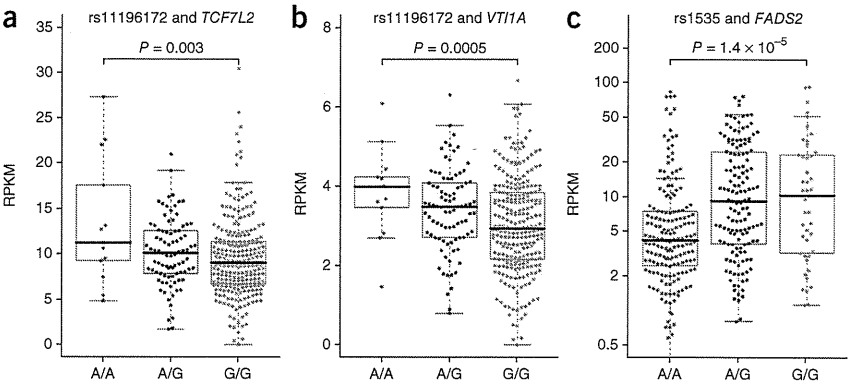

Figure 2 Association of selected risk variants identified in this study with gene expression in colon tumor tissue. (**a**) rs11196172 and *TCF7L2*. (**b**) rs11196172 and *VTI1A*. (**c**) rs1535 and *FADS2*. Gene expression levels are represented by reads per kilobase of exon per million mapped reads (RPKM) values based on the three genotypes of each SNP shown in red, blue and green. The median RPKM values and interquartile ranges (IQRs) for each SNP are presented in the overlaid box plots, and whiskers represent 1.5 times the IQR of the lower quartile to 1.5 times the IQR of the upper quartile. In **a** and **b**, RPKM values are shown at normal scale, whereas RPKM values in **c** are shown with a logarithmic scale owing to departure from a normal distribution. *P* values for associations between SNP genotypes and gene RPKM values were tested using a linear regression model.

whereas the minor allele frequency (MAF) is <0.02 in individuals of European descent. These differences might in part 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 were considerably different in East Asians and individuals of European descent (**Supplementary Fig. 5**). Large-scale fine-mapping of these loci will be helpful in identifying 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 (ref. 23), the 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], the Expression Atlas[31], PubMed and Online Mendelian Inheritance in Man (OMIM) (Online Methods). We summarize the results below for each locus.

At the 10q25.2 locus, rs11196172 is located in intron 4 of the *TCF7L2* gene. This SNP and other correlated SNPs ($r^2 > 0.5$) fall within a region with strong enhancer activity and a DNase I hypersensitivity site annotated by ENCODE (**Supplementary Table 14**), suggesting a potential functional role for these SNPs. We found that the risk-associated allele of rs11196172 was significantly associated with higher 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 a key transcription factor in the Wnt signaling pathway. Aberrant activation of Wnt signaling is found in more than 90% of CRCs[30], and TCF7L2 is a known tumor suppressor in 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 8–12.5% (refs. 29,30). Although *TCF7L2* is the only gene in this locus (**Supplementary Fig. 4**), we also found that the risk-associated allele of rs11196172 was significantly associated with higher expression of the *VTI1A* gene ($P = 5.1 \times 10^{-4}$) in colon tumor tissue (**Fig. 2**). The *VTI1A* gene is located approximately 131 kb upstream of the *TCF7L2* gene, and mRNA levels for these two genes are highly correlated in colon tumor tissue ($r = 0.71$; $P < 0.0001$). Recently, a recurrent gene fusion connecting the first three exons of *VTI1A* to the fourth exon of *TCF7L2* was identified in approximately 3% of colorectal tumors[34]. It is possible that the *VTI1A* gene might 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 these, rs1800469 has previously been investigated with respect to CRC risk in many small candidate gene association studies, with conflicting results[5]. We herein provide for the first time, to our knowledge, convincing evidence of association for rs1800469 through our GWAS analysis. SNP rs1800469 maps to the promoter of the *TGFB1* gene, and rs2241714 is a nonsynonymous SNP that results in an amino acid substitution at residue 11 of the B9D2 protein. The A allele of rs1800469 has been related to higher transcriptional activity for the *TGFB1* gene and higher circulating levels of the transforming growth factor (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 residue 10 of the TGF-β1 protein. Although the two nonsynonymous SNPs are predicted to be tolerated[36] or benign[37], the Pro10 variant encoded by rs1800470 has also been associated with an increase in *TGFB1* gene expression, TGF-β1 protein secretion and circulating levels of TGF-β1 protein[38–40]. Whereas 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, we suggest that many highly correlated SNPs located in the *TGFB1* gene might potentially have regulatory functions (**Supplementary Table 14**). The TGF-β1 protein is a member of the TGF-β signaling pathway. Somatic alterations of certain components in this pathway (*TGFBR2*, *SMAD2*, *SMAD3* and *SMAD4*) are estimated to be present in almost half of CRCs[41]. High-penetrance germline mutations in the *SMAD4* gene are known to cause juvenile polyposis, an autosomal dominant polyposis syndrome linked to 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]. Thus far, 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 these SNPs, rs4246215 is an eQTL for the *FEN1* gene in normal colorectal tissue[44] and is predicted to affect microRNA (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 ($P = 1.4 \times 10^{-5}$) in colon tumor tissue (**Fig. 2**). These findings suggest that the potential functions of these SNPs might be mediated through their effects 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 that 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 for *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 the metabolism of polyunsaturated fatty acids. Of these proteins, 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-associated variants for CRC identified in this study are strongly associated with circulating arachidonic acid levels[49]. We have shown previously that high levels of the PGE_2 metabolite in urine, a marker of endogenous PGE_2 production, are strongly related to higher risk of CRC[51]. Because the LD block of approximately 190 kb tagged by the four risk-associated variants covers many putatively functional SNPs that are located in the *FEN1*, *FADS1* and *FADS2* genes (**Supplementary Fig. 6** and **Supplementary Table 14**), it is difficult to pinpoint a single SNP or gene that might be responsible for the association with CRC risk in this locus. Nevertheless, our study provides evidence of a potentially important role for 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 identified 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, whose expression is downregulated in the Caco-2 and HT-29 CRC cell lines[31]. In line with these observations, we found in TCGA data that *ZMIZ1* gene expression is lower in colon tumor tissue compared with normal colon tissue ($P = 3.28 \times 10^{-6}$). In addition, somatic mutations in the *ZMIZ1* gene have been reported in more than 2% of colon tumors[29]. Whereas *ZMIZ1-AS1* is a miscellaneous RNA (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 might have 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-associated variant we identified for CRC, was not associated with CRC in this study (data not shown). Given the biological function of the *ZMIZ1* gene, it is possible that this gene is involved in the association observed in this locus.

In the 12p13.31 locus, rs10849432 maps to an LD block of approximately 52 kb with no known genes. ENCODE data suggest that rs4764551 and rs4764552, perfectly correlated with rs10849432, might 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 for 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*) all have downregulated expression 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 has a critical role in the suppression of cancer cell motility and metastasis[54], and overexpression of the *CD9* gene is associated with favorable prognosis for patients with CRC[55]. CD9 is also involved in suppressing Wnt signaling[56]. Although the 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 (TNF)-α and is known to be involved in cytokine-induced senescence in cancer[57]. In addition to evidence for the three nearby genes, we also found that rs4764552 is an eQTL for the *LTBR* gene (**Supplementary Table 15**). The LTβR protein has an essential role in lymphoid organ formation and has also been linked to cancer[58], including CRC[59]. On the basis of these data, we propose that the *CD9* gene is the most likely candidate to explain the association identified in this locus. However, potential roles for other genes cannot be excluded.

At the 17p13.3 locus, rs12603526 lies in intron 1 of the *NXN* gene, in 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 lower in the colon tumor tissue samples included in TCGA ($P = 2.83 \times 10^{-5}$). 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 nucleoredoxin dysfunction might cause activation of the transcription factor TCF (T cell factor), accelerated cell proliferation and enhancement of oncogenicity[61]. Further research is needed to determine the causal variant and biological mechanism for the association at this locus.

Previously reported CRC-associated loci in East Asians

We evaluated association evidence for 31 SNPs in 25 established CRC susceptibility loci[7–20] by analyzing data from stages 1–3 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 associations of the four loci identified previously in our GWAS conducted among East Asians ($P = 1.40 \times 10^{-10}$ to 3.05×10^{-15}). Of the 23 SNPs in the 18 susceptibility loci previously identified by GWAS of individuals of European descent, 20 showed association with CRC risk at $P < 0.05$ in East Asians in the same direction as reported in the original studies[7–17]. These signals included 6 SNPs in 4 loci (1q41, 8q24.21, 10p14 and 18q21.1) with association at $P < 5 \times 10^{-8}$, 6 SNPs in 6 loci with association at $P < 0.002$ (significance level adjusted for multiple comparisons of 25 independent loci) and 8 SNPs in 8 additional loci with 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 substantial risk of CRC in East Asian populations. In general, loci initially identified in individuals of European descent 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 individuals of European descent, are not polymorphic in East Asians, and SNP rs5934683 is located on the X chromosome. We did not have data to evaluate the associations of these three SNPs with CRC risk in this study.

Familial relative risk explained by CRC-associated 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 individuals of European descent and confirmed in this study. On the basis of per-allele OR values derived

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_{het}[f]
Loci initially identified in East Asians												
5q31.1	rs647161	*PITX1*	Intergenic	134,526,991	**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	4,238,613	**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	7,760,350	**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	44,704,974	**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 individuals of European descent												
1q41	rs6687758	*DUSP10*	Intergenic	220,231,571	**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	192,295,449	**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	170,974,795	**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	36,730,878	**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	128,476,625	**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	128,482,487	**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	128,493,974	**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	8,741,225	**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	74,023,198	**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	110,676,919	**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	49,166,483	**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	49,441,930	**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	53,480,669	**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	53,629,768	**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	30,780,403	**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	30,782,048	**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	30,791,539	**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	67,378,447	**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	44,707,461	**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	38,224,140	**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
20p12.3	rs961253	*BMP2*	Intergenic	6,352,281	**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	6,647,595	**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	60,354,439	**C**/T	37,780	0.77	1.05 (1.01–1.10)	0.01	0.68	1.08 (1.05–1.10)	0.38

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

[a]Closest gene(s). [b]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, OR and 95% CI values) from the original studies (refs. 7–19). [f]The *P* value for heterogeneity between this study and published studies was calculated using a Cochran's *Q* test.

from previously published GWAS[7–18] and this study, we estimate that the SNPs in the 31 loci identified thus far explain approximately 9% of the familial relative risk of CRC in individuals of European descent (**Supplementary Table 19**), a level slightly higher than the 7.7% explained in East Asians.

DISCUSSION

In the largest GWAS conducted thus far among East Asians, we identified six new genetic loci associated with CRC risk and provided suggestive evidence for three additional previously unreported 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-associated variants at some loci fall within potential functional regions, and two are associated with the expression levels of the *TCF7L2* and *FADS2* genes. This study expands current understanding of the genetic basis of CRC risk and provides evidence for new genes and biological pathways that might be involved in colorectal tumorigenesis.

On the basis of a large twin study conducted in Sweden, Denmark and Finland[2], the heritabilities estimated for CRC, breast cancer and prostate cancer were 35%, 27% and 42%, respectively. Thus far, 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 these cancers in individuals of European descent. For CRC, however, only 31 low-penetrance susceptibility loci have been identified, explaining approximately 9% of the familial relative risk of CRC in individuals of European descent. Compared with GWAS of breast cancer and prostate cancer, studies conducted for CRC have been relatively small. Our study, in which we evaluated approximately 7,000 promising variants identified by GWAS in the replication stages, represents one of the largest efforts thus far to follow up genetic variants identified by GWAS. We identified six new loci, representing the largest number of new loci identified for CRC risk in a single study. Although multiple GWAS with sample sizes larger than the one in this study have been conducted among individuals of European descent[13,14,16], we were still able to identify risk-associated variants with relatively large effect sizes. Our study further highlights the value of conducting GWAS in non-European populations to discover new 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 functions that have not yet been well linked to CRC. Our study considerably expands knowledge of the genetic landscape of CRC and provides direction for future studies to characterize the causal variants and functional mechanisms of these GWAS-identified loci.

URLs. 1000 Genomes Browser, http://browser.1000genomes.org/index.html; BioBank Japan (in Japanese), http://biobankjp.org/; Blood eQTL browser, http://genenetwork.nl/bloodeqtlbrowser/; Catalogue of Somatic Mutations in Cancer (COSMIC), http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/; database of Genotypes and Phenotypes (dbGaP), http://www.ncbi.nlm.nih.gov/gap; EIGENSTRAT, http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm; eQTL Browser from the University of Chicago, http://eqtl.uchicago.edu/Home.html; GTEx eQTL Browser, http://www.ncbi.nlm.nih.gov/projects/gap/eqtl/index.cgi/; 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/; Illumina HumanExome-12v1_A BeadChip, International Mouse Phenotyping Consortium (IMPC), https://www.mousephenotype.org/; LocusZoom, http://csg.sph.umich.edu/locuszoom/; http://genome.sph.umich.edu/wiki/Exome_Chip_Design; 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/; Multiple Tissue Human Expression Resource (MuTHER) Project, http://www.muther.ac.uk/; Online Mendelian Inheritance in Man (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/; R version 3.0.0, http://www.r-project.org/; SAS version 9.2, http://www.sas.com/; SIFT, SNAP, http://www.broadinstitute.org/mpg/snap/; http://sift.jcvi.org/; The Cancer Genome Atlas (TCGA), http://cancergenome.nih.gov/; TRANSFAC, http://www.gene-regulation.com/pub/databases.html; UCSC Genome Browser, http://genome.ucsc.edu/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

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 or 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, R. Courtney for DNA preparation, J. He for data processing and analyses, X. Guo for suggestions on bioinformatics analysis, and M.J. Daly and B.J. Rammer for editing and preparing the manuscript. The work at the Vanderbilt University School of Medicine was supported by US National Institutes of Health (NIH) grants R37CA070867, R01CA082729, R01CA124558, R01CA148667 and R01CA122364, as well as by Ingram Professorship and Research Reward funds from the Vanderbilt University School of Medicine. Studies (grant support) participating in the Asia Colorectal Cancer Consortium include the Shanghai Women's Health Study (US NIH, R37CA070867), the Shanghai Men's Health Study (US NIH, R01CA082729), the Shanghai Breast and Endometrial Cancer Studies (US NIH, R01CA064277 and R01CA092585; contributing only controls), Shanghai Colorectal Cancer Study 3 (US NIH, R37CA070867 and Ingram Professorship funds), the Guangzhou Colorectal Cancer Study (National Key Scientific and Technological Project, 2011ZX09307-001-04; the National Basic Research Program, 2011CB504303, contributing only controls; the Natural Science Foundation of China, 81072383, contributing only controls), the Japan BioBank Colorectal Cancer Study (grant from the Ministry of Education, Culture, Sports, Science and Technology of the Japanese government), the Hwasun Cancer Epidemiology Study–Colon and Rectum Cancer (HCES-CRC; grants from the Korea Center for Disease Control and Prevention and the Jeonnam Regional Cancer Center), the 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, 17015018 and 221S0001), the Korea-NCC (National Cancer Center) Colorectal Cancer Study (Basic Science Research Program through the National Research Foundation of Korea, 2010-0010276; National Cancer Center Korea, 0910220), the Korea-Seoul Colorectal Cancer Study (none reported) and the KCPS-II Colorectal Cancer Study (National R&D Program for Cancer Control, 1220180; Seoul R&D Program, 10526).

We also thank all participants, staff and investigators from the GECCO, CORECT and CCFR consortia for making it possible to present results from populations of European ancestry for the new CRC-associated loci identified among East Asians. GECCO, CORECT and CCFR are directed by U. Peters, S. Gruber and G. Casey, respectively. Complete lists of investigators from the GECCO, CORECT and CCFR consortia are provided below.

Investigators (institution and location) in the GECCO consortium include (in alphabetical order) John A. Baron (Division of Gastroenterology and Hepatology, University of North Carolina School of Medicine, Chapel Hill, North Carolina, USA), Sonja I. Berndt (Division of Cancer Epidemiology and Genetics, National Cancer Institute, US NIH, Bethesda, Maryland, USA), Stéphane Bezieau (Service de Génétique Médicale, Centre Hospitalier Universitaire (CHU) Nantes, Nantes, France), Hermann Brenner (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, California, USA), Christopher S. Carlson (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, School of Public Health, University of Washington, Seattle, Washington, USA), Graham Casey (Department of Preventive Medicine, University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA), Andrew T. Chan (Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA and Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, 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, US NIH, Bethesda, Maryland, USA), David V. Conti (Department of Preventive Medicine, University of Southern California, Los Angeles, California, USA), Keith Curtis (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA), David Duggan (Translational Genomics Research Institute, Phoenix, Arizona, USA), Charles S. Fuchs (Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA and Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA), Steven Gallinger (Department of Surgery, Mount Sinai Hospital, Toronto, Ontario, Canada and Samuel Lunenfeld Research Institute, Toronto, Ontario, Canada), Edward L. Giovannucci (Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA, Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA and Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA), Stephen B. Gruber (University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA), Robert W. Haile (Department of Preventive Medicine, University of Southern California, Los Angeles, California, USA), Tabitha A. Harrison (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA), Richard B. Hayes (Division of Epidemiology, Department of Environmental Medicine, New York University School of Medicine, New York, New York, USA), Michael Hoffmeister (Division of Clinical Epidemiology and Aging Research, German Cancer Research Center, Heidelberg, Germany), John L. Hopper (Melbourne School of Population Health, The University of Melbourne, Melbourne, Victoria, Australia), Li Hsu (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA and Department of Biostatistics, University of Washington, Seattle, Washington, USA), Thomas J. Hudson (Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada and Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada), David J. Hunter (Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA), Carolyn M. Hutter (Division of Cancer Control and Population Sciences, National Cancer Institute, US NIH, Bethesda, Maryland, USA), Rebecca D. Jackson (Division of Endocrinology, Diabetes and Metabolism, Ohio State University, Columbus, Ohio, USA), Mark A. Jenkins (Melbourne School of Population Health, The University of Melbourne, Melbourne, Victoria, Australia), Shuo Jiao (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA), Sébastien Küry (Service de Génétique Médicale, CHU Nantes, Nantes, France), Loic Le Marchand (Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, USA), Mathieu Lemire (Ontario Institute for Cancer Research, Toronto, Ontario, Canada), Noralane M. Lindor (Department of Health Sciences Research, Mayo Clinic, Scottsdale, Arizona, USA), Jing Ma (Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA), Polly A. Newcomb (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA and Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, USA), Ulrike Peters (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA and Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, USA), John D. Potter (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA, Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, USA and Centre for Public Health Research, Massey University, Palmerston North, New Zealand), Conghui Qu (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA), Thomas Rohan (Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Yeshiva University, Bronx, New York, USA), Robert E. Schoen (Department of Medicine and Epidemiology, University of Pittsburgh Medical Center, Pittsburgh, Pennsylvania, USA), Fredrick R. Schumacher (Department of Preventive Medicine, University of Southern California, Los Angeles, California, USA), Daniela Seminara (Division of Cancer Control and Population Sciences, National Cancer Institute, US NIH, Bethesda, Maryland, USA), Martha L. Slattery (Department of Internal Medicine, University of Utah Health Sciences Center, Salt Lake City, Utah, USA), Stephen N. Thibodeau (Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA and Department of Laboratory Genetics, Mayo Clinic, Rochester, Minnesota, USA), Emily White (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA and Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, USA) and Brent W. Zanke (Clinical Epidemiology Program, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada).

Investigators (institution and location) from the CORECT consortium include (in alphabetical order) Kendra Blalock (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA), Peter T. Campbell (Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA), Graham Casey (Department of Preventive Medicine, University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA), David V. Conti (Department of Preventive Medicine, University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA), Christopher K. Edlund (Department of Preventive Medicine, University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA), Jane Figueiredo (Department of Preventive Medicine, University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA), W. James Gauderman (Department of Preventive Medicine, University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA), Jian Gong (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA), Roger C. Green (Faculty of Medicine, Memorial University of Newfoundland, St. John's, Newfoundland, Canada), Stephen B. Gruber (University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA), John F. Harju (University of Michigan Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan, USA), Tabitha A. Harrison (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA), Eric J. Jacobs (Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA), Mark A. Jenkins (Melbourne School of Population Health, The University of Melbourne, Melbourne, Victoria, Australia), Shuo Jiao (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA), Li Li (Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, Ohio, USA), Yi Lin (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA), Frank J. Manion (University of Michigan Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan, USA), Victor Moreno (Institut d'Investigació Biomèdica de Bellvitge, Institut Catala d'Oncologia, Hospitalet, Barcelona, Spain), Bhramar Mukherjee (University of Michigan Comprehensive Cancer Center, University of Michigan, Ann Arbor, Michigan, USA), Ulrike Peters (Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA), Leon Raskin (University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA), Fredrick R. Schumacher (Department of Preventive Medicine, University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA), Daniela Seminara (Division of Cancer Control and Population Sciences, National Cancer Institute, US NIH, Bethesda, Maryland, USA), Gianluca Severi (Melbourne School of Population Health, The University of Melbourne, Melbourne, Victoria, Australia), Stephanie L. Stenzel (Department of Preventive Medicine, University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA) and Duncan C. Thomas (Department of Preventive Medicine, University of Southern

California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA).

The CCFR consortium is represented by Graham Casey (Department of Preventive Medicine, University of Southern California Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California, USA).

We also thank B. Buecher of ASTERISK; U. Handte-Daub, M. Celik, R. Hettler-Jensen, U. Benscheid and U. Eilber of DACHS; and P. Soule, H. Ranu, I. Devivo, D.J. Hunter, Q. Guo, L. Zhu and H. Zhang of HPFS, NHS and PHS, as well as the following state cancer registries for their help: Alabama, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Idaho, Illinois, Indiana, Iowa, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Nebraska, New Hampshire, New Jersey, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, Tennessee, Texas, Virginia, Washington and Wyoming. We thank C. Berg and P. Prorok of PLCO; T. Riley of Information Management Services, Inc.; B. O'Brien of Westat, Inc.; B. Kopp and W. Shao of SAIC-Frederick; the WHI investigators (see https://www.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 (US NIH, U01CA137088 and R01CA059045), DALS (US NIH, R01CA048998), DACHS (German Federal Ministry of Education and Research, BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, CH 117/1-1, 01KH0404 and 01ER0814), HPFS (US NIH, P01CA055075, UM1CA167552, R01137178 and P50CA127003), NHS (US NIH, R01137178, P50CA127003 and P01CA087969), OFCCR (US NIH, U01CA074783), PMH (US NIH, R01CA076366), PHS (US NIH, R01CA042182), VITAL (US NIH, K05CA154337), WHI (US NIH, HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, HHSN271201100004C and 268200764316C) and PLCO (US NIH, Z01CP 010200, U01HG004446 and U01HG 004438). CORECT is supported by the National Cancer Institute as part of the GAME-ON consortium (US NIH, U19CA148107) with additional support from National Cancer Institute grants (R01CA81488 and P30CA014089), the National Human Genome Research Institute at the US NIH (T32HG000040) and the National Institute of Environmental Health Sciences at the US NIH (T32ES013678). CCFR is supported by the National Cancer Institute, US NIH under RFA CA-95-011 and through cooperative agreements with members of the Colon Cancer Family Registry and principal investigators of the Australasian Colorectal Cancer Family Registry (US NIH, U01CA097735), the Familial Colorectal Neoplasia Collaborative Group (US NIH, U01CA074799) (University of Southern California), the Mayo Clinic Cooperative Family Registry for Colon Cancer Studies (US NIH, U01CA074800), the Ontario Registry for Studies of Familial Colorectal Cancer (US NIH, U01CA074783), the Seattle Colorectal Cancer Family Registry (US NIH, U01CA074794) and the University of Hawaii Colorectal Cancer Family Registry (US NIH, U01CA074806). The GWAS work was supported by a National Cancer Institute grant (US NIH, U01CA122839). OFCCR was supported by a GL2 grant from the Ontario Research Fund, Canadian Institutes of Health Research and a Cancer Risk Evaluation (CaRE) Program grant from the Canadian Cancer Society Research Institute. T.J. Hudson and B.W. 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 the Regional Council of Pays de la Loire, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (GEFLUC), the Association Anne de Bretagne Génétique and the Ligue Régionale Contre le Cancer (LRCC). PLCO data sets were accessed with approval through dbGaP (CGEMS prostate cancer scan, phs000207.v1.p1; CGEMS pancreatic cancer scan, phs000206.v4.p3; and GWAS of Lung Cancer and Smoking, phs000093.v2.p2, which was funded by Z01CP 010200, U01HG004446 and U01HG 004438 from the US NIH).

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.; and D.-H.K. directed CRC projects for the Guangzhou Colorectal Cancer Study, the BioBank Japan Colorectal Cancer Study, the Hwasun Cancer Epidemiology Study–Colon and Rectum Cancer (HCES-CRC), the Aichi Colorectal Cancer Study, the Shanghai studies, the Korea-NCC (National Cancer Center) Colorectal Cancer Study, the KCPS-II Colorectal Cancer Study and the Korea-Seoul Colorectal Cancer Study, respectively. B.Z., Q.C. and W.W. coordinated the project. Q.C. directed laboratory operations. J.S. performed the genotyping experiments. B.Z. performed the statistical and bioinformatics analyses. W.W. contributed to the statistical analyses and data interpretation. A.T. conducted the statistical analyses and imputation for BioBank Japan. B.Z., W.W. and J.L. managed the data. Y.Z. and B.Z. performed the expression analysis for TCGA data. B.Z. and W.Z. wrote the manuscript 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 to manuscript revision. All authors have reviewed and approved the content of the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Jemal, A. *et al.* Global cancer statistics. *CA Cancer J. Clin.* **61**, 69–90 (2011).
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.* **343**, 78–85 (2000).
3. de la Chapelle, A. Genetic predisposition to colorectal cancer. *Nat. Rev. Cancer* **4**, 769–780 (2004).
4. Aaltonen, L., Johns, L., Jarvinen, H., Mecklin, J.P. & Houlston, R. Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. *Clin. Cancer Res.* **13**, 356–361 (2007).
5. Ma, X., Zhang, B. & Zheng, W. Genetic variants associated with colorectal cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *Gut* **63**, 326–336 (2014).
6. Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat. Genet.* **45**, 136–144 (2013).
7. Zanke, B.W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989–994 (2007).
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.* **39**, 984–988 (2007).
9. Broderick, P. *et al.* A genome-wide association study shows that common alleles of *SMAD7* influence colorectal cancer risk. *Nat. Genet.* **39**, 1315–1317 (2007).
10. Jaeger, E. *et al.* Common genetic variants at the *CRAC1* (*HMPS*) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat. Genet.* **40**, 26–28 (2008).
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.* **40**, 631–637 (2008).
12. Tomlinson, I.P. *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.* **40**, 623–630 (2008).
13. Houlston, R.S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40**, 1426–1435 (2008).
14. Houlston, R.S. *et al.* Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat. Genet.* **42**, 973–977 (2010).
15. Tomlinson, I.P. *et al.* Multiple common susceptibility variants near BMP pathway loci *GREM1*, *BMP4*, and *BMP2* explain part of the missing heritability of colorectal cancer. *PLoS Genet.* **7**, e1002105 (2011).
16. Dunlop, M.G. *et al.* Common variation near *CDKN1A*, *POLD3* and *SHROOM2* influences colorectal cancer risk. *Nat. Genet.* **44**, 770–776 (2012).
17. Peters, U. *et al.* Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. *Gastroenterology* **144**, 799–807 (2013).
18. Jia, W.H. *et al.* Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nat. Genet.* **45**, 191–196 (2013).
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* doi:10.1002/ijc.28733 (21 January 2014).
20. Cui, R. *et al.* Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. *Gut* **60**, 799–805 (2011).
21. Figueiredo, J.C. *et al.* Genotype-environment interactions in microsatellite stable/microsatellite instability-low colorectal cancer: results from a genome-wide association study. *Cancer Epidemiol. Biomarkers Prev.* **20**, 758–766 (2011).
22. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
23. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
24. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
25. Westra, H.J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
26. Degner, J.F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
27. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
28. Grundberg, E. *et al.* Mapping *cis*- and *trans*-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).
29. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
30. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).

31. Kapushesky, M. *et al.* Gene Expression Atlas update—a value-added database of microarray and sequencing-based functional genomics experiments. *Nucleic Acids Res.* **40**, D1077–D1081 (2012).
32. Tang, W. *et al.* A genome-wide RNAi screen for Wnt/β-catenin pathway components identifies unexpected roles for TCF transcription factors in cancer. *Proc. Natl. Acad. Sci. USA* **105**, 9697–9702 (2008).
33. Angus-Hill, M.L., Elbert, K.M., Hidalgo, J. & Capecchi, M.R. T-cell factor 4 functions as a tumor suppressor whose disruption modulates colon cell proliferation and tumorigenesis. *Proc. Natl. Acad. Sci. USA* **108**, 4914–4919 (2011).
34. Bass, A.J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent *VTI1A-TCF7L2* fusion. *Nat. Genet.* **43**, 964–968 (2011).
35. Grainger, D.J. *et al.* Genetic control of the circulating concentration of transforming growth factor type β1. *Hum. Mol. Genet.* **8**, 93–97 (1999).
36. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
37. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
38. Dunning, A.M. *et al.* A transforming growth factor β1 signal peptide variant increases secretion *in vitro* and is associated with increased incidence of invasive breast cancer. *Cancer Res.* **63**, 2610–2615 (2003).
39. Suthanthiran, M. *et al.* Transforming growth factor-β_1 hyperexpression in African-American hypertensives: a novel mediator of hypertension and/or target organ damage. *Proc. Natl. Acad. Sci. USA* **97**, 3479–3484 (2000).
40. Yamada, Y. *et al.* Association of a polymorphism of the transforming growth factor-β1 gene with genetic susceptibility to osteoporosis in postmenopausal Japanese women. *J. Bone Miner. Res.* **13**, 1569–1576 (1998).
41. Markowitz, S.D. & Bertagnolli, M.M. Molecular origins of cancer: molecular basis of colorectal cancer. *N. Engl. J. Med.* **361**, 2449–2460 (2009).
42. Howe, J.R. *et al.* Mutations in the *SMAD4/DPC4* gene in juvenile polyposis. *Science* **280**, 1086–1088 (1998).
43. Valle, L. *et al.* Germline allele-specific expression of *TGFBR1* confers an increased risk of colorectal cancer. *Science* **321**, 1361–1365 (2008).
44. Liu, L. *et al.* Functional *FEN1* genetic variants contribute to risk of hepatocellular carcinoma, esophageal cancer, gastric cancer and colorectal cancer. *Carcinogenesis* **33**, 119–123 (2012).
45. Xu, Z. & Taylor, J.A. SNPinfo: integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.* **37**, W600–W605 (2009).
46. Zheng, L. *et al.* Functional regulation of FEN1 nuclease and its link to cancer. *Nucleic Acids Res.* **39**, 781–794 (2011).
47. Zheng, L. *et al. Fen1* mutations result in autoimmunity, chronic inflammation and cancers. *Nat. Med.* **13**, 812–819 (2007).
48. Kucherlapati, M. *et al.* Haploinsufficiency of Flap endonuclease (*Fen1*) leads to rapid tumor progression. *Proc. Natl. Acad. Sci. USA* **99**, 9924–9929 (2002).
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.* **15**, 1745–1756 (2006).
50. Castellone, M.D., Teramoto, H., Williams, B.O., Druey, K.M. & Gutkind, J.S. Prostaglandin E_2 promotes colon cancer cell growth through a G_s-axin-β-catenin signaling axis. *Science* **310**, 1504–1510 (2005).
51. Cai, Q. *et al.* Prospective study of urinary prostaglandin E_2 metabolite and colorectal cancer risk. *J. Clin. Oncol.* **24**, 5010–5016 (2006).
52. Rogers, L.M., Riordan, J.D., Swick, B.L., Meyerholz, D.K. & Dupuy, A.J. Ectopic expression of Zmiz1 induces cutaneous squamous cell malignancies in a mouse model of cancer. *J. Invest. Dermatol.* **133**, 1863–1869 (2013).
53. Turnbull, C. *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* **42**, 504–507 (2010).
54. Ovalle, S. *et al.* The tetraspanin CD9 inhibits the proliferation and tumorigenicity of human colon carcinoma cells. *Int. J. Cancer* **121**, 2140–2152 (2007).
55. Mori, M. *et al.* Motility related protein 1 (MRP1/CD9) expression in colon cancer. *Clin. Cancer Res.* **4**, 1507–1510 (1998).
56. Lee, J.H. *et al.* Glycoprotein 90K, downregulated in advanced colorectal cancer tissues, interacts with CD9/CD82 and suppresses the Wnt/β-catenin signal via ISGylation of β-catenin. *Gut* **59**, 907–917 (2010).
57. Braumüller, H. *et al.* T-helper-1-cell cytokines drive cancer into senescence. *Nature* **494**, 361–365 (2013).
58. Wolf, M.J., Seleznik, G.M., Zeller, N. & Heikenwalder, M. The unexpected role of lymphotoxin β receptor signaling in carcinogenesis: from lymphoid tissue formation to liver and prostate cancer development. *Oncogene* **29**, 5006–5018 (2010).
59. Lukashev, M. *et al.* Targeting the lymphotoxin-β receptor with agonist antibodies as a potential cancer therapy. *Cancer Res.* **66**, 9617–9624 (2006).
60. Funato, Y. & Miki, H. Nucleoredoxin, a novel thioredoxin family member involved in cell growth and differentiation. *Antioxid. Redox Signal.* **9**, 1035–1057 (2007).
61. Funato, Y., Michiue, T., Asashima, M. & Miki, H. The thioredoxin-related redox-regulating protein nucleoredoxin inhibits Wnt–β-catenin signalling through Dishevelled. *Nat. Cell Biol.* **8**, 501–508 (2006).
62. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45**, 353–361 (2013).
63. Eeles, R.A. *et al.* Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat. Genet.* **45**, 385–391 (2013).

ONLINE METHODS

Study participants. This GWAS was conducted as part of the Asia Colorectal Cancer Consortium, comprising a total of 14,963 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 5 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 4 studies were the same as we reported in our previous study[18]. Stage 2 consisted of 3 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 1 study: the BioBank Japan CRC Study (BBJ; n = 14,172). Stage 4 consisted of 5 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 greater at $P < 5 \times 10^{-8}$ for SNPs with a MAF of as low as 0.30. We evaluated the generalizability of the newly identified associations with CRC risk in individuals of European descent in data from 3 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 the **Supplementary Note**. Study protocols were approved by the relevant review boards in the respective institutions, and informed consent was obtained from all study participants.

Laboratory procedures. Genotyping of samples in stage 1 was conducted as described previously[18,64–69] using the following platforms: the Affymetrix Genome-Wide Human SNP Array 6.0, the Illumina HumanOmniExpress BeadChip, the Illumina Infinium HumanHap550 BeadChip, the Illumina 660W-Quad BeadChip, the Illumina Human610-Quad BeadChip, the Illumina Infinium HumanHap610 BeadChip and the Affymetrix Genome-Wide Human SNP Array 5.0. We used a uniform quality control protocol as recently described[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 calling and quality control are provided in the **Supplementary Note**. Samples were excluded according to the following criteria: (i) genotype call rate of <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) ancestry outliers or (vi) heterozygosity outliers. Genetic markers were excluded using the following criteria: (i) MAF = 0, (ii) genotype call rate of <98%, (iii) consistency rate of <98% in positive quality control samples, (iv) P for Hardy-Weinberg equilibrium $< 1 \times 10^{-5}$ in controls or (v) caution SNPs revealed by the Exome Chip Design group (see URLs). We obtained a final data set including 6,899 SNPs genotyped in 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 data set comprising 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, USA). Details of genotyping and quality control are provided in the **Supplementary Note**. We filtered out SNPs with (i) genotype call rate of <95%, (ii) genotyping consistency rate of <95% in positive control samples, (iii) an unclear genotype call or (iv) P for Hardy-Weinberg equilibrium of $< 1 \times 10^{-5}$ in controls. The average consistency rate of these SNPs passing quality control filters was 99.9% with a median value of 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 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 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 the HapMap Asian population), (vi) SNPs uncorrelated with other SNPs identified in this project ($r^2 < 0.2$) and (vii) data available in at least two studies (**Supplementary Note**). We included multiple SNPs in some regions with a prior association 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, 8,570 unique SNPs were selected. Of these, 7,113 SNPs were successfully designed. For stage 3 replication, we selected 559 SNPs according to the following criteria: (i) $P < 0.005$ in the 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 the following criteria: (i) $P < 0.0001$ in the meta-analysis of stages 1–3, (ii) $P < 0.01$ in the meta-analysis of stages 2 and 3, (iii) association in the same direction in the three stages and (iv) P for heterogeneity > 0.0001.

Statistical and bioinformatics analysis. Details of imputation and population substructure evaluation are provided in the **Supplementary Note**. Briefly, stage 1 imputation was performed with the CHB (Han Chinese in Beijing, China) and JPT (Japanese in Tokyo, Japan) HapMap 2 panel as the reference using the MACH v1.0 program[70] (see URLs). Stage 3 imputation was conducted with phased data for JPT, CHS (Southern Han Chinese, China) and CHD (Chinese in Metropolitan Denver, Colorado) participants from 1000 Genomes Project phase 1 release v3 as the reference using MACH v1.0 and Minimac[71] (see URLs). Regional imputation of genotype data from TCGA[30] (see URLs) was performed with the GIANT ALL reference panel from 1000 Genomes Project phase 1 release v3 using MACH v1.0 and Minimac (see URLs). To evaluate imputation quality in our study, we directly genotyped the 10 newly identified risk variants in the 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 data from this study were therefore 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). On the basis of 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 value 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, $\lambda_{1,000}$, representing the equivalent value for a study with 1,000 cases and 1,000 controls using the formula[73] $\lambda_{1,000} = 1 + 500 \times (\lambda - 1) \times (1/N_{cases} + 1/N_{controls})$ was 1.01 in both stages 1 and 2. These findings show 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, see URLs). Per-allele OR estimates and 95% 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 the Metal[75]

program. SNPs showing an association at $P < 5 \times 10^{-8}$ in the combined analysis of all studies were considered genome-wide significant. We also performed stratified analyses for the top SNPs by tumor anatomical site (colon or rectum), population (Chinese, Korean or Japanese) and sex (male or female). We estimated heterogeneity across studies and subgroups with a Cochran's Q test[76], with P for heterogeneity < 0.008 set as statistically significant when considering multiple comparisons of six independent loci. Independent signals in a locus were identified using stepwise logistic regression models conditioning on the top risk-associated 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 6 top risk-associated variants in the newly identified loci with association $P < 5 \times 10^{-8}$ and also between these 6 SNPs and the risk-associated variants in 25 previously reported loci were evaluated using the maximum-likelihood ratio test with inclusion of interaction terms in logistic regression models. Interactions with $P < 0.00028$ were considered statistically significant with adjustment for multiple comparisons of 180 tests.

The familial relative risk (λ) for the 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 $\log(\lambda)/\log(\lambda_0)$, where λ_0 is the overall familial relative risk. λ_0 is assigned to be 2.2 for CRC risk estimated from a meta-analysis[79]. Assuming that the risks associated with individual loci combine multiplicatively, the familial relative risks also multiply. Thus, the combined contribution of the familial relative risks from multiple loci is equal to

$$\ln\left(\prod_i \lambda_i\right) / \ln(\lambda_0)$$

We generated forest plots and quantile-quantile 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). LD structure between SNPs was determined on the basis of data from 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 genomic coordinates are based on NCBI Build 36.

To find putative functional variants for newly identified loci, we identified all SNPs in LD ($r^2 > 0.5$) with the risk-associated variants using data from the 1000 Genomes Project[22] and HapMap 2 (ref. 23). We mapped the genomic locations of these SNPs to nonsynonymous sites, splice sites, promoters, nearGene-3 regions, nearGene-5 regions, 3′ UTRs, 5′ UTRs, introns and intergenic regions. We evaluated the potential functional effect of nonsynonymous SNPs using the prediction algorithms SIFT[36] and PolyPhen-2 (ref. 37) (see URLs). We predicted the putative function of SNPs in promoters, nearGene-3 regions, nearGene-5 regions, 3′ UTRs and 5′ UTRs with the SNPinfo Web Server[45] (see URLs). We conducted analyses to evaluate the potential regulatory effect of SNPs in noncoding regions on transcription using the 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 modification, proteins bound to these regulatory sites, *cis*-eQTL and transcription factor binding motifs. We obtained additional functional evidence for these SNPs from the published literature.

We identified all genes that localize to 1-Mb windows centered on the top risk-associated variants in our newly identified loci, including SNPs correlated ($r^2 > 0.5$) with the top risk variants. To determine whether these genes might explain the observed associations 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 (GTEx) Project[27] and the Multiple Tissue Human Expression Resource (MuTHER) Project[28]. The significance threshold for these analyses was set to $P < 0.008$ to account for six tests. Somatic mutations of these genes were evaluated using data from COSMIC[29] (see URLs). Expression levels of these genes in CRC cell lines were assessed using data from the Expression Atlas[31] (see URLs). To correct for multiple comparisons of the 11 key genes, associations with $P < 0.0045$ were considered to be statistically significant. We searched the published literature for these genes with respect to CRC in 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 RPKM (reads per kilobase of exon per million mapped reads) values as previously described[83]. Expression differences between tumor and normal samples for each gene were evaluated on the basis of RPKM values with the Wilcoxon rank-sum test. Associations between gene RPKM values 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.

64. Abnet, C.C. *et al.* A shared susceptibility locus in *PLCE1* at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nat. Genet.* **42**, 764–767 (2010).
65. Amundadottir, L. *et al.* Genome-wide association study identifies variants in the *ABO* locus associated with susceptibility to pancreatic cancer. *Nat. Genet.* **41**, 986–990 (2009).
66. Bei, J.X. *et al.* A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. *Nat. Genet.* **42**, 599–603 (2010).
67. Nakata, I. *et al.* Association between the *SERPING1* gene and age-related macular degeneration and polypoidal choroidal vasculopathy in Japanese. *PLoS ONE* **6**, e19108 (2011).
68. Jee, S.H. *et al.* Adiponectin concentrations: a genome-wide association study. *Am. J. Hum. Genet.* **87**, 545–552 (2010).
69. Zheng, W. *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.* **41**, 324–328 (2009).
70. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34**, 816–834 (2010).
71. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
72. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
73. Freedman, M.L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* **36**, 388–393 (2004).
74. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
75. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
76. Lau, J., Ioannidis, J.P. & Schmid, C.H. Quantitative synthesis in systematic reviews. *Ann. Intern. Med.* **127**, 820–826 (1997).
77. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
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.* **22**, 2539–2550 (2013).
79. Johns, L.E. & Houlston, R.S. A systematic review and meta-analysis of familial colorectal cancer risk. *Am. J. Gastroenterol.* **96**, 2992–3003 (2001).
80. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).
81. Johnson, A.D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938–2939 (2008).
82. Ward, L.D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
83. Yan, G. *et al.* Genome sequencing and comparison of two nonhuman primate animal models, the cynomolgus and Chinese rhesus macaques. *Nat. Biotechnol.* **29**, 1019–1023 (2011).

Identification of novel epigenetically inactivated gene *PAMR1* in breast carcinoma

PAULISALLY HAU YI LO[1], CHIZU TANIKAWA[1], TOYOMASA KATAGIRI[2], YUSUKE NAKAMURA[3] and KOICHI MATSUDA[1]

[1]Laboratory of Molecular Medicine, Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo; [2]Division of Genome Medicine, Institute for Genome Research, The University of Tokushima, Tokushima, Japan; [3]Departments of Medicine and Surgery, and Center for Personalized Therapeutics, The University of Chicago, Chicago, IL, USA

DOI: 10.3892/or_xxxxxxxx

Abstract. Development of cancer is a complex process involving multiple genetic and epigenetic alterations. In our microarray analysis of 81 breast carcinoma specimens, we identified *peptidase domain containing associated with muscle regeneration 1 (PAMR1)* as being frequently suppressed in breast cancer tissues. *PAMR1* expression was also reduced in all tested breast cancer cell lines, while *PAMR1* was expressed moderately in normal breast tissues and primary mammary epithelial cells. DNA sequencing of the *PAMR1* promoter after sodium bisulfite treatment revealed that CpG sites were hypermethylated in the breast cancer tissues and cell lines. *PAMR1* expression was restored by 5-aza-2' deoxycytidine treatment, demonstrating that promoter hypermethylation contributed to *PAMR1* inactivation in the breast cancer cells. In addition, ectopic expression of PAMR1 markedly suppressed cancer cell growth. In summary, our study identified PAMR1 as a putative tumor suppressor which was frequently inactivated by promoter hypermethylation in breast cancer tissues.

Introduction

Cancer is the leading cause of death in most developed countries, and breast cancer is one of the leading causes of cancer-related mortality among women (1). Although surgery and follow-up treatment have been successful in improving the prognosis of breast cancer patients, patients with metastatic tumors still suffer from poor prognosis. Therefore, developing novel therapeutics for breast cancer is an absolute necessity. For this purpose, understanding the molecular mechanism of breast carcinogenesis is essential. Microarray technology which provides quantitative genome-wide gene expression profiling has been widely used to analyze the pathways associated with cancer development and progression. (2). Through the screening of genes which showed enhanced expression in breast cancer tissues, we identified several molecular targets that are essential for breast cancer cell proliferation (3-6). For example, brefeldin A-inhibited guanine nucleotide-exchange protein 3 (BIG3), which was found to be frequently upregulated in breast cancer tissues, interacts with prohibitin 2/repressor of estrogen receptor activity (PHB2/REA) protein. This binding inhibits PHB2/REA nuclear translocation and subsequently activates ERα signaling pathways (7). In addition, a synthesized peptide which inhibits the interaction between BIG3 and PHB2/REA is able to suppress E2-dependent breast cancer cell growth (8).

Similarly, identification of genes which exhibit low expression in cancer tissues is also important for the understanding of human carcinogenesis. Tumor-suppressor genes (TSGs) act as guardians against malignant transformation. Genomic alteration or promoter hypermethylation are common causes of TSG inactivation. In breast cancer tissues, hypermethylation of TSGs is considered to be an early event during tumorigenesis (9). Overexpression of DNA methyltransferase *(DNMT)* 1, 3a, and 3b is frequently observed in breast fibroadenoma (22-44%) (10), which may result in TSG promoter hypermethylation including *APC*, *BRCA1*, *p16*, *p21* and *TIMP3* (11-13). Several studies have demonstrated that hypermethylated DNA of TSGs in serum could be a potential biomarker for disease prediction and therapeutic response in breast cancer (14). In addition, DNMT inhibitors are used for the treatment of myelodysplastic syndrome and solid cancers (15-17). Therefore, identification of novel TSGs would not only provide a fundamental understanding of cancer biology, but may also contribute to breast cancer diagnosis or more effective therapeutics. Recently, we reported a TSG candidate, *HSPB7*, which was found to be downregulated in renal cancer samples by epigenetic abnormalities (18). In the present study, we used microarray technology and identified *peptidase domain containing associated with muscle regeneration 1 (PAMR1)* whose expression was frequently suppressed in breast cancer tissues by promoter hypermethylation.

Correspondence to: Dr Koichi Matsuda, Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
E-mail: koichima@ims.u-tokyo.ac.jp

Key words: breast carcinoma, epigenetically inactivated gene, PAMR1, microarray

Table I. List of primers used in the present study.

	Forward primer	Reverse primer
Cell line RT-PCR		
C2orf88	GCTTAATCACAATGCCCTCAAC	CTGAACTAATGCCCACAGCTC
CSRNP3	AGTGGGGACAGTGTCAATCC	CCTTGCCTCCTGGTGAAGTA
PAMR1	CCTTCTCATCTCGTCCTTGC	AACCCACGACTTCCCTCTTT
PDLIM3	CTCAGGGGGCATAGACTTC	ATCTCCAGGACACAGGTTGG
PPP1R12B	TGACCAGCCGTGTAGAAGAAG	CTGGGCTTCCTGAAGTTTTG
SAMD5	GCTACCCCAAACTGAAGCTG	AGCGGCTCTGTGATGACTTC
Tissue RT-PCR	AGGGAAGATCTGGGCTTCATG	GGGAAGGAAAAGGACCAGAC
Cloning PCR	TTAAGAATTCGCGGCAAGGATGGAGCTGGG	CGCGCTCGAGTTTCATATTTCTTTCAATCC
Isoform PCR	TTACAAGTGTGCCTGCTTGG	GCCCCCTGTTATTTTCTGGT
Bisulfite sequencing	TTAATTTGTGATTATTTGGAGTAAA	CTCATCTAAAAAAAACCACCTTCAA

Materials and methods

Breast cancer cell lines and clinical cancer samples. Human breast cancer cell lines including BSY1, BT-20, BT-474, BT-549, HBC4, HBC5, HBL-100, HCC1143, HCC1395, HCC1500, HCC1599, HCC1937, MCF7, MDA-MB-231, MDA-MB-435s, MDA-MB-453, OCUB-F, SK-BR-3, T-47D, YMB-1 and ZR-75-1 were obtained and cultured as previously reported (4). The cell lines BST1, HBC4 and HBC5 were kindly provided by Dr Takao Yamori of the Division of Molecular Pharmacology, Cancer Chemotherapy Center, Japanese Foundation for Cancer Research. The other cell lines were purchased from the American Type Culture Collection (ATCC, USA). Human mammary epithelial cells (HMECs) were purchased from Lonza Switzerland and were cultured in mammary epithelial cell growth medium supplemented with bovine pituitary extract, hEGF, hydrocortisone, GA-1000 and insulin (Lonza). The HMECs used for all experiments were under passage 15. All cells were maintained at 37˚C in an atmosphere of humidified air with 5% CO_2 except for MDA-MB-231 and MDA-MB-435s which were maintained at 37˚C in an atmosphere of humidified air without CO_2. Primary breast normal and cancer tissues were obtained with informed consent from patients who received treatment at the Department of Breast Surgery, Cancer Institute Hospital, Tokyo. All tissue samples underwent laser-microbeam microdissection (19).

Plasmid construction. The two *PAMR1* isoforms were amplified from HMEC cDNA by KOD plus DNA polymerase (Toyobo, Japan). The sequences of the cloning primers are listed in Table I. The amplified DNAs were then subsequently cloned into the pCAGGS vector with HA-tagged at the C-terminal.

cDNA microarray. cDNA microarray analysis was performed as previously described (19). In brief, tumor cells obtained from 81 breast cancer patients (12 ductal carcinomas *in situ* and 69 T2 invasive ductal carcinomas) underwent laser microbeam microdissection. The total RNAs were extracted using the RNeasy Mini kit (Qiagen, Germany) and treated with DNase I digestion according to the manufacturer's manual. The RNAs were then reverse-transcribed and hybridized with the microarray slide. The microarray slide contained 23,040 cDNAs selected from the UniGene database (build #131), including 52 housekeeping genes and two types of negative control genes. A mixture of normal breast ductal cell RNAs isolated from 15 pre-menopausal breast cancer patients was used as the normal control.

Real-time quantitative PCR. The mRNAs of human normal tissues were purchased from Takara (Takara Bio, Japan). Total RNAs from the cell lines were extracted using the RNeasy Mini kit and reverse transcribed into cDNA by SuperScript III (Life Technologies, USA) according to the manufacturer's instructions. Real-time quantitative PCR (qPCR) was performed using SYBR-Green I Master Mix on LightCycler 480 (Roche, Germany). The primer sequences are listed in Table I.

DNA isolation, sodium bisulfite treatment and DNA sequencing. Genomic DNAs were isolated by DNeasy Blood & Tissue kit (Qiagen) according to the instruction manual. Bisulfite treatment and DNA sequencing was performed as previously reported (20). In brief, 2 μg of DNA was digested by *Xho*I for 16 h at 37˚C. The digested DNA was then denatured by 0.3 M NaOH and treated with 3.12 M sodium bisulfite and 0.5 mM hydroquinone for 16 h at 55˚C. Following incubation, DNA was purified and desulfoned by 0.3 M of NaOH at 37˚C for 20 min, followed by ethanol precipitation. Finally the DNA was amplified by PCR with the specific primers (Table I) and subcloned into the pCR 2.1 vector by TA cloning kit (Invitrogen, USA). The cloned plasmids were transformed into competent cells. For each treated DNA, 10 individual colonies were chosen and plasmid extractions were performed. DNA sequencing of the isolated plasmids was performed by the ABI sequencing system (Applied Biosystems, USA) according to the manufacturer's instructions.

Demethylation drug treatment. The demethylation drug 5-aza-2' deoxycytidine (5-aza-dC) was purchased from Sigma (Sigma-Aldrich, USA). The drug was dissolved in

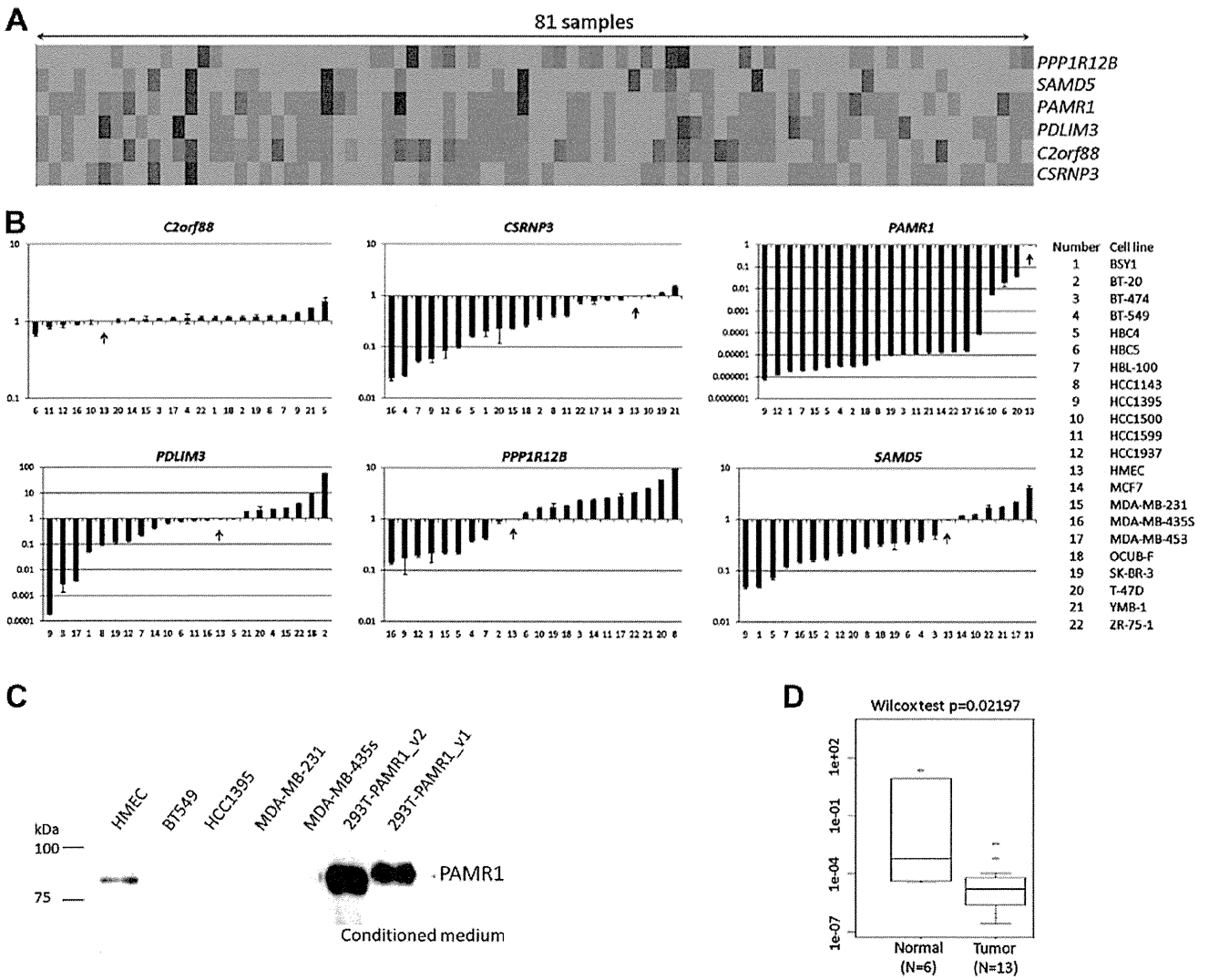

Figure 1. Microarray analysis identified *PAMR1* which exhibited low expression in breast cancer. (A) Heatmap showing the microarray results of the 6 down-regulated genes in the breast cancer samples. (B) qPCR results showing the relative expression of the 6 candidate genes compared with HMECs (indicated by an arrow) in 21 breast cancer cell lines. Data represent means ± SD. The list of cell lines are shown in the right panel. (C) Endogenous PAMR1 expression in HMECs and breast cancer cell lines. Conditioned medium was collected and separated by 8% SDS-PAGE. PAMR1 secreted protein was detected by sheep anti-PAMR1 antibody. (D) Boxplot representing the expression of *PAMR1* in normal and tumor specimens from breast cancer patients.

dimethyl sulphoxide (DMSO) and freshly prepared before use. Breast cancer cells were cultured in 6-well plates one day before drug treatment. Fresh medium containing various concentrations of 5-aza-dC was replaced daily for 3 consecutive days. The RNA from each treated cell line was isolated 72 h post drug treatment. Cells treated with DMSO served as the negative controls.

Western blotting. Breast cancer cells ($5x10^5$) were cultured in a 60-mm dish under normal conditions and allowed to attach for 24 h. The culture medium was then removed and the cells were washed twice by PBS. A total of 2 ml of fresh medium without FBS was then replaced, and the cells were allowed to grow for another 24 h. After incubation, 1 ml of conditioned medium was collected from each sample, followed by centrifugation at 15,000 rpm for 15 min at 4°C twice to remove all floating cells. The conditioned medium was then mixed with an equal volume of ice-cold acetone and stored at -80°C for 1 h. The protein was harvested by centrifugation at 15,000 rpm for 15 min at 4°C. The precipitated protein was dissolved using Laemmli sample buffer and analyzed by western blotting following standard protocols (Bio-Rad, USA). Rat anti-HA antibody (Roche) and sheep anti-PAMR1 antibody (R&D Systems, USA) were used to detect PARM1 protein in the conditioned medium. Mouse anti-β-actin antibody (Santa Cruz, USA) was used as the loading control.

Colony formation assay. Breast cancer cells were cultured in 6-well plates for 24 h before transfection. One hundred and fifty million copies of plasmid from the vector alone control (pCAGGS), and two variants of *PAMR1* were transfected into each well individually by FuGene HD (Roche) in a 1:3 (μg:μl) ratio. Transfection was performed according to the user manual. G418 (Life Technologies) was added to the cells one day after transfection. The drug-resistant cells were allowed to grow for three weeks until colonies formed.

Table II. Microarray study results of the 6 novel candidate genes with downregulated expression in breast cancer tissues.

Gene	Valid sample (n)	Ratio <0.2 (n)	Downregulated (%)
C2orf88	54	54	100
CSRNP3	46	45	98
PAMR1	46	42	91
PDLIM3	44	42	95
PPP1R12B	67	63	94
SAMD5	70	66	94

Finally the cells were fixed by 10% formamide and stained with 0.1% crystal violet solution. The number of colonies was counted by Image J software.

Results

Identification of genes frequently downregulated in breast cancer tissues. We previously performed cDNA microarray analyses of 81 breast tumor samples (19). All the tumor cells and normal breast epithelial cells were purified by laser microbeam microdissection. In order to identify novel genes which are commonly downregulated in breast cancer tissues, we screened the cDNA microarray database consisting of 23,040 probes using the following criteria: i) genes for which we were able to obtain expression signal in >50% of total examined samples; ii) genes whose expression ratio (cancer/normal) was <0.2 in more than 90% of informative samples; iii) genes whose association with human carcinogenesis had not been reported to date. Finally, we selected 6 candidate genes, namely *chromosome 2 open reading frame 88 (C2orf88), cysteine-serine-rich nuclear protein 3 (CSRNP3), PAMR1, PDZ and LIM domain 3 (PDLIM3), protein phosphatase 1 regulatory subunit 12B (PPP1R12B)* and *sterile α motif domain containing 5 (SAMD5)* (Fig. 1A, Table II).

Downregulation of PAMR1 in breast cancer cell lines and tissues. We then examined the expression of these genes in 21 breast cancer cell lines by qPCR analyses (Fig. 1B). Human mammary epithelial cells (HMECs) served as a normal control. Among the 6 candidate genes, *PAMR1* expression was reduced in all breast cancer cell lines. To confirm this result, we conducted western blot analysis using conditioned

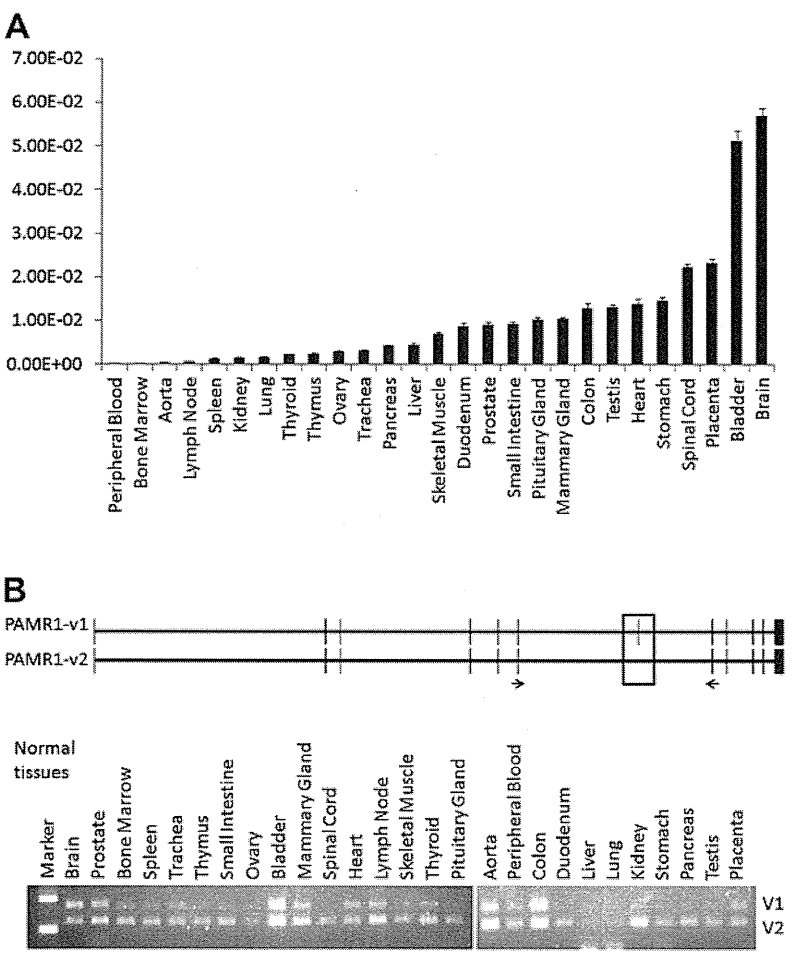

Figure 2. *PAMR1* expression in normal tissues. (A) qPCR results showing the expression of *PAMR1* in 27 normal human tissues. Data represent means ± SD. (B) Genomic structure of *PAMR1* variant 1 (v1) and variant 2 (v2) (upper panel). The exon 7 of variant 1 is absent in variant 2. A pair of primers (arrow) flanking exons 6 and 8 of variant 1 was designed to distinguish each variant. The result of gel electrophoresis indicating the expression of *PAMR1* variants in 27 different normal tissues (lower panel).

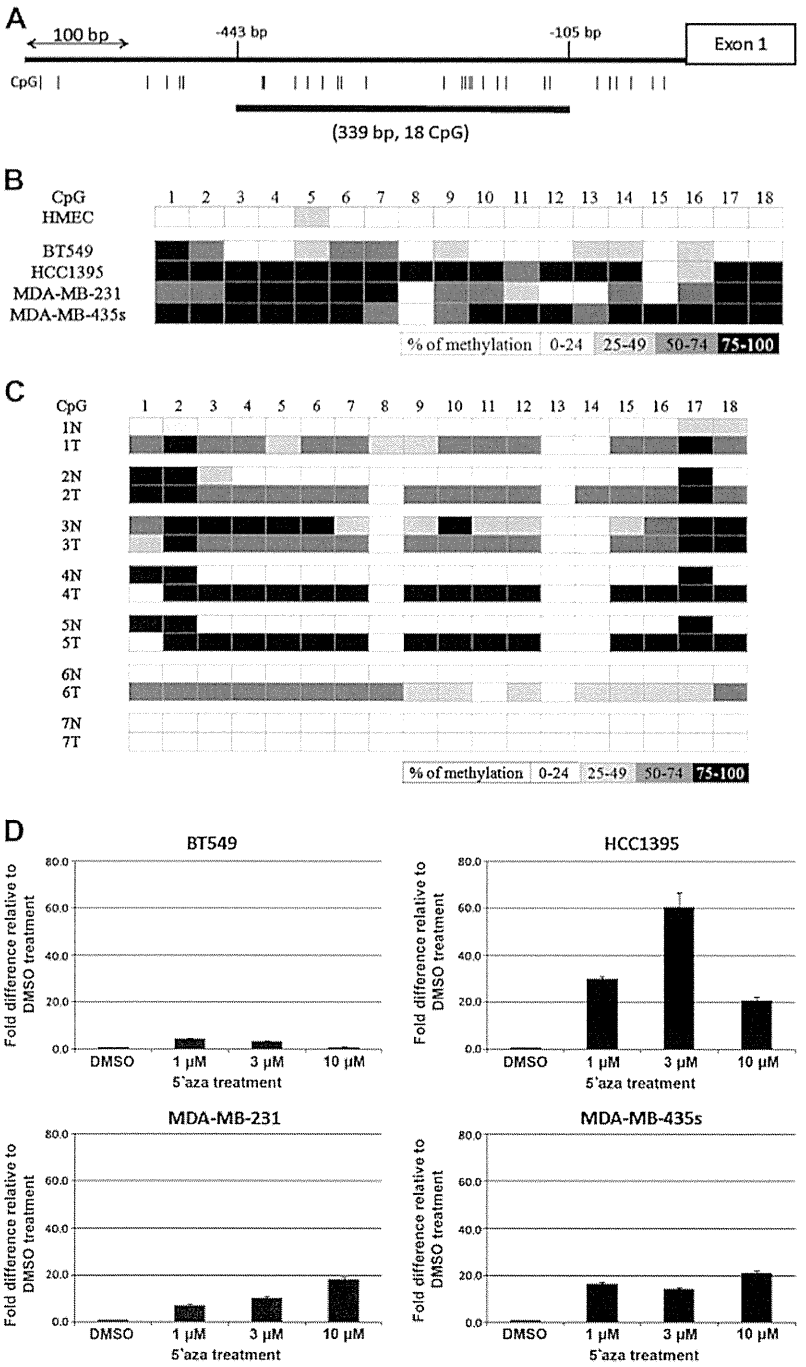

Figure 3. Hypermethylation of the *PAMR1* promoter in breast cancer. (A) *In-silico* prediction of 18 CpG sites located at the promoter region. (B and C) Methylation status of CpG in the *PAMR1* promoter. A 339-bp fragment including 18 CpG sites was analyzed by bisulfite sequencing in HMECs and 4 breast cancer cell lines (B) as well as 7 breast primary tissues and the corresponding normal tissues (C). (D) qPCR analysis of *PAMR1* expression after 3 days of 5-aza-dC treatment. Data represent means ± SD.

medium from the cultured cell lines, as PAMR1 was shown to be a secreted protein (21). As a result, PAMR1 protein was detectable only in the culture medium of HMECs but not in those of the cancer cell lines (Fig. 1C). The conditioned media from HEK293T cells transfected with the plasmid designed to express PAMR1 were used as a positive control. We also examined *PAMR1* expression in 13 breast cancer tissues and 6 normal breast tissues by qPCR analysis. The cancer tissues showed reduced expression of *PAMR1*, concordant with the result of the cDNA microarray analysis (Fig. 1D).

Expression of PAMR1 in mammary gland. PAMR1 was originally identified as a regulator of muscle regeneration. *PAMR1* was found to be downregulated in the muscles of Duchenne muscular dystrophy (DMD) patients and DMD mice (22). Our qPCR analysis revealed that *PAMR1* showed the highest expression in brain tissue and moderate expression in breast and skeletal muscle tissues among the 27 normal human tissues (Fig. 2A), concordant with a previous report (22). Therefore, we hypothesized that *PAMR1* may have unique functions in different tissues. *PAMR1* has two isoforms, and

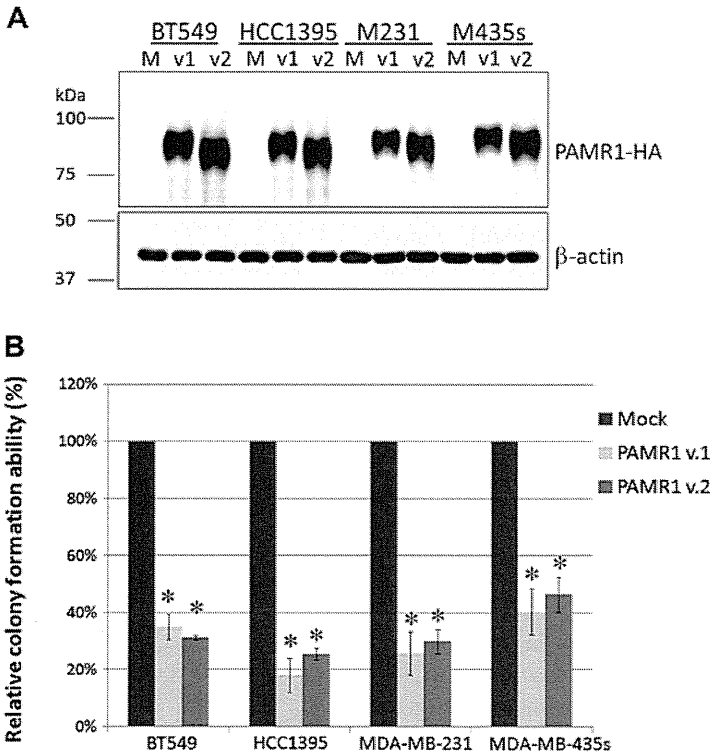

Figure 4. Suppression of breast cancer cell growth by PAMR1. (A) Expression of PAMR1 in cells transfected with the plasmid encoding HA-tagged PAMR1. (B) Relative number of colonies in cells transfected with the plasmid encoding PAMR1 or mock. Data represent means ± SD. *P<0.05 by Student's t-test.

variant 2 which lacks exon 7 (51 bp) encodes a 17-amino acid shorter protein compared with variant 1. To investigate expression of the two isoforms in each tissue, we designed a pair of primers flanking exons 6 and 8. After PCR amplification and gel electrophoresis, DNA fragments corresponding to variant 1 and variant 2 showed similar intensity in the brain, prostate, bladder, heart, colon and placenta, while the intensity of the DNA fragment corresponding to variant 2 was dominant in the other tissues including the mammary gland (Fig. 2B).

Promoter hypermethylation of PAMR1 in breast cancer tissues and cell lines. To further investigate the molecular mechanism of *PAMR1* inactivation in breast cancer tissues, we sequenced all exons of *PAMR1* in 21 breast cancer cell lines. However, we did not identify any mutations in our tested samples. We, then, considered whether epigenetic inactivation could cause *PAMR1* downregulation. Although we were not able to identify any CpG island within the *PAMR1* locus including a 10-kb region encompassing its 5' flanking region by *in-silico* analysis (23), a CpG-rich region was found within -443 to -105 bp of the *PAMR1* promoter region (Fig. 3A). From the result of the bisulfite treated DNA sequencing analysis, hypermethylation was found in 3 cancer cell lines, namely HCC1395, MDA-MB-231, and MDA-MB-435s among the 4 cancer cell lines examined. Moreover, the *PAMR1* promoter was also found to be moderately methylated in the BT549 cancer cells but not in normal HMECs (Fig. 3B). We, then, analyzed 7 pairs of normal and tumor tissues from breast cancer patients and found tumor-specific promoter hypermethylation in 5/7 tumor samples (Fig. 3C).

We treated the breast cancer cell lines with demethylating agent 5-aza-2' deoxycytidine (5-aza-dC) and examined *PAMR1* expression by qPCR analysis. The expression of *PAMR1* was recovered after drug treatment by 4.2-, 62.7-, 18.1- and 20.8-fold in the BT549, HCC1395, MDA-MB-231 and MDA-MB-435s cells, respectively (Fig. 3D). The expression of *PAMR1* in the BT549 cells showed the least degree of restoration compared to the other cell lines, concordant with the low degree of DNA methylation in the BT549 cells (Fig. 3B). Taken together, promoter hypermethylation is one of the mechanisms contributing to the inactivation of *PAMR1* in both breast cancer cell lines and tumor tissues.

Suppression of tumor cell growth by ectopic expression of PAMR1. To investigate the role of *PAMR1* in breast carcinogenesis, we constructed plasmids expressing variant 1 or variant 2 of PAMR1. We confirmed the expression of PAMR1 protein in all cancer cell lines examined (Fig. 4A). We next conducted colony formation assays and observed a significant decrease in colony number (18-46%) for all *PAMR1*-introduced cells (Fig. 4B), indicating the growth-suppressive function of PAMR1.

Discussion

In the present study, we identified *PAMR1* as a putative breast cancer tumor suppressor by a screening of the gene expression profiling of 81 breast cancer tissues. Although we did not find mutations of *PAMR1* in 21 breast cancer cell lines, promoter hypermethylation was frequently observed in both breast cancer tissues and cell lines. The *PAMR1* gene is located at

chromosome 11p13, which is frequently lost in breast cancer samples (20.8-58.3%) (24-26). Therefore, both genetic and epigenetic inactivation would contribute to the downregulation of *PAMR1* in breast cancer.

PAMR1 was first identified as a gene which was downregulated in myoblastic cells isolated from DMD mice. The expression of *PAMR1* was induced in gastrocnemius muscle cells after crush injury, reaching the highest expression on day 4 and was reduced to a normal level on day 14. *PAMR1* induction was only observed in the regenerating muscle fibers by *in situ* hybridization but not in normal muscle cells. Thus, PAMR1 is considered to be involved in the regeneration of skeletal muscles (22). *PAMR1* was expressed in various tissues such as skeletal muscle, brain, and mammary gland. Moreover, our microarray analyses indicated that *PAMR1* expression was reduced in several types of cancers including breast, bladder, liver cancers and osteosarcoma (data not shown). Therefore, *PAMR1* may have as yet unidentified roles other than muscle regeneration.

Although the molecular mechanism whereby *PAMR1* suppresses tumor cell growth has not yet been clarified, PAMR1 contains putative signal peptides at the N-terminal, a CUB domain, two EGF domains, two Sushi domains and an inactive trypsin-like serine protease domain. The secreted *signal peptide CUB-EGF domain-containing protein 2 (SCUBE2)* which contains CUB and EGF domains was shown to suppress breast cancer cell growth (27). Functional domain analysis revealed that the CUB domain bound to bone morphogenetic protein (BMP) and antagonized BMP signaling to suppress cell differentiation and proliferation. Moreover, the EGF-like repeats of SCUBE2 interact with E-cadherin to inhibit the β-catenin pathway (27-29). Since overexpression of *PAMR1* in breast cancer cell lines significantly suppressed cancer cell growth, secreted PAMR1 might exert a tumor-suppressive function by antagonizing growth signals through the interaction with growth factors or their receptors.

In conclusion, our study demonstrated that *PAMR1* may be a novel TSG for breast cancer. We provide evidence that promoter hypermethylation plays an important role in *PAMR1* inactivation during breast carcinogenesis. Although further functional studies and pathway analyses are necessary, identification of its downstream pathway would lead to the development of novel breast cancer therapy by using recombinant soluble PAMR1 protein.

References

1. Jemal A, Bray F, Center MM, Ferlay J, Ward E and Forman D: Global cancer statistics. CA Cancer J Clin 61: 69-90, 2011.
2. Campo E: Whole genome profiling and other high throughput technologies in lymphoid neoplasms - current contributions and future hopes. Mod Pathol 26: S97-S110, 2013.
3. Ajiro M, Katagiri T, Ueda K, *et al*: Involvement of RQCD1 overexpression, a novel cancer-testis antigen, in the Akt pathway in breast cancer cells. Int J Oncol 35: 673-681, 2009.
4. Kim JW, Fukukawa C, Ueda K, Nishidate T, Katagiri T and Nakamura Y: Involvement of C12orf32 overexpression in breast carcinogenesis. Int J Oncol 37: 861-867, 2010.
5. Shimo A, Nishidate T, Ohta T, Fukuda M, Nakamura Y and Katagiri T: Elevated expression of protein regulator of cytokinesis 1, involved in the growth of breast cancer cells. Cancer Sci 98: 174-181, 2007.
6. Shimo A, Tanikawa C, Nishidate T, *et al*: Involvement of kinesin family member 2C/mitotic centromere-associated kinesin overexpression in mammary carcinogenesis. Cancer Sci 99: 62-70, 2008.
7. Kim JW, Akiyama M, Park JH, *et al*: Activation of an estrogen/estrogen receptor signaling by BIG3 through its inhibitory effect on nuclear transport of PHB2/REA in breast cancer. Cancer Sci 100: 1468-1478, 2009.
8. Yoshimaru T, Komatsu M, Matsuo T, *et al*: Targeting BIG3-PHB2 interaction to overcome tamoxifen resistance in breast cancer cells. Nat Commun 4: 2443, 2013.
9. Agrawal A, Murphy RF and Agrawal DK: DNA methylation in breast and colorectal cancers. Mod Pathol 20: 711-721, 2007.
10. Yu Z, Xiao Q, Zhao L, *et al*: DNA methyltransferase 1/3a overexpression in sporadic breast cancer is associated with reduced expression of estrogen receptor-alpha/breast cancer susceptibility gene 1 and poor prognosis. Mol Carcinog: Jan 25, 2014 (Epub ahead of print). doi: 10.1002/mc.22133.
11. Radpour R, Kohler C, Haghighi MM, Fan AX, Holzgreve W and Zhong XY: Methylation profiles of 22 candidate genes in breast cancer using high-throughput MALDI-TOF mass array. Oncogene 28: 2969-2978, 2009.
12. Niwa Y, Oyama T and Nakajima T: BRCA1 expression status in relation to DNA methylation of the BRCA1 promoter region in sporadic breast cancers. Jpn J Cancer Res 91: 519-526, 2000.
13. Barekati Z, Radpour R, Lu Q, *et al*: Methylation signature of lymph node metastases in breast cancer patients. BMC Cancer 12: 244, 2012.
14. Van De Voorde L, Speeckaert R, Van Gestel D, *et al*: DNA methylation-based biomarkers in serum of patients with breast cancer. Mutat Res 751: 304-325, 2012.
15. Medina-Franco JL and Caulfield T: Advances in the computational development of DNA methyltransferase inhibitors. Drug Discov Today 16: 418-425, 2011.
16. Schrump DS, Fischette MR, Nguyen DM, *et al*: Phase I study of decitabine-mediated gene expression in patients with cancers involving the lungs, esophagus, or pleura. Clin Cancer Res 12: 5777-5785, 2006.
17. Issa JP, Kantarjian HM and Kirkpatrick P: Azacitidine. Nature Rev Drug Discov 4: 275-276, 2005.
18. Lin J, Deng Z, Tanikawa C, *et al*: Downregulation of the tumor suppressor HSPB7, involved in the p53 pathway, in renal cell carcinoma by hypermethylation. Int J Oncol 44: 1490-1498, 2014.
19. Nishidate T, Katagiri T, Lin ML, *et al*: Genome-wide gene-expression profiles of breast-cancer cells purified with laser microbeam microdissection: identification of genes associated with progression and metastasis. Int J Oncol 25: 797-819, 2004.
20. Tanikawa C, Furukawa Y, Yoshida N, Arakawa H, Nakamura Y and Matsuda K: XEDAR as a putative colorectal tumor suppressor that mediates p53-regulated anoikis pathway. Oncogene 28: 3081-3092, 2009.
21. Clark HF, Gurney AL, Abaya E, *et al*: The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: a bioinformatics assessment. Genome Res 13: 2265-2270, 2003.
22. Nakayama Y, Nara N, Kawakita Y, *et al*: Cloning of cDNA encoding a regeneration-associated muscle protease whose expression is attenuated in cell lines derived from Duchenne muscular dystrophy patients. Am J Pathol 164: 1773-1782, 2004.
23. Li LC and Dahiya R: MethPrimer: designing primers for methylation PCRs. Bioinformatics 18: 1427-1431, 2002.
24. Gao Y, Niu Y, Wang X, *et al*: Chromosome aberrations associated with centrosome defects: a study of comparative genomic hybridization in breast cancer. Hum Pathol 42: 1693-1701, 2011.
25. Hawthorn L, Luce J, Stein L and Rothschild J: Integration of transcript expression, copy number and LOH analysis of infiltrating ductal carcinoma of the breast. BMC Cancer 10: 460, 2010.
26. Huang Y, Bove B, Wu Y, *et al*: Microsatellite instability during the immortalization and transformation of human breast epithelial cells in vitro. Mol Carcinog 24: 118-127, 1999.
27. Cheng CJ, Lin YC, Tsai MT, *et al*: SCUBE2 suppresses breast tumor cell proliferation and confers a favorable prognosis in invasive breast cancer. Cancer Res 69: 3634-3641, 2009.
28. Lin YC, Lee YC, Li LH, Cheng CJ and Yang RB: Tumor suppressor SCUBE2 inhibits breast-cancer cell migration and invasion through the reversal of epithelial-mesenchymal transition. J Cell Sci 127: 85-100, 2014.
29. Lin YC, Chen CC, Cheng CJ and Yang RB: Domain and functional analysis of a novel breast tumor suppressor protein, SCUBE2. J Biol Chem 286: 27039-27047, 2011.

nature genetics

Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1

Qiuyin Cai[1], Ben Zhang[1], Hyuna Sung[2,3], Siew-Kee Low[4], Sun-Seog Kweon[5,6], Wei Lu[7], Jiajun Shi[1], Jirong Long[1], Wanqing Wen[1], Ji-Yeob Choi[2,8], Dong-Young Noh[8,9], Chen-Yang Shen[10–12], Keitaro Matsuo[13], Soo-Hwang Teo[14,15], Mi Kyung Kim[16], Ui Soon Khoo[17], Motoki Iwasaki[18], Mikael Hartman[19–21], Atsushi Takahashi[4], Kyota Ashikawa[22], Koichi Matsuda[23], Min-Ho Shin[5], Min Ho Park[24], Ying Zheng[7], Yong-Bing Xiang[25], Bu-Tian Ji[26], Sue K Park[2,8,27], Pei-Ei Wu[10,11], Chia-Ni Hsiung[10,11], Hidemi Ito[28], Yoshio Kasuga[29], Peter Kang[14], Shivaani Mariapun[14,15], Sei Hyun Ahn[30], Han Sung Kang[31], Kelvin Y K Chan[17,32], Ellen P S Man[17], Hiroji Iwata[33], Shoichiro Tsugane[34], Hui Miao[19,20], Jiemin Liao[35,36], Yusuke Nakamura[37], Michiaki Kubo[22], DRIVE GAME-ON Consortium[38], Ryan J Delahanty[1], Yanfeng Zhang[1], Bingshan Li[39], Chun Li[40], Yu-Tang Gao[25], Xiao-Ou Shu[1], Daehee Kang[2,8,27] & Wei Zheng[1]

In a three-stage genome-wide association study among East Asian women including 22,780 cases and 24,181 controls, we identified 3 genetic loci newly associated with breast cancer risk, including rs4951011 at 1q32.1 (in intron 2 of the *ZC3H11A* gene; $P = 8.82 \times 10^{-9}$), rs10474352 at 5q14.3 (near the *ARRDC3* gene; $P = 1.67 \times 10^{-9}$) and rs2290203 at 15q26.1 (in intron 14 of the *PRC1* gene; $P = 4.25 \times 10^{-8}$). We replicated these associations in 16,003 cases and 41,335 controls of European ancestry ($P = 0.030$, 0.004 and 0.010, respectively). Data from the ENCODE Project suggest that variants rs4951011 and rs10474352 might be located in an enhancer region and transcription factor binding sites, respectively. This study provides additional insights into the genetics and biology of breast cancer.

Breast cancer is one of the most common malignancies among women worldwide. Genetic factors have a substantial role in breast cancer etiology[1,2]. Thus far, genome-wide association studies (GWAS) have identified approximately 75 genetic loci associated with breast cancer risk[2–5]. With the exception of the studies we have conducted among East Asian women[6–9] and one study conducted among women of African ancestry[10], all other published GWAS have been conducted among women of European ancestry. Genetic risk variants identified thus far from GWAS explain only about 10% of familial risk for breast cancer in East Asian women[3]. Given the differences in genetic architecture and environmental exposures for women of European and East Asian ancestry, additional GWAS need to be conducted among East Asian women to study the genetic basis of breast cancer risk.

The current study was conducted as part of the Asia Breast Cancer Consortium (ABCC) to search for additional susceptibility loci for breast cancer. Included in this study are data obtained from 22,780 breast cancer cases and 24,181 controls who were recruited in 14 studies conducted in multiple Asian countries (**Supplementary Table 1**). The discovery stage (stage 1) included 2 GWAS in which 5,285 Chinese women (SBCGS-1) and 4,777 Korean women (SeBCS1) were scanned primarily using Affymetrix Genome-Wide Human SNP Array 6.0, which consists of 906,602 SNPs. After applying quality control filters described previously[6,9,11], 5,152 Chinese women (2,867 cases and 2,285 controls; 677,157 SNPs) and 4,298 Korean women (2,246 cases and 2,052 controls; 555,117 SNPs) remained in the current analysis. Imputation was conducted for each study following the MACH algorithm[12] using HapMap 2 release 22 CHB (Han Chinese in Beijing, China) and JPT (Japanese in Tokyo, Japan) data (2,416,663 SNPs) as the reference. Only SNPs with a high imputation quality score (RSQR ≥ 0.50) were analyzed for associations with breast cancer risk. In the analyses of data from Chinese and Korean women, a total of 1,930,412 and 1,907,146 SNPs, respectively, were included. A meta-analysis of these GWAS data was conducted using a fixed-effects, inverse variance meta-analysis with the METAL program[13]. There was little evidence of inflation in the association test statistics for the studies included in stage 1 (genomic inflation factors (λ): $\lambda = 1.0426$ for SBCGS-1, $\lambda = 1.0431$ for SeBCS1 and $\lambda = 1.0499$ for both studies combined; **Supplementary Fig. 1**). When scaled to a study of 1,000 cases and 1,000 controls, $\lambda_{1,000}$ values were 1.02, 1.02 and 1.01, respectively.

To select SNPs for stage 2 replication, we used the following criteria: (i) association $P < 0.05$ in the stage 1 meta-analysis results; (ii) the

A full list of affiliations appears at the end of the paper.

Received 4 November 2013; accepted 27 June 2014; published online 20 July 2014; doi:10.1038/ng.3041

Table 1 Associations of breast cancer risk with newly identified risk variants

SNP (alleles[a])	Frequency[b]	Locus (position[c])	Closest gene (annotation)	Stage	Per-allele association OR (95% CI)[d]	Per-allele association P[e]	P for heterogeneity[f]
rs4951011	0.282	1q32.1	*ZC3H11A*	Stage 1	1.09 (1.02–1.17)	0.007	
(**G**/A)		(202,032,954)	(intron 2)	Stage 2	1.10 (1.02–1.18)	0.011	
				Stage 3	1.08 (1.05–1.12)	1.02×10^{-5}	
				Combined	1.09 (1.06–1.12)	8.82×10^{-9}	0.98
rs10474352	0.482	5q14.3	*ARRDC3*	Stage 1	1.09 (1.03–1.17)	0.006	
(**C**/T)		(90,767,981)	(intergenic)	Stage 2	1.12 (1.05–1.20)	7.06×10^{-4}	
				Stage 3	1.08 (1.04–1.12)	1.92×10^{-5}	
				Combined	1.09 (1.06–1.12)	1.67×10^{-9}	0.50
rs2290203	0.504	15q26.1	*PRC1*	Stage 1	1.08 (1.02–1.14)	0.012	
(**G**/A)		(89,313,071)	(intron 14)	Stage 2	1.19 (1.10–1.30)	4.97×10^{-5}	
				Stage 3	1.06 (1.03–1.10)	2.45×10^{-4}	
				Combined	1.08 (1.05–1.11)	4.25×10^{-8}	0.06

OR, odds ratio; CI, confidence interval.
[a]Risk/reference allele; the risk allele is shown in bold. [b]Risk allele frequency in controls from all three stages combined. [c]Chromosome position (bp) based on NCBI Human Genome Build 36. [d]Per-allele OR (95% CI) was adjusted for age and the principal components in each study; summary OR (95% CI) was obtained using fixed-effect meta-analysis in each stage. [e]Derived from a weighted z statistic–based meta-analysis. [f]P for heterogeneity across studies in all stages was calculated using Cochran's Q test.

same direction of association in both stage 1 studies; (iii) no heterogeneity observed between the two stage 1 studies ($P > 0.05$ and $I^2 < 25\%$); (iv) an imputation score of RSQR > 0.5 in both stage 1 studies; (v) a minor allele frequency (MAF) of >0.05 in both stage 1 studies; and (vi) lack of strong LD ($r^2 < 0.5$) with any of the known breast cancer susceptibility loci or any SNPs for which we had previously found evidence of association[3,6–9]. For SNPs that met the above criteria but were in LD ($r^2 > 0.5$) with each other, we selected only one SNP for replication. A total of 4,598 SNPs were selected, and assays for 4,071 SNPs were successfully designed using Illumina Infinium assays as part of a large-scale genotyping effort. Of the 4,071 SNPs, 3,850 were successfully genotyped in an independent set of 3,944 cases and 3,980 controls selected from the Shanghai studies (SBCGS-2). After quality control exclusions, 3,678 SNPs were included in the analyses of 3,472 cases and 3,595 controls.

For stage 3, the top 50 SNPs were selected for further replication in an independent set of 14,195 cases and 16,249 controls from 10 studies participating in ABCC on the basis of the following criteria: (i) association $P < 0.005$ in the meta-analysis of stage 1 and 2 data and (ii) the same direction of association in both stages 1 and 2. Of the 50 SNPs evaluated in stage 3, 11 showed an association with breast cancer risk at $P < 0.05$ (**Supplementary Table 2**). Combined analyses of data from all three stages identified three SNPs that were associated with breast cancer risk at the genome-wide significance level ($P < 5.0 \times 10^{-8}$): rs4951011 at 1q32.1, odds ratio (OR) = 1.09, $P = 8.82 \times 10^{-9}$; rs10474352 at 5q14.3, OR = 1.09, $P = 1.67 \times 10^{-9}$; and rs2290203 at 15q26.1, OR = 1.08, $P = 4.25 \times 10^{-8}$ (**Table 1**). The association between breast cancer risk and each of these three SNPs was consistent across the studies included in ABCC (**Fig. 1**), and none of the tests for heterogeneity gave statistically significant results ($P > 0.05$) (**Table 1**). No significant heterogeneity was found for the association of these three SNPs with breast cancer risk among Chinese, Japanese or Korean women (**Supplementary Table 3**). One additional SNP showed an association with breast cancer risk with a P value near the conventional GWAS significance level (rs11082321 at 18q11.2, OR = 1.08, $P = 6.77 \times 10^{-7}$) (**Supplementary Table 2**).

The associations of SNPs rs10474352 and rs2290203 appeared to be stronger for estrogen receptor (ER)-positive breast cancer than for ER-negative breast cancer, and the heterogeneity test was of borderline significance for rs10474352 ($P = 0.085$) (**Supplementary Table 4**). The associations of rs4951011 with breast cancer risk were similar for ER-positive and ER-negative breast cancer.

We evaluated the three newly identified risk variants for associations with breast cancer risk in women of European ancestry using data from 16,003 cases and 41,335 controls from 12 breast cancer GWAS and included in the DRIVE GAME-ON Consortium. SNPs rs4951011, rs10474352 and rs2290203 were all associated with breast cancer risk in women of European ancestry at $P < 0.05$ with the same direction of association as observed in East Asian women (**Supplementary Table 5**). However, the strength of the associations was weaker in women of European ancestry than in women of East Asian ancestry, and the frequencies of the risk alleles were quite different in these two populations.

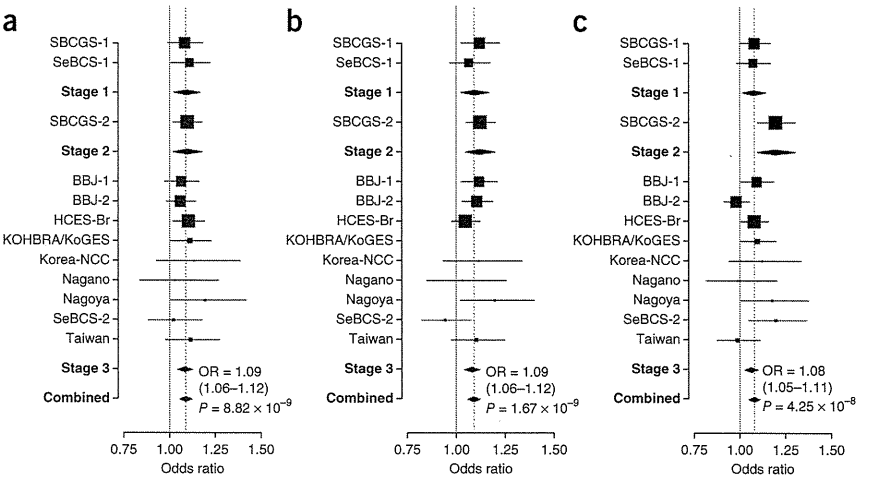

Figure 1 Forest plots for risk variants in the three newly identified breast cancer risk loci by study site and stage. Per-allele OR estimates are presented. The size of the box is proportional to the number of cases and controls in each study. (**a**) rs4951011. (**b**) rs10474352. (**c**) rs2290203.

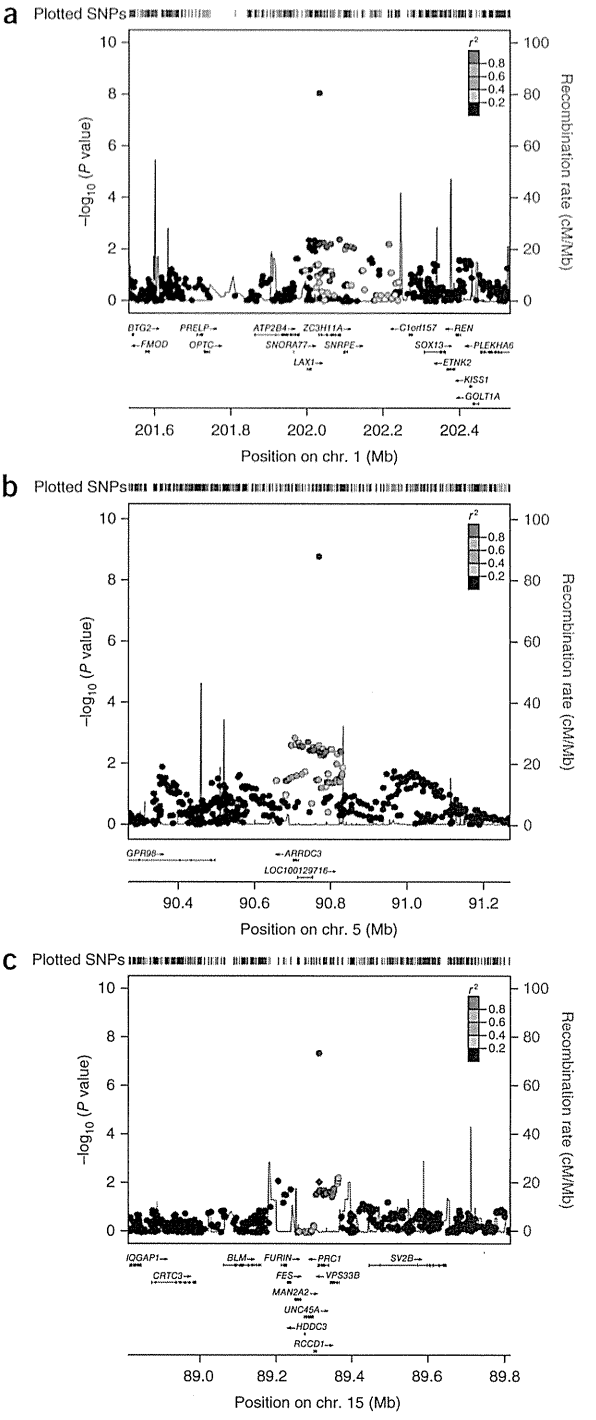

Figure 2 Regional plots of association results for the three newly identified risk loci for breast cancer. For each plot, the $-\log_{10} P$ values (*y* axis) of the SNPs are shown according to their chromosomal positions (*x* axis) in NCBI Build 36. The color of the SNPs represents their LD (r^2; HapMap Asian) with the index SNP at each locus. With the exception of the index SNPs, which are shown as purple diamonds for stage 1 and purple circles for the meta-analyses of all studies, data shown for all other SNPs are from stage 1 only. (**a**) rs4951011 (1q32.1). (**b**) rs10474352 (5q14.3). (**c**) rs2290203 (15q26.1).

We evaluated and annotated putative functional variants and candidate genes in each of the three newly identified loci using data from the Encyclopedia of DNA Elements (ENCODE)[14] Project, The Cancer Genome Atlas (TCGA) breast cancer project[15] and expression quantitative trait locus (eQTL) databases[16] as well as RegulomeDB[17] and HaploReg v2 (ref. 18). We summarize the results below for each locus.

SNP rs10474352 is located on 5q14.3, 53,078 bp upstream of the *ARRDC3* gene (**Fig. 2b**). The *ARRDC3* gene is a member of the arrestin gene family and is suspected of having a role in breast cancer development. A gene cluster at 5q11-q23 that includes *ARRDC3* was found to be deleted in 17% of breast cancer tumor tissues[19]. Upregulation of the *ARRDC3* gene in a breast cancer cell line has been shown to repress cell proliferation, migration, invasion and *in vivo* tumorigenesis[20]. We evaluated *ARRDC3* gene expression in 87 breast cancer cases included in TCGA. The expression level of the *ARRDC3* gene was significantly lower in tumor tissue than in adjacent normal tissue ($P = 1.88 \times 10^{-18}$) (**Supplementary Table 6**). This finding is consistent with a previous study showing that expression levels of the *ARRDC3* gene were lower in breast tumor tissue in comparison to normal tissue and in metastatic tumor tissue in comparison to primary tumor tissue[20]. Furthermore, lower *ARRDC3* expression in tumor tissue has been associated with poorer disease-free survival in individuals with breast cancer[20]. A search of RegulomeDB[17] and HaploReg[18] indicated that rs10474352 might be located in predicted AP-1and VDR motifs (**Supplementary Table 7**), suggesting a potential regulatory role. We evaluated whether SNPs in this locus act as *cis*-eQTLs for other genes by analyzing TCGA breast cancer data. Our analysis showed no evidence that this SNP or its correlated SNPs were *cis*-eQTLs for any genes in this locus. Recently, a SNP located ~596 kb upstream of the *ARRDC3* gene, rs421379, was found to be associated with prognosis for early-onset breast cancer in a GWAS[21]. However, rs421379 is not in LD with rs10474352 ($r^2 = 0$ in both ASN (Asian) and CEU (Utah residents of Northern and Western European ancestry) data), the SNP in close proximity to *ARRDC3* that was identified in our study. Furthermore, in our study, rs421379 had a low MAF (0.03–0.04) and was not associated with breast cancer risk ($P = 0.2484$ in stage 1).

SNP rs2290203 is located in intron 14 of the *PRC1* gene (NM_003981) at 15q26.1 (**Fig. 2c**). This gene encodes the protein regulator of cytokinesis 1 (PRC1) protein, which is involved in cytokinesis and is a substrate for several cyclin-dependent kinases[22]. The *PRC1* gene is downregulated by p53 in MCF-7 and T47D breast cancer cells[23]. Interestingly, the *PRC1* gene is included in a five-gene expression signature that predicted prognosis for individuals with breast cancer in a recent study[24]. The expression level of the *PRC1* gene was significantly higher in tumor tissue than in adjacent normal tissue ($P = 4.62 \times 10^{-30}$) among breast cancer cases included in TCGA (**Supplementary Table 6**). Our *cis*-eQTL analysis using TCGA data showed no association of rs2290203 with *PRC1* gene expression but did show a correlation with expression of the *RCCD1* gene, which is 5,712 bp upstream of rs2290203. An eQTL analysis of human monocytes has also indicated that rs2290203 is a *cis*-eQTL for the *RCCD1*gene[16]. In our study, the rs2290203 risk allele (G) was associated with lower *RCCD1* expression in both tumor ($P = 3.6 \times 10^{-4}$) and adjacent normal tissue ($P = 0.007$) (**Supplementary Fig. 2**). However, these associations were no longer statistically significant after adjusting for the most significant *cis*-eQTL SNPs (rs4544218 for tumor tissue; rs59278520 for normal tissue), which are in strong LD with rs2290203 (**Supplementary Fig. 3**). Variant rs4544218 was not associated with breast cancer risk ($P = 0.8925$), and rs59278520 was marginally associated with breast