HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2014 December 01.

Published in final edited form as:

Nat Genet. 2014 June ; 46(6): 533–542. doi:10.1038/ng.2985.

Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk

Ben Zhang[1], **Wei-Hua Jia**[2], **Koichi Matsuda**[3], **Sun-Seog Kweon**[4,5], **Keitaro Matsuo**[6], **Yong-Bing Xiang**[7], **Aesun Shin**[8,9], **Sun Ha Jee**[10], **Dong-Hyun Kim**[11], **Qiuyin Cai**[1], **Jirong Long**[1], **Jiajun Shi**[1], **Wanqing Wen**[1], **Gong Yang**[1], **Yanfeng Zhang**[1], **Chun Li**[12], **Bingshan Li**[13], **Yan Guo**[14], **Zefang Ren**[15], **Bu-Tian Ji**[16], **Zhi-Zhong Pan**[2], **Atsushi Takahashi**[17], **Min-Ho Shin**[4], **Fumihiko Matsuda**[18], **Yu-Tang Gao**[7], **Jae Hwan Oh**[19], **Soriul Kim**[10], **Yoon-Ok Ahn**[9], **Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO)**[20], **Andrew T Chan**[21,22], **Jenny Chang-Claude**[23], **Martha L. Slattery**[24], **Colorectal Transdisciplinary (CORECT) Study**[20], **Stephen B. Gruber**[25], **Fredrick R. Schumacher**[25], **Stephanie L. Stenzel**[25], **Colon Cancer Family Registry (CCFR)**[20], **Graham Casey**[25], **Hyeong-Rok Kim**[26], **Jin-Young Jeong**[11], **Ji Won Park**[19,27], **Hong-Lan Li**[7], **Satoyo Hosono**[6], **Sang-Hee Cho**[28], **Michiaki Kubo**[17], **Xiao-Ou Shu**[1], **Yi-Xin Zeng**[2], and **Wei Zheng**[1]

[1]Division of Epidemiology, Department of Medicine, Vanderbilt-Ingram Cancer Center, Vanderbilt Epidemiology Center, Vanderbilt University School of Medicine, Nashville, Tennessee, the United States

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

[3]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1, Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

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

[20]A complete list of members is provided in the Acknowledgements.

AUTHOR CONTRIBUTIONS

W.Z. conceived and directed the Asia Colorectal Cancer Consortium and the Shanghai-Vanderbilt Colorectal Cancer Genetics Project. W.H.J. and Y.X.Z., K. Matsuda, S.S.K., K. Matsuo, X.O.S., Y.B.X. and Y.T.G., A.S., S.H.J., D.H.K., U.P., S.B.G, and G.C. directed CRC projects in Guangzhou, BBJ, HCES-CRC, Aichi, Shanghai, Korea-NCC, KCPS-II, Korea-Seoul, GECCO, CORECT, and CCFR, respectively. B.Z., Q.C. and W.W. coordinated the project. Q.C. directed the lab operations. J.S. performed the genotyping experiments. B.Z. performed the statistical and bioinformatic analyses. W.W. contributed to the statistical analyses and data interpretation. A.T. conducted the statistical analyses and imputation for BBJ. B.Z. W.W. and J.L. managed the data. Y.Z. and B.Z. performed the expression analysis of TCGA data. B.Z. and W.Z. wrote the paper with significant contributions from X.O.S., Q.C., J.L., W.W., B.L. and Y.Z. All authors contributed to data and biological sample collection in the original studies included in this project and/or manuscript revision. H.B., J.A.B., K.B., S.B., S.I.B., A.T.C., B.J.C., C.S.C., D.V.C., G.C., K.C., P.T.C., S.J.C., J.C.-C., D.D., C.K.E., C.S.F., J.F., E.L.G., J.G., R.C.G., S.G., S.B.G., W.J.G., C.M.H., D.J.H., J.D.H., J.F.H., J.L.H., L.H., M.H., R.B.H., R.W.H., T.A.H., T.J.H., E.J.J., M.A.J., R.D.J., S.J., S.K., L.L., M.L., N.M.L., Y.L., L.Le Marchand, B.M., F.J.M., J.M., V.M., P.A.N, J.D.P., U.P., C.Q., L.R., T.R., D.S., F.R.S., G.S., M.L.S., R.E.S., S.L.S., D.C.T., S.N.T., E.W., E.W. and B.W.Z. contributed to data and biological sample collection for the studies included in GECCO, CORECT and CCFR. All authors have reviewed and approved the content of the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

[4]Department of Preventive Medicine, Chonnam National University Medical School, Gwangju, South Korea

[5]Jeonnam Regional Cancer Center, Chonnam National University Hwasun Hospital, Hwasun, South Korea

[6]Department of Preventive Medicine, Kyushu University Faculty of Medical Sciences, 3-1-1 Maidashi, Higashi-ku, Fukuoka 812-8582, Japan

[7]Department of Epidemiology, Shanghai Cancer Institute, Shanghai, China

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

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

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

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

[12]Department of Biostatistics, Vanderbilt University School of Medicine, Tennessee, the United States

[13]Department of Molecular Physiology and Biophysics, Vanderbilt University School of Medicine, Tennessee, the United States

[14]Department of Cancer Biology, Vanderbilt University School of Medicine, Nashville, Tennessee, the United States

[15]School of Public Health, Sun Yat-sen University, Guangzhou, China

[16]Division of Cancer Epidemiology & Genetics, National Cancer Institute, Bethesda, Maryland, the United States

[17]Center for Integrative Medical Sciences, The Institute of Physical and Chemical Research (RIKEN), Kanagawa, Japan

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

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

[21]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, the United States

[22]Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, the United States

[23]Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany

[24]Department of Internal Medicine, University of Utah Health Sciences Center, Salt Lake City, Utah, the United States

[25]USC Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, the United States

[26]Department of Surgery, Chonnam National University Medical School, Gwangju, South Korea

[27]Department of Surgery, Seoul National University Hospital, Seoul, South Korea

[28]Department of Hemato-oncology, Chonnam National University Medical School, Gwangju, South Korea

Abstract

Known genetic loci explain only a small proportion of the familial relative risk of colorectal cancer (CRC). We conducted the largest genome-wide association study in East Asians with 14,963 CRC cases and 31,945 controls and identified six new loci associated with CRC risk ($P = 3.42 \times 10^{-8}$ to 9.22×10^{-21}) at 10q22.3, 10q25.2, 11q12.2, 12p13.31, 17p13.3 and 19q13.2. Two of these loci map to genes (*TCF7L2* and *TGFB1*) with established roles in colorectal tumorigenesis. Four other loci are located in or near genes involved in transcription regulation (*ZMIZ1*), genome maintenance (*FEN1*), fatty acid metabolism (*FADS1* and *FADS2*), cancer cell motility and metastasis (*CD9*) and cell growth and differentiation (*NXN*). We also found suggestive evidence for three additional loci associated with CRC risk near genome-wide significance at 8q24.11, 10q21.1 and 10q24.2. Furthermore, we replicated 22 previously reported CRC loci. Our study provides insights into the genetic basis of CRC and suggests new biological pathways.

Colorectal cancer (CRC) is a leading cause of cancer morbidity and mortality worldwide [1]. It is well established that genetic factors play a significant role in the etiology of CRC [2, 3]. Deleterious germline mutations in known susceptibility genes, notably *APC* (adenomatous polyposis coli), *MLH1*, *MSH2*, *MSH6* and *PMS2*, confer high risk of CRC in hereditary cancer syndromes [3–6]. Most sporadic CRC cases, however, do not carry these high-penetrance mutations [3, 4]. Since 2007, genome-wide association studies (GWAS) and subsequent fine-mapping analyses conducted in European descendants have identified 21 low-penetrance susceptibility loci associated with CRC risk [7–17]. Together, these common loci explain less than 10% of the familial relative risk of CRC in European populations [13, 14]. In a GWAS of 7,456 CRC cases and 11,671 controls conducted as part of the Asia Colorectal Cancer Consortium, we identified three new loci at 5q31.1 (near *PITX1*), 12p13.32 (near *CCND2*) and 20p12.3 (near *HAO1*) associated with CRC risk [18]. In addition, we discovered a new risk variant in the *SMAD7* gene associated with CRC among East Asians [19]. Over the past two years, we have doubled the sample size in the Asia Colorectal Cancer Consortium and conducted a four-stage GWAS including 14,963 CRC cases and 31,945 controls to identify additional susceptibility loci for CRC.

RESULTS

We performed a fixed-effects meta-analysis to evaluate approximately 2.4 million genotyped or imputed SNPs in 22 autosomes from five GWAS (stage 1) conducted in China, Japan and South Korea, totaling 2,098 CRC cases and 6,172 cancer-free controls (Supplementary Tables 1 and 2). There was little evidence of population stratification in these studies (Supplementary Figs. 1 and 2), with genomic inflation factor λ <1.04 in any of the five studies and the meta-analysis (λ_{1000} =1.01). We selected 8,539 SNPs showing

evidence of association with CRC risk (P <0.05) according to pre-specified criteria (**ONLINE METHODS**). We also included the 31 risk variants identified by previous GWAS [7–20], resulting in a total of 8,569 SNPs. Of them, 7,113 SNPs were successfully designed using Illumina Infinium assays as part of a large genotyping effort for multiple projects. Using this customized array, we genotyped an independent set of 3,632 CRC cases and 6,404 controls recruited in three studies (stage 2) conducted in China. After quality control exclusions, 6,899 SNPs remained for the analysis in 3,519 cases and 6,275 controls. We evaluated associations between CRC risk and these SNP in each study separately and then performed a fixed-effects meta-analysis to obtain the summary estimates. Again, we observed little evidence of population stratification either in the three studies individually (λ <1.05) or combined (λ = 1.05, λ_{1000} = 1.01) (Supplementary Fig. 3). In a meta-analysis of data from stages 1 and 2, we identified 559 SNPs showing evidence of association at P <0.005. We then evaluated these SNPs using data from a large Japanese CRC GWAS (stage 3) with 2,814 CRC cases and 11,358 controls [20]. Thirty SNPs in 25 new loci were associated with CRC risk at P <0.0001 in the meta-analysis of data from stages 1 to 3 and at P <0.01 in the meta-analysis of stages 2 and 3. Of them, 29 were successfully genotyped in an independent sample of 6,532 CRC cases and 8,140 controls from five additional studies (stage 4) conducted in China, South Korea and Japan.

Newly identified risk loci for CRC

In the meta-analysis of all data for the 29 SNPs from stages 1 to 4 with 14,963 CRC cases and 31,945 controls, signals from ten SNPs, representing six new loci, showed convincing evidence for an association with CRC risk at the genome-wide significance level (P <5×10^{-8}) including: rs704017 at 10q22.3; rs11196172 at 10q25.2; rs174537, rs4246215, rs174550 and rs1535 at 11q12.2; rs10849432 at 12p13.31; rs12603526 at 17p13.3; and rs1800469 and rs2241714 at 19q13.2 (Table 1, Supplementary Tables 3 and 4, and Supplementary Fig. 4). Associations of CRC risk with the top SNPs in each of the six loci were consistent across almost all studies with no evidence of heterogeneity (Fig. 1). With the exception of rs10849432 intergenic to 12p13.31, the remaining nine newly identified risk variants are located in the exonic, promoter, three prime untranslated region (3′-UTR) or intronic regions of known genes (Table 1). The linkage disequilibrium (LD) blocks (r^2 >0.5) tagged by rs704017 (10q22.3), rs174537 (11q12.2), and rs1800469 (19q13.2), each span multiple genes (Supplementary Table 5). The LD blocks tagged by rs11196172 (10q25.2) and rs12603526 (17p13.3), each lie within a single gene. The LD block tagged by rs10849432 (12p13.31) does not contain any known genes. Stratification analyses of the newly identified risk variants by tumor anatomic site (colon, rectum), population (Chinese, Korean, and Japanese), and sex (men, women) did not reveal any significant heterogeneity (Supplementary Tables 6 to 8). In addition to the six newly identified loci, three additional regions also showed an association with CRC risk near genome-wide significance at 8q24.11 (rs6469656, P =5.38×10^{-8}), 10q21.1 (rs4948317, P =7.14×10^{-8}) and 10q24.2 (rs12412391, P =7.41×10^{-7}). Results for all 29 SNPs across stage 1 to stage 4 are presented in Supplementary Table 3.

We performed conditional analyses for SNPs within a 1-mb region centered on the index SNPs in each of the six newly identified loci. No second signal was identified at P <0.01

after adjusting for the respective index SNPs (data not shown). Four SNPs at 11q12.2 and two SNPs at 19q13.2 showed association with CRC risk at $P < 5\times10^{-8}$, and thus we performed haplotype analysis for these two loci using genotype data available for 10,051 CRC cases and 14,415 controls (stages 2 and 4). Two common haplotypes were found in the 11q12.2 locus, accounting for more than 99% of the haplotypes constructed using the four highly correlated SNPs. The haplotype with all four risk alleles (frequency =0.574 in controls) was strongly associated with CRC risk (odds ratio (OR) =1.40, 95% confidence interval (CI): 1.29–1.51; $P = 3.69\times10^{-16}$) (Supplementary Table 9). Similarly, we identified two common haplotypes in the 19q13.2 locus, accounting for more than 99% of the haplotypes constructed using the two highly correlated SNPs. The haplotype with the risk allele in both SNPs (frequency =0.485 in controls) was also associated with increased risk of CRC (OR=1.16, 95% CI: 1.08–1.26; $P = 1.18\times10^{-4}$) (Supplementary Table 10). Therefore, these analyses did not reveal an independent signal in any of the six newly identified loci.

We examined potential SNP-SNP interactions between the six new risk variants (rs704017, rs11196172, rs174537, rs10849432, rs12603526, and rs1800469) identified in this study and also between these six SNPs and the risk variants in 25 previously reported loci (Supplementary Table 11). Multiplicative interactions were found with suggestive evidence ($P < 0.05$) for seven pairs of SNPs. None of these interactions, however, remain statistically significant after correcting for multiple comparisons of 180 tests (adjusted $P = 0.000277$).

We evaluated associations of the ten newly identified SNPs with CRC risk in European descendants using data from three consortia, the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) [17], the Colorectal Transdisciplinary (CORECT) Study and the Colon Cancer Family Registry (CCFR) [21], with a total sample size of 16,984 CRC cases and 18,262 controls (Supplementary Table 12). In a meta-analysis of data from these consortia, all ten SNPs showed associations with CRC risk in the same direction as observed in East Asians (Table 2). Five SNPs in two loci (10q22.3 and 11q12.2) were associated with CRC risk at $P < 0.008$ (corrected for multiple comparisons of six loci). The strength of these associations in Europeans, however, was weaker than in East Asians. Tests for heterogeneity were statistically significant for risk variants in 11q12.2 and 19q13.2 ($P < 0.008$). The frequency of the risk allele also differed considerably between Europeans and East Asians for SNPs in five loci (Supplementary Table 13). For example, rs12603526 is common in East Asians, whereas the minor allele frequency (MAF) is <0.02 in Europeans. These differences may partly reflect distinct patterns of LD between the index SNPs and causal SNPs in these two populations. As expected, LD patterns for most of the newly identified loci differed considerably between Europeans and East Asians (Supplementary Fig. 5). Large-scale fine-mapping of these loci will be helpful to identify causal variants.

Putative functional variants and candidate genes

We evaluated and annotated putative functional variants and candidate genes in each of the six newly identified loci using data from the 1000 Genomes Project [22], HapMap 2 [23], Encyclopedia of DNA Elements (ENCODE) [24], expression quantitative trait locus (eQTL) databases [25–28], the Catalogue of Somatic Mutations in Cancer (COSMIC) [29], The Cancer

Genome Atlas (TCGA) CRC project [30], Gene Expression Atlas [31], PubMed and OMIM (**ONLINE METHODS**). We summarize results below for each locus.

At the 10q25.2 locus, rs11196172 is located in intron 4 of the *TCF7L2* gene. The SNP and other correlated SNPs (r^2 >0.5) fall within a strong enhancer activity region and a DNase I hypersensitivity site annotated by ENCODE (Supplementary Table 14), suggesting a potentially functional role for these SNPs. We found that the risk allele of rs11196172 was significantly associated with increased expression of the *TCF7L2* gene (P =0.003) in colon tumor tissue using TCGA data (Fig. 2). The *TCF7L2* gene encodes TCF7L2 (previously known as TCF4), which is key transcription factor in the Wnt signaling pathway. Aberrant activation of Wnt signaling is found in more than 90% of CRC [30]. TCF7L2 is a known tumor suppressor for CRC. Loss of TCF7L2 function enhances CRC cell growth, whereas gain of function suppresses CRC cell growth [32, 33]. The *TCF7L2* gene is one of the most frequently mutated genes in CRC, with estimated point mutation rates of approximately 8 to 12.5% [29, 30]. Although *TCF7L2* is the only gene in this locus (Supplementary Fig 4), we also found that the risk allele of rs11196172 was significantly associated with increased expression of the *VTI1A* gene (P =5.1×10^{-4}) in colon tumor tissue (Fig. 2). The *VTI1A* gene is located approximately 131 kb upstream of the *TCF7L2* gene and mRNA levels of these two genes are highly correlated in colon tumor tissues (r =0.71, P <0.0001). Recently, a recurrent gene fusion of the first three exons of *VTI1A* to the fourth exon of *TCF7L2* has been found in approximately 3% of colorectal tumors [34]. It is possible that the *VTI1A* gene may also be involved in the association between rs11196172 and CRC risk.

At the 19q13.2 locus, we identified two perfectly correlated SNPs (rs1800469 and rs2241714, r^2 =1) associated with CRC risk. Of them, rs1800469 has been previously investigated in relation to CRC risk in many small candidate-gene association studies with conflicting results [5]. We herein provide, for the first time, convincing evidence for this association through our GWAS. SNP rs1800469 maps to the promoter of the *TGFB1* gene, while rs2241714 is a nonsynonymous SNP that results in an amino acid substitution on codon 11 of the B9D2 protein. The A allele of rs1800469 has been related to higher levels of transcription activity of the *TGFB1* gene and higher circulating levels of the TGF-β1 protein than the G allele [35]. Both rs1800469 and rs2241714 are in perfect LD with another nonsynonymous SNP rs1800470, which causes a proline to leucine substitution at codon 10 of the TGF-β1 protein. Although the two nonsynonymous SNPs are predicted to be tolerant [36] or benign [37], the Pro10 variant of rs1800470 has also been associated with an increase in gene expression of *TGFB1*, TGF-β1 protein secretion and circulating levels of TGF-β1 [38–40]. While rs2241714 is an eQTL for *TGFB1*, both rs1800469 and rs2241714 are also eQTLs for other genes in this locus (Supplementary Table 15). In addition to these three SNPs, many highly correlated SNPs located in the *TGFB1* gene are suggested to have potentially regulatory functions (Supplementary Table 14). The TGF-β1 protein is a major member of the TGF-β signaling pathway. Somatic alterations of certain components (*TGFBR2*, *SMAD4*, *SMAD2* and *SMAD3*) in this pathway are estimated to affect almost half of CRC [41]. High-penetrance germline mutations in the *SMAD4* gene are known to cause juvenile polyposis, an autosomal dominant polyposis syndrome with a high risk of CRC [42]. Germline, allele-specific expression of the *TGFBR1* gene has also been shown to contribute

to increased risk of CRC [43]. To date, GWAS have identified at least six other independent SNPs that are located in or proximal to genes in the TGF-β signaling pathway (*SMAD7*, *GREM1*, *BMP2*, *BMP4* and *RHPN2*) [9, 10, 13, 19]. Our finding of an association between a genetic variant in the *TGFB1* gene and CRC risk adds further evidence for the critical role of this pathway in colorectal tumorigenesis.

At the 11q12.2 locus, the four perfectly correlated SNPs rs174537, rs4246215, rs174550 and rs1535 lie in intron 24 of *MYRF*, the 3′-UTR of *FEN1*, intron 7 of *FADS1* and intron 1 of *FADS2*, respectively. Of them, rs4246215 is an eQTL for the *FEN1* gene in normal colorectal tissue [44] and is predicted to affect miRNA binding site activity [45]. SNP rs174537 is an eQTL for the *FADS1* and *FADS2* genes in whole blood and other types of tissue (Supplementary Table 15). Using data from TCGA, we identified a strong correlation of rs1535 genotypes with *FADS2* gene expression ($P = 1.4\times10^{-5}$) in colon tumor tissue (Fig. 2). These findings suggested that the potential function of these SNPs may be mediated through their effect on their host genes. We also found that the *FEN1*, *FADS1* and *FADS2* genes are all highly expressed in colon tumor tissue compared with normal colon tissue (Supplementary Table 16). The *FEN1* gene encodes flap structure-specific endonuclease 1, a protein that is essential for DNA repair, replication and degradation and has a critical role in maintaining genome stability and protecting against carcinogenesis [46]. *FEN1* mutations have been found in several human cancers [47]. Mouse models with haploinsufficiency of *Fen1* showed rapid progression of CRC and reduced survival [48]. Two other genes in this locus, *FADS1* and *FADS2*, respectively encode delta-5 and delta-6 desaturases, which are key enzymes in polyunsaturated fatty acid metabolism. Of them, delta-6 desaturase is responsible for the synthesis of arachidonic acid [49], the precursor of prostaglandin E_2 (PGE_2), which is a key molecule mediating the effect of cyclooxygenase-2 in colorectal carcinogenesis [50]. Notably, SNPs in perfect LD with the risk variants for CRC identified in this study are strongly associated with circulating arachidonic acid level [49]. We have shown previously that high levels of urinary PGE_2 metabolite, a marker of endogenous PGE_2 production, is strongly related to elevated risk of CRC [51]. Because the LD block of approximately 190 kb tagged by the four risk variants covers many putatively functional SNPs that are located in the *FEN1*, *FADS1* and *FADS2* genes (Supplementary Table 14 and Supplementary Fig. 6), it is difficult to pinpoint a single SNP or gene that may be responsible for the association with CRC risk in this locus. Nevertheless, our study provides evidence for a potentially significant role of the *FEN1*, *FADS1* and *FADS2* genes in the etiology of CRC.

At the 10q22.3 locus, rs704017 is located in intron 3 of the *ZMIZ1-AS1* gene and resides in a strong enhancer region predicted using ENCODE data (Supplementary Fig. 6 and Supplementary Table 14). It also maps to a DNase I hypersensitivity site in the Caco-2 CRC cell line. In addition to the *ZMIZ1-AS1* gene, the LD block tagged by rs704017 also includes the *ZMIZ1* gene, which is down-regulated in the Caco-2 and HT-29 CRC cell lines [31]. In line with this, we found in TCGA data that *ZMIZ1* gene expression is reduced in colon tumor tissue compared with normal colon tissue ($P = 3.28\times10^{-6}$). In addition, somatic mutations in the *ZMIZ1* gene have been reported in more than 2% of colon tumors [29]. While *ZMIZ1-AS1* is a miscRNA gene with unknown function, the *ZMIZ1* gene encodes the

protein ZMIZ1, which regulates the activity of several transcription factors, including AR, SMAD3, SMAD4 and p53. It has been shown that ZMIZ1 may play a broader role in epithelial cancers, including CRC [52]. SNP rs704010, located in intron 1 of the *ZMIZ1* gene, has been associated with breast cancer [53]. However, this SNP, which is in weak LD (r^2 = 0.09) with the risk variant we identified for CRC, was not associated with CRC in this study (data not shown). Given the biologic function of the *ZMIZ1* gene, it is possible that this gene is involved in the association observed in this locus.

At the 12p13.31 locus, rs10849432 maps to a LD block of approximately 52 kb with no known genes. ENCODE data suggest that rs4764551 and rs4764552, perfectly correlated with rs10849432, may be located in a strong enhancer region (Supplementary Table 14). Notably, rs4764551 also maps to a DNase I hypersensitivity site in the HCT-116 CRC cell line and a binding site of the CTCF protein in the Caco-2 CRC cell line. Using data from TCGA, we showed that the closest genes to rs10849432, *CD9*, *PLEKHG6* and *TNFRSF1A* are all down-regulated in colon tumor tissue (Supplementary Table 16). The *CD9* gene encodes the CD9 antigen, which participates in many cellular processes, including differentiation, adhesion and signal transduction. Notably, CD9 plays a critical role in the suppression of cancer cell motility and metastasis [54], and overexpression of the *CD9* gene is associated with favorable prognosis of patients with CRC [55]. CD9 is also involved in suppressing Wnt signaling [56]. While function of the *PLEKHG6* gene is less clear, somatic mutations in this gene were found in approximately 2% of colon tumors [29]. The protein encoded by *TNFRSF1A* is a major receptor for tumor necrosis factor-alpha and is known to be involved in cytokine-induced senescence in cancer [57]. In addition to evidence for the three nearby genes, we found that rs4764552 is an eQTL for the *LTBR* gene (Supplementary Table 15). The LTβR protein plays an essential role in lymphoid organ formation and has also been linked to cancer [58], including CRC [59]. Based on these data, we believe that the *CD9* gene is the most likely candidate to explain the association identified in this locus. However, the potential role of other genes cannot be ruled out.

At the 17p13.3 locus, rs12603526 lies in intron 1 of the *NXN* gene, a region covering several regulatory elements, including a DNase I hypersensitivity site, a strong enhancer region and a site with an effect on regulatory motifs as annotated by ENCODE (Supplementary Table 14). *NXN* gene expression was reduced in colon tumor tissue samples included in TCGA (P =2.83×10^{-5}). Nucleoredoxin, encoded by the *NXN* gene, has functions related to cell growth and differentiation [60]. Overexpression of the *NXN* gene has been found to suppress the Wnt signaling pathway, and dysfunction of nucleoredoxin may cause activation of the transcription factor T cell factor, accelerated cell proliferation and enhancement of oncogenicity [61]. Further research is needed to determine the causal variant and biologic mechanism for the association in this locus.

Previously reported CRC loci in East Asians

We evaluated association evidence for 31 SNPs in 25 established CRC susceptibility loci [7–20] by analyzing data from stages 1 to 3 and our previous GWAS [18, 19] with a total sample size of up to 11,934 CRC cases and 28,282 controls (Table 3 and Supplementary Table 17). We found further evidence to support the association for the four loci identified

previously in our GWAS conducted among East Asians (P =1.40×10^{-10} to 3.05×10^{-15}). Of the 23 SNPs in the 18 susceptibility loci previously identified by GWAS of European descendants, 20 showed associations with CRC risk at P <0.05 among East Asians in the same direction as reported in the original studies [7–17]. These included six SNPs in four loci (1q41, 8q24.21, 10p14 and 18q21.1) with an association at P <5×10^{-8}, six SNPs in six loci with an association at P <0.002 (significance level adjusted for multiple comparisons of 24 independent loci), and eight SNPs in eight additional loci with an association at P <0.05. Three SNPs in three loci were not associated with CRC risk (P >0.05). Given that our study had a statistical power of >80% to identify an association with an OR of 1.05 at P =0.05 for SNPs with a MAF of 0.20, it is unlikely that these three SNPs confer a substantial risk of CRC in East Asian populations. Generally, loci initially identified in Europeans had smaller ORs in East Asians, with evidence of heterogeneity noted for three SNPs (P <0.002). SNPs rs6691170 and rs16892766, identified by previous GWAS of European descendants, are not polymorphic in East Asians and SNP rs5934683 is located in chromosome X. We did not have data to evaluate the association of these three SNPs with CRC risk in this study.

Familial relative risk explained by established CRC loci

The six newly identified loci in this study explain approximately 2.1% of the familial relative risk of CRC in East Asians (Supplementary Table 18). The variants, along with the four SNPs identified in our previous GWAS, explained approximately 4.3% of the familial relative risk of CRC in East Asians. An additional 3.4% of the familial relative risk in East Asians can be explained by 18 independent SNPs initially identified in studies conducted among European descendants and confirmed in this study. Based on per-allele ORs derived from previously published GWAS [7–18] and this study, we estimate that the SNPs in the 31 loci identified to date explain approximately 9% of the familial relative risk of CRC in Europeans (Supplementary Table 19), slightly higher than the 7.7% explained in East Asians.

DISCUSSION

In the largest GWAS conducted to date among East Asians, we identified six new genetic loci associated with CRC risk and provided suggestive evidence for three additional novel loci. In addition, we replicated 22 previously reported CRC susceptibility loci. Of the six newly identified loci, two map to genes (*TCF7L2* and *TGFB1*) that have established roles in colorectal tumorigenesis. The other four loci are located in or proximal to genes that are functionally important in transcription regulation (*ZMIZ1*), genome maintenance (*FEN1*), fatty acid metabolism (*FADS1* and *FADS2*), cancer cell motility and metastasis (*CD9*) and cell growth and differentiation (*NXN*). Risk variants at some loci fall within potentially functional regions and two are associated with expression levels of the *TCF7L2* and *FADS2* genes. This study expands our current understanding of the genetic basis of CRC risk and provides evidence for novel genes and biological pathways that may be involved in colorectal tumorigenesis.

Based on a large twin study conducted in Sweden, Denmark and Finland [2], the heritability estimated for CRC, breast cancer and prostate cancer was 35%, 27% and 42%, respectively.

To date, more than 70 low-penetrance susceptibility loci have been identified in GWAS for breast cancer [62] or prostate cancer [63], and these loci together explain approximately 14% and 30%, respectively, of the familial relative risk of breast cancer and prostate cancer among European descendants. For CRC, however, only 31 low-penetrance susceptibility loci have been identified, explaining approximately 9% of the familial relative risk of CRC among European descendants. Compared with GWAS of breast cancer and prostate cancer, studies conducted for CRC have been relatively small. In our study, we evaluated approximately 7,000 promising variants identified from GWAS in the replication stages, which represents one of the largest efforts made to date to follow-up genetic variants identified by GWAS. Six novel loci were identified, representing the largest number of loci identified for CRC risk in a single study. Although multiple GWAS with sample sizes larger than this study have been conducted among European descendants [13, 14, 16], we were still able to identify risk variants with relatively large effect sizes. Our study further highlights the value of conducting GWAS in non-European populations to discover novel susceptibility loci for CRC.

In summary, we have identified six new loci associated with CRC risk in this large GWAS conducted among East Asians. These new loci contain genes with established connections to colorectal tumorigenesis through major biological pathways such as Wnt and TGF-β signaling, as well as genes with important biological function that have not yet been well linked to CRC. Our study considerably expands our knowledge of the genetic landscape of CRC and provides clues for future studies to characterize the causal variants and functional mechanisms for these GWAS-identified loci.

ONLINE METHODS

Studies participants

This genome-wide association study (GWAS) was conducted as part of the Asia Colorectal Cancer Consortium, including a total of 14,963 colorectal cancer (CRC) cases and 31,945 controls of East Asian ancestry from 14 studies conducted in China, South Korea and Japan (Supplementary Table 1). Specifically, stage 1 (GWAS discovery) consisted of five studies: Shanghai CRC Study 1 (Shanghai-1, *n* = 3,102), Shanghai CRC Study 2 (Shanghai-2, *n* = 908), Guangzhou CRC Study 1 (Guangzhou-1, *n* = 1,603), Aichi CRC Study 1 (Aichi-1, *n* =1,346), and Korean Cancer Prevention Study-II CRC (KCPS-II, *n* = 1,301). With the exception of Shanghai-2 for which we added 423 controls from other studies [64, 65], samples for the remaining four studies were the same as we reported in our previous study [18]. Stage 2 consisted of three studies: Shanghai CRC Study 3 (Shanghai-3, *n* = 6,577), Guangzhou CRC Study 2 (Guangzhou-2, *n* = 809), and Guangzhou CRC Study 3 (Guangzhou-3, *n* = 2,408). Stage 3 included one study: the BioBank Japan CRC Study (BBJ, *n* = 14,172). Stage 4 consisted of five studies: Guangzhou CRC Study 4 (Guangzhou-4, *n* = 1,791), Aichi CRC Study 2 (Aichi-2, *n* = 708), Korean-National Cancer Center CRC Study (Korea-NCC, *n* = 2,721), Seoul CRC Study (Korea-Seoul, *n* = 1,522), and Hwasun Cancer Epidemiology Study-Colon and Rectum Cancer (HCES-CRC, *n* = 7,930). We estimated that our study had a statistical power of >80% to identify an association with an OR of 1.10 or above at $P < 5\times10^{-8}$ for SNPs with a MAF of as low as 0.30. We evaluated generalizability of the newly

identified associations with CRC risk in European descendants in three consortia including 23 studies (Supplementary Table 13) with a total sample size of 16,984 cases and 18,262 controls recruited in the United States, Europe, Canada and Australia: the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) [17], the Colorectal Transdisciplinary (CORECT) Study and the Colon Cancer Family Registry (CCFR) [21]. Summary descriptions of participating studies are presented in Supplementary Note. Study protocols were approved by the relevant review boards in respective institutions and informed consents were obtained from all study participants.

Laboratory procedures

Genotyping of samples in stage 1 was conducted as described previously using the following platforms: Affymetrix Genome-Wide Human SNP Array 6.0, Illumina HumanOmniExpress BeadChip, Illumina Infinium HumanHap550 BeadChip, Illumina 660W-Quad BeadChip, Illumina Human610-Quad BeadChip, Illumina Infinium HumanHap610 BeadChip, and Affymetrix Genome-Wide Human SNP Array 5.0 [18, 64–69]. We used a uniform quality control protocol as described in our recent paper [18] to filter samples and SNPs. Genotyping and quality control methods are also presented in the Supplementary Note. After quality control exclusions, we obtained 502,145 autosomal SNPs for samples in Shanghai-1, 245,961 SNPs in Shanghai-2, 250,612 SNPs in Guangzhou-1, 232,426 SNPs in Aichi-1, and 312,869 SNPs in KCPS-II (Supplementary Table 2).

Genotyping for 3,632 cases and 6,404 controls in stage 2 was completed using Illumina Infinium assays as part of the customer add-on content for multiple projects to the Illumina HumanExome Beadchip (see URLs). Details of array design, genotyping, genotype call and quality control are provided in the Supplementary Note. Samples were excluded according to the following criteria: (i) genotype call rate <98%, (ii) genetically identical or duplicated samples, (iii) sex determined using genetic data inconsistent with epidemiological or clinical data, (iv) first or second degree relatives, (v) ethnic outliers, or (vi) heterozygosity outliers. Genetic markers were excluded using the following criteria: (i) MAF = 0, (ii) genotype call rate <98%, (iii) consistency rate <98% in positive quality control samples, (iv) *P* for Hardy-Weinberg equilibrium (HWE) $<10^{-5}$ in controls or (v) caution SNPs revealed by the Exome Chip Design group (see URLs). We obtained a final dataset including 6,899 SNPs genotyped on 3,519 cases and 6,275 controls for this project.

Cases and controls in stage 3 were genotyped using the Illumina HumanHap610-Quad BeadChip. Quality control filters were based on criteria described previously [20]. Methods of genotyping and quality control procedures are also presented in the Supplementary Note. After sample and SNP exclusions, we generated a dataset including 2,814 cases and 11,358 controls with 460,463 SNPs.

Stage 4 genotyping for 29 SNPs was conducted using the iPLEX Sequenom MassARRAY platform according to manufacturer's protocols at the Vanderbilt Molecular Epidemiology Laboratory (Nashville, Tennessee, United States). Details of genotyping and quality control are provided in the Supplementary Note. We filtered out SNPs with (i) genotype call rate <95%, (ii) genotyping consistency rate <95% in positive control samples, (iii) an unclear genotype call or (iv) *P* for HWE $<10^{-5}$ in controls. The average consistency rate of these

SNPs passing quality control filters was 99.9% with median value 100% in each of the five participating studies included in this stage.

Samples in GECCO, CORECT and CCFR were genotyped with Illumina and Affymetrix Arrays[17, 21]. Genotyping, quality control and imputation have been reported in previously [17, 21] and are described in the Supplementary Note.

SNP selection

Selection of SNPs for stage 2 replication was primarily based on the following criteria: (i) P <0.05 in meta-analysis, (ii) P for heterogeneity >0.0001, (iii) imputation R^2 >0.5 in at each of the included studies, (iv) MAF >0.05 in each of the included studies, (v) SNPs uncorrelated with established CRC SNPs (defined as r^2 <0.2 in HapMap Asian), (vi) SNPs uncorrelated with other SNPs identified in this project (r^2 <0.2) and (vii) data available in at least two studies (see Supplementary Note). We included multiple SNPs in some regions with a prior P value of <0.002 or with genes of interest. Risk variants identified from previously published GWAS were also included in the assay [7–20]. In total, 8,569 unique SNPs were selected. Of them, 7,113 SNPs were successfully designed. For stage 3 replication, we selected 559 SNPs according to criteria: (i) P <0.005 in meta-analysis of data from stages 1 and 2, (ii) association in the same direction in both stages and (iii) P for heterogeneity >0.0001. For stage 4, we selected 30 SNPs on the basis of criteria: (i) P <0.0001 in meta-analysis of stages 1, 2, and 3, (ii) P <0.01 in meta-analysis of stages 2 and 3, (iii) association in the same direction in three stages and (iv) P for heterogeneity >0.0001.

Statistical and bioinformatic analysis

Details of imputation and population substructure evaluation are provided in the Supplementary Note. Briefly, stage 1 imputation was performed with CHB (Han Chinese in Beijing, China) and JPT (Japanese in Tokyo, Japan) HapMap 2 panel as the reference using program MACH v1.0 [70] (see URLs). Stage 3 imputation was conducted with phased data of JPT/CHS/CHD participants from the 1000 Genomes Project phase1 v3 as the reference using program MACH v1.0 [70] and minimac [71] (see URLs). Regional imputation of genotype data from The Cancer Genome Atlas (TCGA) [30] (see URLs) was performed with the GIANT ALL reference panel from the 1000 Genomes Project phase1 release v3 using MACH v1.0 [70] and minimac [71] (see URLs). To evaluate the imputation quality in our study, we directly genotyped the ten newly identified risk variants in approximately 2,800 samples included in stage 1. The concordance between imputed and genotyped data was very high, with mean values ranging from 96.00% to 99.96% for the ten SNPs (Supplementary Table 20). For rs10849432, the imputation quality for the Aichi-1 study was relatively low (R^2 =0.57), and thus data from this study were not included in our final analysis. We evaluated population structure in studies included in stages 1 and 2 using principal components analysis with EIGENSTRAT software [72] (see URLs). Based on adjusted regression models including the first ten principal components, the genomic inflation factor λ was <1.04 in each of the five studies included in stage 1 and 1.0368 in the meta-analysis of all five studies (Supplementary Fig. 2). The λ was <1.05 in each of the three studies included in stage 2 and 1.0525 in the meta-analysis of all three studies (Supplementary Fig. 3). A rescaled inflation statistic λ_{1000}, representing an equivalent value of a study with 1,000 cases and 1,000

controls using the formula: $\lambda_{1000} = 1 + 500 \times (\lambda - 1) \times (1/N_{cases} + 1/N_{controls})$ 73, was 1.01 in both stages 1 and 2. These findings showed little evidence of population stratification in our studies.

Associations between SNPs and CRC risk were evaluated on the basis of the log-additive model using mach2dat [70], PLINK version 1.0.7 [74], R version 3.0.0 and SAS version 9.3 (for all of these see URLs). Per-allele odds ratios (ORs) and 95% confidence intervals (CIs) were derived from logistic regression models, adjusting for age, sex and the first ten principal components when appropriate. Association analysis was conducted for each participating study separately and a fixed-effects meta-analysis was conducted to obtain summary results for each of the four stages and all stages combined with the inverse-variance method using program METAL [75]. SNPs showing an association at $P < 5\times10^{-8}$ in the combined analysis of all studies were considered genome-wide significant. We also performed stratified analyses for the top SNPs by tumor anatomic site (colon and rectum), population (Chinese, Korean and Japanese) and sex (men and women). We estimated heterogeneity across studies and subgroups with a Cochran's *Q* test [76], with *P* for heterogeneity <0.008 as statistically significant considering multiple comparisons of six independent loci. Independent signals in a locus were identified using stepwise logistic regression models conditioning on the top risk variant we identified in each of the new loci using R software (see URLs). We estimated haplotype frequencies using Haploview version 4.2 [77] (see URLs) and conducted haplotype association analysis for two loci (11q12.2 and 19q13.2) where two or more SNPs were identified using SAS Genetics v9.3 with logistic regression models. Pairwise SNP-SNP interactions between six top risk variants in the newly identified loci with $P < 5\times10^{-8}$ and also between these six SNPs and the risk variants in 25 previously reported loci were evaluated using the maximal likelihood ratio test with inclusion of interaction terms into logistic regression models. Interactions with *P* <0.00028 were considered statistically significant with the adjustment of multiple comparisons of 180 tests.

The familial relative risk (λ) to offspring of an affected individual due to a single locus was estimated using a log-additive model: $\lambda=(pr^2+q) / (pr+q)^2$, where *p* is the frequency of the risk allele, $q = 1-p$ is the frequency of the reference allele, and *r* is the per-allele relative risk [78]. The proportion of the familial relative risk explained by this locus, assuming a multiplicative interaction between markers in the locus and other loci, was calculated as: $\log(\lambda)/ \log (\lambda_o)$, where λ_o is the overall familial relative risk. λ_o is assigned to be 2.2 for CRC estimated from a meta-analysis [79]. Assuming that the risk associated with each locus combine multiplicatively, the familial relative risks also multiply. Then the combined contribution of the familial relative risks from multiple loci is equal to: $\ln(\Pi_i\lambda_i)/\ln(\lambda_0)$.

We generated forest plots and Q-Q plots using R software (see URLs). Regional association plots for SNPs in newly identified loci were generated using the website-based tool LocusZoom version 1.1 [80] (see URLs). Linkage equilibrium (LD) structure between SNPs was determined on the basis of data from the 1000 Genomes Project Pilot 1 or HapMap 2 as provided by the website-based tool SNAP [81] (see URLs) and plotted using Haploview, SNAP and the UCSC Genome Browser (see URLs). LD blocks were defined using HapMap recombination rates and hotspots [23]. All the genomic coordinates are based on the National Center for Biotechnology Information (NCBI), Build 36.

To identify putative functional variants for newly identified loci, we identified all SNPs in LD (i.e., $r^2 > 0.5$) with the risk variants using data from the 1000 Genomes Project [22] and HapMap 2 [23]. We mapped the genomic locations of these SNPs to nonsynonymous sites, splice sites, promoters, nearGene-3 regions, nearGene-5 regions, three prime untranslated regions (3′-UTR), five prime untranslated regions (5′-UTR), introns and intergenic regions. We evaluated the potential functional effect of nonsynonymous SNPs using the prediction algorithms SIFT [36] and PolyPhen-2 [37] (see URLs). We predicted the putative function of SNPs in promoters, nearGene-3 regions, nearGene-5 regions, 3′-UTR and 5′-UTR with SNPinfo Web Server [45] (see URLs). We conducted analyses to evaluate the potential regulatory effect of SNPs in non-coding regions on transcription using the Encyclopedia of DNA Elements (ENCODE) tool HaploReg v2 [82] and the UCSC Genome Browser (see URLs) on the basis of their location within regions of promoter or enhancer activity, DNase I hypersensitivity; local histone modifications, proteins bound to these regulatory sites, *cis* expression quantitative trait loci (eQTL) and transcription factor binding motif. We obtained additional functional evidence for these SNPs from the published literature.

We identified all genes that localize in a 1-mb window centered on the top risk variants in our newly identified loci and including SNPs correlated ($r^2 > 0.5$) with the top risk variants. To determine whether these genes may explain the observed association in these loci, we first examined genome-wide *cis* eQTL data in multiple tissues from four major eQTL databases: the Blood eQTL browser [25], the eQTL Browser [26], the Genotype-Tissue Expression project (GTEx) [27] and the Multiple Tissue Human Expression Resource project (MuTHER) [28]. The significance threshold for these analyses was set to $P < 0.008$ to count for six tests. Somatic mutations of these genes were evaluated using data from the Catalogue of Somatic Mutations in Cancer (COSMIC) [29] (see URLs). Expression levels of these genes in CRC cell lines were assessed using data from Gene Expression Atlas [31] (see URLs). To correct for multiple comparisons of the 11 key genes, associations with a $P < 0.0045$ were considered to be statistically significant. We searched the published literature for these genes in relation to CRC from PubMed and OMIM (see URLs).

Expression analysis

We downloaded RNA sequencing (level 1) and SNP array (level 2) data for 364 colon adenocarcinoma and 18 normal colon tissue samples from TCGA [30] (see URLs). To quantify expression levels of candidate genes in the newly identified loci, we normalized gene expression levels using the reads per kilobase of exon per million mapped reads (RPKM) value as previously described [83]. Expression differences between tumor and normal samples for each gene were evaluated on the basis of the RPKM values with the Wilcoxon rank sum test. Associations between gene RPKM value and SNP genotypes were analyzed using a linear regression model including age and sex as covariates. We converted the RPKM value of a gene to log scale for analysis if it was not normally distributed. We considered $P < 0.0045$ to be statistically significant with adjustment for testing of the 11 key genes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are solely responsible for the scientific content of this paper. The sponsors of this study had no role in study design, data collection, analysis, interpretation, writing of the report or the decision for submission. We thank all study participants and research staff of all parent studies for their contributions and commitment to this project, Regina Courtney for DNA preparation, Jing He for data processing and analyses, Xiangyi Guo for suggestions in bioinformatic analysis and Mary Jo Daly and Bethanie J Rammer for editing and preparing the manuscript. The work at the Vanderbilt University School of Medicine was supported by U.S. National Institutes of Health grants R37CA070867, R01CA082729, R01CA124558, R01CA148667 and R01CA122364, as well as Ingram Professorship and Research Reward funds from the Vanderbilt University School of Medicine. Studies (grant support) participating in the Asia Colorectal Cancer Consortium include: Shanghai Women's Health Study (R37CA070867), Shanghai Men's Health Study (R01CA082729), Shanghai Breast and Endometrial Cancer Studies (R01CA064277 and R01CA092585, contributing only controls), Shanghai Colorectal Cancer Study 3 (R37CA070867 and Ingram Professorship funds), Guangzhou Colorectal Cancer Study (National Key Scientific and Technological Project – 2011ZX09307-001-04; the National Basic Research Program – 2011CB504303, contributing only controls; Natural Science Foundation of China – 81072383, contributing only controls), Japan-BioBank Colorectal Cancer Study (grant from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese government), Hwasun Cancer Epidemiology Study-Colon and Rectum Cancer (grants from the Korea Center for Disease Control and Prevention and Jeonnam Regional Cancer Center), Aichi Colorectal Cancer Study (Grant-in-aid for Cancer Research, Grant for the Third Term Comprehensive Control Research for Cancer and Grants-in-Aid for Scientific Research from the Japanese Ministry of Education, Culture, Sports, Science and Technology, Nos. 17015018 and 221S0001), Korea-NCC Colorectal Cancer Study (Basic Science Research Program through the National Research Foundation of Korea, 2010-0010276 and National Cancer Center Korea, 0910220), Korea-Seoul Colorectal Cancer Study (none reported), and KCPS-II Colorectal Cancer Study (National R&D Program for cancer control, 1220180; Seoul R&D Program, 10526).

We wish to thank all participants, staff and investigators of GECCO, CORECT and CCFR for making it possible to present results for populations of European ancestry for the new CRC loci identified among East Asians. Institutions and location (investigators) from GECCO, CORECT and CCFR who provided support to this project include (in alphabetical order): Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, Ohio, USA (Li Li); Centre for Public Health Research, Massey University, Palmerston North, New Zealand (John D. Potter); Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA (Andrew T. Chan, Charles S. Fuchs, Edward L. Giovannucci, Jing Ma); Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, ON, Canada (Brent W. Zanke); Department of Biostatistics, University of Washington, Seattle, WA, USA (Li Hsu); Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Yeshiva University, Bronx, NY, USA (Thomas Rohan); Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA (Edward L. Giovannucci, David J. Hunter); Department of Epidemiology, University of Washington School of Public Health, Seattle, WA, USA (Polly A. Newcomb, Ulrike Peters, John D. Potter, Emily White); Department of Internal Medicine, University of Utah Health Sciences Center, Salt Lake City, UT, USA (Martha L. Slattery); Department of Medical Biophysics and Department Molecular Genetics, University of Toronto, Toronto, ON, Canada (Thomas J. Hudson); Department of Medical Oncology, Dana Farber Cancer Institute, Boston, MA, USA (Charles S. Fuchs); Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, PA, USA (Robert E. Schoen); Department of Nutrition, Harvard School of Public Health, Boston, MA, USA (Edward L. Giovannucci); Department of Preventive Medicine, University of Southern California, Los Angeles, CA, USA (Graham Casey, David V. Conti, Robert W. Haile, Fredrick R. Schumacher); Department of Surgery, Mount Sinai Hospital, Toronto, ON, Canada (Steven Gallinger); Departments of Laboratory Medicine and Pathology and Laboratory Genetics, Mayo Clinic, Rochester, MN, USA (Noralane M. Lindor, Stephen N. Thibodeau); Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA (Carolyn M. Hutter, Daniela Seminara); Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA (Sonja I. Berndt, Stephen J. Chanock); Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany (Jenny Chang-Claude); Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg, Germany (Hermann Brenner, Michael Hoffmeister); Division of Endocrinology, Diabetes and Metabolism, Ohio State University, Columbus, OH, USA (Rebecca D. Jackson); Division of Epidemiology, Department of Environmental Medicine, New York University School of Medicine, New York, NY, USA (Richard B. Hayes); Division of Gastroenterology and Hepatology, UNC School of Medicine, Chapel Hill, NC, USA (John A. Baron); Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA (Andrew T. Chan); Division of Research, Kaiser Permanente Medical Care Program, Oakland, CA, USA (Bette J. Caan); Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA (Loic Le Marchand); Epidemiology Research Program, American

Cancer Society, Atlanta, GA, USA (Peter T. Campbell; Eric J. Jacobs); Faculty of Medicine, Memorial University of Newfoundland, St. John's, NL, Canada (Roger C. Green); Institut d'Investigació Biomèdica de Bellvitge, Institut Catala d'Oncologia, Hospitalet, Barcelona, Spain (Victor Moreno); Melborne School of Population Health, The University of Melbourne, Melbourne, VIC, Australia (John L. Hopper, Mark A. Jenkins, Gianluca Severi); Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA (Graham Casey (PI of CCFR, co-PI of CORECT), David V. Conti, Christopher K. Edlund, Jane Figueiredo, Fredrick R. Schumacher, W. James Gauderman, Stephen B. Gruber (PI of CORECT), Leon Raskin, Stephanie L. Stenzel, Duncan C. Thomas (co-PI of CORECT)); Ontario Institute for Cancer Research, Toronto, ON, Canada (Mathieu Lemire); Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA (Kendra Blalock, Christopher S. Carlson, Keith Curtis, Jian Gong, Tabitha A. Harrison, Li Hsu, Shuo Jiao, Yi Lin, Polly A. Newcomb, Ulrike Peters (PI of GECCO, co-PI of CORECT), John D. Potter, Conghui Qu, Emily White); Samuel Lunenfeld Research Institute, Toronto, ON, Canada (Steven Gallinger); School of Public Health, University of Washington, Seattle, WA, USA (Christopher S. Carlson); Service de Génétique Médicale, CHU Nantes, Nantes, France (Stéphane Bezieau, Sébastien Küry); Translational Genomics Research Institute, Phoenix, Arizona, USA (David Duggan); University of Michigan Comprehensive Cancer Center, University of Michigan, Ann Arbor, MI, USA (John F. Harju, Frank J. Manion, Bhramar Mukherjee). Additionally, we also thank Bruno Buecher of ASTERISK; Ute Handte-Daub, Muhabbet Celik, Renate Hettler-Jensen, Utz Benscheid and Ursula Eilber of DACHS; Patrice Soule, Hardeep Ranu, Immaculata Devivo, David Hunter, Qin Guo, Lixue Zhu and Haiyan Zhang of HPFS, NHS and PHS, 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; Christine Berg and Philip Prorok of PLCO; Tom Riley of Information Management Services Inc.; Barbara O'Brien of Westat, Inc.; Bill Kopp and Wen Shao of SAIC-Frederick; WHI investigators (see https://cleo.whi.org/researchers/SitePages/Write%20a%20Paper.aspx), and the GECCO Coordinating Center. Participating studies (grant support) in the GECCO, CORECT and CCFR GWAS meta-analysis are: GECCO (U01 CA137088 and R01 CA059045), DALS (R01 CA048998), DACHS (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, CH 117/1-1, 01KH0404 and 01ER0814), HPFS (P01 CA055075, UM1 CA167552, R01 137178 and P50 CA127003), NHS (R01 137178, P50 CA127003 and P01 CA087969), OFCCR (U01 CA074783), PMH (R01 CA076366), PHS (R01 CA042182), VITAL (K05 CA154337), WHI (HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, HHSN271201100004C and 268200764316C), PLCO (Z01 CP 010200, U01 HG004446 and U01 HG 004438). CORECT is supported by the National Cancer Institute as part of the GAME-ON consortium (U19 CA148107) with additional support from NCI grants (R01 CA81488, P30 CA014089) and the National Human Genome Research Institute at the National Institutes of Health (T32 HG000040), the National Institute of Environmental Health Sciences at the National Institutes of Health (T32 ES013678). CCFR is supported by the National Cancer Institute, National Institutes of Health under RFA #CA-95-011 and through cooperative agreements with members of the Colon Cancer Family Registry and PIs of the Australasian Colorectal Cancer Family Registry (U01 CA097735), Familial Colorectal Neoplasia Collaborative Group (U01 CA074799) [USC], Mayo Clinic Cooperative Family Registry for Colon Cancer Studies (U01 CA074800), Ontario Registry for Studies of Familial Colorectal Cancer (U01 CA074783), Seattle Colorectal Cancer Family Registry (U01 CA074794) and University of Hawaii Colorectal Cancer Family Registry (U01 CA074806). The GWAS work was supported by a National Cancer Institute grant (U01CA122839). OFCCR was supported by a GL2 grant from the Ontario Research Fund, the Canadian Institutes of Health Research and the Cancer Risk Evaluation (CaRE) Program grant from the Canadian Cancer Society Research Institute. Thomas Hudson and Brent Zanke are recipients of Senior Investigator Awards from the Ontario Institute for Cancer Research, through support from the Ontario Ministry of Economic Development and Innovation. ASTERISK was funded by a Regional Hospital Clinical Research Program (PHRC) and supported by Regional Council of Pays de la Loire, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (GEFLUC), the Association Anne de Bretagne Génétique and the Ligue Régionale Contre le Cancer (LRCC). PLCO datasets were accessed with approval through dbGaP (Cancer Genetic Markers of Susceptibility (CGEMS) prostate cancer scan, phs000207.v1.p1; CGEMS pancreatic cancer scan, phs000206.v3.p2; and GWAS of Lung Cancer and Smoking, phs000093.v2.p2, which was funded by Z01 CP 010200, U01 HG004446, and U01 HG 004438).

URLs

BioBank Japan: http://biobankjp.org/; Blood eQTL browser, http://genenetwork.nl/bloodeqtlbrowser/; CGEMS, http://cgems.cancer.gov/; COSMIC, http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/; dbGaP, http://www.ncbi.nlm.nih.gov/gap; EIGENSTRAT, http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm; eQTL Browser in University of Chicago, http://eqtl.uchicago.edu/Home.html; GTEx eQTL Browser, http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi/ http://www.broadinstitute.org/gtex/; Gene Expression Atlas, http://www.ebi.ac.uk/gxa/; Haploview, http://www.broad.mit.edu/mpg/haploview/; HaploReg v2, http://www.broadinstitute.org/mammals/haploreg/haploreg.php;

HapMap project, http://hapmap.ncbi.nlm.nih.gov/; IKMC, http://www.knockoutmouse.org/; LocusZoom, http://csg.sph.umich.edu/locuszoom/; Illumina HumanExome-12v1_A Beadchip http://genome.sph.umich.edu/wiki/Exome_Chip_Design; IPA, http://www.ingenuity.com/; MACH 1.0, http://www.sph.umich.edu/csg/abecasis/MACH/; mach2dat, http://genome.sph.umich.edu/wiki/Mach2dat:_Association_with_MACH_output; Minimac, http://genome.sph.umich.edu/wiki/Minimac; METAL, http://www.sph.umich.edu/csg/abecasis/Metal/; MuTHER, http://www.muther.ac.uk/; OMIM, http://www.ncbi.nlm.nih.gov/omim/; PLINK version 1.07, http://pngu.mgh.harvard.edu/~purcell/plink/; PolyPhen-2, http://genetics.bwh.harvard.edu/pph2/; PubMed, http://www.ncbi.nlm.nih.gov/pubmed/; 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/; SIFT, http://sift.jcvi.org/; The 1000 Genomes Browser, http://browser.1000genomes.org/index.html; The Cancer Genome Atlas, http://cancergenome.nih.gov/; TRANSFAC, http://www.gene-regulation.com/pub/databases.html; UCSC Genome Browser, http://genome.ucsc.edu/.

References

1. Jemal A, et al. Global cancer statistics. CA Cancer J Clin. 2011; 61:69–90. [PubMed: 21296855]
2. 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. 2000; 343:78–85. [PubMed: 10891514]
3. de la Chapelle A. Genetic predisposition to colorectal cancer. Nat Rev Cancer. 2004; 4:769–780. [PubMed: 15510158]
4. Aaltonen L, Johns L, Jarvinen H, Mecklin JP, Houlston R. Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. Clin Cancer Res. 2007; 13:356–361. [PubMed: 17200375]
5. Ma X, Zhang B, Zheng W. Genetic variants associated with colorectal cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. Gut. 2014; 63:326–336. [PubMed: 23946381]
6. Palles C, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat Genet. 2013; 45:136–144. [PubMed: 23263490]
7. 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]
8. 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]
9. 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]
10. 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]
11. 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]
12. 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]
13. 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]
14. 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]

15. Tomlinson IP, 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. 2011; 7:e1002105. [PubMed: 21655089]
16. Dunlop MG, et al. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. Nat Genet. 2012; 44:770–776. [PubMed: 22634755]
17. Peters U, et al. Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. Gastroenterology. 2013; 144:799–807. [PubMed: 23266556]
18. Jia WH, et al. Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. Nat Genet. 2013; 45:191–196. [PubMed: 23263487]
19. Zhang B, et al. Genome-wide association study identifies a new SMAD7 risk variant associated with colorectal cancer risk in East Asians. Int J Cancer. 2014
20. Cui R, et al. Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. Gut. 2011; 60:799–805. [PubMed: 21242260]
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. Abecasis GR, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]
23. Frazer KA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851–861. [PubMed: 17943122]
24. Bernstein BE, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
25. Westra HJ, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013; 45:1238–1243. [PubMed: 24013639]
26. Degner JF, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012; 482:390–394. [PubMed: 22307276]
27. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013; 45:580–585. [PubMed: 23715323]
28. Grundberg E, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012; 44:1084–1089. [PubMed: 22941192]
29. Forbes SA, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Res. 2011; 39:D945–D950. [PubMed: 20952405]
30. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012; 487:330–337. [PubMed: 22810696]
31. Kapushesky M, et al. Gene Expression Atlas update--a value-added database of microarray and sequencing-based functional genomics experiments. Nucleic Acids Res. 2012; 40:D1077–D1081. [PubMed: 22064864]
32. Tang W, et al. A genome-wide RNAi screen for Wnt/beta-catenin pathway components identifies unexpected roles for TCF transcription factors in cancer. Proc Natl Acad Sci U S A. 2008; 105:9697–9702. [PubMed: 18621708]
33. Angus-Hill ML, Elbert KM, Hidalgo J, Capecchi MR. T-cell factor 4 functions as a tumor suppressor whose disruption modulates colon cell proliferation and tumorigenesis. Proc Natl Acad Sci U S A. 2011; 108:4914–4919. [PubMed: 21383188]
34. Bass AJ, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat Genet. 2011; 43:964–968. [PubMed: 21892161]
35. Grainger DJ, et al. Genetic control of the circulating concentration of transforming growth factor type beta1. Hum Mol Genet. 1999; 8:93–97. [PubMed: 9887336]
36. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4:1073–1081. [PubMed: 19561590]
37. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]

38. Dunning AM, et al. A transforming growth factorbeta1 signal peptide variant increases secretion in vitro and is associated with increased incidence of invasive breast cancer. Cancer Res. 2003; 63:2610–2615. [PubMed: 12750287]
39. Suthanthiran M, et al. Transforming growth factor-beta 1 hyperexpression in African-American hypertensives: A novel mediator of hypertension and/or target organ damage. Proc Natl Acad Sci U S A. 2000; 97:3479–3484. [PubMed: 10725360]
40. Yamada Y, et al. Association of a polymorphism of the transforming growth factor-beta1 gene with genetic susceptibility to osteoporosis in postmenopausal Japanese women. J Bone Miner Res. 1998; 13:1569–1576. [PubMed: 9783545]
41. Markowitz SD, Bertagnolli MM. Molecular origins of cancer: Molecular basis of colorectal cancer. N Engl J Med. 2009; 361:2449–2460. [PubMed: 20018966]
42. Howe JR, et al. Mutations in the SMAD4/DPC4 gene in juvenile polyposis. Science. 1998; 280:1086–1088. [PubMed: 9582123]
43. Valle L, et al. Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. Science. 2008; 321:1361–1365. [PubMed: 18703712]
44. Liu L, et al. Functional FEN1 genetic variants contribute to risk of hepatocellular carcinoma, esophageal cancer, gastric cancer and colorectal cancer. Carcinogenesis. 2012; 33:119–123. [PubMed: 22072618]
45. Xu Z, Taylor JA. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. Nucleic Acids Res. 2009; 37:W600–W605. [PubMed: 19417063]
46. Zheng L, et al. Functional regulation of FEN1 nuclease and its link to cancer. Nucleic Acids Res. 2011; 39:781–794. [PubMed: 20929870]
47. Zheng L, et al. Fen1 mutations result in autoimmunity, chronic inflammation and cancers. Nat Med. 2007; 13:812–819. [PubMed: 17589521]
48. Kucherlapati M, et al. Haploinsufficiency of Flap endonuclease (Fen1) leads to rapid tumor progression. Proc Natl Acad Sci U S A. 2002; 99:9924–9929. [PubMed: 12119409]
49. Schaeffer L, et al. Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids. Hum Mol Genet. 2006; 15:1745–1756. [PubMed: 16670158]
50. Castellone MD, Teramoto H, Williams BO, Druey KM, Gutkind JS. Prostaglandin E2 promotes colon cancer cell growth through a Gs-axin-beta-catenin signaling axis. Science. 2005; 310:1504–1510. [PubMed: 16293724]
51. Cai Q, et al. Prospective study of urinary prostaglandin E2 metabolite and colorectal cancer risk. J Clin Oncol. 2006; 24:5010–5016. [PubMed: 17075120]
52. Rogers LM, Riordan JD, Swick BL, Meyerholz DK, Dupuy AJ. Ectopic expression of Zmiz1 induces cutaneous squamous cell malignancies in a mouse model of cancer. J Invest Dermatol. 2013; 133:1863–1869. [PubMed: 23426136]
53. Turnbull C, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet. 2010; 42:504–507. [PubMed: 20453838]
54. Ovalle S, et al. The tetraspanin CD9 inhibits the proliferation and tumorigenicity of human colon carcinoma cells. Int J Cancer. 2007; 121:2140–2152. [PubMed: 17582603]
55. Mori M, et al. Motility related protein 1 (MRP1/CD9) expression in colon cancer. Clin Cancer Res. 1998; 4:1507–1510. [PubMed: 9626469]
56. Lee JH, et al. Glycoprotein 90K, downregulated in advanced colorectal cancer tissues, interacts with CD9/CD82 and suppresses the Wnt/beta-catenin signal via ISGylation of beta-catenin. Gut. 2010; 59:907–917. [PubMed: 20581239]
57. Braumuller H, et al. T-helper-1-cell cytokines drive cancer into senescence. Nature. 2013; 494:361–365. [PubMed: 23376950]
58. Wolf MJ, Seleznik GM, Zeller N, Heikenwalder M. The unexpected role of lymphotoxin beta receptor signaling in carcinogenesis: from lymphoid tissue formation to liver and prostate cancer development. Oncogene. 2010; 29:5006–5018. [PubMed: 20603617]
59. Lukashev M, et al. Targeting the lymphotoxin-beta receptor with agonist antibodies as a potential cancer therapy. Cancer Res. 2006; 66:9617–9624. [PubMed: 17018619]

60. Funato Y, Miki H. Nucleoredoxin, a novel thioredoxin family member involved in cell growth and differentiation. Antioxid Redox Signal. 2007; 9:1035–1057. [PubMed: 17567240]
61. Funato Y, Michiue T, Asashima M, Miki H. The thioredoxin-related redox-regulating protein nucleoredoxin inhibits Wnt-beta-catenin signalling through dishevelled. Nat Cell Biol. 2006; 8:501–508. [PubMed: 16604061]
62. Michailidou K, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet. 2013; 45:353–2. [PubMed: 23535729]
63. Eeles RA, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet. 2013; 45:385–2. [PubMed: 23535732]
64. Abnet CC, et al. A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. Nat Genet. 2010; 42:764–767. [PubMed: 20729852]
65. Amundadottir L, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. Nat Genet. 2009; 41:986–990. [PubMed: 19648918]
66. 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]
67. 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]
68. Jee SH, et al. Adiponectin concentrations: a genome-wide association study. Am J Hum Genet. 2010; 87:545–552. [PubMed: 20887962]
69. 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]
70. 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]
71. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012; 44:955–959. [PubMed: 22820512]
72. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]
73. Freedman ML, et al. Assessing the impact of population stratification on genetic association studies. Nat Genet. 2004; 36:388–393. [PubMed: 15052270]
74. 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]
75. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010; 26:2190–2191. [PubMed: 20616382]
76. Lau J, Ioannidis JP, Schmid CH. Quantitative synthesis in systematic reviews. Ann Intern Med. 1997; 127:820–826. [PubMed: 9382404]
77. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005; 21:263–265. [PubMed: 15297300]
78. Zheng W, et al. Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls. Hum Mol Genet. 2013; 22:2539–2550. [PubMed: 23535825]
79. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol. 2001; 96:2992–3003. [PubMed: 11693338]
80. Pruim RJ, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010; 26:2336–2337. [PubMed: 20634204]
81. Johnson AD, et al. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008; 24:2938–2939. [PubMed: 18974171]
82. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012; 40:D930–D934. [PubMed: 22064851]

83. Yan G, et al. Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. Nat Biotechnol. 2011; 29:1019–1023. [PubMed: 22002653]

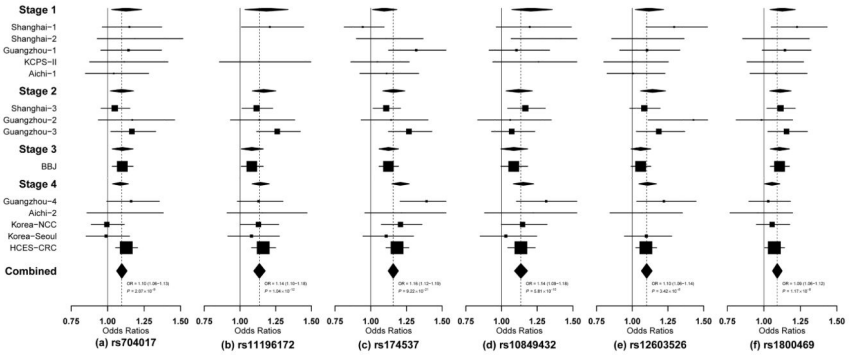

Figure 1. Forest plots for risk variants in the six newly identified loci

The six plots represent (a) rs704017, (b) rs11196172, (c) rs174537, (d) rs10849432, (e) rs12603526 and (f) rs1800469. Per-allele ORs are presented, with the area of each box proportional to the inverse variance weight of the estimate. Horizontal lines represent 95% CIs. Diamonds represent summary ORs generated under a fixed-effects meta-analysis; width of the diamonds corresponds to the 95% CIs. Unbroken vertical lines represent the null value; broken vertical lines represent the summary ORs for all studies for each SNP.

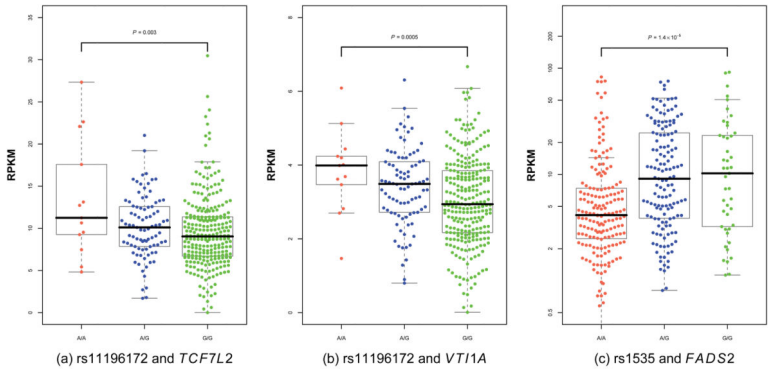

Figure 2. Association of selected risk variants identified in this study with gene expression in colon tumor tissue

The three plots are for (a) rs11196172 and *TCF7L2*, (b) rs11196172 and *VTI1A*, and (c) rs1535 and *FADS2*. Gene expression levels are represented by per kilobase of exon per million mapped reads (RPKM) value based on the three genotypes of each SNP in red, blue and green. The median RPKM values and the interquartile range for each SNP are presented in the overlaid box plot. In (a) and (b), RPKM values are shown at normal scale, whereas RPKM values in (c) are shown with a log scale due to departure from normal distribution.

Table 1

Summary results for risk variants in the six newly identified loci associated with CRC in East Asians

							Stage 1	Stage 2	Stage 3	Stage 4	Stages 1 to 4	
Locus	**SNP**	**Gene**[a]	**Annotation**	**Position**[b]	**Alleles**[c]	**RAF**[d]	***P***	***P***	***P***	***P***	**OR (95% CI)**[e]	***P***[e]
10q22.3	rs704017	*ZMIZ1-AS1*	Intron 3	80489138	**G**/A	0.32	0.01	0.01	0.004	9.99×10^{-4}	1.10 (1.06–1.13)	2.07×10^{-8}
10q25.2	rs11196172	*TCF7L2*	Intron 4	114716833	**A**/G	0.68	0.03	1.82×10^{-5}	0.03	5.18×10^{-7}	1.14 (1.10–1.18)	1.04×10^{-12}
11q12.2	rs174537	*MYRF*	Intron 24	61309256	**G**/T	0.59	0.02	1.33×10^{-5}	1.61×10^{-4}	1.60×10^{-13}	1.16 (1.12–1.19)	9.22×10^{-21}
	rs4246215	*FEN1*	3′-UTR	61320875	**G**/T	0.59	0.02	2.29×10^{-6}	1.83×10^{-4}	1.25×10^{-11}	1.15 (1.12–1.19)	7.65×10^{-20}
	rs174550	*FADS1*	Intron 7	61328054	**T**/C	0.59	0.01	5.71×10^{-6}	1.83×10^{-4}	2.70×10^{-11}	1.15 (1.12–1.19)	1.58×10^{-19}
	rs1535	*FADS2*	Intron 1	61354548	**A**/G	0.59	0.02	7.55×10^{-6}	1.24×10^{-4}	1.20×10^{-11}	1.15 (1.12–1.19)	8.21×10^{-20}
12p13.31	rs10849432	*CD9*	Intergenic	6255988	**T**/C	0.82	0.002	0.007	0.06	6.95×10^{-6}	1.14 (1.09–1.18)	5.81×10^{-10}
17p13.3	rs12603526	*NXN*	Intron 1	747343	**C**/T	0.30	0.02	6.86×10^{-4}	0.08	3.80×10^{-4}	1.10 (1.06–1.14)	3.42×10^{-8}
19q13.2	rs1800469	*TGFB1*	Promoter	46552136	**G**/A	0.48	0.002	0.002	6.74×10^{-4}	0.03	1.09 (1.06–1.12)	1.17×10^{-8}
	rs2241714	*B9D2*	Exon 1	46561232	**C**/T	0.48	0.003	0.002	0.001	0.02	1.09 (1.06–1.12)	1.36×10^{-8}

Abbreviations: RAF, risk allele frequency; OR, odds ratio; CI, confidence interval.

[a] The closest gene(s).

[b] The chromosome position (bp) is based on the National Center for Biotechnology Information (NCBI) database, build 36.

[c] Risk/reference alleles are based on forward allele coding in NCBI, build 36. OR was estimated based on the risk allele (bold).

[d] RAF in controls from all stages combined.

[e] Summary OR (95% CI) and *P* value were obtained from a fixed-effects meta-analysis.

Table 2

Associations of risk variants in the six newly identified loci with CRC in European descendants

Locus	SNP (alleles)[a]	Gene[b]	Position[c]	Cases/controls	RAF[d]	OR (95% CI)[e]	P^e	$P_{heterogeneity}$[f]
10q22.3	rs704017 (**G**/A)	*ZMIZ1-AS1*	80489138	16,984/18,262	0.57	1.06 (1.03–1.10)	4.71×10^{-4}	0.20
10q25.2	rs11196172 (**A**/G)	*TCF7L2*	114716833	7,563/6,328	0.15	1.06 (0.99–1.13)	0.11	0.07
11q12.2	rs174537 (**G**/T)	*MYRF*	61309256	16,984/18,262	0.67	1.07 (1.04–1.11)	7.39×10^{-5}	0.001
	rs4246215 (**G**/T)	*FEN1*	61320875	16,984/18,262	0.65	1.07 (1.03–1.10)	2.71×10^{-4}	8.31×10^{-4}
	rs174550 (**T**/C)	*FADS1*	61328054	16,984/18,262	0.67	1.07 (1.03–1.10)	2.37×10^{-4}	8.87×10^{-4}
	rs1535 (**A**/G)	*FADS2*	61354548	16,984/18,262	0.67	1.07 (1.04–1.11)	4.12×10^{-5}	0.002
12p13.31	rs10849432 (**T**/C)	*CD9*	6255988	7,563/6,328	0.90	1.03 (0.95–1.11)	0.50	0.03
17p13.3	rs12603526 (**C**/T)	*NXN*	747343	16,984/18,262	0.02	1.12 (0.98–1.27)	0.10	0.83
19q13.2	rs1800469 (**G**/A)	*TGFB1*	46552136	16,984/18,262	0.67	1.03 (1.00–1.07)	0.09	0.01
	rs2241714 (**C**/T)	*B9D2*	46561232	16,984/18,262	0.67	1.02 (0.99–1.06)	0.18	0.007

Abbreviations: RAF, risk allele frequency; OR, odds ratio; CI, confidence interval.

[a] Risk/reference alleles for Asians as shown in Table 1. OR was estimated for the risk allele.

[b] The closest gene(s).

[c] The chromosome position (bp) is based on NCBI Build 36.

[d] RAF in controls.

[e] Summary OR (95% CI) and *P* value were obtained from a fixed-effects meta-analysis.

[f] *P* for heterogeneity between Asian and European populations was calculated using a Cochran's *Q* test.

Table 3

Association evidence in East Asians for risk variants in previously reported CRC susceptibility loci

						East Asians combined in this study				Published GWAS		
Locus	**SNP**	**Gene**[a]	**Annotation**	**Position**[b]	**Alleles**[c]	**N**	**RAF**[d]	**OR (95% CI)**	***P***	**RAF**[e]	**OR (95% CI)**[e]	$P_{heterogeneity}$[f]
Loci initially identified in East Asians												
5q31.1	rs647161	*PITX1*	Intergenic	134526991	**A**/C	40,051	0.31	1.15 (1.11–1.19)	1.87×10^{-14}	0.31	1.17 (1.11–1.22)	0.51
12p13.32	rs10774214	*CCND2*	Intergenic	4238613	**T**/C	33,436	0.37	1.14 (1.09–1.18)	1.40×10^{-10}	0.35	1.17 (1.11–1.23)	0.39
20p12.3	rs2423279	*HAO1*	Intergenic	7760350	**C**/T	40,057	0.31	1.13 (1.09–1.17)	3.04×10^{-12}	0.30	1.14 (1.08–1.19)	0.86
18q21.1	rs7229639	*SMAD7*	Intron 3	44704974	**A**/G	39,288	0.16	1.20 (1.16–1.25)	3.05×10^{-15}	0.15	1.22 (1.15–1.29)	0.72
Loci initially identified in Europeans												
1q41	rs6687758	*DUSP10*	Intergenic	220231571	**G**/A	37,803	0.24	1.12 (1.08–1.17)	8.99×10^{-9}	0.20	1.09 (1.06–1.12)	0.23
2q32.3	rs11903757	*NABP1*	Intergenic	192295449	**C**/T	22,442	0.05	1.15 (1.03–1.28)	0.01	0.16	1.16 (1.10–1.22)	0.89
3q26.2	rs10936599	*MYNN*	Exon 2	170974795	**C**/T	37,790	0.39	1.05 (1.01–1.08)	0.01	0.75	1.08 (1.05–1.10)	0.22
6p21.31	rs1321311	*CDKN1A*	Intergenic	36730878	**A**/C	32,236	0.14	1.09 (1.03–1.15)	0.001	0.23	1.10 (1.07–1.13)	0.77
8q24.21	rs10505477	*Unknown*	Intergenic	128476625	**A**/G	32,235	0.38	1.15 (1.11–1.20)	3.43×10^{-13}	0.51	1.17 (1.12–1.23)	0.64
8q24.21	rs6983267	*Unknown*	Intergenic	128482487	**G**/T	37,790	0.38	1.14 (1.10–1.18)	4.85×10^{-14}	0.52	1.21 (1.15–1.27)	0.06
8q24.21	rs7014346	*Unknown*	Intergenic	128493974	**A**/G	32,236	0.27	1.13 (1.08–1.17)	1.96×10^{-8}	0.37	1.19 (1.14–1.24)	0.06
10p14	rs10795668	*Unknown*	Intergenic	8741225	**G**/A	37,789	0.60	1.15 (1.11–1.19)	4.91×10^{-15}	0.67	1.12 (1.09–1.16)	0.30
11q13.4	rs3824999	*POLD3*	Intron 9	74023198	**G**/T	32,236	0.40	1.06 (1.02–1.11)	0.002	0.50	1.08 (1.05–1.10)	0.54
11q23.1	rs3802842	*Unknown*	Intergenic	110676919	**C**/A	37,791	0.38	1.09 (1.05–1.12)	2.57×10^{-7}	0.29	1.11 (1.08–1.15)	0.37
12q13.13	rs7136702	*LARP4*	Intergenic	49166483	**T**/C	37,774	0.51	1.02 (0.98–1.06)	0.31	0.35	1.06 (1.04–1.08)	0.05
12q13.13	rs11169552	*ATF1*	Intergenic	49441930	**C**/T	37,761	0.65	1.05 (1.01–1.09)	0.01	0.72	1.09 (1.06–1.12)	0.11
14q22.2	rs4444235	*BMP4*	Intergenic	53480669	**C**/T	37,785	0.53	1.04 (1.01–1.08)	0.02	0.46	1.11 (1.08–1.15)	0.007
14q22.2	rs1957636	*BMP4*	Intergenic	53629768	**T**/C	32,236	0.62	0.99 (0.95–1.04)	0.77	0.40	1.08 (1.06–1.11)	0.001
15q13.3	rs16969681	*SCG5*	Intergenic	30780403	**T**/C	32,236	0.44	1.07 (1.03–1.12)	0.002	0.09	1.18 (1.11–1.25)	0.01
15q13.3	rs4779584	*SCG5*	Intergenic	30782048	**T**/C	37,795	0.82	1.06 (1.01–1.11)	0.01	0.18	1.26 (1.19–1.34)	5.48×10^{-6}
15q13.3	rs11632715	*GREM1*	Intergenic	30791539	**A**/G	22,442	0.81	0.95 (0.90–1.01)	0.11	0.47	1.12 (1.08–1.16)	4.05×10^{-6}
16q22.1	rs9929218	*CDH1*	Intron 2	67378447	**G**/A	28,806	0.81	1.06 (1.00–1.11)	0.03	0.71	1.10 (1.07–1.13)	0.19
18q21.1	rs4939827	*SMAD7*	Intron 3	44707461	**T**/C	37,796	0.24	1.12 (1.08–1.16)	1.53×10^{-8}	0.52	1.18 (1.12–1.23)	0.11
19q13.11	rs10411210	*RHPN2*	Intron 2	38224140	**C**/T	37,789	0.82	1.12 (1.07–1.17)	3.14×10^{-6}	0.90	1.15 (1.10–1.20)	0.39

						East Asians combined in this study				Published GWAS		
Locus	**SNP**	**Gene**[a]	**Annotation**	**Position**[b]	**Alleles**[c]	**N**	**RAF**[d]	**OR (95% CI)**	***P***	**RAF**[e]	**OR (95% CI)**[e]	$P_{heterogeneity}$[f]
20p12.3	rs961253	*BMP2*	Intergenic	6352281	**A**/C	37,807	0.09	1.10 (1.04–1.17)	7.74×10^{-4}	0.36	1.12 (1.08–1.16)	0.66
20p12.3	rs4813802	*BMP2*	Intergenic	6647595	**G**/T	32,236	0.21	1.12 (1.06–1.17)	9.87×10^{-6}	0.36	1.09 (1.16–1.12)	0.37
20q13.33	rs4925386	*LAMA5*	Intron 10	60354439	**C**/T	37,780	0.77	1.05 (1.01–1.10)	0.01	0.68	1.08 (1.05–1.10)	0.38

Abbreviations: GWAS, genome-wide association study; RAF, risk allele frequency; OR, odds ratio; CI, confidence interval.

[a] The closest gene(s).

[b] The chromosome position (bp) is based on NCBI Build 36.

[c] Risk/reference alleles (in published GWAS) are based on forward allele coding in NCBI Build 36. OR was estimated for the risk allele (bold).

[d] RAF in controls.

[e] Results (RAF, ORs, and 95% CIs) from the original studies (ref. 7–19).

[f] *P* for heterogeneity between this study and published studies was calculated using a Cochran's *Q* test.