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

Qiuyin Cai1, Ben Zhang1, Hyuna Sung2,3, Siew-Kee Low4, Sun-Seog Kweon5,6, Wei Lu7, Jiajun Shi1, Jirong Long1, Wanqing Wen1, Ji-Yeob Choi2,8, Dong-Young Noh8,9, Chen-Yang Shen10–12, Keitaro Matsuo13, Soo-Hwang Teo14,15, Mi Kyung Kim16, Ui Soon Khoo17, Motoki Iwasaki18, Mikael Hartman19–21, Atsushi Takahashi14, Kyota Ashikawa22, Koichi Matsuda23, Min-Ho Shin5, Min Ho Park24, Ying Zheng7, Yong-Bing Xiang25, Bu-Tian Ji26, Sue K Park2,8,27, Pei-Ei Wu10,11, Chia-Ni Hsiung10,11, Hideki Ito28, Yoshio Kasuga29, Peter Kang14, Shivaani Mariapun14,15, Sei Hyun Ahn30, Han Sung Kang31, Kelvin Y K Chan17,32, Ellen P S Man17, Hiroji Iwata33, Shoichiro Tsugane34, Hui Miao19,20, Jiemin Liao35,36, Yusuke Nakamura37, Michiaki Kubo22, DRIVE GAME-ON Consortium38, Ryan J Delahanty1, Yanfeng Zhang1, Bingshan Li39, Chun Li40, Yu-Tang Gao25, Xiao-Ou Shu1, Daehee Kang2,8,27 & Wei Zheng1

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 ARRD3 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 etiology1,2. Thus far, genome-wide association studies (GWAS) have identified approximately 75 genetic loci associated 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. 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 previously6,9,11, 5,152 Chinese women (2,867 cases and 2,052 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 algorithm12 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 program13. There was little evidence of inflation in the association test statistics for the studies included in stage 1 (genomic inflation factors (\( \lambda \)); \( \lambda = 1.0426 \) 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, \( \lambda_{0.05} \) 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
same direction of association in both stage 1 studies; (iii) no heterogeneity observed between the two stage 1 studies (P > 0.05 and I² < 25%); (iv) an imputation score of R2Q > 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 (r² < 0.5) with any of the known breast cancer susceptibility loci or any SNPs for which we had previously found evidence of association. For SNPs that met the above criteria but were in LD (r² > 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 included in the analyses in all stages across studies in stage 1 and 2.

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 ABCG on the basis of the following criteria: (i) association P < 0.05 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 × 10⁻⁸): rs4951011 at 1q32.1, odds ratio (OR) = 1.09, P = 8.82 × 10⁻⁹; rs10474352 at 5q14.3, OR = 1.09, P = 1.67 × 10⁻⁵; and rs2290203 at 15q26.1, OR = 1.08, P = 4.25 × 10⁻⁸ (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 at a P value near the conventional GWAS significance level (rs1082321 at 18q11.2, OR = 1.08, P = 6.77 × 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.

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

| SNP (alleles) | Frequency | Locus (position) | Closest gene (annotation) | Stage | OR (95% CI) | P | for heterogeneity |
|--------------|-----------|-----------------|---------------------------|-------|-------------|---|------------------|
| rs4951011 (G/A) | 0.282  | 1q32.1 (202,032,954) | ZC3H11A (intron 2) | Stage 1 | 1.09 (1.02–1.17) | 0.007 | |
| rs10474352 (C/T) | 0.482  | 5q14.3 (90,767,981) | ARRDC3 (intergenic) | Stage 1 | 1.09 (1.03–1.17) | 0.006 | |
| rs2290203 (G/A) | 0.504  | 15q26.1 (89,313,071) | PRC1 (intron 14) | Stage 1 | 1.08 (1.02–1.14) | 0.012 | |
|               |          |                 |                           | Stage 2 | 1.10 (1.02–1.17) | 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 |
|               |          |                 |                           | Stage 1 | 1.09 (1.03–1.17) | 0.006 | |
|               |          |                 |                           | 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 |
|               |          |                 |                           | Stage 1 | 1.08 (1.02–1.14) | 0.012 | |
|               |          |                 |                           | 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.

*Risk/reference allele: the risk allele is shown in bold. *Risk allele frequency in controls from all three stages combined. *Chromosome position (bp) based on NCBI Human Genome Build 36.

*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. *Derived from a weighted z statistic–based meta-analysis. *P for heterogeneity across studies in all stages was calculated using Cochran’s Q test.

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.
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 project15 and expression quantitative trait locus (eQTL) databases16 as well as RegulomeDB17 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 tissues19. Upregulation of the ARRDC3 gene in a breast cancer cell line has been shown to repress cell proliferation, migration, invasion and in vivo tumorigenesis20. 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 × 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 tissue20. Furthermore, lower ARRDC3 expression in tumor tissue has been associated with poorer disease-free survival in individuals with breast cancer20. A search of RegulomeDB17 and HaploReg18 indicated that rs10474352 might be located in predicted AP-1 and 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 GWAS21. However, rs421379 is not in LD with rs10474352 (r² = 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 kinases22. The PRC1 gene is downregulated by p53 in MCF-7 and T47D breast cancer cells23. Interestingly, the PRC1 gene is included in a five-gene expression signature that predicted prognosis for individuals with breast cancer in a recent study24. The expression level of the PRC1 gene was significantly higher in tumor tissue than in adjacent normal tissue (P = 4.62 × 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 gene16. In our study, the rs2290203 risk allele (G) was associated with lower RCCD1 expression in both tumor (P = 3.6 × 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 cancer risk (P = 0.007).
cancer risk ($P = 0.0518$) in the SBCGS-1 stage 1 samples. The function of the **RCCD1** gene is unknown. SNP rs4951011 is located in intron 2 of the **ZC3H11A** gene (encoding zinc finger CCCH domain-containing protein 11A) (NM_014827) at 1q32.1 (Fig. 2a) and the 5’ UTR of the **ZBED6** gene (encoding zinc finger BED domain-containing protein 6) (NM_001174108) (not shown). The ZBED6 protein has recently been recognized as a novel transcription factor in placental mammals25. The function of the **ZC3H11A** gene is not clear. ChromHMM annotation using human mammary epithelial cell (HMEC) data from ENCODE suggests that rs4951011 might be located in a strong enhancer region marked by peaks of several active histone methylation modifications (monomethylation of histone H3 at lysine 4 (H3K4me1), trimethylation of histone H3 at lysine 4 (H3K4me3), acetylation of histone H3 at lysine 9 (H3K9ac) and acetylation of histone H3 at lysine 27 (H3K27ac)). A search of RegulomeDB and HaploReg indicated that rs4951011 might be located in a predicted HNF1 motif and map to a DNase I hypersensitivity site in the MCF-7 breast cancer cell line (Supplementary Table 7).

Expression levels of the **ZC3H11A** gene were significantly higher in breast tumor tissue than in adjacent normal tissue ($P = 0.0049$) in TCGA data (Supplementary Table 6). Analyses using TCGA data showed no evidence that this SNP or other SNPs correlated with it are cis-eQTLs for any genes in this locus. Recently, SNP rs4245739 in the **MDM4** gene, ~752 kb downstream of rs4951011 ($r^2 = 0$ in both ASN and CEU data), was associated with ER-negative breast cancer risk3. In our study, rs4245739 had a low MAF (0.03–0.05) and was not associated with breast cancer risk ($P = 0.1861$ in stage 1).

In summary, our large GWAS conducted among East Asian women identified three new breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1 and suggested a possible association with a fourth locus at 18q11.2. The associations of these loci with breast cancer risk might be mediated through the regulation of cell growth control, tumor cell migration and invasion, or metastasis. Further studies of possible mechanisms through which these loci and genes are involved in breast tumorigenesis are warranted. Results from this study provide additional insights into the genetics and biology of breast cancer.

**METHODS**

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

**ACKNOWLEDGMENTS**

The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The authors wish to thank the study participants and research staff for their contributions and commitment to this project. We thank R. Courtney, J. Wu, J. He, H. Cai, X. Guo, B. Rammer and K. Kreth for their help with sample preparation and genotyping, statistical and bioinformatics analyses for the project, and editing and preparation of the manuscript at Vanderbilt.

This research was supported in part by US National Institutes of Health grants R01CA124588, R01CA148667 and R37CA070867 (to W.Z.); R01CA092585 and R01CA064277 (to X.-O.S.); R01CA122756 (to Q.C.); and R01CA137013 (to L.S.). US Department of Defense Idea Awards BC011118 (to X.-O.S.) and BC050791 (to Q.C.), and Ingram Professorship and Research Reward funds (to W.Z.). Sample preparation and genotyping assays at Vanderbilt were conducted at the Survey and Biospecimen Shared Resources and the Vanderbilt Microarray Shared Resource, which are supported in part by the Vanderbilt-Ingram Cancer Center (P30CA068485). SeBCS was supported by the BRL (Basic Research Laboratory) program through the National Research Foundation of Korea (to H. Iwata, S.T., H.M., D.-Y.N., C.-Y.S., K. Matsuo, S.-H.T., M.K.K., U.S.K., M.I., M.H., A.T., K.A., I. Soong, A. Chan and T.Y. Leung for their assistance in facilitating recruitment of the breast cancer cases included in the Hong Kong Study. We thank R. Courtney, J. Wu, J. He, H. Cai, X. Guo, B. Rammer and K. Kreth for their help with sample preparation and genotyping, statistical and bioinformatics analyses for the project, and editing and preparation of the manuscript at Vanderbilt.

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.
LETTERS

J. Liao, Y.N., M.K., Y.-T.G., X.-O.S., D.K. and W.Z. contributed to the collection of the data and biological samples for the original studies. All authors have reviewed and approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Mavaddat, N., Antoniou, A.C., Easton, D.F. & Garcia-Closas, M. Genetic susceptibility to breast cancer. Mol. Oncol. 4, 174–191 (2010).
2. Zhang, B., Beeghly-Fadiel, A., Long, J. & Zheng, W. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. Lancet Oncol. 12, 477–488 (2011).
3. Zheng, W. et al. Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 ERBB4 breast cancer cases and 29,979 controls. Hum. Mol. Genet. 22, 2539–2550 (2013).
4. Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat. Genet. 45, 353–361 (2013).
5. Garcia-Closas, M. et al. Genome-wide association studies identify four ER negative–specific breast cancer risk loci. Nat. Genet. 45, 392–398 (2013).
6. Zheng, W. et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat. Genet. 41, 324–328 (2009).
7. Long, J. et al. Identification of a functional genetic variant at 16q12.1 for breast cancer risk: results from the Asia Breast Cancer Consortium. PLoS Genet. 6, e1001002 (2010).
8. Cai, Q. et al. Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium. Hum. Mol. Genet. 20, 4991–4999 (2011).
9. Long, J. et al. Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer. PLoS Genet. 8, e1002532 (2012).
10. Haiman, C.A. et al. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor–negative breast cancer. Nat. Genet. 43, 1210–1214 (2011).
11. Kim, H.C. et al. A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. Breast Cancer Res. 14, R56 (2012).
12. 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).
13. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191 (2010).
14. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012).
15. Genome Registry. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012).
16. Zeller, T. et al. Genetics and beyond—the transcriptome of human monocytocytes and disease susceptibility. PLoS ONE 5, e10693 (2010).
17. Boyle, A.P. et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22, 1790–1797 (2012).
18. Ward, L.D. & Kellis, M. Haplotype: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 40, D930–D934 (2012).
19. Adélaïde, J. et al. Integrated profiling of basal and luminal breast cancers. Cancer Res. 67, 11565–11575 (2007).
20. Draheim, K.M. et al. ARRD35 suppresses breast cancer progression by negatively regulating integrin 8A. Oncogene 29, 5032–5047 (2010).
21. Rafiq, S. et al. Identification of inherited genetic variations influencing prognosis in early-onset breast cancer. Cancer Res. 73, 1883–1891 (2013).
22. Zhu, C. & Jiang, W. Cell cycle–dependent translocation of PRC1 on the spindle by Kif4 is essential for mitosis formation and cytokinesis. Proc. Natl. Acad. Sci. USA 102, 343–348 (2005).
23. Li, C., Lin, M. & Liu, J. Identification of PRC1 as the p53 target gene uncovers a novel function of p53 in the regulation of cytokinesis. Oncogene 23, 9336–9347 (2004).
24. Mustacchi, G. et al. Identification and validation of a new set of five genes for prediction of risk in early breast cancer. Int. J. Mol. Sci. 14, 9686–9702 (2013).
25. Marklund, E. et al. ZBED6, a novel transcription factor derived from a domesticated DNA transposon regulates IGF2 expression and muscle growth. PLOS Biol. 7, e1000256 (2009).

1Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. 2Department of Biomedical Sciences, Seoul National University College of Medicine, Seoul, South Korea. 3Genetic Epidemiology Branch, Division of Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA. 4Laboratory for Statistical Analysis, Center for Integrative Medical Sciences, RIKEN, Yokohama, Japan. 5Department of Preventive Medicine, Chonnam National University Medical School, Gwangju, South Korea. 6Jeonnam Regional Cancer Center, Chonnam National University Hwasun Hospital, Hwasun, South Korea. 7Shanghai Municipal Center for Disease Control and Prevention, Shanghai, China. 8Cancer Research Institute, Seoul National University College of Medicine, Seoul, South Korea. 9Department of Surgery, Seoul National University College of Medicine, Seoul, South Korea. 10Department of Breast Oncology, Aichi Cancer Center Central Hospital, Aichi Cancer Institute, US National Institutes of Health, Bethesda, Maryland, USA. 11Department of Preventive Medicine, Chonnam National Cancer Institute, US National Institutes of Health, US Department of Health and Human Services, Bethesda, Maryland, USA. 12Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, South Korea. 13Department of Epidemiology, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. 14Department of Biostatistics, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. Correspondence should be addressed to Q.C. (qiyin.cai@vanderbilt.edu).
ONLINE METHODS

Study populations. ABCC comprises 22,780 cases and 24,181 controls from 14 studies (Supplementary Table 1), including 15,483 Chinese women, 18,367 Korean women and 13,111 Japanese women. Chinese participants came from 6 studies based in Shanghai (n = 12,219; the Shanghai Breast Cancer Study (SBCS), the Shanghai Breast Cancer Survival Study (SBCSS), the Shanghai Endometrial Cancer Study (SECS; controls only) and the Shanghai Women’s Health Study (SWHS))26–28, Taiwan (n = 2,131)29 and Hong Kong (n = 1,133)30. Korean participants came from 5 studies: the Seoul Breast Cancer Study (SeBCS; n = 6,179)31, the Hwasun Cancer Epidemiology Study-Breast (HCES-B; n = 6,573)31–33, the Korea Genome Epidemiology Study (KoGES; n = 3,209)34, the Korean Hereditary Breast Cancer study (KOHBRA; n = 1,397)35 and the Korean National Cancer Center (n = 1,009). Japanese participants came from three studies: the Biobank Japan Project (BBJ; n = 11,021)36, the Nagoya Study (n = 1,288)37 and the Nagano Breast Cancer Study (n = 802)38 (Supplementary Table 1). Detailed descriptions of these participating studies are presented in the Supplementary Note. The protocols for all participating studies were approved by their relevant institutional review boards, and all participants of the studies provided written informed consent.

We estimated that our study had a statistical power of >80% to identify an association with an OR of 1.09 or greater at P ≤ 5 × 10−8 for SNPs with a MAF as low as 0.05.

Genotyping methods. Stage 1 genotyping. Stage 1 included 2 GWAS in which 5,285 Chinese women and 4,777 Korean women were scanned primarily using Affymetrix Genome-Wide Human SNP Array 6.0. Genotyping protocols for stage 1 have been described elsewhere6–9,11. In the Chinese GWAS (SBCGS-2), the initial 300 samples were genotyped using the Affymetrix GeneChip Mapping 500K Array Set. The remaining 4,985 samples were genotyped using Affymetrix Genome-Wide Human SNP Array 6.0. After quality control exclusions, the final data set included 2,867 cases and 2,285 controls for 677,157 markers. For the Korean GWAS (SeBCS), Affymetrix Genome-Wide Human SNP Array 6.0 Array was also used31. After quality control exclusions, the final data set included 2,246 cases and 2,052 controls for 555,117 markers. Detailed descriptions of quality control and criteria for sample and marker exclusion are presented in the Supplementary Note.

Stage 2 genotyping. Genotyping assays for 3,944 cases and 3,980 controls (SBCGS-2) in stage 2 were completed using Illumina Infinium assays as the add-on content to the Illumina HumanExome BeadChip (see URLs). Genotype calling was carried out using the Illumina GenTrain version 2.0 clustering algorithm in GenomeStudio version 2011.1. Cluster boundaries were determined using study samples. Further quality control procedures were conducted using PLINK (see URLs). Detailed descriptions of quality control and criteria for sample and marker exclusion are presented in the Supplementary Note. Of the 4,598 SNPs selected, assays for 4,071 SNPs were successfully designed using Illumina Infinium assays. A total of 3,850 SNPs were successfully genotyped, and 3,678 SNPs were included in the analyses of 3,472 breast cancer cases and 3,595 controls.

Stage 3 genotyping. Genotyping assays for the 50 SNPs in stage 3 were completed at the Vanderbilt Molecular Epidemiology Laboratory using the iPLEX Sequenom MassArray platform for 19,423 samples from the Taiwan, Hong Kong, HCES-B, KOHBRA/KoGES, SeBCS2, Korea-NCC, Nagoya and Nagano studies. Detailed descriptions of quality control and criteria for sample and marker exclusion are presented in the Supplementary Note. For the BB1 study, the SNP data needed for the study were extracted from either genotyped (n = 8) or imputed (n = 14; mean RSQR = 0.96) data generated using the OmniExpress BeadChip. Breast cancer cases included in the BB2 study were genotyped using multiplex PCR Invader assays. SNP data for the BB2 controls were extracted from data generated using the OmniExpress BeadChip. We also selected 16 SNPs that showed a promising trend in the other studies included in stage 3 for additional genotyping assays among 2,021 cases and 1,958 controls included in a case-control study conducted in Malaysia and Singapore that used the iPLEX Sequenom MassArray platform at the Cancer Research Initiatives Foundation, Sime Darby Medical Centre, Malaysia. However, because of a potential concern about genetic admixture shown in our previous study3 and an unusual pattern of associations observed in these studies (Supplementary Table 8), we did not include these samples in the final analysis.

Statistical analyses. PLINK version 1.07 (see URLs)39 was used to analyze the genome-wide data obtained in stage 1. To evaluate the population structure in the Chinese GWAS (SBCGS-1), we performed principal-component analyses using EIGENSTRAT software40 in a set of approximately 6,000 independent SNPs that met the following criteria: (i) location at a distance of >200 kb from the SNP of interest, (ii) a MAF of >0.2, (iii) ii^2 < 0.1 and (iv) a genotype call rate of >99%. The inflation factor (λ) was estimated to be 1.0426. Similar analyses were performed for the Korean GWAS (SeBCS1) and yielded a λ value of 1.0431 (ref. 11). We also rescaled the inflation statistic to an equivalent value for a study including 1,000 cases and 1,000 controls (λ_1,000) using the formula λ_1,000 = 1 + 500 × (λ − 1) × (1/(n_cases + 1/controls))41. The λ_1,000 value was 1.02 for both studies included in stage 1 and 1.01 in the meta-analysis of both studies. These data suggest that any population structure, if present, should not have any appreciable effect on the results. OR estimates associated with each SNP and 95% CIs were estimated using logistic regression implemented in PLINK with adjustment for age and the first two principal components.

We used the program MACH 1.0 (see URLs)41 to impute genotypes for autosomal SNPs (n = 2,416,663) that were present in the CHB and JPT HapMap Phase 2 release 22 data for samples included in the Chinese and Korean GWAS. Only SNPs with a high imputation quality score (RSQR > 0.50) and MAF > 0.05 in these two GWAS were included in the analyses. Dosage data for imputed SNPs for samples in each GWAS were analyzed using the program Mach2dat (see URLs)12. Associations between genotype dosage (0, 1 or 2) of the effect allele and breast cancer risk were assessed using logistic regression models after adjusting for age and the first two principal components. ORs associated with each SNP and 95% CIs were estimated under a log-additive model. We also used SAS version 9.3 (see URLs) to analyze genotype data, which yielded results virtually identical to those generated with dosage data using Mach2dat. We obtained summary ORs and 95% CIs for SNPs from the two GWAS using METAL software (see URLs)13 to run a fixed-effects inverse variance meta-analysis.

Individual data were obtained from all studies except for the two BioBank Japan studies (BBJ1 and BBJ2). Case-control differences in selected demographic characteristics and major risk factors were evaluated using t tests (for continuous variables) and χ2 tests (for categorical variables). Summary associations between SNPs and breast cancer risk were generated on the basis of a fixed-effects inverse variance meta-analysis conducted using METAL software13. Analyses stratified by ancestry and ER status were also carried out. Heterogeneity across studies, among ancestry groups and according to ER status was assessed with a Cochran’s Q test. P values of <5 × 10−8 in the combined analysis were considered to be statistically significant.

We assessed associations of breast cancer risk with the three newly identified risk variants among women of European ancestry in collaboration with the DRIVE GAME-ON Consortium (see URLs). Included in this analysis were data from 16,003 cases and 41,335 controls recruited in 12 studies. Genome-wide scan data from these studies were imputed and meta-analysis was performed; summary data are presented herein.

We generated forest plots using R version 3.0.0 (see URLs). Regional association plots were drawn using the web-based tool LocusZoom, version 1.1 (see URLs)42. LD matrices used in this study were reported in the following analyses were considered to be statistically significant.

We assessed associations of breast cancer risk with the three newly identified risk variants among women of European ancestry in collaboration with the DRIVE GAME-ON Consortium (see URLs). Included in this analysis were data from 16,003 cases and 41,335 controls recruited in 12 studies. Genome-wide scan data from these studies were imputed and meta-analysis was performed; summary data are presented herein.

We generated forest plots using R version 3.0.0 (see URLs). Regional association plots were drawn using the web-based tool LocusZoom, version 1.1 (see URLs)42. LD matrices used in this study were reported in the following analyses were considered to be statistically significant.

We assessed associations of breast cancer risk with the three newly identified risk variants among women of European ancestry in collaboration with the DRIVE GAME-ON Consortium (see URLs). Included in this analysis were data from 16,003 cases and 41,335 controls recruited in 12 studies. Genome-wide scan data from these studies were imputed and meta-analysis was performed; summary data are presented herein.

We generated forest plots using R version 3.0.0 (see URLs). Regional association plots were drawn using the web-based tool LocusZoom, version 1.1 (see URLs)42. LD matrices used in this study were reported in the following analyses were considered to be statistically significant.
annotation using HaploReg v2 (see URLs)\textsuperscript{18}, a tool for exploring the annotations of the noncoding genome at variants on haplotype blocks.

**eQTL analysis.** We used TCGA breast cancer data (Supplementary Note) to perform an eQTL analysis for normal and tumor tissue samples separately. Detailed descriptions of the eQTL analysis are presented in the Supplementary Note. We focused only on SNPs and genes located within the 1-Mb regions flanking the three newly identified risk loci to identify cis-eQTLs. A significance threshold $P$ value of <0.01 was used to identify candidate cis-eQTLs.

**Differential gene expression analysis.** To identify differentially expressed genes located in the three identified risk loci, we analyzed data from a total of 87 pairs of tumor-normal breast tissue samples included in TCGA (Supplementary Note). Detailed descriptions of the differential gene expression analysis are presented in the Supplementary Note.