Supplementary Materials and Methods

Study Populations Included in GWAS and Follow-up Studies

GWAS in GECCO and CCFR. We describe each study population used in the GWAS. For information on sample sizes and demographic factors please see Supplementary Table 1.

Ontario Familial Colorectal Cancer Registry. In GECCO, a subset of the Assessment of Risk in Colorectal Tumours in Canada from the OFCCR (Ontario Registry for Studies of Familial Colorectal Cancer) was used. Both the case-control study[1] and the OFCCR[2] have been described in detail previously, as have the GWAS results.[3] In brief, cases were confirmed incident colorectal cancer cases if they were ages 20 to 74 years, residents of Ontario, identified through comprehensive registry, and diagnosed between July 1997 and June 2000. Population-based controls were selected randomly among Ontario residents (random-digit dialing and listing of all Ontario residents), and matched by sex and 5-year age groups. A total of 1236 colorectal cancer cases and 1223 controls were genotyped successfully on at least one of the following: Illumina 1536 GoldenGate assay (Illumina, Inc, San Diego, CA), the Affymetrix GeneChip Human Mapping 100K and 500K Array Set (Affymetrix, Inc, Santa Clara, CA), or a 10K nonsynonymous SNP chip. Analysis was based on a set of unrelated subjects who were non-Hispanic, white by self-report, or by investigation of genetic ancestry. We further excluded subjects if there was a sample mix-up, if they were missing epidemiologic questionnaire data, if they were cases with a tumor in the appendix, or if they overlapped with the CCFR. In addition, only samples genotyped on the Affymetrix GeneChip 500K Array were used to avoid coverage issues in imputation.

The french Association STudy Evaluating RISK for sporadic colorectal cancer. Participants were recruited from the Pays de la Loire region in France between December 2002 and March 2006.[4] Eligibility criteria for cases included being Caucasian, age 40 years or older at diagnosis, and having no family history of colorectal cancer or polyps. Cases were patients with first primary colorectal cancer diagnosed in 1 of the 6 public hospitals and 5 clinics located in the Pays de la Loire region that participated in the study. Cases were confirmed based on medical and pathology reports. Controls were recruited at 2 Health Examination Centers of the Pays de la Loire region, and the recruitment of controls age 70 years and older was completed in the Departments of Internal Medicine and Hepatogastroenterology of the University Hospital Center of Nantes, located in the same region. Controls were eligible to participate if they were Caucasian, age 40 years or older, and had no family history of colorectal cancer or polyps. In the presence of the physician, each participant filled out a standardized questionnaire on family information, medical history, lifestyle, and dietary intake. Cases and controls provided a blood sample.

CCFR. The CCFR is a National Cancer Institute–supported consortium consisting of 6 centers dedicated to the establishment of a comprehensive collaborative infrastructure for interdisciplinary studies in the genetic epidemiology of colorectal cancer.[5] The CCFR includes data from approximately 30,500 total subjects (10,500 probands and 20,000 unaffected and affected relatives and unrelated controls). Cases and controls, age 20–74 years, were recruited at the 6 participating centers beginning in 1998. CCFR implemented a standardized questionnaire that was administered to all participants and included established and suspected risk factors for colorectal cancer, which included questions on medical history and medication use, reproductive history (for female participants), family history, physical activity, demographics, alcohol and tobacco use, and dietary factors. The set 1 scan, which has been described previously,[6] included population-based cases and age-matched controls from the 3 population-based centers: Seattle, Toronto, and Australia. Cases were enriched genetically by oversampling those with a young age at onset or positive family history. Controls were matched to cases on age and sex. All cases and controls were self-reported as white, which was confirmed with genotype data. The set 2 scan included population-based cases and matched controls from all 6 colon CFR centers including the Mayo Clinic, Hawaii Cancer Registry, University of Southern California, Fred Hutchinson Cancer Research Center, Ontario Cancer Care, and University of Melbourne. As with set 1, cases were enriched genetically by oversampling those with a young age at onset or positive family history. Controls were same-generation family controls.

Darmkrebs: Chancen der Verhütung durch Screening. This 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.[7,8] Cases with a first diagnosis of invasive colorectal cancer (International Classification of Diseases 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 1-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 after 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 selected randomly from population registries, using 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 telephone calls. The participation rate was 51%. During an in-person interview, data were collected on demographics, medical history, family history of colorectal cancer, and various lifestyle factors, as were blood and mouthwash samples. The set 1 scan consisted of a subset of participants recruited up until 2007, and samples were frequency matched on age and sex. The set 2 scan consisted of additional subjects who were recruited until 2010 as part of this ongoing study.

Diet, Activity, and Lifestyle Study. DALS (Diet, Activity, and Lifestyle Study) was a population-based, case-control study of colon cancer. Participants were recruited between 1991 and 1994 from 3 locations: the Kaiser Permanente Medical Care Program of Northern California, an 8-county area in Utah, and the metropolitan Twin Cities area of Minnesota.[9] Eligibility criteria for cases included age at diagnosis between 30 and 79 years; diagnosis with first primary colon cancer (International Classification of Diseases for Oncology second edition codes 18.0 and 18.2–18.9) between October 1, 1991, and September 30, 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, Crohn's 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 4 months of diagnosis. Controls from the Kaiser Permanente Medical Care Program were selected randomly from membership lists. In Utah, controls younger than 65 years of age were selected randomly through random-digit dialing and driver's license lists. Controls, 65 years of age and older, were selected randomly from Health Care Financing Administration lists. In Minnesota, controls were identified from Minnesota driver's licenses or state identification lists. Controls were matched to cases by 5-year age groups and sex. The set I scan consisted of a subset of the study designed earlier, from Utah, Minnesota, and the Kaiser Permanente Medical Care Program, and was restricted to subjects who self-reported as white non-Hispanic. The set 2 scan consisted of subjects from Utah and Minnesota who were not genotyped in set 1. Set 2 was restricted to subjects who self-reported as white non-Hispanic and those who had appropriate consent to post data to dbGaP.

Hawaii Colorectal Cancer Studies 2 and 3. Patients with colorectal cancer were identified through the rapid reporting system of the Hawaii Surveillance, Epidemiology and End Results 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.[10] Control subjects were selected from participants in an ongoing 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 2 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 (Food Frequency Questionnaire), and postmenopausal hormone use. A blood sample was obtained from 548 (71%) interviewed cases and 662 (86%) interviewed controls. Surveillance, Epidemiology and End Results 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.

Health Professionals Follow-up Study. The HPFS (Health Professionals Follow-up Study) is a parallel prospective study to the NHS (Nurses' Health Study).[11] The HPFS cohort comprised 51,529 men who, in 1986, responded to a mailed questionnaire. The participants were US male dentists, optometrists, osteopaths, podiatrists, pharmacists, and veterinarians born between 1910 and 1946. Participants 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. Colorectal cancer and other outcomes were reported by participants or next-of-kin and were 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. Incident cases were defined as those occurring after the subject provided the blood sample. Prevalent cases were defined as those occurring after enrollment in the study but before the subject provided the blood sample. Follow-up evaluation has been excellent, with 94% of the men responding to date. Colorectal cancer cases were ascertained through January 1, 2008. In 1993–1995, 18,825 men in the HPFS mailed blood samples by overnight courier, which were aliquoted into buffy coat and stored in liquid nitrogen. In 2001–2004, 13,956 men in the HPFS who had not provided a blood sample previously mailed in a swish-and-spit sample of buccal cells. Incident cases were defined as those occurring after the subject provided a blood or buccal sample. Prevalent cases were defined as those occurring after enrollment in the study in 1986, but before the subject provided either a blood or buccal sample. After excluding participants

with histories of cancer (except nonmelanoma skin cancer), ulcerative colitis, or familial polyposis, 2 case-control sets were constructed from which DNA was isolated from either buffy coat or buccal cells for genotyping, as follows: (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; and (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 6 months). Cases were pair-matched 1:1, 1:2, or 1:3 with a control participant(s).

In addition to colorectal cancer cases and controls, a set of adenoma cases and matched controls with available DNA from buffy coat were selected for genotyping. Over the follow-up period, data were collected on endoscopic screening practices and, if individuals had been diagnosed with a polyp, the polyps were confirmed to be adenomatous by medical record review. Adenoma cases were ascertained through January 1, 2008. A separate case-control set was constructed of participants diagnosed with advanced adenoma matched to control participants who underwent a lower endoscopy in the same time period and did not have an adenoma. Advanced adenoma was defined as an adenoma 1 cm or larger in diameter and/or with tubulovillous, villous, or high-grade dysplasia/carcinoma-in-situ histology. Matching criteria included year of birth (within 1 year) and month/year of blood sampling (within 6 months), the reason for their lower endoscopy (screening, family history, or symptoms), and the time period of any prior endoscopy (within 2 years). Controls matched to cases with a distal adenoma either had a negative sigmoidoscopy or colonoscopy examination, and controls matched to cases with proximal adenoma all had a negative colonoscopy.

Multiethnic Cohort study. The MEC (Multiethnic Cohort) was initiated in 1993 to investigate the impact of dietary and environmental factors on major chronic diseases, particularly cancer, in ethnically diverse populations in Hawaii and California.[12] The study recruited 96,810 men and 118,441 women aged 45–75 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 mo) among cases and the participation rate was 74%. A sample of cohort participants was selected randomly 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 were identified through the Rapid Reporting System of the Hawaii Tumor Registry and through quarterly linkage to the Los Angeles County Cancer Surveillance Program. Both registries are members of Surveillance, Epidemiology and End Results. In GECCO, self-reported white subjects from the nested case-control study described earlier with DNA and clinical and epidemiologic data were selected for genotyping.

Nurses' Health Study. The NHS cohort began in 1976 when 121,700 married female registered nurses age 30–55 years returned the initial questionnaire that ascertained a variety of important health-related exposures.[13] Since 1976, follow-up questionnaires have been mailed every 2 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. The rate of follow-up evaluation has been high: as a proportion of the total possible follow-up time, follow-up evaluation has been more than 92%. Colorectal cancer cases were ascertained through June 1, 2008. In 1989–1990, 32,826 women in NHS I mailed blood samples by overnight courier, which were aliquoted into buffy coat and stored in liquid nitrogen. In 2001–2004, 29,684 women in NHS I who did not previously provide a blood sample mailed a swish-and-spit sample of buccal cells. Incident cases were defined as those occurring after the subject provided a blood or buccal sample. Prevalent cases were defined as those occurring after enrollment in the study in 1976 but before the subject provided either a blood or buccal sample. After excluding participants with histories of cancer (except nonmelanoma skin cancer), ulcerative colitis, or familial polyposis, 2 case-control sets were constructed from which DNA was isolated from either buffy 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 the colorectal cancer was diagnosed in the case; and (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 cases. For both case-control sets, matching criteria included year of birth (within 1 year) and month/year of blood or buccal cell sampling (within 6 months). Cases were pair matched 1:1, 1:2, or 1:3 with a control participant(s).

In addition to colorectal cancer cases and controls, a set of adenoma cases and matched controls with available DNA from buffy coat were selected for genotyping. Over the follow-up period, data were collected on endoscopic screening practices and, if individuals had been diagnosed with a polyp, the polyps were confirmed to be adenomatous by medical record review. Adenoma cases

were ascertained through June 1, 2008. A separate case-control set was constructed of participants diagnosed with advanced adenoma matched to control participants who underwent a lower endoscopy in the same time period and did not have an adenoma. Advanced adenoma was defined as an adenoma more than 1 cm in diameter and/or with tubulovillous, villous, or high-grade dysplasia/carcinoma-in-situ histology. Matching criteria included year of birth (within 1 year) and month/year of blood sampling (within 6 months), the reason for their lower endoscopy (screening, family history, or symptoms), and the time period of any prior endoscopy (within 2 years). Controls matched to cases with a distal adenoma either had a negative sigmoidoscopy or colonoscopy examination, and controls matched to cases with proximal adenoma all had a negative colonoscopy.

Physicians' Health Study. The PHS (Physicians' Health Study) was established as a randomized, double-blind, placebo-controlled trial of aspirin and β-carotene among 22,071 healthy US male physicians, between 40 and 84 years of age, in 1982.[14,15] Participants completed 2 mailed questionnaires before being assigned randomly, additional questionnaires at 6 and 12 months, and questionnaires annually thereafter. In addition, participants were sent postcards at 6 months to ascertain status. From August 1982 to December 1984, there were 14,916 baseline blood samples collected from the physicians during the run-in phase before randomization. When participants reported a diagnosis of cancer, medical records and pathology reports were reviewed by study physicians who were blinded to exposure data. Among those who provided baseline blood samples, colorectal cases were ascertained through March 31, 2008, and controls were matched on age (within 1 year for younger participants, up to 5 years for older participants) and smoking status (never, past, current). Cases were pair-matched 1:1, 1:2, or 1:3 with a control participant(s). Because of DNA availability, samples were genotyped in 2 batches on the same platform at the same genotyping center at different time points.

Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. The PLCO (Prostate, Lung, Colorectal Cancer, and Ovarian Cancer Screening Trial) enrolled 154,934 participants (men and women, aged between 55 and 74 y) at 10 centers into a large, randomized, 2-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; the participation rate was 65%. Details of this study have been described previously[16,17] 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 originated from the Cancer Genetic Markers of Susceptibility prostate cancer scan[18,19] (all male) and the GWAS of Lung Cancer and Smoking[20] (enriched for smokers), along with an additional 92 non-Hispanic white female controls. For the set 2 scan, cases were individuals with colorectal cancer from both arms of the trial who were not already included in set 1. Samples were excluded if participants did not sign appropriate consent forms, if DNA was unavailable, if baseline questionnaire data with follow-up evaluation were unavailable, if they had a history of colon cancer before the trial, if they had a rare cancer, if they were already in a 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 (2-year blocks), enrollment date (2-year blocks), sex, race/ethnicity, trial arm, and study year of diagnosis (ie, controls must be cancer free into the case's year of diagnosis).

Postmenopausal Hormones Supplementary Study to the CCFR. Eligible case patients included all female residents, ages 50–74 years, residing in the 13 counties in Washington State, reporting to the Cancer Surveillance, Epidemiology and End Results program, who were newly diagnosed with invasive colorectal adenocarcinoma (ICD-O C18.0, C18.2–C18.9, C19.9, C20.0–C20.9) between October 1998 and February 2002.[21] Eligibility for all individuals was limited to those who were English speaking with available telephone numbers, through which they could be contacted. On average, cases were identified within 4 months of diagnosis. The overall response proportion of eligible cases identified was 73%. Community-based controls were selected randomly according to age distribution (in 5-year age intervals) of the eligible cases by using lists of licensed drivers from the Washington State Department of Licensing for individuals, ages 50–64 years, and rosters from the Health Care Financing Administration (now the Centers for Medicare and Medicaid), for individuals older than age 64. The overall response proportion of eligible controls was 66%. In GECCO, samples with sufficient DNA extracted from blood were genotyped. Only participants who were not part of the CCFR Seattle site were included in the sample set.

VITamins And Lifestyle. The VITAL (VITamins And Lifestyle) cohort comprised 77,721 Washington State men and women aged 50–76 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 were self-collected by 70% of the participants. Subjects were followed up for cancer by linkage to the western Washington Surveillance, Epidemiology and End Results cancer registry and were censored when they moved out of the area covered by the registry or at time of death. Details of this study have been described previously.[22] In 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; and any type of lymphoma, including non-Hodgkin, Mantle cell, large B-cell, or follicular lymphoma. Controls were matched on age at enrollment (within 1 year), enrollment date (within 1 year), sex, and race/ethnicity. One control was selected randomly per case among all controls who matched according to the 4 factors described earlier and for whom the control follow-up time was greater than the follow-up time of the case until diagnosis.

Women's Health Initiative. The WHI (Women's Health Initiative) is a long-term health study of 161,808 post-menopausal women aged 50–79 years at 40 clinical centers throughout the United States. WHI comprised a clinical trial arm, an observational study (OS) arm, and several extension studies. The details of WHI have been described previously[23,24] 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 comprised centrally adjudicated colon cancer cases from the OS arm 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 arm and matched to cases on age (within 3 years), enrollment date (within 365 days), hysterectomy status, and prevalent conditions at baseline. For 37 cases, there was no control match in the hip fracture GWAS. For these participants, we identified a matched control in the WHI OS arm based on the same criteria. In the set 2 scan, cases were selected from the August 2009 database and comprised centrally adjudicated colon and colorectal cancer cases from the OS and clinical trial arms 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, institutional review board approval not available for data submission into dbGaP, and insufficient DNA available. Matching criteria included age (within years), race/ethnicity, WHI date (within 3 years), WHI Calcium and Vitamin D study date (within 3 years), and randomization arms (OS flag, hormone therapy assignments, dietary modification assignments, calcium/vitamin D assignments). In addition, they were matched by the 4 regions of randomization centers. Each case was matched with 1 control (1:1) who met the matching criteria exactly. Control selection was performed in a time-forward manner, selecting one control for each case first from the risk set at the time of the case's event. The matching algorithm was allowed to select the closest match based on a criterion to minimize an overall distance measure.[25] Each matching factor was given the same weight. Additional available controls who were genotyped as part of the hip fracture GWAS were included to improve power.

Follow-up Studies

In the following, we describe each study population used in the follow-up study. For information on sample sizes and demographic factors please see Supplementary Table 1.

Asia Colorectal Cancer Consortium. The study protocols were approved by relevant institutional review boards at all study sites, and all included subjects provided informed consent. Sample size, genotype platform, the number of SNPs used in imputation, and genomic inflation factors in each of the 5 studies are presented in Supplementary Table 8.

Shanghai study 1 and 2. Colorectal cancer cases were derived from the Shanghai Women's Health Study[26] and the Shanghai Men's Health Study,[27] both population-based cohort studies that are being conducted in urban Shanghai, China. A total of 777 pathologically diagnosed colorectal cancer cases with DNA available were identified in participants from the Shanghai Women's Health Study and Shanghai Men's Health Study and included in this study. A total of 758 cancer-free controls were derived from the Shanghai Women's Health Study/Shanghai Men's Health Study and frequency-matched to colorectal cancer cases by age and sex. To increase statistical power, we also included 2131 community female controls who were scanned using the Affymetrix Genome-Wide Human SNP Array 6.0 (Affymetrix 6.0) as part of an ongoing GWAS of breast cancer.[28] A total of 481 cases and 2632 controls were genotyped using Affymetrix 6.0 (Shanghai Study 1). A total of 296 cases and 257 controls were genotyped using Illumina HumanOmniExpress BeadChip (Illumina OmniExpress) (Shanghai Study 2).

Guangzhou study 1. This study contributed 694 cases and 972 controls. Histopathologically diagnosed colorectal cancer cases were recruited from the Sun Yat-Sen University Cancer Center between January 2002 and January 2012. Healthy controls were recruited from physical examination centers of several large general hospitals in Guangdong province communities.[29] At enrollment, controls reported no history of any cancer. All cases and controls were self-reported Han Chinese who lived in Guangdong Province at the time of recruitment. Blood samples from all cases and controls were obtained as the source of genomic DNA for the study.

Aichi study 1. This study is part of the Hospital-based Epidemiologic Research Program at Aichi Cancer

Center in Japan.[30] All first-visit outpatients 20–79 years of age at the Aichi Cancer Center during December 2000 to November 2005 were asked to participate in the Hospital-based Epidemiologic Research Program at Aichi Cancer Center. Of 29,736 eligible patients approached, 28,766 participated in the study, with a response rate of 96.7%. All participants completed self-administered questionnaires about their lifestyle and demographic characteristics and provided blood samples. Case status was confirmed via the Hospital-based Epidemiologic Research Program at Aichi Cancer Center database and the hospital-based cancer registry database. A total of 589 colorectal cancer cases were identified in this cohort and 497 were included in the GWAS. A total of 942 controls without any cancer at recruitment were selected randomly and frequency-matched to cases by age and sex.[31]

Korean Cancer Prevention Study-II. The Korean Cancer Prevention Study-II included 266,258 individuals, 20–77 years of age, who visited 16 health promotion centers nationwide from April 2004 to December 2008 in South Korea.[32] Subjects were interviewed at baseline to obtain exposure data. Cancer diagnoses were identified through 2008 using data from the national cancer registry and hospitalization records. For the study, we selected 325 colorectal cancer patients who provided a blood sample. Cancer-free cohort members (N = 977) were selected randomly as controls.

Tennessee Colorectal Polyp Study. The Tennessee Colorectal Polyp Study was a colonoscopy-based, case-control study conducted in Nashville, Tennessee, from 2003 to 2011.[33] Eligible participants, aged 40–75 years old, were identified from patients at the Vanderbilt Gastroenterology Clinic and the Veteran's Affairs Tennessee Valley Health System Nashville Campus. Participants were excluded if they had genetic colorectal cancer syndromes, a prior history of inflammatory bowel disease, prevalent adenomatous polyps, or any cancer other than nonmelanoma skin cancer. Colonoscopic procedures were performed and reported using standard clinical protocols and all pathology diagnoses were determined by hospital pathologists. Participants provided DNA either before or after colonoscopy (blood and buccal samples were collected). The analysis included only participants of Caucasian race.

Genotyping, Quality Assurance/QC, and Imputation

GWAS in GECCO and CCFR. We conducted a meta-analysis of GWAS from 13 studies within the GECCO consortium (10,729 cases and 13,328 controls) and additional GWAS within the CCFR (1967 cases and 1785 controls). The GWAS from CCFR, which consisted of participants from sites in the United States, Canada, and Australia, included a population-based, case-control set (CCFR set 1, 1171 cases and 983 controls) genotyped using Illumina Human1M or Human1M-Duo,[6] and a sibling-pair set (CCFR set 2, 796 cases and 802 controls) genotyped using Illumina Omni1. The GECCO GWAS consisted of participants within The french Association STudy Evaluating RISK for sporadic colorectal cancer; Hawaiian Colorectal Cancer Studies 2 and 3; DACHS [Darmkrebs: Chancen der Verhütung durch Screening]; DALS; HPFS; MEC; NHS; OFCCR; PHS; Postmenopausal Hormone Study; PLCO; VITAL study; and the WHI. Phase one genotyping of a total of 1709 colon cancer cases and 4214 controls from PLCO, WHI, and DALS (PLCO set 1, WHI set 1, and DALS set 1) was performed using Illumina HumanHap 550K, 610K, or combined Illumina 300K and 240K, and has been described previously.[34] A total of 650 colorectal cancer cases and 522 controls from OFCCR were included in GECCO from previous genotyping using Affymetrix platforms.[3] A total of 5540 colorectal cancer cases and 5425 controls from the The french Association STudy Evaluating RISK for sporadic colorectal cancer, Hawaiian Colorectal Cancer Studies 2 and 3, DACHS set 1, DALS set 2, the MEC, Postmenopausal Hormone study, PLCO set 2, VITAL study, and WHI set 2 were genotyped successfully using Illumina HumanCytoSNP. A total of 2004 colorectal cancer cases and 2244 controls from HPFS, NHS, PHS, and DACHS set 2, as well as a total of 826 advanced adenoma cases and 923 controls from HPFS and NHS were genotyped successfully using Illumina HumanOmniExpress.

DNA was extracted from blood samples or, for a subset of DACHS, HPFS, MEC, NHS, and PLCO samples, and for all VITAL samples, from buccal cells, using conventional methods. All studies included 1%–6% blinded duplicates to monitor the quality of the genotyping. All individual-level genotype data were managed and underwent quality assurance and QC at the University of Southern California (CCFR sets 1 and 2), the OFCCR, the University of Washington Genetics Coordinating Center (HPFS, NHS, PHS, and DACHS set 2), or the GECCO Coordinating Center at the Fred Hutchinson Cancer Research Center (all other studies). Details on the quality assurance/QC can be found in Supplementary Table 2. In brief, samples were excluded based on call rate, heterozygosity, unexpected duplicates, gender discrepancy, and unexpectedly high identity-by-descent or unexpected genotype concordance (>65%) with another individual. All analyses were restricted to samples clustering with the Utah residents with northern and western European ancestry from the CEU population in principal component analysis, including the HapMap II populations as reference. SNPs were excluded if they were triallelic, not assigned an rs number, or were reported or observed as not performing consistently across platforms. In addition, genotyped SNPs were excluded based on call rate (<98%), lack of HWE in controls ($P < 1 \times 10^{-4}$), and MAF (<5% in set 1 for PLCO, WHI, DALS, and OFCCR; <5/number of samples for remaining studies).

Because imputation of genotypes is established as standard practice in the analysis of genotype array data, all autosomal SNPs from all studies were imputed to the CEU population in HapMap II release 24, with the exception of OFCCR, which was imputed to HapMap II release 22. CCFR sets 1 and 2 were imputed using IMPUTE (available at: https://mathgen.stats.ox.ac.uk/impute/impute.html),[35] OFCCR was imputed using BEAGLE (available at: http://faculty.washington.edu/browning/beagle/beagle.html),[36] and all other studies were imputed using MACH (available at: www.sph.umich.edu/csg/abecasis/MACH/tour/).[37] Imputed data were merged with genotype data such that genotype data preferentially were selected if a SNP had both types of data, unless there was a difference in terms of reference allele frequency (>0.1) or position (>100 base pairs), in which case imputed data were used. Given the high agreement of imputation accuracy among MACH, IMPUTE, and BEAGLE,[38] the common practice to use different imputation programs is unlikely to cause heterogeneity[39] and it has become common practice to combine results across SNPs imputed using different programs. As a measurement of imputation accuracy we calculated R^2. Analyses of imputed data had different QC cut-off values than those for directly genotyped SNPs discussed earlier and were restricted to SNPs with either a MAF of 1% or greater or an R^2 value greater than 0.3, with the exception of CCFR set 2, which was restricted to SNPs with both a MAF of 1% or greater and an R^2 value of 0.3 or greater. After imputation and QC, a total of 2,708,280 SNPs were used in the meta-analysis of the GECCO and CCFR studies.

Follow-up studies. We selected the 10 most statistically significant regions (excluding known GWAS loci) based on the *P* value from the GECCO and CCFR meta-analyses for follow-up evaluation in colorectal cancer studies in Asian populations and adenoma studies in populations of European descent.

The Asian colorectal cancer follow-up study comprised a meta-analysis of 5 studies conducted in China, Japan, and South Korea, including 2293 colorectal cancer cases and 5780 controls. Cases and controls were genotyped using multiple SNP arrays, including Affymetrix Genome-Wide Human SNP Array 6.0, Affymetrix Genome-Wide Human SNP Array 5.0, Illumina Infinium HumanHap610 BeadChip, Illumina Human610-Quad BeadChip, and Illumina HumanOmniExpress BeadChip. Samples were excluded based on low call rate (<95%), heterozygosity, unexpected duplicates, gender discrepancy, and outlying population substructure. After quality control exclusions, 2098 cases and 5749 controls remained in the analysis. SNPs were excluded for low call rate (<95%), low genotype concordance (<95%) among positive QC samples, an MAF less than 5%, or an HWE *P* value less than 1×10^{-5} in controls. For each of the 5 studies, SNPs were imputed for autosomal SNPs that were present in HapMap Japanese in Tokyo, Japan+Han Chinese individuals from Beijing, China Phase 2 release 22 using MACH.[37] SNPs with an R^2 value greater than 0.5 were included in the analysis.

The colorectal adenoma follow-up study consisted of a US-based GWAS of 1049 cases and 987 controls.[33] DNA extracted from blood and buccal samples were genotyped using the Affymetrix Genome-Wide Human SNP Array 5.0. Samples were excluded based on low call rate (<95%), heterozygosity, unexpected duplicates, gender discrepancy, identity-by-descent, and outlying population substructure. After quality control exclusions, 958 cases and 909 controls remained in the analysis. SNPs were excluded for low call rate (<95%), MAF less than 1%, or HWE *P* value less than 1×10^{-6}. After quality control exclusions, a total of 402,326 SNPs remained in the analysis. Data were imputed to the 1000 Genomes Project and HapMap Phase 3 using IMPUTE.[35] SNPs with an R^2 value greater than 0.5 were included in the analysis.

Details on Functional Annotation Findings Using Bioinformatic Databases

There are several bioinformatic tools available for the post-GWAS functional characterization of putative disease-causing loci through the University of California, Santa Cruz genome browser.[40] Annotation of non–protein-coding regions operates under the hypothesis that trait-associated alleles exert their effects by influencing transcriptional levels through multiple regulatory mechanisms. The University of California, Santa Cruz genome browser provides several tracks that can be used to annotate enhancers, promoters, insulators, and silencers[40] (for details see Supplementary Table 9). Such tools help expedite the discovery of causal variants by isolating a few likely culprits from a large background of variants in linkage disequilibrium with the surrogate marker (tag SNP). Because distal enhancers often facilitate cell-type–specific expression, it is helpful to look for evidence in a variety of cell lines in addition to those related to the trait. For example, the ENCODE (availble at: http://genome.ucsc.edu/ENCODE/) transcription summary track assayed by RNA-sequencing can be displayed as an overlay of histograms denoting expression levels in various tissues marked by a specific color, thus allowing identification of cell-type–specific expression.

Similarly the histone modification tracks can provide additional evidence for cell-specific regulatory elements when displayed in this configuration. The methylation and acetylation of histone proteins changes chromatin accessibility for transcription and such marks can serve as a powerful tool for identifying both enhancer and promoter regions. There are 3 summary ENCODE tracks available to detect specific chemical modifications and were assayed in 7 different tissues using chromatin immunoprecipitation sequencing methodology. The H3K4me1 histone mark is associated with enhancers

downstream of transcription start sites. The H3k27Ac histone mark is similarly thought to enhance transcription and likely does so through the blocking of the repressive histone mark H3K27Me3. The last histone modification in the summary tracks, H3K4Me3, is associated with active promoters. Additional chemical modifications and cell lines are available under the Broad Institute histone modification track for further interrogation.

Regulatory regions are susceptible to DNase cutting and ENCODE has assayed this hypersensitivity in a large collection of cell types. The precision of the DNase cluster track is somewhat better than that of chromatin modifications. Identification of evolutionarily conserved segments, phylogenetic footprints, has been used to discover functionally important regions. However, histone marks and DNase hypersensitivity tracks are more robust tools for characterizing regulatory regions because these elements are not always constrained across vertebrate evolution. Functional hypotheses around regulatory regions can be strengthened with the ENCODE transcription factor track. By using the chromatin immunoprecipitation sequencing method, this track helps to identify the alteration of transcription factor binding sites, which potentially alter expression levels. As an example, CCCTC-binding factor is a transcription factor that assumes multiple forms and can act as an activator, a repressor/silencer, or an insulator. When binding chromatic insulators, it can prevent interactions between promoters and nearby enhancers or silencers. However, it also mediates long-range chromatin looping, which can bring enhancers in proximity of a gene's promoter. Combining the strengths and weaknesses of each of these tracks can provide in silico evidence for regulatory function, and enables selection of strong candidates for additional functional studies using reporter gene methods.

Supplementary References

1. Cotterchio M, Manno M, Klar N, et al. Colorectal screening is associated with reduced colorectal cancer risk: a case-control study within the population-based Ontario Familial Colorectal Cancer Registry. Cancer Causes Control 2005;16:865–875.
2. Cotterchio M, Keown-Eyssen G, Sutherland H, et al. Ontario familial colon cancer registry: methods and first-year response rates. Chronic Dis Can 2000;21:81–86.
3. 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.
4. Küry S, Buecher B, Robiou-du-Pont S, et al. Combinations of cytochrome P450 gene polymorphisms enhancing the risk for sporadic colorectal cancer related to red meat consumption. Cancer Epidemiol Biomarkers Prev 2007;16:1460–1467.
5. Newcomb PA, Baron J, Cotterchio M, et al. Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. Cancer Epidemiol Biomarkers Prev 2007;16:2331–2343.
6. Figueiredo JC, Lewinger JP, Song C, et al. Genotype-environment interactions in microsatellite stable/microsatellite instability-low colorectal cancer: results from a genome-wide association study. Cancer Epidemiol Biomarkers Prev 2011;20:758–766.
7. Brenner H, Chang-Claude J, Seiler CM, et al. Protection from colorectal cancer after colonoscopy: population-based case-control study. Ann Intern Med 2011;154:22–30.
8. 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.
9. Slattery ML, Potter J, Caan B, et al. Energy balance and colon cancer–beyond physical activity. Cancer Res 1997;57:75–80.
10. 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.
11. Rimm EB, Stampfer MJ, Colditz GA, et al. Validity of self-reported waist and hip circumferences in men and women. Epidemiology 1990;1:466–473.
12. 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.
13. Belanger CF, Hennekens CH, Rosner B, et al. The Nurses' Health Study. Am J Nurs 1978;78:1039–1040.
14. Hennekens CH, Eberlein K. A randomized trial of aspirin and beta-carotene among U.S. physicians. Prev Med 1985;14:165–168.
15. Christen WG, Gaziano JM, Hennekens CH. Design of Physicians' Health Study II–a randomized trial of beta-carotene, vitamins E and C, and multivitamins, in prevention of cancer, cardiovascular disease, and eye disease, and review of results of completed trials. Ann Epidemiol 2000;10:125–134.
16. Prorok PC, Andriole GL, Bresalier RS, et al. Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial. Control Clin Trials 2000;21:273S–309S.
17. Gohagan JK, Prorok PC, Hayes RB, et al. The Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial of the National Cancer Institute: history, organization, and status. Control Clin Trials 2000;21:251S–272S.
18. National Cancer Institute. Cancer Genetic Markers of Susceptibility (CGEMS) data website. Available at: http://cgems.cancer.gov/data_access.html. CGEMS Data Accessed October 5, 2009.
19. 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.
20. 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.
21. Newcomb PA, Zheng Y, Chia VM, et al. Estrogen plus progestin use, microsatellite instability, and the risk of colorectal cancer in women. Cancer Res 2007;67:7534–7539.
22. 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.
23. Hays J, Hunt JR, Hubbell FA, et al. The Women's Health Initiative recruitment methods and results. Ann Epidemiol 2003;13:S18–S77.
24. The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. Control Clin Trials 1998;19:61–109.
25. Bergstralh EJ, Kosanke JL. Computerized matching of cases to controls. 56th ed. Rochester MN: Department of Health Sciences Research, Mayo Clinic, 1995.
26. Zheng W, Chow WH, Yang G, et al. The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. Am J Epidemiol 2005;162:1123–1131.
27. Cai H, Zheng W, Xiang YB, et al. Dietary patterns and their correlates among middle-aged and elderly Chinese men: a report

from the Shanghai Men's Health Study. Br J Nutr 2007;98: 1006–1013.
28. Zheng W, Long J, Gao YT, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 2009;41:324–328.
29. Bei JX, Li Y, Jia WH, et al. A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. Nat Genet 2010;42:599–603.
30. Matsuo K, Suzuki T, Ito H, et al. Association between an 8q24 locus and the risk of colorectal cancer in Japanese. BMC Cancer 2009;9:379.
31. Nakata I, Yamashiro K, Yamada R, et al. Association between the SERPING1 gene and age-related macular degeneration and polypoidal choroidal vasculopathy in Japanese. PLoS One 2011; 6:e19108.
32. Jee SH, Sull JW, Lee JE, et al. Adiponectin concentrations: a genome-wide association study. Am J Hum Genet 2010;87:545–552.
33. Edwards TL, Shrubsole MJ, Cai Q, et al. A study of prostaglandin pathway genes and interactions with current nonsteroidal anti-inflammatory drug use in colorectal adenoma. Cancer Prev Res (Phila) 2012;5:855–863.
34. **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.
35. Marchini J, Howie B, Myers S, et al. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007;39:906–913.
36. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;81:1084–1097.
37. Li Y, Willer CJ, Ding J, et al. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 2010;34:816–834.
38. Nothnagel M, Ellinghaus D, Schreiber S, et al. A comprehensive evaluation of SNP genotype imputation. Hum Genet 2009;125: 163–171.
39. Gogele M, Minelli C, Thakkinstian A, et al. Methods for meta-analyses of genome-wide association studies: critical assessment of empirical evidence. Am J Epidemiol 2012;175:739–749.
40. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Res 2002;12:996–1006.
41. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 2010;26:2336–2337.

Author names in bold designate shared co-first authorship.

Supplementary Table 1. Descriptive Characteristics of Study Populations

Study name	Other name	Design	Country	Cases	Controls	Age range, *y*	Mean age, *y*	Female, %	Covariates used in analysis
GWAS									
Association STudy Evaluating RISK for sporadic colorectal cancer	ASTERISK	Case-control	France	948	947	40–99	65.3	41.3	Age, sex, 3 PCAs, batch
Colorectal Cancer Studies 2&3	Hawaiian Colo2&3	Case-control	United States	87	125	38–86	65.2	44.8	Age, sex, 3 PCAs
Colon Cancer Family Registry[a]	CCFR	Case-control and sib-pair	United States, Canada, Australia	1967	1785	19–88	55.5	51.8	Age, sex, 3 PCAs, center
Darmkrebs: Chancen der Verhütung durch Screening	DACHS	Case-control	Germany	2376	2206	33–98	68.7	39.9	Age, sex, PCAs
Diet, Activity and Lifestyle Study	DALS	Case-control	United States	1116	1174	30–79	65.2	44.9	Age, sex, 3 PCAs, center
Health Professionals Follow-up Study	HPFS	Cohort	United States	403	402	48–83	65.2	0	age, 3PCAs
Multiethnic Cohort Study	MEC	Cohort	United States	328	346	45–76	63.0	46.4	Age, sex, 3 PCAs
Nurses' Health Study	NHS	Cohort	United States	553	955	44–69	59.8	100	Age, 3 PCAs
Ontario Familial Colorectal Cancer Registry	OFCCR	Case-control	Canada	650	522	31–79	64.1	52.0	Age, sex, 3 PCAs
Physicians' Health Study	PHS	Cohort	United States	382	389	40–84	58.4	0	Age, 3 PCAs, smoking
Postmenopausal Hormone study	PMH	Case-control	United States	280	122	50–75	64.8	100	Age, 3 PCAs
Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial	PLCO	Cohort	United States	1019	2391	55–75	64.0	30.8	Age, sex, 3 PCAs, center
VITamins And Lifestyle	VITAL	Cohort	United States	285	288	50–76	66.5	47.6	Age, sex, 3 PCAs
Women's Health Initiative	WHI	Cohort	United States	1476	2538	50–79	67.4	100	Age, 3 PCAs, region
Health Professionals' Follow-up Study, Adenoma Set	HPFS Ad	Cohort	United States	313	345	48–81	60.7	0	Age, 3 PCAs
Nurses' Health Study, Adenoma Set	NHS Ad	Cohort	United States	513	578	44–69	57.0	100	Age, 3 PCAs
Follow-up studies									
Asian Consortium, Colorectal Cancer				2098	5749				
Shanghai-1	Shanghai-1	Cohort	China	474	2628	25–75	53.22	91.62	Age, sex
Shanghai-2	Shanghai-2	Cohort	China	254	231	40–75	60.96	55.67	Age, sex
Guangzhou	Guangzhou	Case-control	China	641	972	14–85	50.36	30.81	Age, sex
Aichi	Aichi-1	Case-control	Japan	404	942	20–79	51.34	44.65	Age, sex
Korean Cancer Prevention Study-II	KCPS-II	Cohort	Korea	325	976	20–88	43.79	39.28	Age, sex
Tennessee Colorectal Polyp Study	TCPS	Case-control	United States	958	909	40–76	58.72	26.65	Age, sex

PCA, principal component analysis.

[a]CCFR is a collaborating study with GECCO. The analysis of set 2 data did not adjust for PCs because of the sibling-pair study design.

Supplementary Table 2. Details on Genotyping Platform and Quality Assurance and Quality Control

Study	Genotyping platform[a]	Duplicate concordance, %	Mean sample call rate, %	SNP exclusions,[b] n	SNPs passing QC, n	Mean SNP call rate, %	Number of imputed SNPs by R^2		
							<0.3	0.3–0.8	>0.8
ASTERISK	300K	100	99.97	30,446	252,176	99.95	76,043	443,302	1,856,490
Colo2&3	300K	100	99.95	40,390	258,978	99.96	71,487	445,613	1,854,778
DACHS Set 1	300K	99.9	99.93	33,588	255,208	99.90	70,989	434,295	1,869,458
DACHS Set 2	730K	100	99.84	32,159	609,115	99.85	18,551	154,813	1,865,294
DALS Set 1	550K, 610K	>97[c]	99.69	34,644	516,631	99.82	20,173	180,322	1,912,832
DALS Set 2	300K	100	99.94	32,885	250,320	99.94	69,289	438,282	1,867,371
HPFS Set 1	730K	99.90	99.93	32,953	612,091	99.93	18,257	150,880	1,857,252
HPFS Set 2	730K	99.9	99.83	51,725	590,132	99.84	20,040	160,464	1,861,553
HPFS Ad	730K	100	99.86	61,201	597,470	99.86	18,610	155,527	1,861,220
MEC	300K	100	99.97	34,494	259,364	99.96	68,634	433,560	1,868,693
NHS Set 1	730K	100	99.93	47,295	628,541	99.93	17,142	147,723	1,855,814
NHS Set 2	730K	100	99.81	53,328	594,015	99.81	19,434	160,804	1,875,767
NHS Ad	730K	100	99.81	35,954	614,357	99.81	17,901	152,373	1,863,872
PHS Sets 1+2	730K	100	99.90	32,088	594,205	99.90	19,387	157,993	1,864,677
PLCO Set 1	300/240S and 610K	>97[c]	99.65	33,342	503,351	99.85	20,855	184,854	1,921,986
PLCO Set 2	300K	99.90	99.80	38,655	253,702	99.90	68,059	434,769	1,870,311
PMH	300K	99.90	99.89	39,275	256,743	99.92	67,818	429,887	1,875,260
VITAL	300K	99.90	99.81	36,805	243,625	99.89	73,966	461,036	1,845,318
WHI set 1	550K, 550Kduo, 610K	>97[c]	99.60	40,276	511,251	99.77	21,655	184,833	1,914,909
WHI set 2	300K	100	99.96	27,392	251,707	99.96	72,272	442,111	1,864,141

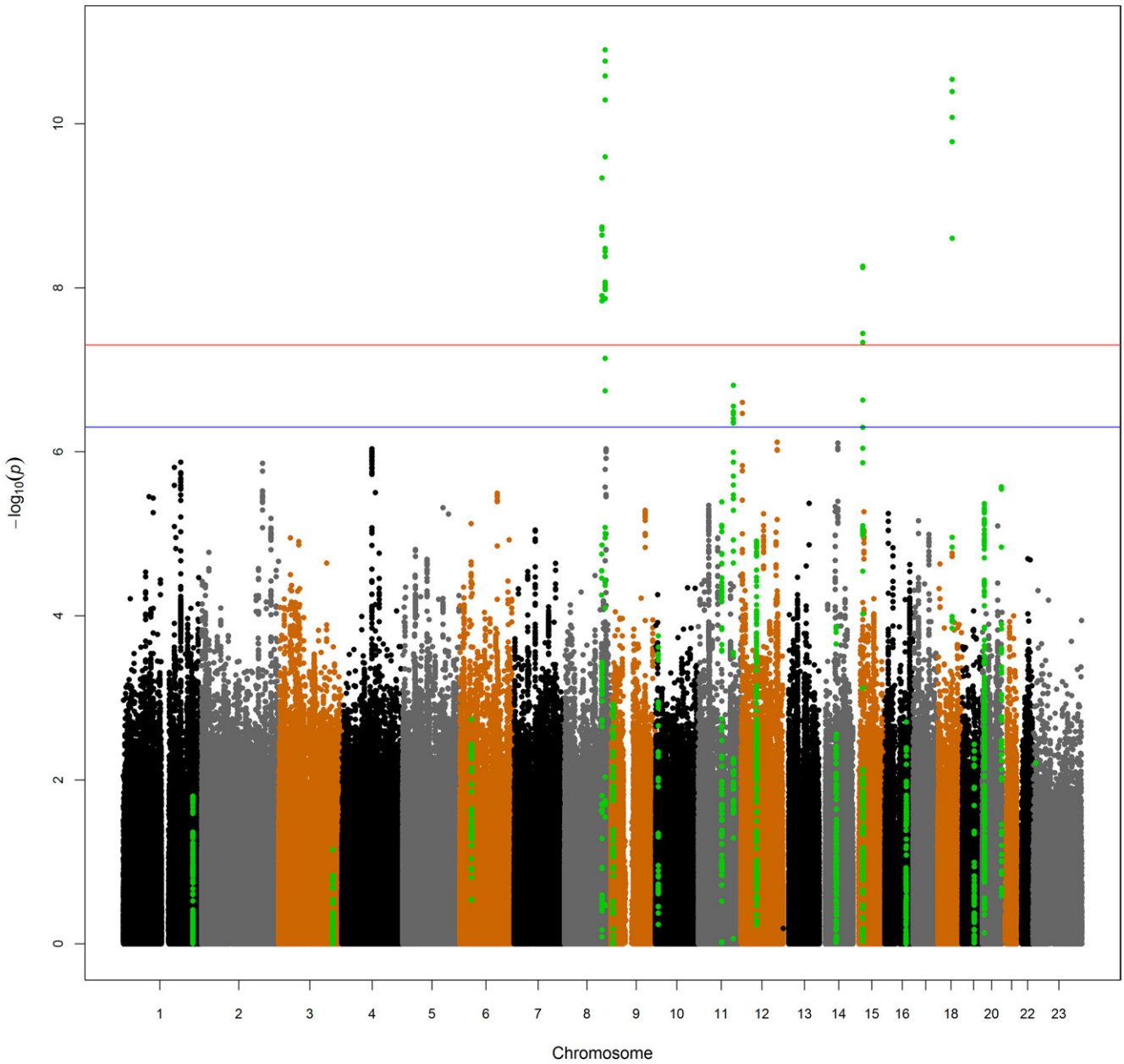
NOTE. CCFR and OFCCR had quality assurance/QC performed separately by OFCCR and CCFR investigators as documented by Zanke et al[3] and Figueiredo et al.[6]

ASTERISK, The french Association STudy Evaluating RISK for sporadic colorectal cancer; Colo2&3, Hawaiian Colorectal Cancer Studies 2 and 3; DACHS, Darmkrebs: Chancen der Verhütung durch Screening; MEC, Multiethnic cohort; PMH, Postmenopausal Hormone study; VITAL, VITamins And Lifestyle.

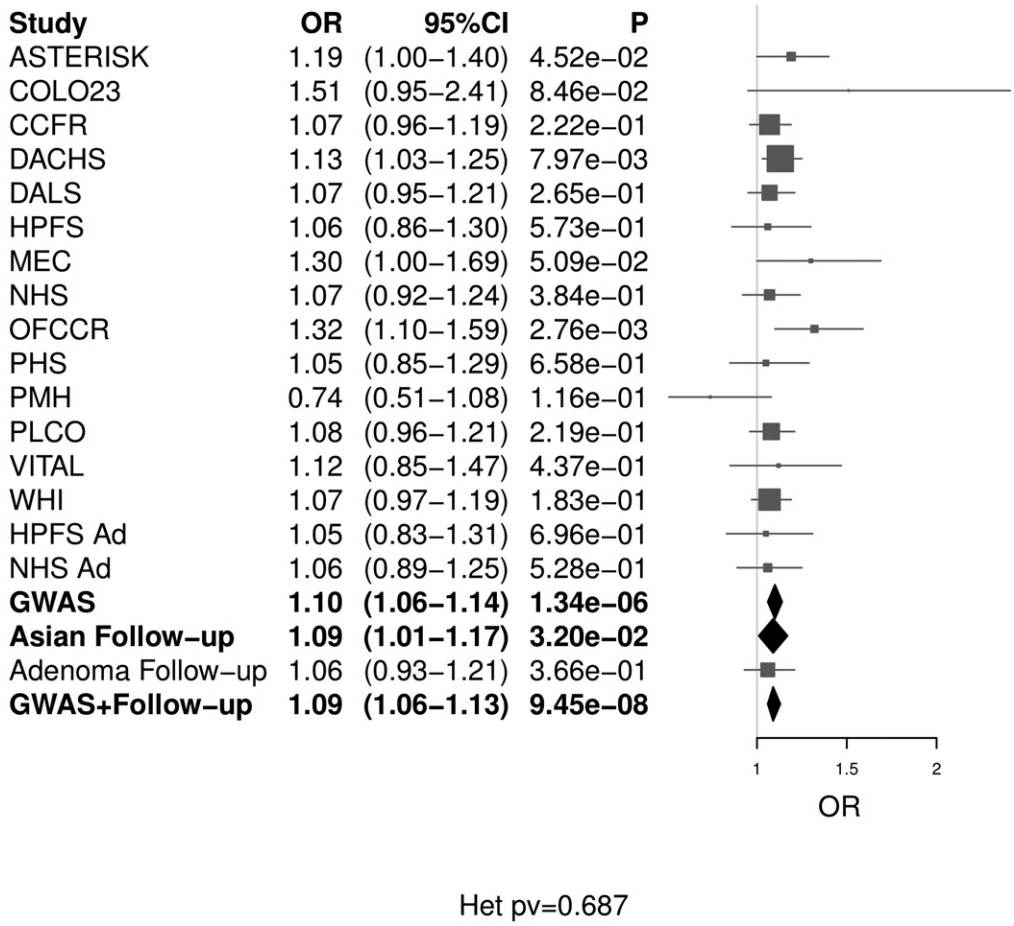
[a]All platforms were Illumina assays, except for OFCCR, which was genotyped using Affymetrix platforms.

[b]Directly genotyped SNPs were excluded for a call rate less than 98%, HWE less than 1×10^{-4}, MAF less than 5 for WHI set 1, PLCO set 1, DALS set 1, and OFCCR set 1; MAF less than 5 per number of samples for remaining studies, and if SNPs reportedly did not perform consistently across platforms.

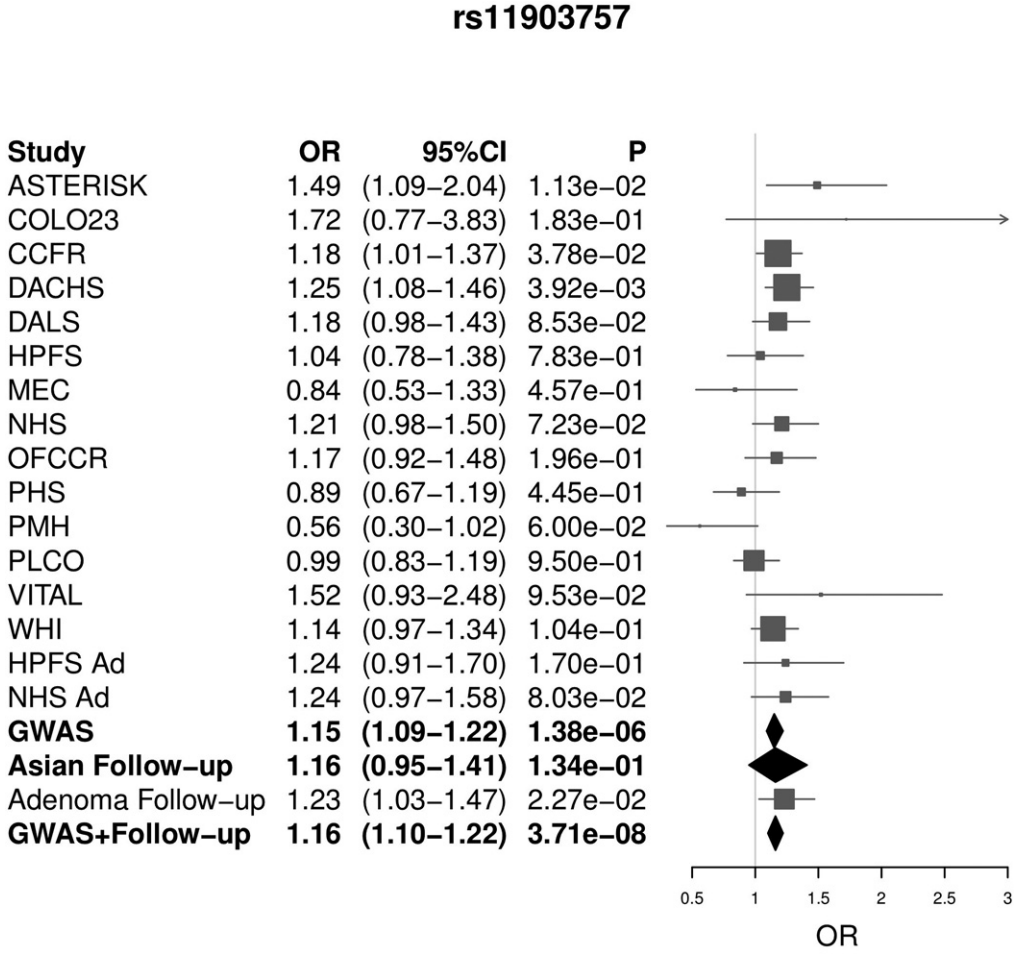
[c]Blinded duplicates were assessed across DALS set 1, PLCO set 1, and WHI set 1; exact concordance was not recorded, but all 98 pairs were identified as having concordance greater than 97%.



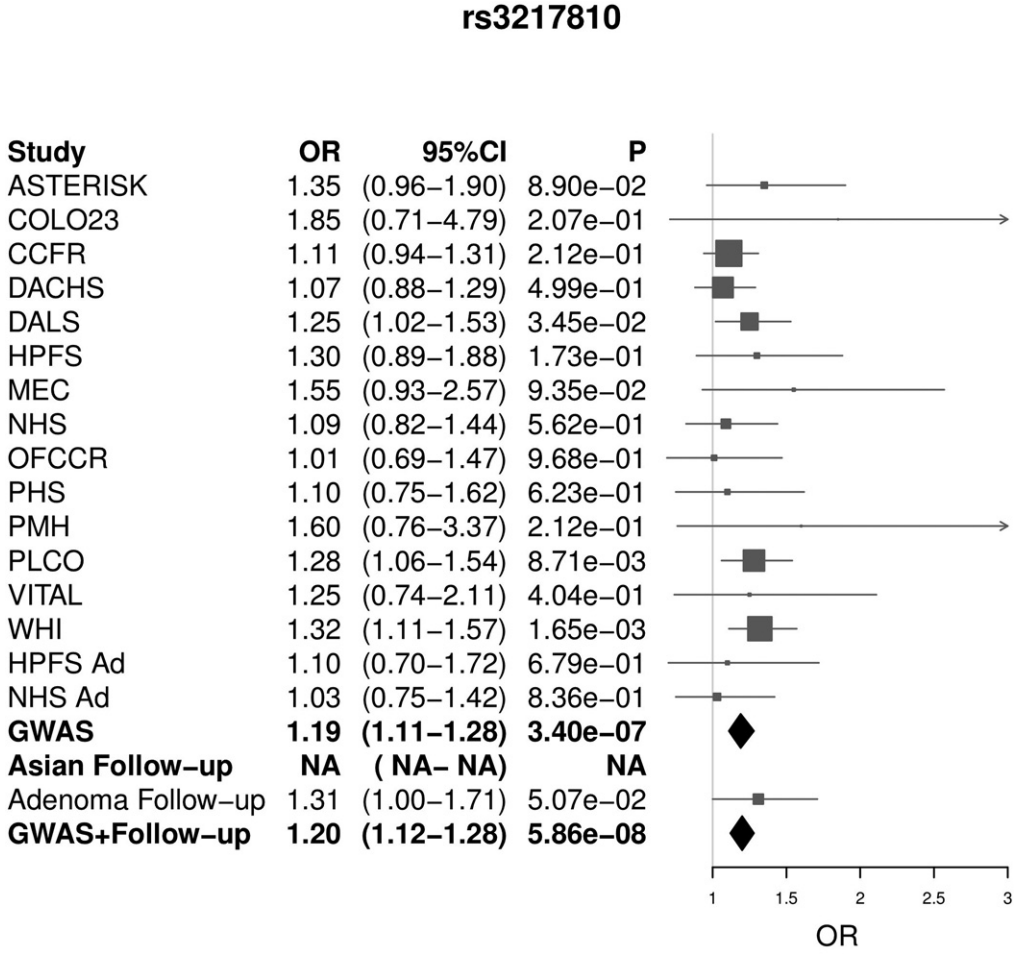
Supplementary Figure 1. Manhattan plot of the GWAS inverse-variance–weighted, fixed-effects meta-analysis, comprising 12,696 cases and 15,113 controls. The $-\log_{10}$ of *P* values for 2,708,280 SNPs plotted against physical chromosomal positions. SNPs above the *blue line* represent those with a *P* value less than 5×10^{-7} whereas SNPs above the *red line* represent those with a *P* value less than 5×10^{-8}. The *green dots* represent previously identified loci as listed in Supplementary Table 4. Chromosome 23 is the X-chromosome. Because we do not have linkage disequilibrium (LD) information for SNPs on the X chromosome, we only show the result of the GWAS SNP on the X chromosome but not SNPs correlated with this GWAS SNP.



Supplementary Figure 2. Forest plot for meta-analysis results for all new findings with a *P* value less than 5×10^{-7} in a combined analysis of GWAS and follow-up studies as listed in Table 1. ORs and 95% confidence intervals (95% CIs) are presented for each additional copy of the minor allele in the multiplicative model. The *grey boxes* are proportional in size to the inverse of the variance for each study, and the *lines* visually depict the confidence interval. Results from the fixed-effects meta-analysis are shown as *diamonds*. The width of the diamond represents the confidence interval.

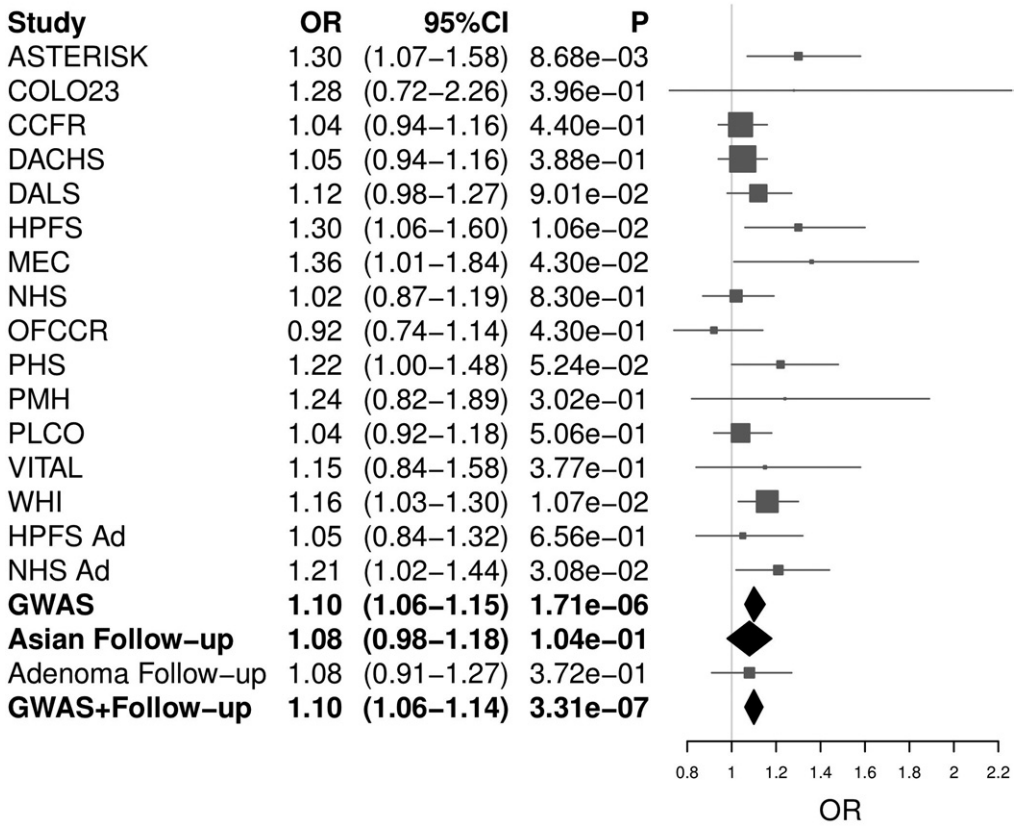


Supplementary Figure 2. Continued



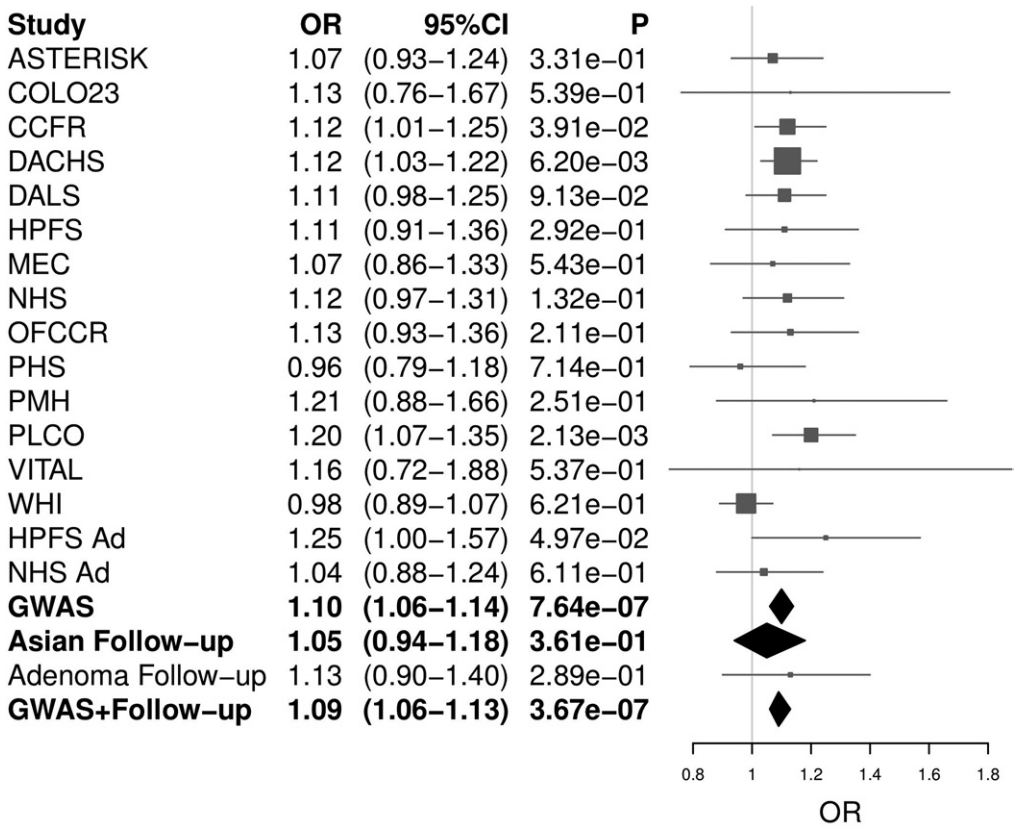
Supplementary Figure 2. Continued

rs3217901

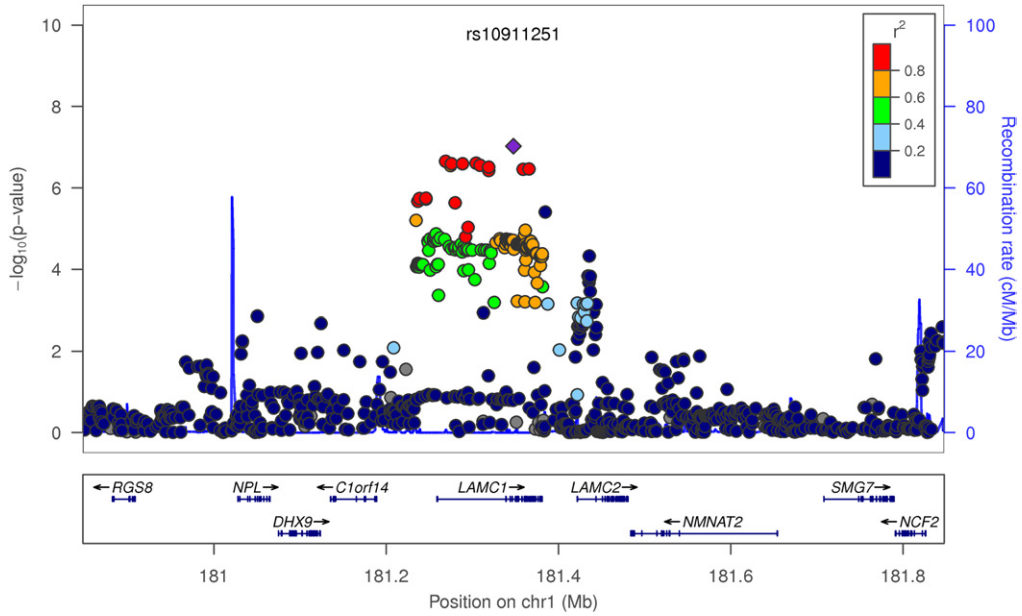


Het pv=0.507

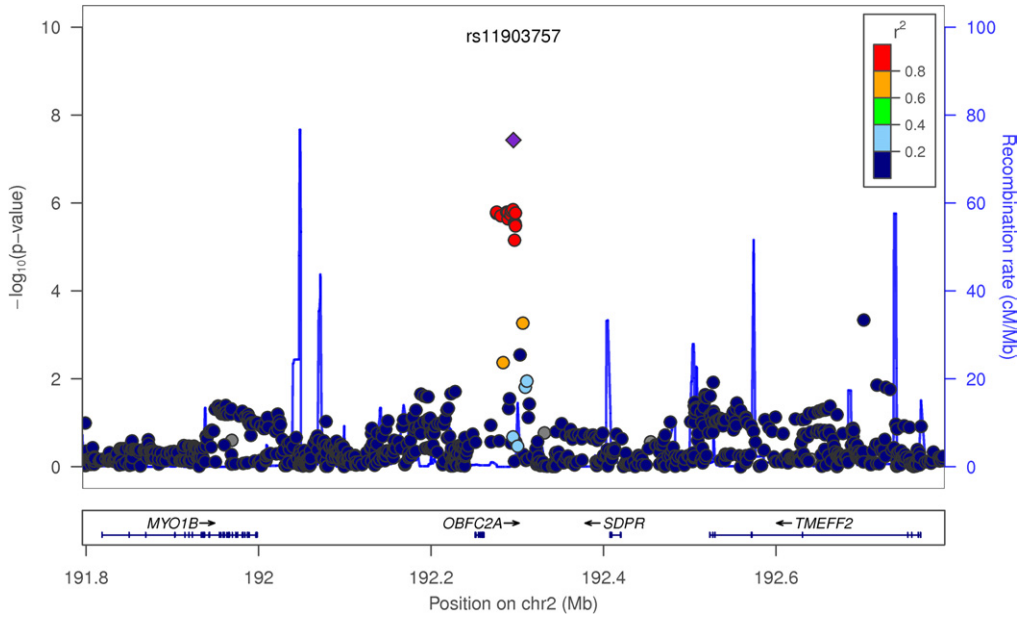
Supplementary Figure 2. Continued



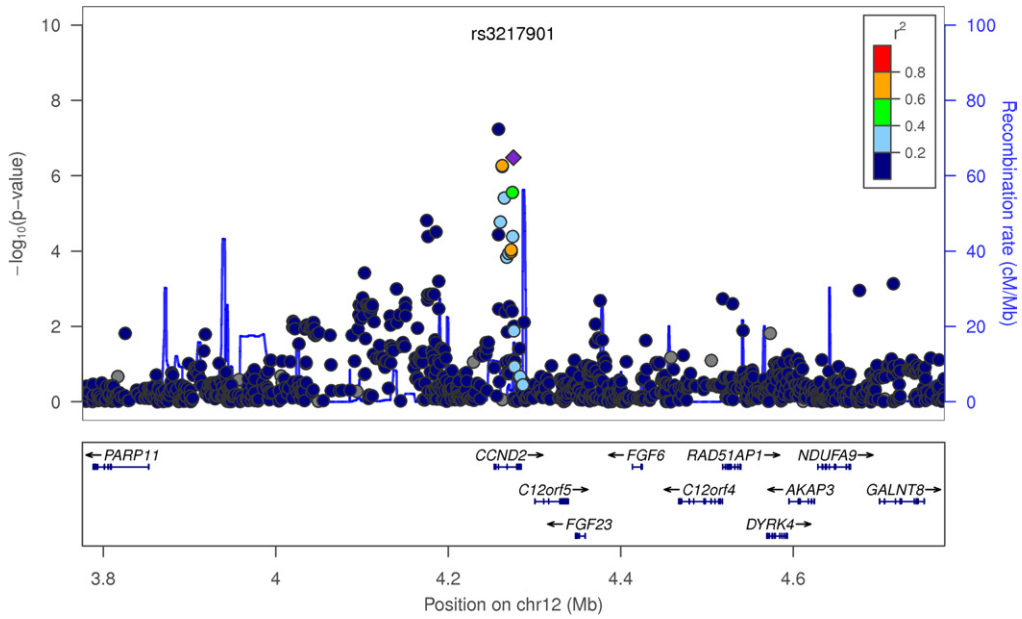
Supplementary Figure 2. Continued



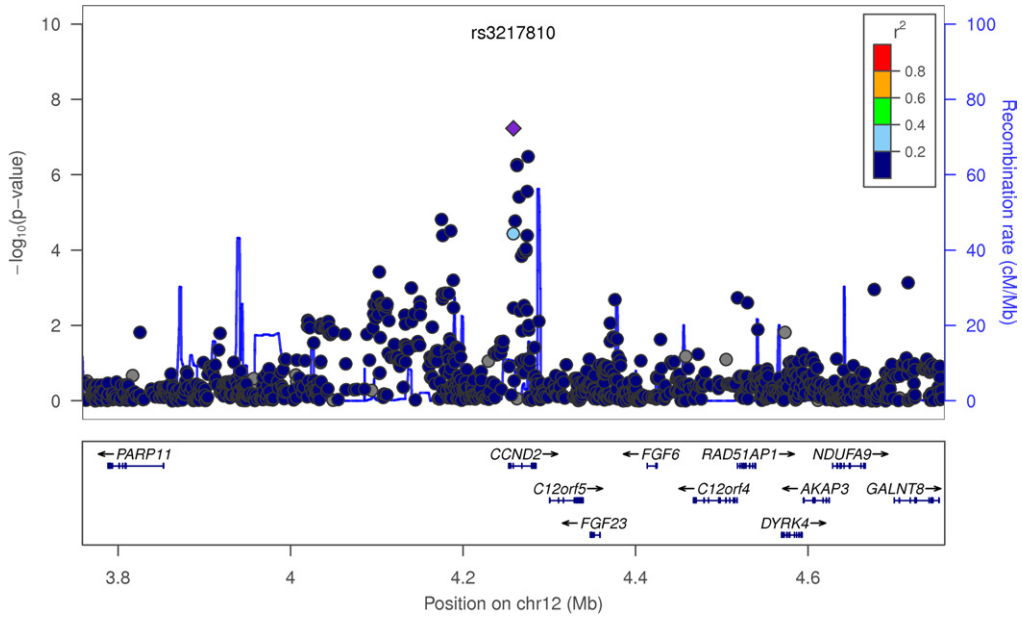
Supplementary Figure 3. Regional association results for all new findings with a *P* value less than 5×10^{-7}, as listed in Table 1. The *top half* of the figure shows the physical position of the SNP on the chromosome along the x-axis, and the -log10 of the meta-analysis *P* value on the y-axis. Each *dot* on the plot represents the *P* value of the association for one SNP with colorectal cancer (allele test) across all studies. The most significant SNP in the region (index SNP) is marked as a *purple diamond*. The color scheme represents the pairwise correlation (r^2) for the SNPs across the region with the index SNP. Correlation was calculated using the HapMap CEU data. The *bottom half* of the figure shows the position of the genes across the region. These regional association plots are also known as LocusZoom plots.[41]



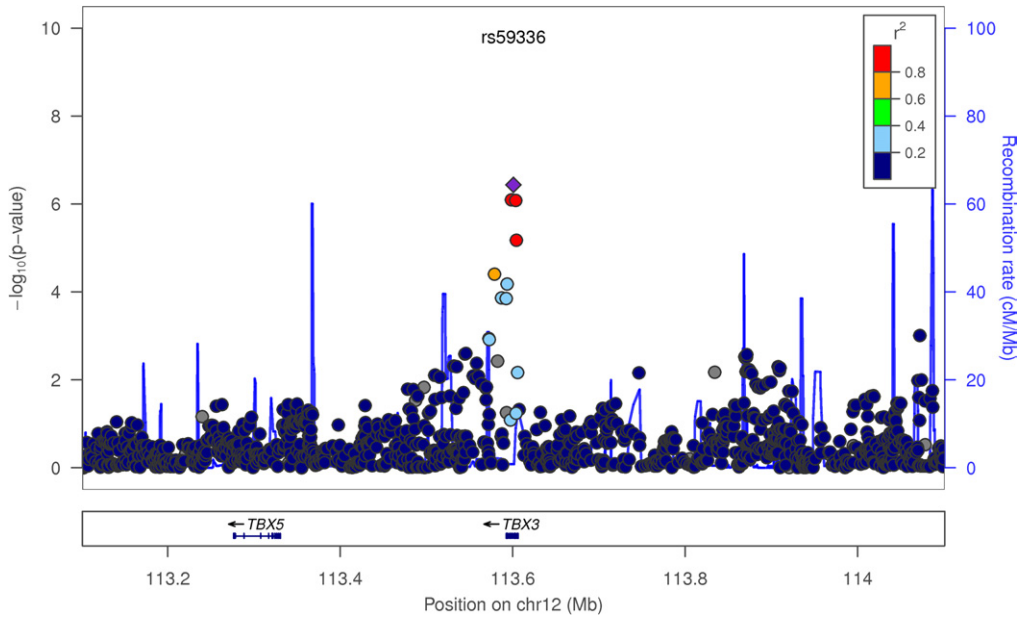
Supplementary Figure 3. Continued



Supplementary Figure 3. Continued



Supplementary Figure 3. Continued



Supplementary Figure 3. Continued

Supplementary Table 5. Risk Estimates for the 2 Top SNPs in 12p13.32/*CCND2* When Both Were Included Simultaneously in the Logistic Regression Analysis

SNP	OR (95% CI)	*P* value
Each SNP analyzed separately		
rs3217901	1.10 (1.06–1.15)	1.71E-06
rs3217810	1.19 (1.11–1.28)	3.40E-07
Both SNPs included simultaneously in the logistic regression analysis		
rs3217901	1.08 (1.03–1.13)	.0008
rs3217810	1.14 (1.06–1.23)	.004

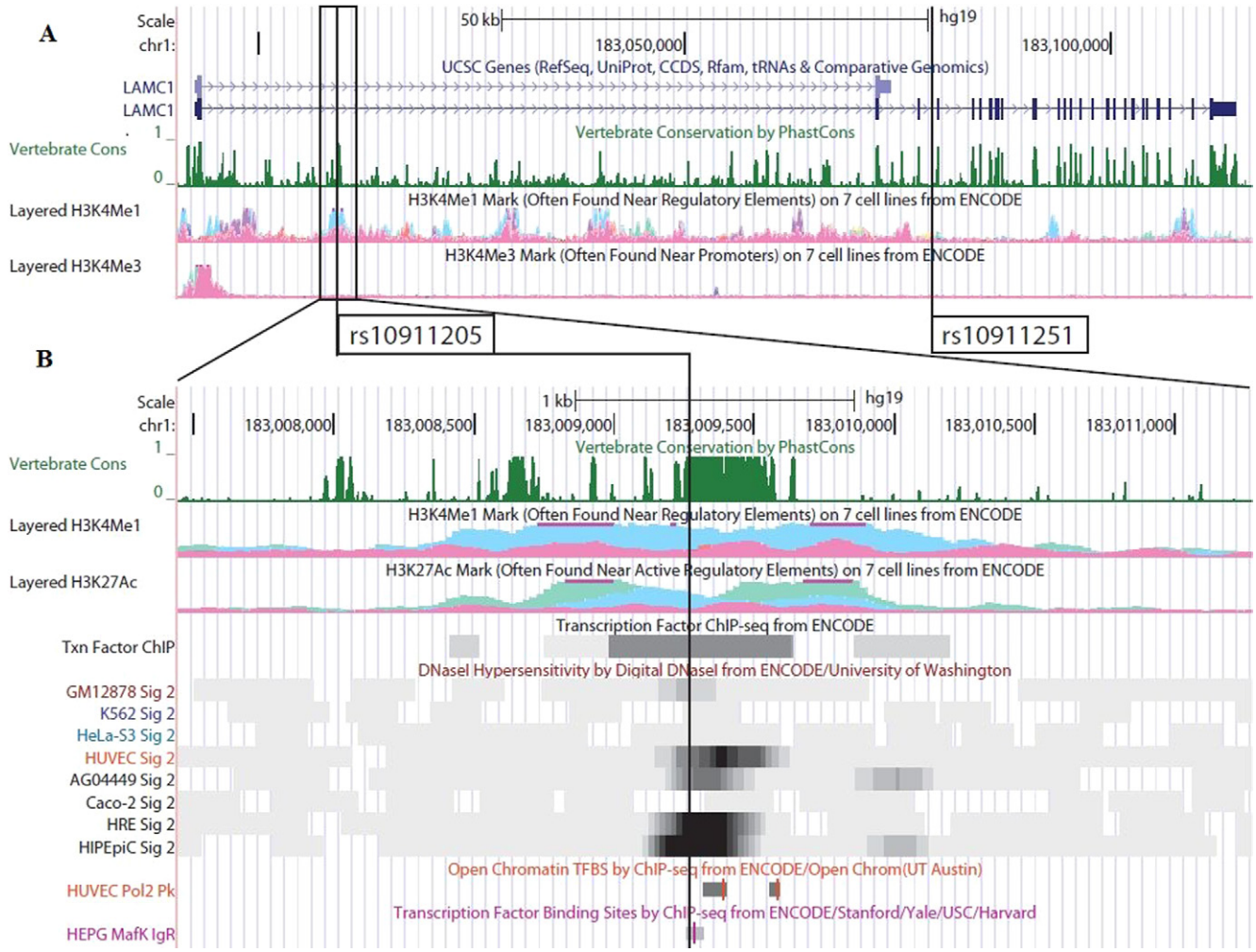
NOTE. Analysis was based on the log-additive model in GWAS of GECCO and CCFR only (12,696 cases and 15,113 controls).
CI, confidence interval.

Supplementary Table 7. Risk Estimates for New Findings With $P < 5 \times 10^{-7}$ Stratified by Colorectal Adenoma and Colorectal Cancer (Log-Additive Model)

SNP	Chromosome (gene)	Cancer/adenoma[a]	OR (95% CI)	*P* value	*P* heterogeneity
SNP with $P < 5 \times 10^{-8}$					
rs11903757	2q32.3	Cancer	1.15 (1.08–1.21)	4.06E-06	.18
	(NABP1)	Adenoma	1.24 (1.08–1.41)	1.46E-03	1.00
		Overall	1.16 (1.10–1.22)	3.71E-08	.27
SNPs with $P < 5 \times 10^{-7}$ and $P > 5 \times 10^{-8}$					
rs10911251	1q25.3	Cancer	1.10 (1.06–1.14)	1.41E-07	.55
	(LAMC1)	Adenoma	1.06 (0.96–1.16)	2.44E-01	.99
		Overall	1.09 (1.06–1.13)	9.45E-08	.69
rs3217810	12p13.32	Cancer	1.20 (1.12–1.29)	2.34E-07	.87
	(CCND2)	Adenoma	1.17 (0.97–1.41)	9.76E-02	.52
		Overall	1.20 (1.12–1.28)	5.86E-08	.91
rs3217901	12p13.32	Cancer	1.10 (1.05–1.14)	2.98E-06	.41
	(CCND2)	Adenoma	1.12 (1.01–1.25)	3.64E-02	.53
		Overall	1.10 (1.06–1.14)	3.31E-07	.51
rs59336	12q24.21	Cancer	1.09 (1.05–1.13)	2.21E-06	.31
	(TBX3)	Adenoma	1.12 (1.00–1.25)	5.73E-02	.44
		Overall	1.09 (1.06–1.13)	3.67E-07	.39

CI, confidence interval.

[a]Cancer (n = 13,968 cases and 19,939 controls, except for rs3217810, which has 11,870 cases and 14,190 controls); adenoma (1784 cases and 1832 controls); overall (15,752 cases and 21,771 controls, except for rs3217810, which has 13,654 cases and 16,022 controls).



Supplementary Figure 4. ENCODE integrate regulation tracks for LAMC1. (*A*) University of California, Santa Cruz (UCSC) genome browser position chr1:182,990,493–183,116,512 (build 37) containing the *LAMC1* protein coding gene. The University of California, Santa Cruz gene track shows 2 variant transcripts for *LAMC1*. Directly beneath the gene track is a histogram of multiple alignments of 46 vertebrate species indicating that there are multiple conserved elements in the gene, primarily concentrated near the 5′ and 3′ regulatory regions. Conservation can help unmask candidate variants that disrupt regulatory regions from other benign associations. The next 2 tracks are transparent overlays from 7 cell lines assayed by the ENCODE project showing the H3K4me1 mark and the H3K4me3 mark associated with active regulatory regions. Peaks in H3K4me3 mark are consistent with the promoter region of *LAMC1*, whereas H3K4Me1 indicates additional enhancer regions in the first intron. The histone marks and pattern of transcription show coordinated, cell-type–specific activity increases in K562 (*blue*) and NHLF (*pink*) cells. (*B*) Focusing on the region containing rs10911205 (chr1:183,007,443–183,011,275), we find that the SNP lies within a strong evolutionarily constrained region. Below this track, evidence from the H3K4me1 and H3K27Ac marks are consistent with rs10911205 falling within a region of coordinated, cell-type–specific activity, most active in K562 (*blue*) cells and human skeletal muscle myoblasts (*green*) cells. The DNase and transcription factor chromatin immunoprecipitation sequencing (ChIP-seq) clusters shown in the last 2 tracks summarize data from a much wider range of cell lines and further supports tissue-specific accessibility for regulatory elements in the region surrounding rs10911205. The r^2 value of rs10911205 with rs10911251 was 0.862. Taken together, evidence provided by the ENCODE integrated regulation tracks is consistent with rs10911205 being a strong functional candidate SNP for the strong rs10911251 association with colorectal cancer.

Supplementary Table 8. Sample Size and Genotyping Methods Used in Asian GWAS

Study	Genotyped		After quality control		Genotyping platform		Number of SNPs[a]	Inflation factor (λ)[b]
	Cases	Controls	Cases	Controls	Cases	Controls		
Shanghai-1	481	2632	474	2628	Affymetrix 6.0	Affymetrix 6.0	502,145	1.03
Shanghai-2	296	257	254	231	Illumina OmniExpress	Illumina OmniExpress	515,701	1.03
Guangzhou-1	694	972	641	972	Illumina OmniExpress	Illumina Human610-Quad	250,612	1.02
Aichi-1	497	942	404	942	Illumina OmniExpress	Illumina HumanHap610	232,426	1.04
KCPS-II	325	977	325	976	Affymetrix 5.0	Affymetrix 5.0	312,869	1.02
Overall	2293	5780	2098	5749				1.01

NOTE. Number of cases and controls differ from Supplementary Table 1 due to quality assurance/quality control exclusions.

[a]Number of SNPs in autosome used for imputation in GWAS.

[b]Genomic inflation factor (lambda) derived from 1,636,780 imputed SNPs with MAF >0.05 and high imputation quality (RSQR >0.50), adjusted with age, sex, and the first 10 principal components.

Supplementary Table 9. Tools for Functional Annotation of Noncoding Variants

UCSC genome browser	Genomic class	Description	Functional evidence
ENCODE transcription	Transcribed region	Transcription levels in 7 cell lines from ENCODE Assayed by high-throughput sequencing of polyadenylated RNA	Variable expression in different tissues provides evidence for cell-type–specific regulation when displayed as transparent overlay of each cell line
ENCODE layered H3K4Me1	Nonpromoter regulatory elements	Uses ChIP-seq method to identify regions of DNA that interact with the mono-methylation of lysine 4 of the H3 histone protein in 7 different cell lines Actual enhancer is likely a small portion of the broad region marked	Methylation of histone proteins changes chromatin accessibility for transcription H3K4Me1 is associated with enhancers downstream of the transcription start site
ENCODE layered H3K4Me3	Promoter regulatory element	Uses ChIP-seq method to identify regions of DNA that interact with the trimethylation of lysine 4 of the H3 histone protein in 7 different cell lines Actual regulatory element is likely a small portion of the broad region marked	H3K4Me3 is associated with promoters that are active or accessible for activation
ENCODE layered H3K27Ac	Nonpromoter regulatory elements	Uses ChIP-seq method to identify regions of DNA that interact with the acetylation of lysine 27 of the H3 histone protein in 7 different cell lines Actual regulatory element is likely a small portion of the broad region marked	H3K27Ac enhances transcription, possibly by blocking the spread of the repressive histone mark H3K27Me3 This mark often is found near active regulatory elements
ENCODE DNase clusters	Regulatory element	Measures digital DNaseI hypersensitivity clusters in a large collection of cell types from ENCODE Greater precision than histone modifications	Regulatory regions and promoters are susceptible DNase cutting Hypersensitivity is used to map chromatin accessibility
ENCODE Txn factor ChIP	Regulatory element	Transcription Factor ChIP-seq from ENCODE is assayed by chromatin immunoprecipitation using antibodies for specific transcription factors and sequencing the precipitated DNA	Marks regions where transcription factors bind DNA and exert specific functions Activators can recruit RNA polymerase, repressors suppress recruitment, and insulators block the activity of nearby activators or repressors
ENCODE UW CTCF binding (within the ENCODE transcription factor binding tracks)	Insulated element	CTCF binding sites are assayed by chromatin immunoprecipitation using antibodies for CTCF and sequencing the precipitated DNA	CTCF can function as a transcriptional activator, a repressor/silencer, or an insulator Binds chromatin insulators to prevent interaction between promoter and nearby enhancers or silencers Also mediates long-range chromatin looping, which can bring enhancers in proximity of a gene's promoter
Vertebrate multi-alignment and conservation (phastCons)	Conserved element	Multiple alignments of 46 vertebrate species Estimates the probability that each nucleotide belongs to a conserved element	Identification of evolutionarily conserved segments of homology, potentially identifying a functionally important region

ChIP-seq, chromatin immunoprecipitation sequencing; CTCF, CCCTC-binding factor; UCSC, University of California, Santa Cruz.