ARTICLE

Genome-wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk

Minta Thomas,[1] Lori C. Sakoda,[1,2] Michael Hoffmeister,[3] Elisabeth A. Rosenthal,[4] Jeffrey K. Lee,[2] Franzel J.B. van Duijnhoven,[5] Elizabeth A. Platz,[6] Anna H. Wu,[7] Christopher H. Dampier,[8] Albert de la Chapelle,[9] Alicja Wolk,[10] Amit D. Joshi,[11,12] Andrea Burnett-Hartman,[13] Andrea Gsur,[14] Annika Lindblom,[15,16] Antoni Castells,[17] Aung Ko Win,[18] Bahram Namjou,[19,20,21] Bethany Van Guelpen,[22,23] Catherine M. Tangen,[24] Qianchuan He,[1] Christopher I. Li,[1] Clemens Schafmayer,[25] Corinne E. Joshu,[6] Cornelia M. Ulrich,[26] D. Timothy Bishop,[27] Daniel D. Buchanan,[28,29,30] Daniel Schaid,[31] David A. Drew,[11] David C. Muller,[32] David Duggan,[33] David R. Crosslin,[34] Demetrius Albanes,[35] Edward L. Giovannucci,[12,36,37] Eric Larson,[38] Flora Qu,[1] Frank Mentch,[39] Graham G. Giles,[18,40,41] Hakon Hakonarson,[39] Heather Hampel,[42] Ian B. Stanaway,[4] Jane C. Figueiredo,[43,44] Jeroen R. Huyghe,[1] Jessica Minnier,[45] Jenny Chang-Claude,[46,47] Jochen Hampe,[48] John B. Harley,[19,20,21] Kala Visvanathan,[6] Keith R. Curtis,[1] Kenneth Offit,[49,50] Li Li,[51] Loic Le Marchand,[52] Ludmila Vodickova,[53,54,55] Marc J. Gunter,[56] Mark A. Jenkins,[18] Martha L. Slattery,[57] Mathieu Lemire,[58] Michael O. Woods,[59] Mingyang Song,[11,60,61,62] Neil Murphy,[56] Noralane M. Lindor,[64] Ozan Dikilitas,[65] Paul D.P. Pharoah,[66] Peter T. Campbell,[67]

(Author list continued on next page)

Summary

Accurate colorectal cancer (CRC) risk prediction models are critical for identifying individuals at low and high risk of developing CRC, as they can then be offered targeted screening and interventions to address their risks of developing disease (if they are in a high-risk group) and avoid unnecessary screening and interventions (if they are in a low-risk group). As it is likely that thousands of genetic variants contribute to CRC risk, it is clinically important to investigate whether these genetic variants can be used jointly for CRC risk prediction. In this paper, we derived and compared different approaches to generating predictive polygenic risk scores (PRS) from genome-wide association studies (GWASs) including 55,105 CRC-affected case subjects and 65,079 control subjects of European ancestry. We built the PRS in three ways, using (1) 140 previously identified and validated CRC loci; (2) SNP selection based on linkage disequilibrium (LD) clumping followed by machine-learning approaches; and (3) LDpred, a Bayesian approach for genome-wide risk prediction. We tested the PRS in an independent cohort of 101,987 individuals with 1,699 CRC-affected case subjects. The discriminatory accuracy, calculated by the age- and sex-adjusted area under the receiver operating characteristics curve (AUC), was highest for the LDpred-derived PRS (AUC = 0.654) including nearly 1.2 M genetic variants (the proportion of causal genetic variants for CRC assumed to be 0.003), whereas the PRS of the 140 known variants identified from GWASs had the lowest AUC (AUC = 0.629). Based on the LDpred-derived PRS, we are able to identify 30% of individuals without a family history as having risk for CRC similar to those with a family history of CRC, whereas the PRS based on known GWAS variants identified only top 10% as having a similar relative risk. About 90% of these individuals have no family history and would have been considered average risk under current screening guidelines, but might benefit from earlier screening. The developed PRS offers a way for risk-stratified CRC screening and other targeted interventions.

Introduction

Colorectal cancer (CRC) is a leading cause of cancer death, yet it is among the most preventable cancers in part because CRC screening is effective for both early detection of treatable cancers and for reducing cancer risk by removing pre-cancerous lesions.[1] Despite improvements in screening and treatment, about 50,000 fatal CRC cases

[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [2]Division of Research, Kaiser Permanente Northern California, Oakland, CA 94612, USA; [3]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; [4]Department of Medicine (Medical Genetics), University of Washington Medical Center, Seattle, WA 98195, USA; [5]Division of Human Nutrition and Health, Wageningen University & Research, Wageningen 176700, the Netherlands; [6]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, and the Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD 21287, USA; [7]University of Southern California, Preventative Medicine, Los Angeles, CA 90089, USA; [8]Department of Surgery, University of Virginia Health System, Charlottesville, VA 22903, USA; [9]Department of Cancer Biology and Genetics and the Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA; [10]Institute of Environmental Medicine, Karolinska Institutet, Stockholm 17177, Sweden; [11]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA; [12]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA; [13]Institute for Health Research, Kaiser Permanente Colorado, Denver, CO 80014, USA; [14]Institute of Cancer Research, Department of Medicine I, Medical University Vienna, Vienna 1090, Austria; [15]Department of Clinical Genetics, Karolinska University Hospital, Stockholm 17177, Sweden; [16]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm 17177, Sweden; [17]Gastroenterology Department, Hospital Clínic, Institut

(Affiliations continued on next page)

Polly A. Newcomb,[1,68] Roger L. Milne,[18,40,41] Robert J. MacInnis,[18,40] Sergi Castellví-Bel,[17] Shuji Ogino,[12,61,69,70] Sonja I. Berndt,[35] Stéphane Bézieau,[71] Stephen N. Thibodeau,[72] Steven J. Gallinger,[73] Syed H. Zaidi,[74] Tabitha A. Harrison,[1] Temitope O. Keku,[75] Thomas J. Hudson,[74] Veronika Vymetalkova,[53,54,55] Victor Moreno,[63,76,77,78] Vicente Martín,[76,79] Volker Arndt,[3] Wei-Qi Wei,[80] Wendy Chung,[81,82] Yu-Ru Su,[1] Richard B. Hayes,[83] Emily White,[1,84] Pavel Vodicka,[53,54,55] Graham Casey,[85] Stephen B. Gruber,[86] Robert E. Schoen,[87] Andrew T. Chan,[11,12,36,60,61,88] John D. Potter,[1,89] Hermann Brenner,[3,90,91] Gail P. Jarvik,[4,92] Douglas A. Corley,[2] Ulrike Peters,[1,84,*] and Li Hsu[1,93,*]

occurred in the United States (US) in 2019.[2] Better treatments have improved survival rates but achieving higher uptake and adherence to CRC screening could more rapidly reduce morbidity and mortality.[2,3] US 5-year relative survival for individuals with advanced stage cancers is below 15%, whereas individuals with cancers detected early have 5-year relative survival approaching 90%.[2] For those detected with adenomas, survival is essentially

d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Centro de Investigación Biomédica en Red de Enfermedades Hepáticas y Digestivas (CIBEREHD), University of Barcelona, Barcelona 08007, Spain; [18]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC 3000, Australia; [19]Center for Autoimmune Genomics and Etiology (CAGE), Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA; [20]University of Cincinnati College of Medicine, Cincinnati, OH 45229, USA; [21]Cincinnati VA Medical Center, Cincinnati, OH 45229, USA; [22]Department of Radiation Sciences, Oncology Unit, Umeå University, Umeå 90187, Sweden; [23]Wallenberg Centre for Molecular Medicine, Umeå University, Umeå 90187, Sweden; [24]SWOG Statistical Center, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; [25]Department of General Surgery, University Hospital Rostock, Rostock 18051, Germany; [26]Huntsman Cancer Institute and Department of Population Health Sciences, University of Utah, Salt Lake City, UT 84112, USA; [27]Leeds Institute of Cancer and Pathology, University of Leeds, Leeds LS2 9JT, UK; [28]University of Melbourne Centre for Cancer Research, Victorian Comprehensive Cancer Centre, Parkville, VIC 3010, Australia; [29]Colorectal Oncogenomics Group, Department of Clinical Pathology, The University of Melbourne, Parkville, VIC 3010, Australia; [30]Genomic Medicine and Family Cancer Clinic, Royal Melbourne Hospital, Parkville, VIC 3010, Australia; [31]Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA; [32]School of Public Health, Imperial College London, London SW7 2AZ, UK; [33]Translational Genomics Research Institute - An Affiliate of City of Hope, Phoenix, AZ 85003, USA; [34]Department of Bioinformatics and Medical Education, University of Washington Medical Center, Seattle, WA 98195, USA; [35]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA; [36]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; [37]Department of Nutrition, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA 02108, USA; [38]Kaiser Permanente Washington Research Institute, Seattle, WA 98101, USA; [39]Center for Applied Genomics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA; [40]Cancer Epidemiology Division, Cancer Council Victoria, 615 St Kilda Road, Melbourne, VIC 3004, Australia; [41]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC 3168, Australia; [42]Division of Human Genetics, Department of Internal Medicine, The Ohio State University Comprehensive Cancer Center, Columbus, OH 43210, USA; [43]Department of Medicine, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA; [44]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA; [45]School of Public Health, Oregon Health & Science University, Portland, OR 97239, USA; [46]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, 69120 Germany; [47]University Medical Centre Hamburg-Eppendorf, University Cancer Centre Hamburg (UCCH), Hamburg 20246, Germany; [48]Department of Medicine I, University Hospital Dresden, Technische Universität Dresden (TU Dresden), Dresden 01062, Germany; [49]Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY 10021, USA; [50]Department of Medicine, Weill Cornell Medical College, NY 10065, USA; [51]Department of Family Medicine, University of Virginia, Charlottesville, VA 22903, USA; [52]University of Hawaii Cancer Center, Honolulu, HI 96813, USA; [53]Department of Molecular Biology of Cancer, Institute of Experimental Medicine of the Czech Academy of Sciences, 142 20 Prague 4, Czech Republic; [54]Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University, 128 00 Prague, Czech Republic; [55]Faculty of Medicine and Biomedical Center in Pilsen, Charles University, 323 00 Pilsen, Czech Republic; [56]Nutrition and Metabolism Section, International Agency for Research on Cancer, World Health Organization, Lyon 69372, France; [57]Department of Internal Medicine, University of Utah, Salt Lake City, UT 84132, USA; [58]PanCuRx Translational Research Initiative, Ontario, Institute for Cancer Research, Toronto, ON M5G0A3, Canada; [59]Memorial University of Newfoundland, Discipline of Genetics, St. John's, NL A1B 3R7, Canada; [60]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114, USA; [61]Broad Institute of Harvard and MIT, Cambridge, MA 02141, USA; [62]Department of Nutrition, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA; [63]Oncology Data Analytics Program, Catalan Institute of Oncology, L'Hospitalet de Llobregat, Barcelona 08908, Spain; [64]Department of Health Science Research, Mayo Clinic, Scottsdale, AZ 85260, USA; [65]Department of Cardiovascular Medicine, Mayo Clinic, Rochester, MN 55905, USA; [66]Department of Public Health and Primary Care, University of Cambridge, Cambridge CB2 0SR, UK; [67]Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, GA 30303, USA; [68]School of Public Health, University of Washington, Seattle, WA 98195, USA; [69]Program in MPE Molecular Pathological Epidemiology, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA; [70]Department of Oncologic Pathology, Dana-Farber Cancer Institute, Boston, MA 02215, USA; [71]Service de Génétique Médicale, Centre Hospitalier Universitaire (CHU) Nantes, Nantes 44093, France; [72]Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 85054, USA; [73]Lunenfeld Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, ON M5G1X5, Canada; [74]Ontario Institute for Cancer Research, Toronto, ON M5G0A3, Canada; [75]Center for Gastrointestinal Biology and Disease, University of North Carolina, Chapel Hill, NC 27599, USA; [76]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid 28029, Spain; [77]Department of Clinical Sciences, Faculty of Medicine, University of Barcelona, Barcelona 08907, Spain; [78]ONCOBEL Program, Bellvitge Biomedical Research Institute (IDIBELL), L'Hospitalet de Llobregat, Barcelona 08908, Spain; [79]Biomedicine Institute (IBIOMED), University of León, León 24071, Spain; [80]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232, USA; [81]Office of Research & Development, Department of Veterans Affairs, Washington, DC 20420, USA; [82]Departments of Pediatrics and Medicine, Columbia University Medical Center, New York, NY 10032, USA; [83]Division of Epidemiology, Department of Population Health, New York University School of Medicine, New York, NY 10016, USA; [84]Department of Epidemiology, University of Washington, Seattle, WA 98195, USA; [85]Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22903, USA; [86]Department of Preventive Medicine, USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA; [87]Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, PA 15219, USA; [88]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA 02115, USA; [89]Centre for Public Health Research, Massey University, Wellington 6140, New Zealand; [90]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg 69120, Germany; [91]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; [92]Genome Sciences, University of Washington Medical Center, Seattle, WA 98195, USA; [93]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA

*Correspondence: upeters@fredhutch.org (U.P.), lih@fredhutch.org (L.H.)
https://doi.org/10.1016/j.ajhg.2020.07.006.

100%. The guidelines for initiating CRC screening are currently based mainly on two risk factors: attained age and family history of CRC.[4] Use of these criteria results in substantial under- and over-utilization of CRC screening with associated harms, because more than 80% of all CRC cases occur in those without a positive family history in first-degree relatives. It is therefore important to improve risk prediction to inform screening and other prevention strategies. Risk prediction using data from genome-wide association studies (GWASs) has been proposed in Kooperberg et al.[5] Polygenic risk scores (PRS), such as those based on LDpred,[6] have shown great promise in improving prediction for complex disease risk. The study from Khera et al.[7] is part of an emerging corpus considering the plausibility of incorporating genome-wide PRS into disease screening within health care systems.[8] For coronary artery diseases, the PRS was able to identify 10 times more people at the same or higher risk than the conventionally used monogenic test that identifies about 2 out of 100 individuals with an OR > 3. They showed similar results for other diseases, such as type 2 diabetes or breast cancer. Those at high risk can potentially benefit from targeted interventions, such as lipid-lowering drugs, dietary interventions, or screening.[7]

Models have been developed and evaluated for prediction of CRC risk using known genetic susceptibility variants identified by GWASs.[9–13] The area under the receiver operating characteristics curve (AUC) has improved as more susceptibility variants are included with the most recent model that includes 63 known variants and family history yielding AUC = 0.59 for both men and women.[9] However, we found known variants identified to date explain only about 10% of the heritable fraction of CRC risk.[14] This suggests that substantial improvement in prediction could be achieved by using a genome-wide approach that includes many more single-nucleotide polymorphisms (SNPs) that, individually, may not reach the stringent threshold for genome-wide significance.[15]

Machine-learning techniques, such as support vector machines, penalized regression, neural networks, random forests, and the extreme gradient tree boosting approaches, have been applied to GWAS data.[16–20] Typically, these approaches require first reducing the number of genetic variants from millions to thousands and then building a risk-prediction model from selected variants with various machine-learning methods. For example, a widely used approach for dimension reduction involves linkage disequilibrium (LD)-based marker pruning or clumping[21] and applying a p value threshold to association statistics. As some of the familial aggregation of CRC is explained by a polygenic component, such dimension reduction based on p values may discard variants that individually have little predictive power but collectively have substantial predictive power. To account for this possibility, the LDpred method employs a Bayesian framework to jointly model all genetic variants of the genome in building the PRS without *a priori* dimension reduction.[6]

Using statistical and machine-learning techniques on GWAS data from more than 120,000 CRC-affected case subjects and control subjects of European ancestry, we address the question of whether a PRS that uses variants beyond known CRC risk-associated variants can improve discriminatory accuracy between CRC-affected case subjects and control subjects. We developed PRS using three different approaches, based on: (1) 140 known GWAS variants as the baseline model; (2) SNP selection followed by machine learning; and (3) LDpred. We then evaluated the performance of these scores externally in an independent contemporary community-based cohort of 101,987 study participants, including 72,791 of European ancestry.

Material and Methods

Datasets

Derivation Datasets

To develop an accurate CRC risk prediction model, we used GWAS data on 55,105 case subjects and 65,079 control subjects of European ancestry from large-scale research studies (~120,000 participants with genotype data on more than 40 million variants), including the Genetics and Epidemiology of CRC Consortium and Colon Cancer Family Registry (GECCO) with 29,864 case subjects and 31,629 control subjects, the CRC Transdisciplinary Study (CORECT) with 19,885 case subjects and 12,043 control subjects, and United Kingdom Biobank (UKB) with 5,356 case subjects and 21,407 control subjects. For more details such as study participant characteristics, genotyping, imputation, quality control, and single-variant association analyses, readers are referred to the Supplemental Material and Methods (Section 3 and Table S1) and Huyghe et al.[14] Briefly, the average age was 62 years (standard deviation [SD] = 11 years). About 52% were men and 11% had a positive family history of CRC in first-degree relatives. Our primary analysis was focused on individuals of European ancestry due to insufficient numbers of CRC cases among other ancestral groups.

Evaluation Dataset

The risk prediction models were externally evaluated in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort, an independent contemporary cohort including 101,987 genotyped participants (≥18 years old) nested within the Kaiser Permanente Northern California (KPNC) integrated healthcare delivery system.[22] Participants provided a saliva sample and broadly consented to the research use of their DNA and mailed survey data, which was then linked to selected data from electronic health records. Of note, this cohort was not used in any prior discovery of CRC risk variants and, hence, provides the opportunity for an independent evaluation. Details on the genotyping array, quality control, and imputation have been described previously[23] and in the Supplemental Material and Methods (Section 4 and Table S3).

As the model building was limited to case and control subjects of European descent defined by genetic clustering with Europeans from HapMap, we also restricted the primary analysis to the genetically defined European subsets (n = 72,791, 42,520 men and 30,271 women), which included 1,311 CRC cases, 3,949 advanced adenoma cases (AA), 13,472 adenoma cases, and 10,730 individuals with hyperplastic polyps. A personal history of cancer was determined from cancer-registry data and

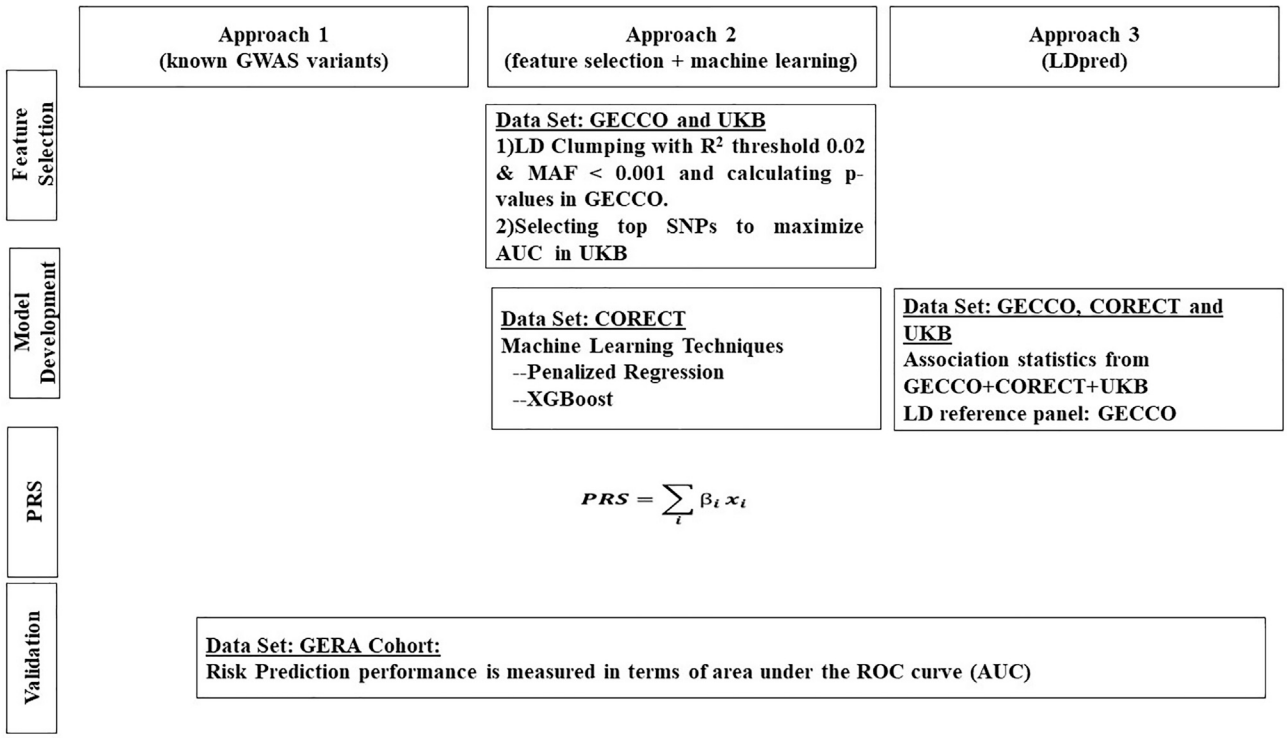

Figure 1. Description of Three Approaches to Derive Polygenic Risk Scores (PRS) for Colorectal Cancer

electronic-health-record data. A family history of CRC was ascertained by integrating data from baseline surveys and electronic health records (i.e., diagnosis codes, family history documentation). About 9.6% of participants (n = 7,029) had a positive family history in first-degree relatives. Hyperplastic polyps, AA, and non-AA were identified using Systematized Nomenclature of Medicine (SNOMED) pathology codes and validated using natural language processing.[24] We defined an AA as any adenoma with villous histology or which was 10 mm in size or greater. The cohort was unselected for any disease phenotype and GERA participants were not asked to engage in specific medical or screening tests for research purposes. However, given the age distribution of the GERA participants (median age at baseline = 52 years with median follow-up 21 years), 70% of population has undergone screening for CRC as part of their usual care, either by fecal immunochemical testing (FIT, 38%) or endoscopy (sigmoidoscopy or colonoscopy, 58%). All study participants provided written informed consent and the study was approved by the KPNC Institutional Review Board.

Validation Dataset

We further validated the models in an independent study, the Electronic Medical Records and Genomics (eMERGE) (n = 83,717). The details of the study were described elsewhere.[25] A brief description of the genotyping array, quality control, and imputation is provided in Supplemental Material and Methods (Section 5). The colorectal cancer case subjects were defined as those who had at least two ICD9/10 codes for CRC. Control subjects had zero ICD9/10 codes for CRC. Participants with a single ICD9/10 code for CRC were excluded from analysis. Adults over age 18 years who had confirmed European ancestry and no missing age were included in the validation dataset, resulting a total of 38,214 participants. The characteristics of these participants are provided in Table S10.

Polygenic Risk Score Derivation

PRS provides a quantitative measure of an individual's inherited risk based on the cumulative impact of many genetic risk variants. Each variant is scored based on the number of variant alleles an individual carries (e.g., zero, one, or two copies). The individual variant scores are then weighted according to the strength and direction of their association with disease and finally summed to give a single risk score. Imputed variants are scored by expected number of variant alleles (i.e., dosage). We studied three approaches for constructing PRS. Figure 1 depicts the summary of these different PRS derivation strategies. The weights for Approach 1 of known loci are provided in Table S4. As the number of variants for the other two approaches are very large, the weights for these variants are available upon request from the authors.

Approach 1: Known GWAS Variants

Using GWAS, we and others have identified 140 SNPs that were independently associated with CRC risk[14] and references therein.[26,27] All but three were present in the GERA dataset. For the three missing SNPs, we selected surrogates based on LD and the p value of univariate association analysis. The surrogates are provided in Table S4.

We calculated the PRS as a weighted sum of risk alleles $\sum_i \widehat{\beta}_i x_i$, where x_i is the expected number of risk alleles and $\widehat{\beta}_i$ is the log-odds ratio (OR) estimate of single-variant association from the previously published results that first reported the variants or meta-analysis results of our datasets. The meta-analysis adjusted for age, sex, study, and principal components (PCs) to account for population substructure. For the SNPs discovered in the data from this consortium, we adjusted for the winner's curse.[28] We provided the details of meta-analysis in Section 3.3, Supplemental Material and Methods.

Approach 2: SNP Selection and Machine Learning

In this approach, we first selected a subset of SNPs using LD clumping and p value thresholding and then built risk-prediction models using machine learning. To avoid overfitting, we divided the derivation datasets into two non-overlapping sets, one for SNP selection and the other for model building.

SNP Selection. We used GWAS data from GECCO (29,864 case subjects and 31,629 control subjects) and performed univariate association analysis, adjusting for age, sex, study, and PCs to account for population substructure. To remove highly correlated SNPs, we performed LD-clumping using the LD-driven p value clumping procedure in PLINK v.1.90b (–clump).[29] In this process, the algorithm generates clumps around index SNPs with p values less than an *a priori* defined threshold. Each clump contains all SNPs that are in LD with the index SNP, within 500 kilobases, as determined by pairwise correlation (R^2) threshold. The algorithm iteratively cycles through all index SNPs, beginning with the smallest p value, only allowing each index SNP to appear in one clump (non-overlapping). The final output contains the most statistically significant disease-associated SNP for each LD-based clump across the genome. To identify the optimal p value cut-off and LD-R^2 value, we chose a wide range of p value thresholds, from 5×10^{-8} to 0.01, and two R^2 values, 0.02 and 0.2, to select SNPs and calculated the corresponding PRS summing these SNPs weighted by the log-OR estimates, where the log-OR is the log-odds ratio estimate of univariate association analysis using GECCO data. We then used the UKB data (5,356 case subjects and 21,407 control subjects) to evaluate the discriminatory accuracy of these PRS (Figure S1). The AUC reached the maximum when $R^2 = 0.02$ and p value $= 1 \times 10^{-3}$. At this threshold, we had about 15,000 SNPs. We then explored further the number of SNPs ranging from 1,000 up to 50,000 and calculated the PRS by adding SNPs in the incremental order of p values. The AUC of the PRS peaked when the number of SNPs was at around 10,000 SNPs, which were used for the subsequent model building.

Model Building. Based on these selected SNPs we developed prediction models using machine-learning algorithms, using data from CORECT on 19,885 case subjects and 12,043 control subjects. We used two complementary machine-learning approaches, penalized generalized linear regression[30] and XGBoost.[31] We obtained the optimal values of the tuning parameters using 10-fold cross validation and re-estimated the regression coefficients using the entire CORECT data at the optimal tuning parameter values.

We performed penalized regression including both the known GWAS variants PRS and top SNPs from the SNP-selection step adjusting for age, sex, genotyping phase, and PCs. The confounders and known GWAS variants PRS were not penalized. We calculated the overall PRS by summing the known loci PRS and $\sum_{i=1}^{N} \widehat{\beta}_i x_i$, where x_i is the i^{th} selected SNP and $\widehat{\beta}_i$ is the corresponding regression coefficient estimate from penalized regression. We performed ridge, lasso and elastic net penalized regression. We used the R package glmnet for the ridge and lasso regression and caret for the elastic net.

XGBoost[31] is based on gradient boosted decision trees, which, in contrast to penalized regression methods, incorporate complex non-linear interactions into prediction models in a non-additive form. Boosting is a powerful ensemble learning algorithm in which weak classifiers are added sequentially to correct the errors made by existing classifiers toward building a strong classifier. As in the penalized regression, we included both the known loci PRS and top SNPs from the SNP-selection step. The PRS from XGBoost is the classifier that gives the smallest misclassification error in cross-validated datasets. We derived the model using the R package XGBoost, a fast and efficient implementation of the gradient tree boosting method.

Approach 3: LDpred

LDpred[6] is a Bayesian genetic risk prediction method, developed for genome-wide genetic risk prediction, which takes into account LD among the markers (SNPs). In an infinitesimal model, all markers are assumed to be causal and the marker effects follow a normal distribution, i.e., $\beta_i \sim N(0, (n^2/M))$, i = 1, …, M, where M is the total number of markers and h^2 is the total heritability explained by the markers. In the non-infinitesimal model, only a fraction of the M markers is assumed to be causal. A Gaussian-mixture prior is assumed in which $\beta_i \sim N(0, (n^2/M_\rho))$ with probability ρ and $\beta_i \sim 0$ with probability $(1 - \rho)$. LDpred computes the posterior mean effects of markers, taking into account the LD structure.

We used summary statistics from all GWASs, including GECCO, CORECT, and UKB, and calculated LD using the genotypes from a subset of our samples (29,305 case subjects and 31,727 control subjects) to reduce computational burden; this far exceeded the at least 2,000 individuals as suggested by LDpred. We further restricted the genetic markers to the HapMap3 panel to circumvent the non-convergence issue from training on summary statistics of very large sample sizes. LDPred requires a prior specification of ρ, the fraction of causal variants. Because ρ is generally unknown, we used a range of values for ρ: 1.0, 0.3, 0.1, 0.03, 0.01, 0.005, 0.003, and 0.001, the default values recommended by LDPred. A total of 8 candidate PRS were derived. The analysis was performed using the software LDpred.

Evaluation of Model Performance in an Independent Cohort

We evaluated the discriminatory accuracy of PRS derived from the three approaches described above in the GERA cohort by calculating the AUC.[32] Our primary outcome was CRC in European ancestry. We compared CRC case subjects with control subjects who did not have CRC or any precursor lesions, including AA, adenomas, or hyperplastic polyps. As a secondary analysis, we evaluated the AUC for AA, non-AA, and hyperplastic polyps, respectively. As sensitivity analyses, we estimated AUC using control subjects who also had precursor lesions in a sequential manner: that is, for the CRC analysis, control subjects included any precursor lesion; for AA, control subjects included adenoma and hyperplastic polyps; and for adenoma, control subjects included hyperplastic polyps. In addition, we estimated the AUCs stratified on first-degree family history (yes/no), sex (men/women), and other race/ethnicity (Asian, Hispanic, and African American). We adjusted for age (at diagnosis for case subjects and at last observation for control subjects) and sex in all AUC estimations and obtained the 95% confidence intervals by bootstrap resampling. The p values for comparing the AUC estimates between different models or groups were also obtained via bootstrap methods. A total of 500 bootstrap datasets were generated.

We performed the Cox proportional hazards model for CRC and obtained estimates of hazard ratios (HRs) and 95% confidence intervals (CI) by comparing the top percentiles (0.5%, 1%, 5%, 10%, 20%, and 30%) with the remaining percentiles (99.5%, 99%, 95%, 90%, 80%, and 70%) of PRS using Cox proportional hazards regression. Observation time was defined as the earliest of the following times: age at CRC diagnosis, death, or last follow-up. The disease status was 1 if the individual developed CRC and

Table 1. AUC Comparisons of CRC versus Control Subjects for PRS Derived via Three Different Approaches in the Independent GERA Cohort

PRS Derivation Strategy		n Variants	AUC (95% CI)
Approach 1: Known GWAS Variants			
Known variants		140	0.629 (0.613–0.645)
Approach 2: SNP Selection and Machine Learning			
Ridge		10,000	0.633 (0.617–0.648)
Lasso		10,000	0.629 (0.601–0.646)
Elastic Net		10,000	0.630 (0.612–0.641)
XGBoost		10,000	0.629 (0.614–0.643)
Approach 3: LDpred			
LDpred	$\rho = 1$	1,180,765	0.620 (0.603–0.637)
	$\rho = 0.3$	1,180,765	0.625 (0.608–0.642)
	$\rho = 0.1$	1,180,765	0.628 (0.611–0.645)
	$\rho = 0.03$	1,180,765	0.635 (0.619–0.651)
	$\rho = 0.01$	1,180,765	0.646 (0.630–0.662)
	$\rho = 0.005$	1,180,765	0.649 (0.633–0.664)
	$\rho = 0.003$	1,180,765	0.654 (0.639–0.669)
	$\rho = 0.001$	1,180,765	0.643 (0.628–0.658)

For LDpred, ρ is the proportion of genetic variants assumed to be causal for CRC.

0 otherwise. As individuals joined GERA at different ages, we treated age at starting membership as left truncated.

We estimated age-dependent disease incidences for CRC and advanced neoplasia (CRC and AA), stratified by the top 5% and bottom 5% of PRS by 1 minus the Kaplan-Meier estimator. For advanced neoplasia, the observation time was defined as the earliest of the following times: age at CRC diagnosis, AA, death, or last follow-up, and the disease status was 1 if the individual developed CRC or AA and 0 otherwise.

To gauge the potential clinical impact of PRS, we calculated the proportion of case subjects and probabilities of developing CRC by age 80, stratified by the deciles of LDpred-derived PRS. In addition, we estimated the proportion of case subjects in the top 10%, 20%, and 30% and the bottom 10%, 20%, and 30% of PRS both alone and together with family history.

We used the R packages survival for the survival analysis and survminer for the plots.

Results

Discriminatory Accuracy of Risk Prediction Models

There were 1,311 CRC case subjects and 53,722 control subjects in the GERA cohort. The AUC estimate for Approach 1 of 140 known GWAS variants was 0.629 with 95% confidence interval (CI): 0.613–0.645 (Table 1). In Approach 2, we selected a total of 10,000 SNPs, based on which we built prediction models using penalized linear regression and XGBoost. Ridge regression produced an AUC estimate of 0.633 (95% CI 0.617–0.648), slightly better than lasso (AUC 0.630, 95% CI 0.601–0.646) and elastic net (AUC 0.629, 95% CI 0.612–0.641). XGBoost had a similar AUC estimate: 0.629 (95% CI 0.614–0.643). Approach 3, LDpred, had the best performance when the fraction of causal variants (ρ) = 0.003, producing an AUC estimate of 0.654 (95% CI 0.639–0.669). This was a substantial improvement (4% increase in AUC) over both Approach 1 (p value = 0.010) and Approach 2 (p value = 0.010 for both ridge regression and XGBoost).

We further calculated the AUC of the best performing model for each approach stratified by family history and sex (Table S5). All models had statistically significantly greater AUC estimates in individuals with a positive family history than those without (the p values are 0.021, 0.020, and 0.021 for Approaches 1, 2, and 3, respectively) and there is no significant difference in AUC estimates between men and women (p values > 0.05 for all models).

In addition to CRC, we evaluated the performance of the models for advanced neoplasia, as well as CRC precursor lesions separately: AA, adenoma, and hyperplastic polyps in Europeans (Table S5). The AUC estimate of LDpred for the advanced neoplasia was 0.629 (95% CI 0.620–0.637), close to the AUC estimate for AA, as it was mainly driven by the large number of AA compared to CRC case subjects. All models showed some discriminatory accuracy between various precursor lesions compared with control subjects; however, the accuracy was sequentially reduced compared with the model for CRC. Again, LDpred had the best performance among the three approaches. As a sensitivity analysis, we assessed the AUC where the control subjects also included precursor lesions (Table S6). The AUC estimates were all reduced, but the reduction was modest ranging from 0.01 to 0.02, and the AUC still showed a sequential decrease across CRC, AA, adenoma, and hyperplastic polyps.

We estimated the AUC of the PRS among Asians (96 CRC case subjects and 5,758 control subjects), Hispanics (70 CRC case subjects and 5,221 control subjects), and African Americans (56 CRC case subjects and 2,409 control subjects). All models performed more poorly for these demographic groups than for Europeans, whether for CRC, AA, adenoma, or hyperplastic polyps (Table S7). For example, the AUC estimates of LDpred for CRC were 0.601 (95% CI 0.538–0.664), 0.602 (95% CI 0.500–0.624), and 0.543 (95% CI 0.542–0.662) for Asians, Hispanics, and African Americans, respectively, which were considerably poorer than for Europeans.

Association of PRS with Age of Diagnosis of CRC

Focusing on the best model for each approach, we estimated the HR and 95% CI for individuals in the top 30%, 20%, 10%, 5%, 1%, and 0.5% of the PRS compared with the remaining individuals (Table 2). Individuals in the top 1% of LDpred-derived PRS distribution had 2.68-fold increased CRC risk (95% CI 1.82–3.96) compared with the remaining 99% of the individuals. In contrast, the PRS from ridge regression identified only 0.5% of

Table 2. Hazard Ratio Estimates (95% Confidence Intervals) of CRC for PRS Derived from Three Different Approaches

	Approach 1		Approach 2		Approach 3	
	HR (95% CI)	p Value	HR (95% CI)	p Value	HR (95% CI)	p Value
Top 30% versus remaining	1.92 (1.75–2.23)	$<2 \times 10^{-16}$	1.94 (1.72–2.19)	$<2 \times 10^{-16}$	2.19 (1.94–2.47)	$<2 \times 10^{-16}$
Top 20% versus remaining	1.96 (1.73–2.23)	$<2 \times 10^{-16}$	2.07 (1.82–2.35)	$<2 \times 10^{-16}$	2.42 (2.14–2.74)	$<2 \times 10^{-16}$
Top 10% versus remaining	2.08 (1.82–2.70)	$<2 \times 10^{-16}$	2.26 (1.95–2.63)	$<2 \times 10^{-16}$	2.54 (2.20–2.95)	$<2 \times 10^{-16}$
Top 5% versus remaining	2.13 (1.63–2.69)	$<2 \times 10^{-16}$	2.36 (1.95–2.86)	4.9×10^{-15}	2.56 (2.12–3.09)	$<2 \times 10^{-16}$
Top 1% versus remaining	2.15 (1.17–2.90)	8.3×10^{-3}	2.34 (1.56–3.51)	3.7×10^{-5}	2.68 (1.82–3.96)	6.6×10^{-07}
Top 0.5% versus remaining	2.21 (1.16–3.81)	1.0×10^{-2}	2.77 (1.64–4.69)	1.5×10^{-3}	2.82 (1.66–4.79)	9.7×10^{-04}

Approach 1: known GWAS variants; Approach 2: SNP selection and machine learning (ridge regression); Approach 3: LDpred with $\rho = 0.003$.

individuals with a similar HR estimate. The estimates for the known GWAS variants were smaller for the same top 0.5%. Furthermore, LDpred identified more than 30% of individuals without a family history of CRC (Table S8) as having about 2.2-fold higher risk of CRC, similar to that of those with a first-degree family history of CRC. In contrast, the ridge regression identified 10%, and the known GWAS variants 5%, of these individuals as being at this level of risk.

Assessing CRC Probabilities for PRS

We estimated age-specific probabilities for developing CRC and advanced neoplasia by age 80 by percentile of PRS (Figure 2). Individuals in the top 5% of PRS (high risk) from LDpred had 7.5% (95% CI 5.6%–8.3%) and 23.5% (95% CI 21.3%–25.7%) probabilities of developing CRC and advanced neoplasia, respectively. In contrast, the probabilities for individuals in the bottom 5% of PRS (low risk) were 0.7% (95% CI: 0.1%–1.0%) and 4.3% (95% CI: 3.3%–5.3%), respectively.

We calculated the proportion of cases stratified by the deciles of LDpred-derived PRS and the corresponding disease probabilities by age 80 (Figure 3). The proportion of cases that fell in the highest decile of PRS was 23.4% (95% CI: 19.8%–27.0%); in contrast, the proportion of cases in the lowest decile was 3.3% (95% CI: 2.0%–4.6%) (Table 3).

We also estimated the disease probabilities stratified by family history of CRC (Figure S2) and advanced neoplasia (Figure S3). There was substantial variation in advanced neoplasia probabilities for top 5% and bottom 5%, even among those with a positive family history. For example, individuals with a positive family history but with LDpred-derived PRS in the low-risk group (bottom 5%) had lower lifetime risk (~8.0% by age 80) than individuals at average risk but without a family history (~12%). On the other hand, individuals with a positive family history and a LD-derived PRS in the high-risk group (top 5%) had a lifetime risk of about 35%. In general, compared with the PRS based on known GWAS variants, the LDpred-derived PRS showed a greater separation in disease probabilities between the high-risk and low-risk group and, among high-risk groups, between those with and without a family history.

Taking into account both PRS and family history simultaneously, 18.0% of individuals were either in the top 10% of PRS or had a positive family history in the cohort but constituted 39.3% of case subjects (95% CI 38.9%–39.8%) (Table 3). On the other hand, 9.1% of individuals were in the bottom 10% of PRS and had no positive family history but constituted only 2.3% of case subjects (95% CI 1.9%–2.8%). The proportion of case subjects with a positive family history was 21.0% (95% CI 19.3%–21.4%).

We further validated the LDpred models using eMERGE data. The pattern of AUC estimates for LDPred models were consistent to the results in GERA cohort; however, the AUC estimates were overall weaker. Specifically, LDpred rho = 0.005 had the best AUC 0.629 followed closely by LDpred rho = 0.003 with AUC 0.628, both of which improved substantially compared to the AUC for the known 140 GWAS loci (AUC = 0.591) (Table S11).

Discussion

It is important to be able to identify individuals at high risk of CRC to enable enhanced screening and other interventions, including dietary recommendations, weight loss, and physical activity. Equally pressing is the need to identify individuals at low risk to prevent unnecessary screening and associated complications. As CRC has a sizable heritable fraction[33] and is polygenic in nature with probably thousands of genetic variants contributing to its development,[34] utilizing genome-wide data to predict risk holds promise for risk stratification for primary and secondary prevention. Our study comprehensively explores the predictive power for CRC of genome-wide genetic data, using the largest available resources including more than 120,000 CRC case subjects and control subjects of European ancestry with individual-level genetic data for model building and an independent cohort study of more than 100,000 genotyped participants for evaluation. We show that the LDpred approach including 1.2 M variants substantially improves the discriminatory accuracy over an approach that includes only 140 known GWAS

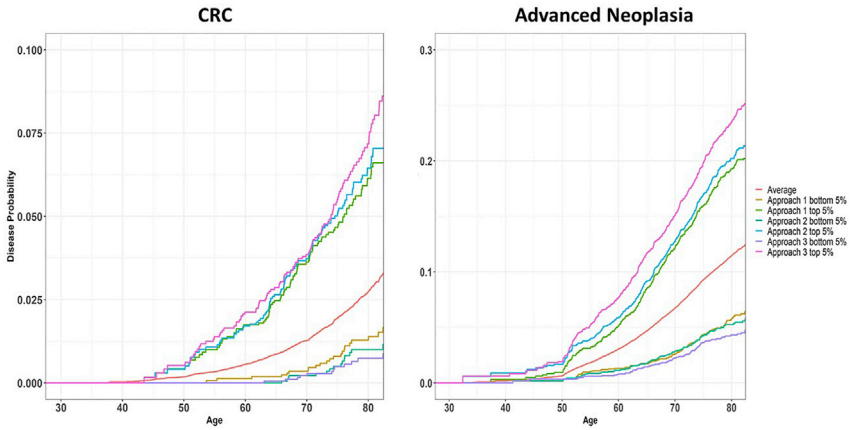

Figure 2. Disease Probabilities for Developing CRC and Advanced Adenoma
Probabilities of developing CRC (left) and advanced neoplasia (right) by age for PRS in the top 5% and bottom 5%, based on models derived from three approaches: known GWAS variants (Approach 1), SNP selection + machine learning with ridge regression (Approach 2), and LDpred with $\rho = 0.003$ (Approach 3). Average is the overall age-specific CRC (left) and advanced neoplasia (right) probabilities for the GERA.

variants. In contrast, using a combination of SNP selection and machine learning shows little improvement over the known GWAS variants. To our knowledge, the LDpred-derived PRS has the best performance of any existing CRC genetic-risk-prediction model.

Although the improvement of the AUC from 0.629 to 0.654 may not appear marked (the improvement is 4%), the AUC is an average measurement and it is critical to evaluate the model with other measures to gauge the clinical impact of the model. For example, the LDpred-derived PRS identified the top 30% of the study population as having a relative risk of ~2.2, which is similar to that associated with having an affected first-degree relative.[14,26] For individuals with an affected first-degree relative, some guidelines recommend initiation of screening with colonoscopy at an earlier age. In contrast, the PRS based on the known GWAS variants identified <5% as having a similar relative risk, demonstrating clearly the substantial improvement of the LDpred-derived PRS. It is important to note that only 10.5% of those individuals who were in the top 30% risk based on LDpred-derived PRS had a family history of CRC, demonstrating that the LDpred-derived PRS can potentially identify a larger fraction of the study population at high risk than family history alone. This means that ~27% (89.5% × 30%) of the population who are classified as average risk based on current guidelines might benefit from earlier screening. As the PRS is a continuous variable, it allows for tailored recommendation, including a specified age of starting screening,[9,26] rather than simply defining a single high-risk group based on family history that, as we show, is itself heterogeneous.

In Approach 2, if we were to use the same dataset for feature selection and model development, there would be overfitting in the model development, which result in a worse performance in an independent dataset (Supplemental Material and Methods Section 6.1 and Table S9). To mitigate this overfitting, we thus split the data in two sets in the training step. The downside is that there is potential power loss for feature selection due to smaller sample size used in calculating the test statistics compared to the entire dataset as used in Approach 3. Nevertheless, we expect that when the sample size of studies continues to rise, Approach 2 will be further improved. Our observations here are not unique to genome-wide risk prediction for colorectal cancer (see Chatterjee et al.,[15] Abraham et al.,[18] Evans et al.,[35] Yang et al.,[36] de Vlaming and Groenen,[37] and Malo et al.[38] for examples).

The LDpred approach, which builds a risk prediction model based on the entire genome, yielded better predictive performance than the approach that initially selected features before applying machine-learning algorithms. It is likely that the derivation dataset that we used for SNP selection is still too small given the large number of features (40M genetic variants) and weak effect sizes. As a result, performing SNP selection may lead to a substantial loss of information that cannot be compensated for, even with machine-learning algorithms like XGBoost. A potential limitation of LDpred is the assumption of additive effects only, whereas machine-learning approaches, such as XGBoost and random forest, can accommodate more complex non-linear effects but are not readily applicable to ultra-high dimensional data. Approaches such as deep learning that can handle ultra-high dimensional data may have potential to further improve the accuracy of prediction.

Including only the known GWAS variants (Approach 1) is simplest computationally. The SNP selection in Approach 2 also reduces computation time substantially. LDpred is the most computationally intensive due to the Monte Carlo Markov Chain (MCMC) procedure. It took ~4 days for LDpred to compute the regression weights for each parameter setting, using our computing infrastructure, which has a node of 20 cores with 768 GB memory across all cores. Although LDpred is more computationally intensive than the other two PRS approaches, the implementation of the LDpred-derived PRS into electronic health record (EHR) data, once genome-wide array or sequencing data are available, will not be much more difficult. For example, it took ~6 h to calculate the LDpred-derived PRS for 100,000 individuals in the GERA cohort. As these scores need to be calculated only once (although updates for improved models are likely), they

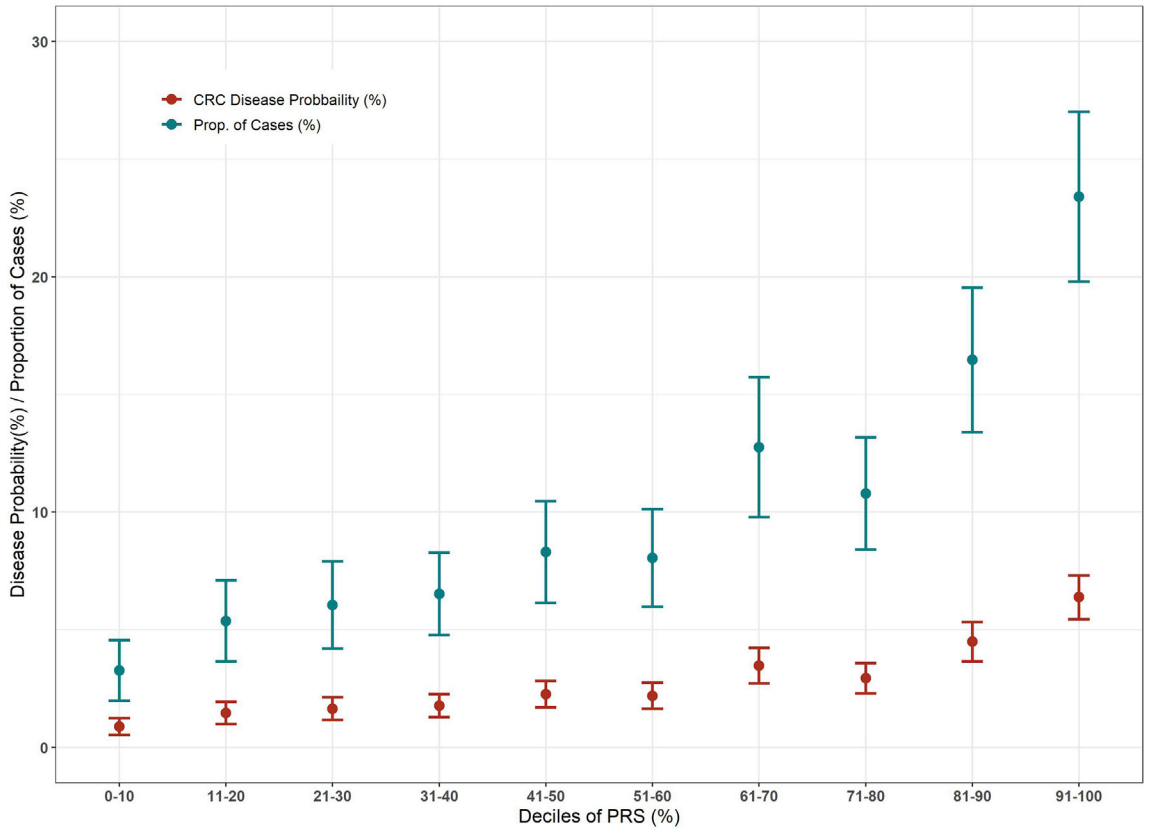

Figure 3. Disease Probabilities and Proportion of Cases (95% CI) Subjects Stratified by the Deciles of LDpred-Derived PRS

can be calculated upfront and stored as part of individual records like any other measurements (e.g., BMI, serum cholesterol). The more substantial challenge to implementation is perhaps the storage of genotype or sequencing data in a structured data object that is readily available to the EHR. To date, this challenge has not been solved in a standardized way;[39,40] however, the increasing clinical utility of PRS may motivate more rapid adoption of standardized integration of genotype and sequencing information into EHRs, which would serve as a foundation for implementation of a wide array of stratified-medicine tools.

Our study's large sample size likely is an important factor for the improved performance of the LDpred approach. Further, having access to an independent cohort that has not been included in any previous discoveries is key to provide an unbiased evaluation of the models.

Ideally, CRC would be detected early, allowing easier removal, perhaps even as a precursor lesion with a lower risk of complications and without the need for additional treatment such as radiation or chemotherapy. Previous work has shown that a PRS with fewer than 50 known loci was associated with increased risk of precursor lesions.[41,42] Consistent with these previous reports, we showed here, in our independent cohort, that all three PRS approaches also predicted AA and, to a lesser extent, adenoma and hyperplastic polyps. It is notable that as not all individuals have had endoscopy (colonoscopy or sigmoidoscopy); some control subjects in this study may have precursor lesions. As a result, the actual AUC is likely to be underestimated. Nevertheless, this decline can be expected, as the disease generally progresses from hyperplastic polyps or non-advanced adenomas to AA to CRC, with only a fraction of the precursor lesions giving rise to CRC.

There are several limitations of our PRS. First, they were built using individuals of European descent; hence, the models show substantially lower performance in other ancestral groups. This is not surprising due to the difference in LD across ancestral groups. To address this important issue, dedicated efforts focused on other major racial/ethnic populations (African Americans, Asians, and Hispanic/Latinos) are needed to develop unbiased PRS for these ancestral groups. Second, as CRCs are heterogenous with different molecularly defined subtypes, another limitation of our study is treating CRC as a single entity. However, this problem is not easy to overcome, given the need for large sample sizes and the limited availability of CRC case subjects with detailed molecular characterization. Third, while we validated that the LDpred model with rho = 0.003 performed among the best models in an independent eMERGE study, the model needs to be further evaluated for calibration as our preliminary evaluation shows (Supplemental Material and Methods Section 6.2

Table 3. Disease Probabilities (%) and Proportion of CRC Case Subjects (%) (95% CI) by Age 80 in High- and Low-Risk Groups

LDPred-Derived PRS			LDPred-Derived PRS + FamilyHx		
PRS (%)	**Disease Prob (95% CI) (%)**	**Prop of Cases (95% CI) (%)**	**PRS or Pos FamHx (%)[a]**	**Disease Prob (95% CI) (%)**	**Prop of Cases (95% CI) (%)**
Top 10	6.4 (5.5–7.3)	23.4 (19.8–27.0)	18.0	5.9 (5.2–6.6)	39.3 (38.9–39.8)
20	5.4 (4.8–6.1)	39.7 (32.7–42.8)	26.7	5.3 (4.7–5.8)	51.7 (49.1–54.2)
30	4.6 (4.1–5.1)	50.3 (46.6–55.6)	35.6	4.7 (4.2–5.1)	60.7 (57.5–63.9)
PRS (%)	**Disease Prob (95% CI) (%)**	**Prop of Cases (95% CI) (%)**	**PRS and Neg FamHx (%)[b]**	**Disease Prob (95% CI) (%)**	**Prop of Cases (95% CI) (%)**
Bottom 10	0.9 (0.5–1.2)	3.3 (2.0–4.6)	9.1	0.7 (0.3–0.9)	2.3 (1.9–2.8)
20	1.1 (0.8–1.5)	8.1 (7.5–8.7)	18.4	0.9 (0.7–1.2)	6.1 (5.4–7.1)
30	1.4 (1.0–1.6)	15.3 (14.3–16.5)	27.6	1.0 (0.9–1.2)	10.1 (8.9–12.0)

[a]PRS or Pos. FamHx: individuals were in the top x% of PRS or had a positive family history.
[b]PRS and negative FamHx: individuals were in the bottom x% and had a negative family history.

and Table S12). Caution must be taken when evaluating the calibration to account for the differences in individual-level characteristics such as screening prevalence and lifestyle risk factors.

An important question remains about how far we can improve the predictive performance using genome-wide genetic data. To this end, we showed that the best normal mixture model for effect-size distribution of our genome-wide data of common variants (allele frequency > 5%) yielded a theoretical maximal AUC of 0.68,[34] suggesting that the AUC can be further improved perhaps by using more complex models, larger number of SNPs, larger sample sizes, or some combination of these. We attempted to use all 40M SNPs imputed to the Haplotype Reference Consortium (HRC) when building LDpred models; however, we ran into convergence problems and hence limited the presentation only to SNPs in HapMap. The maximal theoretical AUC of 0.68 does not include rare variants. Based on our HRC imputed data, we estimated that at least half of CRC heritability is due to variants with an allele frequency < 1% (note this does not include high-penetrance variants as these are too rare to be imputed).[14] Accordingly, it can be expected that incorporation of rare variants can further improve the predictive performance of genome-wide genetic prediction models. This is probably not surprising as hundreds of millions of rare variants exist in the genome.

Work from our group[43–45] and others[45] has demonstrated that functional categories of the genome contribute to the heritability of CRC and that most susceptibility loci are in enhancers that vary between tumor and nonmalignant tissue. Thus, including colorectal tissue-specific functional data, such as transcriptomic or epigenomic data, would allow us to narrow down to the variants that are more likely to influence CRC risk. Our future direction is to develop methods that combine different functional annotation scores enriched for heritability, which will be particularly important as we expand prediction to rare variants. Furthermore, we will combine the PRS with other predictive factors, such as age, sex, screening history, high-penetrance genes, environmental/lifestyle risk factors, or biomarkers of early detection, which we expect, based on our previous analysis,[9] will further substantially improve risk prediction. The modifiable risk factors for the CRC are an important component of risk prediction because the best approach to primary prevention is avoidance or elimination of these risk factors. For secondary prevention, both genetics and modifiable risk factors would be helpful for determining optimal CRC screening timing and frequency.

An aim of precision/stratified medicine is to predict risk of diseases based on an individual's genetic makeup, which could, in principle, be done at birth. An important consequence of genetic risk prediction is the identification of high-risk individuals who would otherwise not be identified as high risk. Such knowledge could result in changes in healthcare management to mitigate risk with relatively low-cost lifestyle changes or preventive therapies for those at greater risk.[46] Additionally, genetic risk prediction can identify individuals at low risk who might otherwise be enrolled unnecessarily in more frequent screening or surveillance programs based on age, family history, or history of polyps. The interval between colonoscopies or the modality of screening or surveillance could be informed by PRS. Although the risk of colonoscopic perforation in the setting of cancer screening is not precisely known, estimates from diagnostic (in which there is a clinical suspicion of colorectal pathology) and therapeutic colonoscopies suggest perforations occur about once per 1,000 procedures.[47–49] Perforations are life threatening and often require laparotomy, suggesting that non-invasive screening modalities such as FIT are attractive alternatives, particularly in low-risk individuals. These are already used in other countries where population-based endoscopy screening is not available. Of course, in the US, endoscopy is not population-wide either, so the capacity to stratify individuals on screening methods appropriate to their risk should improve uptake, reduce costs, and reduce complications.

We expect that our model will be a useful first step toward prioritizing those at high risk for targeted screening or intervention and to design clinical trials to test prevention strategies in the high-risk group, particularly with the eye toward those below the age of 50 years given the rising rates of early-onset CRC. In the future, it is expected that detailed genome-wide genetic information will become part of electronic medical records of all individuals to calculate an individual PRS and identify those at high or low risk for any disease, perhaps as early as at birth. This information will allow targeted interventions such as lifestyle modifications, chemoprevention, and screening to prevent diseases or diagnose them early. Broad accessibility, dropping genotyping costs, and the need to account for an individual's risk factor profile to improve screening have provided transformative opportunities in personalized medicine. However, wide-scale adoption of PRS into clinical practice raises key ethical and scientific challenges. For example, as the current PRS has been developed in Europeans given that most GWASs are done in this population, it is substantially more predictive in Europeans compared to other populations, which will widen the health disparity gap. To overcome this major ethical and scientific challenge, it is critical that researchers invest time and effort in developing unbiased PRS across all major US populations. Furthermore, it is important to evaluate the acceptance and effectiveness of genetic testing for risk-stratified interventions among the broader population and health care providers. Cost effectiveness analysis will provide important insights to guide policies related to personalized medicine. In summary, we developed a PRS with substantially higher ability both to predict CRC risk and to identify those at high and low risk than the other two approaches. The proposed CRC PRS offers a way to improve CRC risk prediction, with the potential for translation to optimize clinical decision making.

Data and Code Availability

The source data for the findings of this study are available as follows. Genotype data for GECCO and CORECT have been deposited in the database of Genotypes and Phenotypes (dbGaP) under accession numbers phs001078.v1.p1, phs001415.v1.p1, and phs001315.v1.p1. The UK Biobank data are publicly available upon successful application from the UK Biobank. Genotype data of GERA participants who consented to having their data shared with dbGaP are available from dbGaP under accession phs000674.v2.p2. The complete GERA data are available upon successful application to the KP Research Bank. Genotype data of eMERGE participants are available from dbGaP under the accession number phs001616.v1.p1.

The codes used for statistical analysis and generation of tables and figures are publicly available.

Supplemental Data

Supplemental Data can be found online at https://doi.org/10.1016/j.ajhg.2020.07.006.

Acknowledgments

A full list of funding and acknowledgments is provided in the Supplemental Data.

Declaration of Interests

The authors declare no competing interests.

Received: November 27, 2019
Accepted: July 13, 2020
Published: August 5, 2020

Web Resources

dbGaP, https://www.ncbi.nlm.nih.gov/gap
Elastic Net, https://cran.r-project.org/web/packages/caret/index.html
KP Research Bank, https://researchbank.kaiserpermanente.org/
LDpred, https://bitbucket.org/bjarni_vilhjalmsson/LDpred
PLINK 1.9, http://www.cog-genomics.org/plink/1.9/
Ridge and Lasso Regression, https://cran.r-project.org/web/packages/glmnet/index.html
ROCt, https://www.rdocumentation.org/packages/ROCt/versions/0.9.5
Survivor, https://www.rdocumentation.org/packages/survival/versions/3.2-3
Survminer, https://www.rdocumentation.org/packages/survminer/versions/0.4.7
UK Biobank, https://www.ukbiobank.ac.uk/
XGBoost, https://www.rdocumentation.org/packages/xgboost/versions/1.1.1.1

References

1. Sandouk, F., Al Jerf, F., and Al-Halabi, M.H.D.B. (2013). Precancerous lesions in colorectal cancer. Gastroenterol. Res. Pract. *2013*, 457901.
2. Howlader, N., Noone, A.M., Krapcho, M., and Miller, D. (2019). SEER Cancer Statistics Review, 1975-2016 (Bethesda, MD: National Cancer Institute). https://seer.cancer.gov/archive/csr/1975_2016/.
3. Vogelaar, I., van Ballegooijen, M., Schrag, D., Boer, R., Winawer, S.J., Habbema, J.D.F., and Zauber, A.G. (2006). How much can current interventions reduce colorectal cancer mortality in the U.S.? Mortality projections for scenarios of risk-factor modification, screening, and treatment. Cancer *107*, 1624–1633.
4. Smith, R.A., Mettlin, C.J., Davis, K.J., and Eyre, H. (2000). American Cancer Society guidelines for the early detection of cancer. CA Cancer J. Clin. *50*, 34–49.
5. Kooperberg, C., LeBlanc, M., and Obenchain, V. (2010). Risk prediction using genome-wide association studies. Genet. Epidemiol. *34*, 643–652.
6. Vilhjálmsson, B.J., Yang, J., Finucane, H.K., Gusev, A., Lindström, S., Ripke, S., Genovese, G., Loh, P.-R., Bhatia, G., Do, R., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study (2015). Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. *97*, 576–592.

7. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nat. Genet. *50*, 1219–1224.
8. Schork, A.J., Schork, M.A., and Schork, N.J. (2018). Genetic risks and clinical rewards. Nat. Genet. *50*, 1210–1211.
9. Jeon, J., Du, M., Schoen, R.E., Hoffmeister, M., Newcomb, P.A., Berndt, S.I., Caan, B., Campbell, P.T., Chan, A.T., Chang-Claude, J., et al.; Colorectal Transdisciplinary Study and Genetics and Epidemiology of Colorectal Cancer Consortium (2018). Determining risk of colorectal cancer and starting age of screening based on lifestyle, environmental, and genetic factors. Gastroenterology *154*, 2152–2164.e19.
10. Hsu, L., Jeon, J., Brenner, H., Gruber, S.B., Schoen, R.E., Berndt, S.I., Chan, A.T., Chang-Claude, J., Du, M., Gong, J., et al.; Colorectal Transdisciplinary (CORECT) Study; and Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) (2015). A model to determine colorectal cancer risk using common genetic susceptibility loci. Gastroenterology *148*, 1330–9.e14.
11. Dunlop, M.G., Tenesa, A., Farrington, S.M., Ballereau, S., Brewster, D.H., Koessler, T., Pharoah, P., Schafmayer, C., Hampe, J., Völzke, H., et al. (2013). Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42,103 individuals. Gut *62*, 871–881.
12. Ibáñez-Sanz, G., Díez-Villanueva, A., Alonso, M.H., Rodríguez-Moranta, F., Pérez-Gómez, B., Bustamante, M., Martin, V., Llorca, J., Amiano, P., Ardanaz, E., et al. (2017). Risk Model for Colorectal Cancer in Spanish Population Using Environmental and Genetic Factors: Results from the MCC-Spain study. Sci. Rep. *7*, 43263.
13. Smith, T., Gunter, M.J., Tzoulaki, I., and Muller, D.C. (2018). The added value of genetic information in colorectal cancer risk prediction models: development and evaluation in the UK Biobank prospective cohort study. Br. J. Cancer *119*, 1036–1039.
14. Huyghe, J.R., Bien, S.A., Harrison, T.A., Kang, H.M., Chen, S., Schmit, S.L., Conti, D.V., Qu, C., Jeon, J., Edlund, C.K., et al. (2019). Discovery of common and rare genetic risk variants for colorectal cancer. Nat. Genet. *51*, 76–87.
15. Chatterjee, N., Wheeler, B., Sampson, J., Hartge, P., Chanock, S.J., and Park, J.-H. (2013). Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat. Genet. *45*, 400–405, e1–e3.
16. Wei, Z., Wang, K., Qu, H.-Q., Zhang, H., Bradfield, J., Kim, C., Frackleton, E., Hou, C., Glessner, J.T., Chiavacci, R., et al. (2009). From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. *5*, e1000678.
17. Moore, J.H., Asselbergs, F.W., and Williams, S.M. (2010). Bioinformatics challenges for genome-wide association studies. Bioinformatics *26*, 445–455.
18. Abraham, G., Kowalczyk, A., Zobel, J., and Inouye, M. (2013). Performance and robustness of penalized and unpenalized methods for genetic prediction of complex human disease. Genet. Epidemiol. *37*, 184–195.
19. Bureau, A., Dupuis, J., Hayward, B., Falls, K., and Van Eerdewegh, P. (2003). Mapping complex traits using Random Forests. BMC Genet. *4* (*Suppl 1*), S64.
20. Goldstein, B.A., Hubbard, A.E., Cutler, A., and Barcellos, L.F. (2010). An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC Genet. *11*, 49.
21. Martin, A.R., Daly, M.J., Robinson, E.B., Hyman, S.E., and Neale, B.M. (2019). Predicting polygenic risk of psychiatric disorders. Biol. Psychiatry *86*, 97–109.
22. Gordon, N.P. (2006). How does the adult Kaiser Permanente membership in Northern California compare with the larger community?. https://divisionofresearch.kaiserpermanente.org/projects/memberhealthsurvey/SiteCollectionDocuments/comparison_kaiser_vs_nonKaiser_adults_kpnc.pdf.
23. Kvale, M.N., Hesselson, S., Hoffmann, T.J., Cao, Y., Chan, D., Connell, S., Croen, L.A., Dispensa, B.P., Eshragh, J., Finn, A., et al. (2015). Genotyping informatics and quality control for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. Genetics *200*, 1051–1060.
24. Lee, J.K., Jensen, C.D., Levin, T.R., Zauber, A.G., Doubeni, C.A., Zhao, W.K., and Corley, D.A. (2019). Accurate identification of colonoscopy quality and polyp findings using natural language processing. J. Clin. Gastroenterol. *53*, e25–e30.
25. Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W.A., Li, R., Manolio, T.A., Sanderson, S.C., Kannry, J., Zinberg, R., Basford, M.A., et al.; eMERGE Network (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet. Med. *15*, 761–771.
26. Law, P.J., Timofeeva, M., Fernandez-Rozadilla, C., Broderick, P., Studd, J., Fernandez-Tajes, J., Farrington, S., Svinti, V., Palles, C., Orlando, G., et al.; PRACTICAL consortium (2019). Association analyses identify 31 new risk loci for colorectal cancer susceptibility. Nat. Commun. *10*, 2154.
27. Lu, Y., Kweon, S.-S., Tanikawa, C., Jia, W.-H., Xiang, Y.-B., Cai, Q., Zeng, C., Schmit, S.L., Shin, A., Matsuo, K., et al. (2019). Large-Scale Genome-Wide Association Study of East Asians Identifies Loci Associated With Risk for Colorectal Cancer. Gastroenterology *156*, 1455–1466.
28. Zhong, H., and Prentice, R.L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics *9*, 621–634.
29. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.
30. Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning, Second Edition (Springer).
31. Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Ann. Stat. *29*, 1189–1232.
32. Heagerty, P.J., Lumley, T., and Pepe, M.S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics *56*, 337–344.
33. Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. (2000). Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. N. Engl. J. Med. *343*, 78–85.
34. Zhang, Y., Wilcox, A.N., Zhang, H., Choudhury, P.P., Easton, D.F., Milne, R.L., Simard, J., Hall, P., Michailidou, K., Dennis, J., et al. (2020). Assessment of Polygenic Architecture and Risk Prediction based on Common Variants Across Fourteen Cancers. Nat. Commun *11*, 3353.
35. Evans, D.M., Visscher, P.M., and Wray, N.R. (2009). Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Hum. Mol. Genet. *18*, 3525–3531.

36. Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., et al. (2010). Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. *42*, 565–569.
37. de Vlaming, R., and Groenen, P.J.F. (2015). The current and future use of ridge regression for prediction in quantitative genetics. BioMed Res. Int. *2015*, 143712.
38. Malo, N., Libiger, O., and Schork, N.J. (2008). Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. Am. J. Hum. Genet. *82*, 375–385.
39. Masys, D.R., Jarvik, G.P., Abernethy, N.F., Anderson, N.R., Papanicolaou, G.J., Paltoo, D.N., Hoffman, M.A., Kohane, I.S., and Levy, H.P. (2012). Technical desiderata for the integration of genomic data into Electronic Health Records. J. Biomed. Inform. *45*, 419–422.
40. Hoffman, J.M., Haidar, C.E., Wilkinson, M.R., Crews, K.R., Baker, D.K., Kornegay, N.M., Yang, W., Pui, C.-H., Reiss, U.M., Gaur, A.H., et al. (2014). PG4KDS: a model for the clinical implementation of pre-emptive pharmacogenetics. Am. J. Med. Genet. C. Semin. Med. Genet. *166C*, 45–55.
41. Weigl, K., Thomsen, H., Balavarca, Y., Hellwege, J.N., Shrubsole, M.J., and Brenner, H. (2018). Genetic risk score is associated with prevalence of advanced neoplasms in a colorectal cancer screening population. Gastroenterology *155*, 88–98.e10.
42. Hang, D., Joshi, A.D., He, X., Chan, A.T., Jovani, M., Gala, M.K., Ogino, S., Kraft, P., Turman, C., Peters, U., et al. (2020). Colorectal cancer susceptibility variants and risk of conventional adenomas and serrated polyps: results from three cohort studies. Int. J. Epidemiol. *49*, 259–269.
43. Bien, S.A., Auer, P.L., Harrison, T.A., Qu, C., Connolly, C.M., Greenside, P.G., Chen, S., Berndt, S.I., Bézieau, S., Kang, H.M., et al.; GECCO and CCFR (2017). Enrichment of colorectal cancer associations in functional regions: Insight for using epigenomics data in the analysis of whole genome sequence-imputed GWAS data. PLoS ONE *12*, e0186518.
44. Su, Y.-R., Di, C., Bien, S., Huang, L., Dong, X., Abecasis, G., Berndt, S., Bezieau, S., Brenner, H., Caan, B., et al. (2018). A Mixed-Effects Model for Powerful Association Tests in Integrative Functional Genomics. Am. J. Hum. Genet. *102*, 904–919.
45. Hu, Y., Lu, Q., Powles, R., Yao, X., Yang, C., Fang, F., Xu, X., and Zhao, H. (2017). Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comput. Biol. *13*, e1005589.
46. De La Vega, F.M., and Bustamante, C.D. (2018). Polygenic risk scores: a biased prediction? Genome Med. *10*, 100.
47. Dafnis, G., Ekbom, A., Pahlman, L., and Blomqvist, P. (2001). Complications of diagnostic and therapeutic colonoscopy within a defined population in Sweden. Gastrointest. Endosc. *54*, 302–309.
48. Gatto, N.M., Frucht, H., Sundararajan, V., Jacobson, J.S., Grann, V.R., and Neugut, A.I. (2003). Risk of perforation after colonoscopy and sigmoidoscopy: a population-based study. J. Natl. Cancer Inst. *95*, 230–236.
49. Arora, N.K. (2009). Importance of patient-centered care in enhancing patient well-being: a cancer survivor's perspective. Qual. Life Res. *18*, 1–4.

The American Journal of Human Genetics, Volume *107*

Supplemental Data

Genome-wide Modeling of Polygenic Risk Score in Colorectal Cancer Risk

Minta Thomas, Lori C. Sakoda, Michael Hoffmeister, Elisabeth A. Rosenthal, Jeffrey K. Lee, Franzel J.B. van Duijnhoven, Elizabeth A. Platz, Anna H. Wu, Christopher H. Dampier, Albert de la Chapelle, Alicja Wolk, Amit D. Joshi, Andrea Burnett-Hartman, Andrea Gsur, Annika Lindblom, Antoni Castells, Aung Ko Win, Bahram Namjou, Bethany Van Guelpen, Catherine M. Tangen, Qianchuan He, Christopher I. Li, Clemens Schafmayer, Corinne E. Joshu, Cornelia M. Ulrich, D. Timothy Bishop, Daniel D. Buchanan, Daniel Schaid, David A. Drew, David C. Muller, David Duggan, David R. Crosslin, Demetrius Albanes, Edward L. Giovannucci, Eric Larson, Flora Qu, Frank Mentch, Graham G. Giles, Hakon Hakonarson, Heather Hampel, Ian B. Stanaway, Jane C. Figueiredo, Jeroen R. Huyghe, Jessica Minnier, Jenny Chang-Claude, Jochen Hampe, John B. Harley, Kala Visvanathan, Keith R. Curtis, Kenneth Offit, Li Li, Loic Le Marchand, Ludmila Vodickova, Marc J. Gunter, Mark A. Jenkins, Martha L. Slattery, Mathieu Lemire, Michael O. Woods, Mingyang Song, Neil Murphy, Noralane M. Lindor, Ozan Dikilitas, Paul D.P. Pharoah, Peter T. Campbell, Polly A. Newcomb, Roger L. Milne, Robert J. MacInnis, Sergi Castellví-Bel, Shuji Ogino, Sonja I. Berndt, Stéphane Bézieau, Stephen N. Thibodeau, Steven J. Gallinger, Syed H. Zaidi, Tabitha A. Harrison, Temitope O. Keku, Thomas J. Hudson, Veronika Vymetalkova, Victor Moreno, Vicente Martín, Volker Arndt, Wei-Qi Wei, Wendy Chung, Yu-Ru Su, Richard B. Hayes, Emily White, Pavel Vodicka, Graham Casey, Stephen B. Gruber, Robert E. Schoen, Andrew T. Chan, John D. Potter, Hermann Brenner, Gail P. Jarvik, Douglas A. Corley, Ulrike Peters, and Li Hsu

1. Supplemental Figures

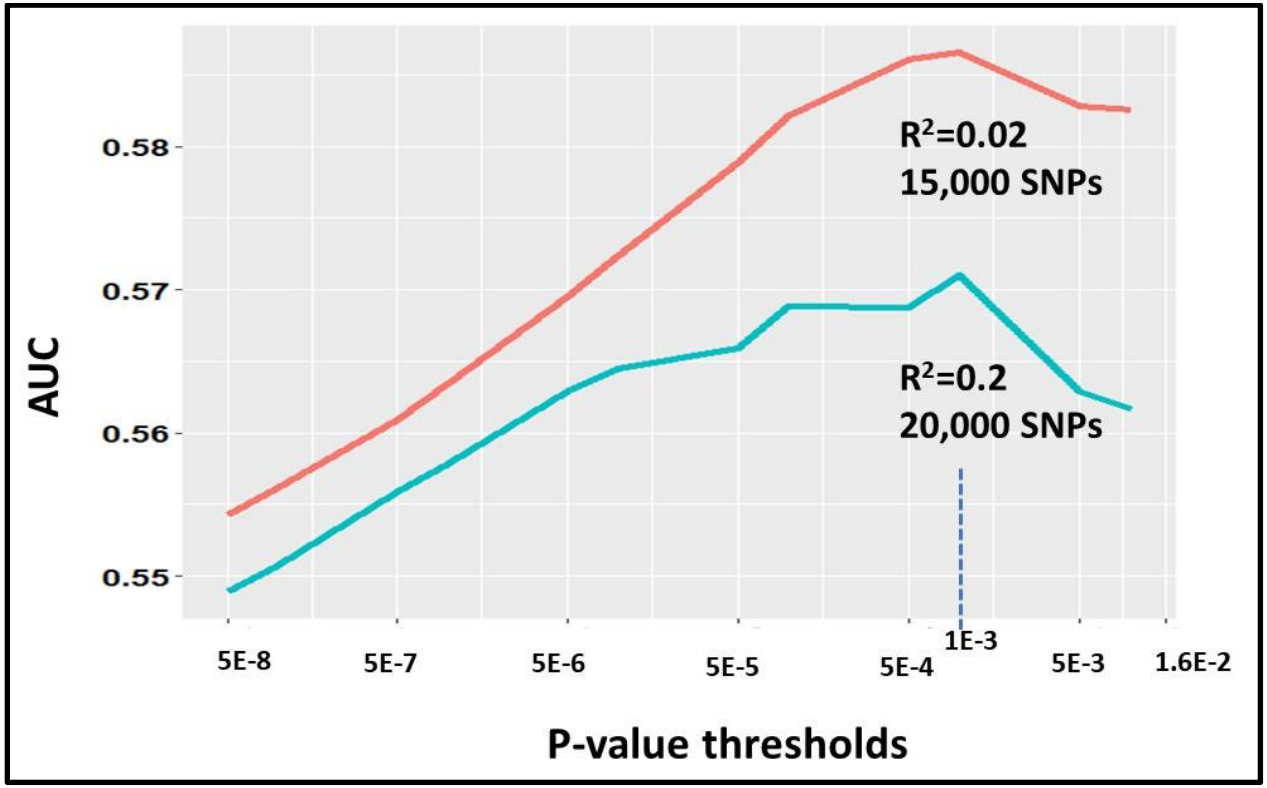

Figure S1: Feature selection with various significance thresholds and LD clumping R^2 values. Data are based on UK Biobank data after LD clumping with R^2 at 0.2 and 0.02

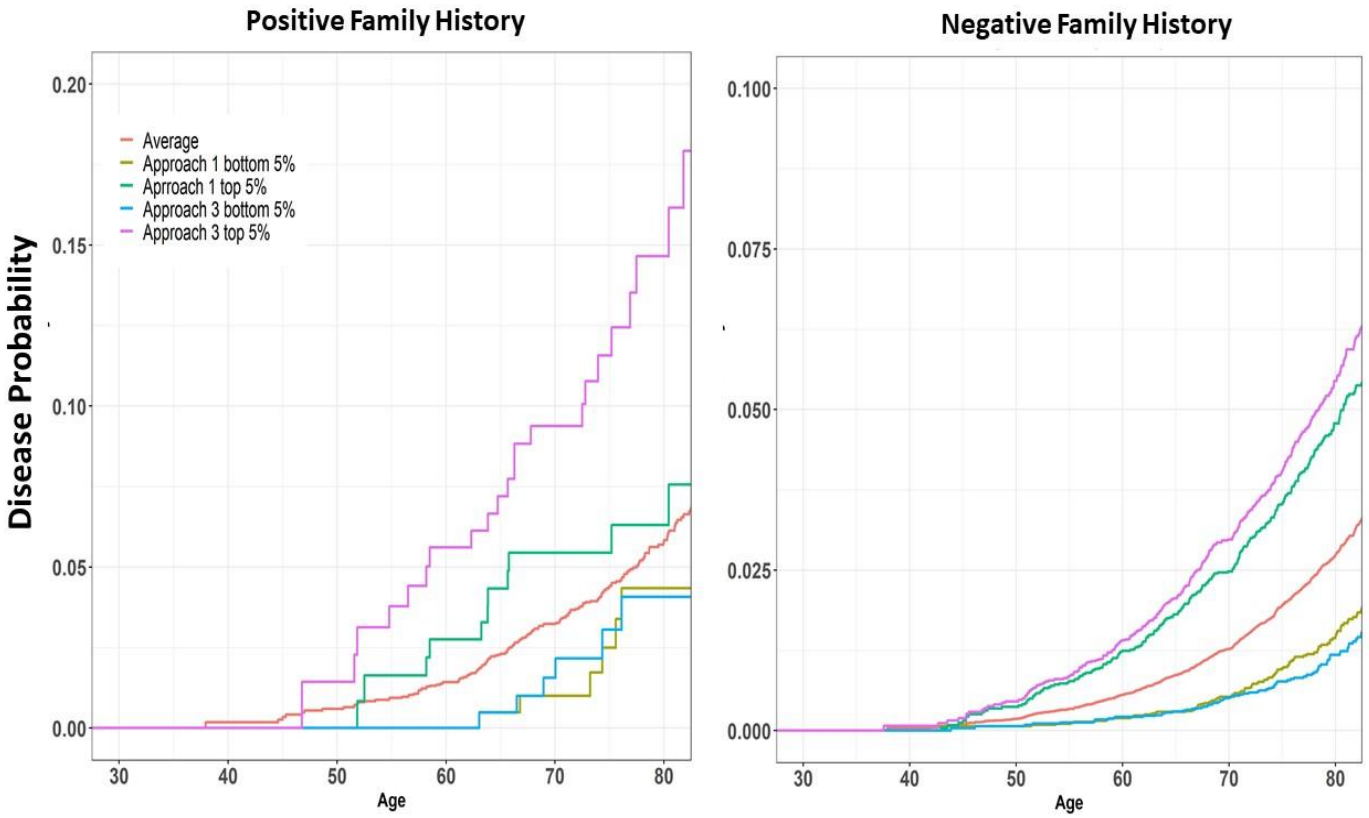

Figure S2: **Probabilities of developing colorectal cancer:** Probabilities of developing colorectal cancer by age for PRS in the top 5% and bottom 5%, based on two approaches: known GWAS variants (Approach 1) and LDpred with $\rho = 0.003$ (Approach 3), stratified by family history. Average is the overall age-specific CRC probabilities in the GERA cohort stratified by family history.

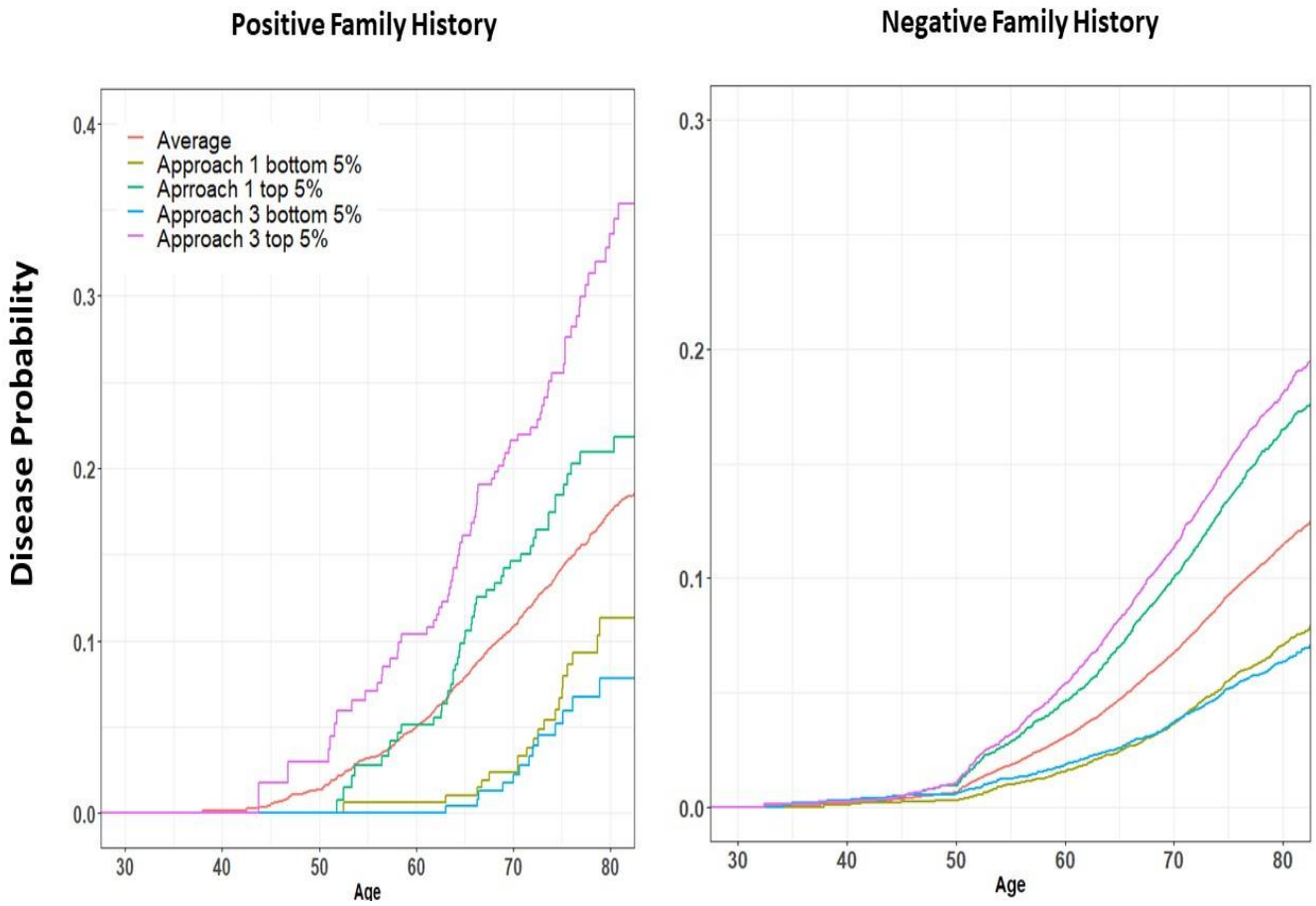

Figure S3: Probabilities of developing advanced neoplasia: Probabilities of developing advanced neoplasia by age for PRS in the top 5% and bottom 5%, based on two approaches: known GWAS variants (Approach 1) and LDpred with $\rho = 0.003$ (Approach 3), stratified by family history. Average is the overall age-specific probabilities of developing advanced neoplasia in the GERA cohort stratified by family history

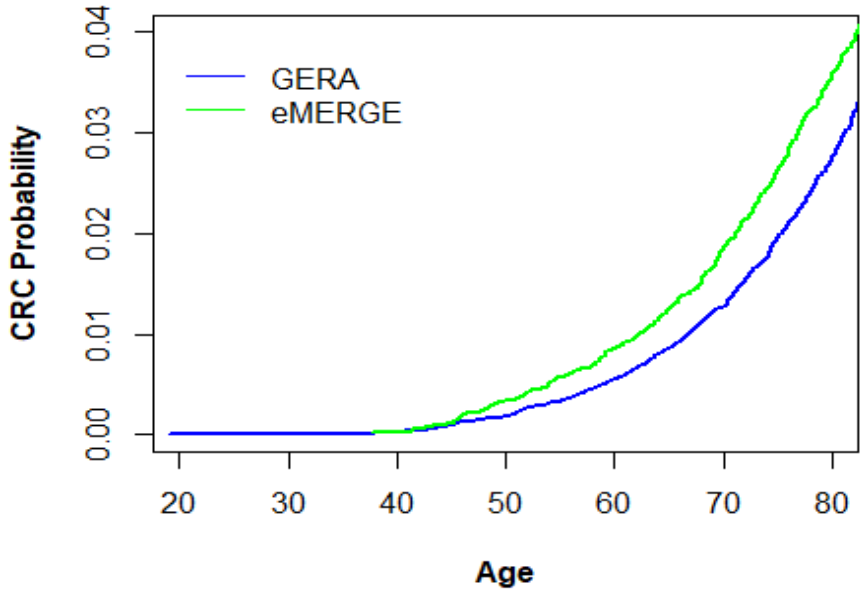

Figure S4: Age-specific probabilities of developing CRC in GERA and eMERGE

2. Supplemental Tables

Data Set	Study (Acronym)[a]	Covariates [b]
Pooled data set 1	ASTERISK, CCFR_1, CCFR_2, Colo2&3, DACHS_1, DACHS_2, DALS_1, DALS_2, HPFS_1, HPFS_2, HPFS_3_AD, MEC_1, NHS_1, NHS_2, NHS_3_AD, OFCCR, PHS, , PLCO_1, PLCO_2, PMH-CCFR, VITAL, WHI_1, WHI_2	Age, Sex, Study, Genotyping batch, 3 PCs
Pooled data set 2	ATBC, CCFR_3, CCFR_4, ColoCare_Heidelberg, ColoCare_Seattle, CPSII_1, CRCGEN, ESTHER_VERDI, Kentucky, MCCS, MECC_1, MECC_2, MECC_3, MSKCC, NFCCR, NGCCS, NHSII, SEARCH, SLRCCS, SMC_COSM, USC_HRT_CRC	Age, Sex, Genotyping platform, 23 PCs [c]
Illumina Oncoarray+custom iSelect	CLUEII, CORSA_2, CPSII_2, Czech, EDRN, EPICOLON, HawaiiCCS_AD, HPFS_5_AD, LCCS, NCCCSI, NCCCSII, NHS_5_AD, NSHDS, OSUMC, PLCO_4_AD, SELECT, SMS_AD, WHI_3	Age, Sex, Study, 10 PCs
Illumina OmniExpressExome	COLON, DACHS_3, EPIC, HPFS_4, NHS_4	Age, Sex, Study, 13 PCs
UK Biobank	UK Biobank	Age, Sex, 7 PCs
CORSA_1	CORSA_1	Age, Sex

[a]Full study names are given in Supplementary Table 1.

[b]To determine the number of genotype PCs to use as covariates in the model, we either regressed PCs on disease status and kept PC1 plus all PCs up to the highest PC with a significant *P*-value, or we visually inspected pairwise scatter plots of PCs.

[c]For pooled data set 2, we calculated PCs separately in three subsets of the data. In the pooled data set 2 mega-analysis, PCs were set to zero for participants not included in a given subset.

Table S2: Covariates included in the association analysis: The detailed description of the covariates associated in association analysis.

Risk factors	**Europeans**	**African American** [b]	**East Asians** [b]	**Hispanic**[b]
Total N [a]	72,791	5,249	6,966	6,660
CRC	1,311	56	96	70
Advanced Adenoma	3,949	198	287	320
Adenoma	13,472	556	1,195	1,080
Polyps	10,730	395	810	705
Healthy Controls	53,722	2,409	2,927	2,579
Female (%)	42,520 (58.4%)	1,866 (35.5%)	4,039 (57.9%)	4,081 (61.3%)
With endoscopy history (%)	39,020 (54.0%)	2,858 (54.4%)	3,134 (44.9%)	2,724 (41.0%)
With family history (%)	7,029 (9.6%)	456 (8.9%)	636 (9.1%)	543 (8.2%)
Age distribution at survey: Mean (min-max, Q1, Median, Q3)	62.3 (20-90, 54, 63, 71)	61.6 (20-90, 53, 62, 71)	55.8 (20-90, 47, 56, 66)	55 (20-90, 46, 56, 65)

[a] As many of the participants have multiple outcomes (e.g., polyps and adenoma), the sum of participants with colorectal cancer, advanced adenoma, adenoma, polyps, and healthy controls do not equal to the total sample size

[b]genotyping-defined ancestry

Table S3: Kaiser GERA Cohort: Descriptive statistics of Kaiser GERA study by ancestry ethnic groups.

Stratification	Cases	Controls	Approach 1[a]	Approach 2[a]	Approach 3[a]
			AUC (95% CI)		
Colorectal Cancer					
Positive Family History	277	4,066	0.636 (0.599-0.672)	0.635 (0.599-0.671)	0.667 (0.623-0.697)
Negative Family History	1,034	49,656	0.624 (0.606-0.641)	0.629 (0.612-0.646)	0.650 (0.633-0.666)
Sex: Men	637	20,628	0.628 (0.607-0.649)	0.631 (0.610-0.652)	0.670 (0.649-0.691)
Sex: Women	674	33,094	0.628 (0.606-0.649)	0.632 (0.617-0.653)	0.650 (0.636-0.676)
Types of Precursor Lesions					
Advanced Neoplasia[b]	4,852	53,722	0.606 (0.598-0.615)	0.607 (0.598-0.615)	0.629 (0.620-0.637)
Advanced Adenoma	3,949	53,722	0.601 (0.592-0.610)	0.602 (0.592-0.611)	0.626 (0.617-0.635)
Adenoma	13,472	53,722	0.572 (0.566-0.577)	0.572 (0.567-0.577)	0.595 (0.590-0.600)
Hyperplastic Polyps	10,730	53,722	0.558 (0.552-0.563)	0.558 (0.552-0.564)	0.579 (0.573-0.585)

[a] Approach 1: known GWAS variants; Approach 2: SNP selection and machine learning (ridge regression); Aapproach 3: LDpred with $\rho = 0.003$; [b] advanced neoplasia: Colorectal cancer and advanced adenoma

Table S5: AUC estimates (95% CI) of PRS stratified by family history and sex for CRC, and for various types of precursor lesions: AUC estimates (95% CI) of PRS for the best performing model for each of the three approaches stratified by family history and sex for CRC, and for various types of precursor lesions.

Stratified by CRC and Polyps	CRC[a]	Advanced adenoma[b]	Adenoma [c]	Hyperplastic polyps[d]	Advanced neoplasm [e]
Approach 1	0.615(0.600-0.615)	0.585(0.575-0.595)	0.553(0.546-0.559)	0.558(0.552-0.563)	0.595(0.587-0.603)
Approach 2	0.621(0.606-0.636)	0.586(0.576-0.596)	0.554(0.547-0.560)	0.558(0.552-0.564)	0.598(0.589-0.606)
Approach 3	0.640(0.628-0.656)	0.608(0.598-0.619)	0.566(0.559-0.572)	0.579(0.573-0.585)	0.620(0.618-0.629)

Approach 1: Known GWAS variants; Approach 2: SNP selection and machine learning (ridge regression); Approach 3: LDpred with ρ = 0.003.

[a] *CRC cases vs. All non-CRC polyps (advanced adenoma, adenoma, hyperplastic polyps and healthy controls)*

[b] *Advanced Adenoma vs. non-Advanced Adenoma (adenoma, hyperplastic polyps and healthy controls)*

[c] *Adenoma vs. Polyps (hyperplastic polyps and healthy controls)*

[d] *Hyperplastic polyps vs. non-Polyps (healthy controls)*

[e] *Advanced neoplasm (CRC cases and advanced adenoma) vs. non-Advanced Neoplasms (adenoma, hyperplastic polyps and healthy controls)*

Table S6: AUC estimates (95% confidence intervals) of PRS, comparing cases and controls who do not have CRC but not excluding any precursor lesions: AUC estimates of CRC and for various types of precursor lesions using approach1, approach 2 and approach 3.

Ethnicity			Asian	Black/African American	Hispanic
CRC	Cases /Controls(N)		96/5,758	56/2,409	70/5,221
	Approaches	1	0.591(0.536-0.625)	0.581(0.500-0.645)	0.592(0.531-0.652)
		2	0.563(0.523-0.617)	0.571(0.500-0.635)	0.564(0.504-0.625)
		3	0.601(0.538-0.664)	0.543(0.500-0.6241)	0.602(0.542-0.662)
Advanced Adenoma	Cases /Controls(N)		287/5,758	198/2,409	320/5,221
	Approaches	1	0.583(0.55-0.617)	0.539(0.500-0.595)	0.581(0.543-0.618)
		2	0.573(0.538-0.608)	0.548(0.500-0.604)	0.531(0.500-0.568)
		3	0.591(0.556-0.626)	0.579(0.520-0.638)	0.589(0.550-0.627)
Adenoma	Cases /Controls(N)		1,195/5,758	556/2,409	1,080/5,221
	Approaches	1	0.558(0.540-0.576)	0.541(0.510-0.561)	0.578(0.560-0.597)
		2	0.553(0.534-0.571)	0.539(0.508-0.569)	0.532(0.513-0.551)
		3	0.567(0.547-0.587)	0.552(0.523-0.581)	0.571(0.552-0.590)
Polyps	Cases /Controls(N)		810/5,758	395/2,409	705/5,221
	Approaches	1	0.544(0.521-0.567)	0.531(0.500-0.567)	0.572(0.549-0.595)
		2	0.550(0.527-0.573)	0.530(0.500-0.566)	0.523(0.500-0.546)
		3	0.563(0.541-0.585)	0.532(0.500-0.568)	0.557(0.535-0.579)

Table S7: Minorities AUC Estimates: AUC estimate of PRS (95% confidence intervals), stratified by minorities, for the best performing model

	Approach 1[a]		Approach 2[a]		Approach 3[a]	
	HR (95% CI)	**p-value**	**HR (95% CI)**	**p-value**	**HR (95% CI)**	**p-value**
Top 30 % vs. Remaining	1.97(1.73-2.25)	<2e-16	1.93(1.69-2.20)	<2e-16	2.26(1.98-2.58)	<2e-16
Top 20 % vs. Remaining	2.04 (1.80-3.70)	<2e-16	2.08 (1.81-2.40)	<2e-16	2.43(2.2-2.78)	<2e-16
Top 10 % vs. Remaining	2.35 (2.00-2.79)	<2e-16	2.29 (1.94-2.70)	<2e-16	2.55 (2.2-3.03)	<2e-16
Top 5 % vs. Remaining	2.19 (1.79-2.76)	6.7×10^{-13}	2.57 (2.09-3.16)	1.2×10^{-3}	2.67 (2.17-3.29)	<2e-16
Top 1 % vs. Remaining	2.37 (1.52-3.70)	1.4×10^{-4}	2.50 (1.35-3.44)	1.2×10^{-3}	2.91 (1.94-4.36)	$2.46e^{-07}$
Top 0.5% vs. Remaining	2.72 (1.50-4.93)	9.6×10^{-4}	2.91 (1.64-5.15)	2.3×10^{-4}	3.13 (1.81-5.41)	$4.42e^{-05}$
[a] Approach 1: Known GWAS variants; Approach 2: SNP selection and machine learning (ridge regression); Approach 3: LDpred with ρ = 0.003.						

Table S8: Hazard Ratio Estimates (95% Confidence Intervals) of CRC: Hazard Ratio Estimates (95% Confidence Intervals) of CRC for PRS Derived from Three Different Approaches – PRS Stratified without Family History

	Known loci	Ridge	Lasso	Elastic Net	XGBoost
	AUC Estimate				
140 known loci	0.629				
140 known loci+1000 SNPs		0.629	0.610	0.618	0.587
140 known loci +5000 SNPs		0.627	0.605	0.615	0.586
140 known loci +10000 SNPs		0.625	0.586	0.611	0.578
140 known loci +15000 SNPs		0.601	0.586	0.609	0.576
140 known loci +20000 SNPs		0.592	0.582	0.603	0.562

Table S9. AUC estimates with varying number of genetic variants using machine learning approaches based on GERA validation cohort: Estimation of AUC with varying number of genetic variants for ridge, lasso and elastic net penalized regression, and XGBoost using the entire derivation data sets for feature selection and model development**.**

	Controls	Cases
Number of Participants	37,641	573
Number of Female (%)	20,265 (54)	278 (49)
Median Entry Age (Range) (years)	52 (18, 90)	61 (24, 89)
Median Follow up (Range) (years)	13 (0, 43)	8 (0, 34)

Table S10: Characteristics of eMERGE participants.

PRS Derivation Strategy		**N Variants**	**AUC (95% CI)**
Approach 1: Known GWAS variants			
Known variants		140	0.591
Approach 3: LDpred			
LDpred	$\rho = 1$	1,180,765	0.594
	$\rho = 0.3$	1,180,765	0.601
	$\rho = 0.1$	1,180,765	0.611
	$\rho = 0.03$	1,180,765	0.623
	$\rho = 0.01$	1,180,765	0.628
	$\rho = 0.005$	1,180,765	0.629
	$\rho = 0.003$	1,180,765	0.628
	$\rho = 0.001$	1,180,765	0.623

Table S11: LDpred Results – eMERGE: AUC estimation of known loci PRS and LDpred PRRs using eMERGE data

PRS (%)	Observed Cases	Expected Cases	E/O Ratio (CI)
1-10	27	16.0	0.59 (0.41-0.87)
11-20	29	23.0	0.79 (0.55-1.14)
21-30	31	26.8	0.86 (0.61-1.23)
31-40	60	30.6	0.51(0.40-0.66)
41-50	47	33.2	0.71 (0.53-0.94)
51-60	56	39.1	0.70(0.54-0.91)
61-70	63	44.4	0.71 (0.55-0.90)
71-80	72	50.7	0.71 (0.56-0.89)
81-90	83	58.0	0.70 (0.56-0.87)
91-100	105	88.5	0.84 (0.70-1.02)

Table S12: Model calibration in eMERGE: The columns are PRS (%) in 10 equally sized groups, observed number of CRC cases, expected number of CRC cases based on the model derived from GERA, and ratio of expected and observed cases with 95% confidence intervals.

3. Derivation Data Sets

3. 1 Study description

French Association Study Evaluating RISK for sporadic colorectal cancer (ASTERISK)

Participants were recruited from the Pays de la Loire region in France between December 2002 and March 2006. Eligibility criteria for cases included being of Caucasian origin, being greater than or 40 years of age at diagnosis and having no family history of colorectal cancer or polyps. Cases were patients with first primary colorectal cancer diagnosed in one of the six public hospitals and five clinics located in the Pays de la Loire region which participated in the study. Cases were confirmed based on medical and pathology reports. Controls were recruited at two Health Examination Centers of the Pays de la Loire region, and the recruitment of controls greater than or 70 years 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, aged greater than or 40 years, 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.

Alpha-Tocopherol, Beta-Carotene Cancer Prevention (ATBC)

The ATBC Study was conducted in Finland as a joint project between the National Institute for Health and Welfare of Finland and the US National Cancer Institute. The overall design, rationale, objectives, and initial results of this intervention trial have been published. Briefly, it was a randomized, double-blind, placebo-controlled primary prevention trial testing whether daily supplementation with alpha-tocopherol, beta-carotene, or both would reduce the incidence of lung or other cancers among male smokers. The trial was registered as ClinicalTrials.gov number, NCT00342992. A total of 29,133 50-69-year-old male smokers of at least five cigarettes daily were recruited from southwestern Finland between 1985 and 1988, and randomly assigned to one of four intervention groups based on a 2 x 2 factorial design. Participants received either alpha-tocopherol (50 mg/day) as dl-alpha-tocopheryl acetate, beta-carotene (20 mg/day) as all-trans-beta-carotene, both vitamins, or placebo capsules for 5-8 years (median 6.1 years) until trial closure (April 30, 1993). Men with a prior cancer or serious illness, or who reported current use of vitamins E (>20mg/day), A (>20,000 IU/day), or beta-carotene (>6 mg/day) were ineligible. At baseline, study subjects completed a general risk factor, smoking, and medical history questionnaire, along with a food frequency (use) questionnaire, which consisted of a modified diet history, including

both portion size and frequency of consumption for 203 food items and 73 mixed dishes. Follow-up consisted of three visits annually to the local field center, during which the men were asked about their health, use of non-trial vitamin supplements, and smoking habits since the last visit. Height, weight, blood pressure, and heart rate were measured. Whole blood samples were collected from subjects close to trial closure. Incident cancer cases were identified through the Finnish Cancer Registry which provides almost 100% coverage. Between 1992 and 1993, whole blood samples were collected from approximately 20,000 participants, which were later used as the source of germline DNA. Post-intervention follow-up continues through linkage with the Finnish Cancer Registry and Register of Causes of Death. The analytic dataset from the ATBC study included in the Discovery GWAS consisted of 151 CRC cases and 32 controls.

Colon Cancer Family Registry (CCFR, www.coloncfr.org)

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. The CCFR includes data from approximately 42,500 total subjects (10,500 probands and 26,800 unaffected and affected relatives,4,276 unrelated population-based controls and 923 spouse controls). Colorectal cancer (CRC) cases and controls, age 20 –74 years, were recruited at the 6 participating centers beginning in 1998. All participants completed a standardized questionnaire that asked about 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 CCFR set 1 scan (Illumina Human 1M or Human 1M-Duo), included population-based cases and unrelated population-based controls from the 3 population-based centers: Seattle Familial Colon Cancer Family Registry (SFCCR) at Fred Hutchinson Cancer Research Center, Ontario Familial Colorectal Cancer Registry (OFCCR) at Mount Sinai Hospital (previously at Cancer Care Ontario), and the Australasian Colorectal Cancer Family Registry (ACCFR) at the University of Melbourne). Cases were genetically enriched by oversampling those with a young age at onset or a positive family history of CRC. Controls were matched to cases on age and sex. The Set 2 scan includes population-based cases and matched controls from all six Colon CFR centers including Mayo Clinic, Rochester, Minnesota; the University of Hawaii, Honolulu, Hawaii; University of Southern California consortium, Los Angeles, California; Fred Hutchinson Cancer Research Center, Seattle, Washington; Mount Sinai Hospital, Toronto, Ontario; and The University of Melbourne, Victoria, Australia. As with Set 1, cases were genetically enriched by over-sampling those with a young age at onset or positive family history. Controls were same generation family controls. The CCFR set 3 scan (Affymetrix Axiom CORECT Set array) included CRC-affected population based probands and clinic-based cases and matched controls from 5 Colon CFR centers (excluding HCCFR). Controls were related family controls or unrelated population-based controls. All participants selected for

CCFR sets-1, -2 and -3 were non-Hispanic White or of European ancestral heritage, which was confirmed with genotype data. The analytic dataset from CCFR included in the Discovery GWAS consisted of 1,972 CRC cases and 651 controls.

Hawaii Colorectal Cancer Studies 2 & 3 (Colo2&3)

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. self-reported Caucasian subjects with DNA, and clinical and epidemiologic data were selected for genotyping.

ColoCare Consortium (ColoCare)

The ColoCare Study (ClinicalTrials.gov Identifier: NCT02328677) is a prospective cohort study of newly diagnosed colorectal cancer (CRC) patients. The ColoCare Consortium is a multicenter initiative establishing an international cohort of colorectal cancer (CRC) patients for interdisciplinary studies of CRC prognosis and outcomes with sites at the Fred Hutchinson Cancer Research Center, Seattle (Washington, USA), H. Lee Moffitt Cancer Center and Research Institute, Tampa (Florida, USA), the University Hospital Heidelberg (Germany), and the Huntsman Cancer Institute (Utah, USA). The ColoCare Study investigates clinical outcomes, including disease-free and overall survival, predictors of cancer recurrence, health-related quality-of-life, and treatment toxicities. In addition, cross-sectional analyses of biomarkers and/or health behaviors are undertaken. Patients are recruited at baseline (time of first diagnosis) and followed for up to 5 years at regular time points (3 months (m), 6m, 12m, 24m, 36m, 48m, 60m). The cohort includes a comprehensive collection of specimens and data. Patients included in the CORECT project were recruited at the following ColoCare sites: Fred Hutchinson Cancer Research Center (FHCRC) and the German Cancer Research Center (DKFZ,

Heidelberg, HBG). CRC patients were recruited at the ColoCare Consortium sites when consulting with a colorectal surgeon or their staff as soon as possible after their diagnosis. Inclusion criteria for the ColoCare cohort are: (1) age 18-89 years, (2) newly-diagnosed CC (stages I-III), (3) English (FHCRC, Moffitt) or German (DKFZ) speaking, and (4) mentally/physically able to consent and participate. Pregnant women and prisoners are excluded. All activities including patient identification and recruitment, administration of health behavior questionnaires, specimen collection, medical record abstraction, biospecimen and data analysis are conducted according to IRB-approved protocols. Procedures and protocols for ColoCare FHCRC are currently approved under FHCRC IRB File 6407 and ColoCare Heidelberg (HBG) IRB approval has also been obtained (University of Heidelberg, 3/10/2010). The analytic dataset from the ColoCare study included in the Discovery GWAS consisted of 364 CRC cases and 39 controls.

Cancer Prevention Study II (CPS II)

The CPS II Nutrition cohort is a prospective study of cancer incidence and mortality in the United States, established in 1992 and described in detail elsewhere (Calle et al., 2002) [PMID: 12015775]. At enrollment, participants completed a mailed self-administered questionnaire including information on demographic, medical, diet, and lifestyle factors. Follow-up questionnaires to update exposure information and to ascertain newly diagnosed cancers were sent biennially starting in 1997. Reported cancers were verified through medical records, state cancer registry linkage, or death certificates. The Emory University Institutional Review Board approves all aspects of the CPS II Nutrition Cohort. Set 1, Set 2. A total of 360 cases and 359 controls were selected for this study.

Colorectal Cancer Genetics & Genomics (CRCGEN)

The Spanish study combines data of three case-control studies. The first one, performed in University Hospital of Bellvitge, L'Hospitalet, Barcelona, recruited 304 incident pathology- confirmed CRC cases and 293 age and sex frequency-matched hospital controls during the period 1996-1998. The control group consisted of patients without previous colorectal cancer who had been randomly selected among those admitted to the same hospital during the same period. To avoid selection bias, the criterion of inclusion in the control group was a new diagnosis. The second study, performed in the same hospital during the period 2007-2015, included a total of 324 cases and 376 population controls. The control group consisted of subjects invited to participate and selected from the primary health care lists of the hospital's referral area, frequency matched by age and sex. The third study was conducted in Hospital of Leon, Leon, during 2008-2013. A total of 325 incident CRC cases and 407 population controls were included. The control population consisted of subjects invited to participate and selected from the primary health care lists, frequency matched by

age and sex. Written informed consent was required from all participants. Each hospital's ethics committees (Bellvitge and Leon) approved the protocols of the study.

Darmkrebs: Chancen der Verhütung durch Screening (DACHS)

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 CRC risk and to investigate etiologic determinants of disease, particularly lifestyle/environmental factors and genetic factors[1,2]. Cases with a first diagnosis of invasive CRC (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 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 CRC. All hospitals treating CRC 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 CRC 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. This analysis includes participants recruited up to 2010 in this ongoing study. In total 1,268 cases and 634 matched controls were sent for genotyping using the HumanOmniExpressExome-8v1-2 array (referred to as DACHS_3).

Diet, Activity, and Lifestyle Study (DALS)

DALS was a population-based, case–control study of colon cancer.3 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. Eligibility criteria for cases included age at diagnosis between 30 and 79 years, diagnosis with first primary colon cancer (International Classification of Disease for Oncology, Second Edition, 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 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 license or state identification lists. Cases and controls were matched by 5-year age groups and sex. The set I scan consisted of a subset of the study, 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 the database of Genotypes and Phenotypes.

Epidemiologische Studie zu Chancen der Verhutung, Fruherkennung und optimierten THerapie chronischer ERkrankungen in der alteren Bevolkerungstudy (ESTHER_VERDI)

In the ESTHER/VERDI study, patients diagnosed with various forms of cancer at ages 50-75, including patients with colorectal cancer (n=420), were recruited statewide in Saarland, Germany between 1996-1998 and 2001-2003. Controls, who were frequency matched by sex and age, were randomly drawn from women and men who were recruited for a statewide cohort study in Saarland, Germany when undergoing a health check-up with their general practitioners in 2000-2002 (n=437). Blood samples were drawn by the treating physicians who also provided medical data from their records. Risk factor information was collected by self-administered standardized questionnaires. The analytic dataset from the ESTHER/VERDI study included in the Discovery GWAS consisted of 420 CRC cases and 437 controls.

Kentucky Case-Control Study (Kentucky)

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)

The MCCS is a prospective study that recruited 41,514 healthy adult volunteers (17,045 men) aged between 27 and 76 years (99% aged 40-69) from the Melbourne metropolitan area between 1990 and 1994. All CRC cases eligible for this study were selected based on the availability of a blood sample, were not genotyped previously, had no pre-baseline history of Victorian Cancer Registry (VCR) confirmed CRC or pre-baseline history of another primary cancer, excluding non-melanocytic skin cancer. Incident cases of invasive (including metastatic) adenocarcinoma of the colon or rectum were identified through the VCR up to 31st December 2012. Germline DNA was extracted from blood samples. Study participants provided written, informed consent in accordance with the Declaration of Helsinki. The study was approved by Cancer Council Victoria's Human Research Ethics Committee and performed in accordance with the institution's ethical guidelines. Set 1 consisted of 576 incident cases diagnosed during follow-up from baseline (1990-1994) till mid-2010 and 576 individually matched population-based controls. 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. Samples were genotyped on the Affymetrix Axiom CORECT Set array. Set 2 consisted of 238 CRC cases met our eligibility criteria and were matched to a control using risk set sampling with age as the time variable. Controls were matched to cases based on sex, year of baseline attendance and country of birth (Australia/New Zealand/United Kingdom/Greece/Italy/other). Samples were genotyped on OncoArray.

Multiethnic Cohort Study (MEC)

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. Set 1, 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. Set 2 were genotyped on OncoArrary.

Molecular Epidemiology of Colorectal Cancer Study (MECC)

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. Participant recruitment began in 1998 and remains on-going. Individually-matched controls with no prior history of CRC are selected from the same source population that gave rise to cases using the Clalit Health Services database. Matching factors include age, sex, Jewish ethnicity (Jew versus non-Jew), and primary clinic site. Subjects are interviewed for demographic and clinical information, family history, and dietary habits, 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. Set 1, a case-control set consisted of 484 cases and 498 controls genotyped on the Illumina Omni 2.5 array. Case selection for genotyping in MECC1 enriched for colon cancer, enriched for a specific stage distribution for a separate GWAS study of stage and prognosis, and excluded cases with microsatellite instable (MSI-H) tumors. Set 2 utilizes genotypes from 1,120 cases and 1156 controls on the Affymetrix Axiom CORECT Set array. Cases were unselected for cancer site, stage, or MSI. In addition to self-reported ancestral heritage (Ashkenazi / Sephardi), PCA analysis was used to examine the correspondence between self-reported ancestry and genotypic classification. Set 3 consisted of 3,591 cases of pathologically-confirmed adenocarcinoma and 2,848 controls on the OncoArray.

Memorial Sloan Kettering cohort (MSKCC)

The Memorial Sloan Kettering (MSK) cohort consisted of 126 individuals of Ashkenazi Jewish descent with a diagnosis of colorectal cancer and no known germline mutations in colon cancer predisposition genes. Eligible patients were ascertained between 2001–2013 under three existing MSK IRB-approved protocols allowing for tumor/germline biospecimen collection and germline analysis for cancer susceptibility. Two of the protocols specifically focused on ascertainment of patients with either early-onset (age $\leq$ 50 at diagnosis) colorectal cancer or familial colorectal cancer with no identifiable germline mutations, while the third study included colorectal cancer patients irrespective of age or family cancer history. Patient data extracted from medical records included information on stage, tumor location, chemotherapy regimen received, history of medication

use (HRT NSAIDs), endoscopy results, and metachronous or synchronous colorectal or other primary cancer diagnoses. The analytic dataset from the MSKCC study included in the Discovery GWAS consisted of 78 CRC cases.

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. Set 1, cases only genotyped on Illumina OmniQuad. Set 2 were genotyped on the Affymetrix Axiom CORECT Set array.

North German Case-Control Study (NGCCS)

All samples used were collected through the PopGen Biobank. The CRC cases were members of a patient cohort from the Kiel area, described in detail elsewhere. Briefly, CRC patients who had been diagnosed or operated on between 2002 and 2005 were identified through the cancer registry of Schleswig-Holstein or one of 25 surgical departments in Northern Germany and were contacted by mail between August 2004 and December 2006. A total of 2,715 patients agreed to participate (response rate: *40%). All cases eventually included in the study had histologically proven CRC, a primary CRC diagnosis, and no previous cancer. Venous EDTA blood samples were collected at baseline, either at the PopGen facility or by local general practitioners. Genomic DNA (600–1,000 lg) was extracted by standard methods, using the Blood Gigakit (Invitek, Berlin, Germany), and stored under quality-controlled conditions at -20 °C. For the purposes of this collaborative study, only study participants who explicitly consented to deposition of genotype data in scientific databases upon re-consent were included. The analytic dataset from the Kiel study included in the Discovery GWAS consisted of 1,119 CRC cases.

Nurses' Health Study II (NHSII)

The Nurses' Health Study II (NHSII) is an ongoing cohort of 116,430 female registered nurses in the US, aged 25-42 years at baseline in 1989. Demographic, lifestyle and health-related information were obtained from participants at baseline and updated every 2 years using self-administered questionnaires. The follow-up rate in each cycle has been over 90% to date. Study participants who had not previously reported a diagnosis of cancer and had responded to the 1995 NHSII study questionnaire were invited to provide blood samples between 1996 and 1999. Blood samples were collected from 29,611 NHSII participants, aged 32 to 54 years at the time of blood draw. Similarly, between 2004 and 2006, active study participants who had not previously provided a blood sample were invited to provide buccal samples. Swish-and-spit sample of buccal cells were received from 29,859 participants. Cases and controls selected for genotyping were nested within the subcohort of participants who provided a blood or a buccal sample. Participants with a prior history of any cancer (except non-melanoma skin cancer), ulcerative colitis, or familial polyposis syndromes were excluded. Incident cases of colorectal adenocarcinoma were ascertained first by self-report and later confirmed by reviewing medical records and pathology reports within each follow up cycle. Deaths due to colorectal cancer were identified through family or next of kin or by querying the National Death Index. Controls were randomly selected among participants in the subcohort provided they were free of colorectal cancer and matched to a corresponding case by both age (within 1 year) and sample collection date (month/year of blood or buccal sampling). Overall, 133 cases and 132 matched controls were selected for OncoArray genotyping, and 109 cases and 102 controls with ≥80% estimated European ancestry based on STRUCTURE were included in the Discovery GWAS.

Ontario Familial Colorectal Cancer Registry (OFCCR)

In GECCO, a subset of the Assessment of Risk in Colorectal Tumours in Canada (ARCTIC) from the Ontario Registry for Studies of Familial Colorectal Cancer (OFCCR) was used. Both the case-control study and the OFCCR have been described in detail previously, as have GWAS results. In brief, cases were confirmed incident colorectal cancer (CRC) cases ages 20 to 74 years, residents of Ontario identified through comprehensive registry and diagnosed between July 1997 and June 2000. Population-based controls were randomly selected among Ontario residents (random-digit-dialing and listing of all Ontario residents) and matched by sex and 5-year age groups. A total of 1,236 CRC cases and 1,223 controls were successfully genotyped on at least one of the Illumina 1536 GoldenGate assay, the Affymetrix GeneChip® Human Mapping 100K and 500K Array Set, and a 10K non-synonymous 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 appendix cases, or if they were overlapped with the Colon Cancer Family Registry. Additionally, only samples genotyped on the Affymetrix GeneChip® 500K Array were utilized to avoid coverage issues in imputation.

Physician's Health Study (PHS)

The PHS was established as a randomized, double-blind, placebo-controlled trial of aspirin and ß-carotene among 22,071 healthy U.S. male physicians, between 40 and 84 years of age in 1982. Participants completed two mailed questionnaires before being randomly assigned, additional questionnaires at six and 12 months, and questionnaires annually thereafter. In addition, participants were sent postcards at six months to ascertain status. From August 1982 to December 1984, 14,916 baseline blood samples were collected from the physicians during the run-in phase before randomization. When participants report a diagnosis of cancer, medical records and pathology reports are reviewed by study physicians who are 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 one year for younger participants, up to five 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). Due to DNA availability samples were genotyped in two batches on the same platform at the same genotyping center at different time points.

Studies of Epidemiology and Risk Factors in Cancer Heredity (SEARCH)

The study started recruitment on March 1, 2001 and all CRC cases diagnosed between the ages of 18 and 69 since January 1, 1996 in the regions served by the Eastern Cancer Registration and Information Centre were eligible for inclusion. Recruitment continued until the end 2010. Sex and age (in 5-year age bands) frequency matched controls were identified from the registration lists of ten representative general practices across East Anglia (England). Controls were matched to cases participating in SEARCH breast, colorectal, prostate, ovarian and endometrial cancer studies. All participants completed an epidemiological questionnaire, provided a blood sample for DNA and provided written informed consent. Genotyping was carried out on all SEARCH CRC cases and controls that had provided a blood sample and returned a completed consent form. SEARCH is approved by the Cambridgeshire 4 Research Ethics Committee.

The Swedish Low-Risk Colorectal Cancer Study (SLRCCS)

During the years 2004-2009 more than 3300 consecutive patients operated on for colorectal cancer (CRC) in 14 hospitals in and around Stockholm and Uppsala were included in the Swedish Colorectal Cancer Low-risk study and gave informed consent and blood for genetic studies. All cases were interviewed by the same person about their family history of colorectal cancer and other malignancies. Cancer in first- and second-degree

relatives and cousins was recorded, and pedigrees for the families of the index-person (the patient) were constructed. All diagnoses in family members which could have been CRC were verified using medical records or death certificates. Other diagnoses were coded as stated by the index case. All hematological malignancies were coded as one entity as well as all gynecological cancers because of difficulties in defining the exact diagnosis. Cases with no relative diagnosed with CRC were considered sporadic. Familial CRC was defined as cases with at least one relative with CRC in the family as defined above. All patients where relatives were at increased risk because of the family history were offered genetic counselling. Sex, age and tumor location of the index-patients were recorded based on the medical records. Tumors were assigned locations in caecum, ascending colon, hepatic flexure, transverse colon, splenic flexure, descending colon, the sigmoid or rectum. All tumors underwent evaluation directly after surgery by a local pathologist. The tumors were staged both according to the AJCC classification and the TNM-system. Some cases had two or more tumors and when tumors were located within the same segment they could be classified. As controls were used samples from 2,300 blood donors from the same region and 700 spouses to CRC patients, who did not have cancer and no family history of cancer. No information except gender was available for blood donors. For the spouses', information on gender, age, height, weight was obtained. All patients gave written informed consents in accordance with Swedish legislation and the study was approved by the Regional research ethics committee, Dnr: 02-489. The analytic dataset from this study included in the Discovery GWAS consisted of 2,667 CRC cases and 1,643 controls.

Swedish Mammography Cohort (SMC) and Swedish Men Cohort (SMC_COSM)

The Swedish Mammography Cohort (SMC) and the Cohort of Swedish Men (COSM) are two large population-based prospective cohorts from central Sweden. The SMC was initiated between 1987 and 1990 when all women born in 1914-1948 and residing in Uppsala and Västmanland counties were invited; response rate 74% (n=66,651). The COSM started in late 1997, with the invitation of all men born in 1918-1952 and residing in Västmanland and Örebro county; response rate 49% (n=48,850). Questionnaire data on diet and other lifestyle factors was collected at the start of the studies and has been updated repeatedly during follow-up. Further, biological samples (saliva, blood) have been collected together with signed informed consent and are available for DNA extraction. The cohorts are annually matched to the Swedish Cancer Register for ascertainment of incident cancer cases. For the CORECT study, follow-up through 2011 was available. The Regional Ethical Review Board at Karolinska Institutet in Stockholm approved genetic studies of CRC based on the cohorts. The analytic dataset from this study included in the Discovery GWAS consisted of 580 CRC cases and 859 controls.

USC-HRT-CRC

Observational epidemiological studies and randomized trials have reported a protective effect of estrogen and progestin therapy (EPT) on the risk of colorectal cancer, but the findings on estrogen-alone therapy (ET) are less consistent. To further investigate the relationship between menopausal hormones and risk of colon cancer, we conducted a population-based case–control study in Los Angeles County involving 831 women with newly diagnosed colon cancer and 755 population-based control women. The cases were identified by the Los Angeles County Cancer Surveillance Program, part of the National Cancer Institute's Surveillance, Epidemiology and End Results Program. Eligible subjects were English-speaking women with a histologically confirmed primary colon cancer diagnosed between the ages of 55 and 74 years on or after January 1998 through December 2002 and who were residents of Los Angeles County. Race/ ethnicity and aged matched female controls were identified through a well-established neighborhood recruitment algorithm. In-person interviews were conducted using a structured questionnaire that covered medical, menstrual, and reproductive history, use of select hormonal and non-hormonal medications, body size, physical activity, and other lifestyle factors. Interviewed participants were asked to donate a blood specimen. DNA from buffy coats of peripheral blood samples were used for OncoArray genotyping. The analytic dataset from the USC-HRT-CRC study included in the Discovery GWAS consisted of 346 CRC cases and 409 controls.

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 competed 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 22. In GECCO, a nested case-control set was genotyped. Samples included, colorectal cancer cases with DNA, excluding subject 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. Controls were matched on age at enrollment (within one year), enrollment date (within one year), sex, and race / ethnicity. One control was randomly selected per case among all controls that matched on the four factors above and where the control follow-up time was greater than follow-up time of the case until diagnosis.

Campaign against Cancer and Heart Disease II (CLUE II)

The Campaign Against Cancer and Heart Disease, is a prospective cohort designed to identify biomarkers and other factors associated with risk of cancer, heart disease, and other conditions[3]. 32,894 participants were recruited from May through October 1989 from Washington County, Maryland and surrounding communities. Colorectal cancer cases (297) and matched controls (296) were identified between 1989 and 2000 among participants in the CLUE II cohort of Washington County, Maryland, and sent for genotyping using the OncoArray+custom iSelect array.

Colorectal Cancer: Longitudinal Observational study on Nutritional and lifestyle factors that influence colorectal tumor recurrence, survival and quality of life (COLON)

The COLON study is a multi-center prospective cohort study to assess the role of diet and other lifestyle factors in cancer recurrence and survival among incident colorectal cancer patients in the Netherlands. Patients with colorectal cancer from 11 hospitals were invited upon diagnosis. Patients with a history of colorectal cancer or (partial) bowel resection, chronic inflammatory bowel disease, hereditary colorectal cancer syndromes, or dementia were excluded from the study. At diagnosis and at several time points during follow-up, patients donated a blood sample and filled out questionnaires about diet and other lifestyle factors. Blood samples are stored in a biobank to facilitate future analyses. Information on vital status is retrieved by linkage with national registries. Information on clinical characteristics is gathered from linkage with the Netherlands Cancer Registry and with hospital databases. Matching controls were selected from the Nutrition Questionnaires plus (NQplus) study. NQplus is a longitudinal observational study on diet and health in the general Dutch population. A total of 2,048 participants were recruited by inviting randomly selected inhabitants of the neighboring cities Wageningen, Ede, Renkum and Arnhem. In Veenendaal, another neighboring city, one individual of each household was invited to participate in the NQplus study. Baseline measurements consisted of a fasting venipuncture, dietary assessment, a physical examination, 24-h urine collection and general and lifestyle questionnaires. After excluding subjects with a history of colorectal cancer, chronic inflammatory bowel disease, or dementia, 692 controls were included in this study that were selected from the remaining participants and matched to 643 CRC cases of the COLON study by age and gender. All participants were genotyped using the HumanOmniExpressExome-8v1-2 array.

Colorectal Cancer Study of Austria (CORSA)

In the ongoing CRC study of Austria (CORSA), more than 13,000 Caucasian participants have been recruited within the province-wide screening project “Burgenland Prevention Trial of Colorectal Disease with Immunological Testing” (B-PREDICT) since 2003[4]. All inhabitants of the Austrian

province Burgenland aged between 40 and 80 years are annually invited to participate in fecal immunochemical testing and haemoccult positive screening participants are invited for colonoscopy. CORSA includes genomic DNA and plasma from CRC cases, low-risk and high-risk adenomas, and colonoscopy-negative controls. Controls received a complete colonoscopy and were free of CRC or polyps. CORSA participants have been recruited in the four KRAGES hospitals in Burgenland, Austria, and additionally, at the Medical University of Vienna (Department of Surgery), the Viennese hospitals "Rudolfstiftung" and the "Sozialmedizinisches Zentrum Süd", and at the Medical University of Graz (Department of Internal Medicine). Distribution of factors sex and age (5 year strata) were evenly matched between cases and controls. This study includes 1,460 CRC (941) or advanced adenoma (519) cases and 774 matched controls genotyped using the Affymetrix Axiom Genome-Wide Human Origins 1 Array, and 1210 CRC (523) or advanced adenoma (687) cases and 1273 matched controls genotyped using the OncoArray+custom iSelect array.

Cancer Prevention Study II (CPS-II)

The CPS-II Nutrition Survey cohort is a prospective study of cancer incidence and mortality in the United States, established in 1992 and described in detail elsewhere[5,6]. At enrollment, participants completed a mailed self-administered questionnaire including information on demographic, medical, diet, and lifestyle factors. Follow-up questionnaires to update exposure information and to ascertain newly diagnosed cancers were sent biennially starting in 1997. Reported cancers were verified through medical records, state cancer registry linkage, or death certificates. The Emory University Institutional Review Board approves all aspects of the CPS II Nutrition Cohort. A total of 360 cases and 359 controls were selected for genotyping using the OncoArray+custom iSelect array.

Czech Republic Colorectal Cancer Study (Czech Republic CCS)

Cases with positive colonoscopy results for malignancy, confirmed by histology as colon or rectal carcinomas, were recruited between September 2003 and May 2012 in several oncological departments in the Czech Republic (Prague, Pilsen, Benesov, Brno, Liberec, Ples, Pribram, Usti and Labem, and Zlin). Two control groups, sampled at the same time of cases recruitment, were included in the study. The first group consisted of hospital-based individuals with a negative colonoscopy result for malignancy or idiopathic bowel diseases. The reasons for the colonoscopy were: i) positive fecal occult blood test, ii) hemorrhoids, iii) abdominal pain of unknown origin, and iv) macroscopic bleeding. The second control group consisted of healthy blood donor volunteers from a blood donor center in Prague. All individuals were subjected to standard examinations to verify the health status for blood donation and were cancer-free at the time of the sampling. Details of CRC cases and controls have been reported previously[7–9]. All subjects were informed and provided written consent to participate in the study. They approved the use of their biological samples

for genetic analyses, according to the Declaration of Helsinki. The design of the study was approved by the Ethics Committee of the Institute of Experimental Medicine, Prague, Czech Republic. All subjects included in this study were Caucasians and comprised 1792 cases and 1764 matched controls. Controls were matched to CRC cases as 1:1 ratio. Matching was done on age and sex. Age was matched on ±5 years, whereas sex was matched exactly. For the cases without matched controls, matching was done only on sex. All subjects were genotyped using the OncoArray+custom iSelect array.

Early Detection Research Network (EDRN)

The aim of the EDRN initiative is to develop and sustain a biorepository for support of translational research[10]. High-quality biospecimens were accrued and annotated with pertinent clinical, epidemiologic, molecular and genomic information. A user-friendly annotation tool and query tool was developed for this purpose. The various components of this annotation tool include common data elements (CDEs) are developed from the College of American Pathologists (CAP) Cancer Checklists and North American Association of Central Cancer Registries (NAACR) standards. The CDEs provides semantic and syntactic interoperability of the data sets by describing them in the form of metadata or data descriptor. A total of 352 colorectal case samples and 399 controls were matched based on age and sex and sent for genotyping using the OncoArray+custom iSelect array.

European Prospective Investigation into Cancer and Nutrition (EPIC)

EPIC is an ongoing multicenter prospective cohort study designed to investigate the associations between diet, lifestyle, genetic and environmental factors and various types of cancer. In summary, 521,448 participants (~70% women) mostly aged 35 years or above were recruited between 1992 and 2000. Participants were recruited from 23 study centers in ten European countries. The current study included participants from France, Germany, Greece, Italy, the Netherlands, Spain, Sweden, and United Kingdom (UK). Blood samples were collected at baseline according to standardized procedures and stored at the International Agency for Research on Cancer (IARC; -196°C, liquid nitrogen) for all countries except Sweden (-80°C freezers). All study participants provided written informed consent. Ethical approval for the EPIC study was obtained from the review boards of IARC and local participating centers. Incident cancer cases were identified using population cancer registries in Italy, the Netherlands, Spain, and the United Kingdom. In Sweden, cases were identified by linkage with the essentially complete Cancer Registry of Northern Sweden and were verified by a gastrointestinal pathologist. In France, Germany and Greece, cancer cases were identified during follow-up by a combination of methods including: health insurance records, cancer and pathology registries, and by active follow-up directly through study participants or through next-of-kin. Controls were selected from the full cohort of individuals who were alive and free of cancer (except non-

melanoma skin cancer) at the time of diagnoses of the cases, using incidence density sampling and matched by: age (±6 months at recruitment), sex, study center, follow-up time since blood collection, time of day at blood collection (±4 hours), fasting status, menopausal status, and phase of menstrual cycle at blood collection. In total, 2,095 incident colorectal cancer cases, and 2,306 matched controls, genotyped using the HumanOmniExpressExome-8v1-2 array, were included in the analyses.

The EPICOLON Consortium (EPICOLON)

The EPICOLON Consortium comprises a prospective, multicentre and population-based epidemiology survey of the incidence and features of CRC in the Spanish population [11]. Cases were selected as patients with *de novo* histologically confirmed diagnosis of colorectal adenocarcinoma. Patients with familial adenomatous polyposis, Lynch syndrome or inflammatory bowel disease-related CRC, and cases where patients or family refused to participate in the study were excluded. Hospital-based controls were recruited through the blood collection unit of each hospital, together with cases. All of the controls were confirmed to have no history of cancer or other neoplasm and no reported family history of CRC. Controls were randomly selected and matched with cases for hospital, sex and age (± 5 years). A total of 370 cases and 370 controls were selected for genotyping using the OncoArray+custom iSelect array.

Hawaiian Adenoma Study

For this adenoma study, two flexible-sigmoidoscopy screening clinics were first used to recruit participants on Oahu, Hawaii. Adenoma cases were identified either from the baseline examination at the Hawaii site of the Prostate Lung Colorectal and Ovarian cancer screening trial during 1996–2000 or at the Kaiser Permanente Hawaii's Gastroenterology Screening Clinic during 1995–2007. In addition, starting in 2002 and up to 2007, we also approached for recruitment all eligible patients who underwent a colonoscopy in the Kaiser Permanente Hawaii Gastroenterology Department. Cases were patients with histologically confirmed first-time adenoma(s) of the colorectum and were of Japanese, Caucasian or Hawaiian race/ethnicity. Controls were selected among patients with a normal colorectum and were individually matched to the cases on age at exam, sex, race/ethnicity, screening date (±3 months) and clinic and type of examination (colonoscopy or flexible sigmoidoscopy). We recruited 1016 adenoma cases (67.8% of all eligible) and 1355 controls (69.2% of all eligible); 889 cases and 1169 controls agreed to give a blood and 29 cases and 34 controls, a mouthwash sample. A total of 989 cases and 1185 controls were selected for genotyping using the OncoArray+custom iSelect array. The analyses described here only included the subset of European-ancestry individuals.

Health Professionals Follow-Up Study (HPFS)

HPFS is a parallel prospective study to the NHS. The HPFS cohort comprised 51,529 men aged 40-75 who, in 1986, responded to a mailed questionnaire. 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. Follow-up evaluation has been excellent, with 94% of the men responding to date. 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. Sample selection for the stage 1 case-control sets has been described in detail previously[12,13]. For one case-control set included in stage 2 (HPFS_4), CRC cases were ascertained through January 1, 2010 and excluded cases included in stage 1[13]. Participants with histories of cancer (except nonmelanoma skin cancer), ulcerative colitis, or familial polyposis were excluded. CRC cases matched to randomly selected controls who provided a blood or buccal sample and were free of colorectal cancer at the same time the CRC was diagnosed in the cases. Matching criteria included year of birth (within 1 year) and month/year of blood or buccal cell sampling (within 1 year). If no control could be matched for a case using the initial stringent criteria, age criteria were relaxed to <5 years to find an eligible control. A total of 183 CRC cases and 197 controls genotyped using the HumanOmniExpressExome-8v1-2 array were included. In addition to the CRC case-control set, a separate case-control set (HPFS_5_AD) 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. In total, 159 advanced adenoma cases and 109 controls were selected for genotyping using the OncoArray+custom iSelect array.

Leeds Colorectal Cancer Study (LCCS)

Following local ethical approval, colorectal cancer cases were recruited from 1997 until 2012 in Leeds, UK through surgical clinics. Initially, funding was provided by the UK Ministry of Agriculture, Farming and Fisheries (subsequently the Food Standards Agency) and Imperial Cancer Research Fund (subsequently Cancer Research UK). Recruitment also occurred similarly in Dundee, Perth and York between the periods of 1997 and 2001 using the same protocol[14,15] and the data and samples were combined. Pathologically confirmed cases were consented at outpatient clinics, providing information on known and postulated risk factors for colorectal cancer (diet, lifestyle and family history) as well as providing a blood sample for DNA. Exclusion criteria included pre-existing diverticular disease and an inability to complete the questionnaire. The General Practitioners of cases (all UK residents have a nominated General Practitioner to whom to refer initial medical queries) and these GPs were asked to send letters to other persons on their patient list of the same gender and born within 5 years of the case. Subsequently to enhance the number of controls, we systematically invited patients from selected GP practices. Diet was assessed in cases and controls using an extensive dietary and lifestyle questionnaire modified by that produced by the European Prospective Investigation in Cancer (EPIC). The frequency that each specific food items were eaten was recorded and we also obtained average fruit and vegetable consumption as a cross-check. In total, 1591 cases and 739 controls provided a DNA sample for genotyping using the OncoArray+custom iSelect array.

North Carolina Colon Cancer Studies (NCCCS I/II)

The North Carolina Colon Cancer Studies (NCCCS I-colon and NCCCS II-rectal) were population-based case-control studies conducted in 33 counties of North Carolina. Cases were identified using the rapid case ascertainment system of the North Carolina Central Cancer Registry. Patients with a first diagnosis of histologically confirmed invasive adenocarcinoma of the colon (cecum through sigmoid colon) between October 1996 and September 2000 were classified as potential cases in the NCCCS I. The NCCCS II included patients with a first diagnosis of histologically confirmed invasive adenocarcinoma of the sigmoid colon, rectosigmoid, or rectum (hereafter collectively referred to as rectal cancer) between May 2001 and September 2006. Additional eligibility requirements were: aged 40–80 years, residence in one of the 33 counties, ability to give informed consent and complete an interview, had a driver's license or identification card issued by the North Carolina Department of Motor Vehicles (if under the age of 65), and had no objections from the primary physician in regards to contacting the individual. Controls, identified and sampled during the respective study dates, were selected from two sources. Potential controls under the age of 65 were identified using the North Carolina Department of Motor Vehicles records. For those 65 years and older, records from the Center for Medicare and Medicaid Services were used. Controls were matched to cases using randomized recruitment strategies. Recruitment probabilities were done using strata of 5-year age, sex, and race groups.

Dietary information was collected using a modified version of the semiquantitative food frequency questionnaire developed at the National Cancer Institute. In addition, participants were asked about vitamin and mineral supplementation, special diets, restaurant eating, sodium use, and fats used in cooking. In NCCCS I, 515 colorectal cases and 687 matched controls were sent for genotyping. In NCCCS II, 796 colorectal cases and 823 controls were sent from the NCCCS II for genotyping. Controls were matched to CRC cases as 1:1 ratio. Matching was done on age, race, and sex. Age was matched on ±5 years. Race and sex were matched exactly. For the cases without matched controls, matching was done only on sex and race. All individuals were genotyped using the OncoArray+custom iSelect array and the analyses reported here only included the subset of European-ancestry individuals.

Nurses' Health Study (NHS)

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 [16]. 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, case-control sets were previously constructed from which DNA was isolated from either buffy coat or buccal cells for genotyping. Sample selection for these case-control sets that were included in the stage 1 meta-analysis, has been described in detail previously[12,13]. For one case-control set included in stage 2 (NHS_4), colorectal cancer cases were ascertained through June 1, 2012 and excluded cases included in stage 1[13]. Participants with histories of cancer (except nonmelanoma skin cancer), ulcerative colitis, or familial polyposis were excluded. CRC cases matched to randomly selected controls who provided a blood or buccal sample and were free of colorectal cancer at the same time the colorectal cancer was diagnosed in the cases. Matching criteria included year of birth (within 1 year) and month/year of blood or buccal cell sampling (within 1 year). If no control could be matched for a case using the initial stringent criteria, age criteria were relaxed to <5 years to find

an eligible control. A total of 308 CRC cases and 303 controls genotyped using the HumanOmniExpressExome-8v1-2 array were included. In addition to the CRC case-control set, a separate case-control set (NHS_5_AD) 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. In total, 272 advanced adenoma cases and 236 matched controls were selected for genotyping using the OncoArray+custom iSelect array.

The Northern Sweden Health and Disease Study (NSHDS)

This study comprises over 110,000 participants, including approximately one third with repeated sampling occasions, from three population-based cohorts[17,18]. The largest is the ongoing Västerbotten Intervention Programme, in which all residents of Västerbotten County are invited to a health examination upon turning 30 (some years), 40, 50 and 60 years of age. Extensive measured and self-reported health and lifestyle data, as well as blood samples for central biobanking in Umeå, Sweden, are collected at the health exam. In total 878 leukocyte DNA samples for a 1:1-matched CRC case-control set were selected for genotyping using the OncoArray+custom iSelect array.

Columbus-area HNPCC study (HNPCC), Ohio Colorectal Cancer Prevention Initiative (OCCPI), Ohio State University Medical Center (OSUMC)

Patients with colorectal adenocarcinoma diagnosed at six participating hospitals were eligible for this study, regardless of age at diagnosis or family history of cancer. Patients with a clinical diagnosis of familial adenomatous polyposis were not eligible for this study. These six hospitals perform the vast majority of all operations for CRC in the Columbus metropolitan area (population 1.7 million). The institutional review board at all participating hospitals approved the research protocol and consent form in accordance with assurances filed with and approved by the United States Department of Health and Human Services. Briefly, during the period of January 1999 through August 2004, 1,566 eligible patients with CRC were accrued to the study[19]. A total of 1472 colorectal cancer samples had enough blood DNA remaining to be sent for genotyping. OCCPI (ClinicalTrials.gov identifier: NCT01850654) is a population-based study of colorectal cancer patients diagnosed in one of 51 hospitals throughout the state of Ohio from January 1, 2013 through December 31, 2016. The OCCPI was created to decrease CRC incidence in Ohio by identifying

patients with hereditary predisposition (statewide universal tumor screening for newly diagnosed CRC patients), increase colonoscopy compliance for first-degree relatives of CRC patients, and encourage future research through the creation of a biorepository. The 51 Ohio hospitals participating in the OCCPI were selected to represent a cross-section of clinical centers in the state based on high reported volume of CRC patients, affiliation with a high volume hospital, or interest in participation. Institutional Review Board (IRB) approval was obtained by the individual hospitals, Community Oncology Programs, or by ceding review to the OSU IRB. Written informed consent was obtained. A total of 2139 colorectal cases were genotyped. Patients were considered eligible for this study if they were age 18 or older at the time of enrollment, if they had a surgical resection (or biopsy if unresectable) in the state of Ohio demonstrating an adenocarcinoma of the colorectum from 1/1/13–12/31/16. Cases from the HNPCC and OCCPI studies were matched to controls selected from the Ohio State University Medical Center's (OSUMC) Human Genetics Sample Bank. The Columbus Area Controls Sample Bank is a collection of control samples for use in human genetics research that includes both donors' anonymized biological specimens and linked phenotypic data. The data and samples are collected under the protocol "Collection and Storage of Controls for Genetics Research Studies", which is approved by the Biomedical Sciences Institutional Review Board at OSUMC. Recruitment takes place in OSUMC primary care and internal medicine clinics. If individuals agree to participate, they provide written informed consent, complete a questionnaire that includes demographic, medical and family history information, and donate a blood sample. 4-7 ml of blood is drawn into each of 3 ACD Solution A tubes and is used for genomic DNA extraction and the establishment of an EBV-transformed lymphoblastoid cell culture, cell pellet in Trizol, and plasma. Controls were matched to CRC cases as 1:1. Matching was done on age at reference time (age_ref), race, and sex. Age_ref was matched on ±5 years. Sex and race were matched exactly. For the cases without matched controls, matching was done only on sex and race with 1:1 ratio. Since controls are fewer than cases, one control is matched on 2 cases at most. All samples were genotyping using the OncoArray+custom iSelect array.

Postmenopausal Hormones - Colon Cancer Family Registry (PMH-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.

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 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 [20] 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[21] (all male) and the GWAS of Lung Cancer and Smoking[22] (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). For this study 1651 advanced adenoma cases and 1392 controls were selected for genotyping using the OncoArray+custom iSelect array.

Selenium and Vitamin E Prevention Trial (SELECT)

The Selenium and Vitamin E Cancer Prevention Trial (SELECT) was a double-blind, placebo controlled clinical trial which explored using selenium and vitamin E alone and in combination to prevent prostate cancer in healthy men[23]. Secondary endpoints included the prevention of colorectal and lung cancers. SELECT was conducted at 427 sites and centers in the United States, Canada and Puerto Rico; 35,533 men 55 years and older (50 or older if African American) were randomized beginning August 22, 2001. Supplementation was discontinued on October 23, 2008 due to futility. 308 colorectal cancer cases and 308 matched controls were selected from the SELECT population and sent for genotyping using the OncoArray+custom iSelect array.

Screening Markers for Colorectal Cancer Study (advanced adenomas) (SMS_AD)

Details on this study population were previously reported[24]. Participants were enrollees in an integrated health-care delivery system in western Washington State (Group Health Cooperative, Seattle, Washington) aged 24–79 years who underwent an index colonoscopy for any indication between 1998 and 2007 and donated a buccal-cell or blood sample for genotyping analysis. Study recruitment took place in 2 phases, with phase 1 occurring in 1998–2003 and phase 2 occurring in 2004–2007. Persons who had undergone a colonoscopy less than 1 year prior to the index colonoscopy, persons with inadequate bowel preparation for the index colonoscopy, and persons with a prior or new diagnosis of colorectal cancer, a familial colorectal cancer syndrome (such as familial adenomatous polyposis), or another colorectal disease were ineligible. Patients diagnosed with adenomas or serrated polyps and persons who were polyp-free at the index colonoscopy (controls) were systematically recruited during both phases of recruitment. Approximately 75% agreed to participate and provided written informed consent. Based on medical records, persons who agreed to participate and those who refused study participation were similar with respect to age, sex, and colorectal polyp status. Study protocols were approved by the institutional review boards of the Group Health Cooperative and the Fred Hutchinson Cancer Research Center (Seattle, Washington). A total of 575 cases and 508 matched were selected for the OncoArray+custom iSelect array genotyping project. Controls were matched to CRC cases as 1:1 ratio. Matching was done on age_ref, race, and sex. Age_ref was matched on ±5 years.

UK Biobank (UKB)

We constructed a CRC and advanced adenoma nested case-control dataset from the UK Biobank resource which was accessed through application number 8614. CRC cases were defined as subjects with primary invasive CRC diagnosed, or who died from CRC according to ICD9 (1530-1534, 1536-1541) or ICD10 (C180, C182-C189, C19, C20) codes. Appendix cases, non-invasive (*in situ*) CRC cases, cases with histology of tumor as

carcinoid, and related tumors and lymphomas (ICD-O-3 tumor histology codes 8240-8249, 9590-9729) were excluded. Advanced adenoma cases were defined as primary in situ CRC cases according to ICD9 (2303, 2304) or ICD10 (D010-D012) codes, or benign neoplasms according to ICD10 codes (D120, D122, D123, D124-D128, D374, D375) with ICD-O-3 tumor histology codes 8210, 8211, 8220, 8221, or 8261-8263. Incident and prevalent CRC or advanced adenoma cases were defined based on date of diagnosis and date of enrollment. Eligible control participants were required to be free of invasive colorectal cancer, non-invasive (*in situ*) CRC, appendix, anus, anal canal, and overlapping lesion of rectum, anus and anal canal cancer, or advanced adenoma. For incident cases, each case was matched with 4 controls that exactly matched the following matching criteria: age at enrollment, year at enrollment, race/ethnicity, and sex. 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 event, and then multiple passes were made to match second, third and fourth controls. For prevalent cases, each case was matched with 4 controls that exactly matched the following matching criteria: year at enrollment, race/ethnicity, and sex. The risk set was then defined as controls who were at risk at the age when the cases were diagnosed. For matching of both incident and prevalent cases, the matching algorithm selected the closest match based on criteria to minimize an overall distance measure[25] in total, 5,356 CRC (5,004) or advanced adenoma (352) cases and 21,407 matched controls were included in the stage 2 analysis. All participants were genotyped using the Affymetrix UK Biobank Axiom Array[25,26].

Women's Health Initiative Study (WHI)

WHI is a long-term national health study that has focused on strategies for preventing heart disease, breast and colorectal cancer, and osteoporotic fractures in postmenopausal women. The original WHI study included 161,808 postmenopausal women enrolled between 1993 and 1998. The Fred Hutchinson Cancer Research Center in Seattle, WA serves as the WHI Clinical Coordinating Center for data collection, management, and analysis of the WHI. The WHI has two major parts: a partial factorial randomized Clinical Trial (CT) and an Observational Study (OS); both were conducted at 40 Clinical Centers nationwide. The CT enrolled 68,132 postmenopausal women between the ages of 50-79 into trials testing three prevention strategies. If eligible, women could choose to enroll in one, two, or all three of the trial components. The components are: **Hormone Therapy Trials (HT)**: This double-blind component examined the effects of combined hormones or estrogen alone on the prevention of coronary heart disease and osteoporotic fractures, and associated risk for breast cancer. Women participating in this component with an intact uterus were randomized to estrogen plus progestin (conjugated equine estrogens [CEE], 0.625 mg/d plus medroxyprogesterone acetate [MPA] 2.5 mg/d] or a matching placebo. Women with prior hysterectomy were randomized to CEE or placebo. Both trials were stopped early, in July 2002 and March 2004, respectively, based on adverse effects. All HT participants continued to be followed without intervention until close-out. **Dietary**

Modification Trial (DM): The Dietary Modification component evaluated the effect of a low-fat and high fruit, vegetable and grain diet on the prevention of breast and colorectal cancers and coronary heart disease. Study participants were randomized to either their usual eating pattern or a low-fat dietary pattern. **Calcium/Vitamin D Trial (CaD)**: This double-blind component began 1 to 2 years after a woman joined one or both of the other clinical trial components. It evaluated the effect of calcium and vitamin D supplementation on the prevention of osteoporotic fractures and colorectal cancer. Women in this component were randomized to calcium (1000 mg/d) and vitamin D (400 IU/d) supplements or a matching placebo. The **Observational Study (OS)** examines the relationship between lifestyle, environmental, medical and molecular risk factors and specific measures of health or disease outcomes. This component involves tracking the medical history and health habits of 93,676 women not participating in the CT. Recruitment for the observational study was completed in 1998 and participants were followed annually for 8 to 12 years. All centrally confirmed cases of invasive colorectal cancers, or deaths from colorectal cancer were selected as potential cases from September 30, 2015 database. Controls were participants free of colorectal cancer (invasive or in situ) as of September 30, 2015. Potential cases and controls were excluded if they (1) were non-White; (2) had history of colorectal cancers at baseline; (3) lost to follow-up after enrollment; (4) dbGaP ineligible; (5) had <1.25ug of DNA; (6) selected for WHI study M26 Phase I or II; (7) selected for WHI study AS224 and also included in the imputation project. A total of 578 cases and 104,429 controls met the eligibility criteria. Each case was matched with 1 control (1:1) that exactly met the following matching criteria: age (±5 years), 40 randomization centers (exact), WHI date (±3 years), CaD date (±3 years), OS flag (exact), HRT assignments (exact), DM assignments (exact), and CaD assignments (exact). Control selection was done in a time-forward manner, selecting one control for each case 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 criteria to minimize an overall distance measure. Each matching factor was given the same weight. When exact matches could not be found, the matching criteria were gradually relaxed among unmatched cases and controls until all cases had found matched controls. Using the matching criteria specified above, 559 of the 578 eligible cases found exact matches. The matching criteria was then relaxed to: Age±5, randomization centers, WHI date ± 3 years, CaD date ± 3 years, OS flag, HRT flag, DM flag, CaD flag. 17 of the remaining 19 unmatched cases found matched controls. By matching on Age±5, randomization centers, WHI date ± 3 years, CaD date ± 3 years, OS flag, HRT flag, the remaining 2 unmatched cases found their matches. All subjects were genotyped using the OncoArray+custom iSelect array genotyping project.

3.2. GWAS genotype quality control, imputation, and principal component analysis

Table S1 provides details of genotyping platforms used in these studies. Details of genotyping and QC for studies included in the meta-analysis are described elsewhere [13.12,27] .

Principal component analysis

After excluding close relatives, we performed PCA using PLINK1.9 [28] on LD-pruned sets of autosomal SNPs obtained by removing regions with extensive long-range LD[29,30], SNPs with MAF < 5%, HWE $P < 1 \times 10^{-4}$, or any genotype missingness, and carrying out LD pruning using the PLINK option "-indep-pairwise 50 5 0.2." To identify population outliers we merged in 1,092 individuals from 1000 Genomes Project Phase III and performed PCA using the intersection of variants[31].

Genotype imputation

To improve imputation accuracy we performed phasing and imputation separately for each genotyping project data set and imputed to the Haplotype Reference Consortium (HRC)) panel (39.2 million variants) using the University of Michigan Imputation Server[32].

3.3. Statistical analyses

We analyzed each dataset separately (Table S2), and combined association summary statistics across analyses via fixed-effects inverse variance–weighted meta-analysis using METAL[33]. Within each data set, variants with an imputation accuracy $r^2 \geq 0.3$ and MAC ≥ 50 were tested for CRC association using the imputed genotype dosage in a logistic regression model adjusted for age, sex and study or genotyping project–specific covariates, including principal components to adjust for population structure. To account for residual confounding within CORSA, we tested association with each variant using a linear mixed model and kinship matrix calculated from the data, as implemented in EMMAX[34]. To enable meta-analysis, we then calculated approximate allelic log[OR] estimates and corresponding standard errors as described in Cook et al.[35]

4. Evaluation Data Sets

Genetic Epidemiology Research on Adult Health and Aging (GERA)

The Genetic Epidemiology Research on Adult Health and Aging (GERA) resource is a cohort of more than 100,000 subjects who are participants in the Kaiser Permanente Medical Care Plan, Northern California Region (KPNC), Research Program on Genes, Environment and Health (RPGEH) Genome-wide genotyping was targeted for this cohort to enable large-scale genome-wide association studies by linkage to comprehensive longitudinal clinical data derived from extensive KPNC electronic health record databases. The cohort is multi-ethnic, with ∼20% minority representation (African American, East Asian, and Latino or mixed), and the remaining 80% non-Hispanic white. For this project, four ethnic-

specific arrays were designed based on the Affymetrix Axiom Genotyping System. Imputation was performed using 1000 Genomes data on an array-wise basis.

Study population

The GERA cohort, is an independent contemporary cohort, nested within the Kaiser Permanente Northern California integrated healthcare delivery system[36]. Construction of this cohort began in 2007, when a six-page survey was mailed to 1.9 million individuals, ages 18 and older, who had been previously enrolled health plan members for at least two years. This survey was designed to ascertain data on demographic and lifestyle characteristics, including race/ethnicity, education, income, marital status, self- and family history for 35 selected conditions, diet and physical activity, smoking and alcohol consumption, and reproductive history and health. In July 2008, approximately 400,000 survey respondents were asked to sign a consent form authorizing broad use of their survey data, longitudinal electronic health record data, and biospecimens in conducting research on genetic and environmental factors associated with health and disease. Those who provided consent were mailed saliva DNA collection (Oragene) kits. In 2009, over 40,000 men ages 45 to 69 years, who were KPNC health plan members and had enrolled in the California Men's Health Study (CMHS) in 2002–2003, were similarly asked to provide saliva samples and added to expand the GERA cohort. At CMHS enrollment, men completed mailed surveys to ascertain data on demographic and lifestyle factors, akin to that of GERA.

In total, 110,266 consenting participants who provided saliva samples were selected for genotyping. All racial and ethnic minority participants with saliva samples (n = 20,935, 19%) were included to maximize diversity. Four custom Affymetrix Axiom arrays were designed for genotyping, one for each major ancestral group represented in the GERA cohort: European, African, East Asian, and Latino. As detailed elsewhere[37], the selected number of SNPs and SNP content varied by array in order to maximize coverage of the whole genome, along with common and low frequency SNPs specific to race/ethnicity and known SNPs associated with disease phenotypes. Details on the calling and quality control have been described previously[38].

A personal history of cancer was determined from cancer registry and electronic health record data, medical record coding and electronic pathology report coding. A family history of CRC was ascertained through a baseline study questionnaire, diagnostics codes, and the family history table of our electronic health record.by integrating data from baseline surveys and electronic health records (i.e., diagnosis codes, family history documentation). Hyperplastic polyps, advanced adenomas, and non-advanced adenomas were identified using Systematized Nomenclature of

Medicine (SNOMED) pathology codes and a validated natural language processing tool[39]. We defined an advanced adenoma as any adenoma with villous histology or which was 10 mm in size or greater. All study participants provided written informed consent, and the study was approved by the Kaiser Permanente Northern California Institutional Review Board. As a cohort unselected on any disease phenotype, GERA participants were not asked to engage in specific medical or screening tests for research purposes. Therefore, although the majority (69%) of GERA participants were age 55 and older at baseline, most but not all have undergone screening for colorectal cancer, either by fecal immunochemical testing (FIT) or endoscopy (sigmoidoscopy or colonoscopy). At their baseline questionnaire, 70% were up to date for colorectal cancer screening.

Colorectal cancer status was determined for study participants from their initiation of Kaiser health plan membership by linkage to the KPNC Cancer Registry, a current database of all patients with newly diagnosed cancers at KPNC facilities that adheres to the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) Program standards. The observed time was defined from the age of initial KPNC enrollment to the earliest of age at CRC diagnosis, advanced adenoma, death or end of follow-up (the GERA cohort was followed until December 31, 2016) and the CRC incidence is measured from 2007 to 2015.

5 Validation Data Set

Electronic Medical Records and Genomics (eMERGE)

The eMERGE network has been developing a unified genome-wide single-nucleotide variant (SNV) genotype array-based association platform for analysis of electronic medical record (EMR)-derived phenotypes for approximately 10 years [40–42] . In the first phase, eMERGE 1, discovery efforts were based on the Illumina 660k genotype array with ~20,000 participants being enrolled through five medical centers. In eMERGE 2 ~30,000 more individuals with high-density genotype data were ascertained resulting in analyses with ~50,000 individuals. In eMERGE 3, genotype and clinical EMR data of ~33,000 additional participants have been added to the resources available for analysis. The case control status for CRC was defined by an algorithm based on ICD9 codes 153 -153.9, 154 - 154.2,154.8 and ICD10 codes C18-C18.9, C19, C20, C21-C21.2, C21.8. .

To implement the minimac3 missing genotype variant imputation statistical model, the MIS guidelines[32,43,44] has been followed and imputed each genotype array batch independently. The MIS used the HRC1.1 variation reference in genome build 37 (hg19) coordinates to impute the missing variants across samples in genotype array batches of up to 15,000 samples**.** The hard genotype call merged sample set of unique imputed samples was analyzed by PCA using the plink2–pca approx fast pca method for large sample sizes and participant groupings were compared to the self-reported or observed-reported ancestry. PCA is performed on the 83,717 participant multisample with variants MAF > 5%, variant missingness of 0.1, and LD-pruned to an *R*-squared threshold of 0.7. After removal of low-call-rate samples (>2% missingness) and duplicated samples, the data set resulted in 83,717 unique imputed participants based on the eMERGE subject IDs from 77 imputation batches [45].

6 Additional Analysis

6.1 Additional Analysis of Approach 2 Feature Selection and Model Development using Training Data

Using the entire data set of 120,000 samples, we performed feature selection (marginal association) and model development for ridge, lasso, elastic net regression, and xgboost. The penalty parameter was chosen based on 10-fold cross validation. Models were developed in a progressive manner, first only 140 known loci, then 140 known loci + top 1000 top variants based on marginal association, + 5000 top variants, +10,000 variants, so on and so forth. The results based on GERA validation cohort are summarized in Table S9.

We noted several observations. First, the AUC of the model does not improve as we include more variants. In fact, there is a steady decline as we include more variants. Second, these top SNPs are not pruned based on linkage disequilibrium. As it can be seen here, for top 10,000 SNPs without LD pruning, ridge regression does not perform as well as ridge regression with LD pruning. This can perhaps be explained by the fact that SNPs in strong LD contribute very similar information. It is true penalized regression can handle this situation in principle; however, when the number of predictors is very high, and the signal-to-noise ratio is low, penalized regression cannot discern the signals from noise well. This makes it very important to select (independent) variants that most likely contribute to CRC prediction.

6.2 Model Calibration

We assessed model calibration in the eMERGE study, for which we obtained the hazard ratio and baseline hazard function estimates from GERA for the LDPred model with rho=0.003, which was the best performing model. Based on these estimates, we predicted the risk of developing colorectal cancer (CRC) during the follow-up time for each individual in eMERGE given the PRS. Table S12 shows the observed number of CRC cases, expected number, ratio of expected and observed cases with 95% confidence interval stratified by 10 equally sized groups of PRS in eMERGE. There is an increasing trend for the observed and expected number of CRC cases with PRS. However, the model consistently under-estimates the number of cases. A closer examination of the probabilities of developing CRC for eMERGE and GERA shows that eMERGE has a higher overall disease probability (Figure S4). There might be several reasons contributing to this discrepancy such as a higher CRC screening rate in GERA than the general population, CRC case definition where some in-situ CRC may be included in eMERGE, or healthy volunteer effect in GERA.

7. Funding and Acknowledgements

Funding:
Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO): National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services (U01 CA164930, U01 CA137088, R01 CA059045, R01 CA195789, R01 CA224097).
Genotyping/Sequencing services were provided by the Center for Inherited Disease Research (CIDR) (X01-HG008596 and X-01-HG007585). CIDR is fully funded through a federal contract from the National Institutes of Health to The Johns Hopkins University, contract number HHSN268201200008I. This research was funded in part through the NIH/NCI Cancer Center Support Grant P30 CA015704.

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

The ATBC Study is supported by the Intramural Research Program of the U.S. National Cancer Institute, National Institutes of Health, and by U.S. Public Health Service contract HHSN261201500005C from the National Cancer Institute, Department of Health and Human Services.

CLUE II: Funding to support CLUE II, for studies on colorectal cancer in the cohort, and for the investigators were from grants from the National Cancer Institute (U01 CA86308, Early Detection Research Network; P30 CA006973), National Institute on Aging (U01 AG18033), and the American Institute for Cancer Research. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US government.

COLO2&3: National Institutes of Health (R01 CA60987).

ColoCare: This work was supported by the National Institutes of Health (grant numbers R01 CA189184 (Li/Ulrich), U01 CA206110 (Ulrich/Li/Siegel/Figueireido/Colditz, 2P30CA015704- 40 (Gilliland), R01 CA207371 (Ulrich/Li)), the Matthias Lackas-Foundation, the German Consortium for Translational Cancer Research, and the EU TRANSCAN initiative.

The Colon Cancer Family Registry (CFR) Illumina GWAS was supported by funding from the National Cancer Institute, National Institutes of Health (grant numbers U01 CA122839, R01 CA143247). The Colon CFR/CORECT Affymetrix Axiom GWAS and OncoArray GWAS were supported by funding from National Cancer Institute, National Institutes of Health (grant number U19 CA148107 to S Gruber). The Colon CFR participant recruitment and collection of data and biospecimens used in this study were supported by the National Cancer Institute, National Institutes of Health (grant number UM1 CA167551) and through cooperative agreements with the following Colon CFR centers: Australasian Colorectal Cancer Family Registry (NCI/NIH grant numbers U01 CA074778 and U01/U24 CA097735), USC Consortium Colorectal Cancer Family Registry (NCI/NIH grant numbers U01/U24 CA074799), Mayo Clinic Cooperative Family Registry for Colon Cancer Studies (NCI/NIH grant number U01/U24 CA074800), Ontario Familial Colorectal Cancer Registry (NCI/NIH grant number U01/U24 CA074783), Seattle Colorectal Cancer Family Registry (NCI/NIH grant number U01/U24 CA074794), and University of Hawaii Colorectal Cancer Family Registry (NCI/NIH grant number U01/U24 CA074806), Additional support for case ascertainment was provided from the Surveillance, Epidemiology

and End Results (SEER) Program of the National Cancer Institute to Fred Hutchinson Cancer Research Center (Control Nos. N01-CN-67009 and N01-PC-35142, and Contract No. HHSN2612013000121), the Hawai'i Department of Health (Control Nos. N01-PC- 67001 and N01-PC-35137, and Contract No. HHSN26120100037C, and the California Department of Public Health (contracts HHSN261201000035C awarded to the University of Southern California, and the following state cancer registries: AZ, CO, MN, NC, NH, and by the Victoria Cancer Registry and Ontario Cancer Registry.

COLON: The COLON study is sponsored by Wereld Kanker Onderzoek Fonds, including funds from grant 2014/1179 as part of the World Cancer Research Fund International Regular Grant Programme, by Alpe d'Huzes and the Dutch Cancer Society (UM 2012–5653, UW 2013-5927, UW2015-7946), and by TRANSCAN (JTC2012-MetaboCCC, JTC2013-FOCUS). The NQplus study is sponsored by a ZonMW investment grant (98-10030); by PREVIEW, the project PREVention of diabetes through lifestyle intervention and population studies in Europe and around the World (PREVIEW) project which received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant no. 312057; by funds from TI Food and Nutrition (cardiovascular health theme), a public–private partnership on precompetitive research in food and nutrition; and by FOODBALL, the Food Biomarker Alliance, a project from JPI Healthy Diet for a Healthy Life.

Colorectal Cancer Transdisciplinary (CORECT) Study: The CORECT Study was supported by the National Cancer Institute, National Institutes of Health (NCI/NIH), U.S. Department of Health and Human Services (grant numbers U19 CA148107, R01 CA81488, P30 CA014089, R01 CA197350,; P01 CA196569; R01 CA201407) and National Institutes of Environmental Health Sciences, National Institutes of Health (grant number T32 ES013678).

CORSA: "Österreichische Nationalbank Jubiläumsfondsprojekt" (12511) and Austrian Research Funding Agency (FFG) grant 829675.

CPS-II: The American Cancer Society funds the creation, maintenance, and updating of the Cancer Prevention Study-II (CPS-II) cohort. This study was conducted with Institutional Review Board approval.

CRCGEN: Colorectal Cancer Genetics & Genomics, Spanish study was supported by Instituto de Salud Carlos III, co-funded by FEDER funds –a way to build Europe– (grants PI14-613 and PI09-1286), Agency for Management of University and Research Grants (AGAUR) of the Catalan Government (grant 2017SGR723), and Junta de Castilla y León (grant LE22A10-2). Sample collection of this work was supported by the Xarxa de Bancs de Tumors de Catalunya sponsored by Pla Director d'Oncología de Catalunya (XBTC), Plataforma Biobancos PT13/0010/0013 and ICOBIOBANC, sponsored by the Catalan Institute of Oncology.

Czech Republic CCS: This work was supported by the Czech Science Foundation (grants: 17-16857S, 18-09709S) and by the Czech Health Research Council Ministry of Health, Czech Rep. (grants 15-27580A, 17-30920A and NV18/03/00199).

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

DALS: National Institutes of Health (R01 CA48998 to M. L. Slattery).

EDRN: This work is funded and supported by the NCI, EDRN Grant (U01 CA 84968-06).

eMERGE: The eMERGE Network was initiated and funded by NHGRI through the following grants:

Phase III: U01HG008657 (Group Health Cooperative/University of Washington); U01HG008685 (Brigham and Women's Hospital); U01HG008672 (Vanderbilt University Medical Center); U01HG008666 (Cincinnati Children's Hospital Medical Center); U01HG006379 (Mayo Clinic); U01HG008679 (Geisinger Clinic); U01HG008680 (Columbia University Health Sciences); U01HG008684 (Children's Hospital of Philadelphia); U01HG008673 (Northwestern University); U01HG008701 (Vanderbilt University Medical Center serving as the Coordinating Center); U01HG008676 (Partners Healthcare/Broad Institute); and U01HG008664 (Baylor College of Medicine)

Phase II: U01HG006828 (Cincinnati Children's Hospital Medical Center/Boston Children's Hospital); U01HG006830 (Children's Hospital of Philadelphia); U01HG006389 (Essentia Institute of Rural Health, Marshfield Clinic Research Foundation and Pennsylvania State University); U01HG006382 (Geisinger Clinic); U01HG006375 (Group Health Cooperative/University of Washington); U01HG006379 (Mayo Clinic); U01HG006380 (Icahn School of Medicine at Mount Sinai); U01HG006388 (Northwestern University); U01HG006378 (Vanderbilt University Medical Center); and U01HG006385 (Vanderbilt University Medical Center serving as the Coordinating Center).

If the project includes data from the eMERGE imputed merged Phase I and Phase II dataset, please also add U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers. and/or The PGRNSeq dataset (eMERGE PGx), please also add U01HG004438 (CIDR) serving as a Sequencing Center.

Phase I: U01-HG-004610 (Group Health Cooperative/University of Washington); U01-HG-004608 (Marshfield Clinic Research Foundation and Vanderbilt University Medical Center); U01-HG-04599 (Mayo Clinic); U01HG004609 (Northwestern University); U01-HG-04603 (Vanderbilt University Medical Center, also serving as the Administrative Coordinating Center); U01HG004438 (CIDR) and U01HG004424 (the Broad Institute) serving as Genotyping Centers.

EPIC: The coordination of EPIC is financially supported by the European Commission (DGSANCO) and the International Agency for Research on Cancer. The national cohorts are supported by Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); German Cancer Aid, German Cancer Research Center (DKFZ), Federal Ministry of Education and Research (BMBF), Deutsche Krebshilfe, Deutsches Krebsforschungszentrum and Federal Ministry of Education and Research (Germany); the Hellenic Health Foundation (Greece); Associazione Italiana per la Ricerca sul Cancro-AIRCItaly and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), Statistics Netherlands (The Netherlands); ERC-2009-AdG 232997 and Nordforsk, Nordic Centre of Excellence programme on Food, Nutrition and Health (Norway); Health Research Fund (FIS), PI13/00061 to Granada, PI13/01162 to EPIC-Murcia, Regional Governments of Andalucía, Asturias, Basque Country, Murcia and Navarra, ISCIII RETIC (RD06/0020) (Spain); Swedish Cancer Society, Swedish Research Council and County Councils of Skåne and Västerbotten (Sweden); Cancer Research UK (14136 to EPIC-Norfolk;

C570/A16491 and C8221/A19170 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk, MR/M012190/1 to EPICOxford) (United Kingdom).

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

EPICOLON: This work was supported by grants from Fondo de Investigación Sanitaria/FEDER (PI08/0024, PI08/1276, PS09/02368, P111/00219, PI11/00681, PI14/00173, PI14/00230, PI17/00509, 17/00878, Acción Transversal de Cáncer), Xunta de Galicia (PGIDIT07PXIB9101209PR), Ministerio de Economia y Competitividad (SAF07-64873, SAF 2010-19273, SAF2014-54453R), Fundación Científica de la Asociación Española contra el Cáncer (GCB13131592CAST), Beca Grupo de Trabajo "Oncología" AEG (Asociación Española de Gastroenterología), Fundación Privada Olga Torres, FP7 CHIBCHA Consortium, Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR, Generalitat de Catalunya, 2014SGR135, 2014SGR255, 2017SGR21, 2017SGR653), Catalan Tumour Bank Network (Pla Director d'Oncologia, Generalitat de Catalunya), PERIS (SLT002/16/00398, Generalitat de Catalunya), CERCA Programme (Generalitat de Catalunya) and COST Action BM1206 and CA17118. CIBERehd is funded by the Instituto de Salud Carlos III.

ESTHER/VERDI. This work was supported by grants from the Baden-Württemberg Ministry of Science, Research and Arts and the German Cancer Aid.

Harvard cohorts (HPFS, NHS, PHS): HPFS is supported by the National Institutes of Health ((P01 CA055075, UM1 CA167552, U01 CA167552, R01 CA137178, R01 CA151993 and R35 CA197735), NHS by the National Institutes of Health (R01 CA137178, P01 CA087969, M1 CA186107, R01 CA151993 and R35 CA197735) and PHS by the National Institutes of Health (R01 CA042182). Hawaii Adenoma Study: NCI grants R01 CA72520.

Kentucky: This work was supported by the following grant support: Clinical Investigator Award from Damon Runyon Cancer Research Foundation (CI-8); NCI R01CA136726.

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

MCCS cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further augmented by Australian National Health and Medical Research Council grants 209057, 396414 and 1074383 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry and the Australian Institute of Health and Welfare, including the National Death Index and the Australian Cancer Database.

MEC: National Institutes of Health (U01 CA164973, R37 CA54281, P01 CA033619, and R01 CA063464).

MECC: This work was supported by the National Institutes of Health, U.S. Department of Health and Human Services (R01 CA81488

to SBG and GR).

MSKCC: The work at Sloan Kettering in New York was supported by the Robert and Kate Niehaus Center for Inherited Cancer Genomics and the Romeo Milio Foundation. Moffitt: This work was supported by funding from the National Institutes of Health (grant numbers R01 CA189184, P30 CA076292), Florida Department of Health Bankhead-Coley Grant 09BN-13, and the University of South Florida Oehler Foundation. Moffitt contributions were supported in part by the Total Cancer Care Initiative, Collaborative Data Services Core, and Tissue Core at the H. Lee Moffitt Cancer Center & Research Institute, a National Cancer Institute-designated Comprehensive Cancer Center (grant number P30 CA076292).

NCCCS I & II: We acknowledge funding support for this project from the National Institutes of Health, R01 CA66635 and P30 DK034987.

NFCCR: This work was supported by an Interdisciplinary Health Research Team award from the Canadian Institutes of Health Research (CRT 43821); the National Institutes of Health, U.S. Department of Health and Human Serivces (U01 CA74783); and National Cancer Institute of Canada grants (18223 and 18226). The authors wish to acknowledge the contribution of Alexandre Belisle and the genotyping team of the McGill University and Génome Québec Innovation Centre, Montréal, Canada, for genotyping the Sequenom panel in the NFCCR samples. Funding was provided to Michael O. Woods by the Canadian Cancer Society Research Institute.

NSHDS: Research in this study was supported in part by Swedish Cancer Society; Swedish Research Council; Faculty of Medicine, Umeå University, Umeå, Sweden; and Region Västerbotten Cutting-Edge Research Grant from the County Council of Västerbotten, Sweden.

OFCCR: National Institutes of Health, through funding allocated to the Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA074783); see CCFR section above. Additional funding toward genetic analyses of OFCCR includes the Ontario Research Fund, the Canadian Institutes of Health Research, and the Ontario Institute for Cancer Research, through generous support from the Ontario Ministry of Research and Innovation.

OSUMC: OCCPI funding was provided by Pelotonia and HNPCC funding was provided by the NCI (CA16058 and CA67941).

PLCO: Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH,

PMH: National Institutes of Health (R01 CA076366 to P.A. Newcomb).

RPGEH: Data used in this study were generated by the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH): Genetic Epidemiology Research on Adult Health and Aging (GERA), funded by the National Institutes of Health [RC2 AG036607 (Schaefer and Risch)], the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, The Ellison Medical Foundation, and the Kaiser Permanente Community Benefits Program. Access to RPGEH data used in this study may be obtained by application to the Kaiser Permanente Research Bank (KPRB) via ResearchBankAcccss@kp.org. A subset of the GERA cohort consented for public use can be found at NIH/dbGaP: phs000674.vl.pl. The work reported in this publication was also supported in part by the

National Cancer Institute (K07 CA212057 to J.K. Lee and K07 CA188142 to L.C. Sakoda).

SEARCH: The University of Cambridge has received salary support in respect of PDPP from the NHS in the East of England through the Clinical Academic Reserve. Cancer Research UK (C490/A16561); the UK National Institute for Health Research Biomedical Research Centres at the University of Cambridge.

SELECT: Research reported in this publication was supported in part by the National Cancer Institute of the National Institutes of Health under Award Numbers U10 CA37429 (CD Blanke), and UM1 CA182883 (CM Tangen/IM Thompson). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

SMS: This work was supported by the National Cancer Institute (grant P01 CA074184 to J.D.P. and P.A.N., grants R01 CA097325, R03 CA153323, and K05 CA152715 to P.A.N., and the
National Center for Advancing Translational Sciences at the National Institutes of Health (grant KL2 TR000421 to A.N.B.-H.)

The Swedish Low-risk Colorectal Cancer Study: The study was supported by grants from the Swedish research council; K2015-55X-22674-01-4, K2008-55X-20157-03-3, K2006-72X-
20157-01-2 and the Stockholm County Council (ALF project).

Swedish Mammography Cohort and Cohort of Swedish Men: This work is supported by the Swedish Research Council /Infrastructure grant, the Swedish Cancer Foundation, and the Karolinska Institute´s Distinguished Professor Award to Alicja Wolk.

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

VITAL: National Institutes of Health (K05 CA154337).

WHI: The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C.

Acknowledgements:

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

CLUE II: We appreciate the continued efforts of the staff members at the Johns Hopkins George W. Comstock Center for Public Health Research and Prevention in the conduct of the CLUE II study. Cancer incidence data for CLUE were provided by the Maryland Cancer Registry, Center for Cancer Surveillance and Control, Maryland Department of Health, 201 W. Preston Street, Room 400, Baltimore, MD 21201, http://phpa.dhmh.maryland.gov/cancer, 410-767-4055. We acknowledge the State of Maryland, the Maryland Cigarette Restitution

Fund, and the National Program of Cancer Registries of the Centers for Disease Control and Prevention for the funds that support the collection and availability of the cancer registry data.

COLON and NQplus: the authors would like to thank the COLON and NQplus investigators at Wageningen University & Research and the involved clinicians in the participating hospitals.

CORSA: We kindly thank all those who contributed to the screening project Burgenland against CRC. Furthermore, we are grateful to Doris Mejri and Monika Hunjadi for laboratory assistance.

CPS-II: The authors thank the CPS-II participants and Study Management Group for their invaluable contributions to this research. The authors would also like to acknowledge the contribution to this study from central cancer registries supported through the Centers for Disease Control and Prevention National Program of Cancer Registries, and cancer registries supported by the National Cancer Institute Surveillance Epidemiology and End Results program.

Czech Republic CCS: We are thankful to all clinicians in major hospitals in the Czech Republic, without whom the study would not be practicable. We are also sincerely grateful to all patients participating in this study.

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

EDRN: We acknowledge all the following contributors to the development of the resource: University of Pittsburgh School of Medicine, Department of Gastroenterology, Hepatology and Nutrition: Lynda Dzubinski; University of Pittsburgh School of Medicine, Department of Pathology: Michelle Bisceglia; and University of Pittsburgh School of Medicine, Department of Biomedical Informatics.

EPICOLON: We are sincerely grateful to all patients participating in this study who were recruited as part of the EPICOLON project. We acknowledge the Spanish National DNA Bank, the Barcelona CRC screening program, the Biobank core facility of Hospital Clínic–IDIBAPS and the Biobanco Vasco para la Investigación/O+ehun-Hospital Donostia for the availability of the samples. The work was carried out (in part) at the Esther Koplowitz Centre, Barcelona.

Harvard cohorts (HPFS, NHS, PHS): The study protocol was approved by the institutional review boards of the Brigham and Women's Hospital and Harvard T.H. Chan School of Public Health, and those of participating registries as required. We would like to thank the participants and staff of the HPFS, NHS and PHS for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. The authors assume full responsibility for analyses and interpretation of these data. Kentucky: We would like to acknowledge the staff at the Kentucky Cancer Registry.

LCCS: We acknowledge the contributions of Jennifer Barrett, Robin Waxman, Gillian Smith and Emma Northwood in conducting this study.

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

NSHDS: We thank all participants in the NSHDS cohorts and the staff at the Department of Biobank Research, Umeå University, as well as Kerstin Näslund, formerly of the Department of Medical Biosciences, and Inger Cullman, Department of Chemistry, both at Umeå University for excellent technical assistance.

PLCO: The authors thank the PLCO Cancer Screening Trial screening center investigators and the staff from Information Management Services Inc and Westat Inc. Most importantly, we thank the study participants for their contributions that made this study possible.

PMH: The authors would like to thank the study participants and staff of the Hormones and Colon Cancer study.

RPGEH: We thank the RPGEH study team and participants for their contributions to support this research.

SEARCH: We thank the SEARCH team.

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

WHI: The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf

8. References

1. Brenner, H., Chang-Claude, J., Jansen, L., Knebel, P., Stock, C., and Hoffmeister, M. (2014). Reduced risk of colorectal cancer up to 10 years after screening, surveillance, or diagnostic colonoscopy. Gastroenterology *146*, 709–717. DOI: 10.1053/j.gastro.2013.09.001

2. Jia, M., Jansen, L., Walter, V., Tagscherer, K., Roth, W., Herpel, E., Kloor, M., Bläker, H., Chang-Claude, J., Brenner, H., et al. (2016). No association of CpG island methylator phenotype and colorectal cancer survival: population-based study. Br. J. Cancer *115*, 1359–1366. DOI: 10.1038/bjc.2016.361

3. Kakourou, A., Koutsioumpa, C., Lopez, D.S., Hoffman-Bolton, J., Bradwin, G., Rifai, N., Helzlsouer, K.J., Platz, E.A., and Tsilidis, K.K. (2015). Interleukin-6 and risk of colorectal cancer: results from the CLUE II cohort and a meta-analysis of prospective studies. Cancer Causes Control *26*, 1449–1460. DOI: 10.1007/s10552-015-0641-1

4. Hofer, P., Baierl, A., Feik, E., Führlinger, G., Leeb, G., Mach, K., Holzmann, K., Micksche, M., and Gsur, A. (2011). MNS16A tandem repeats minisatellite of human telomerase gene: a risk factor for colorectal cancer. Carcinogenesis *32*, 866–871. DOI: 10.1093/carcin/bgr053

5. Calle, E.E., Rodriguez, C., Jacobs, E.J., Almon, M.L., Chao, A., McCullough, M.L., Feigelson, H.S., and Thun, M.J. (2002). The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. Cancer *94*, 2490–2501. DOI: 10.1002/cncr.101970

6. Campbell, P.T., Deka, A., Briggs, P., Cicek, M., Farris, A.B., Gaudet, M.M., Jacobs, E.J., Newton, C.C., Patel, A.V., Teras, L.R., et al. (2014). Establishment of the cancer prevention study II nutrition cohort colorectal tissue repository. Cancer Epidemiol. Biomarkers Prev. *23*, 2694–2702. DOI: 10.1158/1055-9965.EPI-14-0541

7. Vymetalkova, V., Pardini, B., Rosa, F., Di Gaetano, C., Novotny, J., Levy, M., Buchler, T., Slyskova, J., Vodickova, L., Naccarati, A., et al. (2014). Variations in mismatch repair genes and colorectal cancer risk and clinical outcome. Mutagenesis *29*, 259–265. DOI: 10.1093/mutage/geu014

8. Naccarati, A., Rosa, F., Vymetalkova, V., Barone, E., Jiraskova, K., Di Gaetano, C., Novotny, J., Levy, M., Vodickova, L., Gemignani, F., et al. (2016). Double-strand break repair and colorectal cancer: gene variants within 3' UTRs and microRNAs binding as modulators of cancer risk and clinical outcome. Oncotarget *7*, 23156–23169. DOI: 10.18632/oncotarget.6804

9. Vymetalkova, V., Pardini, B., Rosa, F., Jiraskova, K., Di Gaetano, C., Bendova, P., Levy, M., Veskrnova, V., Buchler, T., Vodickova, L., et al. (2017). Polymorphisms in microRNA binding sites of mucin genes as predictors of clinical outcome in colorectal cancer patients. Carcinogenesis *38*, 28–39. DOI: 10.1093/carcin/bgw114,

10. Amin, W., Singh, H., Dzubinski, L.A., Schoen, R.E., and Parwani, A.V. (2010). Design and utilization of the colorectal and pancreatic neoplasm virtual biorepository: An early detection research network initiative. J. Pathol. Inform. *1*, 22. DOI: 10.4103/2153-3539.70831

11. Fernandez-Rozadilla, C., Cazier, J.-B., Tomlinson, I.P., Carvajal-Carmona, L.G., Palles, C., Lamas, M.J., Baiget, M., López-Fernández, L.A., Brea-Fernández, A., Abulí, A., et al. (2013). A colorectal cancer genome-wide association study in a Spanish cohort identifies two variants associated with colorectal cancer risk at 1p33 and 8p12. BMC Genomics *14*, 55. DOI: 10.1186/1471-2164-14-55

12. Schumacher, F.R., Schmit, S.L., Jiao, S., Edlund, C.K., Wang, H., Zhang, B., Hsu, L., Huang, S.-C., Fischer, C.P., Harju, J.F., et al. (2015). Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nat. Commun. *6*, 7138. DOI: 10.1038/ncomms8138

13. Peters, U., Jiao, S., Schumacher, F.R., Hutter, C.M., Aragaki, A.K., Baron, J.A., Berndt, S.I., Bézieau, S., Brenner, H., Butterbach, K., et al. (2013). Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. Gastroenterology *144*, 799–807.e24. DOI: 10.1053/j.gastro.2012.12.020

14. Lightfoot, T.J., Barrett, J.H., Bishop, T., Northwood, E.L., Smith, G., Wilkie, M.J.V., Steele, R.J.C., Carey, F.A., Key, T.J., Wolf, R., et al. (2008). Methylene tetrahydrofolate reductase genotype modifies the chemopreventive effect of folate in colorectal adenoma, but not colorectal cancer. Cancer Epidemiol. Biomarkers Prev. *17*, 2421–2430. DOI: 10.1158/1055-9965.EPI-08-0058

15. Turner, F., Smith, G., Sachse, C., Lightfoot, T., Garner, R.C., Wolf, C.R., Forman, D., Bishop, D.T., and Barrett, J.H. (2004). Vegetable, fruit and meat consumption and potential risk modifying genes in relation to colorectal cancer. Int. J. Cancer *112*, 259–264. DOI: 10.1002/ijc.20404

16. Belanger, C.F., Hennekens, C.H., Rosner, B., and Speizer, F.E. (1978). The nurses' health study. Am. J. Nurs. *78*, 1039–1040. PMID: 248266

17. Dahlin, A.M., Palmqvist, R., Henriksson, M.L., Jacobsson, M., Eklöf, V., Rutegård, J., Oberg, A., and Van Guelpen, B.R. (2010). The role of the CpG island methylator phenotype in colorectal cancer prognosis depends on microsatellite instability screening status. Clin. Cancer Res. *16*, 1845–1855. DOI: 10.1158/1078-0432.CCR-09-2594

18. Myte, R., Gylling, B., Schneede, J., Ueland, P.M., Häggström, J., Hultdin, J., Hallmans, G., Johansson, I., Palmqvist, R., and Van Guelpen, B. (2016). Components of One-carbon Metabolism Other than Folate and Colorectal Cancer Risk. Epidemiology *27*, 787–796. DOI: 10.1097/EDE.0000000000000529

19. Hampel, H., Frankel, W.L., Martin, E., Arnold, M., Khanduja, K., Kuebler, P., Clendenning, M., Sotamaa, K., Prior, T., Westman, J.A., et al. (2008). Feasibility of screening for Lynch syndrome among patients with colorectal cancer. J. Clin. Oncol. *26*, 5783–5788. DOI: 10.1200/JCO.2008.17.5950

20. Huang, J., Mondul, A.M., Weinstein, S.J., Koutros, S., Derkach, A., Karoly, E., Sampson, J.N., Moore, S.C., Berndt, S.I., and Albanes, D. (2016). Serum metabolomic profiling of prostate cancer risk in the prostate, lung, colorectal, and ovarian cancer screening trial. Br. J. Cancer *115*, 1087–1095. DOI: 10.1038/bjc.2016.305

21. Yeager, M., Chatterjee, N., Ciampa, J., Jacobs, K.B., Gonzalez-Bosquet, J., Hayes, R.B., Kraft, P., Wacholder, S., Orr, N., Berndt, S., et al. (2009). Identification of a new prostate cancer susceptibility locus on chromosome 8q24. Nat. Genet. *41*, 1055–1057. DOI: 10.1038/ng.444

22. Landi, M.T., Chatterjee, N., Yu, K., Goldin, L.R., Goldstein, A.M., Rotunno, M., Mirabello, L., Jacobs, K., Wheeler, W., Yeager, M., et al. (2009). A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. Am. J. Hum. Genet. *85*, 679–691. DOI: 10.1016/j.ajhg.2009.09.012

23. Lippman, S.M., Klein, E.A., Goodman, P.J., Lucia, M.S., Thompson, I.M., Ford, L.G., Parnes, H.L., Minasian, L.M., Gaziano, J.M., Hartline, J.A., et al. (2009). Effect of selenium and vitamin E on risk of prostate cancer and other cancers: the Selenium and Vitamin E Cancer Prevention Trial (SELECT). JAMA *301*, 39–51. DOI: 10.1001/jama.2008.864

24. Burnett-Hartman, A.N., Newcomb, P.A., Hutter, C.M., Peters, U., Passarelli, M.N., Schwartz, M.R., Upton, M.P., Zhu, L.-C., Potter, J.D., and Makar, K.W. (2014). Variation in the association between colorectal cancer susceptibility loci and colorectal polyps by polyp type. Am. J. Epidemiol. *180*, 223–232. DOI: 10.1093/aje/kwu114,

25. Walter, S.D. (1977). Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. Am. J. Epidemiol. *105*, 387–397. DOI: 10.1093/oxfordjournals.aje.a112395

26. Gail, M., Williams, R., Byar, D.P., and Brown, C. (1976). How many controls? J. Chronic Dis. *29*, 723–731. DOI: 10.1016/0021-9681(76)90073-4

27. Schmit, S.L., Edlund, C.K., Schumacher, F.R., Gong, J., Harrison, T.A., Huyghe, J.R., Qu, C., Melas, M., Van Den Berg, D.J., Wang, H., et al. (2019). Novel common genetic susceptibility loci for colorectal cancer. J. Natl. Cancer Inst. *111*, 146–157. DOI: 10.1093/jnci/djy099

28. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7. DOI: 10.1186/s13742-015-0047-8

29. Price, A.L., Weale, M.E., Patterson, N., Myers, S.R., Need, A.C., Shianna, K.V., Ge, D., Rotter, J.I., Torres, E., Taylor, K.D., et al. (2008). Long-range LD can confound genome scans in admixed populations. Am. J. Hum. Genet. *83*, 132–135. DOI: 10.1016/j.ajhg.2008.06.005

30. Laurie, C.C., Doheny, K.F., Mirel, D.B., Pugh, E.W., Bierut, L.J., Bhangale, T., Boehm, F., Caporaso, N.E., Cornelis, M.C., Edenberg, H.J., et al. (2010). Quality control and quality assurance in genotypic data for genome-wide association studies. Genet. Epidemiol. *34*, 591–602. DOI: 10.1002/gepi.20516

31. 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65. DOI: 10.1038/nature11632

32. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A.E., Kwong, A., Vrieze, S.I., Chew, E.Y., Levy, S., McGue, M., et al. (2016). Next-generation genotype imputation service and methods. Nat. Genet. *48*, 1284–1287. DOI: 10.1038/ng.3656

33. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191. DOI: 10.1093/bioinformatics/btq340

34. Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S.-Y., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat. Genet. *42*, 348–354. DOI: 10.1038/ng.548

35. Cook, J.P., Mahajan, A., and Morris, A.P. (2017). Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. Eur. J. Hum. Genet. *25*, 240–245. DOI: 10.1038/ejhg.2016.150

36. Gordon, N.P. (2006). How does the adult Kaiser Permanente membership in Northern California compare with the larger community? Http://www. Dor. Kaiser. org/external/uploadedFiles ….

37. Hoffmann, T.J., Kvale, M.N., Hesselson, S.E., Zhan, Y., Aquino, C., Cao, Y., Cawley, S., Chung, E., Connell, S., Eshragh, J., et al. (2011). Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics *98*, 79–89. DOI: 10.1016/j.ygeno.2011.04.005

38. Kvale, M.N., Hesselson, S., Hoffmann, T.J., Cao, Y., Chan, D., Connell, S., Croen, L.A., Dispensa, B.P., Eshragh, J., Finn, A., et al. (2015). Genotyping informatics and quality control for 100,000 subjects in the genetic epidemiology research on adult health and aging (GERA) cohort. Genetics *200*, 1051–1060. DOI: 10.1534/genetics.115.178905

39. Lee, J.K., Jensen, C.D., Levin, T.R., Zauber, A.G., Doubeni, C.A., Zhao, W.K., and Corley, D.A. (2019). Accurate identification of colonoscopy quality and polyp findings using natural language processing. J. Clin. Gastroenterol. *53*, e25–e30. DOI: 10.1097/MCG.0000000000000929

40. Chisholm, R.L. (2013). At the Interface between Medical Informatics and Personalized Medicine: The eMERGE Network Experience. Healthc. Inform. Res. *19*, 67–68. DOI: 10.4258/hir.2013.19.2.67

41. Crawford, D.C., Crosslin, D.R., Tromp, G., Kullo, I.J., Kuivaniemi, H., Hayes, M.G., Denny, J.C., Bush, W.S., Haines, J.L., Roden, D.M., et al. (2014). eMERGEing progress in genomics-the first seven years. Front. Genet. *5*, 184. DOI: 10.3389/fgene.2014.00184

42. Gottesman, O., Kuivaniemi, H., Tromp, G., Faucett, W.A., Li, R., Manolio, T.A., Sanderson, S.C., Kannry, J., Zinberg, R., Basford, M.A., et al. (2013). The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genet. Med. *15*, 761–771. DOI: 10.1038/gim.2013.72

43. Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., A Reshef, Y., K Finucane, H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G.R., et al. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. Nat. Genet. *48*, 1443–1448. DOI: 10.1038/ng.3679

44. McCarty, C.A., Chisholm, R.L., Chute, C.G., Kullo, I.J., Jarvik, G.P., Larson, E.B., Li, R., Masys, D.R., Ritchie, M.D., Roden, D.M., et al. (2011). The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med. Genomics *4*, 13. DOI: 10.1186/1755-8794-4-13

45. Stanaway, I.B., Hall, T.O., Rosenthal, E.A., Palmer, M., Naranbhai, V., Knevel, R., Namjou-Khales, B., Carroll, R.J., Kiryluk, K., Gordon, A.S., et al. (2019). The eMERGE genotype set of 83,717 subjects imputed to ~40 million variants genome wide and association with the herpes zoster medical record phenotype. Genet. Epidemiol. *43*, 63–81. DOI: 10.1002/gepi.22167