Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer

Wei-Hua Jia[1,16], **Ben Zhang**[2,16], **Keitaro Matsuo**[3], **Aesun Shin**[4], **Yong-Bing Xiang**[5], **Sun Ha Jee**[6], **Dong-Hyun Kim**[7], **Zefang Ren**[1], **Qiuyin Cai**[2], **Jirong Long**[2], **Jiajun Shi**[2], **Wanqing Wen**[2], **Gong Yang**[2], **Ryan J. Delahanty**[2], **GECCO and CCFR**[8], **Bu-Tian Ji**[9], **Zhi-Zhong Pan**[1], **Fumihiko Matsuda**[10], **Yu-Tang Gao**[5], **Jae Hwan Oh**[11], **Yoon-Ok Ahn**[12], **Eun Jung Park**[6], **Hong-Lan Li**[5], **Ji Won Park**[11], **Jaeseong Jo**[6], **Jin-Young Jeong**[7], **Satoyo Hosono**[3], **Graham Casey**[13], **Ulrike Peters**[14,15], **Xiao-Ou Shu**[2], **Yi-Xin Zeng**[1,17], and **Wei Zheng**[2,17]

[1]State Key Laboratory of Oncology in South China, Cancer Center, Sun Yat-sen University, Guangzhou, China

[2]Division of Epidemiology, Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA

[3]Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Japan

[4]Molecular Epidemiology Branch, National Cancer Center, Goyang-si, Korea

[5]Department of Epidemiology, Shanghai Cancer Institute, Shanghai, P.R. China

[6]Institute for Health Promotion, Department of Epidemiology and Health Promotion, Graduate School of Public Health, Yonsei University, Seoul, Korea

[7]Department of Social and Preventive Medicine, Hallym University College of Medicine, Okcheon-dong, Republic of Korea

Corresponding author contact information: Wei Zheng, M.D., Ph.D., Vanderbilt Epidemiology Center and Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, 2525 West End Avenue, 8th Floor, Nashville, TN 37203-1738, USA, Phone: (615) 936-0682 Fax: (615) 936-8241, wei.zheng@vanderbilt.edu.

[8]A complete list of investigators in GECCO (Genetics and Epidemiology of Colorectal Cancer Consortium) and CCFR (Colon Cancer Family Registry) is provided in the Acknowledgement section.

[16]These authors contributed equally to this work.

[17]These authors jointly directed this work.

AUTHOR CONTRIBUTIONS

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

[9]Division of Cancer Epidemiology & Genetics, National Cancer Institute, Bethesda, MD, USA

[10]Center for Genomic Medicine, Graduate School of Medicine, Kyoto University, Kyoto, Japan

[11]Center for Colorectal Cancer, National Cancer Center, Goyang-si, Korea

[12]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Korea

[13]Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA

[14]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[15]Department of Epidemiology, University of Washington School of Public Health, Seattle, WA, USA

Abstract

To identify novel genetic factors for colorectal cancer (CRC), we conducted a genome-wide association study in East Asians. By analyzing genome-wide data in 2,098 cases and 5,749 controls, we selected 64 promising SNPs for replication in an independent set of samples including up to 5,358 cases and 5,922 controls. We identified four SNPs with a P-value of 8.58×10^{-7} to 3.77×10^{-10} in the combined analysis of all East Asian samples. Three of the four SNPs were replicated in a study conducted among 26,060 European descendants with a combined P-value of 1.22×10^{-10} for rs647161 (5q31.1), 6.64×10^{-9} for rs2423279 (20p12.3), and 3.06×10^{-8} for rs10774214 (12p13.32 near the *CCND2* gene), respectively, derived from the meta-analysis of data from both East Asian and European populations. This study identified three new CRC susceptibility loci and provides additional insight into the genetics and biology of CRC.

Colorectal cancer (CRC) is one of the most commonly diagnosed malignancies in East Asia and many other parts of the world [1]. Genetic factors play an important role in the etiology of both sporadic and familial CRC [2]. However, less than 6% of CRC cases can be explained by rare, high-penetrance variants in the CRC susceptibility genes identified to date, such as the *APC*, *SMAD4*, *AXIN2*, *BMPR1A*, *POLD1*, *STK11*, *MUTYH*, and DNA mismatch repair genes [2]. Over the past two decades, many candidate gene studies have evaluated common genetic risk factors for CRC; only a few of them have been replicated in subsequent studies [3]. Recent genome-wide association studies (GWAS) have identified approximately 15 common genetic susceptibility loci for CRC [4–12]. However, these newly identified genetic factors, along with known high-penetrance CRC susceptibility genes, explain less than 15% of the heritability for this common malignancy [10, 11]. Furthermore, with the exception of a small study conducted in Japan [12], all other GWAS were conducted among European-ancestry populations which differ from other ethnic groups in certain genetic architecture. Many of the variants discovered in European-ancestry populations show only a weak or no association with CRC in other ethnic groups [13]. Therefore, additional GWAS are needed, particularly in non-European-ancestry populations, to fully uncover the genetic basis for CRC susceptibility.

In 2009, we initiated the Asia Colorectal Cancer Consortium (ACCC), a GWAS in East Asians, to search for novel genetic risk factors for CRC. The discovery stage (Stage 1)

consisted of five GWAS conducted in China, Korea, and Japan, including 2,293 CRC patients and 5,780 controls (Supplementary Table 1). Cases and controls were genotyped using several SNP arrays, including Affymetrix Genome-Wide Human SNP Array 6.0 (906,602 SNPs), Affymetrix Genome-Wide Human SNP Array 5.0 (443,104 SNPs), Illumina Infinium HumanHap610 BeadChip (592,044 SNPs), Illumina Human610-Quad BeadChip (620,901 SNPs), and Illumina HumanOmniExpress BeadChip (729,462 SNPs) (Supplementary Table 1). After quality control (QC) exclusions as described previously [14–17], 2,098 cases and 5,749 controls remained for this study (Supplementary Tables 1 and 2). Also excluded from the analyses were SNPs with a call rate < 95%, genotype concordance rate < 95% among positive QC samples, minor allele frequency (MAF) < 5%, or P-value for Hardy-Weinberg equilibrium $< 1.0 \times 10^{-5}$ in controls for each study. Imputation was conducted for each study following the MACH algorithm [18] using phased HapMap 2 CHB and JPT samples as the reference. No apparent genetic admixture was identified except for one sample from KCPS-II (Supplementary Fig. 1). Associations between CRC risk and each of the genotyped and imputed SNPs were evaluated using logistic regression within each study after adjusting for age, sex, and the first ten principal components using mach2dat [18]. Meta-analyses were conducted under a fixed-effects model using the METAL program [19]. There was little evidence for inflation in the association test statistics for any of the five studies (genomic inflation factor (λ) range: 1.02 to 1.04) or for all studies combined (λ= 1.01) (Supplementary Table 1 and Supplementary Fig. 2). The observed number of SNPs with a small P-value was slightly larger than that expected by chance (Supplementary Fig. 2).

Multiple genomic locations were revealed as potentially related to CRC risk (Supplementary Fig. 3). Nine SNPs identified from published GWAS conducted in European-ancestry populations showed an association with CRC risk at P< 0.05 in Stage 1 (data not shown). To improve the statistical power for evaluating these SNPs, we genotyped 6,476 additional samples to bring the total sample size to 5,252 cases and 9,071 controls. Except for the two SNPs (rs6691170 and rs16892766) that are monomorphic in East Asians, all 16 of the other SNPs identified from published GWAS conducted in European-ancestry populations showed an association with CRC risk in the same direction as reported previously (Supplementary Table 3). A significant association with CRC risk at $P < 0.05$ was found for 13 SNPs, including rs6687758, rs10936599, rs10505477, rs6983267, rs7014346, rs10795668, rs3802842, rs4444235, rs4779584, rs9929218, rs4939827, rs10411210, and rs961523. Except for two SNPs (rs6983267 and rs4779584), no statistically significant heterogeneity at P< 0.05 was observed between East-Asian- and European-ancestry populations (Supplementary Table 3).

To identify novel genetic factors for CRC, we selected 64 SNPs for replication in an independent set of 5,358 cases and 5,922 controls recruited in five studies conducted in China, Korea, and Japan (Supplementary Table 2). SNPs were selected from among those with 1) MAF > 5%; 2) no heterogeneity across studies ($P_{\text{heterogeneity}}$> 0.05 and $I^2 < 25\%$); 3) not in linkage disequilibrium (LD) (r^2< 0.2) with any known CRC risk variants reported from previous GWAS; 4) high imputation quality in each of the five studies (RSQ > 0.5); 5)

$P < 0.01$ in the combined analysis of all five studies included in Stage 1. These criteria were used to prioritize SNPs for replication for this study.

Of the 64 SNPs evaluated in Stage 2, seven SNPs showed an association with CRC risk at $P < 0.05$ with a direction of association consistent with that observed in Stage 1 (Table 1 and Supplementary Table 4). In the combined analysis of data from both Stages 1 and 2, P-values for the association with two SNPs (rs647161 at 5q31.1, OR=1.17, $P = 3.77 \times 10^{-10}$ and rs10774214 at 12p13.32, OR=1.17, $P = 5.48 \times 10^{-10}$) were lower than the conventional genome-wide significance level of 5.0×10^{-8}, providing convincing evidence for an association of these SNPs with CRC risk (Table 1). An additional SNP, rs2423279, showed a significant association in Stage 2 after Bonferroni correction (corrected $P < 7.8 \times 10^{-4}$), but did not reach the conventional GWAS significance level for association with CRC risk in the combined analysis of all samples (OR=1.14, $P = 2.29 \times 10^{-7}$). The association between CRC risk and each of these three SNPs was consistent across most studies (Fig. 1). Results for the other four SNPs (rs1665650, rs2850966, rs1580743, and rs4503064) replicated in Stage 2 at $P < 0.05$ are also presented in Supplementary Table 4, including one SNP (rs1665650) with a P-value of 8.58×10^{-7} in the combined analysis of all data from both stages (Table 1).

We next evaluated these top four SNPs shown in Table 1 using data from GWAS in the Genetics and Epidemiology of Colorectal Cancer Consortium and the Colon Cancer Family Registry (GECCO and CCFR), which include 11,870 cases and 14,190 controls of European ancestry [4, 20, 21]. Three of the four SNPs were replicated in the GECCO and CCFR, although the strength of the association was weaker than that found in East Asians (Table 2). These results provide independent support of our findings in the East Asian population. Meta-analyses of data from both East Asian and European populations provided strong evidence for associations of CRC risk with three SNPs with P-values all exceeding the genome-wide significant threshold of 5×10^{-8} (Table 2). The weaker associations observed in European-ancestry populations could be explained in part by differences in LD patterns for these loci for East Asians and Europeans (Supplementary Fig. 4). It is possible that causal variants in these regions are tagged by different SNPs in these two populations or there is allelic heterogeneity, in which different underlying causal variants exist in Asian- and European-ancestry populations. The difference in LD structure between Asian and European descendants and possible allelic heterogeneity in these two populations may explain, in part, why these loci were not discovered in previous studies conducted in European descendants. The fourth SNP, rs1665650 evaluated in in the GECCO and CCFR, however, was not replicated in European-ancestry populations (OR = 0.96, $P = 0.05$).

Stratification analyses showed that the associations of CRC risk with each of these three replicated SNPs were generally consistent in Chinese, Korean, and Japanese ($P_{heterogeneity} > 0.05$), although the association with rs2423279 was not statistically significant in the Japanese, perhaps due to a small sample size (Supplementary Table 5). Associations of these three SNPs with CRC risk were similar for men and women ($P_{heterogeneity} > 0.05$) (Supplementary Table 6).

SNP rs10774214 is located just 15 kb upstream of *CCND2*, the gene encoding cyclin D2 (Figure 2), a member of the D-type cyclin family, which also includes cyclins D1 and D3. These cyclins play a critical role in cell cycle control (from G1 phase to S phase) through activation of cyclin-dependent kinases (CDK), primarily CDK4 and CDK6 [22]. *CCND2* is closely related to *CCND1*, a well-established human oncogene [22, 23]. Although *CCND2* has been less well studied than *CCND1*, several studies including The Cancer Genome Atlas (TCGA) have shown *CCND2* to be overexpressed in a substantial proportion of human colorectal tumors [22–25]. Overexpression of this cyclin may be an independent predictor of survival in CRC patients [24]. Several other genes, including *PARP11*, *FGF23, FGF6, C12orf5,* and *RAD51AP1*, are also in close proximity to the SNP identified in our study, of which both *C12orf5* (also known as *TIGAR,* TP53-induced glycolysis and apoptosis regulator) and *RAD51AP1* were found to be overexpressed in CRC tissue included in TCGA [25]. SNP rs10774214 is in strong LD with several SNPs that are located in potential transcription sites as determined by the TRANSFAC database [26]. Additional research may be warranted regarding possible mechanisms by which this SNP is related to CRC risk.

SNP rs647161 is located on chromosome 5q31.1, where a cluster of SNPs were associated with CRC risk (Figure 2). Of the genes in this region (including *PITX1*, *CATSPER3*, *PCBD2*, *MIR4461*, and *H2AFY)*, *PITX1* is the closest to rs647161 (approximately 129 kb upstream). The *PITX1* (*paired-like homeodomain 1*) gene has been described as a tumor suppressor gene and may be involved in the tumorigenesis of multiple human cancers [27–31], including CRC [27, 32]. *PITX1* has been reported to suppress tumorigenicity by down-regulating the RAS pathway, which is frequently altered in colorectal tumors [27]. Inhibition of *PITX1* induces the RAS pathway and tumorigenicity, and restoring *PITX1* in colon cancer cells inhibits tumorigenicity [27]. It also has been reported that *PITX1* may activate *P53*[33] and regulate telomerase activity [34]. Consistent with the role of a tumor suppressor, this gene has been found to be down-regulated in human cancer tissue samples and cell lines [27–30, 32]. CRC tissue expressing wild-type KRAS showed significantly lower expression levels of *PITX1* than tissue with mutant KRAS [32]. Most recently, low *PITX1* expression was found to be associated with poor survival in CRC patients [35]. In addition, rs6596201 in moderate LD with rs647161 (r^2=0.25), is an eQTL (P=2.42×10^{-28}) for the *PITX1* gene [36]. Several other genes, including *C5orf24*, *H2AFY*, and *NEUROG1*, at this locus were also found to be highly expressed in colorectal tumors included in the TCGA (P<0.001) [25]. Additional studies are warranted to explore any possible role of these genes in the etiology of CRC.

SNP rs2423279 is located on chromosome 20p12.3, close to the *HAO1* and *PLCB1* genes (Figure 2). *HAO1* encodes hydroxyacid oxidase, which has 2-hydroxyacid activity. *PLCB1* encodes phospholipase C beta 1, which plays an important role in the intracellular transduction of many extracellular signals. Overexpression of the *PLCB1* gene has been observed in CRC tissue [25]. Possible mechanisms by which these genes are involved in CRC carcinogenesis are unknown. SNP rs2423279 is 1,408,069 bp downstream of rs961253, a SNP previously identified in a European GWAS to be associated with CRC risk [10]. However, these two SNPs are not correlated in East Asians (r^2=0) or in Europeans (r^2=0). Adjustment for rs961253 did not change the results for rs2423279 (data not shown).

To our knowledge, this is the largest GWAS performed for CRC in East Asians, a population that differs from the European-ancestry population in CRC risk and certain aspects of genetic architecture. Our study, along with data from a large study conducted in a European-ancestry population, provides convincing evidence of association with CRC risk for three novel independent susceptibility loci at 5q31.1, 12p13.32, and 20p12.3. Results from this study provide new insights into the genetics and biology of CRC.

URLs

CGEMS, http://cgems.cancer.gov/; dbGaP, http://www.ncbi.nlm.nih.gov/gap; EIGENSTRAT, genepath.med.harvard.edu/~reich/EIGENSTRAT.htm; eqtl.uchicago.edu, http://eqtl.uchicago.edu/Home.html; GTEx eQTL Browser, http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi; Haploview, http://www.broad.mit.edu/mpg/haploview/; HapMap project, http://hapmap.ncbi.nlm.nih.gov/; IntOGen, http://www.intogen.org/home; LocusZoom, http://csg.sph.umich.edu/locuszoom/; MACH 1.0, http://www.sph.umich.edu/csg/abecasis/MACH/; mach2dat, http://www.sph.umich.edu/csg/abecasis/MACH/; METAL, http://www.sph.umich.edu/csg/abecasis/Metal/; PLINK version 1.07, http://pngu.mgh.harvard.edu/~purcell/plink/; R version 2.13.0, http://www.r-project.org/; SAS version 9.2, http://www.sas.com/; SNAP, http://www.broadinstitute.org/mpg/snap/; TRANSFAC, http://www.gene-regulation.com/pub/databases.html; UCSC Genome Browser, http://genome.ucsc.edu/.

ONLINE METHODS

Study populations

After quality control (QC), 7,456 cases and 11,671 controls from ten studies were included in this consortium (Supplementary Table 2). Detailed descriptions of participating studies and demographic characteristics of study participants are provided in Supplementary Note. Briefly, the consortium included 10,730 Chinese participants, 5,544 Korean participants, and 2,853 Japanese participants. Chinese participants were from five studies: Shanghai Study 1 (Shanghai-1, $n = 3,102$), Shanghai Study 2 (Shanghai-2, $n = 485$), Guangzhou Study 1 (Guangzhou-1, $n = 1,613$), Guangzhou Study 2 (Guangzhou-2, $n = 2,892$), and Guangzhou Study 3 (Guangzhou-3, $n = 2,638$). Korean participants were from three studies: the Korean Cancer Prevention Study-II (KCPS-II, $n = 1,301$), the Seoul Study ($n = 1,522$), and the Korea-National Cancer Center (Korea-NCC) Study ($n = 2,721$). Japanese participants were from two studies: Aichi Study 1 (Aichi-1, $n = 1,346$) and Aichi Study 2 (Aichi-2, $n = 1,507$). We also evaluated associations for the top four SNPs using data from 11,870 CRC cases and 14,190 controls of European ancestry included in the Genetics and Epidemiology of Colorectal Cancer Consortium and the Colon Cancer Family Registry (GECCO and CCFR), which include 14 studies from the USA, Europe, Canada and Australia [4, 20, 21]. Approval was granted from the relevant institutional review boards at all study sites, and all included participants gave informed consent.

Genotyping and QC procedures

For detailed descriptions of genotyping and QC procedures, and design for plates and QC samples, see the Supplementary Note. Briefly, in Stage 1, 481 cases and 2,632 controls from

Shanghai-1 were genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 as described previously [14]. The average concordance percentage of QC samples was 99.7% with a median value of 100% in Shanghai-1 [14, 37, 38]. Stage 1 genotyping for 296 cases and 257 controls in Shanghai-2 was performed using Illumina HumanOmniExpress BeadChips. The same method was used to genotype cases from the Guangzhou-1 (n= 694) and Aichi-1 (n = 497) studies in Stage 1. The positive QC samples in these studies had an average concordance percentage of 99.41% and a median value of 99.97%. Cases and controls in KCPS-II were genotyped using the Affymetrix Genome-Wide Human SNP Array 5.0 [16]. Controls for the Guangzhou-1 and Aichi-1 studies were genotyped previously using the Illumina Human610-Quad [15] and Illumina Infinium HumanHap610 BeadChip [17] platforms, respectively. Details of QC procedures for these samples have been described previously [15–17]. Excluded from the analysis were samples that were genetically identical or duplicated, had a genotype-determined sex inconsistent with self-reported data, had unclear population structure, had close relatives with a PI-HAT estimate greater than 0.25 or had a call rate < 95%. Within each study, SNPs were excluded if: 1) MAF < 5%, 2) call rate < 95%, 3) genotyping concordance percentage < 95% in QC samples, 4) P-value for Hardy-Weinberg equilibrium $< 1.0 \times 10^{-5}$ in controls, or 5) SNPs not in the 22 autosomes. The final numbers of cases, controls, and SNPs remaining for analysis in each participating study are presented in Supplementary Table 1.

Genotyping for Stage 2 was completed using the iPLEX Sequenom MassARRAY platform as described previously [14, 39]. With the exception of some samples from Guangzhou study, which were genotyped at Fudan University (Shanghai, China), all other samples were genotyped at the Vanderbilt Molecular Epidemiology Laboratory. The average concordance percentage of the genotyping data for positive QC samples was > 99% with a median value of 100% for each of the five studies. SNPs were excluded from the analysis if: 1) call rate < 95%, 2) genotyping concordance percentage < 95% in QC samples, 3) unclear genotyping cluster, or 4) P-value for Hardy-Weinberg equilibrium $< 7.8 \times 10^{-4}$. The numbers of SNPs remaining for analysis in each participating study in Stage 2 are presented in the Supplementary Note.

Genotyping for samples included in the GECCO and CCFR GWAS was conducted using Illumina BeadChip arrays, with the exception of the Ontario Familial Colorectal Cancer Registry study, for which Affymetrix arrays were used [4, 20, 21]. Details of the QC procedures for these samples are presented in the Supplementary Note.

SNP selection for replication

SNPs were selected for Stage 2 replication based on the following criteria: 1) data available in each of the five Stage 1 studies; 2) MAF > 5% in each Stage 1 study; 3) no heterogeneity across the five studies included in Stage 1 ($P_{\text{heterogeneity}}$> 0.05 and I^2< 25%); 4) not in LD (r^2< 0.2) with any known risk variants reported from previous GWAS; 5) not in LD (r^2< 0.2) with each other; 6) high imputation quality in each of the five studies (RSQ > 0.5), and 7) P< 0.01 in combined analysis of all Stage 1 studies.

Evaluation of population structure

We evaluated population structure in each of the five participating studies included in Stage 1 by using principal components analysis (PCA). Genotyping data for uncorrelated, genome-wide SNPs were pooled with data from HapMap to generate the first ten principal components using EIGENSTRAT software [40] (see **URLs**). The first two principal components for each sample were plotted using R (see **URLs**). We identified and excluded one participant of KCPS-II who was more than 6 σ away from the means of PC1 and PC2 (Supplementary Fig. 1). The remaining 7,847 samples showed clear East Asian origin, and these samples were included in the final genome-wide association analysis. Cases and controls in each of the five studies were in the same cluster as HapMap Asian samples. The estimated inflation factor λ ranged from 1.02 to 1.04 in these studies after adjusting for age, sex, and the first ten principal components with a λ of 1.01 for combined Stage 1 data (Supplementary Table 1 and Supplementary Fig. 2).

Imputation

We used the program MACH 1.0 [18](see **URLs**)to impute genotypes for autosomal SNPs which were present in HapMap Phase II release 22 separately for each of the five studies included in Stage 1. Genotype data from the 90 Asian subjects from HapMap were used as reference. For Guangzhou-1 and Aichi-1, cases and controls were genotyped using different platforms. To improve imputation quality [41], we identified SNPs shared between cases and controls (250,612 SNPs in Guangzhou-1 and 232,426 SNPs in Aichi-1) and used them to impute genotyping data. A total of 1,636,380 genotyped SNPs or imputed SNPs with high imputation quality (RSQ > 0.50) in all the five studies were tested for association with CRC. To directly evaluate the imputation quality for the top four SNPs identified in our study, we genotyped them in approximately 2,500 samples included in Stage 1. The agreement of genotype calls derived from direct genotyping and imputation was very high, with a mean value of 98.05%, 95.61%, 99.84%, and 97.90% for rs647161, rs10774214, rs2423279, and rs1665650, respectively (Supplementary Table 7).

Statistical analyses

Dosage data for genotyped and imputed SNPs for participants in each Stage 1 study were analyzed using the program mach2dat [18](see **URLs**). We coded 0, 1, or 2 copies of the effect allele as dosage for genotyped SNPs, and for imputed SNPs, we used the expected number of copies of the effect allele as dosage score. This approach has been shown to give unbiased estimates in meta-analyses [42]. Associations between SNPs and CRC risk were assessed using odds ratios (ORs) and 95% confidence intervals (CIs) derived from logistic regression models. ORs were estimated based on the log-additive model and adjusted for age, sex, and the first ten principal components. PLINK version 1.07 (see **URLs**) also was used to analyze genotype data [43] and yielded results virtually identical to those derived from dosage data using mach2dat [18]. Meta-analyses were performed using the inverse-variance method, assuming a fixed-effects model, and calculations were implemented in the METAL package [19] (see **URLs**).

Similar to Stage 1, we used logistic regression models to derive ORs and 95% CIs for the 64 selected SNPs in Stage 2, assuming a log-additive model with adjustment for age and sex.

We performed joint analyses to generate summary results for combined samples from all studies with additional adjustment for study site. We also conducted stratification analysis for the top four SNPs by population ethnicity (Chinese, Korean, and Japanese) and by sex. We used Cochran's Q statistic to test for heterogeneity [44] and I^2 statistic to quantify heterogeneity [45] across studies as described elsewhere in detail [46]. Analyses for Stage 2, as well as combined Stages 1 and 2 data were conducted using SAS, version 9.2(see **URLs**), with the use of two-tailed tests. *P*-value 5×10^{-8} in the combined analysis was considered statistically significant.

We used Haploview version 4.2 [47](see **URLs**)to generate a genome-wide Manhattan plot for results from the Stage 1 meta-analysis. Forest plots and quantile-quantile (Q-Q) plots were drawn using R. We drew regional association plots using the website-based tool LocusZoom, version 1.1 [48] (see **URLs**). LD plots were generated using Haploview [47] and UCSC Genome Browser (see **URLs**).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Jemal A, et al. Global cancer statistics. CA Cancer J Clin. 2011; 61:69–90. [PubMed: 21296855]

2. de la Chapelle A. Genetic predisposition to colorectal cancer. Nat Rev Cancer. 2004; 4:769–780. [PubMed: 15510158]

3. Dong LM, et al. Genetic susceptibility to cancer: the role of polymorphisms in candidate genes. JAMA. 2008; 299:2423–2436. [PubMed: 18505952]

4. Zanke BW, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet. 2007; 39:989–994. [PubMed: 17618283]

5. Tomlinson I, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet. 2007; 39:984–988. [PubMed: 17618284]

6. Broderick P, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet. 2007; 39:1315–1317. [PubMed: 17934461]

7. Jaeger E, et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nat Genet. 2008; 40:26–28. [PubMed: 18084292]

8. 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. 2008; 40:631–637. [PubMed: 18372901]

9. Tomlinson IP, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet. 2008; 40:623–630. [PubMed: 18372905]

10. Houlston RS, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet. 2008; 40:1426–1435. [PubMed: 19011631]

11. Houlston RS, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. Nat Genet. 2010; 42:973–977. [PubMed: 20972440]

12. Cui R, et al. Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. Gut. 2011

13. He J, et al. Generalizability and epidemiologic characterization of eleven colorectal cancer GWAS hits in multiple populations. Cancer Epidemiol Biomarkers Prev. 2011; 20:70–81. [PubMed: 21071539]

14. Zheng W, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet. 2009; 41:324–328. [PubMed: 19219042]

15. Bei JX, et al. A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. Nat Genet. 2010; 42:599–603. [PubMed: 20512145]

16. Jee SH, et al. Adiponectin concentrations: a genome-wide association study. Am J Hum Genet. 2010; 87:545–552. [PubMed: 20887962]

17. Nakata I, et al. Association between the SERPING1 gene and age-related macular degeneration and polypoidal choroidal vasculopathy in Japanese. PLoS One. 2011; 6:e19108. [PubMed: 21526158]

18. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010; 34:816–834. [PubMed: 21058334]

19. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010; 26:2190–2191. [PubMed: 20616382]

20. Peters U, et al. Meta-analysis of new genome-wide association studies of colorectal cancer risk. Hum Genet. 2012; 131:217–234. [PubMed: 21761138]

21. Figueiredo JC, et al. Genotype-environment interactions in microsatellite stable/microsatellite instability-low colorectal cancer: results from a genome-wide association study. Cancer Epidemiol Biomarkers Prev. 2011; 20:758–766. [PubMed: 21357381]

22. Musgrove EA, Caldon CE, Barraclough J, Stone A, Sutherland RL. Cyclin D as a therapeutic target in cancer. Nat Rev Cancer. 2011; 11:558–572. [PubMed: 21734724]

23. Mermelshtein A, et al. Expression of D-type cyclins in colon cancer and in cell lines from colon carcinomas. Br J Cancer. 2005; 93:338–345. [PubMed: 16012517]

24. Sarkar R, et al. Expression of cyclin D2 is an independent predictor of the development of hepatic metastasis in colorectal cancer. Colorectal Dis. 2010; 12:316–323. [PubMed: 19508551]

25. Gundem G, et al. IntOGen: integration and data mining of multidimensional oncogenomic data. Nat Methods. 2010; 7:92–93. [PubMed: 20111033]

26. Matys V, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006; 34:D108–D110. [PubMed: 16381825]

27. Kolfschoten IG, et al. A genetic screen identifies PITX1 as a suppressor of RAS activity and tumorigenicity. Cell. 2005; 121:849–858. [PubMed: 15960973]

28. Chen Y, et al. Decreased PITX1 homeobox gene expression in human lung cancer. Lung Cancer. 2007; 55:287–294. [PubMed: 17157953]

29. Chen YN, Chen H, Xu Y, Zhang X, Luo Y. Expression of pituitary homeobox 1 gene in human gastric carcinogenesis and its clinicopathological significance. World J Gastroenterol. 2008; 14:292–297. [PubMed: 18186570]

30. Lord RV, et al. Increased CDX2 and decreased PITX1 homeobox gene expression in Barrett's esophagus and Barrett's-associated adenocarcinoma. Surgery. 2005; 138:924–931. [PubMed: 16291394]

31. Nagel S, et al. Activation of Paired-homeobox gene PITX1 by del(5)(q31) in T-cell acute lymphoblastic leukemia. Leuk Lymphoma. 2011; 52:1348–1359. [PubMed: 21425961]

32. Watanabe T, et al. Differential gene expression signatures between colorectal cancers with and without KRAS mutations: crosstalk between the KRAS pathway and other signalling pathways. Eur J Cancer. 2011; 47:1946–1954. [PubMed: 21531130]

33. Liu DX, Lobie PE. Transcriptional activation of p53 by Pitx1. Cell Death Differ. 2007; 14:1893–1907. [PubMed: 17762884]

34. Qi DL, et al. Identification of PITX1 as a TERT suppressor gene located on human chromosome 5. Mol Cell Biol. 2011; 31:1624–1636. [PubMed: 21300782]

35. Knosel T, et al. Loss of desmocollin 1–3 and homeobox genes PITX1 and CDX2 are associated with tumor progression and survival in colorectal carcinoma. Int J Colorectal Dis. 2012

36. Zeller T, et al. Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. PLoS One. 2010; 5:e10693. [PubMed: 20502693]

37. Long J, et al. Genome-wide association study in East asians identifies novel susceptibility Loci for breast cancer. PLoS Genet. 2012; 8:e1002532. [PubMed: 22383897]

38. Shu XO, et al. Identification of new genetic risk variants for type 2 diabetes. PLoS Genet. 2010; 6

39. Zheng W, et al. Genetic and clinical predictors for breast cancer risk assessment and stratification among Chinese women. J Natl Cancer Inst. 2010; 102:972–981. [PubMed: 20484103]

40. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]

41. Sinnott JA, Kraft P. Artifact due to differential error when cases and controls are imputed from different platforms. Hum Genet. 2012; 131:111–119. [PubMed: 21735171]

42. Jiao S, Hsu L, Hutter CM, Peters U. The use of imputed values in the meta-analysis of genome-wide association studies. Genet Epidemiol. 2011; 35:597–605. [PubMed: 21769935]

43. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]

44. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. Ann Intern Med. 1997; 127:820–826. [PubMed: 9382404]

45. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002; 21:1539–1558. [PubMed: 12111919]

46. Zhang B, Beeghly-Fadiel A, Long J, Zheng W. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. Lancet Oncol. 2011; 12:477–488. [PubMed: 21514219]

47. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005; 21:263–265. [PubMed: 15297300]

48. Pruim RJ, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010; 26:2336–2337. [PubMed: 20634204]
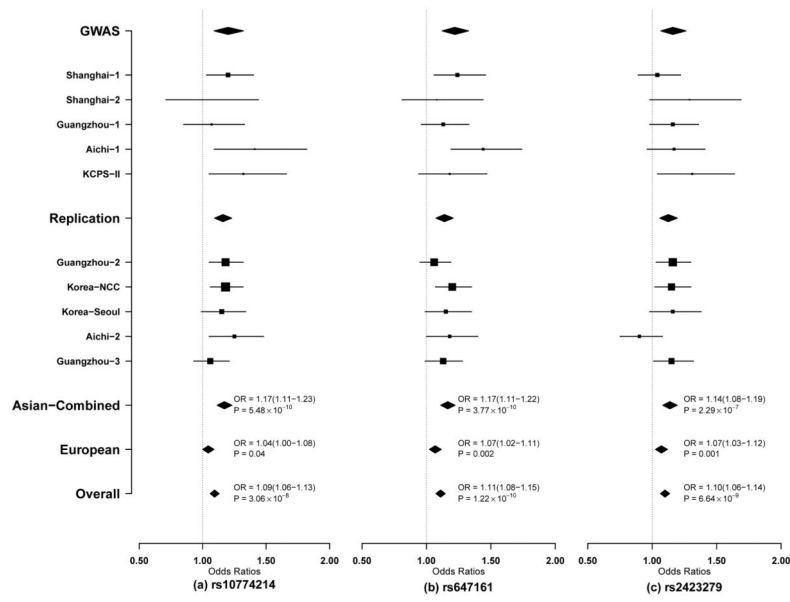
Appendix

(Service de Génétique Médicale, CHU Nantes, Nantes, France), Hermann Brenner (Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg, Germany), Katja Butterbach (Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg, Germany), Bette J. Caan (Division of Research, Kaiser Permanente Medical Care Program, Oakland, CA USA), Christopher S. Carlson (Public Health Sciences Division, Fred Hutchinson Cancer Research Center; School of Public Health, University of Washington, Seattle, WA, USA), Graham Casey (Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA), Andrew T. Chan (Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School; Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA), Jenny Chang-Claude (Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany), Stephen J. Chanock (Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA), Lin S. Chen (Department of Health Studies, University of Chicago, Chicago, IL, USA), Gerhard A. Coetzee (Keck School of Medicine, University of Southern California, Los Angeles, CA, USA), Simon G. Coetzee (Keck School of Medicine, University of Southern California, Los Angeles, CA, USA), David V. Conti (Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA), Keith Curtis (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA), David Duggan (Translational Genomics Research Institute, Phoenix, Arizona, USA), Todd L. Edwards (Division of Epidemiology, Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN, USA), Charles S. Fuchs (Department of Medical Oncology, Dana Farber Cancer Institute; Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA), Steven Gallinger (Department of Surgery, Mount Sinai Hospital; Samuel Lunenfeld Research Institute, Toronto, ON, Canada), Edward L. Giovannucci (Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School; Departments of Epidemiology and Nutrition, Harvard School of Public Health, Boston, MA, USA), Stephanie M. Gogarten (School of Public Health, University of Washington, Seattle, WA, USA), Stephen B. Gruber (Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA), Robert W. Haile (Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA), Tabitha A. Harrison (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA), Richard B. Hayes (Division of Epidemiology, Department of Environmental Medicine, New York University School of Medicine, New York, NY, USA), Michael Hoffmeister (Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg, Germany), John L. Hopper (Melborne School of Population Health, The University of Melborne, Melbourne, VIC, Australia), Li Hsu (Public Health Sciences Division, Fred Hutchinson Cancer Research Center; Department of Biostatistics, University of Washington, Seattle, WA, USA), Thomas J. Hudson (Department of Medical Biophysics and Department Molecular Genetics, University of Toronto, Toronto, ON, Canada), David J. Hunter (Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA), Carolyn M. Hutter (Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA), Rebecca D. Jackson (Division of

Endocrinology, Diabetes, and Metabolism, Ohio State University, Columbus, OH, USA), Mark A. Jenkins (Melborne School of Population Health The University of Melbourne, Melbourne, VIC, Australia), Shuo Jiao (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA), Charles Kooperberg (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA), Sébastien Küry (Service de Génétique Médicale, CHU Nantes, Nantes, France), Andrea Z. LaCroix (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA), Cathy C. Laurie (Department of Biostatistics, University of Washington, Seattle, WA, USA), Cecelia A. Laurie (Department of Biostatistics, University of Washington, Seattle, WA, USA), Loic Le Marchand (Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA), Mathieu Lemire (Ontario Institute for Cancer Research, Toronto, ON, Canada), David Levine (School of Public Health, University of Washington, Seattle, WA, USA), Noralane M. Lindor (Department of Health Sciences Research, Mayo Clinic, Scottsdale, AZ, USA), Yan Liu (Stephens and Associates, Carrollton, TX, USA), Jing Ma (Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA), Karen W. Makar (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA), Polly A. Newcomb (Public Health Sciences Division, Fred Hutchinson Cancer Research Center; Department of Epidemiology, University of Washington School of Public Health, Seattle, WA, USA), Ulrike Peters (Public Health Sciences Division, Fred Hutchinson Cancer Research Center; Department of Epidemiology, University of Washington School of Public Health, Seattle, WA, USA), John D. Potter (Public Health Sciences Division, Fred Hutchinson Cancer Research Center; Department of Epidemiology, Univer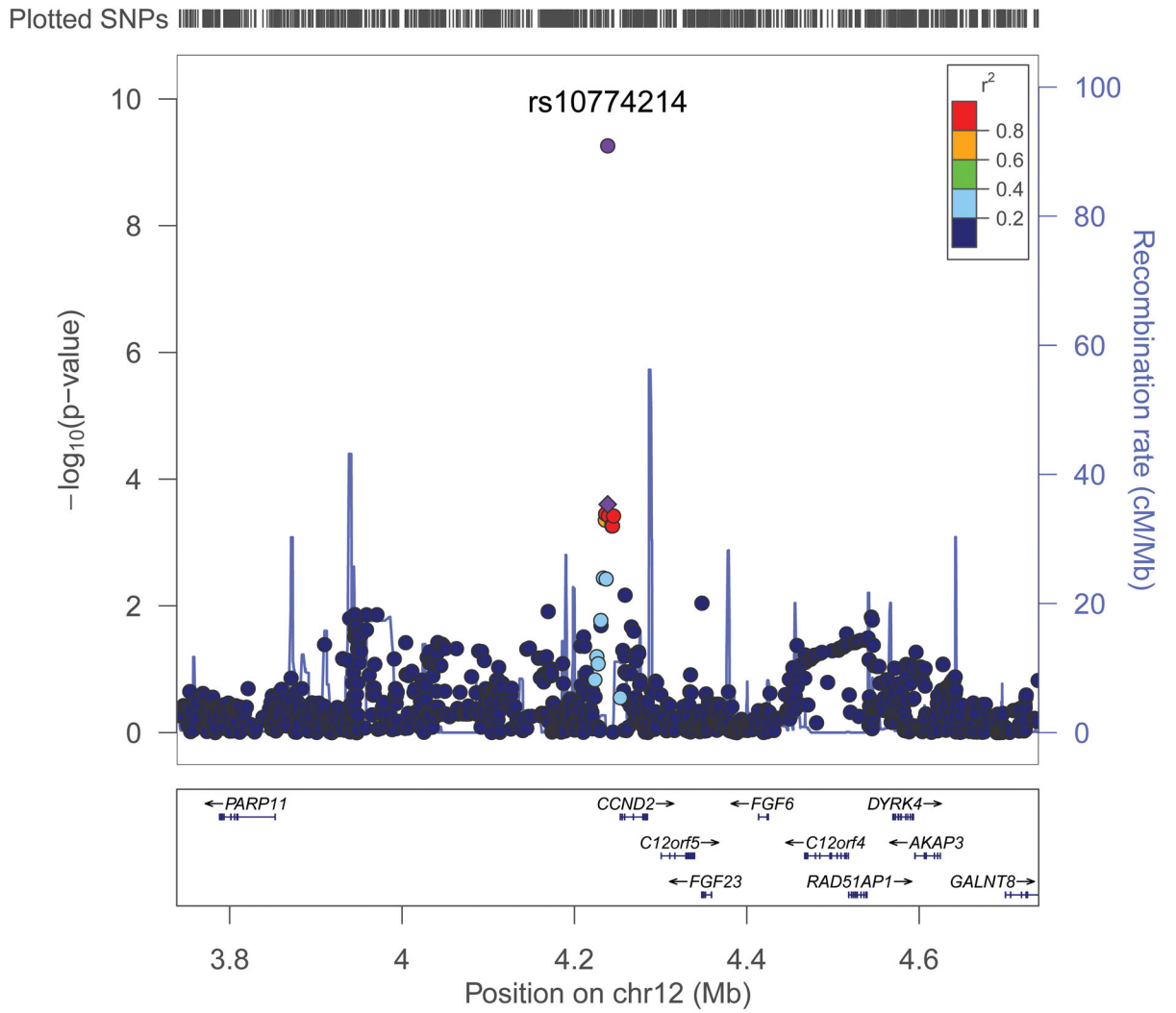sity of Washington School of Public Health, Seattle, WA, USA; Centre for Public Health Research, Massey University, Palmerston North, New Zealand), Ross L. Prentice (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA), Conghui Qu (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA), Thomas Rohan (Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Yeshiva University, Bronx, NY, USA), Robert E. Schoen (Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, PA), Fredrick R. Schumacher (Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA), Daniela Seminara (Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA), Martha L. Slattery (Department of Internal Medicine, University of Utah Health Sciences Center, Salt Lake City, UT, USA), Darin Taverna (Translational Genomics Research Institute, Phoenix, Arizona, USA), Stephen N. Thibodeau (Departments of Laboratory Medicine and Pathology and Laboratory Genetics, Mayo Clinic, Rochester, MN, USA), Cornelia M. Ulrich (Division of Preventive Oncology, German Cancer Research Center, Heidelberg, Germany; Public Health Sciences Division, Fred Hutchinson Cancer Research Center; Department of Epidemiology, University of Washington School of Public Health, Seattle, WA, USA), Raakhee Vijayaraghavan(Genetic Basis of Human Disease Division, Translational Genomics Research Institute, Phoenix, AZ, USA), Bruce Weir (Department of Biostatistics, University of Washington, Seattle, WA, USA), Emily White (Public Health Sciences Division, Fred Hutchinson Cancer Research Center; Department of Epidemiology, University of Washington School of Public Health, Seattle, WA, USA), and
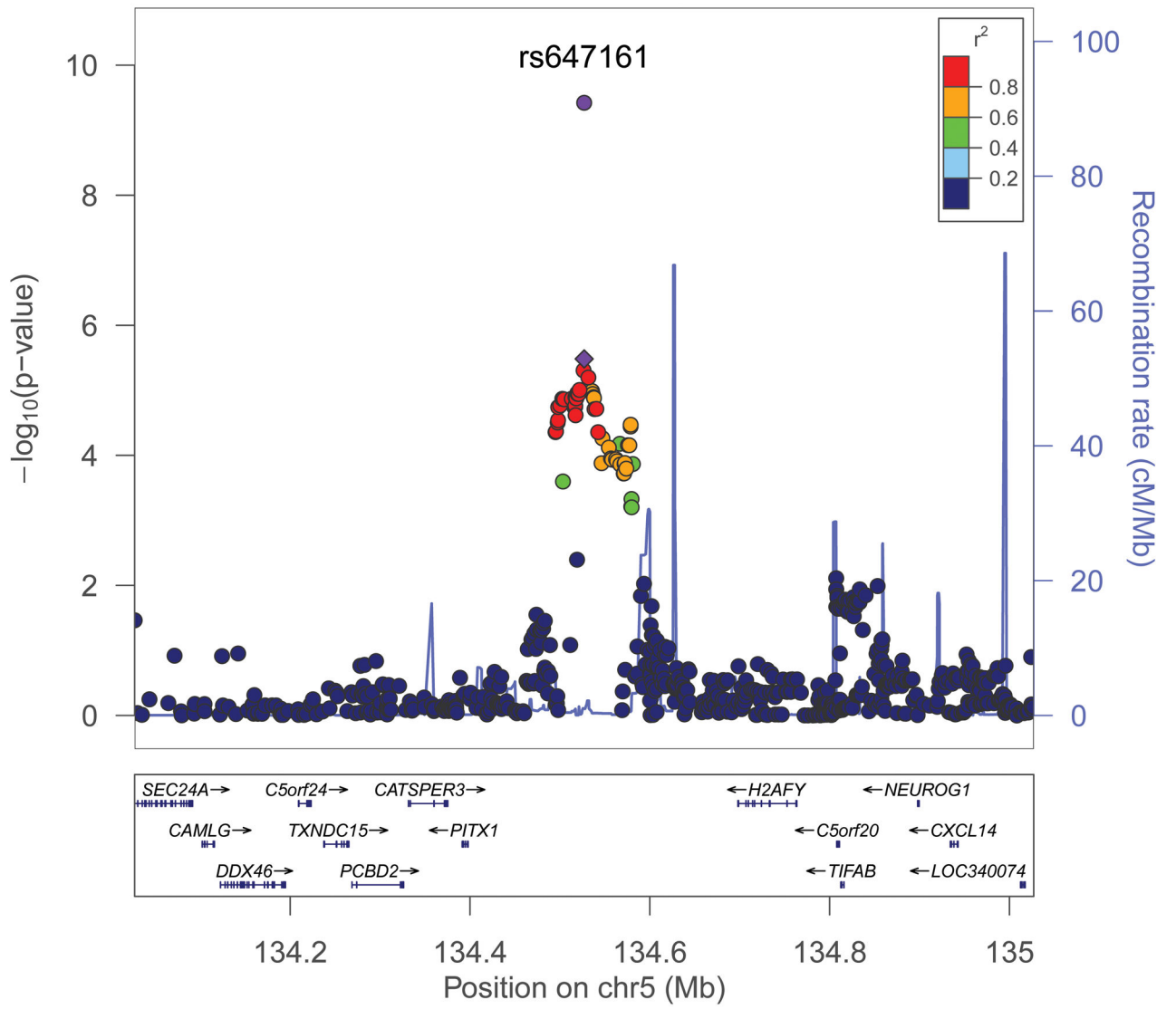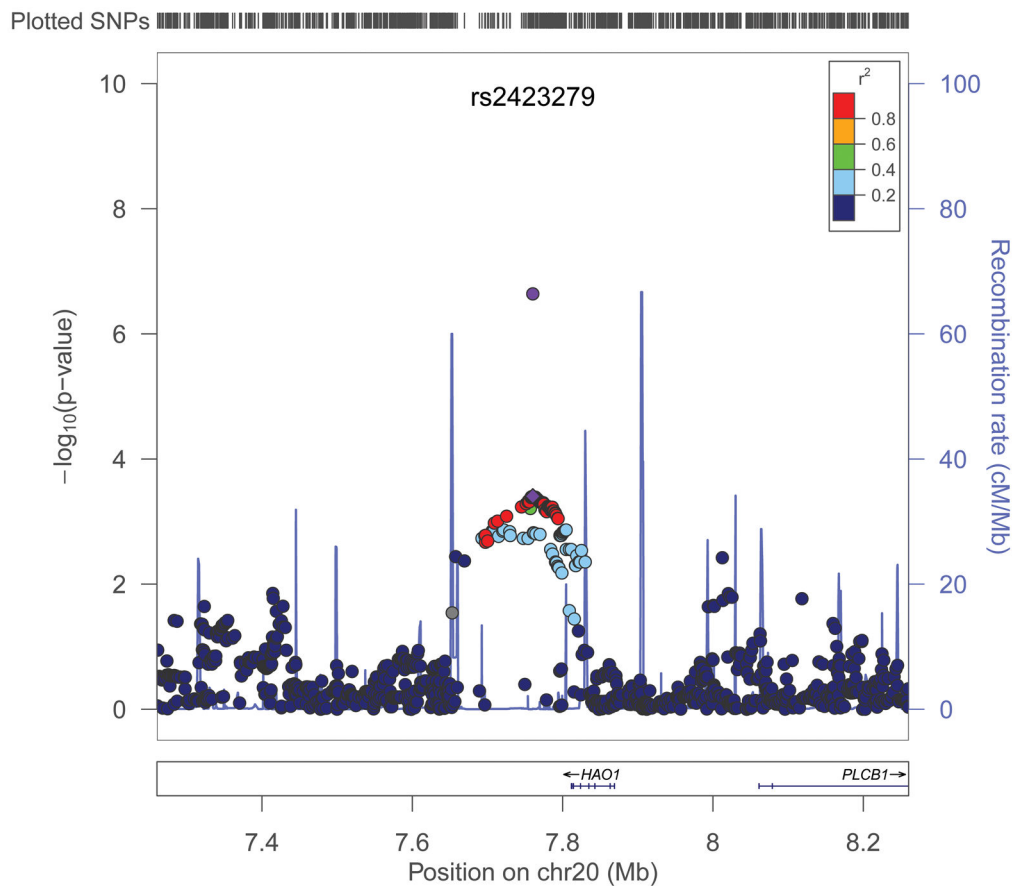
Figure 1. Forest plots for the three SNPs showing evidence of an association with CRC risk
Per-allele ORs are presented with the area of the box proportional to the inverse variance weight of the estimate. Horizontal lines represent 95% CIs.

Figure 2. Regional plots of association results and recombination rates for the three SNPs showing evidence of an association with CRC risk
Genotyped and imputed data from GWAS samples are plotted based on their chromosomal position in NCBI Human Genome Build 36.3. For each region, the SNP selected for Stage 2 replication is denoted with a diamond, and *P*-value from the combined analysis of Stages 1 and 2 data is provided. Data are shown for (a) rs10774214, (b) rs647161, and (c) rs2423279.

Table 1

Association of colorectal cancer risk with the top four risk variants identified in East Asian samples

SNP (alleles)[a]	Chr. (gene)[b]	Location (bp)[c]	Stage	Cases Sample size	Cases MAF	Controls Sample size	Controls MAF	Per-allele association OR (95% CI)[d]	Per-allele association P_{trend}	Heterogeneity P^{e}	Heterogeneity I^2
rs10774214 (T/C)	12p13.32 (CCND2)	4,238,613	GWAS	2,098	0.373	5,749	0.348	1.20 (1.09–1.32)	2.03×10^{-4}		
			Replication	5,197	0.381	5,797	0.355	1.16 (1.09–1.23)	5.80×10^{-7}		
			Overall	7,295	0.379	11,546	0.352	1.17 (1.11–1.23)	5.48×10^{-10}	0.615	0%
rs647161 (A/C)	5q31.1 (PITX1)	134,526,991	GWAS	2,098	0.353	5,749	0.308	1.22 (1.12–1.33)	3.29×10^{-6}		
			Replication	5,217	0.344	5,815	0.319	1.14 (1.07–1.21)	1.15×10^{-5}		
			Overall	7,315	0.347	11,564	0.313	1.17 (1.11–1.22)	3.77×10^{-10}	0.444	0%
rs2423279 (C/T)	20p12.3 (HAO1)	7,760,350	GWAS	2,098	0.339	5,749	0.307	1.16 (1.07–1.26)	4.96×10^{-4}		
			Replication	5,227	0.315	5,811	0.297	1.13 (1.06–1.19)	1.22×10^{-4}		
			Overall	7,325	0.322	11,560	0.302	1.14 (1.08–1.19)	2.29×10^{-7}	0.331	12%
rs1665650 (T/C)	10q26.12 (HSPA12A)	118,477,090	GWAS	2,098	0.346	5,749	0.310	1.20 (1.10–1.31)	3.88×10^{-5}		
			Replication	5,192	0.328	5,808	0.320	1.10 (1.04–1.17)	0.0018		
			Overall	7,290	0.333	11,557	0.315	1.13 (1.08–1.19)	8.58×10^{-7}	0.404	4%

Abbreviations: Chr., Chromosome; MAF, minor allele frequency; OR, odds ratio; CI, confidence interval.

[a] Minor/major allele for East Asians, OR was estimated for the minor allele.

[b] The closest gene.

[c] Location based on NCBI Human Genome Build 36.3.

[d] Adjusted for age, sex, the first ten principal components (Stage 1) and study site.

[e] P for heterogeneity across studies in GWAS and Replication was calculated using a Cochran's Q test.

Table 2

Association of colorectal cancer risk with the three newly-identified risk variants in European-ancestry populations and the meta-analyses of East Asians and Europeans

SNP	Alleles[a]	MAF[b] Cases	Controls	Europeans[c] Cases/controls	OR (95% CI)	P_{meta}	East Asians and Europeans combined[c] Cases/controls	OR (95% CI)	P_{meta}
rs10774214	T/C	0.385	0.379	11,870/14,190	1.04 (1.00–1.09)	0.040	19,165/25,736	1.09 (1.06–1.13)	3.06×10^{-8}
rs647161	A/C	0.680	0.667	11,870/14,190	1.07 (1.02–1.11)	0.002	19,185/25,754	1.11 (1.08–1.15)	1.22×10^{-10}
rs2423279	C/T	0.263	0.252	11,870/14,190	1.07 (1.03–1.12)	0.001	19,195/25,750	1.10 (1.06–1.14)	6.64×10^{-9}

[a] Alelles (minor/major) as shown in Table 2 for East Asians.

[b] Minor allele frequency (MAF) in European-ancestry populations.

[c] Summary statistics were generated using inverse-variance weighted, fixed-effects meta-analysis.