

Supplementary Note

Determining Risk of Colorectal Cancer and Starting Age of Screening Based on Lifestyle, Environmental, and Genetic Factors

Description of Study Populations

The study design and characteristics for studies that was included in the development of our risk prediction models are described in the following. Table S1 summarizes the sample sizes and demographic factors in these studies.

Hawaii Colorectal Cancer Studies 2 and 3 (Colo2&3)¹: Patients with colorectal cancer were identified through the rapid reporting system of the Hawaii SEER registry and consisted of all Japanese, Caucasian, and Native Hawaiian residents of Oahu who were newly diagnosed with an adenocarcinoma of the colon or rectum between January 1994 and August 1998. Control subjects were selected from participants in an on-going population-based health survey conducted by the Hawaii State Department of Health and from Health Care Financing Administration participants. Controls were matched to cases by sex, ethnicity, and age (within two years). Personal interviews were obtained from 768 matched pairs, resulting in a participation rate of 58.2% for cases and 53.2% for controls. A questionnaire, administered during an in-person interview, included questions about demographics, lifetime history of tobacco, alcohol use, aspirin use, physical activity, personal medical history, family history of colorectal cancer, height and weight, diet (FFQ), and postmenopausal hormone use. A blood sample was obtained from 548 (71%) of interviewed cases and 662 (86%) of interviewed controls. SEER staging information was extracted from the Hawaii Tumor Registry. In GECCO, self-reported Caucasian subjects with DNA, and clinical and epidemiologic data were selected for genotyping.

Cancer Prevention Study II (CPS2): Men and women in the CPS-II Nutrition Cohort (N = 184,194) were recruited from among the 1.2 million U.S. adults enrolled in the CPS-II Baseline Cohort, a study of cancer mortality that was initiated in 1982.² In 1992 and 1993, a detailed questionnaire was mailed to a subgroup of the Baseline Cohort. Respondents to this questionnaire were enrolled into the CPS-II Nutrition Cohort.² Participants in CPS-II Nutrition are followed for cancer incidence and mortality; they have received additional mailed questionnaires in 1997 and every 2 years thereafter to update exposure information

and to obtain self-reported cancer diagnoses. We ask each participant who self-reports a cancer diagnosis to grant us permission to obtain her/his medical records to verify the diagnosis. Fatal cases are also identified through linkage with the National Death Index. When medical records cannot be obtained, often because the participants died before being able to self-report their cancer or provide consent for access to medical records, cancer diagnoses are verified through computerized linkage with state cancer registries. Blood samples were collected from 39,380 members of the CPS-II Nutrition Cohort from 1998 to 2001, and buccal cell samples were collected from an additional 67,000 cohort members in 2001 and 2002. All aspects of the CPS-II study are approved by the Emory University Institutional Review Board (IRB).

Darmkrebs: Chancen der Verhütung durch Screening (DACHS)^{3,4}: This German study was initiated as a large population-based case-control study in 2003 in the Rhine-Neckar-Odenwald region (southwest region of Germany) to assess the potential of endoscopic screening for reduction of colorectal cancer risk and to investigate etiologic determinants of disease, particularly lifestyle/environmental factors and genetic factors. Cases with a first diagnosis of invasive colorectal cancer (ICD-10 codes C18-C20) who were at least 30 years of age (no upper age limit), German speaking, a resident in the study region, and mentally and physically able to participate in a one-hour interview, were recruited by their treating physicians either in the hospital a few days after surgery, or by mail after discharge from the hospital. Cases were confirmed based on histologic reports and hospital discharge letters following diagnosis of colorectal cancer. All hospitals treating colorectal cancer patients in the study region participated. Based on estimates from population-based cancer registries, more than 50% of all potentially eligible patients with incident colorectal cancer in the study region were included. Community-based controls were randomly selected from population registries, employing frequency matching with respect to age (5-year groups), sex, and county of residence. Controls with a history of colorectal cancer were excluded. Controls were contacted by mail and follow-up calls. The participation rate was 51%. During an in-person interview, data were collected on demographics, medical history, family history of CRC, and various life-style factors, as were blood and mouthwash samples. The Set 1 scan consisted of a subset of participants recruited up to 2007, and samples were frequency matched on age and sex. The Set 2 scan consisted of additional subjects that were recruited up to 2010 as part of this ongoing study.

Diet, Activity and Lifestyle Study (DALIS)⁵: DALIS is a population-based case-control study of colon cancer. Participants were recruited between 1991 and 1994 from three locations: the Kaiser Permanente Medical Care Program (KPMCP) of Northern California, an eight-county area in Utah, and the metropolitan Twin Cities area of Minnesota. Eligibility criteria for cases included age at diagnosis between 30 and 79 years, diagnosis with first primary colon cancer (ICD-O-2 codes 18.0 and 18.2-18.9) between October 1st 1991 and September 30th 1994, English speaking, and competency to complete the interview. Individuals with cancer of the rectosigmoid junction or rectum were excluded, as were those with a pathology report noting familial adenomatous polyposis, Crohns disease, or ulcerative colitis. A rapid-reporting system was used to identify all incident cases of colon cancer resulting in the majority of cases being interviewed within four months of diagnosis. Controls from KPMCP were randomly selected from membership lists. In Utah, controls under 65 years of age were randomly selected through random-digit dialing and driver license lists. Controls, 65 years of age and older, were randomly selected from Health Care Financing Administration lists. In Minnesota, controls were identified from Minnesota drivers license or state ID lists. Cases and controls were matched by 5-year age groups and sex. The Set 1 scan consisted of a subset of the study designed above, from Utah, Minnesota, and KPMCP, and was restricted to subjects who self-reported as White non-Hispanic. The Set 2 scan consisted of subjects from Utah and Minnesota that were not genotyped in Set 1. Set 2 was restricted to subjects who self-reported as White non-Hispanic and those that had appropriate consent to post data to dbGaP.

Health Professionals Follow-up Study (HPFS)⁶: The HPFS is a parallel prospective study to the Nurses Health Study (NHS). The HPFS cohort comprises 51,529 men who, in 1986, responded to a mailed questionnaire. The participants are U.S. male dentists, optometrists, osteopaths, podiatrists, pharmacists, and veterinarians born between 1910 and 1946. Participants have provided information on health related exposures, including: current and past smoking history, age, weight, height, diet, physical activity, aspirin use, and family history of colorectal cancer. Follow-up has been excellent, with 94% of the men responding to date. Colorectal cancer and other outcomes were reported by participants or next-of-kin and followed up through review of the medical and pathology record by physicians. Overall, more than 97% of self-reported colorectal cancers were confirmed by medical record review. Colorectal cancer cases were ascertained through January 1, 2008. In 1993-95, 18,825 men in HPFS mailed in blood samples by overnight courier which were aliquoted into buoy coat

and stored in liquid nitrogen. In 2001-04, 13,956 men in HPFS who had not previously provided a blood sample mailed in a “swish-and-spit” sample of buccal cells. Incident cases are defined as those occurring after the subject provided a blood or buccal sample. Prevalent cases are defined as those occurring after enrollment in the study in 1986, but prior to the subject providing either a blood or buccal sample. After excluding participants with histories of cancer (except non-melanoma skin), ulcerative colitis, or familial polyposis, two case-control sets were constructed from which DNA was isolated from either buccal cells or buccal cells for genotyping: 1) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a blood sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the cases; 2) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a buccal sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the case. For both case-control sets, matching criteria included year of birth (within 1 year) and month/year of blood or buccal cell sampling (within six months). Cases were pair matched 1:1, 1:2, or 1:3 with a control participant(s).

Kentucky Case-Control Study (Kentucky): The Kentucky Case-Control study was initiated in July 2003 through the University of Kentucky Cancer Center. A web-based reporting system implemented by the Kentucky Cancer Registry in 2003 has facilitated rapid report of cases state-wide, with approximately 76.8% of all cases reported to the registry within 6 months of diagnosis. Cases (>21 years) diagnosed with histologically confirmed colon cancer and entered into the registry within 6 months of their diagnoses are invited to join the study. Population-based unrelated controls are recruited through random digit dialing and are frequency matched to the cases by age (± 5 years), gender, and race. Excluded from the study are those individuals who have been diagnosed with colon cancer because of known hereditary forms of colon cancer or polyposis such as familial adenomatous polyposis (FAP), hereditary non-polyposis colorectal cancer (HNPCC), Peutz-Jeghers, and Cowden disease. Currently there are more than 1,040 incident population-based cases of colorectal cancer and 1,750 population-based controls fully recruited, with comprehensive epidemiologic data, pathology data, and DNA from cases and controls

Melbourne Collaborative Cohort Study (MCCS): A case-control study nested in the Melbourne Collaborative Cohort Study, a prospective cohort of 41,514 volunteers of which 24,469 women was designed including 576 incident cases diagnosed during follow-up from

baseline (1990-1994) till mid-2010 and 576 individually matched population-based controls.⁷ The matching factors were sex, country of birth (Australia/UK, Italy and Greece), and year of baseline attendance. Cases were all incident cases in the cohort ascertained through linkage to the Victorian Cancer Registry and other State cancer registries in Australia. For the GWAS, cases with only DNA extracted from Guthrie cards available were excluded.

Multi-Ethnic Cohort Study (MEC)⁸: MEC was initiated in 1993 to investigate the impact of dietary and environmental factors on major chronic diseases, particularly cancer, in ethnically diverse populations in Hawai'i and California. The study recruited 96,810 men and 118,441 women aged 45 to 75 years between 1993 and 1996. Incident colorectal cancer cases occurring since January 1995, and controls were contacted for blood or saliva samples. The median interval between diagnosis and blood draw was 14 months (interquartile range, 10-19) among cases and the participation rate 74%. A sample of cohort participants was randomly selected to serve as controls at the onset of the nested case-control study (participation rate 66%). The selection was stratified by sex, age, and race/ethnicity. Colorectal cancer cases are identified through the Rapid Reporting System of the Hawai'i Tumor Registry and through quarterly linkage to the Los Angeles County Cancer Surveillance Program. Both registries are members of SEER. In GECCO, self-reported White subjects from the nested case-control study described above with DNA, and clinical and epidemiologic data were selected for genotyping.

Molecular Epidemiology of Colorectal Cancer (MECC)⁹: The Molecular Epidemiology of Colorectal Cancer Study (MECC) is a population-based case-control study of colorectal cancer (CRC). Incident, pathologically-confirmed CRC cases and controls were recruited from a specific region of northern Israel. Newly-diagnosed CRC cases beginning March 31, 1998, who agreed to participate, were interviewed, gave a venous blood sample, and provided permission for tumor tissue retrieval. Written, informed consent was obtained according to Institutional Review Board-approved protocols at Carmel Medical Center in Haifa and the University of Southern California (HS-12-00324, HS-12-00672, and HS-08-00378). Germline DNA was extracted from whole blood for genotyping. The analytic dataset from the MECC study genotyped on the OncoArray and included in the CORECT Phase 2 European GWAS consisted of 3,591 cases of pathologically-confirmed adenocarcinoma and 2,848 controls. In addition, previously genotyped cases and controls were included in the Phase 1 GWAS: these consisted of 484 cases and 498 controls genotyped on the Illumina

Omni 2.5 array, and 1,120 cases and 820 controls were genotyped on the Affymetrix Axiom CORECT Set array. Thus, the total number of cases and controls from the MECC study included in Phases 1 and 2 (after quality control for genotyping) was 5,195 cases and 4,166 controls.

Nurses' Health Study (NHS)¹⁰: The NHS cohort began in 1976 when 121,700 married female registered nurses aged 30 to 55 years returned the initial questionnaire that ascertained a variety of important health-related exposures. Since 1976, follow-up questionnaires have been mailed every two years. Colorectal cancer and other outcomes were reported by participants or next-of-kin and followed up through review of the medical and pathology record by physicians. Overall, more than 97% of self-reported colorectal cancers were confirmed by medical-record review. Information was abstracted on histology and primary location. Follow-up has been high: as a proportion of the total possible follow-up time, follow-up has been over 92%. Colorectal cancer cases were ascertained through June 1, 2008. In 1989-90, 32,826 women in NHS I, mailed in blood samples by overnight courier which were aliquoted into buccal coat and stored in liquid nitrogen. In 2001-04, 29,684 women in NHS I who did not previously provide a blood sample mailed in a "swish-and-spit" sample of buccal cells. Incident cases are defined as those occurring after the subject provided a blood or buccal sample. Prevalent cases are defined as those occurring after enrollment in the study in 1976, but prior to the subject providing either a blood or buccal sample. After excluding participants with histories of cancer (except non-melanoma skin), ulcerative colitis, or familial polyposis, we constructed two case-control sets from which DNA was isolated from either buccal coat or buccal cells for genotyping: 1) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a blood sample and were free of colorectal cancer at the same time that the colorectal cancer was diagnosed in the case; 2) a case-control set with cases of colorectal cancer matched to randomly selected controls who provided a buccal sample and were free of colorectal cancer at the same time that the colorectal cancer was diagnosed in the cases. For both case-control sets, matching criteria included year of birth (within one year) and month / year of blood or buccal cell sampling (within six months). Cases were pair matched 1:1, 1:2, or 1:3 with a control participant(s).

Newfoundland Case-Control Study (NFCCR): The NFCCR is a case-control study that includes pathology confirmed CRC cases less than 75 years of age diagnosed between

January 1999 and December 2003, as identified from the Newfoundland Cancer Registry. The Newfoundland Cancer Registry registers all cases of invasive cancer diagnosed among residents of the province of Newfoundland and Labrador. Consenting patients received a family history questionnaire and were asked to provide a blood sample and to permit access to tumor tissue and medical records. If a patient was deceased, we sought the participation of a close relative for the purposes of obtaining the family history and for permission to access tissue blocks and medical records. Use of proxies in this way removes the bias of excluding advanced stage patients who die before they can give consent. Population-based controls were identified by random digit dialing from the residents of the province, and matched to the cases on sex and five-year age groups. Controls provided a blood sample and filled out a risk factor questionnaire.

Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO):

PLCO enrolled 154,934 participants (men and women, aged between 55 and 74 years) at ten centers from 1993 to 2001 into a large, randomized, two-arm trial to determine the effectiveness of screening to reduce cancer mortality. Sequential blood samples were collected from participants assigned to the screening arm. Participation was 93% at the baseline blood draw. In the observational (control) arm, buccal cells were collected via mail using the “swish-and-spit” protocol and participation rate was 65%. Details of this study have been previously described^{10, 11} and are available online (<http://dcp.cancer.gov/plco>).

The Set 1 scan included a subset of 577 colon cancer cases self-reported as being non-Hispanic White with available DNA samples, questionnaire data, and appropriate consent for ancillary epidemiologic studies. Cases were excluded if they had a history of inflammatory bowel disease, polyps, polyposis syndrome or cancer (excluding basal or squamous cell skin cancer). Controls come from the Cancer Genetic Markers of Susceptibility (CGEMS) prostate cancer scan^{11,12} (all male) and the GWAS of Lung Cancer and Smoking¹³ (enriched for smokers) along with an additional 92 non-Hispanic White female controls. For the Set 2 scan, cases were colorectal cancers from both arms of the trial, which were not already included in Set 1. Samples were excluded if participants did not sign appropriate consents, if DNA was unavailable, if baseline questionnaire data with follow-up were unavailable, if they had a history of colon cancer prior to the trial, if they had a rare cancer, and if they were already in colon GWAS, or if they were a control in the prostate or lung populations. Controls were frequency matched 1:1 to cases without

replacement, and cases were not eligible to be controls. Matching criteria were age at enrollment (two year blocks), enrollment date (two year blocks), sex, race / ethnicity, trial arm, and study year of diagnosis (i.e. controls must be cancer free into the case's year of diagnosis).

VITamins And Lifestyle (VITAL): The VITamins And Lifestyle (VITAL) cohort comprises of 77,721 Washington State men and women aged 50 to 76 years, recruited from 2000 to 2002 to investigate the association of supplement use and lifestyle factors with cancer risk. Subjects were recruited by mail, from October 2000 to December 2002, using names purchased from a commercial mailing list. All subjects completed a 24 page questionnaire and buccal-cell specimens for DNA was self-collected by 70% of the participants. Subjects are followed for cancer by linkage to the western Washington SEER cancer registry and are censored when they move out of the area covered by the registry or at time of death. Details of this study have been previously described.¹⁴ In Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), a nested case-control set was genotyped. Samples included colorectal cancer cases with DNA, excluding subjects with colorectal cancer before baseline, in situ cases, (large cell) neuroendocrine carcinoma, squamous cell carcinoma, carcinoid tumor, Goblet cell carcinoid, any type of lymphoma, including non-Hodgkin, Mantle cell, large B-cell, or follicular lymphoma. One control was randomly selected per case among all controls whose follow-up time were greater than the follow-up time of the case until diagnosis and who were matched on age at enrollment (within one year), enrollment date (within one year), sex, and race / ethnicity.

Women's Health Initiative (WHI): WHI is a long-term health study of 161,808 post-menopausal women aged 50 to 79 years recruited from 1993 to 1998 at 40 clinical centers throughout the U.S. WHI comprises a Clinical Trial (CT) arm, an Observational Study (OS) arm, and several extension studies. The details of WHI have been previously described^{15,16} and are available online (<https://cleo.whi.org/SitePages/Home.aspx>). In GECCO, Set 1 cases were selected from the September 12, 2005 database and were comprised of centrally adjudicated colon cancer cases from the Observational Study (OS) who self-reported as White. Controls were first selected among controls previously genotyped as part of a Hip Fracture GWAS conducted within the WHI OS and matched to cases on age (within three years) enrollment date (within 365 days), hysterectomy status, and prevalent conditions at baseline. For 37 cases, there was not a control match in the Hip Fracture GWAS. For these participants, we identified a matched control in the WHI OS based on

same criteria. In the Set 2 scan, cases were selected from the August 2009 database and were comprised of centrally adjudicated colon and colorectal cancer cases from the OS and CT who were not genotyped in Set 1. In addition, case and control participants were subject to the following exclusion criteria: a prior history of colorectal cancer at baseline, IRB approval not available for data submission into dbGaP, and not sufficient DNA available. Matching criteria included age (within years), race/ethnicity, WHI date (within three years), WHI Calcium and Vitamin D study date (within three years), and randomization arms (OS flag, hormone therapy assignments, dietary modification assignments, calcium/vitamin D assignments). In addition, they were matched on the four regions of randomization centers. Each case was matched with one control (1:1) that exactly met the matching criteria. Control selection was done in a time-forward manner, selecting one control for each case first from the risk set at the time of the case's diagnosis. The matching algorithm was allowed to select the closest match based on a criterion to minimize an overall distance measure.¹⁷ Each matching factor was given the same weight. Additional available controls that were genotyped as part of the Hip Fracture GWAS were included to improve power.

Acknowledgements

CPS2: The American Cancer Society funds the creation, maintenance, and updating of the Cancer Prevention Study II (CPS-II) cohort. 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's National Program of Cancer Registries, and cancer registries supported by the National Cancer Institute's Surveillance Epidemiology and End Results program.

DACHS: We thank all participants and cooperating clinicians, and Ute Handte-Daub, Renate Hettler-Jensen, Utz Benschaid, Muhabbet Celik and Ursula Eilber for excellent technical assistance.

NHS and HPFS: We would like to acknowledge Patrice Soule and Hardeep Ranu of the Dana Farber Harvard Cancer Center High-Throughput Polymorphism Core who assisted in the genotyping for NHS and HPFS under the supervision of Dr. Immaculata Devivo and Dr. David Hunter, Qin (Carolyn) Guo and Lixue Zhu who assisted in programming for NHS and HPFS. We would like to thank the participants and staff of the Nurses' Health Study and the Health Professionals Follow-Up Study, 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. In addition, this study was approved by the Connecticut Department of Public Health (DPH) Human Investigations Committee. Certain data used in this publication were obtained from the DPH. The authors assume full responsibility for analyses and interpretation of these data.

Kentucky: The authors would like to acknowledge the staff and at Kentucky Cancer Registry.

MCCS: This study was made possible by the contribution of many people, including the original investigators and the diligent team who recruited participants and continue to work on follow-up. We would also like to express our gratitude to the many thousands of Melbourne residents who took part in the study and provided blood samples.

NFCCR: We would like to express our deepest thanks to the families and patients who have so generously provided their time and biological samples to the Newfoundland Colorectal Cancer Registry (NFCCR). Supported by the Canadian Institutes of Health Research grant CRT-43821; and the genotyping of the samples was conducted by Dr. Stephen B Gruber, USC Norris Comprehensive Cancer Center, University of Southern California as part of the ColoRectal Transdisciplinary Study (CORECT).

PLCO: The authors thank Drs. Christine Berg and Philip Prorok, Division of Cancer Prevention, National Cancer Institute, the Screening Center investigators and staff of the Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial, Mr. Tom Riley and staff, Information Management Services, Inc., Ms. Barbara O'Brien and staff, Westat, Inc., and Drs. Bill Kopp, Wen Shao, and staff, SAIC-Frederick. Most importantly, we acknowledge the study participants for their contributions to making this study possible. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

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://cleo.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf>

GECCO: The authors would like to thank all those at the GECCO Coordinating Center for helping bring together the data and people that made this project possible.

Harmonization of lifestyle and environmental data

For the development of our risk prediction models, we have used many studies in two large multidisciplinary national consortia; Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), and Colorectal Transdisciplinary (CORECT) Study in the Genetic Associations and Mechanisms in Oncology (GAME-ON). To ensure all the variables are comparable across studies, all data were harmonized at the coordinating center at the Fred Hutch using a standardized protocol (Figure S2).

Information on basic demographics lifestyle and environmental risk factors was collected by self-report using in-person interviews and/or structured questionnaires. Individual level data of all studies were centrally harmonized at the data coordinating center. We carried out a multi-step data harmonization procedure, reconciling each study's unique protocols and data-collection instruments (Figure S2). First, we defined common data elements (CDEs). We examined the questionnaires and data dictionaries for each study to identify study specific data elements that could be mapped to the CDEs. Through an iterative process, we communicated with each data contributor to obtain relevant data and coding information. The data elements were written to a common data platform, transformed, and combined into a single dataset with common definitions, standardized permissible values, and standardized coding. The mapping and resulting data were reviewed for quality assurance, and range and logic checks were performed to assess data and data distributions within and between studies. Outlying samples were truncated to the minimum or maximum value of established range for each variable. All variables were collected at the study reference time, which was defined as study entry or blood collection for cohort studies and one to two years before sample ascertainment for case-control studies to ensure exposures assessed before cancer diagnoses. Age at referent time was defined in years and modeled continuously.

Height^{18,19} and *Body mass index*^{18,20-22}

Height was defined in centimeters based on self-reports or direct measures. Body mass index (BMI) was calculated from self-reports or direct measures of body weight (kg) divided by height (m²). In our analysis, the BMI/5 was used, and if the BMI<18.5, it was set as missing.

*Family history*²³

Family history was a yes/no variable for presence or absence of a first-degree relative with colorectal cancer.

*Endoscopy history*²³

Endoscopy history was coded as yes, no, or missing, depending on whether a participant had sigmoidoscopy or colonoscopy screening before the study reference time, or such information was missing.

Education^{24,25}

Highest level of education was defined in four categories: less than high school degree, high school degree or completed GED, some college or technical school, and college or graduate degree.

*Diabetes*²⁶ and *Physical activity*²⁷

Self-reported type 2 diabetes was categorized as yes/no. Physical activity was calculated by summing hours per week of leisure-time or undifferentiated activities and categorized as active (≥ 1 hour/week) or inactive (< 1 hour/week).

Smoking Status^{18,21,28-31}

Smoking status was defined as never- and ever-smoking; it was defined as “yes” for current or former smokers and “no” for never smokers. Pack-years of smoking were calculated by multiplying the average number of packs of cigarettes smoked per day by smoking duration (years). Smoking pack-years among ever smokers was harmonized across studies by sex- and study-specific quartiles with quartile cutoffs determined within the controls of each study and sex. We assigned values 1 for lesser than equal to the first quartile, 2 for between the first and second quartiles, 3 for the second and the third quartiles, 4 for greater than the third quarter. For never smokers, it was assigned as “0”. This variable was treated as continuous variable in the analysis.

Alcohol consumption^{18,21,31,32}

We converted consumption of alcoholic beverages into grams of alcohol per day (g/day) by summing the alcohol content of each beverage consumed per day. We grouped study participants as non-/occasional drinkers (drinking < 1 g/day); light-to-moderate drinkers (drinking 1-28 g/day); and heavy drinkers (drinking > 28 g/day, one standard drinking is

approximately equal to 14 grams of alcohol).

Aspirin and NSAIDs use^{18,21,33,34}

We used dichotomous variable for regular use of aspirin and/or NSAIDs (yes or no). Aspirin use is defined as “yes” if a person used aspirin regularly in reference time period and “no” otherwise. NSAIDs use is defined as “yes” if a person used non-aspirin NSAIDs regularly in reference time period and “no” otherwise.

Post-menopausal hormone use^{21,35,36}

Menopausal hormone therapy (MHT) use was considered either as any MHT use, Estrogen-only use or Estrogen+Progesterone use at reference time.

Dietary variables^{18,21,37-46}

Intake of dietary factors was assessed using food frequency questionnaires (FFQs) or diet history (DALIS only). The dietary variables fruits, vegetables, and red or processed meats were measured in servings/day; fiber as g/day; calcium as mg/day; and folate as µg/day. The dietary variables were coded as sex- and study-specific quartiles with quartile cutoffs determined within the controls of each study and sex. Some studies with less variation in dietary intake (primarily because of fewer questions) had less than 4 intake categories for certain factors. In these instances, we assigned intake to only the 2nd and 3rd quartiles. Calcium intake, measured in mg/day was determined from calcium in foods (i.e., dietary) or supplements (single + multivitamins + antacids) when available. Total calcium intake was calculated as dietary + supplemental calcium. For studies that entered supplement data as regular user vs. nonuser, we assumed regular use was 500 mg/day, 500 mg/single tablet, or 130 mg/multivitamin tablet [the generic dose in supplements³⁹]. Folate and folic acid intake in each study was 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. To account for the higher bioavailability of synthetic folic acid vs. food folate, we calculated total folate intake as dietary folate equivalents (DFE): total mcg DFE = mcg of dietary folate + 1.7 x mcg folic acid from supplements.³⁸ Because the times 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 x mcg folic acid from fortified food. If studies entered supplement data as regular user vs. nonuser, we assumed regular

use was 400 mcg/day or 400 mcg/tablet (for multivitamins), which corresponds to the generic dose in supplements.^{37,43} Total energy consumption was calculated in kcal/day and modeled as a continuous variable scaled by its standard error.

Harmonization for E-score

Even though all individual risk factors were harmonized across studies through multiple steps, the distribution of risk factors could vary between studies. We therefore recoded the E-score as a sex- and study-specific percentile. As the population attributable risk (PAR) is a central measure in calculating the absolute risk of CRC, we evaluate the PAR estimates by studies for the E-score in percentile. The PAR estimates were generally consistent across studies (Figure S1), suggesting that our approach is robust.

Table S1. Descriptive characteristics of study populations in the model building dataset

Study Name	Short Name	Design	Country	Cases	Controls	Mean Age (range)	Female (%)
Hawaii Colorectal Cancer Studies 2 and 3	Colo2&3	Case/Control	U.S.	48	70	64 (38-84)	49
Cancer Prevention Study II	CPS2	Cohort Study	U.S.	279	275	75 (58-94)	49
Darmkrebs: Chancen der Verhütung durch Screening	DACHS	Case/Control	Germany	1187	1092	68 (33-99)	40
Diet, Activity and Lifestyle Study	DALS	Case/Control	U.S.	579	570	65 (30-79)	44
Health Professionals Follow-up Study	HPFS	Cohort Study	U.S.	74	57	70 (50-89)	0
Kentucky Case-Control Study	Kentucky	Case/Control	U.S.	505	586	65 (21-95)	53
Melbourne Collaborative Cohort Study	MCCS	Cohort Study	Australia	125	226	68 (45-87)	48
Multi-Ethnic Cohort	MEC	Cohort Study	U.S.	120	187	70 (47-89)	46
Molecular Epidemiology of Colorectal Cancer	MECC	Case/Control	Israel	540	393	72 (25-95)	48
Newfoundland Case-Control Study	NFCCR	Case/Control	Canada	99	234	62 (23-76)	38
Nurses' Health Study	NHS	Cohort Study	U.S.	118	273	66 (46-83)	100
Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial	PLCO	Cohort Study	U.S.	426	414	70 (55-87)	42
VITamins And Lifestyle	VITAL	Cohort Study	U.S.	151	138	70 (51-83)	49
Women's Health Initiative	WHI	Cohort Study	U.S.	624	776	72 (52-89)	100

Table S2. Descriptive characteristics of study populations in the validation dataset

Study Name	Short Name	Design	Country	Cases	Controls	Mean Age (range)	Female (%)
Hawaii Colorectal Cancer Studies 2 and 3	Colo2&3	Case/Control	U.S.	39	54	66 (41-86)	39
Cancer Prevention Study II	CPS2	Cohort Study	U.S.	261	261	75 (58-89)	49
Darmkrebs: Chancen der Verhütung durch Screening	DACHS	Case/Control	Germany	1186	1108	69 (33-98)	40
Diet, Activity and Lifestyle Study	DALS	Case/Control	U.S.	533	600	65 (30-79)	46
Health Professionals Follow-up Study	HPFS	Cohort Study	U.S.	79	69	72 (50-90)	0
Kentucky Case-Control Study	Kentucky	Case/Control	U.S.	530	546	64 (30-90)	48
Melbourne Collaborative Cohort Study	MCCS	Cohort Study	Australia	142	237	69 (45-85)	48
Multi-Ethnic Cohort	MEC	Cohort Study	U.S.	125	159	69 (47-86)	46
Molecular Epidemiology of Colorectal Cancer	MECC	Case/Control	Israel	553	415	73 (26-98)	52
Newfoundland Case-Control Study	NFCCR	Case/Control	Canada	94	233	62 (35-76)	43
Nurses' Health Study	NHS	Cohort Study	U.S.	141	271	66 (46-86)	100
Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial	PLCO	Cohort Study	U.S.	368	447	70 (55-87)	40
VITamins And Lifestyle	VITAL	Cohort Study	U.S.	130	148	71 (52-83)	46
Women's Health Initiative	WHI	Cohort Study	U.S.	692	751	71 (50-91)	100

Table S3. Descriptive characteristics of environmental and lifestyle risk factors in study population of the validation dataset.

Variable	Men		Women	
	Cases (N=2229)	Controls (N=2429)	Cases (N=2644)	Controls (N=2870)
Age				
Mean (SD)	68.2 (9.8)	68.6 (9.8)	68.8 (9.8)	69.4 (8.8)
Height (cm)				
Mean (SD)	176.1 (7.5)	175.9 (7.5)	162.5 (6.5)	162.1 (6.4)
BMI (kg/m ²)				
Mean (SD)	27.7 (4.2)	27.0 (3.8)	27.6 (5.4)	26.8 (5.0)
Family history				
Yes (%)	306 (13.7)	243 (10.0)	414 (15.7)	372 (13.0)
Endoscopy history ¹				
Yes (%)	451 (20.2)	898 (37.0)	698 (26.4)	1039 (36.2)
Education ²				
Cat1 (%)	332 (14.9)	341 (14.0)	496 (18.8)	439 (15.3)
Cat2 (%)	699 (31.4)	623 (25.6)	686 (25.9)	730 (25.4)
Cat3 (%)	515 (23.1)	576 (23.7)	682 (25.8)	719 (25.1)
Cat4 (%)	649 (29.1)	861 (35.4)	745 (28.2)	948 (33.0)
Diabetes				
Yes (%)	285 (12.8)	251 (10.3)	307 (11.6)	191 (6.7)
Lifestyle factors				
Physical activity				
Yes (%)	1107 (49.7)	1250 (51.5)	1039 (39.3)	1094 (38.1)
Smoking status				
Ever smoker (%)	1461 (65.5)	1494 (61.5)	1180 (44.6)	1243 (43.3)
Smoking pack-years ³				
Mean (SD)	1.8 (1.5)	1.6 (1.4)	1.2 (1.5)	1.0 (1.4)
Alcohol consumption				
< 1g/day (%)	726 (32.6)	789 (32.5)	1508 (57.0)	1550 (54.0)
1-28 g/day (%)	916 (41.1)	1090 (44.9)	846 (32.0)	1075 (37.5)
>28 g/day (%)	376 (16.9)	396 (16.3)	105 (4.0)	105 (3.7)
Pharmaceutical factors				
Aspirin use				
Yes (%)	695 (31.2)	889 (36.6)	583 (22.0)	744 (25.9)
NSAIDs use				
Yes (%)	130 (5.8)	218 (9.0)	459 (17.4)	580 (20.2)
Post-menopausal hormone use				
Yes (%)	-	-	701 (26.5)	1026 (35.7)
Dietary factors				
Fiber ⁴				
Mean (SD)	1.6 (0.9)	1.5 (0.9)	1.6 (1.0)	1.5 (1.0)
Calcium ⁴				
Mean (SD)	1.6 (1.1)	1.5 (1.1)	1.6 (1.1)	1.5 (1.1)
Folate ⁴				
Mean (SD)	1.6 (0.9)	1.5 (0.9)	1.5 (1.0)	1.5 (1.0)
Processed meat ⁴				
Mean (SD)	1.5 (1.0)	1.4 (1.0)	1.5 (1.0)	1.4 (1.0)
Red meat ⁴				
Mean (SD)	1.4 (1.0)	1.3 (1.0)	1.5 (1.1)	1.4 (1.1)
Fruit ⁴				
Mean (SD)	1.7 (0.9)	1.6 (0.9)	1.7 (1.0)	1.6 (1.0)
Vegetable ⁴				
Mean (SD)	1.7 (0.9)	1.6 (0.9)	1.6 (1.0)	1.6 (1.0)
Total energy				
Mean (SD)	2297.8 (766.9)	2267.2 (708.8)	1715.8 (568.9)	1693.2 (563.1)
Combined risk scores				
E-score				
Mean (SD)	59.2 (28.4)	49.6 (29.1)	58.6 (28.4)	49.6 (29.1)
G-score				
Mean (SD)	57.0 (29.2)	49.4 (29.0)	57.1 (28.6)	49.6 (28.8)

¹ Endoscopy history was entirely missing in five studies (MCCS, MECC, Kentucky, NFCCR, Colo2&3), so these studies were excluded when we compute the summary statistic for endoscopy history.

² Education variable has four categories. Cat1: less than high school graduate, Cat2: high school graduate or completed GED, Cat3: some college or technical school, Cat4: college graduate or more.

³ Smoking pack-years among ever smokers was harmonized across studies by sex- and study-specific quartiles, and assigned values 1,2,3,4. For never smokers, it was assigned as "0". This variable was treated as continuous variable in the analysis.

⁴ Dietary variables (fiber, calcium, folate, processed meat, red meat, fruit, vegetable) were harmonized across studies by sex- and study-specific quartiles, and assigned values 0,1,2,3 in the order of increasing risk marginally. These variables were treated as continuous variables in the analysis.

Table S4. Associations between 19 environmental/lifestyle variables and risk of colorectal cancer in the study population of the model building dataset: Estimated weights, log(ORs), for the 19 variables to construct the environmental risk score (E-score) from a multivariable logistic regression analysis. The reference category (coded as 0) for each factor, except education, was that associated with the lowest CRC risk in published studies.

Parameter	Men		Women	
	Estimate log(OR)	P-value	Estimate log(OR)	P-value
Height (per 10cm)	-0.008	0.858	0.074	0.110
BMI (per 5kg/m ²)	0.226	< 10 ⁻⁵	0.107	3X10 ⁻⁴
Education ¹				
Reference (Cat1)	0		0	
Cat2	0.012	0.913	-0.064	0.501
Cat3	-0.230	0.037	-0.139	0.177
Cat4	-0.294	0.007	-0.129	0.218
Diabetes (Yes vs. No)	0.103	0.299	0.330	0.002
Physical activity (No vs. Yes)	-0.015	0.888	-0.215	0.024
Ever smoking (Yes vs. No)	-0.025	0.825	-0.057	0.611
Pack-years ²	0.056	0.115	0.067	0.079
Alcohol				
Reference (1-28g/day)	0		0	
1 g/day	0.080	0.288	0.120	0.071
>28 g/day	0.356	2X10 ⁻⁴	0.102	0.475
Aspirin (No vs. Yes)	0.359	< 10 ⁻⁵	0.258	2X10 ⁻⁴
NSAIDs (No vs. Yes)	0.206	0.075	0.116	0.143
Postmenopausal hormone use (No vs. Yes)	-	-	0.336	< 10 ⁻⁵
Fiber ³	0.045	0.419	0.009	0.844
Calcium ³	0.043	0.210	0.063	0.041
Folate ³	-0.020	0.619	-0.023	0.487
Processed meat ³	0.060	0.157	-0.052	0.160
Red meat ³	0.056	0.198	0.163	1X10 ⁻⁵
Fruit ³	-0.033	0.416	-0.003	0.923
Vegetable ³	0.097	0.022	0.045	0.207
Total energy ⁴ (per 742.7 kcal/day)	0.066	0.202	-0.090	0.112

¹ Education variable has four categories. Cat1: less than high school graduate, Cat2: high school graduate or completed GED, Cat3: some college or technical school, Cat4: college graduate or more.

² Smoking pack-years among ever smokers was harmonized across studies by sex- and study-specific quartiles, and assigned values 1,2,3,4. For never smokers, it was assigned as "0". This variable was treated as continuous variable in the analysis.

³ Dietary variables (fiber, calcium, folate, processed meat, red meat, fruit, vegetable) were harmonized across studies by sex- and study-specific quartiles, and assigned values 0,1,2,3 in the order of increasing risk marginally. These variables were treated as continuous variables in the analysis.

⁴ Total energy was scaled by dividing with its standard error 742.7 kcal/day for convenience.

Table S5. Associations between 63 previously reported common genetic variants (SNPs) and risk of colorectal cancer in the study population: Estimated weights, log(ORs), for each genetic variant from a multivariable logistic regression analysis^a.

SNP	Locus	Genetic Region	Risk/ other allele	Risk allele freq	Mean Imputation Rsq	Estimate log (OR)	Ref ^b
rs10911251	1q25.3	<i>LAMC1</i>	A/C	0.56	0.97	0.076	1,21
rs6687758	1q41	<i>DUSP10/CICP13</i>	G/A	0.20	1.00	0.056	2
rs6691170	1q41	<i>DUSP10/CICP13</i>	T/G	0.37	1.00	0.001	2
rs72647484	1p36.12	<i>WNT4/CDC42</i>	T/C	0.91	0.98	0.001	22
rs11903757	2q32.3	<i>NABP1/SDPR</i>	C/T	0.15	0.95	0.072	1
rs812481	3p14.1	<i>LRIG1</i>	G/C	0.50	1.00	0.065	19
rs35360328	3p22.1	<i>RP11/761N21.1</i>	A/T	0.14	0.97	0.127	19
rs10936599	3q26.2	<i>MYNN</i>	C/T	0.76	1.00	0.015	2
rs647161	5q31.1	<i>PITX1/H2AFY</i>	A/C	0.68	1.00	0.053	3
rs202110856 ^c	5q15	<i>ERAP1</i>	G/C	0.99	1.00	0.020	22
rs1321311	6p21.2	<i>SRSF3/CDKN1A</i>	A/C	0.25	1.00	0.078	4
rs4711689	6p21.1	<i>TFEB</i>	A/G	0.56	1.00	0.005	23
rs7758229	6q25.3	<i>SLC22A3</i>	T/G	0.32	1.00	0.039	20
rs16892766	8q23.3	<i>TRPS1/EIF3H</i>	C/A	0.08	1.00	0.200	5
rs2450115	8q23.3	<i>EIF3H</i>	T/C	0.81	1.00	0.042	23
rs6469656	8q23.3	<i>EIF3H</i>	A/G	0.88	1.00	-0.005	23
rs10505477	8q24	<i>MYC</i>	A/G	0.50	0.99	-0.008	7
rs6983267	8q24	<i>SRRM1P1/POU5F1B/MYC</i>	G/T	0.51	1.00	0.073	6-9
rs7014346	8q24	<i>MYC</i>	A/G	0.36	1.00	0.096	11
rs719725	9p24	<i>TPD52L3/UHRF2/GLDC</i>	A/C	0.62	1.00	0.035	7,10
rs10904849	10p13	<i>CUBN</i>	G/T	0.67	1.00	0.010	22
rs10795668	10p14	<i>KRT8P16/TCEB1P3</i>	G/A	0.69	1.00	0.017	5
rs704017	10q22.3	<i>ZMIZ1/AS1</i>	G/A	0.58	0.97	0.072	17
rs1035209	10q24.2	<i>snoU13</i>	T/C	0.18	1.00	-0.011	21
rs11190164	10q24.2	<i>snoU13</i>	G/A	0.25	1.00	0.087	19
rs4919687	10q24.3	<i>CYP17A1</i>	G/A	0.70	1.00	0.043	23
rs12241008	10q25.2	<i>VTI1A</i>	C/T	0.08	1.02	0.116	18
rs10506868	10q25.2	<i>VTI1A</i>	T/C	0.03	1.04	-0.056	23
rs11196172	10q25.2	<i>TCF7L2</i>	A/G	0.14	1.12	0.062	17
rs1535	11q12.2	<i>FADS2</i>	A/G	0.67	1.00	0.184	17
rs174537	11q12.2	<i>MYRF</i>	G/T	0.67	0.99	-0.095	17
rs174550	11q12.2	<i>FADS1</i>	T/C	0.67	0.99	0.134	17
rs4246215	11q12.2	<i>FEN1</i>	G/T	0.66	0.98	-0.165	17
rs3824999	11q13.4	<i>POLD3</i>	G/T	0.51	1.00	0.079	4

rs3802842	11q23	<i>C11orf93</i>	C/A	0.28	0.99	0.098	11
rs10849432	12p13.31	<i>CD9</i>	T/C	0.89	1.00	0.069	17
rs3217810	12p13.32	<i>CCND2</i>	T/C	0.12	0.90	0.143	1,21
rs10774214	12p13.32	<i>RPL18P9/CCND2</i>	T/C	0.37	0.99	0.040	3
rs11064437	12p13.32	<i>SPSB2</i>	C/T	0.99	0.97	0.218	23
rs11169552	12q13.13	<i>DIP2B/ATF1/LIMA1</i>	C/T	0.73	1.00	0.014	2
rs7136702	12q13.13	<i>LARP4/DIP2B/LIMA1</i>	T/C	0.34	0.96	0.034	2
rs3184504	12q24.12	<i>SH2B3</i>	C/T	0.49	1.00	0.089	19
rs73208120	12q24.22	<i>NOS1</i>	G/T	0.08	0.97	0.188	19
rs1957636	14q22.2	<i>BMP4/ATP5C1P1/CDKN3</i>	T/C	0.40	1.00	0.047	12
rs4444235	14q22.2	<i>BMP4/MIR5580</i>	C/T	0.46	1.00	0.086	12,13
rs4779584	15q13	<i>SCG5/GREM1</i>	T/C	0.19	1.00	0.105	12,14
rs11632715	15q13	<i>SCG5/GREM1</i>	A/G	0.47	1.00	0.025	12
rs16969681	15q13	<i>SCG5/GREM1</i>	T/C	0.09	1.00	0.046	12
rs79900961	16p13.2	<i>C16orf72</i>	G/A	0.98	0.73	0.030	22
rs9929218	16q22.1	<i>CDH1</i>	G/A	0.69	1.00	0.080	13
rs16941835	16q24.1	<i>FOXL1</i>	C/G	0.19	0.99	0.054	22
rs12603526	17p13.3	<i>NXN</i>	C/T	0.02	1.00	0.068	17
rs4939827	18q21.1	<i>SMAD7</i>	T/C	0.52	1.00	0.111	11,15
rs7229639	18q21.1	<i>SMAD7</i>	A/G	0.09	1.00	0.023	17
rs10411210	19q13.1	<i>RHPN2</i>	C/T	0.89	1.00	0.064	13
rs1800469	19q13.2	<i>TGFB1</i>	G/A	0.68	1.00	0.151	17
rs2241714	19q13.2	<i>B9D2</i>	C/T	0.67	1.00	-0.124	17
rs2423279	20p12.3	<i>BMP2/HAO1</i>	C/T	0.26	1.00	0.066	3
rs4813802	20p12.3	<i>FERMT1/BMP2</i>	G/T	0.34	0.99	0.076	12,16
rs961253	20p12.3	<i>FERMT1/BMP2</i>	A/C	0.35	1.00	0.094	13
rs6066825	20q13.1	<i>PREX1</i>	A/G	0.63	1.00	0.079	19
rs4925386	20q13.33	<i>LAMA5</i>	C/T	0.69	0.98	-0.001	2,16
rs6061231	20q13.33	<i>RPS21</i>	C/A	0.73	0.99	0.042	2

Abbreviations: SNP, single nucleotide polymorphism; freq., frequency

^aAdjusted for age in years, sex, genotype platforms, principal components

^bPrevious genome-wide association studies:

1) Peters et al. *Gastroenterology*, 2013.⁴⁷, 2) Houlston et al. *Nat Genet*, 2010.⁴⁸, 3) Jia et al. *Nat Genet*, 2013.⁴⁹, 4) Dunlop et al. *Nat Genet*, 2012.⁵⁰, 5) Tomlinson et al. *Nat Genet*, 2008.⁵¹, 6) Tomlinson et al. *Nat Genet*, 2007.⁵², 7) Zanke et al. *Nat Genet*, 2007.⁵³, 8) Haiman et al. *Nat Genet*, 2007.⁵⁴, 9) Hutter et al. *BMC Cancer*, 2010.⁵⁵, 10) Kocarnik et al. *CEBP*, 2010.⁵⁶, 11) Tenesa et al. *Nat Genet*, 2008.⁵⁷, 12) Tomlinson et al. *PLoS Genet*, 2011.⁵⁸, 13) COGENT. *Nat Genet*, 2008.⁵⁹, 14) Jaeger et al. *Nat Genet*, 2008.⁶⁰, 15) Broderick et al. *Nat Genet*, 2007.⁶¹, 16) Peters et al. *Hum Genet*, 2012.⁶², 17) Zhang et al. *Nat Genet*, 2014.⁶³, 18) Wang et al. *Nat Commun*, 2014.⁶⁴, 19) Schumacher et al. *Nat Commun*, 2015.⁶⁵, 20) Cui et al. *Gut*, 2011.⁶⁶, 21) Whiffin et al. *HMG*, 2014.⁶⁷, 22) Al-Tassan et al. *Sci Rep*, 2015.⁶⁸, 23) Zeng et al. *Gastroenterology*, 2016.⁶⁹

^c"rs202110856" is an indel, and "rs186474654" was used as a proxy when construct a genetic risk score.

Table S6. Colorectal cancer (CRC) incidence (cases per 100,000) and mortality (deaths per 100,000) for Whites in SEER13 (1992-2005). The other-cause mortality was computed by subtracting the CRC mortality from the all-cause mortality.

Age (years)	White Men			White Women		
	Incidence*	Mortality*		Incidence*	Mortality*	
	CRC	CRC	Other-causes	CRC	CRC	Other-causes
0-4	0.03	0.00	648.96	0.01	0.00	529.79
5-9	0.05	0.00	16.09	0.02	0.00	12.42
10-14	0.15	0.02	22.39	0.07	0.00	14.16
15-19	0.33	0.08	91.65	0.20	0.00	34.22
20-24	0.73	0.17	125.52	0.65	0.14	39.28
25-29	1.68	0.41	121.43	1.58	0.29	43.38
30-34	3.60	0.93	152.05	3.13	0.70	58.14
35-39	7.03	1.64	204.99	5.68	1.36	87.92
40-44	13.01	3.46	283.79	11.64	2.73	135.14
45-49	25.89	7.00	390.72	21.66	5.06	205.89
50-54	53.35	13.88	550.94	38.26	9.39	320.86
55-59	91.06	26.23	822.97	62.37	16.51	507.52
60-64	145.61	43.97	1323.97	98.50	28.16	835.83
65-69	224.75	69.98	2087.97	146.94	43.06	1326.94
70-74	299.24	95.29	3305.94	205.40	60.31	2137.19
75-79	379.69	126.55	5187.57	269.28	86.09	3402.75
80-84	453.80	177.16	8486.57	335.04	121.69	5774.61
≥ 85	464.39	241.93	16681.69	357.54	202.17	13808.75

* Age-adjusted rates using 2000 US standard population.

Table S7. Odds ratio (95% CI) of risk factors associated with CRC risk in the model building dataset. A G-score was constructed based on only 54 known GWAS SNPs which were not identified by our consortia.

Variable	Men (N=4,666)	Women (N=5,500)
	OR (95% CI)	OR (95% CI)
E-score*	1.36 (1.29 to 1.44)	1.34 (1.28 to 1.41)
G-score*. ¹	1.30 (1.23 to 1.37)	1.26 (1.19 to 1.31)
Family history		
No	1.00	1.00
Yes	1.67 (1.37 to 2.02)	1.46 (1.24 to 1.72)
Endoscopy history		
No	1.00	1.00
Yes	0.29 (0.24 to 0.34)	0.54 (0.47 to 0.62)
Missing	0.47 (0.30 to 0.73)	0.68 (0.50 to 0.93)

The logistic regression model includes study, age, E-score, G-score, family history, endoscopy history, missing indicator for endoscopy history, genotype platform, and PCs. * ORs for E-score and G-score per quartile increase. ¹G-score was constructed by using 54 known GWAS SNPs after excluding 9 SNPs discovered by GECCO, CCFR, CORECT: rs11903757, rs10911251, rs3217810, rs35360328, rs812481, rs11190164, rs3184504, rs73208120, rs6066825.

Table S8. Odds ratio (95% CI) of risk factors associated with CRC risk in the model building dataset. Three studies (MECC, Kentucky, NFCCR) were excluded.

Variable	Men (N=3,462)	Women (N=4,347)
	OR (95% CI)	OR (95% CI)
E-score*	1.37 (1.29 to 1.46)	1.32 (1.25 to 1.40)
G-score*	1.38 (1.30 to 1.47)	1.31 (1.24 to 1.39)
Family history		
No	1.00	1.00
Yes	1.71 (1.36 to 2.16)	1.48 (1.23 to 1.78)
Endoscopy history		
No	1.00	1.00
Yes	0.28 (0.24 to 0.33)	0.53 (0.46 to 0.61)
Missing	0.46 (0.30 to 0.71)	0.67 (0.49 to 0.91)

The logistic regression model includes study, age, E-score, G-score, family history, endoscopy history, genotype platform, and PCs.

* ORs for E-score and G-score per quartile increase

Table S9. Risk of colorectal cancer according to risk factors included in the risk model in the study population of the model building dataset

Parameter	Men	Women
	OPERA ⁴ (95% CI)	OPERA ⁴ (95% CI)
Height (per 10cm)	1.00 (0.94 to 1.07)	1.05 (0.99 to 1.11)
BMI (per 5kg/m ²)	1.19 (1.12 to 1.26)	1.11 (1.05 to 1.17)
Education	1.12 (1.05 to 1.19)	1.04 (0.98 to 1.10)
Diabetes	1.03 (0.97 to 1.10)	1.09 (1.04 to 1.16)
Physical activity	1.01 (0.95 to 1.07)	1.06 (1.01 to 1.12)
Ever smoking	1.01 (0.95 to 1.07)	1.02 (0.96 to 1.07)
Pack-years ¹	1.05 (0.99 to 1.12)	1.05 (0.99 to 1.11)
Alcohol	1.12 (1.05 to 1.19)	1.05 (0.99 to 1.11)
Aspirin	1.17 (1.10 to 1.24)	1.11 (1.05 to 1.17)
NSAIDs	1.06 (0.99 to 1.12)	1.04 (0.99 to 1.10)
Postmenopausal hormone use	-	1.15 (1.09 to 1.22)
Fiber ²	1.03 (0.97 to 1.09)	1.00 (0.95 to 1.06)
Calcium ²	1.04 (0.98 to 1.11)	1.06 (1.00 to 1.12)
Folate ²	1.02 (0.96 to 1.08)	1.02 (0.96 to 1.08)
Processed meat ²	1.05 (0.99 to 1.11)	1.04 (0.99 to 1.10)
Red meat ²	1.04 (0.98 to 1.11)	1.13 (1.07 to 1.20)
Fruit ²	1.02 (0.96 to 1.09)	1.00 (0.95 to 1.06)
Vegetable ²	1.07 (1.01 to 1.14)	1.04 (0.98 to 1.10)
Total energy ³ (per 742.7 kcal/day)	1.04 (0.98 to 1.11)	1.05 (0.99 to 1.11)

All of the estimates from the logistic regression analysis were adjusted for study, age, family history, endoscopy history, and the variables included into the model.

¹ Smoking pack-years among ever smokers was harmonized across studies by sex- and study-specific quartiles, and assigned values 1,2,3,4. For never smokers, it was assigned as "0". This variable was treated as continuous variable in the analysis.

² Dietary variables (fiber, calcium, folate, processed meat, red meat, fruit, vegetable) were harmonized across studies by sex- and study-specific quartiles, and assigned values 0,1,2,3 in the order of increasing risk marginally. These variables were treated as continuous variables in the analysis.

³ Total energy was scaled by dividing with its standard error 742.7 kcal/day for convenience.

⁴ Odds per adjusted standard deviation (Hopper, 2015⁷⁰)

Table S10. Odds ratio (95% CI) of risk factors associated with CRC risk in the subset of training data.

Variable	Men (N= 3,755)	Women (N= 2,988)
	OR (95% CI)	OR (95% CI)
E-score*	1.40 (1.32 to 1.49)	1.46 (1.36 to 1.56)
G-score*	1.34 (1.26 to 1.43)	1.30 (1.21 to 1.39)
Number of first-degree relatives	1.00 (0.97 to 1.03)	1.01 (0.98 to 1.04)
Number of first-degree relatives with CRC	1.77 (1.07 to 2.91)	1.78 (1.07 to 2.95)
Youngest diagnosis age among first-degree CRC relatives	1.00 (0.99 to 1.01)	1.00 (0.99 to 1.01)
Endoscopy history		
No	1.00	1.00
Yes	0.27 (0.22 to 0.32)	0.30 (0.24 to 0.37)
Missing	0.75 (0.43 to 1.29)	0.60 (0.37 to 0.99)

The logistic regression model includes study, age, E-score, G-score, family history, endoscopy history, genotype platform, and PCs.

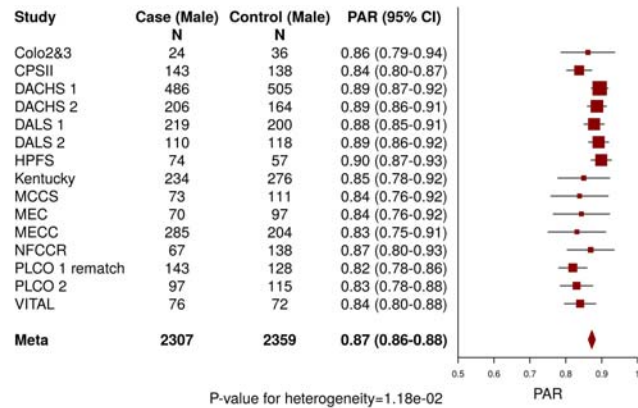
* ORs for E-score and G-score per quartile increase

Table S11. AUC comparisons between risk prediction models in the subset of validation data.

	Men (N= 3,755)	Women (N= 2,988)
	AUC (95% CI)	AUC (95% CI)
Model I		
Family variables & E-score	0.61 (0.60 to 0.62)	0.61 (0.60 to 0.63)
Model II		
Family variables & G-score	0.59 (0.58 to 0.61)	0.60 (0.59 to 0.62)
Model III		
Family variables & E-score & G-score	0.63 (0.62 to 0.64)	0.63 (0.62 to 0.65)

The analyses were adjusted for study, age, and endoscopy history.

(a) Men



(b) Women

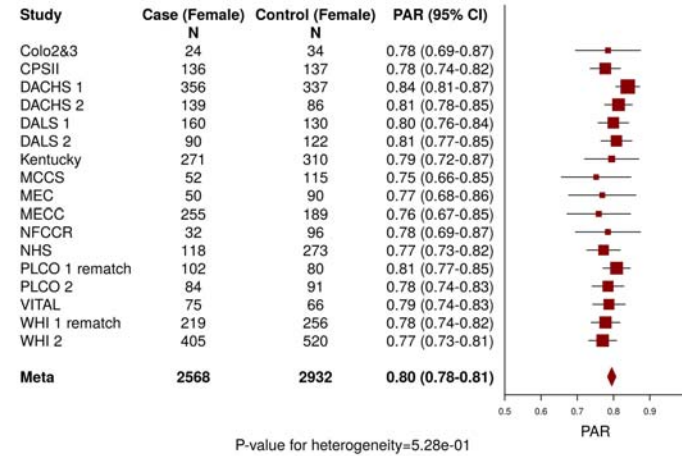


Figure S1. Study-specific population attributable risk (PAR) estimates.

Process of data harmonization

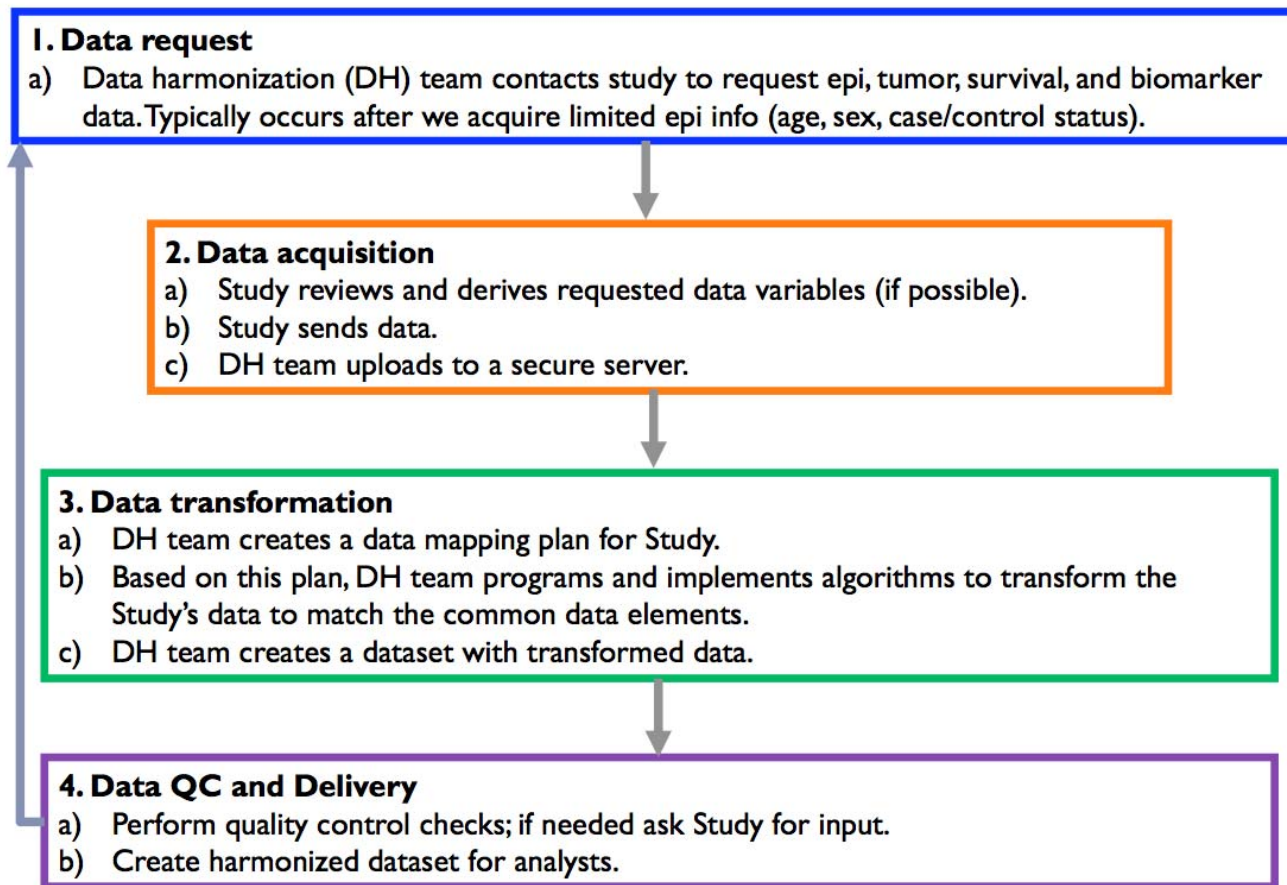


Figure S2. Flowchart of the data harmonization process.

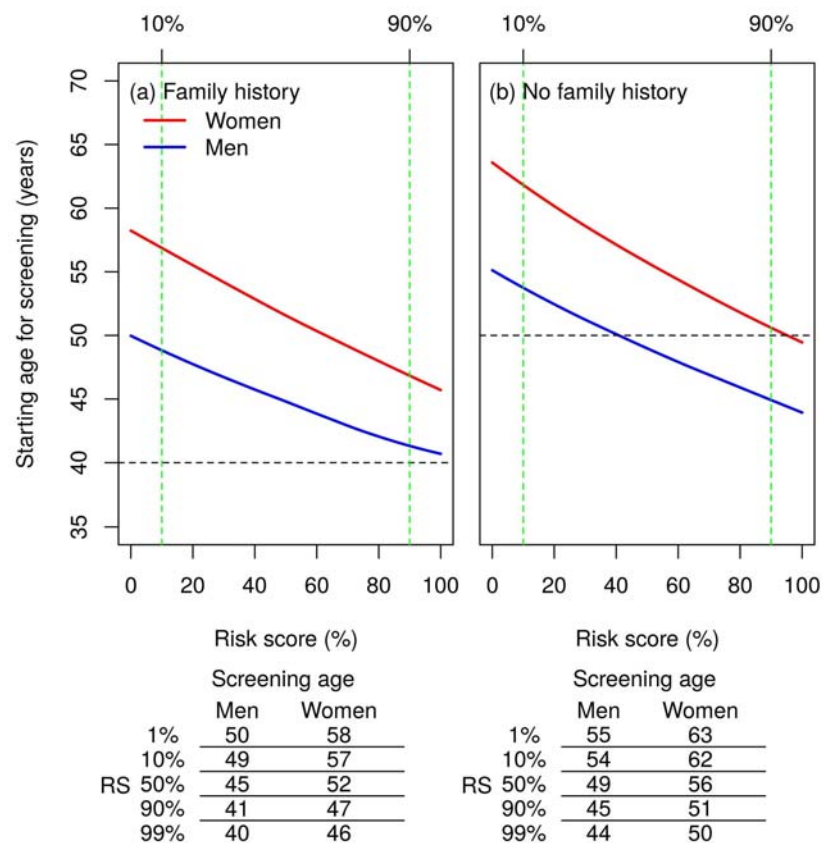


Figure S3. Recommended age to start CRC screening by E-score. The horizontal lines represent the recommended age for the first endoscopy depending on family history in the current screening guideline for CRC. The risk threshold to determine the age for the first screening was set as the average of 10-year CRC risks for a 50-year-old man (1.25%) and woman (0.68%), i.e., $(1.25\%+0.68\%)/2=0.97\%$, who have not previously received an endoscopy.

References


1. Le Marchand L, Hankin JH, Wilkens LR, et al. Combined effects of well-done red meat, smoking, and rapid N-acetyltransferase 2 and CYP1A2 phenotypes in increasing colorectal cancer risk. *Cancer Epidemiol Biomarkers Prev* 2001;10:1259-1266.
2. Calle EE, Rodriguez C, Jacobs EJ, et al. The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer* 2002;94:2490-2501.
3. Brenner H, Chang-Claude J, Seiler CM, Rickert A, Hoffmeister M. Protection from colorectal cancer after colonoscopy: a population-based, case-control study. *Ann Intern Med* 2011;154:22-30.
4. Lilla C, Verla-Tebit E, Risch A, et al. Effect of NAT1 and NAT2 genetic polymorphisms on colorectal cancer risk associated with exposure to tobacco smoke and meat consumption. *Cancer Epidemiol Biomarkers Prev* 2006;15:99-107.
5. Slattery ML, Potter J, Caan B, et al. Energy balance and colon cancer--beyond physical activity. *Cancer Res* 1997;57:75-80.
6. Rimm EB, Stampfer MJ, Colditz GA, Chute CG, Litin LB, Willett WC. Validity of self-reported waist and hip circumferences in men and women. *Epidemiology* 1990;1:466-473.
7. Giles GG, English DR. The Melbourne Collaborative Cohort Study. *IARC Sci Publ* 2002;156:69-70.
8. Kolonel LN, Henderson BE, Hankin JH, et al. A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am J Epidemiol* 2000;151:346-357.
9. Poynter JN, Gruber SB, Higgins PD, et al. Statins and the risk of colorectal cancer. *N Engl J Med* 2005;352:2184-2192.
10. Belanger CF, Hennekens CH, Rosner B, Speizer FE. The nurses' health study. *Am J Nurs* 1978;78:1039-1040.
11. National Cancer Institute. Cancer Genetic Markers of Susceptibility (CGEMS) data website.; 2010. Available at: <http://soc.ics.uci.edu/Resources/collab.php?650>. Accessed 4/27, 2017.
12. Yeager M, Chatterjee N, Ciampa J, et al. Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat Genet* 2009;41:1055-1057.
13. Landi MT, Chatterjee N, Yu K, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet* 2009;85:679-691.

14. White E, Patterson RE, Kristal AR, et al. VITamins And Lifestyle cohort study: study design and characteristics of supplement users. *Am J Epidemiol* 2004;159:83-93.
15. Hays J, Hunt JR, Hubbell FA, et al. The Women's Health Initiative recruitment methods and results. *Ann Epidemiol* 2003;13:S18-77.
16. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials* 1998;19:61-109.
17. Bergstralh E.J. KJL. Computerized matching of cases to controls. 1995; Technical Report Number 56.
18. Hutter CM, Chang-Claude J, Slattery ML, et al. Characterization of gene-environment interactions for colorectal cancer susceptibility loci. *Cancer Res* 2012;72:2036-2044.
19. Thrift AP, Gong J, Peters U, et al. Mendelian randomization study of height and risk of colorectal cancer. *Int J Epidemiol* 2015;44:662-672.
20. Moghaddam AA, Woodward M, Huxley R. Obesity and risk of colorectal cancer: a meta-analysis of 31 studies with 70,000 events. *Cancer Epidemiol Biomarkers Prev* 2007;16:2533-2547.
21. Kantor ED, Hutter CM, Minnier J, et al. Gene-environment interaction involving recently identified colorectal cancer susceptibility Loci. *Cancer Epidemiol Biomarkers Prev* 2014;23:1824-1833.
22. Thrift AP, Gong J, Peters U, et al. Mendelian Randomization Study of Body Mass Index and Colorectal Cancer Risk. *Cancer Epidemiol Biomarkers Prev* 2015;24:1024-1031.
23. American Cancer Society. *Colorectal Cancer Facts & Figures 2014-2016*. 2014.
24. Leufkens AM, Van Duijnhoven FJ, Boshuizen HC, et al. Educational level and risk of colorectal cancer in EPIC with specific reference to tumor location. *Int J Cancer* 2012;130:622-630.
25. Doubeni CA, Laiyemo AO, Major JM, et al. Socioeconomic status and the risk of colorectal cancer: an analysis of more than a half million adults in the National Institutes of Health-AARP Diet and Health Study. *Cancer* 2012;118:3636-3644.
26. Larsson SC, Orsini N, Wolk A. Diabetes mellitus and risk of colorectal cancer: a meta-analysis. *J Natl Cancer Inst* 2005;97:1679-1687.
27. Slattery ML. Physical activity and colorectal cancer. *Sports Med* 2004;34:239-252.
28. Botteri E, Iodice S, Bagnardi V, Raimondi S, Lowenfels AB, Maisonneuve P. Smoking and colorectal cancer: a meta-analysis. *JAMA* 2008;300:2765-2778.

29. Liang PS, Chen TY, Giovannucci E. Cigarette smoking and colorectal cancer incidence and mortality: systematic review and meta-analysis. *Int J Cancer* 2009;124:2406-2415.
30. Gong J, Hutter C, Baron JA, et al. A pooled analysis of smoking and colorectal cancer: timing of exposure and interactions with environmental factors. *Cancer Epidemiol Biomarkers Prev* 2012;21:1974-1985.
31. Gong J, Hutter CM, Newcomb PA, et al. Genome-Wide Interaction Analyses between Genetic Variants and Alcohol Consumption and Smoking for Risk of Colorectal Cancer. *PLoS Genet* 2016;12:e1006296.
32. Cho E, Smith-Warner SA, Ritz J, et al. Alcohol intake and colorectal cancer: a pooled analysis of 8 cohort studies. *Ann Intern Med* 2004;140:603-613.
33. Chubak J, Kamineni A, Buist DSM, Anderson ML, Whitlock EP. Aspirin Use for the Prevention of Colorectal Cancer: An Updated Systematic Evidence Review for the U.S. Preventive Services Task Force [Internet]. 2015;Report No.: 15-05228-EF-1.
34. Nan H, Hutter CM, Lin Y, et al. Association of aspirin and NSAID use with risk of colorectal cancer according to genetic variants. *JAMA* 2015;313:1133-1142.
35. Grodstein F, Newcomb PA, Stampfer MJ. Postmenopausal hormone therapy and the risk of colorectal cancer: a review and meta-analysis. *Am J Med* 1999;106:574-582.
36. Garcia-Albeniz X, Rudolph A, Hutter C, et al. CYP24A1 variant modifies the association between use of oestrogen plus progestogen therapy and colorectal cancer risk. *Br J Cancer* 2016;114:221-229.
37. Giovannucci E, Stampfer MJ, Colditz GA, et al. Multivitamin use, folate, and colon cancer in women in the Nurses' Health Study. *Ann Intern Med* 1998;129:517-524.
38. Sutor CW, Bailey LB. Dietary folate equivalents: interpretation and application. *J Am Diet Assoc* 2000;100:88-94.
39. Cho E, Smith-Warner SA, Spiegelman D, et al. Dairy foods, calcium, and colorectal cancer: a pooled analysis of 10 cohort studies. *J Natl Cancer Inst* 2004;96:1015-1022.
40. Park Y, Hunter DJ, Spiegelman D, et al. Dietary fiber intake and risk of colorectal cancer: a pooled analysis of prospective cohort studies. *JAMA* 2005;294:2849-2857.
41. Chao A, Thun MJ, Connell CJ, et al. Meat consumption and risk of colorectal cancer. *JAMA* 2005;293:172-182.
42. Koushik A, Hunter DJ, Spiegelman D, et al. Fruits, vegetables, and colon cancer risk in a pooled analysis of 14 cohort studies. *J Natl Cancer Inst* 2007;99:1471-1483.

43. Kim DH, Smith-Warner SA, Spiegelman D, et al. Pooled analyses of 13 prospective cohort studies on folate intake and colon cancer. *Cancer Causes Control* 2010;21:1919-1930.
44. Figueiredo JC, Hsu L, Hutter CM, et al. Genome-wide diet-gene interaction analyses for risk of colorectal cancer. *PLoS Genet* 2014;10:e1004228.
45. 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;23:2971-2976.
46. Ananthakrishnan AN, Du M, Berndt SI, et al. Red meat intake, NAT2, and risk of colorectal cancer: a pooled analysis of 11 studies. *Cancer Epidemiol Biomarkers Prev* 2015;24:198-205.
47. Peters U, Jiao S, Schumacher FR, et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology* 2013;144:799-807.e24.
48. Houlston RS, Cheadle J, Dobbins SE, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat Genet* 2010;42:973-977.
49. Jia WH, Zhang B, Matsuo K, et al. Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nat Genet* 2013;45:191-196.
50. Dunlop MG, Dobbins SE, Farrington SM, et al. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet* 2012;44:770-776.
51. Tomlinson IP, Webb E, Carvajal-Carmona L, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat Genet* 2008;40:623-630.
52. Tomlinson I, Webb E, Carvajal-Carmona L, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat Genet* 2007;39:984-988.
53. Zanke BW, Greenwood CM, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat Genet* 2007;39:989-994.
54. Haiman CA, Le Marchand L, Yamamoto J, et al. A common genetic risk factor for colorectal and prostate cancer. *Nat Genet* 2007;39:954-956.
55. Hutter CM, Slattery ML, Duggan DJ, et al. Characterization of the association between 8q24 and colon cancer: gene-environment exploration and meta-analysis. *BMC Cancer* 2010;10:670.

56. Kocarnik JD, Hutter CM, Slattery ML, et al. Characterization of 9p24 risk locus and colorectal adenoma and cancer: gene-environment interaction and meta-analysis. *Cancer Epidemiol Biomarkers Prev* 2010;19:3131-3139.
57. Tenesa A, Farrington SM, Prendergast JG, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat Genet* 2008;40:631-637.
58. Tomlinson IP, Carvajal-Carmona LG, Dobbins SE, et al. Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet* 2011;7:e1002105.
59. COGENT Study, Houlston RS, Webb E, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* 2008;40:1426-1435.
60. Jaeger E, Webb E, Howarth K, et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat Genet* 2008;40:26-28.
61. Broderick P, Carvajal-Carmona L, Pittman AM, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet* 2007;39:1315-1317.
62. Peters U, Hutter CM, Hsu L, et al. Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum Genet* 2012;131:217-234.
63. Zhang B, Jia WH, Matsuda K, et al. Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat Genet* 2014;46:533-542.
64. Wang H, Burnett T, Kono S, et al. Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. *Nat Commun* 2014;5:4613.
65. Schumacher FR, Schmit SL, Jiao S, et al. Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nat Commun* 2015;6:7138.
66. Cui R, Okada Y, Jang SG, et al. Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. *Gut* 2011;60:799-805.
67. Whiffin N, Hosking FJ, Farrington SM, et al. Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Hum Mol Genet* 2014;23:4729-4737.
68. Al-Tassan NA, Whiffin N, Hosking FJ, et al. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Sci Rep* 2015;5:10442.



69. Zeng C, Matsuda K, Jia WH, et al. Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. *Gastroenterology* 2016;150:1633-1645.

70. Hopper JL. Odds per adjusted standard deviation: comparing strengths of associations for risk factors measured on different scales and across diseases and populations. *Am J Epidemiol* 2015;182:863-867.

Author names in bold designate shared co-first authorship.