Research in context

Evidence before this study

Before the era of genome-wide association studies (GWAS), candidate gene study was a powerful approach to study complex diseases. Its major limitation is the very small sample size and low statistical power. In 2011, we conducted a systematic review of 1059 publications and investigated 279 genetic variants in 128 candidate genes and found moderate to strong evidence of an association with breast cancer for 51 of those variants. In our previous review with all available data pooled together, the median sample size for each genetic variant was only 4334 breast cancer cases and 5213 controls. In past years, GWAS data have been generated for hundreds of thousands of breast cancer cases and controls, and four variants identified in our previous candidate gene study were found to reach genome-wide significance. However, other variants suggested in our previous candidate gene study have not been systemically investigated in large GWAS.

Added value of this study

To our knowledge, this study is, to date, the largest candidate gene study to evaluate genetic variants identified in candidate gene studies for their association with breast cancer risk. In the present study, we have increased the sample size by a median of 18-fold (range of 3-451) and substantially improved the statistical power, compared with the sample size in the previous combined candidate gene studies. We found 12 variants from the original investigation in 10 candidate genes that were associated with breast cancer risk at a Bonferroni-corrected threshold. In our previous system review of candidate gene studies, only four of these 12 variants showed moderate/strong evidence of associations. Further investigating these 10 genes, we found two

additional variants showing associations at genome-wide significance. Among these 14 variants, only four have been reported in previous GWAS. Our findings suggest that some of the variants in candidate gene studies were associated with disease risk, and the inconclusive results from previous candidate genes studies were due to low statistical power.

Implications of all of the available evidence

By using large GWAS data, we found 14 variants in 10 candidate genes associated with breast cancer risk. Meanwhile, a null association was established for a large majority of variants in previous candidate gene studies. A functional investigation of the variants identified in the present study may provide insight into the biological and genetic etiology of breast cancer.

Re-evaluating Genetic Variants Identified in Candidate Gene Studies of Breast Cancer Risk Using Data from Nearly 280,000 Women of Asian and European Ancestry

Brief title: Re-evaluating Genetic Variants Identified in Candidate Gene Studies of Breast Cancer Risk

Yaohua Yang, PhD[1], Xiang Shu, PhD[1], Xiao-ou Shu, MD[1], Manjeet K. Bolla, MSc[2], Sun-Seog Kweon, PhD[3], Qiuyin Cai, MD[1], Kyriaki Michailidou, PhD[2], Qin Wang, MSc[2], Joe Dennis, MSc[2], Boyoung Park, PhD[4], Keitaro Matsuo, PhD[5,6], Ava Kwong, PhD[7,8,9], Sue Kyung Park, PhD[10,11,12], Anna H. Wu, PhD[13], Soo Hwang Teo, PhD[14,15], Motoki Iwasaki, PhD[16], Ji-Yeob Choi, PhD[10,11], Jingmei Li, PhD[17,18], Mikael Hartman, PhD[18,19], Chen-Yang Shen, PhD[20,21], Kenneth Muir, PhD[22,23], Artitaya Lophatananon, PhD[22,23], Bingshan Li, PhD[24], Wanqing Wen, PhD[1], Yu-Tang Gao, PhD[25], Yong-Bing Xiang, PhD[26], Kristan J. Aronson, PhD[27], John J. Spinell, PhD[28,29], Manuela Gago-Dominguez, MD[30,31], Esther M. John, PhD[32,33,34], Allison W. Kurian, PhD[33,35], Jenny Chang-Claude, PhD[36,37], Shou-Tung Chen, PhD[38], Thilo Dörk, PhD[39], D. Gareth R. Evans, PhD[40,41], Marjanka K. Schmidt, PhD[42,43], Min-Ho Shin, PhD[3], Graham G. Giles, PhD[44,45,46], Roger L. Milne, PhD[44,45,47], Jacques Simard, PhD[48], Michiaki Kubo, PhD[49], Peter Kraft, PhD[50,51], Daehee Kang, PhD[10,11,12], Douglas F. Easton, PhD[2], Wei Zheng, MD[1], Jirong Long, PhD[1]

[1] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, Tennessee, USA

[2] Center for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

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

[4] Department of Medicine, College of Medicine, Hanyang University, Seoul, South Korea

[5] Department of Cancer Epidemiology, Nagoya University Graduate School of Medicine, Nagoya, Japan

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

[7] Hong Kong Hereditary Breast Cancer Family Registry, Happy Valley, Hong Kong

[8] Department of Surgery, The University of Hong Kong, Pok Fu Lam, Hong Kong

[9] Department of Surgery, Hong Kong Sanatorium and Hospital, Happy Valley, Hong Kong

[10] Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, South Korea

[11] Cancer Research Institute, Seoul National University, Seoul, South Korea

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

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

[14] Cancer Research Malaysia, Subang Jaya, Selangor, Malaysia

[15] Breast Cancer Research Unit, Cancer Research Institute, University Malaya Medical Centre, Kuala Lumpur, Malaysia

[16] Division of Epidemiology, Center for Public Health Sciences, National Cancer Center, Tokyo, Japan

[17] Human Genetics, Genome Institute of Singapore, Singapore

[18] Department of Surgery, National University Hospital, Singapore

[19] Saw Swee Hock School of Public Health, National University of Singapore, Singapore

[20] School of Public Health, China Medical University, Taichung, Taiwan

[21] Taiwan Biobank, Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan

[22] Division of Health Sciences, Warwick Medical School, Warwick University, Coventry, UK

[23] Institute of Population Health, University of Manchester, Manchester, UK

[24] Department of Molecular Physiology & Biophysics, Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, USA

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

[26] State Key Laboratory of Oncogene and Related Genes & Department of Epidemiology, Shanghai Cancer Institute, Renji Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China

[27] Department of Public Health Sciences, and Cancer Research Institute, Queen's University, Kingston, Ontario, Canada

[28] Cancer Control Research, BC Cancer Agency, Vancouver, British Columbia, Canada

[29] School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada

[30] Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago De Compostela, Spain

[31] Moores Cancer Center, University of California San Diego, La Jolla, California, USA

[32] Department of Epidemiology, Cancer Prevention Institute of California, Fremont, California, USA

[33] Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California, USA

[34] Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California, USA

[35] Department of Medicine, Stanford University School of Medicine, Stanford, California, USA

[36] Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

[37] University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[38] Division of General Surgery, Changhua Christian Hospital, Changhua, Taiwan

[39] Gynaecology Research Unit, Hannover Medical School, Hannover, Germany

[40] Manchester Centre for Genomic Medicine, Central Manchester University Hospitals NHS Foundation Trust, Manchester, UK

[41] Manchester Centre for Genomic Medicine, University of Manchester, Manchester, UK

[42] Division of Molecular Pathology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands

[43] Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands

[44] Cancer Epidemiology and Intelligence Division, Cancer Council Victoria, 615 St Kilda Road, Melbourne, Victoria 3004, Australia

[45] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia

[46] Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia

[47] Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia

[48] Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Laval University, Québec City, Quebec, Canada

[49] RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

[50] Program in Genetic Epidemiology and Statistical Genetics, Harvard T. H. Chan School of Public Health, Boston, MA, USA

[51] Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA

[*]**Corresponding Author:** Jirong Long, PhD, Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, 2525 West End Ave, Suite 800, Nashville, Tennessee, 37203, USA. Email: jirong.long@vanderbilt.edu. Phone: (615) 343-6741.

Word count: Abstract, 250 words; Body, 3363 words.

Abstract

Background: We previously conducted a systematic field synopsis of 1059 breast cancer candidate gene studies and investigated 279 genetic variants, 51 of which showed associations. The major limitation of this work was the small sample size, even pooling data from all 1059 studies. Thereafter, genome-wide association studies (GWAS) have accumulated data for hundreds of thousands of subjects. It's necessary to re-evaluate these variants in large GWAS datasets.

Methods: Of these 279 variants, data were obtained for 228 from GWAS conducted within the Asian Breast Cancer Consortium (24 206 cases and 24 775 controls) and the Breast Cancer Association Consortium (122 977 cases and 105 974 controls of European ancestry). Meta-analyses were conducted to combine the results from these two datasets.

Findings: Of those 228 variants, an association was observed for 12 variants in 10 genes at a Bonferroni-corrected threshold of $P<2{\cdot}19\times10^{-4}$. The associations for four variants reached $P<5\times10^{-8}$ and have been reported by previous GWAS, including rs6435074 and rs6723097 (*CASP8*), rs17879961 (*CHEK2*) and rs2853669 (*TERT*). The remaining eight variants were rs676387 (*HSD17B1*), rs762551 (*CYP1A2*), rs1045485 (*CASP8*), rs9340799 (*ESR1*), rs7931342 (*CHR11*), rs1050450 (*GPX1*), rs13010627 (*CASP10*) and rs9344 (*CCND1*). Further investigating these 10 genes identified associations for two additional variants at $P<5\times10^{-8}$, including rs4793090 (near *HSD17B1*), and rs9210 (near *CYP1A2*), which have not been identified by previous GWAS.

Interpretation: Though most candidate gene variants were not associated with breast cancer risk, we found 14 variants showing an association. Our findings warrant further functional investigation of these variants.

Funding

National Institutes of Health

Keywords

Re-evaluation, Genetic variants, Candidate gene studies, Breast cancer risk

Research in context

Evidence before this study

Before the era of genome-wide association studies (GWAS), candidate gene study was a powerful approach to study complex diseases. Its major limitation is the very small sample size and low statistical power. In 2011, we conducted a systematic review of 1059 publications and investigated 279 genetic variants in 128 candidate genes and found moderate to strong evidence of an association with breast cancer for 51 of those variants. In our previous review with all available data pooled together, the median sample size for each genetic variant was only 4334 breast cancer cases and 5213 controls. In past years, GWAS data have been generated for hundreds of thousands of breast cancer cases and controls, and four variants identified in our previous candidate gene study were found to reach genome-wide significance. However, other variants suggested in our previous candidate gene study have not been systemically investigated in large GWAS.

Added value of this study

To our knowledge, this study is, to date, the largest candidate gene study to evaluate genetic variants identified in candidate gene studies for their association with breast cancer risk. In the present study, we have increased the sample size by a median of 18-fold (range of 3-451) and substantially improved the statistical power, compared with the sample size in the previous combined candidate gene studies. We found 12 variants from the original investigation in 10 candidate genes that were associated with breast cancer risk at a Bonferroni-corrected threshold. In our previous system review of candidate gene studies, only four of these 12 variants showed moderate/strong evidence of associations. Further investigating these 10 genes, we found two

additional variants showing associations at genome-wide significance. Among these 14 variants, only four have been reported in previous GWAS. Our findings suggest that some of the variants in candidate gene studies were associated with disease risk, and the inconclusive results from previous candidate genes studies were due to low statistical power.

Implications of all of the available evidence

By using large GWAS data, we found 14 variants in 10 candidate genes associated with breast cancer risk. Meanwhile, a null association was established for a large majority of variants in previous candidate gene studies. A functional investigation of the variants identified in the present study may provide insight into the biological and genetic etiology of breast cancer.

Introduction

Breast cancer is the most commonly diagnosed cancer among women globally [1]. Genetic factors contribute significantly to breast cancer etiology. Since 2005, genome-wide association studies (GWAS) have identified common genetic variants at approximately 170 risk loci for this malignancy [2]. Before the era of GWAS, a large number of candidate gene studies had been conducted to identify genetic variants for the risk of breast cancer. The genes were selected based on prior knowledge and biology. Within each gene, only a few genetic variants were investigated based on their potential function and the availability of genotyping assays, e.g., a recognition site for enzyme digestion. In addition, all of these studies were conducted on a limited number of participants, hence these studies had inadequate statistical power to detect the small risks commonly associated with breast cancer susceptibility variants.

In 2011, we conducted a systematic field synopsis of candidate gene studies of breast cancer [3]. Data from 1059 publications for 279 genetic variants in 128 candidate genes were included in the analyses. For those variants with an association with breast cancer risk at $P<0.05$, the epidemiological credibility of meta-analysis was defined as strong, moderate, or weak based on three grades, i.e. A, B, or C, in three categories: sum of test alleles among cases and controls, heterogeneity statistic, and protection from bias [3]. The evidence for significant associations in meta-analyses were defined as strong when grades of all three categories were A, moderate when grades of all three categories were A or B, and weak when grades of any categories C [3]. Using these criteria, we found 10 variants with strong evidence, four variants with moderate evidence, and 37 variants with weak evidence of association with breast cancer risk. Of these 51 variants, four reached genome-wide significance, i.e. $P<5\times10^{-8}$, in subsequent studies, including

rs6723097 and rs6435074 in *CASP8* [4], rs17879961 in *CHEK2* [2], and rs2853669 in *TERT* [5]. These results indicate that the candidate-gene approach is capable of identifying true associations. In addition, in our previous investigation of 279 genetic variants [3], convincing evidence of no association was identified for 45 variants, and no conclusion could be determined for the remaining 183. One of the major limitations of this work was the small sample size. Of the 1059 publications included in our previous analyses [3], the median study sample size was 461 cases and 503 controls. The median pooled sample size for each genetic variant was 4334 cases and 5213 controls. To date, GWAS data have been generated using much larger sample sizes [2,6], which have provided an unprecedented opportunity to re-evaluate genetic variants in candidate genes. Here, we re-evaluated the variants included in our previous investigation for their associations with breast cancer risk, using data from ~270 000 cases and controls.

Materials and Methods

Selection of candidate gene variants for re-evaluation

In the present study, of the 279 genetic variants included in our previous synopsis [3], we re-evaluated the association with breast cancer risk for 228 single nucleotide polymorphisms (SNPs), with data available from a much larger sample size. Among these 228 SNPs, in our previous synopsis [3], four, three and 34 showed an association with strong, moderate and weak evidence, respectively. A null association was found for another 144 SNPs and a null association with convincing evidence was found for the remaining 43 SNPs.

Data source and statistical analyses

Data were available for 213 of the 228 SNPs in the Asian Breast Cancer Consortium (ABCC), which includes 24 206 breast cancer cases and 24 775 controls of Asian ancestry. Detailed information of the ABCC has been described elsewhere [7]. Briefly, participants in the ABCC were originally from seven studies, including the Asian ExomeChip Project (*N*=3959), the Japanese Breast Cancer GWAS (*N*=4741), the Korean Breast Cancer GWAS (*N*=4298), the Breast Cancer Association Consortium (BCAC) OncoArray-Asian study (*N*=14 337), the BCAC iCOGS-Asian study (*N*=10 716), the Shanghai Breast Cancer GWAS (*N*=4646) and the Multi-Ethnic Genotyping Array (MEGA Project, *N*=6284, three sub-studies involved). Genotyping was conducted on multiple arrays and each dataset was imputed with the 1000 Genomes Phase 3 as reference. To estimate potential population structures, principal components (PCs) analyses were performed within each dataset. Then, logistic regression analyses were conducted within each dataset using PLINK2.0 [8] to estimate per-allele odds ratios (ORs) and standard errors (SEs) for SNPs, with age and the top two PCs additionally adjusted. Meta-analyses were conducted to combine the results from all seven datasets via the fixed-effects inverse-variance model implemented in METAL [9].

Data were also available for 222 of the 228 SNPs from the most recent analysis of the European-ancestry component of the BCAC (http://bcac.ccge.medschl.cam.ac.uk). The details of the BCAC dataset can be found elsewhere [2]. Briefly, genetic data were generated for 122 977 breast cancer cases and 105 974 control participants from three datasets. The first dataset included 46 785 cases and 42 892 controls that were genotyped using the iCOGS array [10]. The second dataset included 61 282 cases and 45 494 controls that were genotyped using the OncoArray [11]. The third dataset included 14 910 cases and 17 588 controls genotyped using various GWAS arrays.

All three datasets were also imputed using the 1000 Genomes Phase 3 as reference. PCs analyses were conducted within each of these three datasets to estimate the potential population structure. SNPTEST [12] and in-house software were used to perform logistic regression analyses within each dataset to estimate per-allele ORs and SEs for SNPs [2]. In all of the regression models, the top ten PCs additionally adjusted [2], and for the iCOGS and OncoArray data, country and study sites were also adjusted, respectively [2]. Finally, ORs and SEs of all SNPs were combined through a fixed-effects, inverse-variance meta-analysis using METAL [9].

Statistical analyses

For variants with data available in either ABCC or BCAC, the ORs and SEs for their associations with breast cancer risk were combined with a fixed-effects model using METAL [9]. Altogether, 228 variants in 117 candidate genes were included in the analyses of the present study. A Bonferroni-corrected threshold of $P<2{\cdot}19\times10^{-4}$ (0·05/228) was used to determine associations in the combined data from ABCC and BCAC. For variants that were associated with breast cancer risk, we further investigated the association results stratified by estrogen receptor (ER) status and racial group. The Cochran's Q test was used to evaluate the heterogeneity. For both the AABC and the BCAC, all participating studies were approved by their appropriate ethics review boards and all subjects provided informed consent.

Results

Genetic variants associated with breast cancer risk

As shown in **Table 1**, of the 228 genetic variants investigated, 12 variants in 10 genes were associated with breast cancer risk at a Bonferroni-corrected threshold of $P<2{\cdot}19\times10^{-4}$. Of these,

four variants reached the genome-wide significance threshold ($P<5\times10^{-8}$), including rs6723097 and rs6435074 in the *CASP8* gene, rs17879961 in the *CHEK2* gene and rs2853669 in the *TERT* gene. These four variants have been reported by previous GWAS [2,4,5].

The remaining eight variants were rs676387 (*HSD17B1*, $P=3\cdot78\times10^{-6}$), rs762551 (*CYP1A2*, $P=4\cdot50\times10^{-5}$), rs1045485 (*CASP8*, $P=7\cdot46\times10^{-6}$), rs9340799 (*ESR1*, $P=1\cdot33\times10^{-4}$), rs7931342 (*CHR11*, $P=2\cdot10\times10^{-4}$), rs1050450 (*GPX1*, $P=2\cdot13\times10^{-4}$), rs13010627 (*CASP10*, $P=6\cdot74\times10^{-7}$) and rs9344 (*CCND1*, $P=8\cdot14\times10^{-5}$) (**Table 1**). We further evaluated other variants which are in moderate linkage disequilibrium (LD) with these eight variants ($r^2>0\cdot50$) in either Asians or Europeans in the 1000 Genomes phase 3 data. We found two additional variants, rs4793090 (*HSD17B1*) and rs9210 (*CYP1A2*), that reached genome-wide significance, with *P* values of $5\cdot58\times10^{-9}$ and $4\cdot70\times10^{-8}$, respectively (**Table 1**). The variant rs9210 (*CYP1A2*) is in moderate LD with the originally investigated variant rs762551 (*CYP1A2*) in Europeans ($r^2=0\cdot58$) and in Asians ($r^2=0\cdot20$). The association of rs9210 with breast cancer risk attenuated drastically ($P=0\cdot03$) when conditioning on rs762551. These results indicate that rs9210 and rs762551 represent a single association signal.

The variant rs4793090 (*HSD17B1*) is in LD with the originally investigated variant, rs676387 (*HSD17B1*), in both Asians ($r^2=0\cdot89$) and Europeans ($r^2=0\cdot71$). After adjusting for rs676387, only a nominal association ($P=0\cdot04$) was observed for rs4793090, indicating that these two variants represent a single association signal. Approximately 150 kilobase (Kb) away from these two variants, the variant rs72826962 was reported to be associated with breast cancer at genome-wide significance level in the BCAC [2]. This variant is monomorphic in Asians and rare in

Europeans, and it is not in LD with either rs676387 or rs4793090. In the BCAC, after adjusting for rs72826962, the associations of rs676387 and rs4793090 with breast cancer didn't change materially, with *P* values of $3{\cdot}77\times10^{-4}$ and $1{\cdot}11\times10^{-5}$, respectively. Similarly, after adjusting for rs676387 and rs4793090, the variant rs72826962 was still associated with breast cancer risk with a $P=1{\cdot}31\times10^{-6}$. These results suggest that the associations of rs676387 and rs4793090 observed in the present study were independent of the previously identified GWAS-significant signal.

CASP8 variants rs6723097 and rs6435074 are in moderate LD with an r^2 of 0·35 in Asians and 0·56 in Europeans. After a mutual adjustment, the association for rs6435074 persisted in both Asians and Europeans, although attenuated, but the association for the rs6723097 disappeared in both racial groups. Thus, these two variants represented one association signal. Another variant in *CASP8*, rs1045485, was rare in Asians, with a minor allele frequency (MAF) of 0·0001 in gnomAD (https://gnomad.broadinstitute.org/), and was not investigated in women of Asian ancestry in the present study. The association was only observed for women of European ancestry. It is in weak LD with rs6723097 ($r^2=0{\cdot}08$) and rs6435074 ($r^2=0{\cdot}05$) in Europeans. However, the association for rs1045485 was not totally independent of rs6435074 and rs6723097. After adjusting for rs6723097 and rs6435074, the association for rs1045485 was substantially attenuated ($P=0{\cdot}048$).

Comparing with results from the previous candidate gene study [3]

In 2011, we conducted a systematic field synopsis for candidate gene studies using data from 1,059 publications [3]. For the 12 originally investigated variants that showed associations with breast cancer risk in the present study, only three (rs6723097 and rs1045485 in *CASP8*, and

rs17879961 in *CHEK2*) showed strong evidence of association, and only one variant (rs2853669 in *TERT*) showed moderate evidence in our previous investigation [3] (**Table 1**). Weak evidence of association was observed for four variants, including rs6435074 in *CASP8*, rs9340799 in *ESR1*, rs7931342 in *CHR11*, and rs676387 in *HSD17B1* [3]. The remaining four variants, rs13010627 in *CASP10*, rs9344 in *CCND1*, rs1050450 in *GPX1* and rs762551 in *CYP1A2*, were claimed to be not associated with breast cancer risk [3].

On the other hand, of the 10 variants that showed a strong evidence of association in our previous candidate gene study [3], data were available for four in the present study. Of these four variants, rs231775 in *CTLA4* was not associated with breast cancer risk in the present study (P=0·47; **Supplementary Table**). Of those four variants that showed moderate evidence of association in our previous candidate gene study [3], data were available for three in the present study. The variant rs2853669 in *TERT* showed a genome-wide significant association (P=1·54×10^{-23}; **Table 1**) and rs861539 in *XRCC3* showed a suggestive association (P=4·47×10^{-4}; **Supplementary Table**). The variant rs1800057 in *ATM* was not associated with breast cancer risk in the present study with a P=0·83 (**Supplementary Table**).

Stratified analyses by ER status and racial group

As shown in **Table 2**, all of the 14 variants that were associated with overall breast cancer risk showed nominal associations (P<0·05) for both ER-positive and ER-negative disease, except for rs17879961 in *CHEK2*, which was only associated with ER-positive disease ($P_{heterogeneity}$=3·42×10^{-3}). Three other variants showed a stronger association with ER-negative than ER-positive disease with $P_{heterogeneity}\leq$0·05, including rs2853669 in *TERT*, rs9340799 in

ESR1 and rs1050450 in *GPX1*. In our previous candidate gene study [3], no data were available regarding ER status.

Of the 14 variants associated with breast cancer risk, 12 reached a Bonferroni-corrected threshold ($P<2{\cdot}19\times10^{-4}$) and the remaining two had a $P\leq9{\cdot}02\times10^{-4}$ for women of European ancestry (**Table 3**). Of these 14 variants, three were very rare in East Asians, with a MAF from the 1000 Genomes Project of 0·0001, 0·00, and 0·001 for rs1045485 (*CASP8)*, rs17879961 (*CHEK2)*, and rs13010627 (*CASP10)*, respectively. Data were not available in the ABCC for these three variants. Of the remaining 11 variants, seven showed a nominal association ($P<0{\cdot}05$) in the ABCC. Of those, two variants in the *CASP8* gene, rs6723097 and rs6435074, reached the Bonferroni-corrected threshold of $P<2{\cdot}19\times10^{-4}$. Of these 11 variants tested in both racial groups, only two showed a difference in association between the two racial groups, with a $P_{heterogeneity}\leq0{\cdot}05$. The variant rs6435074 in *CASP8* had a larger effect size for Asians than for Europeans, while the variant rs7931342 in *CHR11* showed an association only for Europeans (**Table 3**). Forest plots showing associations of these 14 variants with breast cancer risk among Asians, Europeans and combined data, as well as in our previous candidate gene study, are presented in **Figure 1**.

Discussion

In the present study, we found 12 originally investigated variants in 10 candidate genes that were associated with breast cancer risk at a Bonferroni-corrected threshold. Four of these 12 variants reached genome-wide significance and had been reported by previous GWAS. Further investigating these candidate genes, we found two additional variants, rs4793090 (*HSD17B1*)

and rs9210 (*CYP1A2*), that showed associations at genome-wide significance. These two variants had not been reported by previous GWAS.

The four variants reported by previous GWAS in Europeans were rs6435074 and rs6723097 in *CASP8* [4], rs17879961 in *CHEK2* [2], and rs2853669 in *TERT* [5]. Of these four variants, rs17879961 (*CHEK2*) is extremely rare in Asians, with a MAF of <0.001 in sequencing data from ~10 000 East Asians in gnomAD. The other three variants showed consistent associations for Asians and Europeans. The two *CASP8* intronic variants, rs6435074 and rs6723097, showed similar associations in Europeans. However, in Asians, the variant rs6435074 showed a stronger association, reaching genome-wide significance. After a mutual adjustment, the association for rs6435074 persisted in both Asians and Europeans, although attenuated, but the association for rs6723097 disappeared in both racial groups. We further checked the GTEx data (https://gtexportal.org/home/) [13] and found that both of these variants were expression quantitative trait loci (eQTL) for the *CASP8* gene, with a stronger effect observed for rs6435074. Together, these results suggest that rs6435074 may be a more interesting variant for further investigation in this locus.

In the present study, we found an association with breast cancer risk for the intronic variant rs676387 in the *HSD17B1* gene. Upon further investigation of this locus, we found that another variant, rs4793090, which is in LD with rs676387 in both Asians and Europeans, was associated with breast cancer risk at genome-wide significance. After mutual adjustment, a nominal association was observed for rs4793090, and the association for rs676387 disappeared. These two variants are not in LD with the previously reported breast cancer susceptibility variant

rs72826962, which is located at ~130Kb from the *HSD17B1* gene [2]. Analyses conditioning on rs72826962 indicated that associations of these two *HSD17B1* variants with breast cancer risk were independent of that of rs72826962. Furthermore, the results from a most recent fine-mapping investigation [14] also showed that the genomic region in which these two variants are located represents an independent association signal from the GWAS-identified variant rs72826962. All of these indicated that rs4793090 and rs676387 represent a single association signal, which is independent from the GWAS-identified variant in this locus. The variant rs4793090 is located at ~15Kb from the *HSD17B1* gene and ~1.8Kb from the *NAGLU* gene. The *NAGLU* gene encodes an enzyme that degrades heparan sulfate by the hydrolysis of terminal N-acetyl-D-glucosamine residues in N-acetyl-alpha-D-glucosaminides. No published evidence has demonstrated a potential link between the *NAGLU* gene and breast cancer. On the other hand, the *HSD17B1* gene encodes the enzyme 17β-Hydroxysteroid dehydrogenase 1 (17β-HSD1), which is responsible for the interconversion between estrone and estradiol, and between androstenedione and testosterone [15]. In breast cancer cells, the expression level of the *HSD17B1* gene was positively correlated with estrone reduction and cell proliferation, but negatively correlated with levels of dihydrotestosterone, which has an antiproliferative effect on breast cancer cell growth [16]. Due to the important role of estrogen in breast cancer etiology, the *HSD17B1* gene has been one of the most commonly studied candidate genes. However, in all of these studies, there is no consistent evidence of association between genetic variants in this gene and breast cancer risk. Even after combining the data from these studies, only weak evidence of an association was observed [3]. To the best of our knowledge, our present study is the first to confirm associations of variants around the *HSD17B1* gene and risk of breast cancer.

For the *CYP1A2* gene, we found the originally investigated variant rs762551 showed an association with breast cancer risk. In our previous investigation, based on data from candidate gene studies, no association was observed for this variant [3]. In another, more recent, meta-analysis of candidate gene studies, a weak association was observed [17]. We further investigated variants around the *CYP1A2* gene and found a variant, rs9210, that showed an association at genome-wide significance. The variant rs9210 is in moderate LD in Europeans and borderline LD in Asians with rs762551. After a mutual adjustment, a nominal association was observed for rs9210 and but not for rs762551. All of these results suggests that rs9210 and rs762551 constitute a single in this locus, which has not been identified as a breast cancer susceptibility locus via previous GWAS. The variant rs9210 is located at the 3'-UTR of the *ULK3* gene and 87.3 Kb from the *CYP1A2* gene. The *ULK3* gene, encoding a serine/threonine protein kinase, was reported to be down-regulated during breast tumor progression [18]. The ULK3 protein was reported to regulate the Hedgehog signaling [19] and to function as a tumor suppressor [20]. The *CYP1A2* gene encodes a member of the cytochrome P450 superfamily of enzymes. The CYP1A2 protein catalyzes the metabolic activation of a variety of aryl- and heterocyclic amines, and also metabolizes some polycyclic aromatic hydrocarbons (PAHs) into carcinogenic intermediates [21]. The variant rs762551 is one of the most commonly studied variants in this gene in relation to breast cancer risk, but the findings were inconsistent [17,22,23]. Our present study provided strong evidence for an association of this variant with breast cancer risk, as well as a stronger association of another neighbor variant with breast cancer risk.

The variant rs13010627 in the *CASP10* gene showed no association in our previous candidate gene study [3]. However, in the present study, this variant was associated with breast cancer risk.

This variant is very rare in Asians; hence it could not be investigated in the ABCC. This variant was located at ~107Kb upstream of a previously GWAS-identified breast cancer risk variant, rs1830298, in Europeans [4]. However, there is no LD between these two variants. The variant rs13010627 represents an independent association signal at this locus.

The strengths of our study include its large sample size, even for the breast cancer sub-type, to evaluate the genetic variants in candidate genes with breast cancer risk. With data combined from women of European and Asian ancestry, we have unprecedent statistical power to detect true associations. For example, the rs9340799 in the *ESR1* gene and rs676387 in the *HSD17B1* gene did not reach the Bonferroni-corrected threshold in either racial group individually, but showed an association using the combined data. Similarly, the variant rs4793090, close to the *HSD17B1* gene, reached genome-wide significance only when using the combined data. In addition, we were able to evaluate the generalizability of the associations for these two racial groups. Furthermore, apart from the originally investigated variants in the candidate gene studies, we were able to investigate variants in LD with them, and found two more variants around the *HSD17B1* and *CYP1A2* genes that showed genome-wide significant associations. The main limitation of our study is that we only investigated common SNPs, since rare variants and indels could not be imputed well. Another limitation is that only women of Asian ancestry and European ancestry were included. Further large studies that include other racial/ethnic groups, such as women of African ancestry, may be helpful to better understand these genetic variants in relation to breast cancer risk.

In summary, using a large amount of GWAS data, we found 14 variants in 10 candidate genes associated with breast cancer risk. Further functional investigations of these variants may provide insight into the biological and genetic etiology of breast cancer.

Acknowledgements

The authors thank Jing He, and Marshal S. Younger of the Vanderbilt Epidemiology Center for their help. The authors would also like to thank all individuals who participated in the parent studies and all the researchers, clinicians, technicians and administrative staff for their contributions. The data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University. The eQTL results were accessed from the website of the GTEx project.

Funding sources

This project was supported in part by grants R01CA158473 and R01CA148677 from the U.S. National Institutes of Health, as well as funds from the Anne Potter Wilson endowment. This project was also supported by development funds from the Department of Medicine at the Vanderbilt University Medical Center. Kenneth Muir and Artitaya Lophatananon are supported by the NIHR Manchester Biomedical Research Centre and by the ICEP, which is supported by CRUK (C18281/A19169). Jingmei Li is supported by a National Research Foundation Singapore Fellowship (NRF-NRFF2017-02).

For studies participating in the ABCC, the BBJ1 was supported by the Ministry of Education, Culture, Sports, Sciences and Technology from the Japanese Government. The SeBCS was

supported by the BRL (Basic Research Laboratory) program through the National Research Foundation of Korea, funded by the Ministry of Education, Science and Technology (2011-0001564). The biospecimens and data of the Hwasun Cancer Epidemiology Study-Breast were provided by the Biobank of Chonnam National University Hwasun Hospital, a member of the Korea Biobank Network (07SA2014020). The Shanghai Breast Cancer GWAS was supported by the U.S. NIH grant R01CA064277.

The BCAC European data were generated with the support by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the 'Ministère de l'Économie, de la Science et de l'Innovation du Québec' through Genome Québec and grant PSR-SIIRI-701, The National Institutes of Health (U19 CA148065, X01HG007492), Cancer Research UK (C1287/A10118, C1287/A16563, C1287/A10710) and The European Union (HEALTH-F2-2009-223175 and H2020 633784 and 634935). The Canadian Breast Cancer Study (CBCS) was funded by the Canadian Institutes of Health Research, and the Canadian Breast Cancer Foundation/ Canadian Cancer Society. All studies and funders of BCAC are listed in Michailidou et al. 2017 [2].

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

Kristan J. Aronson reports grants from Canadian Institutes of Health Research and grants from Canadian Breast Cancer Foundation/ Cancer Society during the conduct of the present study. Gareth R. Evans reports personal fees from Astrazeneca, outside the present study. Allison W. Kurian reports grants from Myriad Genetics, outside the present study. Jacques Simard reports grants from Government of Canada, through Genome Canada and the Canadian Institutes of Health Research, the Ministère de l'Économie, de la Science et de l'Innovation du Québec through Genome Québec and grant PSR-SIIRI-70, during the conduct of the present study. All the authors declare no competing financial interests.

Author contributions

J.Long and W.Z. conceived the study. Y.Y. performed statistical analyses. Y.Y. and J.Long wrote the manuscript with significant contributions from W.Z., X.O.S., and Q.C. X.S., W.W. and B.L. contributed to data analyses. M.K.B., K.Michailidou., Q.W., J.D., J.S., R.L.M., P.K., M.K.S. and D.F.E. contributed to BCAC data management, statistical analyses and/or manuscript revision. S.S.K., B.P., K.Matsuo, A.K., S.K.P., A.H.W., S.H.T., M.I., J.Y.C., J.Li, M.H., C.Y.S., K.Muir, A.L., Y.T.G., Y.B.X., K.J.A., J.J.S., M.G.D., E.M.J., A.W.K., J. C.C., S.T.C., T.D., D.G.R.E., M.K.S., M.H.S., G.G.G., M.K. and D.K. contributed to the collection of the data and biological samples for the original studies in ABCC and BCAC. All authors have reviewed and approved the final manuscript.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 2018.
2. Michailidou K, Lindström S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* 2017; **551**(7678): 92.
3. Zhang B, Beeghly-Fadiel A, Long J, Zheng W. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *The lancet oncology* 2011; **12**(5): 477-88.
4. Michailidou K, Beesley J, Lindstrom S, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature genetics* 2015; **47**(4): 373.
5. Bojesen SE, Pooley KA, Johnatty SE, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nature genetics* 2013; **45**(4): 371.
6. Cai Q, Zhang B, Sung H, et al. Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32. 1, 5q14. 3 and 15q26. 1. *Nature genetics* 2014; **46**(8): 886-90.
7. Zheng W, Zhang B, Cai Q, 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. *Human molecular genetics* 2013; **22**(12): 2539-50.
8. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015; **4**(1): 7.
9. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26**(17): 2190-1.
10. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics* 2013; **45**(4): 353.
11. Amos CI, Dennis J, Wang Z, et al. The OncoArray Consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiology and Prevention Biomarkers* 2017; **26**(1): 126-35.
12. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics* 2007; **39**(7): 906.
13. Consortium G. Genetic effects on gene expression across human tissues. *Nature* 2017; **550**(7675): 204.
14. Fachal L, Aschard H, Beesley J, et al. Fine-mapping of 150 breast cancer risk regions identifies 178 high confidence target genes. *bioRxiv* 2019: 521054.
15. He W, Gauri M, Li T, Wang R, Lin S-X. Current knowledge of the multifunctional 17β-hydroxysteroid dehydrogenase type 1 (HSD17B1). *Gene* 2016; **588**(1): 54-61.
16. Aka JA, Mazumdar M, Chen C-Q, Poirier D, Lin S-X. 17β-hydroxysteroid dehydrogenase Type 1 stimulates breast cancer by dihydrotestosterone inactivation in addition to estradiol production. *Molecular endocrinology* 2010; **24**(4): 832-45.
17. Tian Z, Li Y-L, Zhao L, Zhang C-L. Role of CYP1A2* 1F polymorphism in cancer risk: Evidence from a meta-analysis of 46 case–control studies. *Gene* 2013; **524**(2): 168-74.
18. Vargas AC, Reed AEM, Waddell N, et al. Gene expression profiling of tumour epithelial and stromal compartments during breast cancer progression. *Breast cancer research and treatment* 2012; **135**(1): 153-65.

19. Maloverjan A, Piirsoo M, Michelson P, Kogerman P, Østerlund T. Identification of a novel serine/threonine kinase ULK3 as a positive regulator of Hedgehog pathway. *Experimental cell research* 2010; **316**(4): 627-37.
20. Liang C, Jung JU. Autophagy genes as tumor suppressors. *Current opinion in cell biology* 2010; **22**(2): 226-33.
21. Zhou S-F, Wang B, Yang L-P, Liu J-P. Structure, function, regulation and polymorphism and the clinical significance of human cytochrome P450 1A2. *Drug metabolism reviews* 2010; **42**(2): 268-354.
22. Wang H, Zhang Z, Han S, Lu Y, Feng F, Yuan J. CYP1A2 rs762551 polymorphism contributes to cancer susceptibility: a meta-analysis from 19 case-control studies. *BMC cancer* 2012; **12**(1): 528.
23. Ayari I, Fedeli U, Saguem S, Hidar S, Khlifi S, Pavanello S. Role of CYP1A2 polymorphisms in breast cancer risk in women. *Molecular medicine reports* 2013; **7**(1): 280-6.

Figure legends

Figure 1. Forest plot of fourteen genetic variants that showed an association with breast cancer risk in meta-analyses of 24 206 cases and 24 775 controls. AABC**,** Asian Breast Cancer Consortium, 24 206 cases and 24 775 controls; BCAC, the Breast Cancer Association Consortium, 122 977 cases and 105 974 controls of European ancestry. Logistic regression was used to estimate per-allele odds ratio and standard error for each variant, within the AABC and the BCAC. Meta-analyses were performed to combine the results from the AABC and the BCAC. All statistical tests were two-sided. Associations at a Bonferroni-corrected threshold of $P<2{\cdot}19\times10^{-4}$ were considered as significant.

Re-evaluating Genetic Variants Identified in Candidate Gene Studies of Breast Cancer Risk Using Data from Nearly 280,000 Women of Asian and European Ancestry

Brief title: Re-evaluating Genetic Variants Identified in Candidate Gene Studies of Breast Cancer Risk

Yaohua Yang, PhD[1], Xiang Shu, PhD[1], Xiao-ou Shu, MD[1], Manjeet K. Bolla, MSc[2], Sun-Seog Kweon, PhD[3], Qiuyin Cai, MD[1], Kyriaki Michailidou, PhD[2], Qin Wang, MSc[2], Joe Dennis, MSc[2], Boyoung Park, PhD[4], Keitaro Matsuo, PhD[5,6], Ava Kwong, PhD[7,8,9], Sue Kyung Park, PhD[10,11,12], Anna H. Wu, PhD[13], Soo Hwang Teo, PhD[14,15], Motoki Iwasaki, PhD[16], Ji-Yeob Choi, PhD[10,11], Jingmei Li, PhD[17,18], Mikael Hartman, PhD[18,19], Chen-Yang Shen, PhD[20,21], Kenneth Muir, PhD[22,23], Artitaya Lophatananon, PhD[22,23], Bingshan Li, PhD[24], Wanqing Wen, PhD[1], Yu-Tang Gao, PhD[25], Yong-Bing Xiang, PhD[26], Kristan J. Aronson, PhD[27], John J. Spinell, PhD[28,29], Manuela Gago-Dominguez, MD[30,31], Esther M. John, PhD[32,33,34], Allison W. Kurian, PhD[33,35], Jenny Chang-Claude, PhD[36,37], Shou-Tung Chen, PhD[38], Thilo Dörk, PhD[39], D. Gareth R. Evans, PhD[40,41], Marjanka K. Schmidt, PhD[42,43], Min-Ho Shin, PhD[3], Graham G. Giles, PhD[44,45,46], Roger L. Milne, PhD[44,45,47], Jacques Simard, PhD[48], Michiaki Kubo, PhD[49], Peter Kraft, PhD[50,51], Daehee Kang, PhD[10,11,12], Douglas F. Easton, PhD[2], Wei Zheng, MD[1], Jirong Long, PhD[1]

[1] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, Tennessee, USA

[2] Center for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

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

[4] Department of Medicine, College of Medicine, Hanyang University, Seoul, South Korea

[5] Department of Cancer Epidemiology, Nagoya University Graduate School of Medicine, Nagoya, Japan

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

[7] Hong Kong Hereditary Breast Cancer Family Registry, Happy Valley, Hong Kong

[8] Department of Surgery, The University of Hong Kong, Pok Fu Lam, Hong Kong

[9] Department of Surgery, Hong Kong Sanatorium and Hospital, Happy Valley, Hong Kong

[10] Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, South Korea

[11] Cancer Research Institute, Seoul National University, Seoul, South Korea

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

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

[14] Cancer Research Malaysia, Subang Jaya, Selangor, Malaysia

[15] Breast Cancer Research Unit, Cancer Research Institute, University Malaya Medical Centre, Kuala Lumpur, Malaysia

[16] Division of Epidemiology, Center for Public Health Sciences, National Cancer Center, Tokyo, Japan

[17] Human Genetics, Genome Institute of Singapore, Singapore

[18] Department of Surgery, National University Hospital, Singapore

[19] Saw Swee Hock School of Public Health, National University of Singapore, Singapore

[20] School of Public Health, China Medical University, Taichung, Taiwan

[21] Taiwan Biobank, Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan

[22] Division of Health Sciences, Warwick Medical School, Warwick University, Coventry, UK

[23] Institute of Population Health, University of Manchester, Manchester, UK

[24] Department of Molecular Physiology & Biophysics, Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, USA

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

[26] State Key Laboratory of Oncogene and Related Genes & Department of Epidemiology, Shanghai Cancer Institute, Renji Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China

[27] Department of Public Health Sciences, and Cancer Research Institute, Queen's University, Kingston, Ontario, Canada

[28] Cancer Control Research, BC Cancer Agency, Vancouver, British Columbia, Canada

[29] School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada

[30] Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago De Compostela, Spain

[31] Moores Cancer Center, University of California San Diego, La Jolla, California, USA

[32] Department of Epidemiology, Cancer Prevention Institute of California, Fremont, California, USA

[33] Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California, USA

[34] Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California, USA

[35] Department of Medicine, Stanford University School of Medicine, Stanford, California, USA

[36] Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany

[37] University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[38] Division of General Surgery, Changhua Christian Hospital, Changhua, Taiwan

[39] Gynaecology Research Unit, Hannover Medical School, Hannover, Germany

[40] Manchester Centre for Genomic Medicine, Central Manchester University Hospitals NHS Foundation Trust, Manchester, UK

[41] Manchester Centre for Genomic Medicine, University of Manchester, Manchester, UK

[42] Division of Molecular Pathology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands

[43] Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands

[44] Cancer Epidemiology and Intelligence Division, Cancer Council Victoria, 615 St Kilda Road, Melbourne, Victoria 3004, Australia

[45] Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia

[46] Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Australia

[47] Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia

[48] Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Laval University, Québec City, Quebec, Canada

[49] RIKEN Center for Integrative Medical Sciences, Yokohama, Japan

[50] Program in Genetic Epidemiology and Statistical Genetics, Harvard T. H. Chan School of Public Health, Boston, MA, USA

[51] Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA

[*]**Corresponding Author:** Jirong Long, PhD, Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, 2525 West End Ave, Suite 800, Nashville, Tennessee, 37203, USA. Email: jirong.long@vanderbilt.edu. Phone: (615) 343-6741.

Word count: Abstract, 250 words; Body, 3363 words.

Abstract

Background: We previously conducted a systematic field synopsis of 1059 breast cancer candidate gene studies and investigated 279 genetic variants, 51 of which showed associations. The major limitation of this work was the small sample size, even pooling data from all 1059 studies. Thereafter, genome-wide association studies (GWAS) have accumulated data for hundreds of thousands of subjects. It's necessary to re-evaluate these variants in large GWAS datasets.

Methods: Of these 279 variants, data were obtained for 228 from GWAS conducted within the Asian Breast Cancer Consortium (24 206 cases and 24 775 controls) and the Breast Cancer Association Consortium (122 977 cases and 105 974 controls of European ancestry). Meta-analyses were conducted to combine the results from these two datasets.

Findings: Of those 228 variants, an association was observed for 12 variants in 10 genes at a Bonferroni-corrected threshold of $P<2{\cdot}19\times10^{-4}$. The associations for four variants reached $P<5\times10^{-8}$ and have been reported by previous GWAS, including rs6435074 and rs6723097 (*CASP8*), rs17879961 (*CHEK2*) and rs2853669 (*TERT*). The remaining eight variants were rs676387 (*HSD17B1*), rs762551 (*CYP1A2*), rs1045485 (*CASP8*), rs9340799 (*ESR1*), rs7931342 (*CHR11*), rs1050450 (*GPX1*), rs13010627 (*CASP10*) and rs9344 (*CCND1*). Further investigating these 10 genes identified associations for two additional variants at $P<5\times10^{-8}$, including rs4793090 (near *HSD17B1*), and rs9210 (near *CYP1A2*), which have not been identified by previous GWAS.

Interpretation: Though most candidate gene variants were not associated with breast cancer risk, we found 14 variants showing an association. Our findings warrant further functional investigation of these variants.

Funding

National Institutes of Health

Keywords

Re-evaluation, Genetic variants, Candidate gene studies, Breast cancer risk

Research in context

Evidence before this study

Before the era of genome-wide association studies (GWAS), candidate gene study was a powerful approach to study complex diseases. Its major limitation is the very small sample size and low statistical power. In 2011, we conducted a systematic review of 1059 publications and investigated 279 genetic variants in 128 candidate genes and found moderate to strong evidence of an association with breast cancer for 51 of those variants. In our previous review with all available data pooled together, the median sample size for each genetic variant was only 4334 breast cancer cases and 5213 controls. In past years, GWAS data have been generated for hundreds of thousands of breast cancer cases and controls, and four variants identified in our previous candidate gene study were found to reach genome-wide significance. However, other variants suggested in our previous candidate gene study have not been systemically investigated in large GWAS.

Added value of this study

To our knowledge, this study is, to date, the largest candidate gene study to evaluate genetic variants identified in candidate gene studies for their association with breast cancer risk. In the present study, we have increased the sample size by a median of 18-fold (range of 3-451) and substantially improved the statistical power, compared with the sample size in the previous combined candidate gene studies. We found 12 variants from the original investigation in 10 candidate genes that were associated with breast cancer risk at a Bonferroni-corrected threshold. In our previous system review of candidate gene studies, only four of these 12 variants showed moderate/strong evidence of associations. Further investigating these 10 genes, we found two

additional variants showing associations at genome-wide significance. Among these 14 variants, only four have been reported in previous GWAS. Our findings suggest that some of the variants in candidate gene studies were associated with disease risk, and the inconclusive results from previous candidate genes studies were due to low statistical power.

Implications of all of the available evidence

By using large GWAS data, we found 14 variants in 10 candidate genes associated with breast cancer risk. Meanwhile, a null association was established for a large majority of variants in previous candidate gene studies. A functional investigation of the variants identified in the present study may provide insight into the biological and genetic etiology of breast cancer.

Introduction

Breast cancer is the most commonly diagnosed cancer among women globally [1]. Genetic factors contribute significantly to breast cancer etiology. Since 2005, genome-wide association studies (GWAS) have identified common genetic variants at approximately 170 risk loci for this malignancy [2]. Before the era of GWAS, a large number of candidate gene studies had been conducted to identify genetic variants for the risk of breast cancer. The genes were selected based on prior knowledge and biology. Within each gene, only a few genetic variants were investigated based on their potential function and the availability of genotyping assays, e.g., a recognition site for enzyme digestion. In addition, all of these studies were conducted on a limited number of participants, hence these studies had inadequate statistical power to detect the small risks commonly associated with breast cancer susceptibility variants.

In 2011, we conducted a systematic field synopsis of candidate gene studies of breast cancer [3]. Data from 1059 publications for 279 genetic variants in 128 candidate genes were included in the analyses. For those variants with an association with breast cancer risk at $P<0.05$, the epidemiological credibility of meta-analysis was defined as strong, moderate, or weak based on three grades, i.e. A, B, or C, in three categories: sum of test alleles among cases and controls, heterogeneity statistic, and protection from bias [3]. The evidence for significant associations in meta-analyses were defined as strong when grades of all three categories were A, moderate when grades of all three categories were A or B, and weak when grades of any categories C [3]. Using these criteria, we found 10 variants with strong evidence, four variants with moderate evidence, and 37 variants with weak evidence of association with breast cancer risk. Of these 51 variants, four reached genome-wide significance, i.e. $P<5\times10^{-8}$, in subsequent studies, including

rs6723097 and rs6435074 in *CASP8* [4], rs17879961 in *CHEK2* [2], and rs2853669 in *TERT* [5]. These results indicate that the candidate-gene approach is capable of identifying true associations. In addition, in our previous investigation of 279 genetic variants [3], convincing evidence of no association was identified for 45 variants, and no conclusion could be determined for the remaining 183. One of the major limitations of this work was the small sample size. Of the 1059 publications included in our previous analyses [3], the median study sample size was 461 cases and 503 controls. The median pooled sample size for each genetic variant was 4334 cases and 5213 controls. To date, GWAS data have been generated using much larger sample sizes [2,6], which have provided an unprecedented opportunity to re-evaluate genetic variants in candidate genes. Here, we re-evaluated the variants included in our previous investigation for their associations with breast cancer risk, using data from ~270 000 cases and controls.

Materials and Methods

Selection of candidate gene variants for re-evaluation

In the present study, of the 279 genetic variants included in our previous synopsis [3], we re-evaluated the association with breast cancer risk for 228 single nucleotide polymorphisms (SNPs), with data available from a much larger sample size. Among these 228 SNPs, in our previous synopsis [3], four, three and 34 showed an association with strong, moderate and weak evidence, respectively. A null association was found for another 144 SNPs and a null association with convincing evidence was found for the remaining 43 SNPs.

Data source and statistical analyses

Data were available for 213 of the 228 SNPs in the Asian Breast Cancer Consortium (ABCC), which includes 24 206 breast cancer cases and 24 775 controls of Asian ancestry. Detailed information of the ABCC has been described elsewhere [7]. Briefly, participants in the ABCC were originally from seven studies, including the Asian ExomeChip Project (*N*=3959), the Japanese Breast Cancer GWAS (*N*=4741), the Korean Breast Cancer GWAS (*N*=4298), the Breast Cancer Association Consortium (BCAC) OncoArray-Asian study (*N*=14 337), the BCAC iCOGS-Asian study (*N*=10 716), the Shanghai Breast Cancer GWAS (*N*=4646) and the Multi-Ethnic Genotyping Array (MEGA Project, *N*=6284, three sub-studies involved). Genotyping was conducted on multiple arrays and each dataset was imputed with the 1000 Genomes Phase 3 as reference. To estimate potential population structures, principal components (PCs) analyses were performed within each dataset. Then, logistic regression analyses were conducted within each dataset using PLINK2.0 [8] to estimate per-allele odds ratios (ORs) and standard errors (SEs) for SNPs, with age and the top two PCs additionally adjusted. Meta-analyses were conducted to combine the results from all seven datasets via the fixed-effects inverse-variance model implemented in METAL [9].

Data were also available for 222 of the 228 SNPs from the most recent analysis of the European-ancestry component of the BCAC (http://bcac.ccge.medschl.cam.ac.uk). The details of the BCAC dataset can be found elsewhere [2]. Briefly, genetic data were generated for 122 977 breast cancer cases and 105 974 control participants from three datasets. The first dataset included 46 785 cases and 42 892 controls that were genotyped using the iCOGS array [10]. The second dataset included 61 282 cases and 45 494 controls that were genotyped using the OncoArray [11]. The third dataset included 14 910 cases and 17 588 controls genotyped using various GWAS arrays.

All three datasets were also imputed using the 1000 Genomes Phase 3 as reference. PCs analyses were conducted within each of these three datasets to estimate the potential population structure. SNPTEST [12] and in-house software were used to perform logistic regression analyses within each dataset to estimate per-allele ORs and SEs for SNPs [2]. In all of the regression models, the top ten PCs additionally adjusted [2], and for the iCOGS and OncoArray data, country and study sites were also adjusted, respectively [2]. Finally, ORs and SEs of all SNPs were combined through a fixed-effects, inverse-variance meta-analysis using METAL [9].

Statistical analyses

For variants with data available in either ABCC or BCAC, the ORs and SEs for their associations with breast cancer risk were combined with a fixed-effects model using METAL [9]. Altogether, 228 variants in 117 candidate genes were included in the analyses of the present study. A Bonferroni-corrected threshold of $P<2{\cdot}19\times10^{-4}$ (0·05/228) was used to determine associations in the combined data from ABCC and BCAC. For variants that were associated with breast cancer risk, we further investigated the association results stratified by estrogen receptor (ER) status and racial group. The Cochran's Q test was used to evaluate the heterogeneity. For both the AABC and the BCAC, all participating studies were approved by their appropriate ethics review boards and all subjects provided informed consent.

Results

Genetic variants associated with breast cancer risk

As shown in **Table 1**, of the 228 genetic variants investigated, 12 variants in 10 genes were associated with breast cancer risk at a Bonferroni-corrected threshold of $P<2{\cdot}19\times10^{-4}$. Of these,

four variants reached the genome-wide significance threshold ($P<5\times10^{-8}$), including rs6723097 and rs6435074 in the *CASP8* gene, rs17879961 in the *CHEK2* gene and rs2853669 in the *TERT* gene. These four variants have been reported by previous GWAS [2,4,5].

The remaining eight variants were rs676387 (*HSD17B1*, $P=3\cdot78\times10^{-6}$), rs762551 (*CYP1A2*, $P=4\cdot50\times10^{-5}$), rs1045485 (*CASP8*, $P=7\cdot46\times10^{-6}$), rs9340799 (*ESR1*, $P=1\cdot33\times10^{-4}$), rs7931342 (*CHR11*, $P=2\cdot10\times10^{-4}$), rs1050450 (*GPX1*, $P=2\cdot13\times10^{-4}$), rs13010627 (*CASP10*, $P=6\cdot74\times10^{-7}$) and rs9344 (*CCND1*, $P=8\cdot14\times10^{-5}$) (**Table 1**). We further evaluated other variants which are in moderate linkage disequilibrium (LD) with these eight variants ($r^2>0\cdot50$) in either Asians or Europeans in the 1000 Genomes phase 3 data. We found two additional variants, rs4793090 (*HSD17B1*) and rs9210 (*CYP1A2*), that reached genome-wide significance, with *P* values of $5\cdot58\times10^{-9}$ and $4\cdot70\times10^{-8}$, respectively (**Table 1**). The variant rs9210 (*CYP1A2*) is in moderate LD with the originally investigated variant rs762551 (*CYP1A2*) in Europeans ($r^2=0\cdot58$) and in Asians ($r^2=0\cdot20$). The association of rs9210 with breast cancer risk attenuated drastically ($P=0\cdot03$) when conditioning on rs762551. These results indicate that rs9210 and rs762551 represent a single association signal.

The variant rs4793090 (*HSD17B1*) is in LD with the originally investigated variant, rs676387 (*HSD17B1*), in both Asians ($r^2=0\cdot89$) and Europeans ($r^2=0\cdot71$). After adjusting for rs676387, only a nominal association ($P=0\cdot04$) was observed for rs4793090, indicating that these two variants represent a single association signal. Approximately 150 kilobase (Kb) away from these two variants, the variant rs72826962 was reported to be associated with breast cancer at genome-wide significance level in the BCAC [2]. This variant is monomorphic in Asians and rare in

Europeans, and it is not in LD with either rs676387 or rs4793090. In the BCAC, after adjusting for rs72826962, the associations of rs676387 and rs4793090 with breast cancer didn't change materially, with *P* values of $3{\cdot}77\times10^{-4}$ and $1{\cdot}11\times10^{-5}$, respectively. Similarly, after adjusting for rs676387 and rs4793090, the variant rs72826962 was still associated with breast cancer risk with a $P=1{\cdot}31\times10^{-6}$. These results suggest that the associations of rs676387 and rs4793090 observed in the present study were independent of the previously identified GWAS-significant signal.

CASP8 variants rs6723097 and rs6435074 are in moderate LD with an r^2 of 0·35 in Asians and 0·56 in Europeans. After a mutual adjustment, the association for rs6435074 persisted in both Asians and Europeans, although attenuated, but the association for the rs6723097 disappeared in both racial groups. Thus, these two variants represented one association signal. Another variant in *CASP8*, rs1045485, was rare in Asians, with a minor allele frequency (MAF) of 0·0001 in gnomAD (https://gnomad.broadinstitute.org/), and was not investigated in women of Asian ancestry in the present study. The association was only observed for women of European ancestry. It is in weak LD with rs6723097 ($r^2=0{\cdot}08$) and rs6435074 ($r^2=0{\cdot}05$) in Europeans. However, the association for rs1045485 was not totally independent of rs6435074 and rs6723097. After adjusting for rs6723097 and rs6435074, the association for rs1045485 was substantially attenuated ($P=0{\cdot}048$).

Comparing with results from the previous candidate gene study [3]

In 2011, we conducted a systematic field synopsis for candidate gene studies using data from 1,059 publications [3]. For the 12 originally investigated variants that showed associations with breast cancer risk in the present study, only three (rs6723097 and rs1045485 in *CASP8*, and

rs17879961 in *CHEK2*) showed strong evidence of association, and only one variant (rs2853669 in *TERT*) showed moderate evidence in our previous investigation [3] (**Table 1**). Weak evidence of association was observed for four variants, including rs6435074 in *CASP8*, rs9340799 in *ESR1*, rs7931342 in *CHR11*, and rs676387 in *HSD17B1* [3]. The remaining four variants, rs13010627 in *CASP10*, rs9344 in *CCND1*, rs1050450 in *GPX1* and rs762551 in *CYP1A2*, were claimed to be not associated with breast cancer risk [3].

On the other hand, of the 10 variants that showed a strong evidence of association in our previous candidate gene study [3], data were available for four in the present study. Of these four variants, rs231775 in *CTLA4* was not associated with breast cancer risk in the present study (P=0·47; **Supplementary Table**). Of those four variants that showed moderate evidence of association in our previous candidate gene study [3], data were available for three in the present study. The variant rs2853669 in *TERT* showed a genome-wide significant association (P=1·54×10^{-23}; **Table 1**) and rs861539 in *XRCC3* showed a suggestive association (P=4·47×10^{-4}; **Supplementary Table**). The variant rs1800057 in *ATM* was not associated with breast cancer risk in the present study with a P=0·83 (**Supplementary Table**).

Stratified analyses by ER status and racial group

As shown in **Table 2**, all of the 14 variants that were associated with overall breast cancer risk showed nominal associations (P<0·05) for both ER-positive and ER-negative disease, except for rs17879961 in *CHEK2*, which was only associated with ER-positive disease ($P_{heterogeneity}$=3·42×10^{-3}). Three other variants showed a stronger association with ER-negative than ER-positive disease with $P_{heterogeneity}\leq$0·05, including rs2853669 in *TERT*, rs9340799 in

ESR1 and rs1050450 in *GPX1*. In our previous candidate gene study [3], no data were available regarding ER status.

Of the 14 variants associated with breast cancer risk, 12 reached a Bonferroni-corrected threshold ($P<2{\cdot}19\times10^{-4}$) and the remaining two had a $P\leq9{\cdot}02\times10^{-4}$ for women of European ancestry (**Table 3**). Of these 14 variants, three were very rare in East Asians, with a MAF from the 1000 Genomes Project of 0·0001, 0·00, and 0·001 for rs1045485 (*CASP8)*, rs17879961 (*CHEK2)*, and rs13010627 (*CASP10)*, respectively. Data were not available in the ABCC for these three variants. Of the remaining 11 variants, seven showed a nominal association ($P<0{\cdot}05$) in the ABCC. Of those, two variants in the *CASP8* gene, rs6723097 and rs6435074, reached the Bonferroni-corrected threshold of $P<2{\cdot}19\times10^{-4}$. Of these 11 variants tested in both racial groups, only two showed a difference in association between the two racial groups, with a $P_{heterogeneity}\leq0{\cdot}05$. The variant rs6435074 in *CASP8* had a larger effect size for Asians than for Europeans, while the variant rs7931342 in *CHR11* showed an association only for Europeans (**Table 3**). Forest plots showing associations of these 14 variants with breast cancer risk among Asians, Europeans and combined data, as well as in our previous candidate gene study, are presented in **Figure 1**.

Discussion

In the present study, we found 12 originally investigated variants in 10 candidate genes that were associated with breast cancer risk at a Bonferroni-corrected threshold. Four of these 12 variants reached genome-wide significance and had been reported by previous GWAS. Further investigating these candidate genes, we found two additional variants, rs4793090 (*HSD17B1*)

and rs9210 (*CYP1A2*), that showed associations at genome-wide significance. These two variants had not been reported by previous GWAS.

The four variants reported by previous GWAS in Europeans were rs6435074 and rs6723097 in *CASP8* [4], rs17879961 in *CHEK2* [2], and rs2853669 in *TERT* [5]. Of these four variants, rs17879961 (*CHEK2*) is extremely rare in Asians, with a MAF of <0.001 in sequencing data from ~10 000 East Asians in gnomAD. The other three variants showed consistent associations for Asians and Europeans. The two *CASP8* intronic variants, rs6435074 and rs6723097, showed similar associations in Europeans. However, in Asians, the variant rs6435074 showed a stronger association, reaching genome-wide significance. After a mutual adjustment, the association for rs6435074 persisted in both Asians and Europeans, although attenuated, but the association for rs6723097 disappeared in both racial groups. We further checked the GTEx data (https://gtexportal.org/home/) [13] and found that both of these variants were expression quantitative trait loci (eQTL) for the *CASP8* gene, with a stronger effect observed for rs6435074. Together, these results suggest that rs6435074 may be a more interesting variant for further investigation in this locus.

In the present study, we found an association with breast cancer risk for the intronic variant rs676387 in the *HSD17B1* gene. Upon further investigation of this locus, we found that another variant, rs4793090, which is in LD with rs676387 in both Asians and Europeans, was associated with breast cancer risk at genome-wide significance. After mutual adjustment, a nominal association was observed for rs4793090, and the association for rs676387 disappeared. These two variants are not in LD with the previously reported breast cancer susceptibility variant

rs72826962, which is located at ~130Kb from the *HSD17B1* gene [2]. Analyses conditioning on rs72826962 indicated that associations of these two *HSD17B1* variants with breast cancer risk were independent of that of rs72826962. Furthermore, the results from a most recent fine-mapping investigation [14] also showed that the genomic region in which these two variants are located represents an independent association signal from the GWAS-identified variant rs72826962. All of these indicated that rs4793090 and rs676387 represent a single association signal, which is independent from the GWAS-identified variant in this locus. The variant rs4793090 is located at ~15Kb from the *HSD17B1* gene and ~1.8Kb from the *NAGLU* gene. The *NAGLU* gene encodes an enzyme that degrades heparan sulfate by the hydrolysis of terminal N-acetyl-D-glucosamine residues in N-acetyl-alpha-D-glucosaminides. No published evidence has demonstrated a potential link between the *NAGLU* gene and breast cancer. On the other hand, the *HSD17B1* gene encodes the enzyme 17β-Hydroxysteroid dehydrogenase 1 (17β-HSD1), which is responsible for the interconversion between estrone and estradiol, and between androstenedione and testosterone [15]. In breast cancer cells, the expression level of the *HSD17B1* gene was positively correlated with estrone reduction and cell proliferation, but negatively correlated with levels of dihydrotestosterone, which has an antiproliferative effect on breast cancer cell growth [16]. Due to the important role of estrogen in breast cancer etiology, the *HSD17B1* gene has been one of the most commonly studied candidate genes. However, in all of these studies, there is no consistent evidence of association between genetic variants in this gene and breast cancer risk. Even after combining the data from these studies, only weak evidence of an association was observed [3]. To the best of our knowledge, our present study is the first to confirm associations of variants around the *HSD17B1* gene and risk of breast cancer.

For the *CYP1A2* gene, we found the originally investigated variant rs762551 showed an association with breast cancer risk. In our previous investigation, based on data from candidate gene studies, no association was observed for this variant [3]. In another, more recent, meta-analysis of candidate gene studies, a weak association was observed [17]. We further investigated variants around the *CYP1A2* gene and found a variant, rs9210, that showed an association at genome-wide significance. The variant rs9210 is in moderate LD in Europeans and borderline LD in Asians with rs762551. After a mutual adjustment, a nominal association was observed for rs9210 and but not for rs762551. All of these results suggests that rs9210 and rs762551 constitute a single in this locus, which has not been identified as a breast cancer susceptibility locus via previous GWAS. The variant rs9210 is located at the 3'-UTR of the *ULK3* gene and 87.3 Kb from the *CYP1A2* gene. The *ULK3* gene, encoding a serine/threonine protein kinase, was reported to be down-regulated during breast tumor progression [18]. The ULK3 protein was reported to regulate the Hedgehog signaling [19] and to function as a tumor suppressor [20]. The *CYP1A2* gene encodes a member of the cytochrome P450 superfamily of enzymes. The CYP1A2 protein catalyzes the metabolic activation of a variety of aryl- and heterocyclic amines, and also metabolizes some polycyclic aromatic hydrocarbons (PAHs) into carcinogenic intermediates [21]. The variant rs762551 is one of the most commonly studied variants in this gene in relation to breast cancer risk, but the findings were inconsistent [17,22,23]. Our present study provided strong evidence for an association of this variant with breast cancer risk, as well as a stronger association of another neighbor variant with breast cancer risk.

The variant rs13010627 in the *CASP10* gene showed no association in our previous candidate gene study [3]. However, in the present study, this variant was associated with breast cancer risk.

This variant is very rare in Asians; hence it could not be investigated in the ABCC. This variant was located at ~107Kb upstream of a previously GWAS-identified breast cancer risk variant, rs1830298, in Europeans [4]. However, there is no LD between these two variants. The variant rs13010627 represents an independent association signal at this locus.

The strengths of our study include its large sample size, even for the breast cancer sub-type, to evaluate the genetic variants in candidate genes with breast cancer risk. With data combined from women of European and Asian ancestry, we have unprecedent statistical power to detect true associations. For example, the rs9340799 in the *ESR1* gene and rs676387 in the *HSD17B1* gene did not reach the Bonferroni-corrected threshold in either racial group individually, but showed an association using the combined data. Similarly, the variant rs4793090, close to the *HSD17B1* gene, reached genome-wide significance only when using the combined data. In addition, we were able to evaluate the generalizability of the associations for these two racial groups. Furthermore, apart from the originally investigated variants in the candidate gene studies, we were able to investigate variants in LD with them, and found two more variants around the *HSD17B1* and *CYP1A2* genes that showed genome-wide significant associations. The main limitation of our study is that we only investigated common SNPs, since rare variants and indels could not be imputed well. Another limitation is that only women of Asian ancestry and European ancestry were included. Further large studies that include other racial/ethnic groups, such as women of African ancestry, may be helpful to better understand these genetic variants in relation to breast cancer risk.

In summary, using a large amount of GWAS data, we found 14 variants in 10 candidate genes associated with breast cancer risk. Further functional investigations of these variants may provide insight into the biological and genetic etiology of breast cancer.

Acknowledgements

The authors thank Jing He, and Marshal S. Younger of the Vanderbilt Epidemiology Center for their help. The authors would also like to thank all individuals who participated in the parent studies and all the researchers, clinicians, technicians and administrative staff for their contributions. The data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University. The eQTL results were accessed from the website of the GTEx project.

Funding sources

This project was supported in part by grants R01CA158473 and R01CA148677 from the U.S. National Institutes of Health, as well as funds from the Anne Potter Wilson endowment. This project was also supported by development funds from the Department of Medicine at the Vanderbilt University Medical Center. Kenneth Muir and Artitaya Lophatananon are supported by the NIHR Manchester Biomedical Research Centre and by the ICEP, which is supported by CRUK (C18281/A19169). Jingmei Li is supported by a National Research Foundation Singapore Fellowship (NRF-NRFF2017-02).

For studies participating in the ABCC, the BBJ1 was supported by the Ministry of Education, Culture, Sports, Sciences and Technology from the Japanese Government. The SeBCS was

supported by the BRL (Basic Research Laboratory) program through the National Research Foundation of Korea, funded by the Ministry of Education, Science and Technology (2011-0001564). The biospecimens and data of the Hwasun Cancer Epidemiology Study-Breast were provided by the Biobank of Chonnam National University Hwasun Hospital, a member of the Korea Biobank Network (07SA2014020). The Shanghai Breast Cancer GWAS was supported by the U.S. NIH grant R01CA064277.

The BCAC European data were generated with the support by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the 'Ministère de l'Économie, de la Science et de l'Innovation du Québec' through Genome Québec and grant PSR-SIIRI-701, The National Institutes of Health (U19 CA148065, X01HG007492), Cancer Research UK (C1287/A10118, C1287/A16563, C1287/A10710) and The European Union (HEALTH-F2-2009-223175 and H2020 633784 and 634935). The Canadian Breast Cancer Study (CBCS) was funded by the Canadian Institutes of Health Research, and the Canadian Breast Cancer Foundation/ Canadian Cancer Society. All studies and funders of BCAC are listed in Michailidou et al. 2017 [2].

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Declaration of interests

Kristan J. Aronson reports grants from Canadian Institutes of Health Research and grants from Canadian Breast Cancer Foundation/ Cancer Society during the conduct of the present study. Gareth R. Evans reports personal fees from Astrazeneca, outside the present study. Allison W. Kurian reports grants from Myriad Genetics, outside the present study. Jacques Simard reports grants from Government of Canada, through Genome Canada and the Canadian Institutes of Health Research, the Ministère de l'Économie, de la Science et de l'Innovation du Québec through Genome Québec and grant PSR-SIIRI-70, during the conduct of the present study. All the authors declare no competing financial interests.

Author contributions

J.Long and W.Z. conceived the study. Y.Y. performed statistical analyses. Y.Y. and J.Long wrote the manuscript with significant contributions from W.Z., X.O.S., and Q.C. X.S., W.W. and B.L. contributed to data analyses. M.K.B., K.Michailidou., Q.W., J.D., J.S., R.L.M., P.K., M.K.S. and D.F.E. contributed to BCAC data management, statistical analyses and/or manuscript revision. S.S.K., B.P., K.Matsuo, A.K., S.K.P., A.H.W., S.H.T., M.I., J.Y.C., J.Li, M.H., C.Y.S., K.Muir, A.L., Y.T.G., Y.B.X., K.J.A., J.J.S., M.G.D., E.M.J., A.W.K., J. C.C., S.T.C., T.D., D.G.R.E., M.K.S., M.H.S., G.G.G., M.K. and D.K. contributed to the collection of the data and biological samples for the original studies in ABCC and BCAC. All authors have reviewed and approved the final manuscript.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians* 2018.
2. Michailidou K, Lindström S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* 2017; **551**(7678): 92.
3. Zhang B, Beeghly-Fadiel A, Long J, Zheng W. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *The lancet oncology* 2011; **12**(5): 477-88.
4. Michailidou K, Beesley J, Lindstrom S, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature genetics* 2015; **47**(4): 373.
5. Bojesen SE, Pooley KA, Johnatty SE, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nature genetics* 2013; **45**(4): 371.
6. Cai Q, Zhang B, Sung H, et al. Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32. 1, 5q14. 3 and 15q26. 1. *Nature genetics* 2014; **46**(8): 886-90.
7. Zheng W, Zhang B, Cai Q, 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. *Human molecular genetics* 2013; **22**(12): 2539-50.
8. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015; **4**(1): 7.
9. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 2010; **26**(17): 2190-1.
10. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics* 2013; **45**(4): 353.
11. Amos CI, Dennis J, Wang Z, et al. The OncoArray Consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiology and Prevention Biomarkers* 2017; **26**(1): 126-35.
12. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature genetics* 2007; **39**(7): 906.
13. Consortium G. Genetic effects on gene expression across human tissues. *Nature* 2017; **550**(7675): 204.
14. Fachal L, Aschard H, Beesley J, et al. Fine-mapping of 150 breast cancer risk regions identifies 178 high confidence target genes. *bioRxiv* 2019: 521054.
15. He W, Gauri M, Li T, Wang R, Lin S-X. Current knowledge of the multifunctional 17β-hydroxysteroid dehydrogenase type 1 (HSD17B1). *Gene* 2016; **588**(1): 54-61.
16. Aka JA, Mazumdar M, Chen C-Q, Poirier D, Lin S-X. 17β-hydroxysteroid dehydrogenase Type 1 stimulates breast cancer by dihydrotestosterone inactivation in addition to estradiol production. *Molecular endocrinology* 2010; **24**(4): 832-45.
17. Tian Z, Li Y-L, Zhao L, Zhang C-L. Role of CYP1A2* 1F polymorphism in cancer risk: Evidence from a meta-analysis of 46 case–control studies. *Gene* 2013; **524**(2): 168-74.
18. Vargas AC, Reed AEM, Waddell N, et al. Gene expression profiling of tumour epithelial and stromal compartments during breast cancer progression. *Breast cancer research and treatment* 2012; **135**(1): 153-65.

19. Maloverjan A, Piirsoo M, Michelson P, Kogerman P, Østerlund T. Identification of a novel serine/threonine kinase ULK3 as a positive regulator of Hedgehog pathway. *Experimental cell research* 2010; **316**(4): 627-37.
20. Liang C, Jung JU. Autophagy genes as tumor suppressors. *Current opinion in cell biology* 2010; **22**(2): 226-33.
21. Zhou S-F, Wang B, Yang L-P, Liu J-P. Structure, function, regulation and polymorphism and the clinical significance of human cytochrome P450 1A2. *Drug metabolism reviews* 2010; **42**(2): 268-354.
22. Wang H, Zhang Z, Han S, Lu Y, Feng F, Yuan J. CYP1A2 rs762551 polymorphism contributes to cancer susceptibility: a meta-analysis from 19 case-control studies. *BMC cancer* 2012; **12**(1): 528.
23. Ayari I, Fedeli U, Saguem S, Hidar S, Khlifi S, Pavanello S. Role of CYP1A2 polymorphisms in breast cancer risk in women. *Molecular medicine reports* 2013; **7**(1): 280-6.

Figure legends

Figure 1. Forest plot of fourteen genetic variants that showed an association with breast cancer risk in meta-analyses of 24 206 cases and 24 775 controls. AABC**,** Asian Breast Cancer Consortium, 24 206 cases and 24 775 controls; BCAC, the Breast Cancer Association Consortium, 122 977 cases and 105 974 controls of European ancestry. Logistic regression was used to estimate per-allele odds ratio and standard error for each variant, within the AABC and the BCAC. Meta-analyses were performed to combine the results from the AABC and the BCAC. All statistical tests were two-sided. Associations at a Bonferroni-corrected threshold of $P<2{\cdot}19\times10^{-4}$ were considered as significant.

Figure 1

Click here to download high resolution image

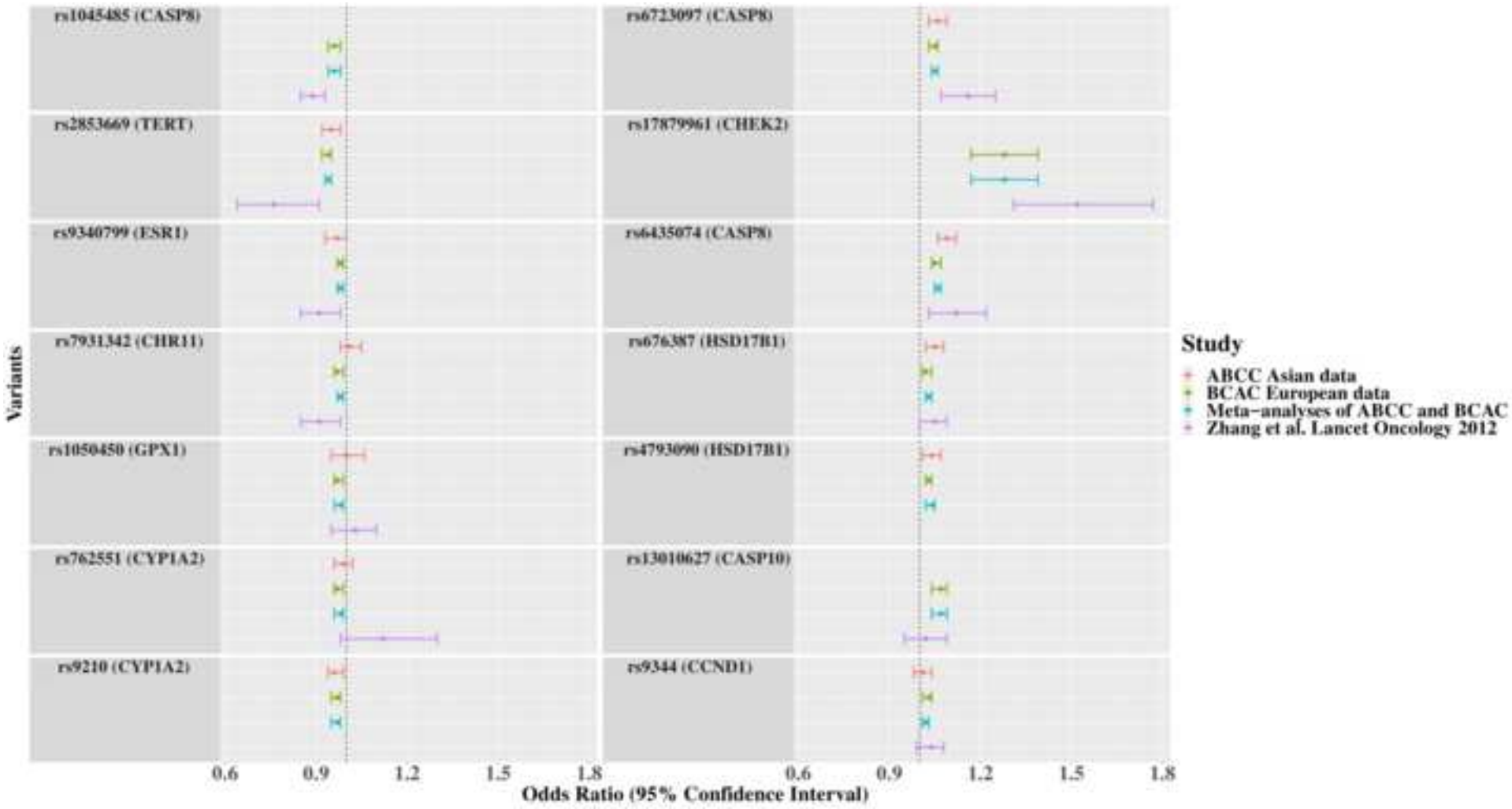

Table 1. Genetic variants in candidate genes showing an association with breast cancer risk

Gene	Variant	Chr	Position (hg19)	Alleles [a]	OR (95% CI)	*P* value	OR (95% CI) [b]	*P* value [b]	Cumulative evidence of association [b]
CASP8	rs6723097	2	202,128,618	A/C	1·05 (1·04-1·06)	$6{\cdot}87\times10^{-17}$	1·16 (1·07-1·25)	$1{\cdot}91\times10^{-4}$	Strong
CASP8	rs1045485	2	202,149,589	C/G	0·96 (0·94-0·98)	$7{\cdot}46\times10^{-6}$	0·89 (0·85-0·93)	$4{\cdot}65\times10^{-8}$	Strong
CHEK2	rs17879961	22	29,121,087	G/A	1·28 (1·17-1·39)	$9{\cdot}66\times10^{-9}$	1·52 (1·31-1·77) [c]	$4{\cdot}76\times10^{-8}$ [c]	Strong
TERT	rs2853669	5	1,295,349	C/T	0·94 (0·93-0·95)	$1{\cdot}54\times10^{-23}$	0·76 (0·64-0·91) [d]	$2{\cdot}00\times10^{-3}$ [d]	Moderate
CASP8	rs6435074	2	202,127,947	A/C	1·06 (1·05-1·07)	$8{\cdot}70\times10^{-21}$	1·12 (1·03-1·22)	0·01	Weak
ESR1	rs9340799	6	152,163,381	G/A	0·98 (0·97-0·99)	$1{\cdot}33\times10^{-4}$	0·91 (0·85-0·98) [d]	0·01 [d]	Weak
CHR11	rs7931342	11	68,994,497	T/G	0·98 (0·97-0·99)	$2{\cdot}10\times10^{-4}$	0·91 (0·85-0·98) [d]	0·01 [d]	Weak
HSD17B1	rs676387	17	40,706,273	A/C	1·03 (1·02-1·04)	$3{\cdot}78\times10^{-6}$	1·05 (1·00-1·09)	0·05	Weak
HSD17B1	rs4793090 [e]	17	40,686,342	G/A	1·04 (1·02-1·05)	$5{\cdot}58\times10^{-9}$	NA	NA	NA
GPX1	rs1050450	3	49,394,834	T/C	0·98 (0·96-0·99)	$2{\cdot}13\times10^{-4}$	1·03 (0·95-1·10)	0·52	No association
CYP1A2	rs762551	15	75,041,917	C/A	0·98 (0·96-0·99)	$4{\cdot}50\times10^{-5}$	1·12 (0·98-1·30)	0·15	No association
CYP1A2	rs9210 [f]	15	75,128,501	T/C	0·97 (0·95-0·98)	$4{\cdot}70\times10^{-8}$	NA	NA	NA
CASP10	rs13010627	2	202,074,098	A/G	1·07 (1·04-1·09)	$6{\cdot}74\times10^{-7}$	1·02 (0·95-1·09)	0·61	No association, convincing evidence
CCND1	rs9344	11	69,462,910	A/G	1·02 (1·01-1·03)	$8{\cdot}14\times10^{-5}$	1·04 (0·99-1·08)	0·12	No association, convincing evidence

Chr=chromosome. OR=odds ratio. CI=confidence interval.

[a] Effect allele vs other allele.

[b] Results from previous meta-analyses in Zhang et al. *Lancet Oncology*, *2011.*

[c] Dominant model.

[d] Recessive model.

[e] The variant rs4793090 was ~18Kb from *HSD17B1*, in LD with rs676387 and showing a genom×10-wide significant association, but not tested in Zhang et al. *Lancet Oncology*, 2011.

[f] The variant rs9210 was ~80Kb from *CYP1A2*, in LD with rs762551 and showing a genom×10-wide significant association, but not tested in Zhang et al. *Lancet Oncology*, 2011.

Table 2. Association results stratified by estrogen receptor (ER) status

Gene	Variant	Chr	Alleles [a]	ER-positive		ER-negative		Heterogeneity	
				OR (95% CI)	*P* value	OR (95% CI)	*P* value	*P* value	I^2 (%)
CASP8	rs6723097	2	A/C	1·05 (1·03-1·06)	$1{\cdot}20\times10^{-10}$	1·06 (1·04-1·09)	$1{\cdot}47\times10^{-9}$	0·16	49·81
CASP8	rs1045485	2	C/G	0·97 (0·95-0·99)	0·01	0·95 (0·92-0·98)	$3{\cdot}42\times10^{-3}$	0·26	20·00
CHEK2	rs17879961	22	G/A	1·35 (1·18-1·54)	$9{\cdot}82\times10^{-6}$	0·95 (0·81-1·12)	0·55	$1{\cdot}10\times10^{-3}$	90·65
TERT	rs2853669	5	C/T	0·96 (0·95-0·97)	$3{\cdot}29\times10^{-8}$	0·89 (0·87-0·91)	$3{\cdot}03\times10^{-24}$	$<2{\cdot}20\times10^{-16}$	96·67
CASP8	rs6435074	2	A/C	1·06 (1·04-1·07)	$9{\cdot}91\times10^{-14}$	1·06 (1·04-1·08)	$1{\cdot}73\times10^{-7}$	0·88	0·00
ESR1	rs9340799	6	G/A	0·98 (0·97-1·00)	0·02	0·95 (0·93-0·97)	$8{\cdot}07\times10^{-6}$	0·01	83·60
CHR11	rs7931342	11	T/G	0·98 (0·96-0·99)	$9{\cdot}92\times10^{-4}$	0·97 (0·95-0·99)	$9{\cdot}32\times10^{-3}$	0·74	0·00
HSD17B1	rs676387	17	A/C	1·02 (1·01-1·04)	$1{\cdot}98\times10^{-3}$	1·04 (1·02-1·06)	$8{\cdot}33\times10^{-4}$	0·31	4·97
HSD17B1	rs4793090 [b]	17	G/A	1·03 (1·02-1·05)	$8{\cdot}40\times10^{-6}$	1·03 (1·01-1·06)	$1{\cdot}54\times10^{-3}$	0·91	0·00
GPX1	rs1050450	3	T/C	0·98 (0·96-0·99)	$2{\cdot}15\times10^{-3}$	0·95 (0·93-0·97)	$1{\cdot}28\times10^{-5}$	0·05	74·11
CYP1A2	rs762551	15	C/A	0·97 (0·96-0·98)	$4{\cdot}72\times10^{-5}$	0·98 (0·96-1·00)	0·04	0·56	0·00
CYP1A2	rs9210 [c]	15	T/C	0·96 (0·95-0·98)	$2{\cdot}63\times10^{-7}$	0·96 (0·94-0·98)	$6{\cdot}21\times10^{-4}$	0·96	0·00
CASP10	rs13010627	2	A/G	1·07 (1·03-1·10)	$3{\cdot}01\times10^{-5}$	1·06 (1·01-1·11)	0·01	0·90	0·00
CCND1	rs9344	11	A/G	1·03 (1·01-1·04)	$1{\cdot}25\times10^{-4}$	1·02 (1·00-1·04)	0·03	0·76	0·00

Chr=chromosome. ER=estrogen receptor. OR=odds ratio. CI=confidence interval.

[a] Effect allele vs. other allele.

[b] The variant rs4793090 was ~18Kb from *HSD17B1*, in LD with rs676387 and showing a genome-wide significant association, but not tested in Zhang et al. *Lancet Oncology*, 2011.

[c] The variant rs9210 was ~80Kb from *CYP1A2*, in LD with rs762551 and showing a genome-wide significant association, but not tested in Zhang et al. *Lancet Oncology*, 2011.

Table 3. Association results stratified by ethnic group

Gene	Variant	Chr	Alleles [a]	Asian			European			Heterogeneity	
				EAF (%)	OR (95% CI)	*P* value	EAF (%)	OR (95% CI)	*P* value	*P* value	I^2 (%)
CASP8	rs6723097	2	A/C	52·18	1·06 (1·03-1·09)	$3·98\times10^{-5}$	40·46	1·05 (1·03-1·06)	$3·85\times10^{-13}$	0·49	0·00
CASP8	rs1045485	2	C/G	0·10	NA	NA	12·03	0·96 (0·94-0·98)	$7·46\times10^{-6}$	NA	NA
CHEK2	rs17879961	22	G/A	0·00	NA	NA	0·50	1·28 (1·17-1·39)	$9·66\times10^{-9}$	NA	NA
TERT	rs2853669	5	C/T	37·70	0·95 (0·92-0·98)	$7·92\times10^{-4}$	28·83	0·94 (0·92-0·95)	$4·05\times10^{-21}$	0·59	0·00
CASP8	rs6435074	2	A/C	29·86	1·09 (1·06-1·12)	$1·40\times10^{-8}$	27·34	1·05 (1·04-1·07)	$2·45\times10^{-14}$	0·05	73·21
ESR1	rs9340799	6	G/A	19·35	0·97 (0·93-1·00)	0·04	30·82	0·98 (0·97-0·99)	$9·02\times10^{-4}$	0·46	0·00
CHR11	rs7931342	11	T/G	76·69	1·01 (0·98-1·05)	0·36	48·61	0·97 (0·96-0·99)	$1·45\times10^{-5}$	0·02	82·63
HSD17B1	rs676387	17	A/C	43·55	1·05 (1·02-1·08)	$9·41\times10^{-4}$	26·64	1·02 (1·01-1·04)	$4·63\times10^{-4}$	0·16	49·21
HSD17B1	rs4793090 [b]	17	G/A	67·69	1·04 (1·01-1·07)	$3·92\times10^{-3}$	32·31	1·03 (1·02-1·04)	$4·32\times10^{-7}$	0·60	0·00
GPX1	rs1050450	3	T/C	7·24	1·00 (0·95-1·06)	0·92	33·60	0·97 (0·96-0·99)	$1·41\times10^{-4}$	0·31	1·34
CYP1A2	rs762551	15	C/A	32·74	0·99 (0·96-1·02)	0·44	32·01	0·97 (0·96-0·99)	$3·49\times10^{-5}$	0·26	20·90
CYP1A2	rs9210 [c]	15	T/C	26·39	0·96 (0·94-0·99)	0·02	31·01	0·97 (0·95-0·98)	$7·14\times10^{-7}$	0·91	0·00
CASP10	rs13010627	2	A/G	0·10	NA	NA	6·16	1·07 (1·04-1·09)	$7·47\times10^{-7}$	NA	NA
CCND1	rs9344	11	A/G	57·14	1·01 (0·98-1·04)	0·67	49·70	1·03 (1·01-1·04)	$4·23\times10^{-5}$	0·23	29·67

Chr=chromosome. EAF= effect allele frequency. OR=odds ratio. CI=confidence interval.

[a] Effect allele vs. other allele.

[b] The variant rs4793090 was ~18Kb from *HSD17B1*, in LD with rs676387 and showing a genome-wide significant association, but not tested in Zhang et al. *Lancet Oncology*, 2011.

[c] The variant rs9210 was ~80Kb from *CYP1A2*, in LD with rs762551 and showing a genome-wide significant association, but not tested in Zhang et al. *Lancet Oncology*, 2011.

e-component

Click here to download e-component: Yang et al. Supplementary Table.pdf