SUPPLEMENTARY MATERIALS

Novel common genetic susceptibility loci for colorectal cancer (Schmit SL *et al*)

Supplementary Methods

Discovery

Part 1 (CORECT, CCFR, and GECCO GWAS)

Study populations, sample collection, genotyping, QC, imputation procedures, and statistical analysis have been described in detail previously [1, 2]. In brief, Part 1 (CORECT, CCFR, and GECCO) included a total of 18,299 CRC cases and 19,656 controls of European ancestral heritage from 19 observational studies. Genotype data was generated from germline DNA on the Affymetrix Axiom CORECT Set (CORECT); Illumina 1M, 1M-Duo, or Omni1 (CCFR); Illumina 300K, Illumina OmniExpress, Illumina 550K/610K, or Affymetrix 100K/500K (GECCO); and Illumina Omni 2.5 (Molecular Epidemiology of Colorectal Cancer Study (MECC)). High-density genotype array data was cleaned by applying standard QC filters at both the individual subject and single nucleotide polymorphism (SNP) levels. Quality-controlled genotype data was imputed to the 1,000 Genomes Project Phase 1 multiethnic reference panel (March 2012 release, n=1,092) [3, 4] using SHAPE-IT/IMPUTE2[5, 6] (CORECT, CCFR, MECC) or BEAGLE/Minimac[6, 7] (GECCO). Genotype data for cases and controls were phased and imputed together by array platform thus avoiding any potentially differential imputation error between cases and controls. Stringent imputation quality (info≥0.7, certainty≥0.9, concordance≥0.9 (CORECT, CCFR); Rsq>0.3(GECCO)) and minor allele frequency filters (MAF≥1%) were imposed on variants prior to the analysis phase. Study-specific logistic regression summary statistics were combined using an inverse variance weighted fixed effects meta-analysis with METAL [8].

Part 2 (OncoArray)

Study Populations

Detailed descriptions of studies contributing to the OncoArray Project are included at the end of the Supplementary Methods. Briefly, European-descent participants from 20 observational studies and three clinical trials contributed germline DNA. The characteristics of Europeans (N_{case}=18,649; $N_{control}$=11,208) genotyped on the OncoArray and included in the discovery GWAS are described in Supplementary Table 2. Of note, DNA samples from East Asian descent participants from 4 observational studies were also included in the OncoArray Project (N_{case}=3,855; $N_{control}$=3,081) and underwent genotyping, QC, imputation, and statistical analysis alongside the European ancestry samples (Supplementary Table 4). Genotype data from these East Asian participants are included in the multiethnic follow-up stage described below, but for the purpose of clarity, the methods for genotyping, QC, and data handling are described here.

Genotyping and QC

Genotyping was conducted in five batches across two centers using a common protocol and array: the Center for Inherited Disease Research (CIDR; batches 1-4) and the University of

Southern California Molecular Genomics Core Facility (batch 5). Batches were determined by timing of sample availability from contributing studies, and within batches, cases and controls per study (where both were contributed) were randomly distributed. Raw data (total SNPs/indels, N_{marker}=533,631; N_{sample}=42,270) were imported into Illumina GenomeStudio, and genotypes were called using cluster definitions provided by the OncoArray Consortium. Standard QC filters at both the individual subject and SNP levels were applied. A first round of filtering excluded samples with <80% call rate (N_{sample}=764), then variants with <80% call rate (N_{marker}=7,394). Next, samples with <95% call rate (N_{sample}=1,078) were excluded as well as those marked for removal from various QC checks such as replicate concordance (within and across platforms), unexpected replicate search (within and across platforms), genotyped vs. reported sex concordance, plate mix-ups, and removal due to lack of consent (N_{sample}=978). Markers were then excluded based on the following criteria: 1) <95% call rate (N_{marker}=10,344); 2) duplicate error rate >1% (only in matching reps with call rate >=99%) or heterozygote duplicate error rate >5% and >2 het mismatches (N_{marker}=943), and 3) SNPs with duplicate probes (N_{marker}=778). Lower call rate sample replicates were also removed (N_{sample}=1,565).

Next, we ran STRUCTURE[9] with HapMap3 European, Asian, and African samples as anchors to identify European and East Asian samples within our dataset. This allowed us to exclude markers with low HWE separately in Europeans and East Asians (a marker was flagged if $P<10^{-7}$ in controls or $P<10^{-12}$ in cases) (N_{marker}=3,318). Additional sample level exclusions were applied: 1) one XXY sample, 2) one individual later determined not to have cancer, and 3) two invalid MECC samples (cancelled participation or withdrew consent). OncoArray Consortium-wide failed SNPs/indels were removed (N_{marker}=16,565). Genotype data from additional controls from the SEARCH breast cancer study (N_{sample}=2,672) were added at this stage, and only overlapping markers and non-identical samples with existing data were retained (N_{marker}=278 and N_{sample}=135 removed). After QC, the final N_{marker} was 474,697.

Prior to statistical analysis, additional samples were excluded for the following reasons: 1) relatedness (removed one from each full-sibling and parent-offspring pair, N_{sample}=724); 2) identical to a previous GWAS study participant including CORECT, CCFR and GECCO Phase 1 and HCCS (N_{sample}=277); 3) cases without CRC (i.e. adenoma, appendix cancer, carcinoid tumor, or unknown case status, N_{sample}=2,958); 4) participants later determined not to be healthy controls (N=12); 5) a subset of Swedish controls from Uppsala due to low study-specific SNP call rate (N_{sample}=910); 6) MECC samples without ethnicity data (N_{sample}=8); and 7) East Asian samples from studies other than Korea, Taiwan, and Shanghai (N_{sample}=356).

Imputation

Here we describe the imputation of the OncoArray data. For imputation details for previous GWAS (Part 1), please see prior publications.[1, 2] A series of additional marker- and sample-level QC filters were applied prior to phasing and imputation. We excluded SNPs/indels with a call rate below 98% and with MAF<0.01 in either European or Asian samples as well as those flagged for known issues by the OncoArray Consortium (N_{marker}=11,235). Next, we calculated SNP concordance between HapMap3 samples genotyped on the OncoArray (N=153) and 1KGP Phase 3 genotypes. We used these results to match the strand between Illumina TOP/TOP alleles and the 1KGP forward strand. For SNPs not overlapping with 1KGP Phase 3, we utilized a master strand flip file from the OncoArray Consortium, which was created using BLAST (of the SNP probes) and frequency matching. Subsequently, we removed monomorphic SNPs, those with 1KGP Phase 3 concordance <95% or heterozygote concordance <95% and >2

het mismatches, and XY/MT SNPs. Alleles and positions for SNPs/indels were updated as appropriate.

Genotype data were pre-phased by chromosome using SHAPEIT v2.r837 and imputed to the 1KGP Phase 3 reference panel [downloaded from the IMPUTE2 website (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#reference); October 2014 release for chr1-22 and August 2015 release for chrX] using IMPUTE v2.3.2 [3-6, 10]. All cases and controls were simultaneously imputed together. Genetic markers with imputed data were retained for analysis if strict imputation quality and accuracy thresholds were attained (info≥0.7, certainty ≥0.9, and concordance ≥0.9 for directly measured markers) as well as a MAF filter of ≥0.01. Of note, 7 out of 11 novel variants identified in the Discovery phase were imputed, and all had info scores > 0.99.

Statistical Analysis

We then re-ran STRUCTURE with HapMap3 samples to estimate percent African, Asian, and European ancestry. Our European-specific analysis included individuals with ≥80% estimated European ancestry, and our East Asian-specific analysis included individuals with ≥80% estimated East Asian ancestry (Supplementary Fig. 2). Ancestral exclusions for <80% European or Asian ancestry from each population subgroup resulted in the exclusion of 2,037 samples.

Subsequently, we conducted PCA separately for European and East Asian study participants to generate PCs for global ancestry adjustment in statistical analyses. For Europeans, we utilized a panel of 35,713 ancestry informative markers (AIMs) from the OncoArray Consortium that was designed to capture within-European substructure. For East Asians, we extracted a random set of 8,000 markers from the OncoArray GWAS backbone for our AIM panel. For European studies, we combined individuals from contributing studies into 'aggregate studies' based on similar PC profiles for adjustment, where appropriate (e.g. for case-only studies). Age at diagnosis (for cases) and age at selection (for controls) were imputed where necessary, with a component study-specific average being assigned to individuals with missing data.

We conducted association analyses separately in Europeans and in East Asians as defined by genetic ancestry (Supplementary Fig. 2) using individual-level imputed genotype data with PLINK 1.9 [11]. In each ethnic group, we evaluated the association between allelic dosage for each imputed autosomal variant passing QC (N_{marker}=9,080,086) and allele frequency filters (minor allele frequency (MAF) ≥0.01) and risk of CRC. To examine each association, we utilized logistic regression, assuming a log-additive genetic model, with adjustment for age, sex, global ancestry (PCs 1-4), and 'aggregate-study' (for the European analysis only). The *P* value threshold for establishing statistical significance was $5x10^{-8}$. We created QQ plots to visualize the observed versus expected *P* value distributions and calculated a sample size-corrected genomic control factor (λ), which is equivalent to λ for a study of 1,000 cases and 1,000 controls.

Meta-Analysis of Discovery GWAS Parts 1 and 2

Association summary statistics for autosomal markers from Part 1 (CORECT, CCFR, and GECCO) and Part 2 (OncoArray) Europeans were combined in a fixed-effect inverse variance-weighted meta-analysis. This analysis on the union of markers (N_{marker}= 12,931,465) was implemented in METAL [8]. Heterogeneity of genetic associations between GWAS was examined with Cochran's *Q* test for heterogeneity. Next, we generated a QQ plot, examined the

GC λ, and calculated the sample-size corrected GC λ_{1000} to assess the possibility of population stratification. A Manhattan plot was used to visualize genome-wide association results, with differential shading for previously published risk loci and novel risk loci (including genetic markers within a 1 megabase window and/or r^2>0.2 in 1KGP Phase 3 Europeans). Further, we generated regional zoom plots with biological annotation for the 11 novel susceptibility regions identified using LocusZoom [12]. The complete list of genetic markers with $P<5x10^{-8}$ was pruned to identify the strongest signal at each unique locus. Markers within +/-500 kb on either side of the top SNP/indel and those with r^2>0.2 in 1KGP Europeans were eliminated.

Replication

We carried forward 11 genome-wide significant risk alleles from the pruned discovery GWAS into the independent replication stage. We examined the association between these SNPs/indels in European ancestry participants from MGI, CORSA, DACHS4, COLON, EPIC, HPFS3, NHS3, and the UK Biobank (study population descriptions below).

MGI. Genotypes from the MGI were imputed using the Michigan Imputation Server to the Haplotype Reference Consortium reference panel (release 1.1) and filtered based on quality score at Rsq ≥ 0.3. Sample quality control led to exclusions based on call rate <99%, estimated contamination >2.5 % (BAF Regress; http://genome.sph.umich.edu/wiki/BAFRegress), large chromosomal copy number variants (single chromosome with missingness ≥ five times larger than other chromosomes), lower call rate then its technical duplicate or twin, gonosomal constellations other than XX and XY, and inferred sex not matching the reported sex. Variant quality control led to exclusions based on probes not perfectly mapped or mapped perfectly to multiple positions in the human genome assembly (Genome Reference Consortium Human genome build 37), deviations from Hardy Weinberg equilibrium in reported European ancestry samples (P<0.0001), and call rate <99%. Statistical analysis was conducted using Firth's Bias-Reduced Logistic Regression (https://cran.r-project.org/web/packages/logistf/index.html).Age, sex, genotyping batch, and principal components 1–4 were included as covariates in the model.

CORSA. All samples had a missing genotype call rate <3%. Sample QC led to exclusion of 20 samples with discrepancies between reported and genotypic sex based on X chromosome heterozygosity, and exclusion of 94 close relatives defined as individuals that are second degree or more closely related. We inferred relatedness using the KING-robust procedure[4] and for each pair the sample with the lowest call rate was excluded. We conducted PCA after merging in 1000 Genomes Project data which identified 6 ancestry outliers which were excluded from analyses. Prior to phasing and imputation, we filtered out SNPs with missing call rate >2%, or HWE P <0.0001. We performed pre-phasing using SHAPEIT2[13] and imputed to the HRC.r1-1 panel using the Michigan Imputation Server [14]. To account for relatedness among study participants, we tested association with each variant using a linear mixed model as implemented in EMMAX [15]. We assumed additive allelic effects and included age and sex as covariates in the model. To enable inverse-variance weighted meta-analysis, we calculated approximate allelic log odds ratios and their corresponding standard errors as described in Cook et al. [16].

DACHS4. After filtering out SNPs with missing genotype call rate >2%, median sample missing rate was 0.0000932, and all but 7 samples had a missing rate <3%; the highest rate was 4.52%.

Sample QC resulted in the exclusion of 1 sample with a discrepancy between reported and genotypic sex, and the exclusion of 4 samples with a second degree or more closely related relative in the data. For each relative pair, the sample with the lowest call rate was excluded. PCA did not reveal any ancestry outliers. SNP QC followed the GAME-ON QC guidelines for the OncoArray chip. Prior to imputation, we filtered out SNPs with HWE P <0.0001. We performed pre-phasing using SHAPEIT2 [13] and imputed to the HRC.r1-1 panel using the Michigan Imputation Server [14]. We tested association with each variant using a logistic regression-based Wald test and adjusted for age, sex and the first 3 genotype PCs.

DACHS3, COLON, EPIC, HPFS3, NHS3. These five studies were genotyped together at the Center for Inherited Disease Research (CIDR) at Johns Hopkins University. Genotypic data that passed initial QC at CIDR underwent QC at the University of Washington Genetic Analysis Center (UW GAC) using standardized Quality Assurance/Quality Control (QA/QC) methods detailed in Laurie *et al.[17]*. In brief, a relatively low number of samples (n=30) had a missing call rate (MCR) >2%, with the highest MCR being 3.79%. Samples with discrepancies between reported and genetic sex (n=7) were excluded, and all analyzed samples were unrelated at the third-degree level. Analysis was restricted to European descent individuals that clustered with European 1000 Genomes Project samples in a combined PCA. Prior to imputation, variants that did not pass the UW GAC-recommended QC filters were filtered out. This included variants with HWE P <0.0001 based on a homogeneous subset of European descent study samples. We performed pre-phasing using SHAPEIT2 and imputed to the HRC.r1-1 panel using the Michigan Imputation Server. We conducted a single pooled analysis by performing logistic regression-based Wald tests, adjusting for age, sex, study, and the first 13 genotype PCs. We determined the number of PCs to include as a covariate by selecting all PCs from PC1 up to the highest PC with a significant P-value after regressing PCs on case/control status. The relatively high number of PCs can be explained by the fact that this dataset includes participants that were recruited from over 25 study centers in 10 European countries, as well as the United States.

UK Biobank. Genotype data imputed to the HRC.r1-1 panel were obtained from UK Biobank and QC and imputation are described elsewhere [http://www.ukbiobank.ac.uk/ ; https://www.biorxiv.org/content/early/2017/07/20/166298]. A nested case-control dataset was constructed as described below. European descent individuals were selected using PCA. We randomly dropped one sample from each pair of individuals that were more closely related than third degree relatives as inferred using KING-robust. This resulted in excluding 137 samples. We tested association with each variant using a logistic regression-based Wald test and in the model adjusted for age, sex and the first 7 genotype PCs.

Summary statistics from each of the individual studies were combined using a fixed effects inverse variance weighted meta-analysis approach.

Multiethnic Follow-up

We carried forward 11 genome-wide significant risk alleles from the pruned discovery GWAS into the multiethnic follow-up stage. We examined the association between these SNPs/indels in East Asians from the OncoArray Project, ACCC, and the US-Japan CRC GWAS; African Americans from the African American CRC GWAS; and Hispanics from the HCCS, MEC, and SIGMA. The genotyping, QC, imputation, and statistical analysis procedures for

participating studies have been described previously in detail except for the OncoArray (see above, Discovery: Part 2 (OncoArray)). ACCC participant samples were genotyped on the Affymetrix 6.0, 5.0 Illumina Omni Express, Human610-Quad, or HumanHap610 arrays and imputed to the GIANT ALL panel from the 1KGP (phase 1, release 3) after standard QC [18]. Samples from the US-Japan CRC GWAS were genotyped on the Illumina 1M-Duo or Illumina 660W-Quad, and after thorough QC, genotype data was imputed to the 1KGP East Asian reference panel (phase 1, release 3) [19]. East Asian association results from the OncoArray Project, ACCC, and the US-Japan CRC GWAS were then meta-analyzed using a fixed effects inverse-variance weighted approach using METAL. The African American GWAS samples were genotyped on the Illumina 1M-Duo or Illumina Omni 2.5M arrays and imputed using Europeans and Africans from the 1KGP (phase 1, release 3) as the reference panel [19]. Finally, individuals from the Hispanic CRC GWAS were genotyped on the HumanOmni2.5Exome-8v1.0, HumanOmni2.5Exome-8v1.1, or HumanOmni2.5-4v1 arrays, and missing genotypes were imputed to the multiethnic panel from 1KGP (phase 3, October 2014 release) [20].

Correcting for potential effect estimation bias due to the winner's curse

To correct for potential bias in effect estimation of newly-discovered variants, we implemented a fully Bayesian version of the weighted correction described by Zhong and Prentice, Eq 3.4.[21] Specifically, we placed a normal prior distribution on MLE effect estimates of the form $\beta_m \sim N(\beta_{Cor}, \tau^2)$. Here, β_m is the log odds ratio from the overall meta-analysis, β_{Cor} is the bias-corrected estimate calculated using the expectation-adjusted estimator from Eq. 3.1 in Zhong and Prentice[21], and τ is a pre-specified variance of the effect distribution reflecting the bias and is defined as $\tau = \left|\hat{\beta}_m - \beta_{Cor}\right|$. This correction was performed using the JAGS software [22]. Bias-corrected ORs and 95% credible intervals are presented in Supplementary Table 7. These estimates are used for calculation of the polygenic risk score.

Polygenic Risk Score Analysis

Assuming a log-additive model, we calculated two polygenic risk scores (PRS) for independent replication set participants from the summed risk alleles of 1) known susceptibility variants (N=67) and 2) known and novel susceptibility variants (N=76). The contribution of each risk allele was weighted by the respective bias-corrected log(OR) ($\beta_m$in the equation below) from our European-descent discovery analysis. For 3 novel susceptibility variants, SNP proxies were used as follows: rs1445012 for rs58791712, rs144037597 for rs6906359, and rs12822984 for rs72013726. Samples from MGI were excluded because individual level genotype data were not available for all variants. Samples from CORSA were excluded because no global PCs were calculated for the primary analysis using EMMAX. Thus, we could not properly adjust for distant relatedness and population structure in a pooled analysis. For each individual (*j*), the PRS was calculated as the weighted sum of risk allelic dosages for the relevant SNPs/indels:

$PRS_i = \sum_{m=1}^{M} \beta_m g_{im}$, where *M* is the total number of variants and g_{im} is the allelic dosage from imputation for individual *i* at variant *m*.

Each PRS was then divided into percentile categories based on risk allele counts among controls (<1%, 1-10%, 10-25%, 25-75%, 75-90%, 90-99%, >99%, with 25-75% serving as the reference). Subsequently, we used logistic regression models to examine the CRC risk prediction capabilities of PRS categories (after adjusting for age, sex, PCs (3 for DACHS4; 13 for COLON, DACHS3, EPIC, HPFS3, and NHS3; 7 for UK Biobank), and PC*study for known and

known+novel variants, respectively.[44] We also stratified the PRS at a clinically actionable threshold of OR=2.0, with the top two categories representing 90-95.7% and >95.7% of the PRS spectrum for known+novel variants.

We also directly applied a parallel PRS analysis to East Asian samples genotyped on the OncoArray. For consistency with the weighted analysis described above, we conducted an analysis weighted by log(ORs) from our European-descent discovery set. We chose to weight using these log(ORs) because they are likely the best estimates of a true effect and because use of the same weights facilitates an easier direct comparison to the European PRS (differences really just reflect allele frequency differences). East Asian samples besides those genotyped on the OncoArray contributed to the discovery of several variants in our previously known list, and thus, are excluded because they do not represent an independent sample set.

Proportion of familial risk explained

The contribution and comparison of the known+novel and the known only variants to the familial risk, under a multiplicative model, was computed using the formula

$$\sum_M (log\lambda_m) \Big/ (log\lambda_0)$$

Where λ_0 is the observed familial relative risk to first degree relatives of cases and λ_m is the familial relative risk due to locus m, calculated assuming a per allele effect:

$$\lambda_m = \frac{p_m r_m^2 + q_m}{(p_m r_m + q_m)^2}$$

where p_m is the frequency of the risk allele for locus m, $q_m = 1 - p_m$ and r_m is the estimated per-allele odds ratio [23]. For estimation, we pruned variants to include only independent SNPs, using only variants with r^2<0.2 in 1000 Genomes Phase 3 EUR or from the 1000 Genomes Pilot CEU data (rs16969681 and rs4779584 only). For any pair of correlated variants, we retained the variant with the largest effect size or, if effect sizes were equivalent, then the variant with the more common risk allele.

We performed this analysis using a fully Bayesian hierarchical modeling procedure that estimates the log(r_m) incorporating the uncertainty in estimating the variant-specific per allele odds ratio and the uncertainty in the estimate of λ_0 - rather than pre-specifying a single value. We specified a prior distribution as λ_0 ~ N(2.0,0.14^2), which places a 95% prior density in the range of [1.72, 2.28] to roughly correspond to the range of observed estimates in the literature [24, 25].

eQTL Analysis

The analysis of eQTL used data from the Colonomics study (www.colonomics.org) and the GTEx project [26, 27]. In Colonomics, normal mucosa was analyzed in a series of 50 donors with normal colon recruited during colonoscopy and normal adjacent mucosa of 100 colon cancer patients. Gene expression was analyzed with the Affymetrix Human Genome U219 Array Plate platform. Multiple probes in a gene were summarized using PCA. Genotypes were analyzed with the Affymetrix Genome-Wide Human SNP 6.0 array. Missing genotypes were imputed to 1KGP (phase 1, March 2012 release). For each SNP, partial correlation, adjusted for tissue type (healthy or adjacent to tumor), was used to explore the association of dosage SNP

data with gene expression for genes located within 2MB of the SNP of interest. For GTEx, the laboratory and analytic methods have previously been described elsewhere in detail (http://www.gtexportal.org/home/documentationPage#AboutData) [27].

Functional Annotation

Candidate functional SNPs were identified using published methods [28]. Briefly, SNPs (CEU, 1KGP, June 2014 release) in LD (we report $r^2 \geq 0.6$ except where noted in Supplementary Table 8 as $r^2 \geq 0.2$) were aligned with ChIP-seq tracks for histone methylation and acetylation marks associated with enhancers H3K4me1 and H3K27ac. For this study, we referenced Sigmoid Colon H3K27 acetylation from the Roadmap Epigenomics Consortium[29] as well as CRC cell lines SW480 and HCT-116 H3K4 monomethylation generated in our laboratory (G. Casey) and from the ENCODE project, respectively [30, 31].

Study populations contributing to the OncoArray Project

European Ancestry Study Populations

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.[32, 33] 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.[34, 35] 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.

Colocare Consortium

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.

Colon Cancer Family Registry (CCFR)

The CCFR (www.coloncfr.org) is an NCI-supported consortium consisting of six centers dedicated to the establishment of a comprehensive collaborative infrastructure for interdisciplinary studies in the genetic epidemiology of colorectal cancer. [36] The CCFR includes data from approximately 42,600 total subjects (11,800 case probands and CRC-affected relatives, and 30,800 unaffected relatives and unrelated controls). Cases and controls, age 20 to 74 years, were recruited at the six participating centers beginning in 1998. CCFR implemented a standardized questionnaire that is administered to all participants, and includes established and suspected risk factors for colorectal cancer, which includes questions on medical history and medication use, reproductive history (for female participants), family history, physical activity, demographics, alcohol and tobacco use, and dietary factors. The Set 1 scan, which has been described previously[37] , includes population-based cases and age-matched controls from the three population-based centers: Fred Hutchinson Cancer Research Center, Seattle, Washington; Mount Sinai Hospital, Toronto, Ontario and The University of Melbourne, Victoria, Australia. Cases were genetically enriched by over-sampling those with a young age at onset or positive family history. Controls were matched to cases on age and sex. All cases and controls were self-reported as White, which was confirmed with genotype data. The Set 2 scan 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 analytic dataset from CCFR included in the Discovery GWAS consisted of 1,972 CRC cases and 651 controls.

USC Norris Comprehensive Cancer Center Genetics Registry

The USC Norris Cancer Genetics Registry is a multicenter registry established to improve clinical care and to facilitate research to elucidate the genetic basis of hereditary cancers. Patients and families at risk of cancer are recruited from the University of Southern California Norris Comprehensive Cancer Center and LAC+USC Medical Center, where risk assessment, genetic counseling and clinical management are provided to individuals and families at risk for a variety of hereditary forms of cancer. Families at risk for hereditary forms of cancer are enrolled for continual follow-up to facilitate adherence to screening and management recommendations and to promote communication about risk to other relatives. Established in January, 2013, the USC Norris Cancer Genetics Registry provides the research structure for the development of a biorepository, clinical annotation, family expansion, longitudinal follow-up for research, germline and somatic DNA and RNA analysis, and analysis for the development of biostatistical and genetic models. For the OncoArray study, DNA samples and data from consented individuals from the USC Norris Cancer Genetics Registry with a pathologically confirmed diagnosis of adenocarcinoma of the colon or rectum were included for analysis, representing 234 colorectal cancer cases, and 221 were retained after QC. No controls were included from the registry for OncoArray analyses. The analytic dataset from the Cancer Genetics Registry included in the Discovery GWAS consisted of 159 CRC cases.

ESTHER II/VERDI[38, 39]

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.

FIRE3 (AIO KRK-0306)[40, 41]

In this open-label, randomized, Phase 3 trial, we recruited patients aged 18–75 years with stage IV, histologically confirmed colorectal cancer, an Eastern Cooperative Oncology Group (ECOG) performance status of 0–2, an estimated life expectancy of greater than 3 months, and adequate organ function, from centers in Germany and Austria. Patients were centrally randomized by fax (1:1) to FOLFIRI plus cetuximab or FOLFIRI plus bevacizumab (using permuted blocks of randomly varying size), stratified according to ECOG performance status, number of metastatic sites, white blood cell count, and alkaline phosphatase concentration. The primary endpoint was objective response analyzed by intention to treat. The study has completed recruitment, but follow-up of participants is ongoing. The trial is registered with

ClinicalTrials.gov, number NCT00433927. Between Jan 23, 2007, and Sept 19, 2012, 592 patients with KRAS exon 2 wild-type tumors were randomly assigned and received treatment (297 in the FOLFIRI plus cetuximab group and 295 in the FOLFIRI plus bevacizumab group). The analytic dataset from the FIRE3 study included in the Discovery GWAS consisted of 235 CRC cases.

GALEON

The GALicia Estudio Oncológico de coloN (GALEON) is a population-based case-control study conducted in the cities of Santiago and Vigo in Galicia, Spain. It included 105 incident histologically-confirmed colorectal cancer cases diagnosed from 2010-2015 in the University Hospitals of Santiago (CHUS) and Vigo (CHUVI). The control group included 205 healthy unrelated individuals who had been randomly selected from CHUS and CHUVI hospitals and primary health centers/family clinics from the same geographical area as the cases during the same period. The standardized questionnaires, modeled after the Colon Cancer Family Registry [42], covered information regarding medical history and medication use, reproductive and hormonal history (for female participants), family history, physical activity, dietary factors, lifestyle factors, and demographic information. Informed consent was required from all participants. The Galicia Ethics Committee for Clinical Investigation approved protocols of the study. The analytic dataset from the GALEON study included in the Discovery GWAS consisted of 92 CRC cases.

***Kiel (PopGen Biobank)*[43]**

All samples used were collected through the PopGen Biobank.[44] The CRC cases were members of a patient cohort from the Kiel area, described in detail elsewhere.[45] 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.

MAVERICC[46]

MAVERICC was a global, randomized, open-label, Phase II trial designed to compare first-line treatment with mFOLFOX6/bevacizumab and FOLFIRI/bevacizumab in 376 patients with mCRC. The unique feature of this study is that it leveraged two biomarkers—intratumoral ERCC1 and plasma VEGF-A—to attempt to predict outcomes with oxaliplatin- and bevacizumab-containing treatment, respectively. This is the first prospective study of tumoral ERCC1 (chemo-resistance marker to platinum compounds) and plasma VEGF-A as potential biomarkers for oxaliplatin- and BV-containing regimens, respectively, in an effort to further define the optimal chemotherapy backbone with biologic therapies, including BV, for mCRC. In this randomized, open-label, global, Phase II study, patients (N=376) with histologically or

cytologically confirmed CRC and ≥1 measurable metastatic lesion are stratified at screening by tumoral ERCC1 mRNA expression (high vs low, cutoff of 1.7 [ERCC1/β-actin mRNA]). Eligibility criteria include completion of adjuvant therapy >12 months before screening and an ECOG performance status ≤1. Blood samples are collected to quantify plasma VEGF-A levels. Patients within each ERCC1 stratification group are randomized 1:1 to mFOLFOX6-BV or FOLFIRI-BV administered in 2-week cycles. BV will be given at a dose of 5 mg/kg IV q2w. Patients will remain on study treatment until disease progression (PD) or unacceptable toxicity. Exploratory endpoints include correlative analyses with additional tumor tissue, blood, and SNP markers. The first patient was enrolled in August 2011. Clinicaltrials.gov Identifier: NCT01374425. The analytic dataset from the MAVERICC study included in the Discovery GWAS consisted of 235 CRC cases.

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. [47] 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. A total 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). Germline DNA was extracted from blood samples and 238 cases and 238 matched controls were sent to CIDR for genotyping. 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. The analytic dataset from the MCCS study included in the Discovery GWAS consisted of 212 CRC cases and 205 controls.

Moffitt (Colorectal Cancer Outcomes Prognosis and Epidemiology (COPE) Study and Total Cancer Care (TCC))

Germline DNA samples from Moffitt Cancer Center were collected from two studies. First, the Colorectal Cancer Outcomes Prognosis and Epidemiology study (COPE) is a prospective cohort of colorectal cancer (CRC) cases focused on identifying novel epidemiologic and biologic markers predictive of outcomes which could subsequently be utilized to tailor therapies and treatment of CRC. Second, the Moffitt Cancer Center Total Cancer Care (TCC) protocol, initiated in 2003, is a prospective, observational study permitting: 1) patient follow-up throughout their lifetime with access to medical data for research (demographics, past medical history, risk factors, therapies, outcomes); 2) donation of tissue samples (tumor, blood, and urine) with potential for various molecular analyses; and 3) re-contact of the consented patient if there is a new finding or trial that could benefit the patient. All CRC cases contributing to the CORECT analysis were newly diagnosed and histologically confirmed. Participants provided written, informed consent to participate in either COPE, TCC, or both studies according to Institutional Review Board-approved protocols (MCC 16028; MCC 14690/IRB 104189). The

analytic dataset from the COPE/TCC studies included in the Discovery GWAS consisted of 383 CRC cases.

Molecular Epidemiology of Colorectal Cancer (MECC) Study[24]

The Molecular Epidemiology of Colorectal Cancer Study (MECC) is a population-based case-control study of colorectal cancer (CRC). Incident, pathologically-confirmed CRC cases and controls were recruited from a specific region of northern Israel. Newly-diagnosed CRC cases beginning March 31, 1998, who agreed to participate, were interviewed, gave a venous blood sample, and provided permission for tumor tissue retrieval. Written, informed consent was obtained according to Institutional Review Board-approved protocols at Carmel Medical Center in Haifa and the University of Southern California (HS-12-00324, HS-12-00672, and HS-08-00378). Germline DNA was extracted from whole blood for genotyping. The analytic dataset from the MECC study genotyped on the OncoArray and included in the Discovery European GWAS consisted of 3,591 cases of pathologically-confirmed adenocarcinoma and 2,848 controls. In addition, previously genotyped cases and controls were included in the Discovery GWAS (Part 1): these consisted of 484 cases and 498 controls genotyped on the Illumina Omni 2.5 array, and 1,120 cases and 820 controls were genotyped on the Affymetrix Axiom CORECT Set array. Thus, the total number of cases and controls from the MECC study included in the Discovery GWAS (after quality control for genotyping) was 5,195 cases and 4,166 controls.

Multiethnic Cohort Study (MEC)[48]

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. In 2001-2005, a blood sample was collected from all willing participants to create a prospective biorepository. Colorectal cancer cases are identified through the Rapid Reporting System of the Hawai'i Tumor Registry and through quarterly linkage to the Los Angeles County Cancer Surveillance Program. Both registries are members of SEER. In GECCO, self-reported White subjects from the nested case-control study described above with DNA, and clinical and epidemiologic data were selected for genotyping. The analytic dataset from the MEC study included in the Discovery GWAS consisted of 69 CRC cases and 89 controls.

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.

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. [49] 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. [50] 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.

Puerto Rico Familial Colorectal Cancer Registry

The Puerto Rico Familial Colorectal Cancer Registry (PURIFICAR) is a population-based registry of colorectal cancer (CRC) patients in the island of Puerto Rico (http://purificar.rcm.upr.edu/index_eng.html). Eligibility for enrollment in PURIFICAR includes: 21 years of age or older, have pathology-confirmed CRC and for controls having a negative colonoscopy. Subjects are recruited through Dr. Cruz-Correa's clinic at the University of Puerto Rico Comprehensive Cancer Center, Puerto Rico Oncologic Hospital, Medical Services Association of Puerto Rico Clinics, Puerto Rico's Veteran's Hospital, Ashford Presbyterian Hospital at San Juan and those referred to PURIFICAR by gastroenterologists/surgeons from across the island. Participants in the study complete a questionnaire (in Spanish) modeled after the Collaborative Family Registries for Colorectal Cancer (Colon CFR). This questionnaire covers information regarding, medical history, reproductive history, diet, physical activity, lifestyle factors and demographic information. Furthermore, a detailed family history of cancer is obtained for each subject. Certified nurses conduct the informed consent and interviews at the Puerto Rico Consortium for Clinical & Translational Research (NIMHD funded http://prctrc.rcm.upr.edu). Pathology reports are obtained for all cases from the PR Central Cancer Registry (http://www.rcpr.org); medical records are requested for all cases. Data on tumor location, tumor stage (TNM stage, tumor

differentiation, lymph node metastasis), and number of positive lymph nodes are collected. Blood samples from study participants are collected according to standard operating procedures. Lymphocytes are isolated from whole blood using Ficoll density gradient centrifugation. The analytic dataset from the PURIFICAR study included in the Discovery GWAS consisted of 78 CRC cases and 71 controls.

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

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 analytic dataset from the SEARCH study included in the Discovery GWAS consisted of 4,537 CRC cases and 2,795 controls.

Spain

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. The analytic dataset from the Spain study included in the Discovery GWAS consisted of 932 CRC cases and 1,028 controls.

The Swedish Low-Risk Colorectal Cancer Study

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 were 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.

Sweden Wolk

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.

TRIBE[51, 52]

The TRIBE study accrued 508 patients across 34 Italian centers with un-resectable mCRC who had not received chemotherapy or biologic therapy for their metastatic disease but may have received prior adjuvant chemotherapy. Patients were randomized to receive FOLFOXIRI-bevacizumab (experimental group) or FOLFIRI-bevacizumab (control group) at two-week intervals for twelve cycles, followed by fluorouracil and bevacizumab maintenance therapy until the time of disease progression. The study design, eligibility criteria, treatment details, follow-up and clinical outcomes have been previously reported. Protocols were approved by the respective local ethics committees and conducted in accordance with the Declaration of Helsinki. All patients provided specific written informed consent for blood and tissue specimen collection for translational studies. Archived formalin-fixed paraffin-embedded (FFPE) tumor tissue was obtained from either the primary tumor or metastasis prior to randomization. Patients with

sufficient tissue for examination of all molecular markers were included in the present study. This retrospective correlative analysis conforms to the reporting guidelines established by the REMARK criteria. The analytic dataset from the TRIBE study included in the Discovery GWAS consisted of 320 CRC cases.

USC-HRT-CRC [53]

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.

East Asian Ancestry Study Populations

Korea: The Hwasun Cancer Epidemiology Study-Colon and Rectum Cancer (HCES-CRC) was a hospital-based case-control study conducted in South Korea. Cases were newly diagnosed CRC patients at Chonnam National University Hwasun Hospital, Jeollanam-do, South Korea from April 2004 to February 2013. Cancer-free controls were randomly selected from participants in the Korean Community Health Survey, an annual nationwide health interview survey, conducted from 2010 to 2012 in the Jindo and Bosung counties, Jeollanam-do, South Korea. The analytic dataset from the Korea study included in the multiethnic follow-up stage consisted of 3,130 CRC cases and 2,854 controls.

Shanghai Men's Health Study (SMHS) and Shanghai Women's Health Study (SWHS)

The Shanghai Women's Health Study (SWHS) and the Shanghai Men's Health Study (SMHS) are both population-based cohort studies which are being conducted in urban Shanghai, China. The SWHS includes 75,049 Chinese women who were between the ages of 40 and 70 years at enrollment during 1997 to 2000 and lived in urban Shanghai (1). In-person interviews were conducted to collect exposure information, and anthropometrics were measured. The response rate is 92% for the baseline interview. Approximately 88% of study participants provided biological samples, either a blood sample (n = 56,833) or exfoliated buccal cell sample (n = 8,921). Using similar study protocols, the SMHS enrolled 61,582 men between the ages of 40 and 74 years in urban Shanghai between 2001 and 2006 with an overall response rate of 74% (2).

Approximately 90% of study participants provided either a blood sample (76%) or a buccal cell sample (14%). Ongoing follow-up for cancer incidence and cause-specific mortality is conducted in both the SMHS and SWHS via a combination of periodic in-person surveys and annual linkage with data routinely collected by the population-based Shanghai Cancer Registry and Vital Statistics Unit (for death certificates). A subset of CRC cases with DNA available were identified in participants of the SWHS and SMHS and included, along with their controls, in this study. The analytic dataset from the SMHS and SWHS included in the multiethnic follow-up stage consisted of 121 CRC cases and 124 controls and 118 CRC cases and 103 controls, respectively.

Taiwan Study

Malignant neoplasms remain a leading cause of death in Taiwan. Data from Taiwan's National Cancer Registry reveals colorectal cancer is the most common form of cancer among Taiwan's population. Moreover, the incidence of colorectal cancer has steadily increased over the past years. Currently, the cost and resources associated with cancer place significant socioeconomic strain on patients, providers, and Taiwan's national healthcare system. To investigate and identify potential susceptibility loci for colorectal cancer, Taipei Medical University's Cancer Research Cancer and Taiwan's National Ministry of Health and Welfare launched an island-wide cancer research network to pool genomic data from colorectal cancer patients and identify novel biomarkers. The primary aims of this study were to elucidate the biological basis of inherited susceptibility of colorectal cancer and to understand how genetic variations can be monitored or modified by genetic and environmental factors. The study population consisted of 100 patients (age 45-80 years) diagnosed with stage III-IV colorectal cancer. Male to female ratio was approximately 1:1. After obtaining informed patient consent, DNA samples were obtained from either biopsy or surgically extracted tissue. The analytic dataset from the Taiwan study included in the multiethnic follow-up stage consisted of 486 CRC cases.

Study populations contributing to the independent replication stage

Michigan Genomics Initiative (MGI; http://pheweb.sph.umich.edu/pheno/153)

Participants were recruited through the University of Michigan Health system while awaiting diagnostic or interventional procedures either during a preoperative visit prior to the procedure or on the day of procedure that required anesthesia. Opt-in written informed consent is obtained. In addition to coded biosamples and protected secure health information, participants understand that all electronic health records, claims, and national data sources linkable to the participant may be incorporated into the MGI databank. Each participant donates a blood sample for genetic analysis, undergoes baseline vital signs and a comprehensive history and physical, and completes validated self-report measures of pain, mood and function, including NIH Patient Reported Outcomes Measurement Information System (PROMIS) measures. Data were collected according to Declaration of Helsinki principles. Study participants provided written informed consent, and protocols were reviewed and approved by local ethics committees. The analytic dataset from MGI included in the independent replication set consisted of 684 CRC cases and 21,430 controls defined by ICD codes.

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 [54]. 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). In total, 1,460 CRC (941) or advanced adenoma (519) cases and 774 matched controls were included in this study. Distribution of factors sex and age (5 year strata) were evenly matched between cases and controls. All participants were genotyped using the Affymetrix Axiom Genome-Wide Human Origins 1 Array.

Darmkrebs: Chancen der Verhütung durch Screening (DACHS3 and DACHS4)
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. 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. 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 cancer patients in the study region participated. 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 datasets analyzed in the replication study consist of 1,210 cases and 617 matched controls genotyped using the HumanOmniExpressExome-8v1-2 array (referred to as DACHS3), and an additional 1,013 cases and 657 matched controls genotyped using the Infinium OncoArray-500K (referred to as DACHS4).

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. A total of 643 CRC cases were included in the replication analysis. 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 the 643 CRC cases of the COLON study by age and gender. All participants were genotyped using the HumanOmniExpressExome-8v1-2 array.

European Prospective Investigation into Cancer (EPIC)

EPIC is an on-going 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 (only the Umeå site was included), 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 were included in the replication analysis. All participants were genotyped using the HumanOmniExpressExome-8v1-2 array.

Health Professionals Follow-up Study (HPFS3)

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 of the discovery stage case-control sets has been described in detail previously [1,2]. For the replication set (HPFS3), colorectal cancer cases were ascertained through January 1, 2010 and excluded cases included in the discovery stage [2]. 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 183 CRC cases and 197 controls were included in the replication analysis. All participants were genotyped using the HumanOmniExpressExome-8v1-2 array.

Nurses' Health Study (NHS3)

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 [55]. 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%. 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. Sample selection of the discovery stage case-control sets has been described in detail previously [1,2]. For the replication set (NHS3), colorectal cancer cases were ascertained through June 1, 2012 and excluded cases included in the discovery GWAS [2]. 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 were included in the replication analysis. All participants were genotyped using the HumanOmniExpressExome-8v1-2 array.

UK Biobank

We constructed a CRC and advanced adenoma nested case-control dataset from the UK Biobank resource (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 [50]. In total, 5,356 CRC (5,004) or advanced adenoma (352) cases and 21,407 matched controls were included in the replication analysis. All participants were genotyped using the Affymetrix UK Biobank Axiom Array.

References

1. Schumacher FR, Schmit SL, Jiao S, *et al.* Genome-wide association study of colorectal cancer identifies six new susceptibility loci. Nat Commun 2015;6:7138.
2. Peters U, Jiao S, Schumacher FR, *et al.* Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. Gastroenterology 2013;144(4):799-807 e24.
3. Genomes Project C, Abecasis GR, Altshuler D, *et al.* A map of human genome variation from population-scale sequencing. Nature 2010;467(7319):1061-73.
4. Genomes Project C, Abecasis GR, Auton A, *et al.* An integrated map of genetic variation from 1,092 human genomes. Nature 2012;491(7422):56-65.
5. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009;5(6):e1000529.
6. Howie B, Fuchsberger C, Stephens M, *et al.* Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 2012;44(8):955-9.
7. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 2007;81(5):1084-97.

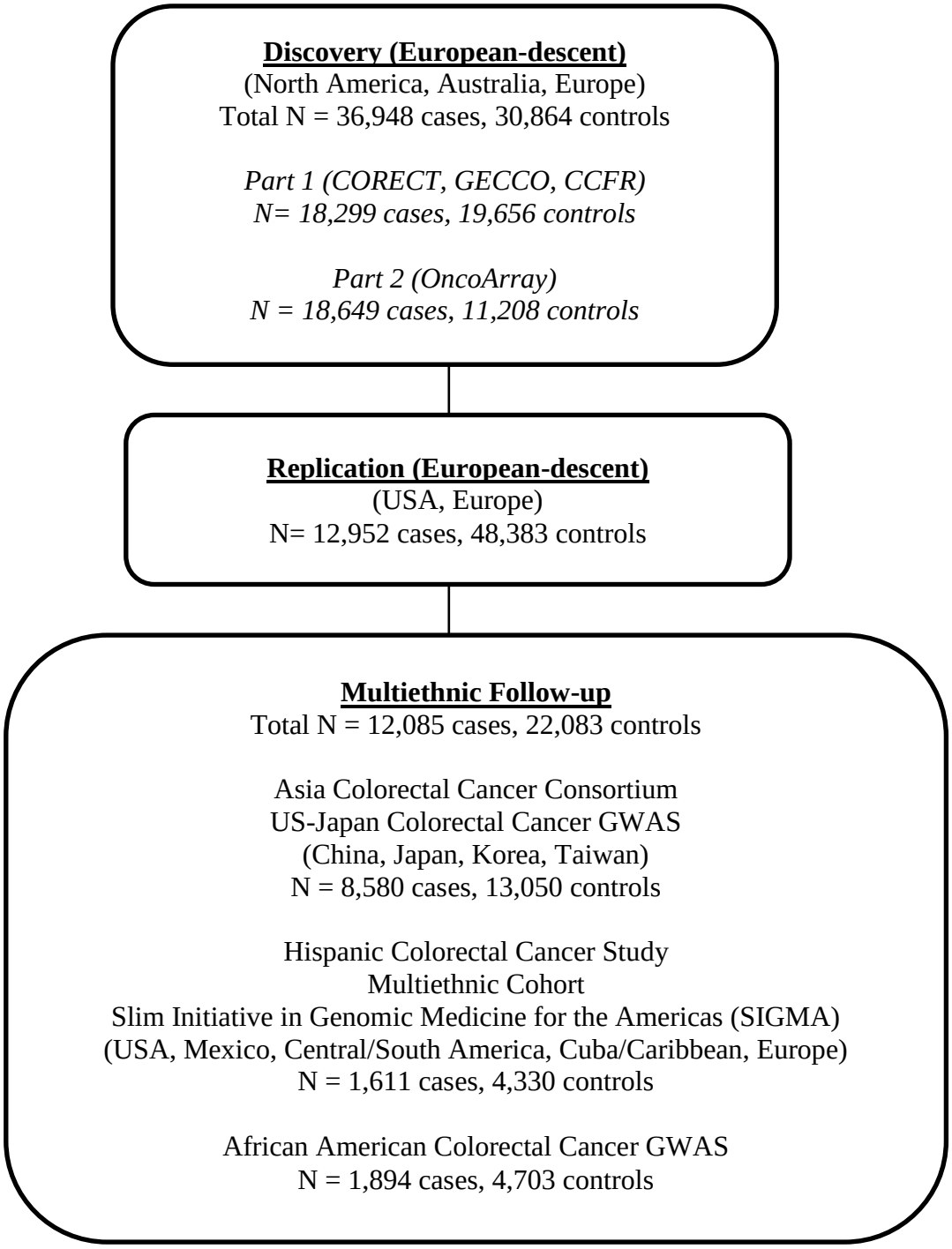
8. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 2010;26(17):2190-1.
9. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155(2):945-59.
10. Delaneau O, Howie B, Cox AJ, *et al.* Haplotype estimation using sequencing reads. Am J Hum Genet 2013;93(4):687-96.
11. Chang CC, Chow CC, Tellier LC, *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience 2015;4:7.
12. Pruim RJ, Welch RP, Sanna S, *et al.* LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 2010;26(18):2336-7.
13. Delaneau O, Zagury JF, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods 2013;10(1):5-6.
14. Das S, Forer L, Schonherr S, *et al.* Next-generation genotype imputation service and methods. Nat Genet 2016;48(10):1284-1287.
15. Kang HM, Sul JH, Service SK, *et al.* Variance component model to account for sample structure in genome-wide association studies. Nat Genet 2010;42(4):348-54.
16. Cook JP, Mahajan A, Morris AP. Guidance for the utility of linear models in meta-analysis of genetic association studies of binary phenotypes. Eur J Hum Genet 2017;25(2):240-245.
17. Laurie CC, Doheny KF, Mirel DB, *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies. Genet Epidemiol 2010;34(6):591-602.
18. Jia WH, Zhang B, Matsuo K, *et al.* Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. Nat Genet 2013;45(2):191-6.
19. Wang H, Burnett T, Kono S, *et al.* Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. Nat Commun 2014;5:4613.
20. Schmit SL, Schumacher FR, Edlund CK, *et al.* Genome-wide association study of colorectal cancer in Hispanics. Carcinogenesis 2016;37(6):547-56.
21. Zhong H, Prentice RL. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics 2008;9(4):621-34.
22. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing. Vienna, Austria, 2003*, https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf.
23. Amin Al Olama A, Dadaev T, Hazelett DJ, *et al.* Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among Europeans. Hum Mol Genet 2015;24(19):5589-602.
24. Poynter JN, Gruber SB, Higgins PD, *et al.* Statins and the risk of colorectal cancer. N Engl J Med 2005;352(21):2184-92.
25. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol 2001;96(10):2992-3003.
26. Closa A, Cordero D, Sanz-Pamplona R, *et al.* Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis. Carcinogenesis 2014;35(9):2039-46.
27. Consortium GT. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science 2015;348(6235):648-60.
28. Fortini BK, Tring S, Plummer SJ, *et al.* Multiple functional risk variants in a SMAD7 enhancer implicate a colorectal cancer risk haplotype. PLoS One 2014;9(11):e111914.

29. Bernstein BE, Stamatoyannopoulos JA, Costello JF, *et al.* The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 2010;28(10):1045-8.
30. O'Geen H, Echipare L, Farnham PJ. Using ChIP-seq technology to generate high-resolution profiles of histone modifications. Methods Mol Biol 2011;791:265-86.
31. Consortium EP. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science 2004;306(5696):636-40.
32. The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group. N Engl J Med 1994;330(15):1029-35.
33. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. The ATBC Cancer Prevention Study Group. Ann Epidemiol 1994;4(1):1-10.
34. Pietinen P, Hartman AM, Haapa E, *et al.* Reproducibility and validity of dietary assessment instruments. I. A self-administered food use questionnaire with a portion size picture booklet. Am J Epidemiol 1988;128(3):655-66.
35. Pietinen P, Hartman AM, Haapa E, *et al.* Reproducibility and validity of dietary assessment instruments. II. A qualitative food frequency questionnaire. Am J Epidemiol 1988;128(3):667-76.
36. Newcomb PA, Baron J, Cotterchio M, *et al.* Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. Cancer Epidemiol Biomarkers Prev 2007;16(11):2331-43.
37. Figueiredo JC, Lewinger JP, Song C, *et al.* Genotype-environment interactions in microsatellite stable/microsatellite instability-low colorectal cancer: results from a genome-wide association study. Cancer Epidemiol Biomarkers Prev 2011;20(5):758-66.
38. Jansen L, Herrmann A, Stegmaier C, *et al.* Health-related quality of life during the 10 years after diagnosis of colorectal cancer: a population-based study. J Clin Oncol 2011;29(24):3263-9.
39. Breitling LP, Raum E, Muller H, *et al.* Synergism between smoking and alcohol consumption with respect to serum gamma-glutamyltransferase. Hepatology 2009;49(3):802-8.
40. Heinemann V, von Weikersthal LF, Decker T, *et al.* FOLFIRI plus cetuximab versus FOLFIRI plus bevacizumab as first-line treatment for patients with metastatic colorectal cancer (FIRE-3): a randomised, open-label, phase 3 trial. Lancet Oncol 2014;15(10):1065-75.
41. Stintzing S, Modest DP, Rossius L, *et al.* FOLFIRI plus cetuximab versus FOLFIRI plus bevacizumab for metastatic colorectal cancer (FIRE-3): a post-hoc analysis of tumour dynamics in the final RAS wild-type subgroup of this randomised open-label phase 3 trial. Lancet Oncol 2016;17(10):1426-1434.
42. Jiang X, Castelao JE, Vandenberg D, *et al.* Genetic variations in SMAD7 are associated with colorectal cancer risk in the colon cancer family registry. PLoS One 2013;8(4):e60464.
43. Siegert S, Hampe J, Schafmayer C, *et al.* Genome-wide investigation of gene-environment interactions in colorectal cancer. Hum Genet 2013;132(2):219-31.
44. Krawczak M, Nikolaus S, von Eberstein H, *et al.* PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. Community Genet 2006;9(1):55-61.
45. Schafmayer C, Buch S, Volzke H, *et al.* Investigation of the colorectal cancer susceptibility region on chromosome 8q24.21 in a large German case-control sample. Int J Cancer 2009;124(1):75-80.

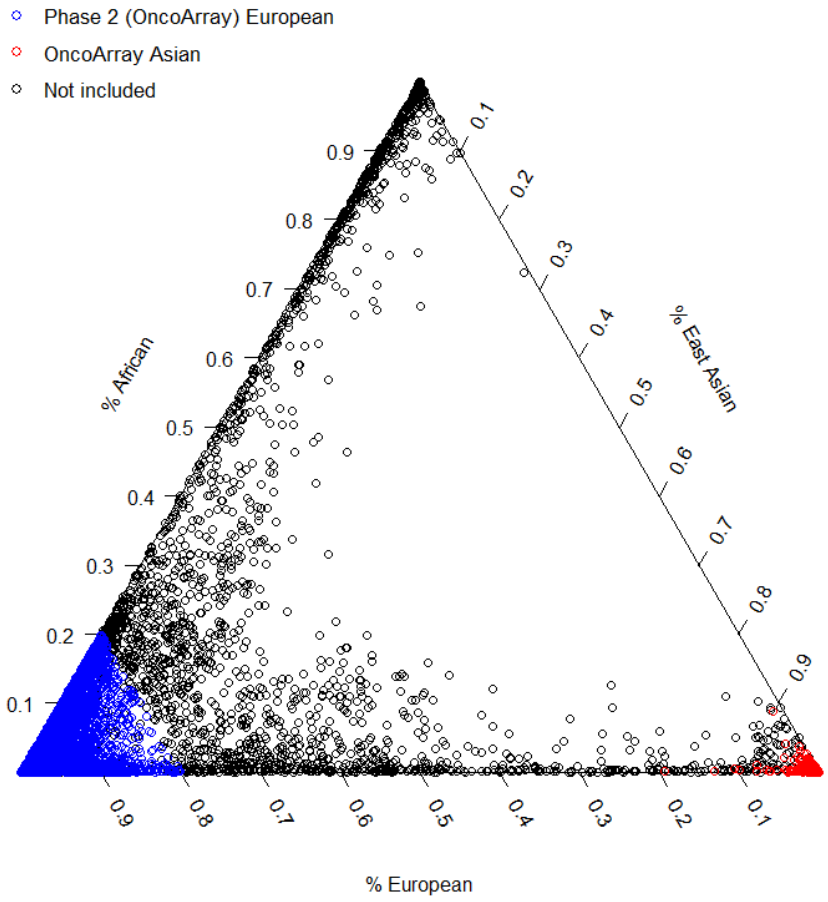
46. H.J. L. MAVERICC, a phase 2 study of mFOLFOX6-bevacizumab (BV) vs FOLFIRI-BV with biomarker stratification as first-line (1L) chemotherapy (CT) in patients (pts) with metastatic colorectal cancer (mCRC). J Clin Oncol 2016;34.
47. Giles GG, English DR. The Melbourne Collaborative Cohort Study. IARC Sci Publ 2002;156:69-70.
48. Kolonel LN, Henderson BE, Hankin JH, *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. Am J Epidemiol 2000;151(4):346-57.
49. Tworoger SS, Sluss P, Hankinson SE. Association between plasma prolactin concentrations and risk of breast cancer among predominately premenopausal women. Cancer Res 2006;66(4):2476-82.
50. Stampfer MJ, Willett WC, Speizer FE, *et al.* Test of the National Death Index. Am J Epidemiol 1984;119(5):837-9.
51. Loupakis F, Cremolini C, Masi G, *et al.* Initial therapy with FOLFOXIRI and bevacizumab for metastatic colorectal cancer. N Engl J Med 2014;371(17):1609-18.
52. McShane LM, Altman DG, Sauerbrei W, *et al.* REporting recommendations for tumor MARKer prognostic studies (REMARK). Nat Clin Pract Urol 2005;2(8):416-22.
53. Wu AH, Siegmund KD, Long TI, *et al.* Hormone therapy, DNA methylation and colon cancer. Carcinogenesis 2010;31(6):1060-7.
54. Hofer P, Baierl A, Feik E, *et al.* MNS16A tandem repeats minisatellite of human telomerase gene: a risk factor for colorectal cancer. Carcinogenesis 2011;32(6):866-71.
55. Belanger CF, Hennekens CH, Rosner B, *et al.* The nurses' health study. Am J Nurs 1978;78(6):1039-40.

Supplementary Figures

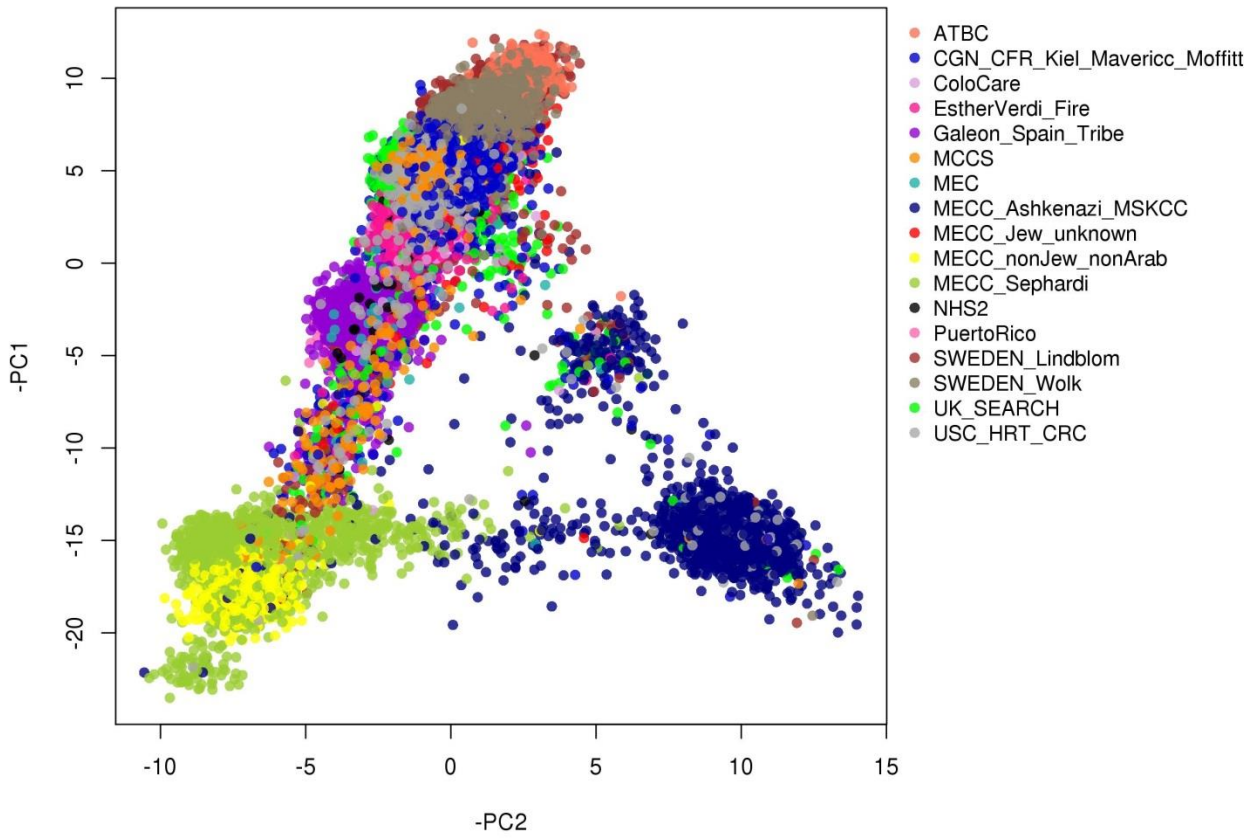
Supplementary Figure 1. Genome-wide association study design: stages and populations.



Supplementary Figure 2. OncoArray Project participants (Discovery Part 2) retained for analysis as determined by ≥80% estimated ancestry from appropriate population subgroups using STRUCTURE. Blue: European-descent in discovery stage (N_{case}=18,649; $N_{control}$=11,208). Red: East Asian-descent in multi-ethnic follow-up (N_{case}=3,855; $N_{control}$=3,081).

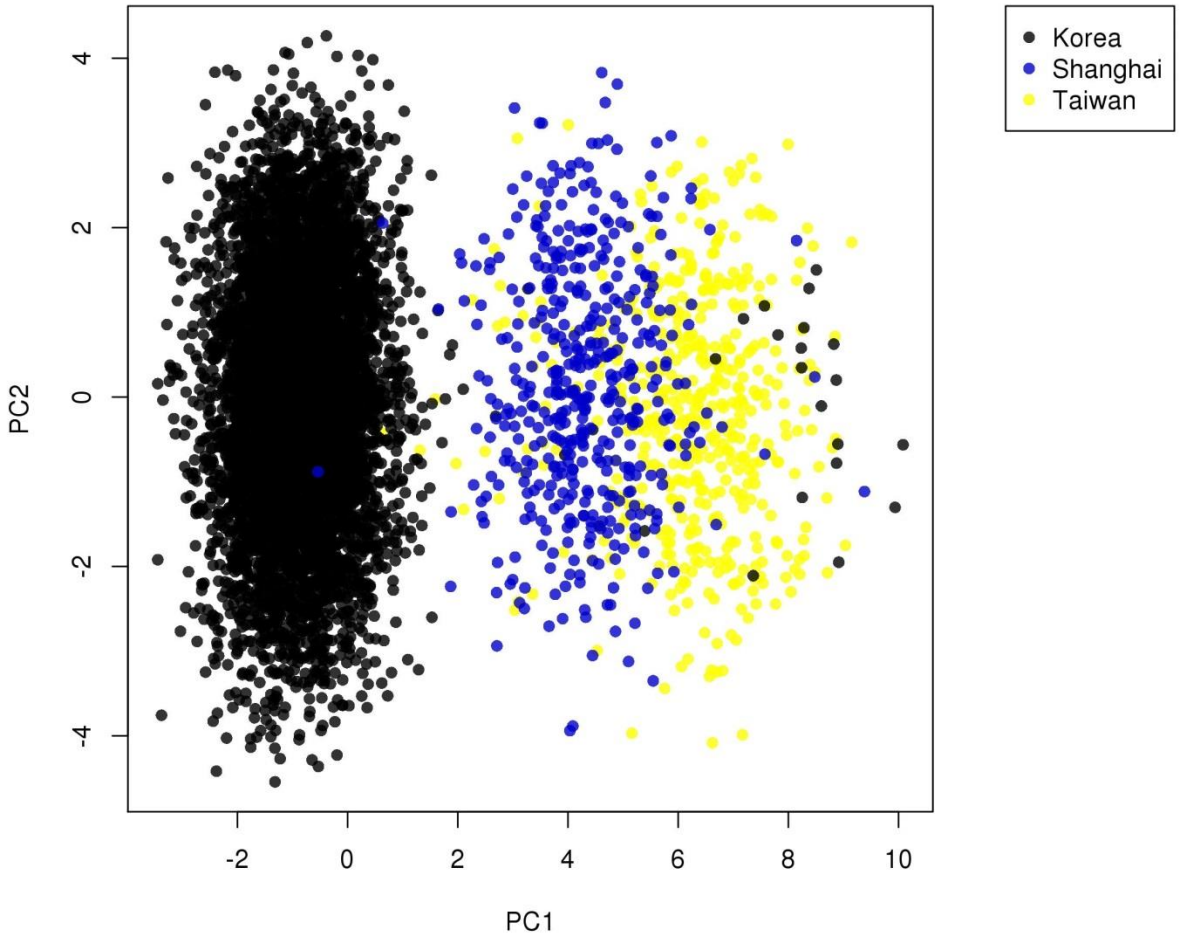


Supplementary Figure 3. Ancestry plot from a principal component analysis (PC1 vs. PC2) of European-descent OncoArray participants included in the discovery stage. PCA yielded the anticipated North to South and East to West clines across Europe and indicated that PCs 1 to 4 for global ancestry were sufficient to control for confounding due to population stratification. Ancestry analyses have been described previously for Part 1 of CORECT, GECCO, and CCFR.[1]

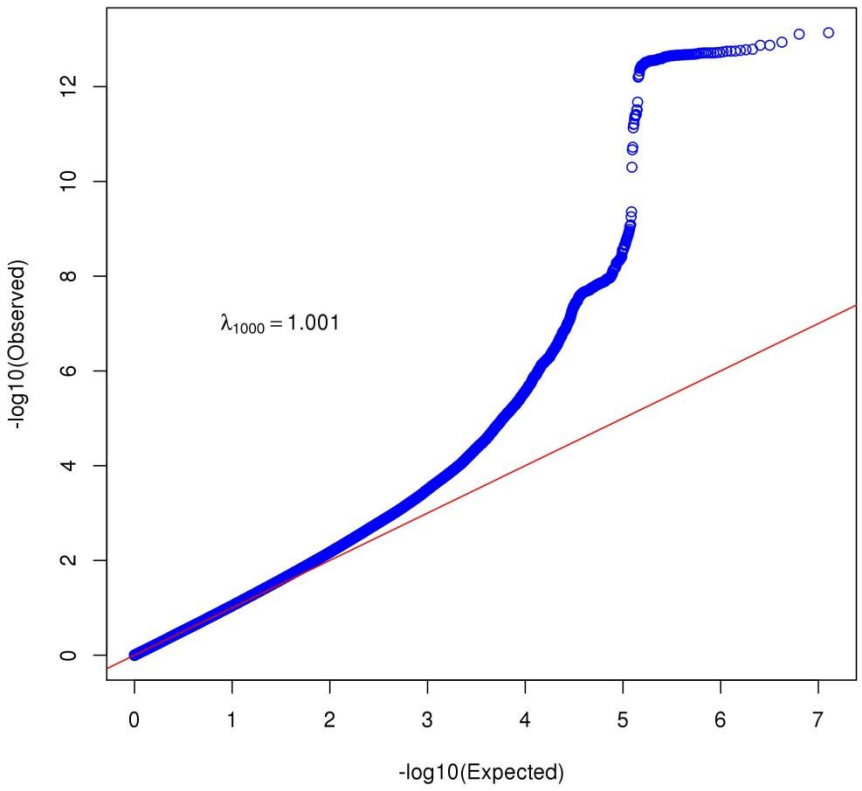


* CGN in this legend refers to the USC Norris Comprehensive Cancer Center Genetics Registry

Supplementary Figure 4. Ancestry plot from a principal component analysis (PC1 vs. PC2) of East Asian-descent OncoArray participants included in the multiethnic follow-up stage.

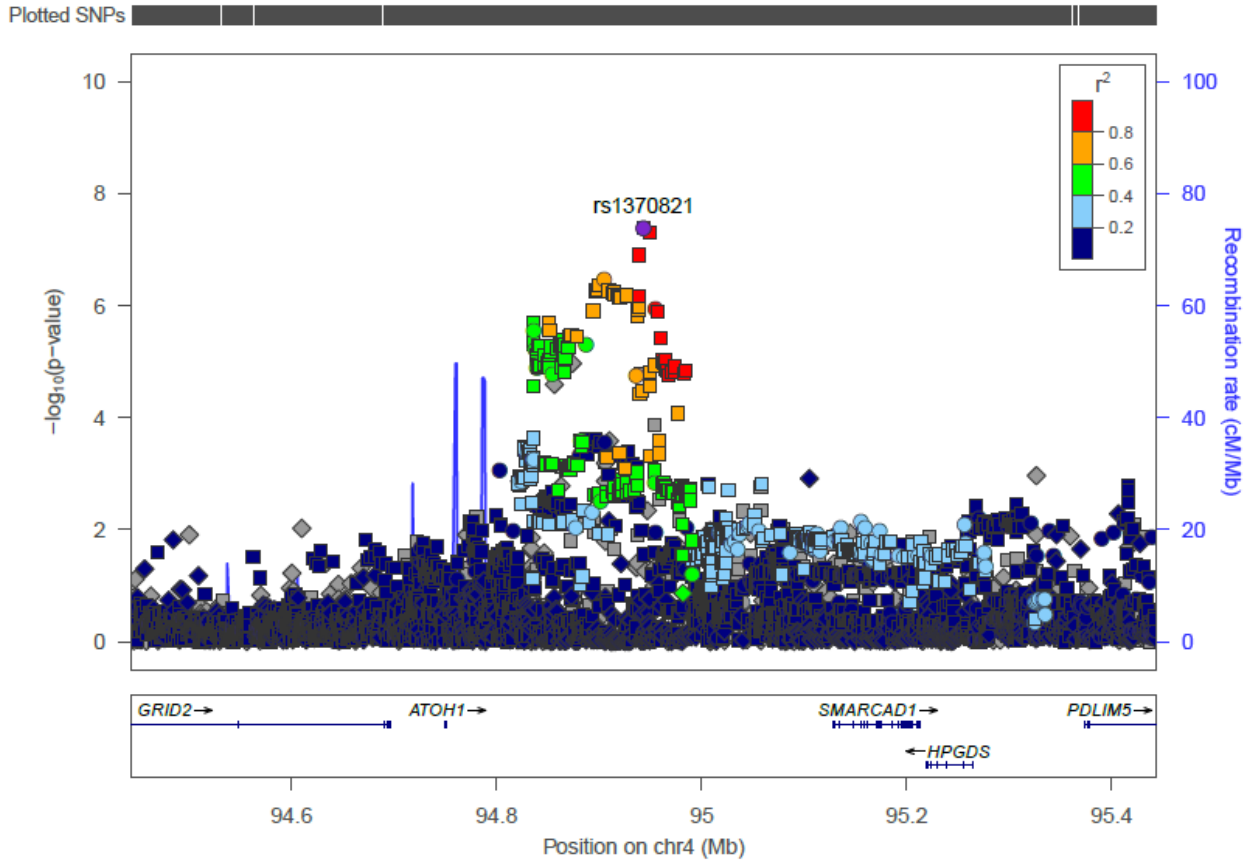


Supplementary Figure 5. Quantile-quantile (QQ) plot for the discovery GWAS (chr1-22), excluding variants in previously known susceptibility regions.

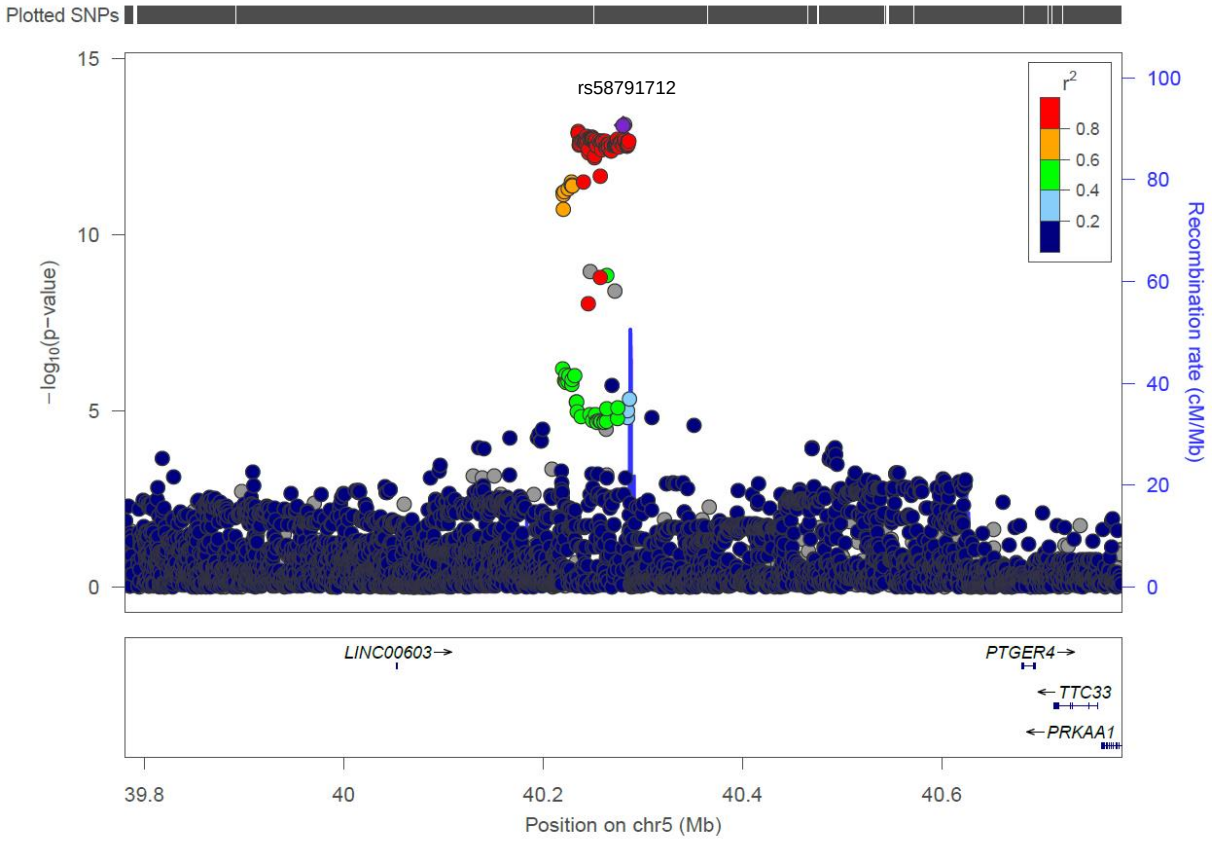


Supplementary Figure 6. Regional plots (500 kb window surrounding the index variant, except where noted) depicting the 11 novel CRC susceptibility loci in the discovery stage. Linkage disequilibrium shading is based on 1KGP Phase 3 Europeans. (A) rs1370821; (B) rs58791712; (C) rs2735940; (D) rs62404968; (E) rs6906359; (F) rs10994860; (G) rs72013726; (H) rs10161980; (I) rs2696839; (J) rs2295444; and (K) rs1810502.

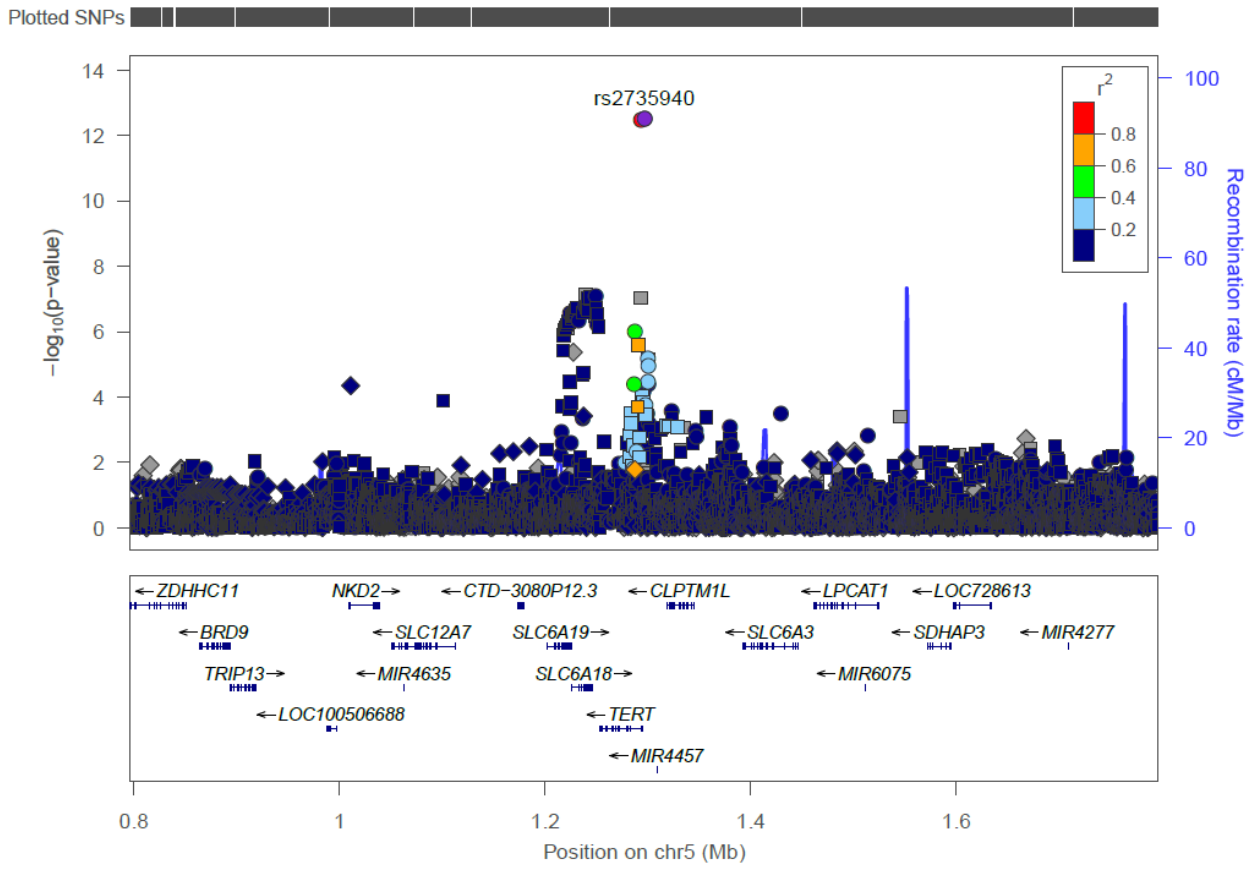
(A) 4q22.2: rs1370821



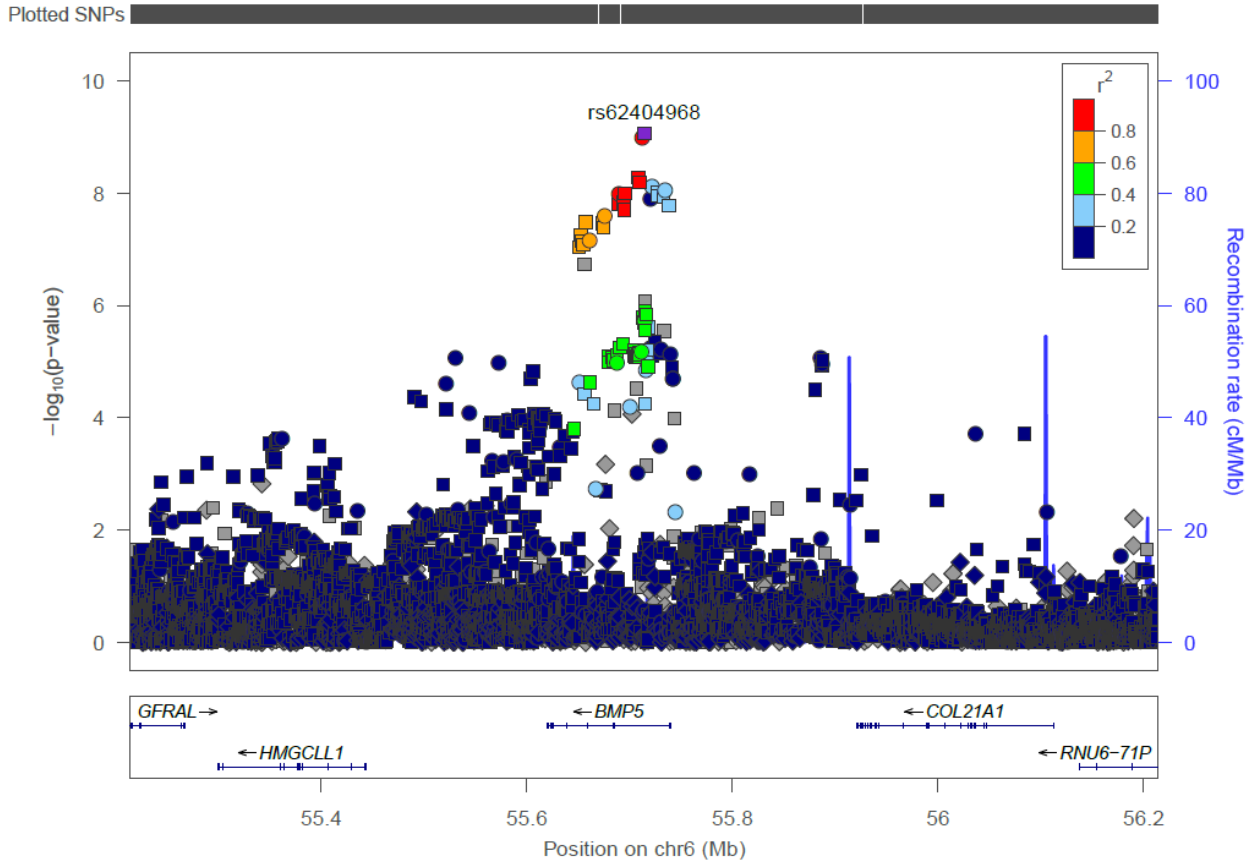
(B) 5p13.1: rs58791712



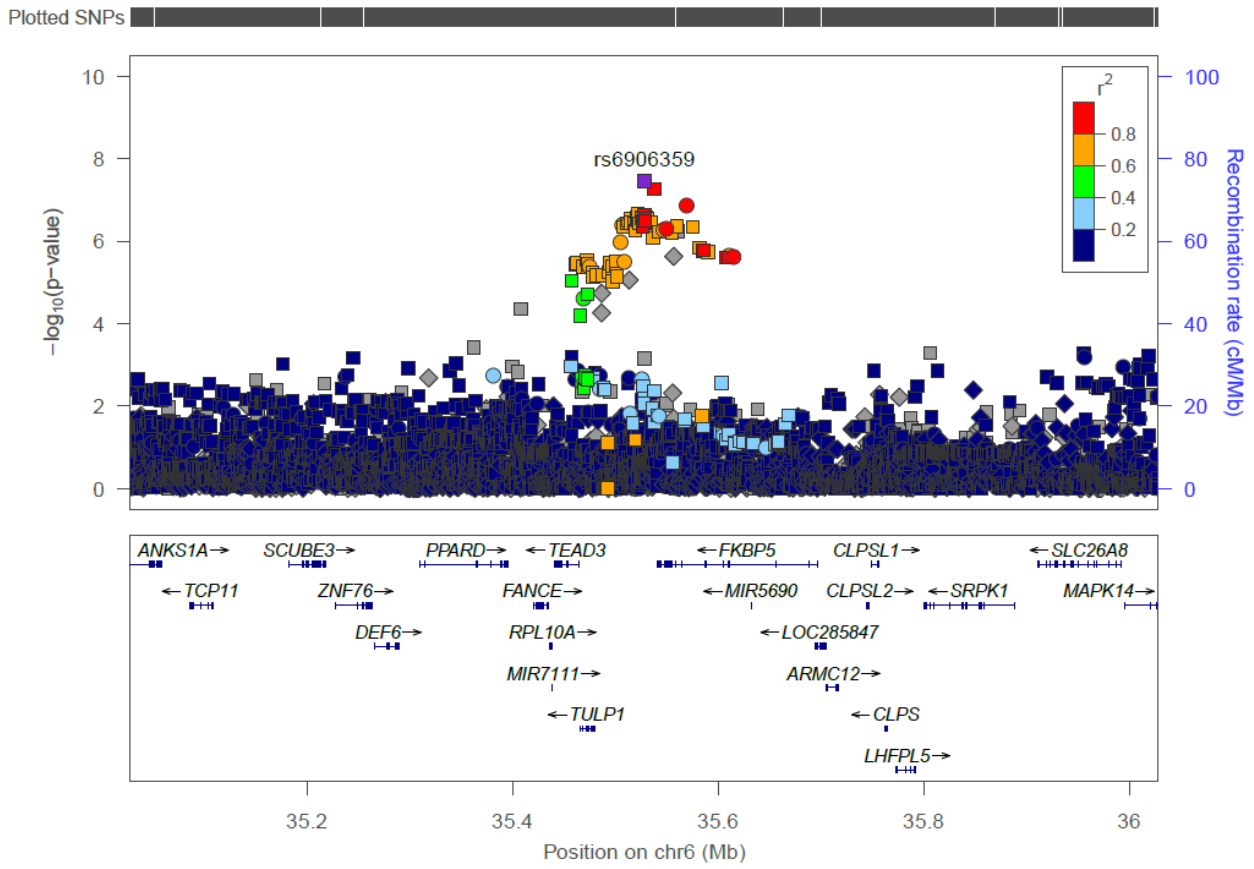
(C) 5p15.33: rs2735940



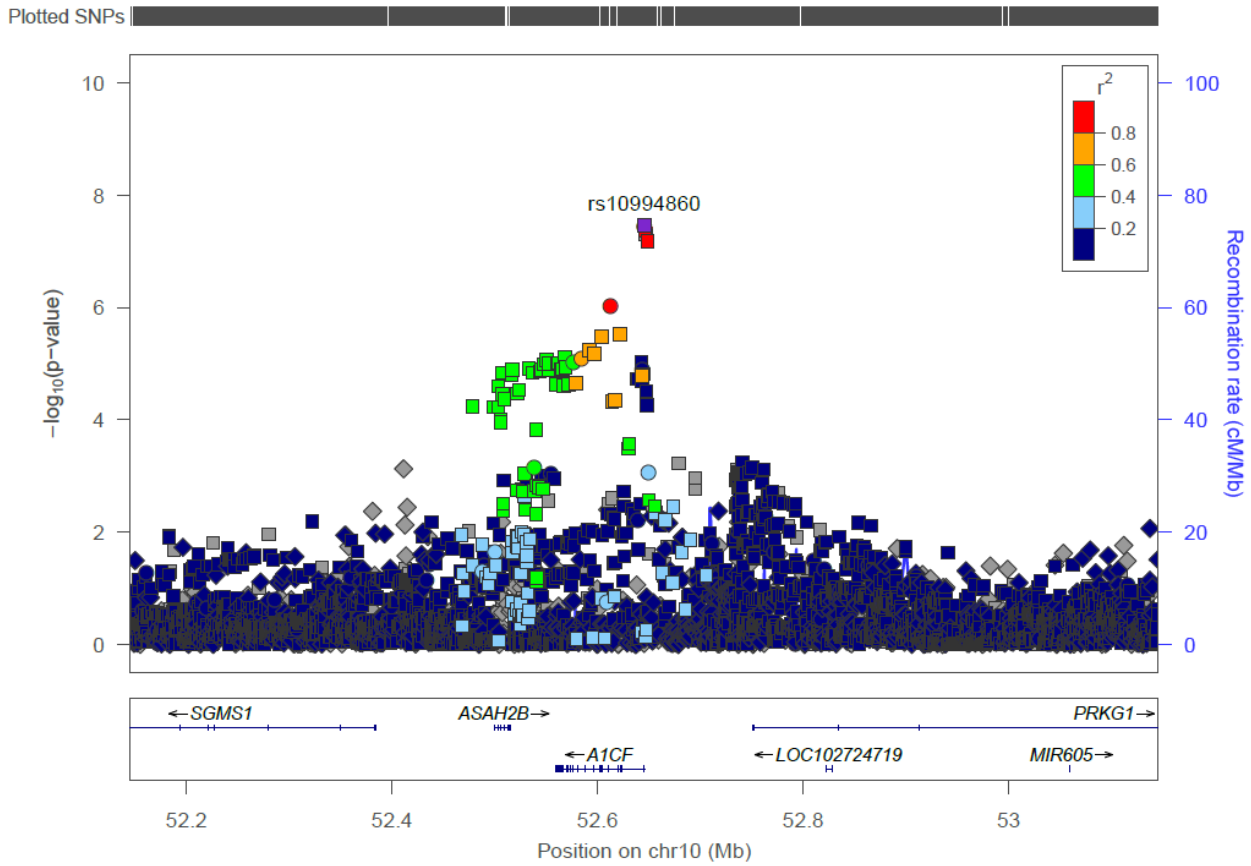
(D) 6p12.1: rs62404968



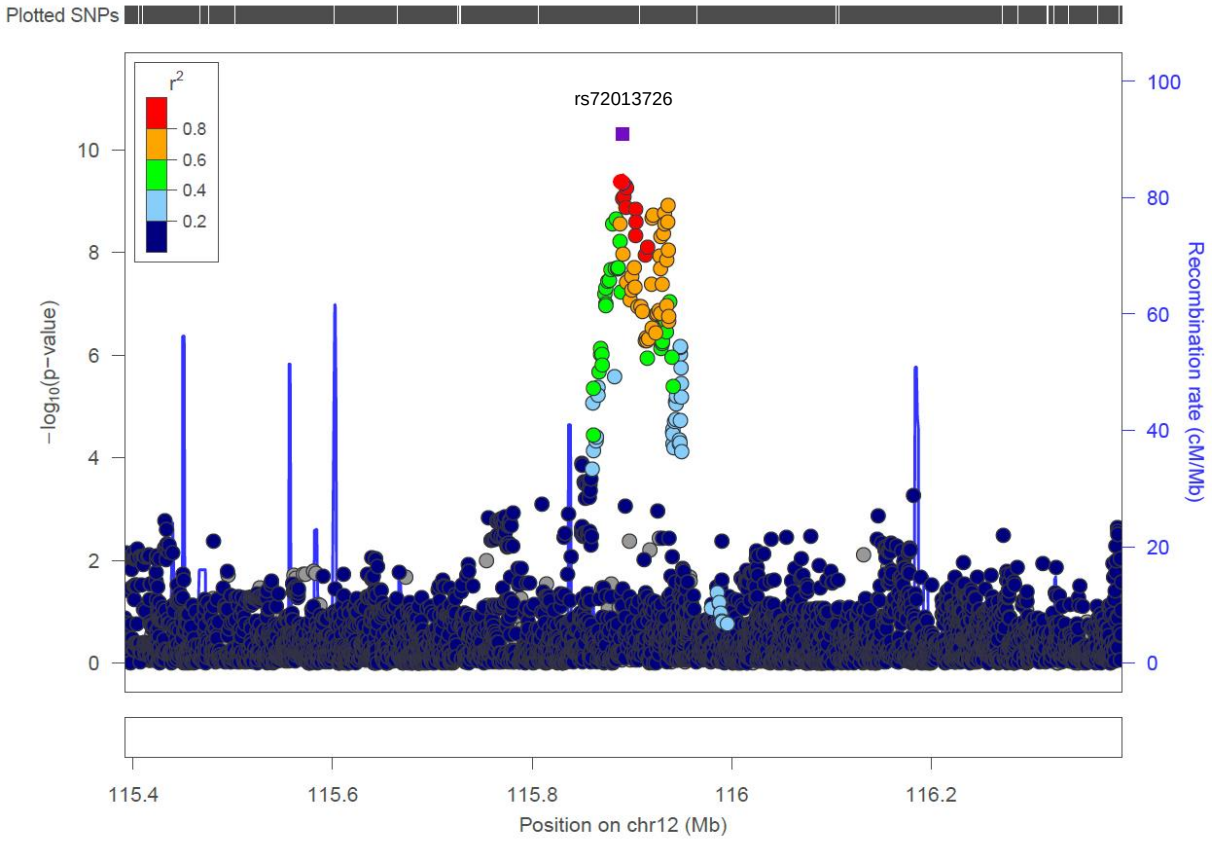
(E) 6p21.31: rs6906359



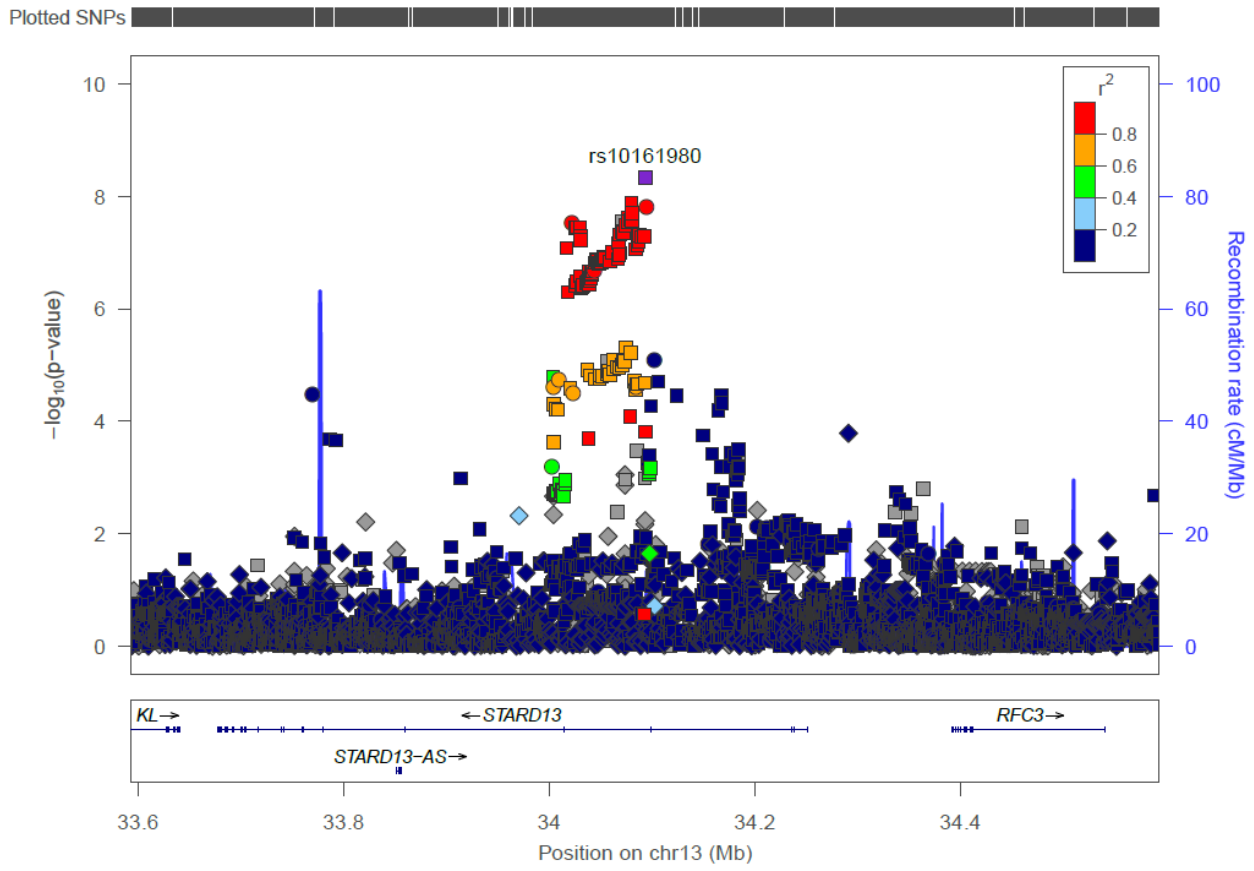
(F) 10q11.23: rs10994860



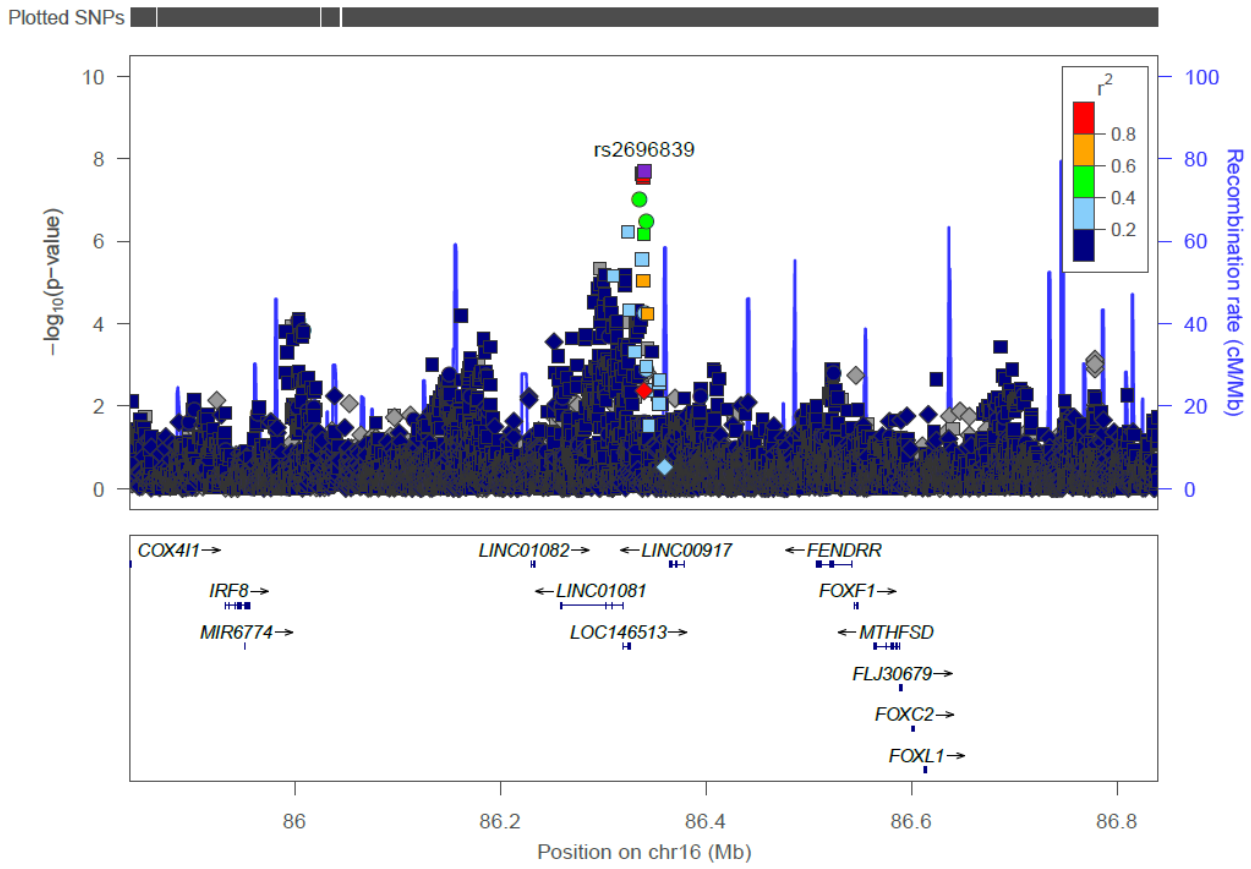
(G) 12q24.21: rs72013726



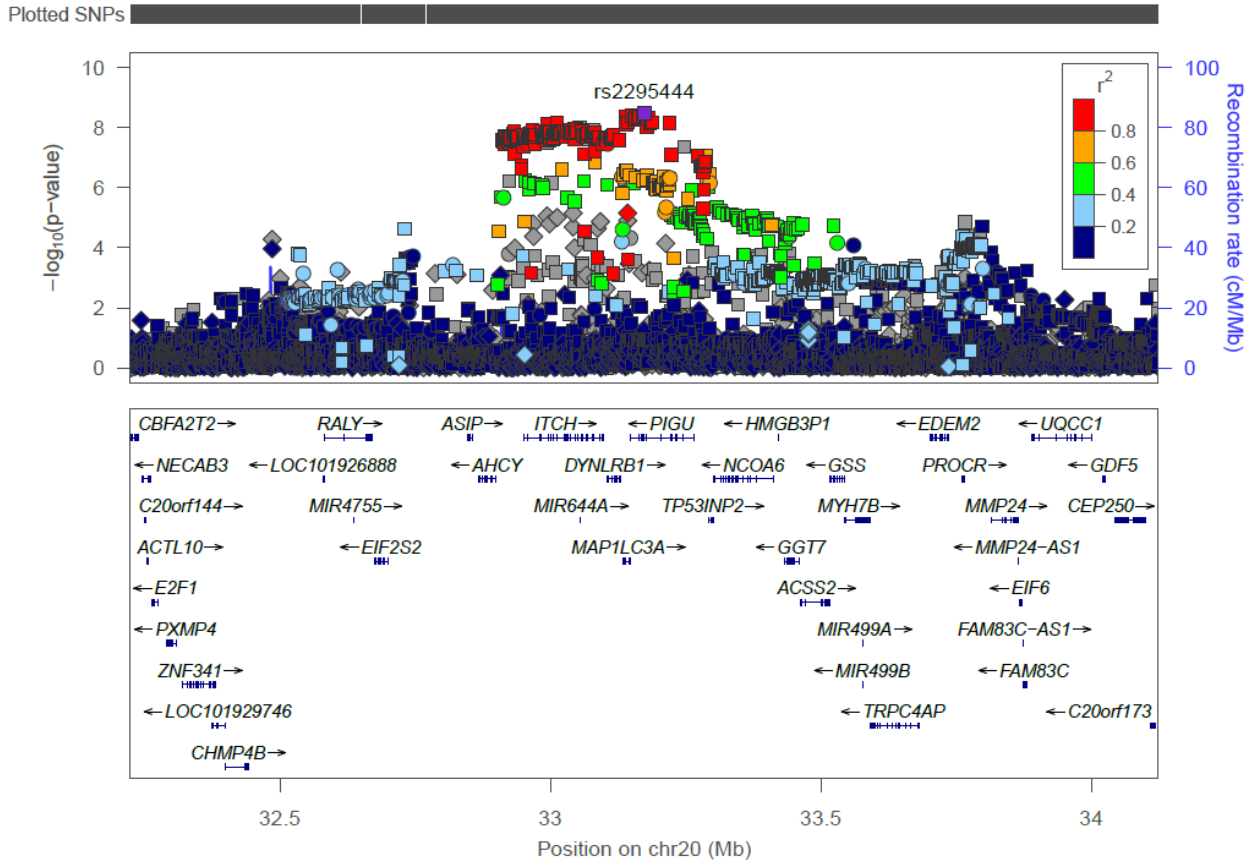
(H) 13q13.2: rs10161980



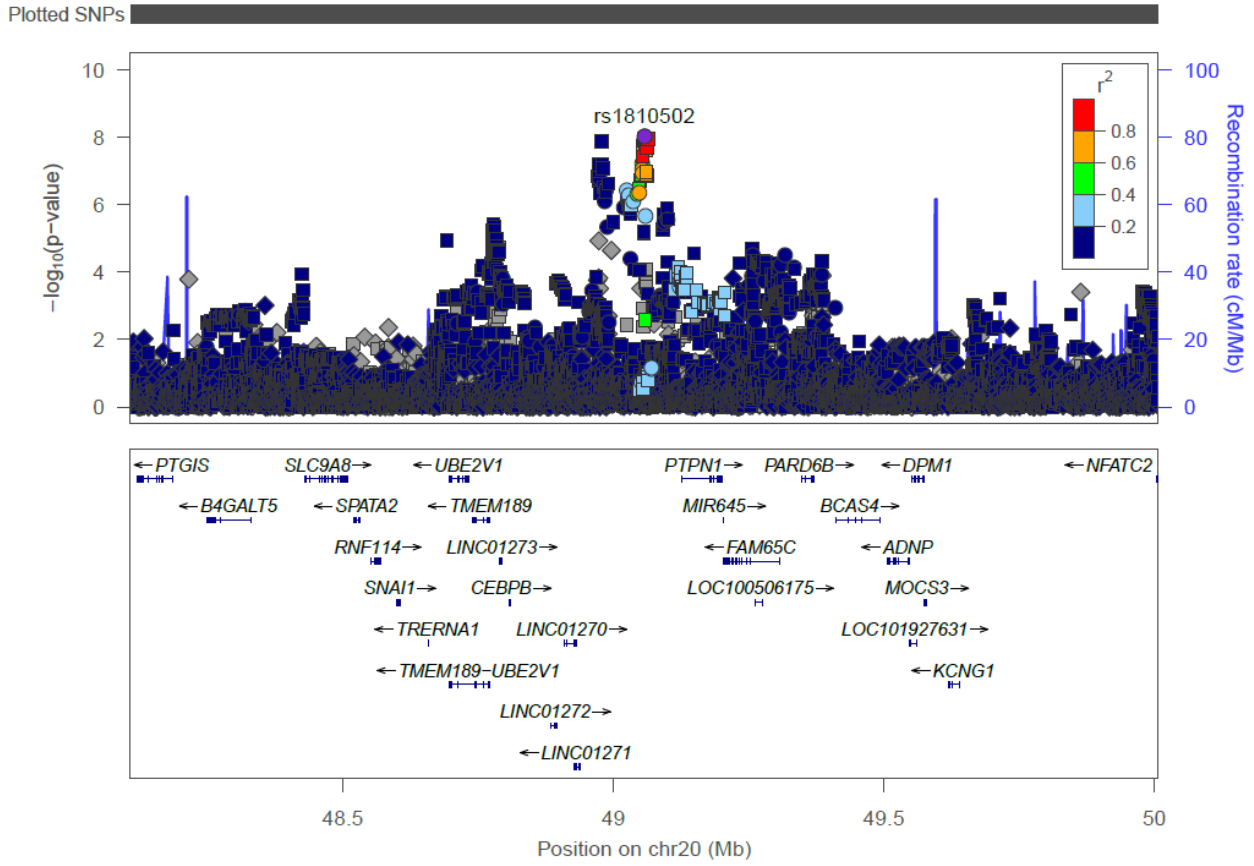
(I) 16q24.1: rs2696839



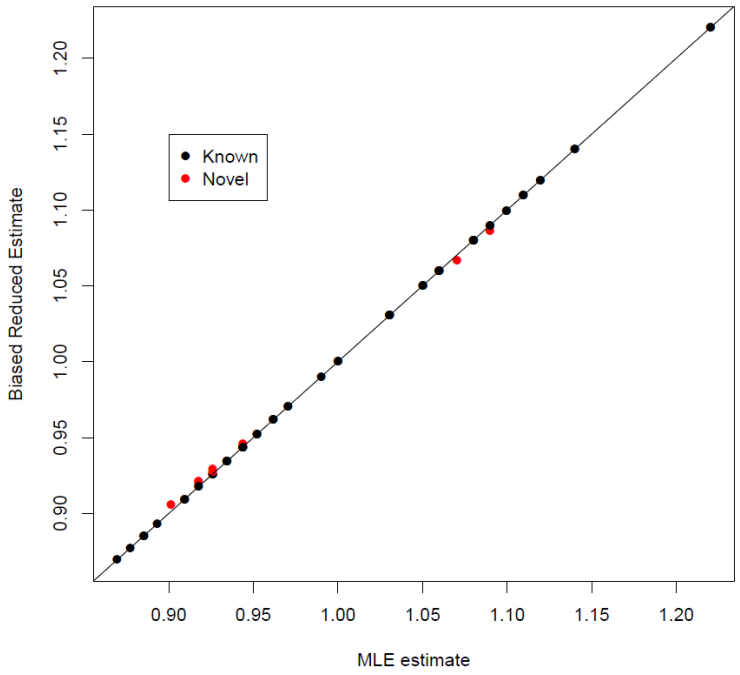
(J) 20q11.22: rs2295444 (+/-950kb window)



(K) 20q13.13: rs1810502 (+/-950kb window)

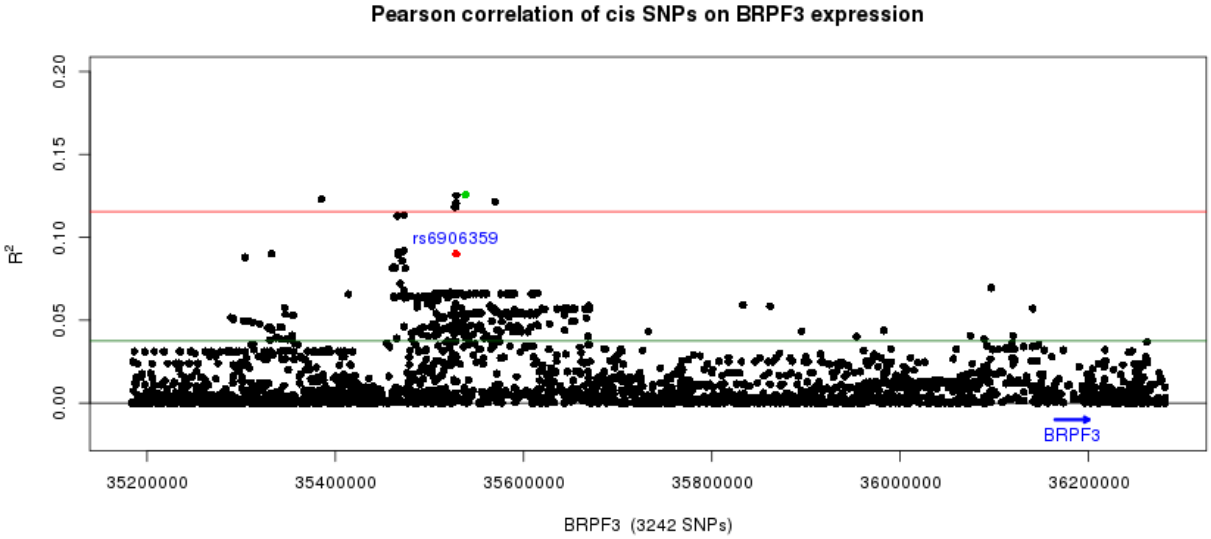


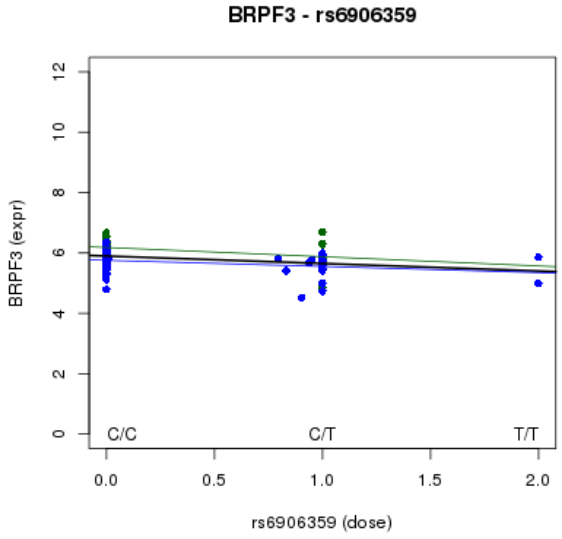
Supplementary Figure 7. Comparison of Maximum Likelihood Estimators (MLE) versus bias-corrected effect estimates for previously published (N=67) and novel (N=9) CRC susceptibility alleles.



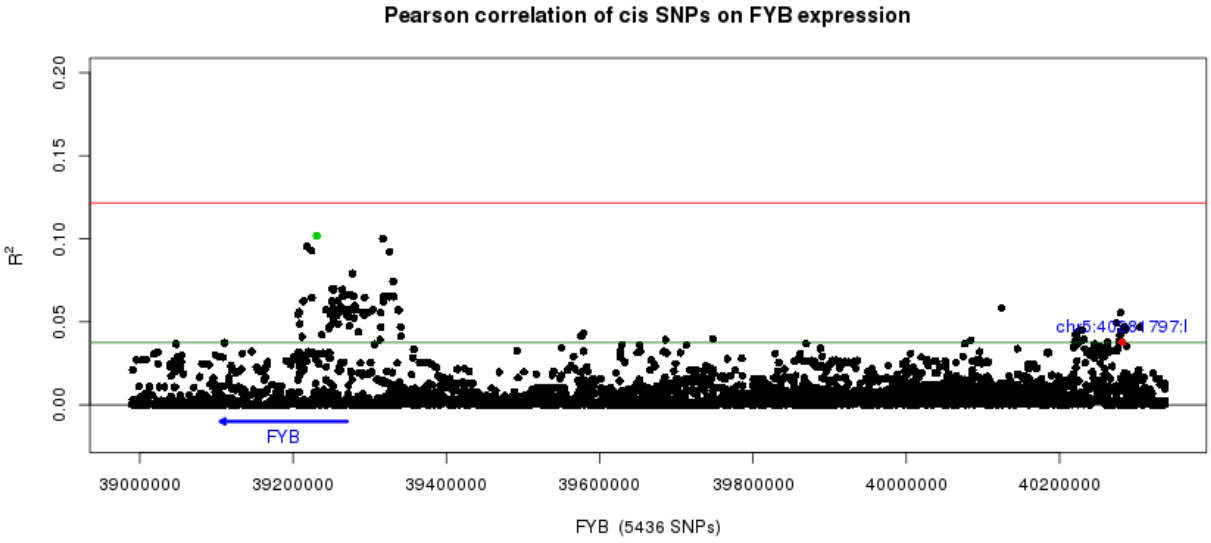
Supplementary Figure 8. Summary of the most statistically significant *cis* eQTLs in each region in the Colonomics dataset with at least 1 variant associated at $P<0.05$ based on Pearson partial correlation adjusted for tissue type (healthy or adjacent normal to tumor. (A) rs6906359 and *BRPF3* (partial $r^2=0.09$, $P=2.6E-04$). (B) rs58791712 and *FYB* (partial $r^2=0.04$, $P=0.02$). (C) rs1370821 and *BMPR1B* (partial $r^2=0.03$, $P=0.04$). (D) rs1810502 and *MOCS3* (partial $r^2=0.03$, $P=0.03$). Green: healthy colonic mucosa. Blue: normal tissue adjacent to CRC. Red line: Bonferroni statistical significance level = 0.05/number of SNPs in the plot. Green line: 0.01 statistical significance level. All statistical tests were two-sided.

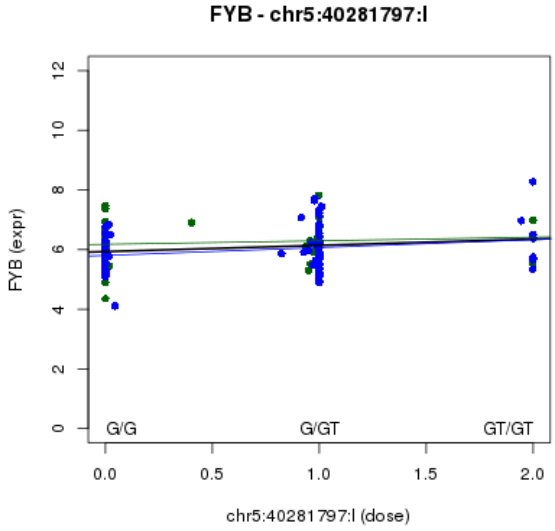
(A)



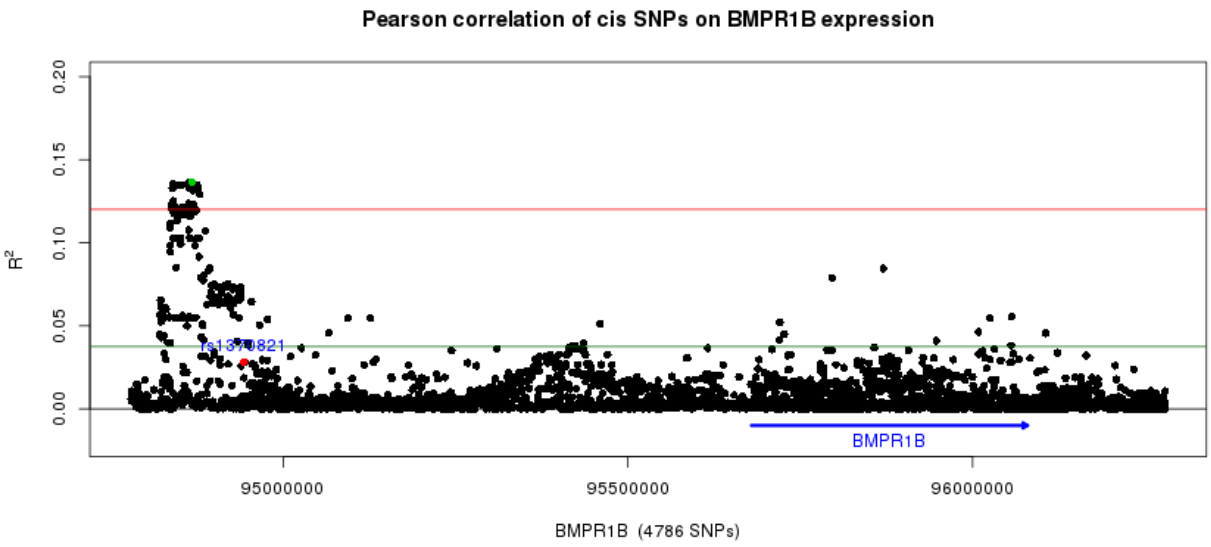


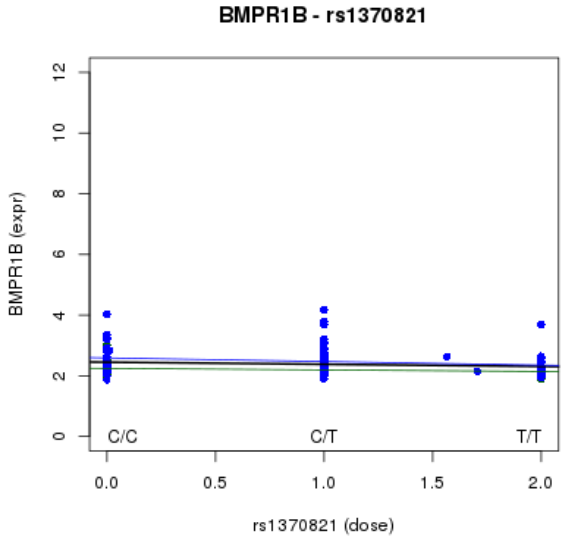
(B)


Pearson correlation of cis SNPs on FYB expression
R2
0.00
0.05
0.10
0.15
0.20
FYB
39000000
39200000
39400000
39600000
39800000
40000000
40200000
FYB (5436 SNPs)

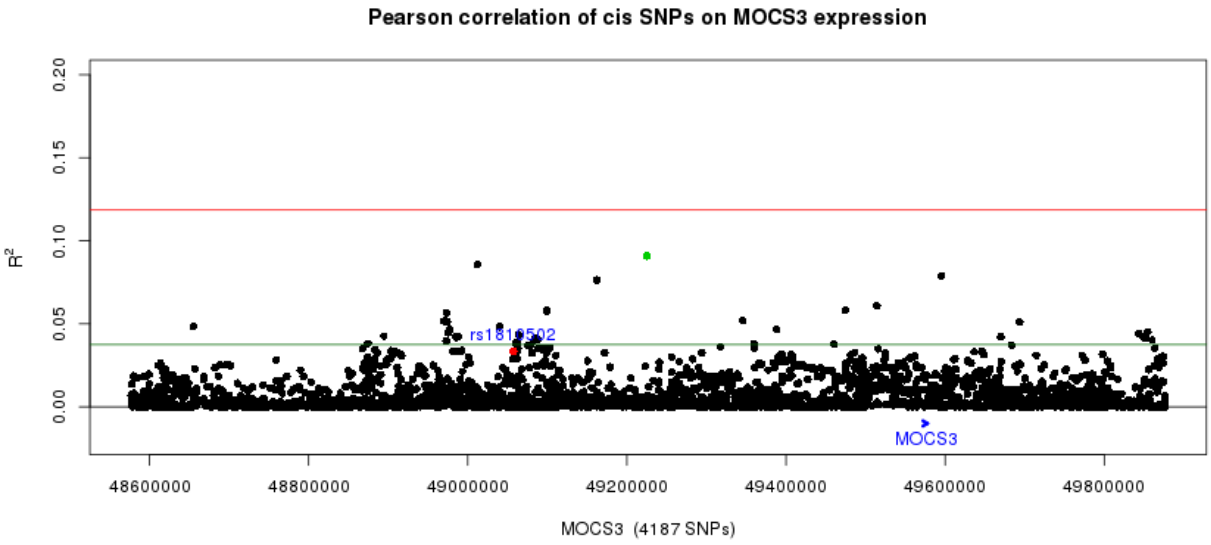

FYB - chr5:40281797:I
FYB (expr)
0
2
4
6
8
10
12
G/G
G/GT
GT/GT
0.0
0.5
1.0
1.5
2.0
chr5:40281797:I (dose)

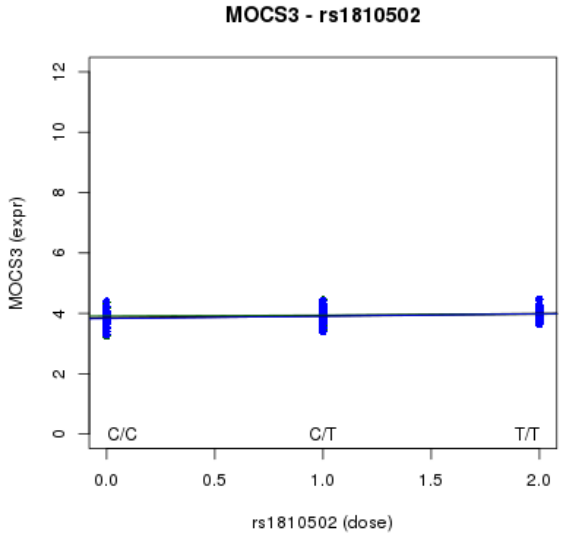
(C)


Pearson correlation of cis SNPs on BMPR1B expression
R²
0.00
0.05
0.10
0.15
0.20
rs1370821
BMPR1B
95000000
95500000
96000000
BMPR1B (4786 SNPs)


BMPR1B - rs1370821
BMPR1B (expr)
0
2
4
6
8
10
12
C/C
C/T
T/T
0.0
0.5
1.0
1.5
2.0
rs1370821 (dose)

(D)


Pearson correlation of cis SNPs on MOCS3 expression
R²
0.00
0.05
0.10
0.15
0.20
rs1810502
MOCS3
48600000
48800000
49000000
49200000
49400000
49600000
49800000
MOCS3 (4187 SNPs)


MOCS3 - rs1810502
MOCS3 (expr)
0
2
4
6
8
10
12
C/C
C/T
T/T
0.0
0.5
1.0
1.5
2.0
rs1810502 (dose)

Supplementary Tables

Supplementary Table 1. List of individual studies contributing to this CRC GWAS*

Study Acronym	Study Name	Consortium	Stage
CCFR	Colon Cancer Family Registry	CCFR, US-Japan CRC GWAS, AA CRC GWAS	D, F
MECC	Molecular Epidemiology of Colorectal Cancer Study	CORECT	D
Kentucky	Kentucky Case-Control Study	CORECT	D
CPS II	American Cancer Society Cancer Prevention Study II nested case-control study	CORECT	D
MCCS	Melbourne Collaborative Cohort Study	CORECT	D
Newfoundland	Newfoundland Case-Control Study	CORECT	D
ASTERISK	French Association Study Evaluating RISK for sporadic colorectal cancer	GECCO	D
COLO2&3	Hawai'i Colorectal Cancer Studies 2 & 3	GECCO, US-Japan CRC GWAS	D, F
DACHS1&2	Darmkrebs: Chancen der Verhütung durch Screening Study	GECCO	D
DALS1&2	Diet, Activity, and Lifestyle Study	GECCO	D
HPFS1, 2, 3, advanced adenomas	Health Professionals Follow-Up Study	GECCO	D, R
MEC	Multiethnic Cohort Study	CORECT, US-Japan CRC GWAS, Hispanic CRC GWAS, AA CRC GWAS	D, F
NHS1, 2, 3, advanced adenomas	Nurses' Health Study	GECCO, CORECT	D, R
OFCCR	Ontario Familial Colorectal Cancer Registry	GECCO	D
PHS	Physicians' Health Study	GECCO	D
PMH	Postmenopausal Hormones Supplementary Study to the Colon Cancer Family Registry	GECCO	D
PLCO1&2	Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial	GECCO, AA CRC GWAS	D, F
VITAL	VITamins And Lifestyle	GECCO	D
WHI1&2	Women's Health Initiative Study	GECCO	D
ATBC	Alpha-Tocopherol, Beta Carotene Cancer Prevention Study	CORECT	D
CGR	USC Norris Comprehensive Cancer Center Genetics Registry	CORECT	D
ColoCare	ColoCare Consortium	CORECT	D
ESTHER	Epidemiologische Studie zu Chancen der Verhütung, Früherkennung und optimierten Therapie chronischer Erkrankungen in der älteren Bevölkerung	CORECT	D
VERDI	Verlauf der diagnotischen Abklärung bei Krebspatienten	CORECT	D
GALEON	GALicia Estudio Oncologico de coloN	CORECT	D
FIRE3	Phase 3 Randomized Clinical Trial (FOLFIRI + cetuximab, FOLFIRI +	CORECT	D

	bevacizumab)		
Kiel	PopGen Biobank	CORECT	D
MAVERICC	Phase 2 Randomized Clinical Trial (mFOLFOX6-bevacizumab, FOLFIRI-bevacizumab)	CORECT	D
Moffitt (COPE and TCC)	Colorectal Cancer Outcomes Prognosis and Epidemiology; Total Cancer Care	CORECT	D
MSKCC	Memorial Sloan Kettering Cancer Center Cohort	CORECT	D
PURIFICAR	The Puerto Rico Familial Colorectal Cancer Registry (PURIFICAR)	CORECT	D
SEARCH	Studies of Epidemiology and Risk Factors in Cancer Heredity	CORECT	D
Spain	The Spanish study (University Hospital of Bellvitge, Hospital of Leon)	CORECT	D
Swedish Low-Risk CRC Study	Swedish Low-Risk Colorectal Cancer Study	CORECT	D
SMC and COSM	Swedish Mammography Cohort and Cohort of Swedish Men	CORECT	D
TRIBE	Phase 3 Randomized Clinical Trial (FOLFOXIRI + bevacizumab, FOLFIRI + bevacizumab)	CORECT	D
USC-HRT-CRC	Los Angeles County Cancer Surveillance Program	CORECT	D
CORSA	Colorectal Cancer Study of Austria	GECCO	R
DACHS3&4	Darmkrebs: Chancen der Verhütung durch Screening Study	GECCO	R
COLON	Colorectal Cancer: Longitudinal Observational study on Nutritional and lifestyle factors that influence colorectal tumor recurrence, survival and quality of life	GECCO	R
EPIC	European Prospective Investigation into Cancer and Nutrition	GECCO	R
UK Biobank	United Kingdom Biobank	N/A	R
HCES-CRC	The Hwasun Cancer Epidemiology Study-Colon and Rectum Cancer	CORECT	F
Shanghai MHS	Shanghai Men's Health Study	ACCC, CORECT	F
Shanghai WHS	Shanghai Women's Health Study	ACCC, CORECT	F
Taiwan	Taiwan Study	ACCC, CORECT	F
Guangzhou	Guangzhou Colorectal Cancer Study 1	ACCC	F
Aichi	Aichi Colorectal Cancer Study 1	ACCC	F
KCPS-II	Korean Cancer Prevention Study-II CRC	ACCC	F
JPHC	Japan Public Health Center-based Prospective Study	US-Japan CRC GWAS	F
Fukuoka	Fukuoka, Japan	US-Japan CRC GWAS	F
Nagano	Nagano, Japan	US-Japan CRC GWAS	F
HCCS	Hispanic Colorectal Cancer Study	Hispanic CRC GWAS	F
SIGMA	Slim Initiative in Genomic Medicine for the Americas Type 2 Diabetes Consortium	Hispanic CRC GWAS	F
SCCS	Southern Community Cohort Study	AA CRC GWAS	F
MD Anderson	MD Anderson Cancer Center	AA CRC GWAS	F
UNC-CanCORS	University of North Carolina CanCORS Study	AA CRC GWAS	F
UNC-Rectal	The North Carolina Rectal Cancer Study	AA CRC GWAS	F

*Abbreviations: CCFR = Colon Cancer Family Registry. CORECT = Colorectal Transdisciplinary Study. GECCO = Genetics and Epidemiology of Colorectal Cancer Consortium. ACCC = Asia Colorectal Cancer Consortium. AA CRC GWAS = African American Colorectal Cancer Genome-wide Association Study. US-Japan CRC GWAS = US-Japan Colorectal Cancer Genome-wide Association Study. D = discovery. R = replication. F = multiethnic follow-up.

Supplementary Table 2. Characteristics of participants from studies contributing to the OncoArray Project (Part 2) of the discovery GWAS*

Study	Population	Design	Genotyping Center	Cases			Controls			Total
				N	% Female	Mean age, y (SD)	N	% Female	Mean age, y (SD)	
ATBC	Finland	Cohort	CIDR	151	0.0%	69.1 (5.4)	32	0.0%	69.7 (4.9)	183
CGR	USA	Case-Series	USC	159	45.9%	48.6 (12.2)	NA	NA	NA	159
ColoCare	USA, Germany	Case-Series	CIDR	364	38.2%	59.6 (12.8)	39	51.3%	55.2 (13.1)	403
CCFR	USA, Canada, Australia	Case-Control	CIDR	1972	48.3%	54.5 (12.4)	651	46.4%	53.5 (13.3)	2623
ESTHER/VERDI	Germany	Case-Control	CIDR	420	34.3%	64.9 (7.3)	437	35.0%	65 (6.4)	857
GALEON	Spain	Case-Control	USC	92	40.2%	69.9 (9.7)	0	NA	NA	92
FIRE3	Germany	Clinical Trial Phase 3	USC	235	28.5%	62.3 (8.7)	0	NA	NA	235
Kiel (PopGen)	Germany	Case-Control	CIDR	1119	44.0%	60 (8.6)	0	NA	NA	1119
MAVERICC	USA	Clinical Trial Phase 2	USC	235	35.7%	60.2 (10.4)	0	NA	NA	235
MCCS	Australia	Cohort	CIDR	212	46.7%	73 (9.1)	205	48.3%	72.6 (9.2)	417
MEC	USA	Cohort	CIDR	69	52.2%	74.2 (9)	89	46.1%	71.8 (8.5)	158
MECC	Israel	Case-Control	CIDR, USC	3591	47.5%	67.3 (12.9)	2848	46.9%	70 (12.2)	6439
Moffitt (COPE + TCC)	USA	Case-Series	USC	383	46.0%	64.3 (12.7)	0	NA	NA	383
MSKCC	USA	Case-Control	CIDR	78	61.5%	59.3 (13.3)	0	NA	NA	78
NHS2	USA	Cohort	CIDR	109	100.0%	52.2 (6.1)	102	100.0%	51.7 (6)	211
PURIFICAR	Puerto Rico	Case-Control	USC	78	53.8%	57.9 (15.2)	71	66.2%	54.8 (12)	149
SEARCH	United Kingdom	Case-Control	CIDR, USC	4537	42.9%	59.6 (7.9)	2795	95.2%	47.1 (14.2)	7332
Spanish Study	Spain	Case-	CIDR, USC	932	35.6%	67.5 (11)	1028	48.1%	64.5 (11.1)	1960

Swedish Low-Risk CRC Study	Sweden	Control Cohort	CIDR	2667	45.0%	68.3 (10.8)	1643	46.8%	63.5 (6.2)	4310
SMC and COSM	Sweden	Cohort	CIDR	580	41.7%	70.8 (9.3)	859	42.50%	63.7 (8.1)	1439
TRIBE	Italy	Clinical Trial	USC	320	39.1%	58.2 (9.3)	0	NA	NA	320
USC – HRT - CRC	USA	Phase 3 Case-Control	CIDR	346	100.0%	66.4 (5.5)	409	100.0%	64.9 (6.7)	755
Total				18,649			11,176			29,857

*Individuals had ≥80% estimated European ancestry based on STRUCTURE analysis. Abbreviations: SD = standard deviation; USC = University of Southern California; CIDR = Center for Inherited Disease Research

Supplementary Table 3. Characteristics of European-descent participants included in the independent replication stage.

Study	Country	Design	Cases			Controls			Total
			N	% Female	Mean age, y (SD)	N	% Female	Mean age, y (SD)	
MGI	USA	Nested case-control	684	45.5	62.0 (12.1)	21,430	54.2	52.5 (16.5)	22,114
CORSA	Austria	Case-control	1,460	37.3	64.5 (10.9)	774	42.9	63.0 (10.4)	2,234
DACHS 4	Germany	Case-control	1,013	37.7	67.0 (11.8)	657	36.2	69.2 (10.7)	1,670
COLON*	Netherlands	Cohort	643	36.5	65.2 (9.4)	692	37.0	61.6 (6.5)	1,335
DACHS 3*	Germany	Case-control	1,210	37.7	68.6 (10.8)	617	39.2	66.6 (11.6)	1,827
EPIC*	Europe[†]	Cohort	2,095	53.3	57.0 (8.1)	2,306	53.4	56.7 (8.1)	4,401
HPFS 3*	USA	Cohort	183	0.0	72.4 (8.6)	197	0.0	72.0 (8.7)	380
NHS 3*	USA	Cohort	308	100.0	70.1 (9.3)	303	100.0	70.3 (9.0)	611
UK Biobank	UK	Cohort	5,356	42.4	61.7 (6.0)	21,407	42.4	59.9 (6.9)	26,783
Total			12,952			48,383			61,355

*5 studies included in a single pooled analysis. Abbreviations: MGI = Michigan Genomics Initiative; CORSA = Colorectal Cancer Study of Austria; DACHS 4 = Darmkrebs: Chancen der Verhütung durch Screening Study 4; COLON = Colorectal Cancer: Longitudinal Observational study on Nutritional and lifestyle factors that influence colorectal tumor recurrence, survival and quality of life; DACHS 3 = Darmkrebs: Chancen der Verhütung durch Screening Study 3; EPIC = European Prospective Investigation into Cancer; HPFS 3 = Health Professionals Follow-Up Study 3; NHS 3 = Nurses' Health Study 3; UK Biobank = United Kingdom Biobank

[†] Countries include Spain, Netherlands, Italy, Greece, Germany, France, Sweden, Greece, and UK.

Supplementary Table 4. Characteristics of East Asian-descent participants newly genotyped on the OncoArray and included in the multiethnic follow-up stage*

Study	Country	Design	Genotyping Center	Cases			Controls			Total
				N	% Female	Mean age, y (SD)	N	% Female	Mean age, y (SD)	
HCES-CRC	Korea	Case-Control	CIDR	3,130	35.5%	62.7 (11.2)	2,854	61.1%	64.2 (14.0)	5,984
Shanghai MHS and WHS†	China	Cohort	CIDR	245	49.4%	69.0 (2.6)	221	45.4%	69.1 (2.4)	466
Taiwan Study	Taiwan	Cohort	USC	486	47.1%	64.1 (14.1)	0	NA	NA	486
Total				3,855			3,081			6,936‡

*Individuals had ≥80% estimated East Asian ancestry based on STRUCTURE analysis. Abbreviations: HCES-CRC = The Hwasun Cancer Epidemiology Study-Colon and Rectum Cancer; CIDR = Center for Inherited Disease Research; USC = University of Southern California

†Shanghai Men's Health Study (MHS; N_{case}=121, $N_{control}$=124) and Shanghai Women's Health Study (WHS; N_{case}=118, $N_{control}$=103)

‡The multiethnic follow-up stage also includes 8,230 cases and 19,002 controls from non-OncoArray genotyping efforts for a total of 12,085 cases and 22,083 controls

Supplementary Table 5. Discovery GWAS results for previously published autosomal CRC susceptibility loci*

LOCUS	rsID:CHR:BP	EFF/ REF	FRQ EFF EUR	FRQ EFF EAS	FRQ EFF AFR	Discovery (Part 1) (CORECT/GECCO/CFR)		Discovery (Part 2) (OncoArray)		Combined (Parts 1 + 2)			
						OR (95%CI)	P†	OR (95%CI)	P†	OR (95%CI)	P†	I^2, %	$P_{heterogeneity}$‡
1q25.3	rs10911251:1: 183081194	A/C	0.547	0.481	0.829	1.07 (1.04 - 1.11)	5.18×10^{-5}	1.07 (1.03 - 1.11)	2.65×10^{-4}	1.07 (1.04 - 1.10)	4.93×10^{-8}	0	0.92
1q41	rs6691170:1: 222045446	T/G	0.401	0.004	0.347	1.02 (0.99 - 1.06)	0.14	1.07 (1.04 - 1.11)	1.23×10^{-4}	1.04 (1.02 - 1.07)	2.98×10^{-4}	73.9	0.05
1q41	rs6687758:1: 222164948	A/G	0.778	0.796	0.821	0.96 (0.92 - 0.99)	0.03	0.87 (0.83 - 0.91)	7.32×10^{-10}	0.92 (0.89 - 0.95)	1.47×10^{-8}	90.7	0.001
2q32.3	rs11903757:2: 192587204	T/C	0.830	0.933	0.860	0.92 (0.88 - 0.96)	3.69×10^{-4}	0.99 (0.94 - 1.04)	0.71	0.95 (0.92 - 0.99)	0.005	79.5	0.03
2q35	rs992157:2: 219154781	A/G	0.583	0.616	0.172	1.06 (1.03 - 1.10)	2.38×10^{-4}	1.07 (1.03 - 1.11)	1.46×10^{-4}	1.07 (1.04 - 1.09)	1.41×10^{-7}	0	0.76
3p22.1	rs35360328:3: 40924962	A/T	0.156	0.084	0.005	1.14 (1.09 - 1.19)	2.39×10^{-8}	1.12 (1.07 - 1.18)	5.10×10^{-6}	1.13 (1.09 - 1.17)	6.21×10^{-13}	0	0.66
3p14.1	rs812481:3: 66442435	C/G	0.441	0.208	0.075	0.92 (0.89 - 0.95)	2.50×10^{-8}	0.98 (0.95 - 1.02)	0.30	0.94 (0.92 - 0.97)	1.11×10^{-6}	87.9	0.004
3q26.2	rs10936599:3: 169492101	T/C	0.243	0.578	0.037	0.99 (0.95 - 1.03)	0.56	0.94 (0.91 - 0.98)	0.006	0.97 (0.94 - 1.00)	0.03	65	0.09
4q26	rs3987:4: 118759055	A/G	0.575	0.377	0.067	1.00 (0.97 - 1.03)	0.85	1.00 (0.96 - 1.03)	0.88	1.00 (0.98 - 1.02)	0.97	0	0.81
4q32.2	rs35509282:4: 163333405	A/T	0.110	0.406	0.171	1.08 (1.03 - 1.14)	0.002	0.97 (0.92 - 1.03)	0.32	1.03 (0.99 - 1.07)	0.11	87	0.005
5q31.1	rs647161:5: 134499092	A/C	0.660	0.302	0.517	1.05 (1.01 - 1.08)	0.006	1.07 (1.03 - 1.11)	4.86×10^{-4}	1.06 (1.03 - 1.09)	1.21×10^{-5}	0	0.44
6p21.2	rs1321311:6: 36622900	A/C	0.217	0.169	0.411	1.05 (1.01 - 1.08)	0.009	1.05 (1.01 - 1.09)	0.02	1.05 (1.02 - 1.08)	3.63×10^{-4}	0	0.91
6p21.1	rs4711689:6: 41692812	A/G	0.539	0.802	0.938	1.01 (0.97 - 1.04)	0.73	1.02 (0.99 - 1.06)	0.18	1.01 (0.99 - 1.04)	0.25	0	0.43
6q25.3	rs7758229:6: 160840252	T/G	0.319	0.246	0.048	0.98 (0.95 - 1.02)	0.33	1.03 (0.99 - 1.07)	0.18	1.00 (0.98 - 1.03)	0.91	63.2	0.10
8q23.3	rs140355816:8: 117574515	C/G	0.988	0.999	1.000	0.85 (0.76 - 0.94)	0.002	0.78 (0.68 - 0.90)	5.47×10^{-4}	0.82 (0.76 - 0.89)	4.97×10^{-6}	0	0.38
8q23.3	rs2450115:8: 117624093	T/C	0.807	0.539	0.821	1.07 (1.03 - 1.11)	5.36×10^{-4}	1.09 (1.04 - 1.14)	1.47×10^{-4}	1.08 (1.05 - 1.11)	3.40×10^{-7}	0	0.55
8q23.3	rs16892766:8: 117630683	A/C	0.910	0.999	0.853	0.83 (0.78 - 0.87)	2.72×10^{-12}	0.80 (0.76 - 0.85)	1.66×10^{-13}	0.82 (0.78 - 0.85)	3.94×10^{-24}	0	0.51
8q23.3	rs6469656:8: 117647788	A/G	0.890	0.673	0.830	1.06 (1.01 - 1.12)	0.01	1.06 (1.00 - 1.12)	0.04	1.06 (1.02 - 1.10)	9.12×10^{-4}	0	0.90
8q24.21	rs10505477:8: 128407443	A/G	0.479	0.389	0.871	1.12 (1.09 - 1.16)	6.89×10^{-14}	1.13 (1.10 - 1.17)	1.98×10^{-12}	1.13 (1.10 - 1.16)	7.89×10^{-25}	0	0.71
8q24.21	rs6983267:8: 128413305	T/G	0.501	0.612	0.048	0.89 (0.86 - 0.92)	3.64×10^{-14}	0.87 (0.84 - 0.90)	3.05×10^{-14}	0.88 (0.86 - 0.90)	7.74×10^{-27}	0	0.46
8q24.21	rs7014346:8: 128424792	A/G	0.335	0.286	0.401	1.12 (1.09 - 1.16)	2.59×10^{-13}	1.15 (1.11 - 1.19)	7.16×10^{-14}	1.13 (1.11 - 1.16)	1.66×10^{-25}	0	0.37
9p24.1	rs719725:9: 6365683	A/C	0.592	0.737	0.769	1.06 (1.02 - 1.09)	5.88×10^{-4}	1.03 (0.99 - 1.06)	0.17	1.04 (1.02 - 1.07)	4.50×10^{-4}	31.4	0.23
10p14	rs10795668:10: 8701219	A/G	0.320	0.368	0.020	0.97 (0.94 - 1.00)	0.05	0.87 (0.84 - 0.91)	6.97×10^{-12}	0.93 (0.90 - 0.95)	2.33×10^{-9}	93.5	9.27×10^{-5}
10p14	rs11255841:10: 8739580	A/T	0.309	0.362	0.092	0.95 (0.92 - 0.98)	0.004	0.86 (0.83, 0.90)	1.68×10^{-13}	0.91 (0.89 - 0.94)	3.16×10^{-12}	92.9	1.69×10^{-4}
10q22.3	rs704017:10: 80819132	A/G	0.438	0.721	0.442	0.94 (0.91 - 0.97)	2.07×10^{-4}	0.92 (0.89 - 0.96)	1.95×10^{-5}	0.93 (0.91 - 0.96)	1.96×10^{-8}	0	0.50
10q24.2	rs1035209:10: 101345366	T/C	0.195	0.179	0.067	1.09 (1.05 - 1.13)	1.26×10^{-5}	1.09 (1.04 - 1.14)	2.21×10^{-4}	1.09 (1.06 - 1.12)	1.03×10^{-8}	0	0.95
10q24.2	rs11190164:10: 101351704	A/G	0.719	0.770	0.966	0.91 (0.88 - 0.95)	8.39×10^{-7}	0.95 (0.91 - 0.99)	0.02	0.93 (0.90 - 0.95)	1.03×10^{-7}	44.4	0.18
10q24.32	rs4919687:10: 104595248	A/G	0.298	0.253	0.113	0.95 (0.92 - 0.99)	0.006	0.98 (0.95 - 1.02)	0.39	0.97 (0.94 - 0.99)	0.008	22.9	0.26
10q25.2	rs12241008:10: 114280702	T/C	0.900	0.700	0.793	0.92 (0.87 - 0.97)	0.002	0.96 (0.91 - 1.02)	0.19	0.94 (0.90 - 0.98)	0.002	13.1	0.28
10q25.2	rs10506868:10: 114319380	T/C	0.032	0.288	0.005	1.04 (0.94 - 1.14)	0.49	1.02 (0.93 - 1.12)	0.64	1.03 (0.96 - 1.10)	0.42	0	0.85
10q25.2	rs11196172:10: 114726843	A/G	0.119	0.647	0.053	1.05 (1.00 - 1.10)	0.04	1.07 (1.02 - 1.13)	0.005	1.06 (1.03 - 1.10)	6.05×10^{-4}	0	0.51
11q12.2	rs174537:11: 61552680	T/G	0.349	0.567	0.025	0.93 (0.90 - 0.96)	5.34×10^{-5}	0.94 (0.90 - 0.97)	5.57×10^{-4}	0.93 (0.91 - 0.96)	1.03×10^{-7}	0	0.93
11q12.2	rs4246215:11: 61564299	T/G	0.370	0.565	0.022	0.94 (0.91 - 0.97)	2.54×10^{-4}	0.94 (0.90 - 0.97)	6.83×10^{-4}	0.94 (0.92 - 0.96)	5.76×10^{-7}	0	0.91
11q12.2	rs174550:11: 61571478	T/C	0.653	0.434	0.978	1.07 (1.04 - 1.11)	4.32×10^{-5}	1.07 (1.03 - 1.11)	7.46×10^{-4}	1.07 (1.04 - 1.10)	1.24×10^{-7}	0	0.87
11q12.2	rs1535:11: 61597972	A/G	0.650	0.434	0.891	1.07 (1.04 - 1.11)	1.91×10^{-5}	1.07 (1.03 - 1.11)	5.76×10^{-4}	1.07 (1.05 - 1.10)	4.15×10^{-8}	0	0.85
11q13.4	rs3824999:11: 74345550	T/G	0.476	0.579	0.893	0.93 (0.90 - 0.96)	1.00×10^{-6}	0.93 (0.90 - 0.97)	9.24×10^{-5}	0.93 (0.91 - 0.95)	3.81×10^{-10}	0	0.82
11q23.1	rs3802842:11: 111171709	A/C	0.731	0.602	0.648	0.89 (0.86 - 0.92)	1.01×10^{-10}	0.88 (0.85 - 0.92)	9.53×10^{-10}	0.89 (0.87 - 0.91)	5.30×10^{-19}	0	0.69
12p13.32	rs10774214:12: 4368352	T/C	0.381	0.321	0.655	1.04 (1.01 - 1.08)	0.008	1.04 (1.01 - 1.08)	0.02	1.04 (1.02 - 1.07)	4.29×10^{-4}	0	0.98
12p13.32	rs3217810:12: 4388271	T/C	0.121	0.011	0.005	1.16 (1.09 - 1.23)	5.25×10^{-6}	1.15 (1.09 - 1.22)	7.03×10^{-7}	1.15 (1.11 - 1.20)	1.61×10^{-11}	0	0.90
12p13.32	rs3217901:12: 4405389	A/G	0.619	0.423	0.912	0.92 (0.89 - 0.95)	2.32×10^{-6}	0.96 (0.93 - 0.99)	0.02	0.94 (0.92 - 0.96)	5.90×10^{-7}	59.1	0.12
12p13.31	rs10849432:12: 6385727	T/C	0.903	0.813	0.703	1.06 (1.01 - 1.12)	0.02	1.12 (1.06 - 1.18)	1.54×10^{-4}	1.09 (1.04 - 1.13)	2.34×10^{-5}	45.5	0.18
12p13.31	rs11064437:12: 6982162	T/C	0.008	0.277	0.310	0.84 (0.63 - 1.13)	0.26	NA	NA	0.84 (0.63 - 1.13)	0.26	0	1.00
12q13.12	rs34245511:12: 50573433	C/G	0.362	0.192	0.475	1.04 (1.01 - 1.08)	0.02	1.05 (1.02 - 1.09)	0.005	1.05 (1.02 - 1.07)	3.00×10^{-4}	0	0.65
12q13.12	rs7136702:12: 50880216	T/C	0.344	0.442	0.648	1.06 (1.02 - 1.10)	0.005	NA	NA	1.06 (1.02 - 1.10)	0.005	0	1.00
12q13.12	rs11169552:12: 51155663	T/C	0.250	0.396	0.056	0.95 (0.92 - 0.98)	0.002	0.96 (0.93 - 1.00)	0.05	0.95 (0.93 - 0.98)	3.61×10^{-4}	0	0.581
12q24.12	rs3184504:12: 111884608	T/C	0.464	0.003	0.019	0.92 (0.89 - 0.94)	1.66×10^{-8}	0.94 (0.91 - 0.98)	0.001	0.93 (0.91 - 0.95)	1.94×10^{-10}	39.3	0.20
12q24.21	rs59336:12: 115116352	A/T	0.510	0.368	0.511	0.94 (0.91 - 0.97)	9.52×10^{-5}	0.96 (0.93 - 1.00)	0.03	0.95 (0.93 - 0.97)	1.34×10^{-5}	0	0.33
12q24.22	rs73208120:12: 117747590	T/G	0.916	1.000	0.994	0.86 (0.81 - 0.90)	2.84×10^{-8}	0.97 (0.90 - 1.03)	0.33	0.90 (0.86 - 0.94)	7.59×10^{-7}	86.5	0.006

14q22.2	rs4444235:14: 54410919	T/C	0.507	0.533	0.707	0.92 (0.89 - 0.95)	5.20×10^{-8}	0.95 (0.91 - 0.98)	0.002	0.93 (0.91 - 0.95)	1.09×10^{-9}	36.7	0.21
14q22.2	rs1957636:14: 54560018	T/C	0.411	0.659	0.790	1.04 (1.01 - 1.07)	0.01	1.06 (1.02 - 1.10)	0.002	1.05 (1.02 - 1.07)	6.42×10^{-5}	0	0.46
14q23.1	rs17094983:14: 59189361	A/G	0.121	0.000	0.143	0.87 (0.83 - 0.91)	6.97×10^{-9}	0.93 (0.88 - 0.98)	0.007	0.90 (0.87 - 0.93)	8.38×10^{-10}	69	0.07
15q13.3	rs16969681:15: 32993111	T/C	0.074	0.371	0.138	1.10 (1.04 - 1.16)	7.10×10^{-4}	1.11 (1.05 - 1.18)	5.01×10^{-4}	1.10 (1.06 - 1.15)	1.27×10^{-6}	0	0.72
15q13.3	rs4779584:15: 32994756	T/C	0.205	0.811	0.626	1.12 (1.08 - 1.16)	1.36×10^{-8}	1.13 (1.08 - 1.18)	1.16×10^{-7}	1.12 (1.09 - 1.16)	8.27×10^{-15}	0	0.79
15q13.3	rs11632715:15: 33004247	A/G	0.461	0.771	0.360	1.05 (1.02 - 1.09)	7.78×10^{-4}	1.07 (1.03 - 1.11)	2.36×10^{-4}	1.06 (1.04 - 1.08)	7.68×10^{-7}	0	0.57
15q13.3	rs73376930:15: 33012502	A/G	0.793	0.487	0.659	0.89 (0.86 - 0.93)	4.05×10^{-9}	0.86 (0.83 - 0.90)	8.37×10^{-11}	0.88 (0.85 - 0.91)	3.57×10^{-18}	13.9	0.28
16q22.1	rs9929218:16: 68820946	A/G	0.294	0.233	0.267	0.95 (0.92 - 0.98)	0.002	0.94 (0.90 - 0.98)	0.001	0.94 (0.92 - 0.97)	7.30×10^{-6}	0	0.70
17p13.3	rs12603526:17: 800593	T/C	0.989	0.801	0.998	0.92 (0.81 - 1.05)	0.20	0.88 (0.78 - 1.00)	0.06	0.90 (0.83 - 0.99)	0.02	0	0.66
18q21.1	rs7229639:18: 46450976	A/G	0.099	0.132	0.210	1.05 (1.00 - 1.11)	0.07	1.14 (1.07 - 1.21)	5.67×10^{-5}	1.08 (1.04 - 1.13)	6.92×10^{-5}	73	0.05
18q21.1	rs4939827:18: 46453463	T/C	0.532	0.311	0.305	1.13 (1.10 - 1.17)	2.04×10^{-15}	1.16 (1.12 - 1.20)	1.38×10^{-16}	1.14 (1.12 - 1.17)	3.41×10^{-30}	8.9	0.30
19q13.11	rs10411210:19: 33532300	T/C	0.097	0.192	0.439	0.94 (0.89 - 0.99)	0.02	0.87 (0.83 - 0.92)	8.02×10^{-7}	0.91 (0.87 - 0.94)	3.44×10^{-7}	74.9	0.05
19q13.2	rs1800469:19: 41860296	A/G	0.312	0.547	0.218	0.98 (0.95 - 1.02)	0.31	0.95 (0.91 - 0.99)	0.006	0.97 (0.94 - 0.99)	0.01	46.8	0.17
19q13.2	rs2241714:19: 41869392	T/C	0.325	0.554	0.139	0.99 (0.95 - 1.02)	0.39	0.96 (0.92 - 0.99)	0.02	0.97 (0.95 - 1.00)	0.03	19.6	0.27
20p12.3	rs961253:20: 6404281	A/C	0.361	0.107	0.367	1.09 (1.05 - 1.12)	1.29×10^{-7}	1.08 (1.04 - 1.12)	1.02×10^{-4}	1.08 (1.06 - 1.11)	5.79×10^{-11}	0	0.63
20p12.3	rs4813802:20: 6699595	T/G	0.682	0.770	0.916	0.92 (0.89 - 0.95)	7.33×10^{-7}	0.93 (0.90 - 0.97)	3.58×10^{-4}	0.93 (0.90 - 0.95)	1.24×10^{-9}	0	0.51
20p12.3	rs2423279:20: 7812350	T/C	0.729	0.667	0.615	0.95 (0.92 - 0.99)	0.007	0.93 (0.90 - 0.97)	5.61×10^{-4}	0.94 (0.92 - 0.97)	1.57×10^{-5}	0	0.49
20q13.13	rs6066825:20: 47340117	A/G	0.622	0.711	0.223	1.07 (1.03 - 1.10)	8.74×10^{-5}	1.05 (1.01 - 1.09)	0.007	1.06 (1.03 - 1.09)	2.29×10^{-6}	0	0.62
20q13.33	rs4925386:20: 60921044	T/C	0.332	0.259	0.825	0.94 (0.91 - 0.97)	3.77×10^{-4}	0.89 (0.85 - 0.92)	5.52×10^{-10}	0.92 (0.89 - 0.94)	1.63×10^{-11}	82.6	0.02
20q13.33	rs6061231:20: 60956917	A/C	0.304	0.156	0.269	0.93 (0.90 - 0.97)	0.001	0.89 (0.86 - 0.93)	1.37×10^{-8}	0.91 (0.89 - 0.94)	2.33×10^{-10}	61.3	0.11
20q13.33	rs2427308:20: 60969451	T/C	0.234	0.150	0.205	0.92 (0.88 - 0.96)	1.14×10^{-4}	0.87 (0.83 - 0.90)	7.88×10^{-11}	0.89 (0.86 - 0.92)	1.75×10^{-13}	67.7	0.08

* References and original ancestral populations where each locus was discovered are described in Supplementary Table 6. Abbreviations: CHR = chromosome; BP = position; EFF = effect allele; REF = reference allele (reference category for the odds ratios); FRQ = frequency; 1KGP = 1000 Genomes; EUR = 1KGP European; EAS = 1KGP East Asian; AFR =1KGP African; CORECT = Colorectal Transdisciplinary Study; GECCO = Genetics and Epidemiology of Colorectal Cancer Consortium; CFR = Colon Cancer Family Registry.
† P values were derived from a fixed-effects inverse variance weighted meta-analysis (Part 1) and unconditional logistic regression (Part 2). All tests were two-sided.
‡ P values were derived from Cochran's *Q* test of heterogeneity. All tests were two-sided.

Supplementary Table 6. Bias-corrected effect estimates for previously published (N=67) and novel (N=9) CRC susceptibility alleles in the discovery GWAS.*

rsID	Known/Novel	Original Estimate	Bias-Corrected Estimate	Reference(s)		Ancestry
		OR (95%CI)	OR (95%CI)	Author/Year	PMID	
rs10911251	Known	0.93 (0.91, 0.96)	0.93 (0.91, 0.96)	Peters et al. 2013; Whiffin et al. 2014	23266556;24737748	EUR
rs6691170†	Known	0.96 (0.94, 0.98)	0.96 (0.94, 0.98)	Houlston et al. 2010	20972440	EUR
rs6687758	Known	1.09 (1.05, 1.13)	1.09 (1.05, 1.13)	Houlston et al. 2010	20972440	EUR
rs11903757	Known	1.05 (1.02, 1.08)	1.05 (1.02, 1.08)	Peters et al. 2013	23266556	EUR
rs992157	Known	0.93 (0.91, 0.96)	0.93 (0.91, 0.96)	Orlando et al. 2016	27005424	EUR
rs35360328	Known	0.89 (0.85, 0.92)	0.89 (0.85, 0.92)	Schumacher et al. 2015	26151821	EUR,ASN
rs812481	Known	1.06 (1.04, 1.08)	1.06 (1.04, 1.08)	Schumacher et al. 2015	26151821	EUR,ASN
rs10936599	Known	1.03 (1.00, 1.06)	1.03 (1.00, 1.06)	Houlston et al. 2010	20972440	EUR
rs1370821	Novel	1.07 (1.04, 1.10)	1.07 (1.04, 1.10)	-	-	-
rs3987	Known	1.00 (0.98, 1.02)	1.00 (0.98, 1.02)	Real et al. 2014	24978480	EUR
rs35509282	Known	0.97 (0.93, 1.01)	0.97 (0.93, 1.01)	Schmit et al. 2014	25023989	EUR
rs2735940	Novel	0.92 (0.89, 0.95)	0.92 (0.89, 0.95)	-	-	-
rs58791712	Novel	0.91 (0.89, 0.93)	0.91 (0.89, 0.93)	-	-	-
rs647161	Known	0.94 (0.92, 0.97)	0.94 (0.92, 0.97)	Jia et al. 2013	23263487	ASN,EUR
rs6906359	Novel	0.91 (0.86, 0.94)	0.91 (0.87, 0.95)	-	-	-
rs1321311	Known	0.95 (0.92, 0.98)	0.95 (0.92, 0.98)	Dunlop et al. 2012	22634755	EUR
rs4711689	Known	0.99 (0.97, 1.01)	0.99 (0.97, 1.01)	Zeng et al. 2016	26965516	ASN,EUR
rs62404968	Novel	0.92 (0.89, 0.95)	0.92 (0.89, 0.95)	-	-	-
rs7758229	Known	1.00 (0.98, 1.02)	1.00 (0.98, 1.02)	Cui et al. 2011	21242260	ASN
rs140355816	Known	1.22 (1.13, 1.32)	1.22 (1.13, 1.32)	Whiffin et al. 2014	24737748	EUR
rs2450115	Known	0.93 (0.90, 0.95)	0.93 (0.90, 0.95)	Tomlinson et al. 2008	18372905	EUR,ASN
rs16892766	Known	1.22 (1.16, 1.28)	1.22 (1.16, 1.28)	Tomlinson et al. 2008	18372905	EUR,ASN
rs6469656†	Known	0.94 (0.91, 0.98)	0.94 (0.91, 0.98)	Tomlinson et al. 2008	18372905	EUR,ASN
rs10505477†	Known	0.89 (0.86, 0.91)	0.89 (0.86, 0.91)	Zanke et al. 2007; Gruber et al. 2007	17618283;17630503	EUR
rs6983267	Known	1.14 (1.11, 1.17)	1.14 (1.11, 1.17)	Zanke et al. 2007; Haiman et al. 2007; Tomlinson et al. 2007; Hutter et al. 2010; Cui et al.2011	17618283;17618282; 17618284;21129217; 21242260	EUR,ASN
rs7014346†	Known	0.89 (0.87, 0.90)	0.89 (0.87, 0.90)	Tenesa et al. 2008; Houlston et al. 2008	18372901;19011631	EUR

rs719725	Known	0.96 (0.94, 0.98)	0.96 (0.94, 0.98)	Zanke et al. 2007	17618283	EUR
rs10795668†	Known	1.08 (1.04, 1.12)	1.08 (1.04, 1.12)	Tomlinson et al. 2008	18372905	EUR
rs11255841	Known	0.91 (0.89, 0.93)	0.91 (0.89, 0.93)	Whiffin et al. 2014	24737748	EUR
rs10994860	Novel	1.09 (1.05, 1.13)	1.09 (1.05, 1.12)	-	-	-
rs704017	Known	0.93 (0.91, 0.95)	0.93 (0.91, 0.95)	Zhang et al. 2014	24836286	ASN,EUR
rs1035209	Known	0.92 (0.89, 0.94)	0.92 (0.89, 0.94)	Whiffin et al. 2014	24737748	EUR,ASN
rs11190164†	Known	1.08 (1.04, 1.12)	1.08 (1.04, 1.12)	Whiffin et al. 2014; Schumacher et al. 2015	24737748;26151821	EUR,ASN
rs4919687	Known	1.03 (1.00, 1.06)	1.03 1.00, 1.06)	Zeng et al. 2016	26965516	ASN,EUR
rs12241008	Known	1.06 (1.02, 1.11)	1.06 (1.01, 1.11)	Wang et al. 2014	25105248	ASN,AA,EUR
rs10506868†	Known	0.97 (0.90, 1.04)	0.97 (0.90, 1.04)	Wang et al. 2014	25105248	ASN,AA,EUR
rs11196172	Known	0.94 (0.92, 0.97)	0.94 (0.92, 0.97)	Zhang et al. 2014	24836286	ASN,EUR
rs174537	Known	1.08 (1.06, 1.10)	1.08 (1.06, 1.10)	Zhang et al. 2014	24836286	ASN,EUR
rs4246215†	Known	1.06 (1.04, 1.08)	1.06 (1.04, 1.08)	Zhang et al. 2014	24836286	ASN,EUR
rs174550†	Known	0.93 (0.91, 0.96)	0.93 (0.91, 0.96)	Zhang et al. 2014	24836286	ASN,EUR
rs1535†	Known	0.93 (0.92, 0.95)	0.93 (0.92, 0.95)	Zhang et al. 2014	24836286	ASN,EUR
rs3824999	Known	1.08 (1.06, 1.10)	1.08 (1.06, 1.10)	Dunlop et al. 2012	22634755	EUR
rs3802842	Known	1.12 (1.09, 1.15)	1.12 (1.09, 1.15)	Tenesa et al. 2008	18372901	EUR
rs10774214	Known	0.96 (0.94, 0.98)	0.96 (0.94, 0.98)	Jia et al. 2013	23263487	ASN,EUR
rs3217810	Known	0.87 (0.84, 0.90)	0.87 (0.84, 0.90)	Peters et al. 2013; Whiffin et al. 2014	23266556;24737748	EUR,ASN
rs3217901	Known	1.06 (1.04, 1.08)	1.06 (1.04, 1.08)	Peters et al. 2013	23266556	EUR,ASN
rs10849432	Known	0.92 (0.88, 0.96)	0.92 (0.88, 0.96)	Zhang et al. 2014	24836286	ASN,EUR
rs34245511	Known	0.95 (0.92, 0.98)	0.95 (0.92, 0.98)	Whiffin et al. 2014	24737748	EUR
rs11169552	Known	1.05 (1.03, 1.07)	1.05 (1.03, 1.07)	Houlston et al. 2010	20972440	EUR
rs3184504	Known	1.08 (1.06, 1.10)	1.08 (1.06, 1.10)	Schumacher et al. 2015	26151821	EUR,ASN
rs59336	Known	1.05 (1.03, 1.07)	1.05 (1.03, 1.07)	Peters et al. 2013	23266556	EUR,ASN
rs72013726	Novel	0.93 (0.90, 0.96)	0.93 (0.90, 0.96)	-	-	-
rs73208120	Known	1.11 (1.06, 1.16)	1.11 (1.06, 1.16)	Schumacher et al. 2015	26151821	EUR,ASN
rs4444235	Known	1.08 (1.06, 1.10)	1.08 (1.06, 1.10)	Houlston et al. 2008; Tomlinson et al. 2011	19011631;21655089	EUR
rs1957636	Known	0.95 (0.92, 0.98)	0.95 (0.92, 0.98)	Tomlinson et al. 2011	21655089	EUR
rs17094983	Known	1.11 (1.07, 1.15)	1.11 (1.07, 1.15)	Lemire et al. 2015	26404086	EUR
rs16969681†	Known	0.91 (0.88, 0.94)	0.91 (0.88, 0.94)	Tomlinson et al. 2008; 2011	18372905;21655089	EUR
rs4779584	Known	0.89 (0.87, 0.92)	0.89 (0.87, 0.92)	Tomlinson et al. 2008	18372905	EUR
rs11632715†	Known	0.94 (0.93, 0.96)	0.94 (0.93, 0.96)	Tomlinson et al. 2008	18372905	EUR
rs73376930†	Known	1.14 (1.10, 1.18)	1.14 (1.10, 1.18)	Whiffin et al. 2014	24737748	EUR
rs9929218	Known	1.06 (1.04, 1.08)	1.06 (1.04, 1.08)	Houlston et al. 2008	19011631	EUR
rs2696839	Novel	0.95 (0.92, 0.96)	0.95 (0.93, 0.97)	-	-	-
rs12603526	Known	1.11 (1.02, 1.20)	1.11 (1.02, 1.20)	Zhang et al. 2014	24836286	ASN.EUR

rs7229639	Known	0.93 (0.89, 0.96)	0.93 (0.89, 0.96)	Zhang et al. 2014	24836286	ASN.EUR
rs4939827	Known	0.88 (0.86, 0.89)	0.88 (0.86, 0.89)	Broderick et al. 2007; Tenesa et al. 2008	17934461;18372901	EUR
rs10411210	Known	1.10 (1.05, 1.15)	1.10 (1.05, 1.15)	Houlston et al. 2008	19011631	EUR
rs1800469	Known	1.03 (1.00, 1.06)	1.03 (1.00, 1.06)	Zhang et al. 2014	24836286	ASN,EUR
rs2241714†	Known	1.03 (1.01, 1.05)	1.03 (1.01, 1.05)	Zhang et al. 2014	24836286	ASN,EUR
rs961253	Known	0.93 (0.91, 0.94)	0.93 (0.91, 0.94)	Houlston et al. 2008	19011631	EUR
rs4813802	Known	1.08 (1.04, 1.12)	1.08 (1.04, 1.12)	Tomlinson et al. 2011; Peters et al. 2012	21655089; 21761138	EUR
rs2423279	Known	1.06 (1.04, 1.08)	1.06 (1.04, 1.08)	Jia et al. 2008	23263487	ASN.EUR
rs6066825	Known	0.94 (0.92, 0.97)	0.94 (0.92, 0.97)	Schumacher et al. 2015	26151821	EUR,ASN
rs1810502	Novel	0.93 (0.91, 0.95)	0.93 (0.91, 0.95)	-	-	-
rs4925386†	Known	1.09 (1.05, 1.13)	1.09 (1.05, 1.13)	Houlston et al. 2010	20972440	EUR
rs6061231†	Known	1.10 (1.08, 1.12)	1.10 (1.08, 1.12)	Zeng et al. 2016	26965516	ASN
rs2427308	Known	1.12 (1.08, 1.16)	1.12 (1.08, 1.16)	Whiffin et al. 2014	24737748	EUR

* rs11064437 and rs7136702 were excluded from known susceptibility alleles list because they were not available in Discovery Part 2. Abbreviations: OR = odds ratio; CI = confidence interval; ASN = East Asian; EUR = European.

† Variants were excluded from the familial relative risk explained analysis according to pruning procedures described in our Supplementary Methods.

Supplementary Table 7. Results from a weighted polygenic risk score analysis of known and novel risk alleles in the European replication dataset and the East Asian OncoArray dataset.

Dataset and Risk Score Category	Known (67 variants)		Known + Novel (76 variants)	
	N (%)	OR (95%CI)	N (%)	OR (95%CI)
European replication dataset*				
25-75%	18369 (49.7)	1.00 (Reference)	18299 (49.5)	1.00 (Reference)
<1%	309 (0.8)	0.40 (0.29, 0.55)	302 (0.8)	0.35 (0.25, 0.50)
1-10%	2908 (7.9)	0.56 (0.50, 0.62)	2917 (7.9)	0.57 (0.51, 0.63)
10-25%	5159 (13.9)	0.76 (0.70, 0.82)	5178 (14.0)	0.79 (0.73, 0.85)
75-90%	5923 (16.0)	1.29 (1.20, 1.38)	5969 (16.1)	1.35 (1.26, 1.44)
90-99%	3855 (10.4)	1.65 (1.53, 1.78)	3844 (10.4)	1.69 (1.56, 1.82)
>99%	464 (1.3)	2.02 (1.66, 2.47)	478 (1.3)	2.18 (1.79, 2.65)
European replication dataset at a clinically actionable threshold of OR ≥ 2.00*				
25-75%	18369 (49.7)	1.00 (Reference)	18299 (49.5)	1.00 (Reference)
<1%	309 (0.8)	0.40 (0.29, 0.55)	302 (0.8)	0.35 (0.25, 0.50)
1-10%	2908 (7.9)	0.56 (0.50, 0.62)	2917 (7.9)	0.57 (0.51, 0.63)
10-25%	5159 (13.9)	0.76 (0.70, 0.82)	5178 (14.0)	0.79 (0.73, 0.85)
75-90%	5923 (16.0)	1.29(1.20, 1.38)	5969 (16.1)	1.35(1.26, 1.44)
90-95.7%	2376 (6.4)	1.52 (1.38, 1.67)	2356 (6.4)	1.54 (1.40, 1.69)
>95.7%	1943 (5.3)	1.91 (1.73, 2.12)	1966 (5.3)	2.00 (1.81, 2.21)
East Asian OncoArray dataset at a clinically actionable threshold of OR ≥ 2.00				
25-75%	3392 (48.9)	1.00 (Reference)	3400 (49.0)	1.00 (Reference)
<1%	55 (0.8)	0.54 (0.29, 1.03)	57 (0.8)	0.51 (0.27, 0.96)
1-10%	515 (7.4)	0.65 (0.53, 0.80)	507 (7.3)	0.63 (0.50, 0.77)
10-25%	903 (13.0)	0.72 (0.61, 0.85)	888 (12.8)	0.69 (0.58, 0.81)
75-90%	1191 (17.2)	1.30 (1.12, 1.51)	1194 (17.2)	1.29 (1.11, 1.49)
90-99.1%	795 (11.5)	1.58 (1.33, 1.88)	789 (11.4)	1.60 (1.35, 1.91)

>99.1%	85 (1.2)	1.62 (1.00, 2.64)	101 (1.5)	2.01 (1.27, 3.21)

*MGI and CORSA participants were excluded from this analysis. OR = odds ratio; CI = confidence interval.

Supplementary Table 8. Summary of functional annotation and eQTL results for 9 novel CRC susceptibility alleles.

rsID	Region	Locus	Gene or Nearest Genes	Index SNP overlaps a histone mark	SNPs in LD overlapping with histone marks*	Significant GTEx eQTL (Transverse Colon)			Significant COLONOMICS eQTL		
						Gene	Effect Size	*P**	Gene	Partial r^2	*P*†
rs1370821‡	intergenic	4q22.2	*ATOH1 SMARCAD1*	No	rs2510787, rs2433324	None	-	-	*BMPR1B*	0.03	0.04
rs2735940‡	intergenic	5p15.33	*TERT CLPTM1L*	No	rs380145, rs426995, rs246994	None	-	-	NA	NA	NA
rs58791712	intergenic	5p13.1	*PTGER4 LINC00603*	No	rs72748452, rs755989, rs4957261	None	-	-	*FYB*	0.04	0.02
									LIFR	0.04	0.03
									FBXO4	0.03	0.03
rs6906359	intergenic	6p21.31	*TULP1 FKBP5*	No	rs72894781, rs72894784, rs16878812, rs45493300	None	-	-	*BRPF3*	0.09	2.60×10^{-4}
									NUDT3	0.04	0.02
									RPS10-NUDT3	0.04	0.02
									SCUBE3	0.05	0.006
									DEF6	0.05	0.008
									CLPSL1	0.03	0.05
									LHFPL5	0.03	0.05
									MAPK13	0.03	0.03
									PXT1	0.03	0.03
									STK38	0.03	0.03
rs62404968	intron	6p12.1	*BMP5*	No	None	None	-	-	None	-	-
rs10994860	exon	10q11.23	*A1CF*	No	rs71457593, rs10994720	*ASAH2*	-0.61	5.7×10^{-5}	None	-	-
rs72013726	intergenic	12q24.21	*TBX3 MED13L*	Yes	NA	None	-	-	None	-	-
rs2696839	intergenic	16q24.1	*LOC146513 LINC00917*	No	rs12932862, rs12149163, rs12149501, rs2665316	None	-	-	None	-	-
rs1810502	intergenic	20q13.13	*LINC01271 PTPN1*	Yes	14 SNPs	None	-	-	*MOCS3*	0.03	0.03
									ADNP	0.03	0.04

* P values were derived from a two-tailed t test. Abbreviations: eQTL = expression quantitative trait locus; SNP = single nucleotide polymorphism; GTEx = Genotype-Tissue Expression program.
† P values were derived from Pearson partial correlation adjusted for tissue type (healthy or adjacent normal to tumor). All tests were two-sided.
‡ r^2>0.6 except for two indicated variants where r^2>0.2 was used.