nature COMMUNICATIONS

ARTICLE

Received 28 Feb 2014 | Accepted 10 Apr 2015 | Published 7 Jul 2015

DOI: 10.1038/ncomms8138

Genome-wide association study of colorectal cancer identifies six new susceptibility loci

Fredrick R. Schumacher[1,*], Stephanie L. Schmit[1,2,*], Shuo Jiao[3,*], Christopher K. Edlund[1], Hansong Wang[4], Ben Zhang[5], Li Hsu[3], Shu-Chen Huang[1], Christopher P. Fischer[6], John F. Harju[6], Gregory E. Idos[1], Flavio Lejbkowicz[7,8], Frank J. Manion[6], Kevin McDonnell[1], Caroline E. McNeil[1], Marilena Melas[1], Hedy S. Rennert[7,8], Wei Shi[9], Duncan C. Thomas[1], David J. Van Den Berg[1], Carolyn M. Hutter[10], Aaron K. Aragaki[3], Katja Butterbach[11], Bette J. Caan[12], Christopher S. Carlson[3], Stephen J. Chanock[13], Keith R. Curtis[3], Charles S. Fuchs[14,15], Manish Gala[16], Edward L. Giocannucci[17,18], Stephanie M. Gogarten[19], Richard B. Hayes[20], Brian Henderson[1], David J. Hunter[21], Rebecca D. Jackson[22], Laurence N. Kolonel[23], Charles Kooperberg[3], Sebastian Kury[24], Andrea LaCroix[3], Cathy C. Laurie[19], Cecelia A. Laurie[19], Mathiew Lemire[25], David Levine[19], Jing Ma[26], Karen W. Makar[3], Conghui Qu[3], Darin Taverna[27], Cornelia M. Ulrich[3,28,29], Kana Wu[30], Suminori Kono[31], Dee W. West[32], Sonja I. Berndt[13], Stéphane Bezieau[33], Hermann Brenner[11], Peter T. Campbell[34], Andrew T. Chan[16,17], Jenny Chang-Claude[35], Gerhard A. Coetzee[1], David V. Conti[1,36], David Duggan[37], Jane C. Figueiredo[1], Barbara K. Fortini[1], Steven J. Gallinger[38], W. James Gauderman[1], Graham Giles[39], Roger Green[40], Robert Haile[41], Tabitha A. Harrison[3], Michael Hoffmeister[11], John L. Hopper[42], Thomas J. Hudson[43], Eric Jacobs[34], Motoki Iwasaki[44], Sun Ha Jee[45], Mark Jenkins[46], Wei-Hua Jia[47], Amit Joshi[48], Li Li[49], Noralene M. Lindor[50], Keitaro Matsuo[31], Victor Moreno[51], Bhramar Mukherjee[52], Polly A. Newcomb[53], John D. Potter[53], Leon Raskin[1,54,55], Gad Rennert[1,7,8,56], Stephanie Rosse[3], Gianluca Severi[39,57], Robert E. Schoen[58], Daniela Seminara[59], Xiao-Ou Shu[55,60], Martha L. Slattery[61], Shoichiro Tsugane[44], Emily White[3], Yong-Bing Xiang[62], Brent W. Zanke[63,64], Wei Zheng[54,55,**], Loic Le Marchand[4,**], Graham Casey[1,**], Stephen B. Gruber[1,2,**] & Ulrike Peters[3,**]

Genetic susceptibility to colorectal cancer is caused by rare pathogenic mutations and common genetic variants that contribute to familial risk. Here we report the results of a two-stage association study with 18,299 cases of colorectal cancer and 19,656 controls, with follow-up of the most statistically significant genetic loci in 4,725 cases and 9,969 controls from two Asian consortia. We describe six new susceptibility loci reaching a genome-wide threshold of $P<5.0E-08$. These findings provide additional insight into the underlying biological mechanisms of colorectal cancer and demonstrate the scientific value of large consortia-based genetic epidemiology studies.

[1] Department of Preventive Medicine, USC Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, California 90089, USA. [2] Department of Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90089, USA. [3] Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98124, USA. [4] Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii 96822, USA. [5] Department of General Surgery, Third Military Medical University Southwest Hospital, Chongqing 400038, China. [6] University of Michigan Comprehensive Cancer Center, Ann Arbor, Michigan 48105, USA. [7] Department of Community Medicine and Epidemiology, Carmel Medical Center, Haifa 34361, Israel. [8] Clalit Health Services National Cancer Control Center, Haifa 34361, Israel. [9] Department of Surgery, Children's Hospital Los Angeles, Los Angeles, California 90027, USA. [10] Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Rockville, Maryland 20892, USA. [11] Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg 69121, Germany. [12] Division of Research, Kaiser Permanente Medical Care Program of Northern California, Oakland, California 94612, USA. [13] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland 20892-9776, USA. [14] Department of Medicine, Brigham and Women's Hospital, Brookline, Massachusetts 02115, USA. [15] Department of Medical Oncology, Dana-Farber Cancer Institute, Brookline, Massachusetts 02115, USA. [16] Division of Gastroenterology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [17] Harvard Medical School, Boston, Massachusetts 02114, USA. [18] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. [19] Department of Biostatistics, University of Washington, Seattle, Washington 98195, USA. [20] Division of Epidemiology, Department of Population Health, New York University School of Medicine, New York, New York 10016, USA. [21] Program in Genetic Epidemiology and Statistical Genetics, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [22] Department of Medicine, Ohio State University, Columbus, Ohio 43210, USA. [23] Office of Public Health Studies, University of Hawaii Manoa, Honolulu, Hawaii 96822, USA. [24] Service de Génétique Médicale, CHU Nantes, Nantes 44093, France. [25] Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada. [26] Harvard School of Public Health, Boston, Massachusetts 02114, USA. [27] Phoenix College, Phoenix, Arizona 85013, USA. [28] Division of Preventive Oncology, German Cancer Research Center, Heidelberg 69120, Germany. [29] Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington 98195, USA. [30] Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [31] Department of Preventive Medicine, Kyushu University, Fukuoka 812-8582, Japan. [32] Cancer Registry of Greater California, Public Health Institute, Sacramento, California 95825, USA. [33] Centre Hospitalier Universitaire Hotel-Dieu, Nantes, 44093, France. [34] Epidemiology Research Program, American Cancer Society, Atlanta, Georgia 30329-4251, USA. [35] Unit of Genetic Epidemiology, Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg 69121, Germany. [36] Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, California 90089, USA. [37] Genetic Basis of Human Disease Division, Translational Genomics Research Institute, Phoenix, Arizona 85004, USA. [38] Zane Cohen Centre for Digestive Diseases, Mount Sinai Hospital, Toronto, Ontario M5T 3L9, Canada. [39] Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Victoria 3004, Australia. [40] Discipline of Genetics, Memorial University of Newfoundland, St. John's, Newfoundland A1B 3V6, Canada. [41] Department of Medicine, Division of Oncology, Stanford University, Stanford, California 94305, USA. [42] Centre for MEGA Epidemiology, The University of Melbourne, Carlton, Victoria 3010, Australia. [43] Department of Genomics, Ontario Institute for Cancer Research, Toronto, Ontario M5G 0A3, Canada. [44] Research Center for Cancer Prevention and Screening, National Cancer Center, Tokyo 104-0045, Japan. [45] Department of Epidemiology and Health Promotion, Graduate School of Public Health, Yonsei University, Seoul 120-749, South Korea. [46] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Parkville, Victoria 3010, Australia. [47] State Key Laboratory of Oncology in South China, Cancer Center, Sun Yat-sen University, Guangzhou 510060, China. [48] Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA. [49] Department of Family Medicine and Community Health, Case Western Reserve University, Cleveland, Ohio 44106, USA. [50] Department of Health Science Research, Mayo Clinic, Scottsdale, Arizona 85259, USA. [51] Cancer Epidemiology Service, Catalan Institute of Oncology, IDIBELL, 08908, Barcelona, Spain. [52] Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA. [53] Cancer Prevention Program, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. [54] Division of Epidemiology, Vanderbilt Epidemiology Center, Vanderbilt University School of Medicine, Nashville, Tennessee 37203-1738, USA. [55] Vanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville, Tennessee 37203-1738, USA. [56] Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology, Haifa 3200003, Israel. [57] Human Genetics Foundation (HuGeF), Torino 10126, Italy. [58] Department of Internal Medicine, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania 15213, USA. [59] Epidemiology and Genomics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. [60] Division of Epidemiology, Vanderbilt University School of Medicine, Nashville, Tennessee 37203-1738, USA. [61] Department of Internal Medicine, University of Utah Health Sciences Center, Salt Lake City, Utah 84132, USA. [62] Department of Epidemiology, Shanghai Cancer Institute, Shanghai 2200-25, China. [63] The University of Ottawa, Ottawa, Ontario K1N 6N5, Canada. [64] Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario K1Y 4E9, Canada. * These authors contributed equally to this work. ** These authors jointly supervised this work. Correspondence and requests for materials should be addressed to S.B.G. (email: sgruber@usc.edu).

The estimated lifetime risk of colorectal cancer (CRC) is 5.2% for men and 4.8% for women in the United States[1]. The narrow-sense heritability estimates based on twin and family studies of CRC range from 12 to 35% (refs 2,3). Although several genome-wide association studies (GWAS) of CRC have successfully identified common single-nucleotide polymorphisms (SNPs) associated with CRC risk[4–21], a large fraction of the heritability still remains elusive[22]. Our GWAS combines data from four large CRC consortia, the Colorectal Cancer Transdisciplinary (CORECT) Study, the Colon Cancer Family Registry (CFR), the Molecular Epidemiology of Colorectal Cancer (MECC) Study and the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) to elucidate previously undiscovered susceptibility loci for CRC. The current meta-analysis identifies novel genome-wide significant risk regions at 3p14.1, 3p22.1, 10q24.2, 12q24.12, 12q24.22 and 20q13.13.

Results

Study Populations and Population Stratification. Data for this discovery analysis focuses on individuals of the European ancestral heritage from North America, Australia and Europe. Our discovery analysis includes 19 observational studies genotyped with high-density SNP arrays and imputed to the 1,000 Genomes Project March 2012 reference panel[23,24] (Supplementary Table 1). We employ an inverse-variance-weighted fixed-effects meta-analysis of study-specific logistic regression results after filtering data for quality control (QC). Quantile–quantile plots show no appreciable evidence of population stratification for the meta-analysis (Supplementary Fig. 1) or by the individual discovery studies (Supplementary Fig. 2) before and after adjustment for principal components (PCs), and the sample size-corrected marginal lambda (equivalent to 1,000 cases and 1,000 controls) measures 1.003 in the discovery meta-analysis. The PC plots for ancestry indicate no difference between cases and controls in the respective discovery GWAS studies (Supplementary Fig. 3).

Confirmation of Prior Studies and Discovery. We evaluate the quality and effectiveness of our study design and analytic methods by assessing previously reported CRC susceptibility loci. We replicate the results for 41 of the 47 ($P<0.05$, unadjusted for multiple testing) published autosomal susceptibility variants for CRC (Supplementary Table 2). We turn our attention to the discovery of new susceptibility loci (Supplementary Fig. 4) by investigating the top 200 independent loci detected in the European discovery phase (Supplementary Table 3) in two separate East Asian consortia. Overall, our combined meta-analysis across European and Asian studies discovers six new susceptibility loci reaching a statistical threshold of $P<5.0E-08$: at chromosome 3p22.1 (rs35360328), 3p14.1 (rs812481), 10q24.2 (rs11190164), 12q24.12 (rs7137828), 12q24.22 (rs73208120) and 20q13.13 (rs6066825; Table 1; Supplementary Table 4). The odds ratios (ORs) across these six loci indicate a range of a 9 to 16% increase in the odds of developing CRC per risk allele, similar to previously reported CRC susceptibility loci. A seventh susceptibility locus tagged by rs4946260 at 6q22.1 approaches genome-wide significance ($P=6.27E-08$). The ORs are consistent across populations and genotyping platforms as shown by chi-square tests for heterogeneity, with only one locus showing marginally significant heterogeneity (rs6066825/20q13.13, $P_{het}=0.04$).

Replication in Asian Populations. Forest plots for the six genome-wide significant loci show that the risk alleles identified in the European populations replicate broadly across the Asian populations even though allele frequencies differ substantially (Fig. 1). Two SNPs were not available for replication in the Asian studies because they are rare in Asians.

Genomic Location and Candidate Genes. Several of the six susceptibility SNPs fall within regions harbouring genes known to be involved in the pathogenesis of CRC (Supplementary Fig. 5). Rs35360328 and a corresponding tagSNP at 3p22.1 (rs35364139, $r^2=0.8$, $P=1.7E-07$) lie in an intergenic region within ~300 kb of *CTNNB1*, the gene that encodes β-catenin. β-Catenin is a key member of the WNT signalling pathway and is commonly mutated in CRC development[25,26]. There are no histone marks in the vicinity of either rs35360328 or rs35364139 in any colon-derived cells in the publicly available ENCODE chromatin immunoprecipitation (ChIP)-seq tracks, making these unlikely to be the functional SNPs in this region (Supplementary Fig. 5). However, there are 26 other SNPs in linkage disequilibrium (LD) with rs35364139 ($r^2>0.5$, CEU population), which may disrupt biofeatures or regulatory elements resulting in the observed CRC risk. Together, the physical proximity of this newly identified susceptibility locus, relevant functional biology and adjacent regulatory marks suggest that *CTNNB1* is an intriguing candidate target gene of a putative enhancer.

The second locus on chromosome 3 is located at 3p14.1 (rs812481) and is intronic of *LRIG1*, a gene encoding a transmembrane protein that interacts with epidermal growth factor receptor-family tyrosine kinase family members[27–29]. LRIG1 has recently been described as a marker of quiescent colon crypt stem cells activated to proliferate following injury[30]. No histone marks are found in the vicinity of rs812481, making it unlikely to be the functional SNP. Notably, rs3856595 ($P=2.4E-07$), in LD ($r^2>0.5$, CEU population) with rs812481, is located in a *LRIG1* intronic active enhancer peak (H3K27ac4) in sigmoid colon epithelium. A second SNP in LD with rs812481 is rs231276 ($P=2.0E-06$), which resides in an

Table 1 | Newly identified genetic susceptibility loci for colorectal cancer.

SNP Chr:Pos (b37)	Alleles Eff/Alt	Effect allele frequency*		Discovery		Asian 1		Asian 2		Combined		
		European	Asian	OR (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value	I^2
rs35360328:40924962	A/T	0.16	0.09	1.14 (1.09-1.19)	2.4E−08	†	†	1.19 (1.01-1.41)	0.04	1.14 (1.09-1.19)	3.1E−09	0
rs8124813:66442435	G/C	0.58	0.79	1.09 (1.05-1.12)	2.5E−08	1.03 (0.94-1.14)	0.47	1.05 (0.97-1.15)	0.21	1.09 (1.05-1.11)	2.0E−08	0
rs1119016410:101351704	G/A	0.29	0.21	1.10 (1.05-1.14)	8.4E−07	1.04 (0.94-1.14)	0.46	1.14 (1.03-1.25)	7.8E−03	1.09 (1.06-1.12)	4.0E−08	0
rs318450412:111884608	C/T	0.53	0.995	1.09 (1.06-1.12)	1.7E−08	†	†	†	†	1.09 (1.06-1.12)	1.7E−08	0
rs7320812012:117747590	G/T	0.11	<0.001	1.16 (1.11-1.23)	2.8E−08	†	†	†	†	1.16 (1.11-1.23)	2.8E−08	0
rs606682520:47340117	A/G	0.64	0.70	1.07 (1.03-1.10)	8.7E−05	1.12 (1.03-1.21)	5.6E−03	1.18 (1.09-1.27)	2.2E−05	1.09 (1.06-1.12)	4.4E−09	0.69‡
rs49462606:117822993	T/C	0.53	0.44	1.08 (1.05-1.11)	4.5E−07	1.06 (0.98-1.14)	0.14	1.05 (0.98-1.13)	0.15	1.07 (1.05-1.10)	6.3E−08	0

Chr, chromosome; CI, confidence interval; OR, odds ratio; Pos, position; SNP, single-nucleotide polymorphism.
*Effect allele frequency (MAF) from 1000 Genomes Project June 2011 release. European frequency is based on CEU + FIN + GBR + IBS + TSI. Asian frequency is based on CHB + CHS + JPT.
†SNP not available.
‡*P* value for Cochran's *Q* test of heterogeneity = 0.038.

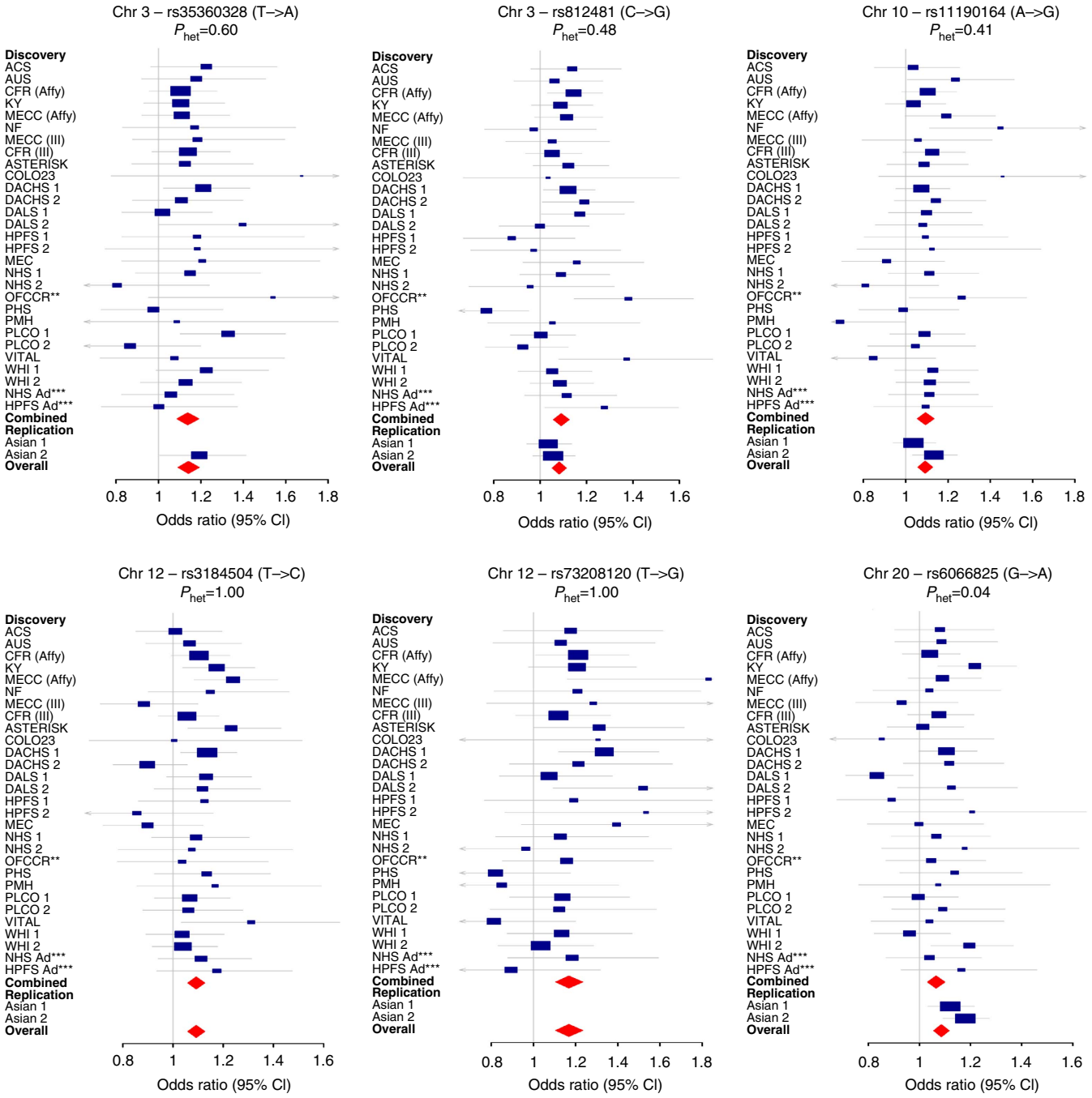

Figure 1 | Forest plots summarizing ORs from studies contributing to colorectal cancer meta-analysis identifying six loci reaching genome-wide significance. The *P* value from the Cochran's *Q* test for heterogeneity (P_{het}) is presented by SNP. ** indicates a subset of the study ARCTIC; *** indicates colorectal polyps. The study specific ORs (blue rectangles) and 95% confidence intervals (CI; horizontal bars) are plotted for each SNP. The red diamonds represent the summary OR and 95% CI for the 'Discovery' series ($N = 18,299$ cases/19,656 controls) and 'Overall'. The 'Replication' series included the Asian 1 ($N = 2,098$ cases/6,172 controls) and Asian 2 ($N = 2,627$ cases/3,797 controls) consortia.

H3K4me1 enhancer peak in a CRC cell (HCT-116). This peak is intronic of *SLC25A26*, a mitochondrial transport protein.

The SNP at 10q24.2 (rs11190164) lies in a genomic region containing multiple genes including *SLC25A28*, *ENTPD7*, *COX15*, *CUTC* and *ABCC2*. Several SNPs in high LD with rs11190164 map to putative enhancers, promoters or 3′ UTRs of genes within the region. A recent study identified rs1035209, 6.3 kb upstream from rs11190164 (CEU $r^2 = 0.4$), to be significantly associated with CRC risk[17]. In addition, rs3740078 (distance to rs11190614 = 93,887 bp, $r^2 = 0.71$, CEU population; $P = 3.2E-05$) causes a synonymous change in the coding sequence of *ENTPD7*. While *ENTPD7* has been linked to intestinal epithelial inflammation in mice and is expressed in normal colonic epithelium[31], a role in CRC has not been previously reported.

Rs3184504 at 12q24.12 implicates *SH2B3* as a putative target gene for CRC susceptibility. SH2B3 is an adaptor protein involved in cytokine signalling and functions as a classic tumour suppressor gene in B-precursor acute lymphoblastic leukaemia that increases STAT3 phosphorylation[32]. Less is known about its

signalling roles in the colon, but rs3184504 is a missense variant (Trp262Arg) that is a known risk allele for coeliac disease and other immune-related disorders[33] and is a well-established risk factor for type 1 diabetes[34] and hypertension[35]. Several other SNPs in LD with rs3184504 also map to putative regulatory regions, but further work is needed to functionally characterize this missense variant or these other SNPs. Other genes within this region, including *CUX2*, *BRAP* and *ACAD10* are also potential candidate genes.

The SNP at 12q24.22 (rs73208120) is independent of rs3184504 at 12q24.12 ($r^2 = 0.002$, CEU population) and lies intronic of *NOS1*. *NOS1* encodes neuronal nitric oxide synthase 1 that generates nitric oxide a reactive free radical involved in several biologic processes, including inflammation, infection and antimicrobial and antitumoral activities[36]. There are several SNPs in LD with rs73208120, but none map to the candidate enhancer regions.

The SNP at 20q13.13 (rs6066825) lies within an intron of the *PREX1* gene that encodes the Rac-guanine nucleotide exchange factor P-Rex1, a signalling protein involved in cell migration and invasion in some cell types[37]. There are 35 SNPs in LD with rs6066825 ($r^2 > 0.5$, CEU population), all intronic or immediately downstream of *PREX1*. The most promising functional candidates are three SNPs, rs2092492 ($r^2 = 0.62$, CEU population), rs6066823 ($r^2 = 0.62$, CEU population) and rs6066825 itself that lie within a putative active enhancer marked by an H3K27ac ChIP-seq peak in sigmoid colon tissue.

Discussion

In conclusion, the combined meta-analysis of 52,649 individuals facilitated the discovery of six new susceptibility loci for CRC. Additional CRC loci remain to be discovered despite the large sample sizes included in our discovery meta-analysis. Although replication of suggestive loci from the discovery phase in similar ancestral populations would be more powerful due to LD and effect allele frequency differences, this study identified six novel CRC risk loci. This study identified opportunities to explore new biologic mechanisms for predisposition to CRC and the potential for translation into improved risk prediction for populations of diverse ancestral heritage.

Methods

Our initial GWAS combined data from three large CRC consortia, the CORECT Study, the CFR, the MECC and the GECCO to elucidate previously undiscovered susceptibility loci for CRC. Data for this discovery analysis focused on individuals of European ancestral heritage from North America, Australia and Europe. Detailed methods are described in the Supplementary Methods. In brief, samples from 19 observational studies genotyped with high-density SNP arrays and imputed to the 1,000 Genomes Project March 2012 reference panel[24] contributed to the discovery meta-analysis. Replication of the top 200 independent SNPs was performed in two additional consortium studies from Asian populations. The studies included in the discovery and replication phases are listed in Supplementary Table 1.

Discovery phase genotyping and QC. The details on study design and characteristics for each study and substudy in the discovery phase are provided in the Supplementary Methods. In brief, the discovery phase consisted of four CRC consortia. The CORECT consortium coordinated genotyping and analysis of six observational studies of CRC for the present analysis: (1) MECC2, (2) CFR2, (3) Kentucky case–control study, (4) American Cancer Society CPS II nested case–control study, (5) Melbourne nested case–control study and (6) Newfoundland case–control study. Genotyping as part of CORECT was conducted using a custom Affymetrix genome-wide platform (the Axiom CORECT Set) with ~1.3 million SNPs and insertions and deletions (indels) on two physical genotyping chips (pegs). In the MECC1 study, germline DNA was extracted from peripheral blood samples and genotyped in two batches using the Illumina HumanOmni 2.5–8 BeadChip, which measures nearly 2.4 million SNPs and indels. Batch 1 (414 cases and 155 controls) was run at the Case Western Reserve University and batch 2 (104 cases and 376 controls) was run at the University of Michigan. Germline DNA for the CFR1 study was extracted from peripheral blood samples and genotyped in two batches using three different platforms—the Illumina Human1M or Human1M-Duo (CFR1-Set1) and the Illumina HumanOmni1-Quad (CFR1-Set 2)—each containing ~1.2 million SNPs and indels. Genotype data were cleaned based on QC metrics at the individual subject and SNP levels. Samples with <95% call rate, sex mismatches (between self-reported and genotypic predicted sex), low concordance with previous genotype data, duplicate samples, unanticipated genotype concordance, identity-by-descent with another sample or ethnic outliers as identified by visual inspection of PCA cluster plots were removed. Before imputation, SNPs with <95% call rate, concordance <95% with 1000 Genomes in samples genotyped for QC, or Hardy–Weinberg equilibrium $P < 10^{-4}$ in controls were excluded. All SNPs overlapping 1000 Genomes were matched to the forward strand.

The GECCO consortium consists of 13 studies. Details are provided in Supplementary Methods and Supplementary Table 1. In brief, DNA was extracted from blood samples or from buccal cells, using conventional methods. Phase one genotyping was done using either Illumina HumanHap 550K, 610K or combined Illumina 300 and 240K, Affymetrix platforms[18], Illumina HumanCytoSNP or Illumina HumanOmniExpress. All studies included 1 to 6% blinded duplicates to monitor quality of the genotyping. All individual-level genotype data were managed and underwent QA/QC at the Ontario Institute for Cancer Research, the University of Washington or at the Fred Hutchinson Cancer Research Center. Details on the QA/QC have previously been described[12]. In brief, samples were excluded based on call rate, heterozygosity, unexpected duplicates, gender discrepancy and unexpectedly high identity-by-descent or unexpected genotype concordance (>65%) with another individual. All analyses were restricted to samples clustering with the Utah residents with Northern and Western European ancestry from the CEPH collection (CEU) population in PC analysis, including the HapMap II populations as reference. SNPs were excluded if they were triallelic, not assigned an rs number, or were reported or observed as not performing consistently across platforms. In addition, genotyped SNPs were excluded based on call rate (<98%), lack of Hardy–Weinberg equilibrium in controls ($P < 1 \times 10^{-4}$) and minor allele frequency (MAF) (<5% in Set 1 for PLCO, WHI, DALS and OFCCR; minor allele count <10 for remaining studies).

Imputation. To meta-analyse genotype data generated from multiple platforms and to increase the coverage of variation that is measurable across the genome, imputation of genotypes was performed for both autosomal (all consortia) and X chromosome (excluding GECCO consortium) markers. Imputing missing genotypes for study samples based on the cosmopolitan panel of reference haplotypes from Phase I of the 1,000 Genomes Project (March 2012 release; $n = 1{,}092$; (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20110521))[23,24] helps improve imputation accuracy of low-frequency variants[38]. The target panel was phased using Beagle[39] (GECCO) or SHAPE-IT[40] (CORECT, MECC1 and CFR1) and the phased target panel was imputed to the 1000 Genomes reference panel using either Minimac[41] (GECCO) or IMPUTE2 (ref. 42) (CORECT, MECC1 and CFR1). Genetic markers retained following imputation had to pass stringent imputation quality and accuracy filters before entering the analysis phase. For GECCO, Rsq was used as the imputation quality measure for imputed SNPs[43], and SNPs were excluded at different Rsq thresholds based on their MAF: for SNPs with MAF > 0.01, we excluded those with Rsq ≤ 0.3; for MAFs of 0.005–0.01, we excluded Rsq < 0.5; and for MAF < 0.005, we excluded Rsq < 0.99. In the remaining studies (CORECT, MECC1 and CFR1) stringent imputation quality and accuracy filters (info ≥0.7, certainty ≥0.9, concordance ≥0.9) were applied between directly measured and imputed genotypes after masking input genotypes (for genotyped markers only) to enter the analysis phase. Further, we restricted the SNP list to those with study-specific MAF ≥ 1%.

Statistical analysis. We utilized PC analysis to assess correspondence between self-reported and genotypic classification of ancestry including unrelated HapMap CEU, YRI and ASN samples as population controls. Ancestral outliers were identified by visual inspection of PC plots for each study and removed. PCs were computed and used for ancestry adjustment. Study-specific association estimates (OR and 95% CI) were obtained employing logistic regression of CRC on allelic dosage adjusting for ancestry and potential confounding variables (for example, age, sex and study site) as defined by the individual studies (Supplementary Methods). The genomic control factor (λ) was estimated by dividing the median χ^2-statistic by 0.456. A sample size-corrected marginal λ, equivalent to studying 1,000 cases and 1,000 controls, was also calculated. Heterogeneity of genetic effects by study was assessed using Cochran's Q test for heterogeneity (P_{het}).

Replication phase. The replication phase was conducted in two Asian consortia (Asian 1 and Asian 2). The Asian Colorectal Cancer Consortium (ACC), Asian 1, consisted of five studies with genome-wide scan data: Shanghai CRC study 1 (Shanghai-1); Shanghai CRC Study 2 (Shanghai-2); Guangzhou CRC Study (Guangzhou); Aichi CRC Study 1 (Aichi-1), and the Korean Cancer Prevention Study-II CRC (KCPS-II). Samples in these studies were genotyped using Affymetrix and Illumina SNP arrays for GWAS (Supplementary Methods)[10,44–48]. A uniform QC protocol (call rates, concordance rates, cryptic relatedness, sex misidentification and ancestry) to filter samples and SNPs was applied[10].

Imputation was performed with the GIANT ALL data panel from the 1,000 Genomes Project phase 1 release v3 as the reference using program MACH v1.0 (ref. 43) and minimac[41]. SNPs with imputation $R^2 > 0.7$ in each of the five studies were included in the final analysis. Associations between SNPs and CRC risk were evaluated based on the log-additive model using mach2dat[43]. Per-allele ORs and 95% confidence intervals (CIs) were derived from logistic regression models, adjusting for age, sex and the first ten PCs when appropriate. Association analysis was conducted for each participating study separately and a fixed-effects meta-analysis was conducted to obtain summary results with the inverse-variance method using program METAL[49].

The Asian 2 consortium was genotyped using the Illumina 1M-duo Array and consisted of studies from the Multiethnic Cohort (MEC; $N = 3,094$), CFR ($N = 285$), Colorectal cancer study on Oahu, Hawaii (CR2 & 3; $N = 134$), Fukuoka, Japan ($N = 1,411$), Nagano, Japan ($N = 207$) and the Japan Public Health Center-based prospective study (JPHC; $N = 1,293$) after QC filtering[50–54]. In general, all genotyped samples were examined and excluded according to the following: (1) call rates < 90, 95 or 97% depending on the batches, (2) missing on basic covariates (age, sex or disease status), (3) gender mismatch, (4) ethnicity outliers and (5) relatedness ($\geq$2nd degree). Prediction of untyped or partly genotyped SNPs was performed with BEAGLE 3.3 (ref. 39) using the 1,000 Genomes Project (phase 1, release 3) East Asians as reference panels. Imputation was performed with all cases and controls combined. Markers with MAF < 0.005 in reference panels were excluded from imputation. Study-specific association statistics were obtained using logistic regression models adjusted for ancestry and potential confounding variables (Supplementary Methods). A fixed-effects meta-analysis was conducted to obtain summary results with the inverse-variance method using programme METAL[49].

All study samples were collected with written informed consent, and procedures were approved by the Human Research institutional review boards (IRBs) of the respective institutions. Specifically, the University of Southern California Health Sciences IRB approved all elements of the CORECT, CFR and MECC studies. The MECC study protocol was also approved by the IRBs at the University of Southern California, University of Michigan, and Carmel Medical Center (Haifa). The Fred Hutchinson Cancer Research Center IRB approved the GECCO contribution. The Asian 1 consortia study protocols were approved by the review board of the Vanderbilt University Medical Center and informed consent was obtained from all study participants. Study protocols of the Asian 2 consortia were approved by the University of Hawaii Human Studies Program and University of Southern California IRB, the IRB in the National Cancer Center, Japan and the Ethics Committee of Kyushu University Faculty of Medical Sciences.

Meta-analysis. A consortia-wide meta-analysis for the discovery and replication phases using fixed-effect models with inverse variance weighting was implemented in METAL. Heterogeneity was evaluated using Cochran's Q test for heterogeneity and the measure I^2. Graphical representation of effect estimates and CIs by study and consortia are presented using forest plots.

References

1. Siegel, R., Ma, J., Zou, Z. & Jemal, A. Cancer statistics, 2014. *CA Cancer J. Clin.* **64,** 9–29 (2014).
2. Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int. J. Cancer.* **99,** 260–266 (2002).
3. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343,** 78–85 (2000).
4. Broderick, P. *et al.* A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat. Genet.* **39,** 1315–1317 (2007).
5. Cui, R. *et al.* Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. *Gut* **60,** 799–805 (2011).
6. Dunlop, M. G. *et al.* Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat. Genet.* **44,** 770–776 (2012).
7. Houlston, R. S. *et al.* Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat. Genet.* **42,** 973–977 (2010).
8. Houlston, R. S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40,** 1426–1435 (2008).
9. Jaeger, E. *et al.* Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat. Genet.* **40,** 26–28 (2008).
10. Jia, W. H. *et al.* Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nat. Genet.* **45,** 191–196 (2013).
11. Peters, U. *et al.* Meta-analysis of new genome-wide association studies of colorectal cancer risk. *Hum. Genet.* **131,** 217–234 (2012).
12. Peters, U. *et al.* Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. *Gastroenterology* **144,** 799–807 e724 (2013).
13. Tenesa, A. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat. Genet.* **40,** 631–637 (2008).
14. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.* **39,** 984–988 (2007).
15. Tomlinson, I. P. *et al.* Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet.* **7,** e1002105 (2011).
16. Tomlinson, I. P. *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.* **40,** 623–630 (2008).
17. Whiffin, N. *et al.* Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Hum. Mol. Genet.* **23,** 4729–4737 (2014).
18. Zanke, B. W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39,** 989–994 (2007).
19. Zhang, B. *et al.* Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat. Genet.* **46,** 533–542 (2014).
20. Schmit, S. L. *et al.* A novel colorectal cancer risk locus at 4q32.2 identified from an international genome-wide association study. *Carcinogenesis* **35,** 2512–2519 (2014).
21. Wang, H. *et al.* Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. *Nat. Commun.* **5,** 4613 (2014).
22. Jiao, S. *et al.* Estimating the heritability of colorectal cancer. *Hum. Mol. Genet.* **23,** 3898–3905 (2014).
23. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
24. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
25. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
26. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499,** 214–218 (2013).
27. Laederich, M. B. *et al.* The leucine-rich repeat protein LRIG1 is a negative regulator of ErbB family receptor tyrosine kinases. *J. Biol. Chem.* **279,** 47050–47056 (2004).
28. Miller, J. K. *et al.* Suppression of the negative regulator LRIG1 contributes to ErbB2 overexpression in breast cancer. *Cancer Res.* **68,** 8286–8294 (2008).
29. Shattuck, D. L. *et al.* LRIG1 is a novel negative regulator of the Met receptor and opposes Met and Her2 synergy. *Mol. Cell. Biol.* **27,** 1934–1946 (2007).
30. Powell, A. E. *et al.* The pan-ErbB negative regulator Lrig1 is an intestinal stem cell marker that functions as a tumor suppressor. *Cell* **149,** 146–158 (2012).
31. Kusu, T. *et al.* Ecto-nucleoside triphosphate diphosphohydrolase 7 controls Th17 cell responses through regulation of luminal ATP in the small intestine. *J. Immunol.* **190,** 774–783 (2013).
32. Perez-Garcia, A. *et al.* Genetic loss of SH2B3 in acute lymphoblastic leukemia. *Blood* **122,** 2425–2432 (2013).
33. Zhernakova, A. *et al.* Evolutionary and functional analysis of celiac risk loci reveals SH2B3 as a protective factor against bacterial infection. *Am. J. Hum. Genet.* **86,** 970–977 (2010).
34. Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41,** 703–707 (2009).
35. Levy, D. *et al.* Genome-wide association study of blood pressure and hypertension. *Nat. Genet.* **41,** 677–687 (2009).
36. Lirk, P., Hoffmann, G. & Rieder, J. Inducible nitric oxide synthase—time for reappraisal. *Curr. Drug Targets Inflamm. Allergy* **1,** 89–108 (2002).
37. Campbell, A. D. *et al.* P-Rex1 cooperates with PDGFRbeta to drive cellular migration in 3D microenvironments. *PLoS ONE* **8,** e53982 (2013).
38. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3* **1,** 457–470 (2011).
39. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81,** 1084–1097 (2007).
40. Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F. & Marchini, J. Haplotype estimation using sequencing reads. *Am. J. Hum. Genet.* **93,** 687–696 (2013).
41. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44,** 955–959 (2012).
42. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
43. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34,** 816–834 (2010).
44. Abnet, C. C. *et al.* A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nat. Genet.* **42,** 764–767 (2010).

45. Amundadottir, L. *et al.* Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat. Genet.* **41,** 986–990 (2009).
46. Bei, J. X. *et al.* A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. *Nat. Genet.* **42,** 599–603 (2010).
47. Nakata, I. *et al.* Association between the SERPING1 gene and age-related macular degeneration and polypoidal choroidal vasculopathy in Japanese. *PLoS ONE* **6,** e19108 (2011).
48. Jee, S. H. *et al.* Adiponectin concentrations: a genome-wide association study. *Am. J. Hum. Genet.* **87,** 545–552 (2010).
49. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).
50. National Cancer Institute, Division of Cancer Epidemiology and Genetics. Cancer Genetic Markers of Susceptibility (CGEMS) Project: Executive Summary. Available at <http://dceg.cancer.gov/research/how-we-study/genomic-studies/cgems-summary> (2009).
51. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39,** 645–649 (2007).
52. Petersen, G. M. *et al.* A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat. Genet.* **42,** 224–228 (2010).
53. Landi, M. T. *et al.* A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am. J. Hum. Genet.* **85,** 679–691 (2009).
54. Lan, Q. *et al.* Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat. Genet.* **44,** 1330–1335 (2012).

Acknowledgements

CORECT: this work was supported by the National Cancer Institute, National Institutes of Health under RFA # CA-09-002, NIH/NCI U19 CA148107. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centres in the CORECT consortium, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government or the CORECT Consortium. ASTERISK: we are very grateful to Dr Bruno Buecher without whom this project would not have existed. We also thank all those who agreed to participate in this study, including the patients and the healthy control persons, as well as all the physicians, technicians and students. DACHS: we thank all participants and cooperating clinicians, and Ute Handte-Daub, Renate Hettler-Jensen, Utz Benscheid, Muhabbet Celik and Ursula Eilber for excellent technical assistance. GECCO: we thank all those at the GECCO Coordinating Center for helping bring together the data and people that made this project possible. HPFS, NHS and PHS: we acknowledge Patrice Soule and Hardeep Ranu of the Dana-Farber Harvard Cancer Center High-Throughput Polymorphism Core who assisted in the genotyping for NHS, HPFS and PHS under the supervision of Dr Immaculata Devivo and Dr David Hunter, Qin (Carolyn) Guo and Lixue Zhu who assisted in programming for NHS and HPFS and Haiyan Zhang who assisted in programming for the PHS. We thank the participants and staff of the Nurses' Health Study and the Health Professionals Follow-Up Study, for their valuable contributions as well as the following state cancer registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. In addition, this study was approved by the Connecticut Department of Public Health (DPH) Human Investigations Committee. Certain data used in this publication were obtained from the DPH. We assume full responsibility for analyses and interpretation of these data. PLCO: we thank Drs Christine Berg and Philip Prorok, Division of Cancer Prevention, National Cancer Institute, the Screening Center investigators and staff or the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial, Mr. Tom Riley and staff, Information Management Services Inc., Ms Barbara O'Brien and staff, Westat Inc. and Drs Bill Kopp, Wen Shao and staff, SAIC-Frederick. Most importantly, we acknowledge the study participants for their contributions for making this study possible. The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI. PMH: we thank the study participants and staff of the Hormones and Colon Cancer study. WHI: we thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at https://cleo.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Short%20List.pdf. ACC: we thank all study participants and research staff of all studies for their contributions and commitment to this project, Regina Courtney for DNA preparation and Jing He for data processing.

Author contributions

S.B.G., G.C. and U.P. contributed to the study concept and design. L.L.M. and W.Z. organized the Asian 1 and Asian 2 consortia. D.J.V.D.B. supervised the genotyping of samples at USC and C.K.E. led the quality control of the CORECT, MECC and CFR GWAS data. F.R.S., S.L.S., S.J., H.W., B.Z. and D.V.C. contributed to the statistical analysis. F.R.S., S.L.S., G.C., U.P. and S.B.G. drafted the manuscript. E.L.G, B.H., R.B.H., L.N.K., R.G., R.H., S.Kury, M.I., P.A.N., D.W.W., S.I.B., B.W.Z., N.M.L., M.J., S.J.G., S.T., W.H.-J., K.M., X.O.S., Y.B.X., S.H.J., G.G., J.L.H., E.J., J.D.P., G.S., W.Z., L.L.M., S.B.G., G.R. and U.P. conducted the epidemiological studies for sample collection. All authors contributed to the writing of the manuscript, interpretation, and discussion of the findings. All authors approved the manuscript.

Additional information

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Schumacher, F. R. *et al.* Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nat. Commun.* 6:7138 doi: 10.1038/ncomms8138 (2015).