Fisher's combination

approach under MiSTi (fMiSTi) was applied to combine both the fixed and random effects for discovering $G \times E$ interactions (42), after adjusting for age at reference time, sex, study, and three population-stratification principal components. Since 25,000 genes were tested, a p-value threshold of 2×10^{-6} ($\alpha = 0.05/25,000$) was considered as statistically significant. The analysis was conducted using the MiSTi R package (42).

Ethics

All participants gave written informed consent and studies were approved by their respective Institutional Review Boards.

Role of funders

The funders had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; or the decision to submit the manuscript for publication.

Results

Consumption of fibre, fruits, and vegetables and CRC risk

Table 1 shows the baseline characteristics of the participants included in the three analyses. Cancer cases were older, had higher BMI and total energy intake, had a greater prevalence of family history of CRC, higher prevalence of type 2 diabetes, and were more likely to have ever smoked cigarettes. Additionally, CRC cases consumed less total dietary fibre (grams/day: 1.49 ± 1.14 versus 1.56 ± 1.14), fruits (servings/day: 1.34 ± 1.04 versus 1.45 ± 1.07), and vegetables (servings/day: 1.40 ± 1.03 versus 1.46 ± 1.08) than controls (p-value <0.001 for all three exposures).

In the meta-analyses, inverse associations were observed between fibre (OR per quartile increase = 0.79; 95% CI 0.74, 0.85), fruits (OR per quartile increase = 0.79; 95% CI 0.72, 0.86), and vegetables (OR per quartile increase = 0.82; 95% CI 0.73, 0.93) and CRC risk (Fig. 1). The inverse associations were similar by sex (minimum P-het: 0.26 for fibre intake) and cancer subsite (minimum P-het: 0.17 for fibre intake) (Fig. 1). There was evidence of heterogeneity in all three analyses which was driven by the case-control studies (minimum $I^2 = 55\%$; P-het <0.001) (Supplemental Figure S1). In general, the inverse associations were stronger for case-control studies (Fruits-OR = 0.69; 95% CI = 0.57, 0.83; Vegetables-OR = 0.66; 95% CI = 0.48, 0.91; Fibre-OR = 0.71; 95% CI = 0.64, 0.79) than cohort studies (Fruits-OR = 0.88; 95% CI = 0.83, 0.92; Vegetables-OR = 0.94; 95% CI = 0.89, 0.99; Fibre-OR = 0.86; 95% CI = 0.81, 0.92) (Supplemental Figure S1).

Interaction analysis results

Genomic control inflation and quantile-quantile (QQ) plots for the SNP-diet interactions for risk of CRC did

not show evidence for residual population stratification (Supplemental Figure S2).

Fibre

The 3-DF joint test revealed a significant hit for rs4730274, which maps upstream of the *SLC26A3* gene (Supplemental Figure S3a; 3-DF p-value: 3.8×10^{-8}). This variant was not directly associated with CRC (G|D p-value: 0.33), but there was a moderate association with fibre intake (G|E p-value: 2.4×10^{-3}), and an interaction with fibre intake on CRC risk ($G \times E$ p-value: 1.8×10^{-7}) (Table 2). Stratifying by genotype of rs4730274 showed that the strength of the inverse association between fibre and CRC increased with every copy of the T allele: OR per quartile of fibre increase = 0.87 for those with CC genotype; OR = 0.80 for those with TC genotype; OR = 0.75 for those with TT genotype (Table 3). The functional annotation analysis showed that rs4730274 and some SNPs in LD with it are in open chromatin suggesting enhancer activity in normal colon, cancer tissues, and CRC cell lines (Supplemental Figure S4a). Furthermore, several SNPs in LD with rs4730274 were eQTLs for *DLD* gene in the BarCUVa-Seq and GTEx transverse colon tissue data but not in GTEx sigmoid colon tissue data (Supplemental Tables S3–S5). We also detected statistically significant interactions between fibre and the expression levels of the *SLC26A3* and *DLD* genes in relation to CRC. For *SLC26A3* a negative interaction (p-value = 0.0052) was observed with the inverse association between fibre intake and CRC risk becoming stronger as the expression levels of *SLC26A3* increased, in contrast a positive interaction was observed for *DLD* (p-value = 0.0021) (Supplemental Table S6).

Fruits

Based on the 3-DF joint test, rs1620977 in *NEGR1* showed a highly significant combined effect (Supplemental Figure S3b; 3-DF p-value: 3.4×10^{-8}) with no direct CRC associations (D|G p-value: 0.77), but with a highly significant association with fruit intake (G|E p-value: 1.0×10^{-8}), and a modest interaction with fruits on risk of CRC ($G \times E$ p-value: 0.03) (Table 2). For this SNP the associations between fruit intake and CRC risk became stronger with every copy of the G allele: OR per quartile of fruit increase = 0.75 for those with AA genotype; OR = 0.65 for those with AG genotype; OR = 0.56 for those with GG genotype (Table 3). There was no evidence for enhancer activity for rs1620977 but a correlated SNP aligned with open chromatin regions (Supplemental Figure S4b). No eQTLs were observed for rs1620977 or other SNPs in LD with it (Supplemental Tables S3–S5).

Vegetables

The analysis for vegetables did not identify any interaction effects for either the 3-DF or the 2-step methods. Several hits were identified in the 3-DF analysis, but

	Fruits		Vegetables		Fibre	
	Controls	Cases	Controls	Cases	Controls	Cases
Age						
Mean (SD)	63.5 (9.4)	64.3 (10.8)	63.5 (9.4)	64.3 (10.8)	64.5 (9.7)	64.8 (10.5)
Sex						
Men	20,208	15,357	20,238	15,396	11,032	9877
Women	19,571	14,463	19,600	14,500	13,109	10,872
BMI						
Mean (SD)	27.0 (4.6)	27.4 (4.9)	27.0 (4.6)	27.4 (4.9)	26.8 (4.7)	27.4 (5.0)
Missing	1460 (3.7%)	1501 (5.0%)	1462 (3.7%)	1508 (5.0%)	473 (2.0%)	582 (2.8%)
Total energy intake Mean (SD)	1900 (719)	1960 (768)	1900 (719)	1970 (769)	1900 (716)	1960 (767)
Family history of CRC						
No	25,180 (63%)	20,612 (69%)	25,219 (63%)	20,676 (69%)	18,361 (76%)	14,613 (70%)
Yes	4002 (10%)	4043 (14%)	4019 (10%)	4062 (14%)	2356 (10%)	2822 (14%)
Missing	10,597 (27%)	5165 (17%)	10,600 (27%)	5158 (17%)	3424 (14.2%)	3314 (16.0%)
Education level (highest completed)						
Less than High School	8084 (20%)	7427 (25%)	8096 (20%)	7455 (25%)	6013 (25%)	6188 (30%)
High School/GED	6139 (15%)	6025 (20%)	6156 (15%)	6054 (20%)	4047 (17%)	3581 (17%)
Some College	9627 (24%)	6542 (22%)	9653 (24%)	6562 (22%)	6122 (25%)	4942 (24%)
College/Graduate School	12,088 (30%)	7968 (27%)	12,118 (30%)	7979 (27%)	7734 (32%)	5701 (27%)
Missing	3841 (11%)	1858 (6%)	3815 (11%)	1846 (6%)	225 (1%)	337 (2%)
Smoking, never/ever						
No	19,298 (49%)	13,525 (45%)	19,326 (49%)	13,530 (45%)	11,460 (47%)	9385 (45%)
Yes	19,869 (50%)	15,656 (53%)	19,896 (50%)	15,727 (53%)	12,224 (51%)	10,953 (53%)
Missing	612 (1%)	639 (2%)	616 (1%)	639 (2%)	457 (2%)	411 (2%)
T2D (ever diagnosed)						
No	34,970 (88%)	24,811 (83%)	35,024 (88%)	24,873 (83%)	20,516 (85%)	16,982 (82%)
Yes	3484 (9%)	3638 (12%)	3489 (9%)	3650 (12%)	2335 (10%)	2490 (12%)
Missing	1325 (3%)	1371 (5%)	1325 (3%)	1373 (5%)	1290 (5%)	1277 (6%)
Red meat (servings/day)						
Mean (SD)	0.55 (0.54)	0.63 (0.58)	0.55 (0.54)	0.63 (0.58)	0.66 (0.63)	0.66 (0.63)
Missing	189 (0.5%)	193 (0.6%)	191 (0.5%)	193 (0.6%)	102 (0.4%)	103 (0.5%)
Processed meat (servings/day)						
Mean (SD)	0.31 (0.35)	0.39 (0.44)	0.31 (0.35)	0.39 (0.44)	0.29 (0.36)	0.35 (0.42)
Missing	2952 (7%)	3211 (11%)	2977 (8%)	3238 (11%)	1047 (5%)	561 (3%)
Alcohol Consumption						
No	14,317 (36%)	12,826 (43%)	14,328 (36%)	12,856 (43%)	11,184 (46%)	10,271 (50%)
Low (1–28 g/day)	20,293 (51%)	12,751 (43%)	20,324 (51%)	12,781 (43%)	10,794 (45%)	8164 (39%)
Moderate (>28 g/day)	4766 (12%)	3867 (13%)	4779 (12%)	3880 (13%)	1906 (8%)	2084 (10%)
Missing	403 (1%)	376 (1%)	407 (1%)	379 (1%)	257 (1%)	230 (1%)
Total fruit intake						
Mean (SD)	1.45 (1.07)	1.34 (1.04)	1.45 (1.07)	1.34 (1.04)	1.51 (1.10)	1.39 (1.11)
Missing			196 (0.5%)	158 (0.5%)	264 (1%)	230 (1%)
Total vegetables intake						
Mean (SD)	1.46 (1.07)	1.40 (1.03)	1.46 (1.08)	1.40 (1.03)	1.53 (1.11)	1.47 (1.11)
Missing	137 (0.3%)	82 (0.3%)			214 (0.9%)	175 (0.8%)
Total fibre intake						
Mean (SD)	1.56 (1.11)	1.47 (1.12)	1.56 (1.11)	1.47 (1.12)	1.56 (1.14)	1.49 (1.14)
Missing	15,902 (40.0%)	9301 (31%)	15,911 (40%)	9322 (31%)		

g, grams; SD, standard deviation.

Table 1: Baseline characteristics of the participants by control-case status in the three analyses.

they were all driven by strong D|G or G|E effects with none showing a significant interaction effect (all $G \times E$ p-value >0.05, data not shown).

Secondary and sensitivity analyses

The interaction analysis for rare variants did not identify any significant interactions for any of the three

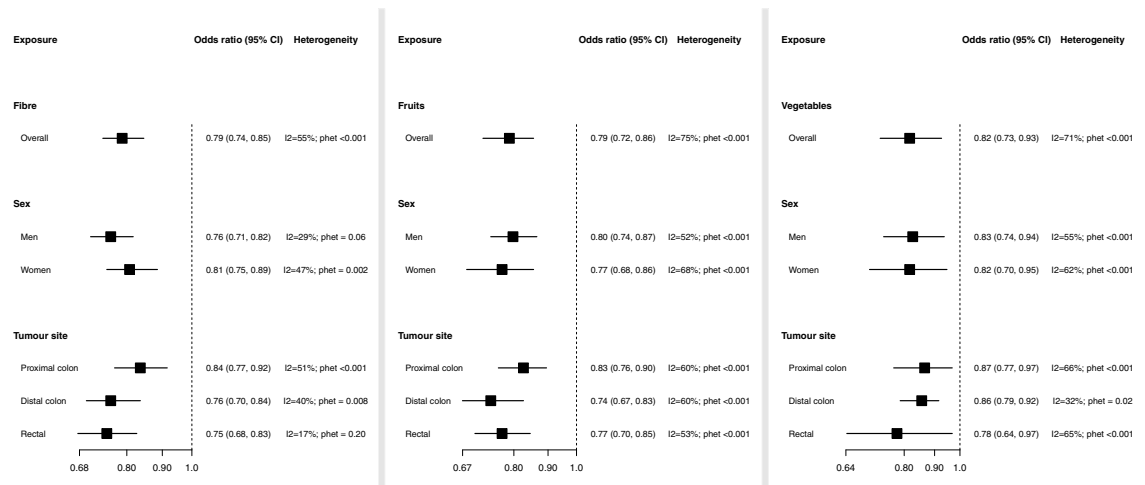


Fig. 1: Results from meta-analysis of association between (a) fibre, fruits, vegetables, and colorectal cancer, overall and stratified by sex and tumour site. Models adjusted for age, sex, and total energy intake. Intake of fibre (g/day), fruits, and vegetables (servings per day) were coded as median of intake sex/study specific quartiles, modelled as continuous variables.

exposures (min p-value = 4.3×10^{-6} in the fibre analysis). Further adjustments for seven additional confounders (i.e., smoking, diabetes, family history of CRC, BMI, alcohol, and red and processed meat) did not change the strength of the interactions (Supplemental Table S7).

Discussion

In this large-scale genome-wide interaction analysis of almost 70,000 participants, we found evidence of interactions between fibre intake and rs4730274 close to *SLC26A3* gene, in relation to CRC risk and an additional hit was observed for fruits intake and rs1620977 in the *NEGR1* gene.

The SNP rs4730274, which maps upstream of the *SLC26A3*, showed the most significant interaction effect in our analysis. This SNP has been linked with ulcerative colitis (UC) and inflammatory bowel disease (IBD) in a previous GWAS study of almost 60,000 participants, with the C allele being associated with a lower risk.⁴¹ Current evidence suggests that dietary fibre can help to maintain remission in patients with IBD and reduce lesions of the intestinal mucosa.⁴² It has also been reported that a high-fibre diet in patients with UC in remission decreased markers of inflammation and reduced intestinal dysbiosis in faecal samples.⁴³ On the

other hand, it is well documented that patients with IBD are at higher risk of CRC.⁴⁴ Therefore, it seems feasible that the protective effect of fibre is stronger in people with genetic liability to these chronic inflammatory intestinal disorders.

The *SLC26A3* gene is located on chromosome 7 and encodes a 764-amino acid protein that is primarily expressed in the digestive tract and more specifically mainly on the cell membrane inside the lumen.⁴⁵ *SLC26A3* is a member of the *SLC26A* transporter family, which includes multifunctional anion exchangers, and is involved in the regulation of Cl^- absorption and HCO_3^- secretion.⁴⁵ Mutations of this gene lead to congenital chloride diarrhoea (CLD), an autosomal recessive disorder that enhances colonic proliferation and up-regulation of ion transporters in the colon.⁴⁵ Down-regulation of *SLC26A3* has been observed in colon adenomas and adenocarcinomas, colon cancer cell lines, and has been linked with higher CRC incidence.^{46–48} Studies in mice provided further support for the tumour-suppressive role with *SLC26A3* knockout mice showing symptoms of CLD like humans and increased proliferation of colonic crypt epithelium denoting an important role of *SLC26A3* in colon tumourigenesis.⁴⁹ Therefore, *SLC26A3* probably functions as a tumour suppressor; however, the exact mechanisms are not well defined although it has been suggested that its ability to export

Exposure	SNP	Chr	Position	Gene	OA	EA	EAF	Method	p-value D G	p-value G E	p-value GxE	p-value 3DF
Fibre	rs4730274	7	107,479,719	upstream <i>SLC26A3</i>	C	T	0.52	Joint test (3DF)	0.33	2.4×10^{-3}	1.8×10^{-7}	3.8×10^{-8}
Fruits	rs1620977	1	72,729,142	<i>NEGR1</i>	A	G	0.71	Joint test (3DF)	0.77	1.0×10^{-8}	2.9×10^{-2}	3.4×10^{-8}

DF, degrees of freedom; SNP, single nucleotide polymorphism; Chr, chromosome; Position, base pair position based on NCBI Build37. ^aNo interaction effects were observed for vegetables.

Table 2: Main results from genome-wide interaction scans of fibre, fruits, and vegetables.^a

Fibre	rs4730274 (7:107,479,719)		
	CC (n = 10,335)	TC (n = 22,310)	TT (n = 12,245)
	0.87 (0.81, 0.94)	0.80 (0.76, 0.84)	0.75 (0.70, 0.80)
Fruits	rs1620977 (1:72,729,142)		
	AA (n = 5715)	AG (n = 28,368)	GG (n = 35,516)
	0.75 (0.59, 0.97)	0.65 (0.59, 0.73)	0.56 (0.50, 0.62)

CRC, colorectal cancer; SNP, single nucleotide polymorphism. ^aNo interaction effects were observed for vegetables.

Table 3: Associations between consumption of fibre, fruits, and vegetables^a and CRC risk, stratified by genotypes of SNPs identified in the interaction analysis.

HCO₃⁻ from the cells might play a role as this reduces intracellular alkalinization, which facilitates cell proliferation.⁵⁰ Finally, downregulation of *SLC26A3* may also be involved in the pathogenesis of UC.^{51,52} Given that chronic inflammation is a hallmark of cancer⁵³ and a prominent driver of CRC development⁵⁴ and the fact that patients with UC have a higher risk of CRC, this might be additional evidence of the tumour suppressive role of *SLC26A3*.

Current evidence suggests there could be a connection between consumption of dietary fibre and the expression levels of *SLC26A3* and CRC. More specifically, fibre promotes the growth of bacterial populations in the gut which in turn increases the expression of *SLC26A3*.^{55,56} Additionally, fibre is metabolized by gut microbiota to short chain fatty acids (SCFA) which have beneficial anti-inflammatory and anti-carcinogenic effects.⁵⁷ However, most SCFAs are ionized and transporters are needed for their absorption.⁵⁸ One of these transporters is *SLC26A3* which show affinities for all three major SCFA (acetate, propionate, and butyrate).⁵⁸ Therefore, there may be synergistic effects between fibre, expression levels of *SLC26A3*, and CRC which are also consistent with our finding that the inverse association between fibre intake and CRC risk is stronger with increased *SLC26A3* expression.

In our analysis we also found some strong normal colon eQTLs for SNPs in LD with rs4730274 and the *DLD* gene, which also showed an interaction effect with fibre on CRC risk. This gene is the third catalytic enzyme of three mitochondrial enzyme complexes: branched-chain alpha-ketoacid dehydrogenase (BCKDH); α -ketoglutarate dehydrogenase (α KGDH); and pyruvate dehydrogenase (PDH).⁵⁹ The main role of PDH is to convert pyruvate to acetyl-CoA as a part of the carbohydrate oxidation pathway.⁶⁰ There is a lack of evidence of changes in the expression of the genes encoding PDH due to cancer but PDH kinase 1 (PDK1), which is often overexpressed in cancer cells and has been strongly implicated in tumourigenesis, phosphorylates and inactivates PDH.⁶⁰ However, it is currently unclear how the *DLD* gene could interact with fibre

consumption and CRC, suggesting that *SLC26A3* is the more likely candidate gene.

We found that the association of fruits with CRC was modified by rs1620977 in the *NEGR1* gene. The SNP rs1620977 has been linked with BMI in previous GWAS with the G allele corresponding to lower values,^{61–63} however our sensitivity analyses after adjusting for additional confounders including BMI did not show any change in the original results. A recent GWAS reported that rs1620977 was associated with fruit consumption at a genome-wide significance level (p-value = 1.2×10^{-18}) with A allele been linked with higher intake.⁶⁴ The *NEGR1* gene is highly expressed in the brain and particularly in the hypothalamus and plays an important role in neuronal outgrowth during neurogenesis.⁶⁵ It has been implicated in regulation of body weight with several SNPs in this gene identified in GWAS of BMI.^{61,66} Furthermore, previous GWAS and individual studies have found associations between SNPs in the same gene in very low to moderate LD ($0.06 < R^2 < 0.59$) with rs1620977 and dietary intake of fruits, meat (beef, pork, processed meat), fat, carbohydrates, and fibre.^{64,67,68} *NEGR1* is an extracellular adhesion protein that binds to cell membrane rafts and promotes cell-to-cell attachment and aggregation, such properties are important in tumour cell migration and invasion during metastasis.⁶⁹ Therefore, *NEGR1* might be involved in malignant transformation through the regulation of intercellular and cell-to-matrix interactions.⁷⁰ *NEGR1* has been shown to be downregulated in various cancer types, including CRC, suggesting a tumour suppression role.⁷¹ Overall, there is some evidence connecting *NEGR1* with diet and CRC since it might also play a role in the choice, preference, and intake of specific type of foods and macronutrients, however, current evidence is limited and further research is needed.⁶⁷

The current study has several strengths. The large sample size gave us increased statistical power to identify G $\times$ E interactions by applying an agnostic genome-wide scan approach. Additionally, we applied a series of new statistical approaches such as the 3-DF joint test, and two-step approach to further increase our power to identify new signals. Finally, the data harmonization and unified quality control across all pooled studies allowed us to evaluate various putative confounders.

A limitation of our study is the fact that information on consumption of fruits, vegetables, and fibre was based on a single measurement through questionnaires which are prone to measurement error and cannot capture long-term consumption efficiently. We also were not able to conduct analyses by fibre type, so it is unknown if the interaction effects we observed differ for soluble and insoluble fibre. Additionally, several of the included studies had a case-control design which are further prone to additional biases like recall and participation bias, and which can lead to biased and inflated results, however the inverse associations were observed for case-control and cohort studies when analyzed separately. Furthermore,

the initial models were adjusted for a small number of confounders which could potentially affect the marginal associations of the three exposures of interest with CRC. Regarding the interaction analyses however, for interaction findings to be confounded the confounder not only has to confound the main effect but also the interactions. Given gene-environment independence, and no interaction of the confounder with the genetic factor it is unlikely that these variables act as confounders.⁷² However, further adjustments for potential confounders yielded similar results with our initial analyses, which gives us confidence for the validity of our results. The two new statistical approaches used in the current study are not without some limitations as well. The two-step approach, even though more statistically efficient, still requires large sample sizes to detect modest-sized interactions. In addition, a significant 3-DF test can occur because of a strong DG effect, $G \times E$ effect, and/or GE correlation and it might not necessarily provide better insights of the mechanisms through which a SNP might affect a trait.^{24,25} Finally, the analysis was conducted in participants of European ancestry and therefore the results may not be generalisable to other populations. However, it is important to follow up in additional population groups where $G \times E$ efforts are limited or underpowered. An important step is the recent colorectal cancer GWAS including individuals of east Asian ancestry.⁸ Additional harmonization of epidemiological data is required however before expanding the $G \times E$ testing.

In summary, we conducted the largest $G \times E$ study to date and we found two $G \times E$ interactions for fibre, fruits, and CRC risk. Our most significant finding was rs4730274 close to the *SLC26A3* gene, which provides supportive evidence for an interaction between fibre consumption, chronic inflammatory intestinal disorders, overall gut function, and CRC development. Additional studies are needed to understand functional implications and to replicate these findings both in European ancestry and other racial/ethnic populations.

Contributors

Conceptualization: Gauderman, Peters, Kim, Kawaguchi, Morrison, Thomas, Lewinger, Data curation: Kim, Kawaguchi, Qu, Diez-Obrero, Morrison, Peters, Harrison, Huyghe, Data verification: Kim, Kawaguchi, Qu, Diez-Obrero, Morrison, Peters, Harrison, Huyghe.

Formal analysis: Kim, Kawaguchi, Diez-Obrero, Morrison.

Methodology: Gauderman, Peters, Kim, Kawaguchi, Morrison, Thomas, Lewinger.

Writing—original draft: Papadimitriou.

Writing—Review and editing: All authors.

Supervision: Peters, Gauderman, Murphy.

All authors read and approved the final version of the manuscript.

Data sharing statement

The dataset used in the current study is available from the corresponding authors on reasonable request.

Declaration of interests

ESK is a co-investigator in a grant from National Institutes of Health (R01CA196569). JW is Stock shareholder of Gilead Sciences Inc. AK has received consulting fees for Illumina Inc., has participated on data safety

monitoring boards or advisory boards of TensorBio, PatchBio, Serimmune, and OpenTargets, and has stock or stock options of Illumina, Freenome, Deep Genomics, Immunai, TensorBio, PatchBio, and Serimmune. MCS was a co-investigator in a grant from National Institutes of Health (R01CA201407). VM has received grant support from Instituto de Salud Carlos III and Fundacion Cientifica Asociación Española Contra el Cáncer. SBG is a co-founder of Brogent international LLC. JPL has received additional grant support (5P01CA196569, 6R01CA201407). The remaining authors declare that they have no conflicts of interest.

Acknowledgements

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.

ESK has received support from the National Institutes of Health (R01CA273198). DAD has received grant support from the National Institutes of Health. SLS has received grant support from the National Institutes of Health. DCT has received grant support from the National Cancer Institute. SIB has received support from the National Cancer Institute and the National Institutes of Health. WJG has received grant support from the National Cancer Institute. DVC has received grant support from the National Cancer Institute (NCI P01CA196569).

Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO): National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (R01 CA059045, U01 CA164930, R01 CA244588, R01 CA201407). This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA015704. Scientific Computing Infrastructure at Fred Hutch funded by ORIP grant S10OD028685. EK is supported by a National Institutes of Health grant (R01CA273198).

ASTERISK: a Hospital Clinical Research Program (PHRC-BRD09/C) from the University Hospital Center of Nantes (CHU de Nantes) and supported by the Regional Council of Pays de la Loire, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (GEFLUC), the Association Anne de Bretagne Génétique and the Ligue Régionale Contre le Cancer (LRCC).

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.

CLUE II funding was from the National Cancer Institute (U01 CA086308, Early Detection Research Network; P30 CA006973), National Institute on Aging (U01 AG018033), and the American Institute for Cancer Research. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government.

The Colon Cancer Family Registry (CCFR, www.coloncfr.org) is supported in part by funding from the National Cancer Institute (NCI), National Institutes of Health (NIH) (award U01 CA167551). Support for case ascertainment was provided in part from the Surveillance, Epidemiology, and End Results (SEER) Program and the following U.S. state cancer registries: AZ, CO, MN, NC, NH; and by the Victoria Cancer Registry (Australia) and Ontario Cancer Registry (Canada). The CCFR Set-1 scan (Illumina 1M/1 M-Duo) scans was supported by NIH awards U01 CA122839 and R01 CA143237 (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 600 K 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 content of this manuscript does not necessarily reflect the views or policies of the NCI, NIH or any of the collaborating centers in the Colon Cancer Family Registry (CCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government, any cancer registry, or the CCFR.

COLO2&3: National Institutes of Health (R01 CA060987).

CPS-II: The American Cancer Society funds the creation, maintenance, and updating of the Cancer Prevention Study-II (CPS-II) cohort. The study protocol was approved by the institutional review boards of Emory University, and those of participating registries as required.

CRCGEN: Colorectal Cancer Genetics & 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), Junta de Castilla y León (grant LE22A10-2), the Spanish Association Against Cancer (AECC) Scientific Foundation grant GCTRA18022MORE and the Consortium for Biomedical Research in Epidemiology and Public Health (CIBER-ESP), action Genrisk. Sample collection of this work was supported by the Xarxa de Bancs de Tumours de Catalunya sponsored by Pla Director d'Oncologia de Catalunya (XBTC), Plataforma Biobancos PT13/0010/0013 and ICOBIOBANC, sponsored by the Catalan Institute of Oncology. We thank CERCA Programme, Generalitat de Catalunya for institutional support.

DACHS: This work was supported by the German Research Council (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, CH 117/1-1, HO 5117/2-1, HE 5998/2-1, KL 2354/3-1, RO 2270/8-1 and BR 1704/17-1), the Interdisciplinary Research Program of the National Center for Tumour Diseases (NCT), Germany, and the German Federal Ministry of Education and Research (01KH0404, 01ER0814, 01ER0815, 01ER1505A and 01ER1505B).

DALS: National Institutes of Health (R01 CA048998 to M. L. Slaterry).

EPIC: The coordination of EPIC is financially supported by International Agency for Research on Cancer (IARC) and also by the Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London which has additional infrastructure support provided by the NIHR Imperial Biomedical Research Centre (BRC). The national cohorts are supported by: Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Éducation Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); German Cancer Aid, German Cancer Research Center (DKFZ), German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Federal Ministry of Education and Research (BMBF) (Germany); Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy, Compagnia di SanPaolo and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKCR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); Health Research Fund (FIS) - Instituto de Salud Carlos III (ISCIII), Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, and the Catalan Institute of Oncology - ICO (Spain); Swedish Cancer Society, Swedish Research Council and Region Skåne and Region Västerbotten (Sweden); Cancer Research UK (14136 to EPIC-Norfolk; C8221/A29017 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk; MR/M012190/1 to EPIC-Oxford). (United Kingdom).

Harvard cohorts: 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 (P01 CA087969, UM1 CA186107, R01 CA137178, 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.

LCCS: The Leeds Colorectal Cancer Study was funded by the Food Standards Agency and Cancer Research UK Programme Award (C588/A19167).

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. BMLynch was supported by MCRF18005 from the Victorian Cancer Agency.

MEC: National Institutes of Health (R37 CA054281, 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 CA081488, R01 CA197350, U19 CA148107, R01 CA242218, and a generous gift from Daniel and Maryann Fong.

NCCCS I & II: We acknowledge funding support for this project from the National Institutes of Health, R01 CA066635 and P30 DK034987.

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 Services (U01 CA074783); and National Cancer Institute of Canada grants (18223 and 18226). The authors wish to acknowledge the contribution of the genotyping team of the McGill University and Genome 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.

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.

SELECT: Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health under Award Numbers U10 CA037429 (CD Blanke), and UM1 CA182883 (CM Tangen/IM Thompson). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

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.

UK Biobank: This research has been conducted using the UK Biobank Resource under Application Number 8614.

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 HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C.

ASTERISK: We are very grateful to those without whom this project would not have existed. We also thank all those who agreed to participate in this study, including the patients and the healthy control persons, as well as all the physicians, technicians and students.

CCFR: The Colon CFR graciously thanks the generous contributions of their 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 authors would like to thank the study participants and staff of the Seattle Colon Cancer Family Registry and the Hormones and Colon Cancer study (CORE Studies).

CLUE II: We thank the participants of Clue II and appreciate the continued efforts of the staff at the Johns Hopkins George W. Comstock Center for Public Health Research and Prevention in the conduct of the Clue II Cohort Study. Cancer data was provided by the Maryland Cancer Registry, Center for Cancer Prevention and Control, Maryland Department of Health, with funding from the State of Maryland and the Maryland Cigarette Restitution Fund. The collection and availability of cancer registry data is also supported by the Cooperative Agreement NU58DP007114, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers for Disease Control and Prevention or the Department of Health and Human Services.

CPS-II: The authors express sincere appreciation to all Cancer Prevention Study-II participants, and to each member of the study and biospecimen management group. The authors would like to acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention's National Program of Cancer Registries and cancer registries supported by the National Cancer Institute's Surveillance Epidemiology and End Results Program. The authors assume full responsibility for all analyses and interpretation of results. The views expressed here are those of the authors and do not necessarily represent the American Cancer Society or the American Cancer Society – Cancer Action Network.

DACHS: We thank all participants and cooperating clinicians, and everyone who provided excellent technical assistance.

EPIC: 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.

EPICOLON: We are sincerely grateful to all patients participating in this study who were recruited as part of the EPICOLON project. We acknowledge the Spanish National DNA Bank, Biobank of Hospital Clínic-IDIBAPS and Biobanco Vasco for the availability of the samples. The work was carried out (in part) at the Esther Koplowitz Centre, Barcelona.

Harvard cohorts: 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 acknowledge Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital as home of the NHS. The authors would like to acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention's National Program of Cancer Registries (NPCR) and/or the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program. Central registries may also be supported by state agencies, universities, and cancer centers. Participating central cancer registries include the following: Alabama, Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Hawaii, Idaho, Indiana, Iowa, Kentucky, Louisiana, Massachusetts, Maine, Maryland, Michigan, Mississippi, Montana, Nebraska, Nevada, New Hampshire, New Jersey, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Puerto Rico, Rhode Island, Seattle SEER Registry, South Carolina, Tennessee, Texas, Utah, Virginia, West Virginia, Wyoming. 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.

LCSS: We acknowledge the contributions of all who conducted this study which was originally reported as 10.1093/carcin/24.2.275.

NCCCS I & II: We would like to thank the study participants, and the NC Colorectal Cancer Study staff.

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. Cancer incidence data have been provided by the District of Columbia Cancer Registry, Georgia Cancer Registry, Hawaii Cancer Registry, Minnesota Cancer Surveillance System, Missouri Cancer Registry, Nevada Central Cancer Registry, Pennsylvania Cancer Registry, Texas Cancer Registry, Virginia Cancer Registry, and Wisconsin Cancer Reporting System. All are supported in part by funds from the Center for Disease Control and Prevention, National Program for Central Registries, local states or by the National Cancer Institute, Surveillance, Epidemiology, and End Results program. The results reported here and the conclusions derived are the sole responsibility of the authors.

SELECT: We thank the research and clinical staff at the sites that participated on SELECT study, without whom the trial would not have been successful. We are also grateful to the 35,533 dedicated men who participated in SELECT.

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: [https://www-whi.org.s3.us-west-2.amazonaws.com/wp-content/uploads/WHI-Investigator-Long-List.pdf](https://www.whi.org.s3.us-west-2.amazonaws.com/wp-content/uploads/WHI-Investigator-Long-List.pdf).

Appendix A. Supplementary data

Supplementary data related to this article can be found at <https://doi.org/10.1016/j.ebiom.2024.105146>.

References

- Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA A Cancer J Clin*. 2021;71(3):209–249.
- World Cancer Research Fund/American Institute for Cancer Research. *Continuous Update Project Expert Report 2018. Diet, nutrition, physical activity and colorectal cancer*. 2018.
- Huyghe JR, Bien SA, Harrison TA, et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet*. 2019;51(1):76–87.
- Law PJ, Timofeeva M, Fernandez-Rozadilla C, et al. Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat Commun*. 2019;10(1):2154.
- Lu Y, Kweon SS, Cai Q, et al. Identification of novel loci and new risk variant in known loci for colorectal cancer risk in east Asians. *Cancer Epidemiol Biomarkers Prev*. 2020;29(2):477–486.
- Huyghe JR, Harrison TA, Bien SA, et al. Genetic architectures of proximal and distal colorectal cancer are partly distinct. *Gut*. 2021;70(7):1325–1334.
- Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N Engl J Med*. 2000;343(2):78–85.
- Fernandez-Rozadilla C, Timofeeva M, Chen Z, et al. Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries. *Nat Genet*. 2023;55(1):89–99.
- Genin E. Missing heritability of complex diseases: case solved? *Hum Genet*. 2020;139(1):103–113.
- Du M, Zhang X, Hoffmeister M, et al. No evidence of gene-calcium interactions from genome-wide analysis of colorectal cancer risk. *Cancer Epidemiol Biomarkers Prev*. 2014;23(12):2971–2976.
- Figueiredo JC, Hsu L, Hutter CM, et al. Genome-wide diet-gene interaction analyses for risk of colorectal cancer. *PLoS Genet*. 2014;10(4):e1004228.
- Hutter CM, Chang-Claude J, Slattery ML, et al. Characterization of gene-environment interactions for colorectal cancer susceptibility loci. *Cancer Res*. 2012;72(8):2036–2044.
- Kantor ED, Hutter CM, Minnier J, et al. Gene-environment interaction involving recently identified colorectal cancer susceptibility Loci. *Cancer Epidemiol Biomarkers Prev*. 2014;23(9):1824–1833.
- Yang T, Li X, Montazeri Z, et al. Gene-environment interactions and colorectal cancer risk: an umbrella review of systematic reviews and meta-analyses of observational studies. *Int J Cancer*. 2019;145(9):2315–2329.
- Gauderman WJ, Mukherjee B, Aschard H, et al. Update on the state of the science for analytical methods for gene-environment interactions. *Am J Epidemiol*. 2017;186(7):762–770.
- Jordahl KM, Shcherbina A, Kim AE, et al. Beyond GWAS of colorectal cancer: evidence of interaction with alcohol consumption and putative causal variant for the 10q24.2 region. *Cancer Epidemiol Biomarkers Prev*. 2022;31(5):1077–1089.
- Carreras-Torres R, Kim AE, Lin Y, et al. Genome-wide interaction study with smoking for colorectal cancer risk identifies novel genetic loci related to tumor suppression, inflammation, and immune response. *Cancer Epidemiol Biomarkers Prev*. 2023;32(3):315–328.
- Peters U, Jiao S, Schumacher FR, et al. Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. *Gastroenterology*. 2013;144(4):799–807 e24.

- 19 Das S, Forer L, Schonherr S, et al. Next-generation genotype imputation service and methods. *Nat Genet.* 2016;48(10):1284–1287.
- 20 McCarthy S, Das S, Kretzschmar W, et al. A reference panel of 64, 976 haplotypes for genotype imputation. *Nat Genet.* 2016;48(10): 1279–1283.
- 21 Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Stat Med.* 2001;20(24):3875–3889.
- 22 Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ.* 2003;327(7414):557–560.
- 23 Morrison J, Kim AE, J G. *GxEScanR: an R package to detect GxE interactions in a genomewide association study.* Los Angeles: Version 10 University of Southern California; 2018.
- 24 Gauderman WJ, Kim A, Conti DV, et al. A unified model for the analysis of gene-environment interaction. *Am J Epidemiol.* 2019;188(4):760–767.
- 25 Gauderman WJ, Zhang P, Morrison JL, Lewinger JP. Finding novel genes by testing G x E interactions in a genome-wide association study. *Genet Epidemiol.* 2013;37(6):603–613.
- 26 Ionita-Laza I, McQueen MB, Laird NM, Lange C. Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *Am J Hum Genet.* 2007;81(3):607–614.
- 27 Kawaguchi ES, Kim AE, Lewinger JP, Gauderman WJ. Improved two-step testing of genome-wide gene-environment interactions. *Genet Epidemiol.* 2023;47(2):152–166.
- 28 Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet Epidemiol.* 2008;32(4):361–369.
- 29 de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet.* 2008;17(R2):R122–R128.
- 30 Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999;55(4):997–1004.
- 31 Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics.* 2010;26(18):2336–2337.
- 32 Consortium GT. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science.* 2020;369(6509):1318–1330.
- 33 Diez-Obrero V, Dampier CH, Moratalla-Navarro F, et al. Genetic effects on transcriptome profiles in colon epithelium provide functional insights for genetic risk loci. *Cell Mol Gastroenterol Hepatol.* 2021;12(1):181–197.
- 34 Cohen AJ, Saiakhova A, Corradin O, et al. Hotspots of aberrant enhancer activity punctuate the colorectal cancer epigenome. *Nat Commun.* 2017;8:14400.
- 35 Lee J, Ottojolkanki, Kim D, Strattan JS, Kundaje A, Nordström K, et al. *ENCODE-DCC/atac-seq-pipeline: v1. 9.1.* v1.9.1 ed. Zenodo; 2020.
- 36 Lee J, Strattan JS, Annashcherbina KM, Maurizio PL. *ENCODE-DCC/chip-seq-pipeline 2: v1.6.1.* v1.6.1 ed. Zenodo; 2020.
- 37 Qunhua L, James BB, Haiyan H, Peter JB. Measuring reproducibility of high-throughput experiments. *Ann Appl Stat.* 2011;5(3):1752–1779.
- 38 Lopez-Delisle L, Rabbani L, Wolff J, et al. pyGenomeTracks: reproducible plots for multivariate genomic datasets. *Bioinformatics.* 2021;37(3):422–423.
- 39 Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. *Curr Protoc Bioinformatics.* 2014;47(11):1–34.
- 40 Su YR, Di CZ, Hsu L, Genetics, Epidemiology of Colorectal Cancer C. A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics.* 2017;18(1):119–131.
- 41 de Lange KM, Moutsianas L, Lee JC, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet.* 2017;49(2):256–261.
- 42 Pituch-Zdanowska A, Banaszkiwicz A, Albrecht P. The role of dietary fibre in inflammatory bowel disease. *Prz Gastroenterol.* 2015;10(3):135–141.
- 43 Fritsch J, Garces L, Quintero MA, et al. Low-fat, high-fiber diet reduces markers of inflammation and dysbiosis and improves quality of life in patients with ulcerative colitis. *Clin Gastroenterol Hepatol.* 2021;19(6):1189–1199 e30.
- 44 Porter RJ, Arends MJ, Churchhouse AMD, Din S. Inflammatory bowel disease-associated colorectal cancer: translational risks from mechanisms to medicines. *J Crohns Colitis.* 2021;15(12):2131–2141.
- 45 Zhang M, Li T, Zhu J, Tuo B, Liu X. Physiological and pathophysiological role of ion channels and transporters in the colorectum and colorectal cancer. *J Cell Mol Med.* 2020;24(17):9486–9494.
- 46 Chapman JM, Knoepp SM, Byeon MK, Henderson KW, Schweinfest CW. The colon anion transporter, down-regulated in adenoma, induces growth suppression that is abrogated by E1A. *Cancer Res.* 2002;62(17):5083–5088.
- 47 Hemminki A, Hoglund P, Pukkala E, et al. Intestinal cancer in patients with a germline mutation in the down-regulated in adenoma (DRA) gene. *Oncogene.* 1998;16(5):681–684.
- 48 Schweinfest CW, Henderson KW, Suster S, Kondoh N, Papas TS. Identification of a colon mucosa gene that is down-regulated in colon adenomas and adenocarcinomas. *Proc Natl Acad Sci U S A.* 1993;90(9):4166–4170.
- 49 Schweinfest CW, Spyropoulos DD, Henderson KW, et al. slc26a3 (dra)-deficient mice display chloride-losing diarrhea, enhanced colonic proliferation, and distinct up-regulation of ion transporters in the colon. *J Biol Chem.* 2006;281(49):37962–37971.
- 50 Bhutia YD, Babu E, Ramachandran S, Yang S, Thangaraju M, Ganapathy V. SLC transporters as a novel class of tumour suppressors: identity, function and molecular mechanisms. *Biochem J.* 2016;473(9):1113–1124.
- 51 Asano K, Matsushita T, Umeno J, et al. A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. *Nat Genet.* 2009;41(12):1325–1329.
- 52 Liu JZ, van Sommeren S, Huang H, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet.* 2015;47(9):979–986.
- 53 Hanahan D. Hallmarks of cancer: new dimensions. *Cancer Discov.* 2022;12(1):31–46.
- 54 Schmitt M, Greten FR. The inflammatory pathogenesis of colorectal cancer. *Nat Rev Immunol.* 2021;21(10):653–667.
- 55 Thomson C, Garcia AL, Edwards CA. Interactions between dietary fibre and the gut microbiota. *Proc Nutr Soc.* 2021:1–11.
- 56 Cresci GA, Thangaraju M, Mellinger JD, Liu K, Ganapathy V. Colonic gene expression in conventional and germ-free mice with a focus on the butyrate receptor GPR109A and the butyrate transporter SLC5A8. *J Gastrointest Surg.* 2010;14(3):449–461.
- 57 Alvandi E, Wong WKM, Joglekar MV, Spring KJ, Hardikar AA. Short-chain fatty acid concentrations in the incidence and risk-stratification of colorectal cancer: a systematic review and meta-analysis. *BMC Med.* 2022;20(1):323.
- 58 Mirzaei R, Dehkhodaie E, Bouzari B, et al. Dual role of microbiota-derived short-chain fatty acids on host and pathogen. *Biomed Pharmacother.* 2022;145:112352.
- 59 Quinonez SC, Thoene JG. Dihydrolipoamide dehydrogenase deficiency. In: Adam MP, Everman DB, Mirzaa GM, et al., eds. *GeneReviews((R))*. Seattle (WA) 1993.
- 60 Olson KA, Schell JC, Rutter J. Pyruvate and metabolic flexibility: illuminating a path toward selective cancer therapies. *Trends Biochem Sci.* 2016;41(3):219–230.
- 61 Pulit SL, Stoneman C, Morris AP, et al. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum Mol Genet.* 2019;28(1):166–174.
- 62 Savage JE, Jansen PR, Stringer S, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet.* 2018;50(7):912–919.
- 63 Lee JJ, Wedow R, Okbay A, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet.* 2018;50(8):1112–1121.
- 64 Niarchou M, Byrne EM, Trzaskowski M, et al. Genome-wide association study of dietary intake in the UK biobank study and its associations with schizophrenia and other traits. *Transl Psychiatry.* 2020;10(1):51.
- 65 Schmid PM, Heid I, Buechler C, et al. Expression of fourteen novel obesity-related genes in Zucker diabetic fatty rats. *Cardiovasc Diabetol.* 2012;11:48.
- 66 Boender AJ, van Gestel MA, Garner KM, Luijckendijk MC, Adan RA. The obesity-associated gene Negr1 regulates aspects of energy balance in rat hypothalamic areas. *Physiol Rep.* 2014;2(7).
- 67 Bauer F, Elbers CC, Adan RA, et al. Obesity genes identified in genome-wide association studies are associated with adiposity

- measures and potentially with nutrient-specific food preference. *Am J Clin Nutr.* 2009;90(4):951–959.
- 68 Rukh G, Sonestedt E, Melander O, et al. Genetic susceptibility to obesity and diet intakes: association and interaction analyses in the Malmo Diet and Cancer Study. *Genes Nutr.* 2013;8(6):535–547.
- 69 Hennig EE, Kluska A, Piatkowska M, et al. GWAS links new variant in long non-coding RNA LINC02006 with colorectal cancer susceptibility. *Biology.* 2021;10(6).
- 70 Okegawa T, Pong RC, Li Y, Hsieh JT. The role of cell adhesion molecule in cancer progression and its application in cancer therapy. *Acta Biochim Pol.* 2004;51(2):445–457.
- 71 Kim H, Hwang JS, Lee B, Hong J, Lee S. Newly identified cancer-associated role of human neuronal growth regulator 1 (NEGR1). *J Cancer.* 2014;5(7):598–608.
- 72 Vanderweele TJ, Ko YA, Mukherjee B. Environmental confounding in gene-environment interaction studies. *Am J Epidemiol.* 2013;178(1):144–152.