The American Journal of Human Genetics, Volume *109*

Supplemental information

Genome- and transcriptome-wide association studies of 386,000 Asian and European-ancestry women provide new insights into breast cancer genetics

Guochong Jia, Jie Ping, Xiang Shu, Yaohua Yang, Qiuyin Cai, Sun-Seog Kweon, Ji-Yeob Choi, Michiaki Kubo, Sue K. Park, Manjeet K. Bolla, Joe Dennis, Qin Wang, Xingyi Guo, Bingshan Li, Ran Tao, Kristan J. Aronson, Tsun L. Chan, Yu-Tang Gao, Mikael Hartman, Weang Kee Ho, Hidemi Ito, Motoki Iwasaki, Hiroji Iwata, Esther M. John, Yoshio Kasuga, Mi-Kyung Kim, Allison W. Kurian, Ava Kwong, Jingmei Li, Artitaya Lophatananon, Siew-Kee Low, Shivaani Mariapun, Koichi Matsuda, Keitaro Matsuo, Kenneth Muir, Dong-Young Noh, Boyoung Park, Min-Ho Park, Chen-Yang Shen, Min-Ho Shin, John J. Spinelli, Atsushi Takahashi, Chiuchen Tseng, Shoichiro Tsugane, Anna H. Wu, Taiki Yamaji, Ying Zheng, Alison M. Dunning, Paul D.P. Pharoah, Soo-Hwang Teo, Daehee Kang, Douglas F. Easton, Jacques Simard, Xiao-ou Shu, Jirong Long, and Wei Zheng

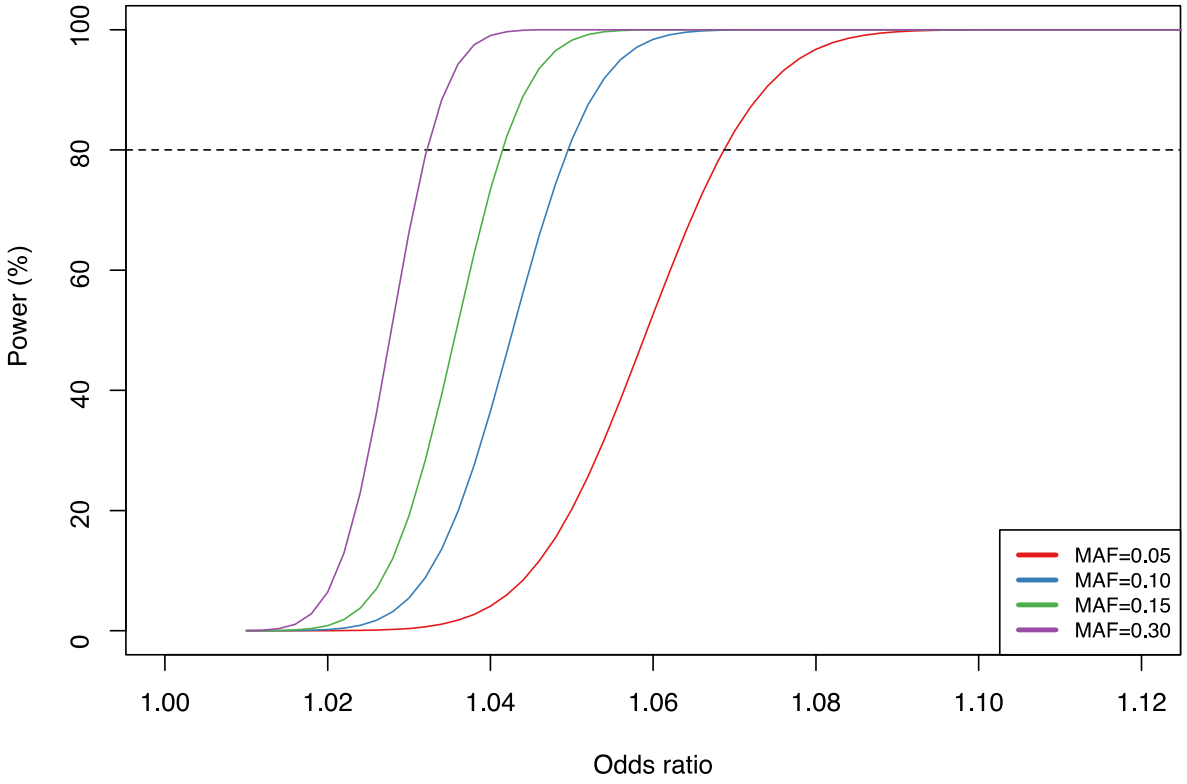

Figure S1. Estimated power of cross-ancestry meta-analysis using samples from ABCC and BCAC.

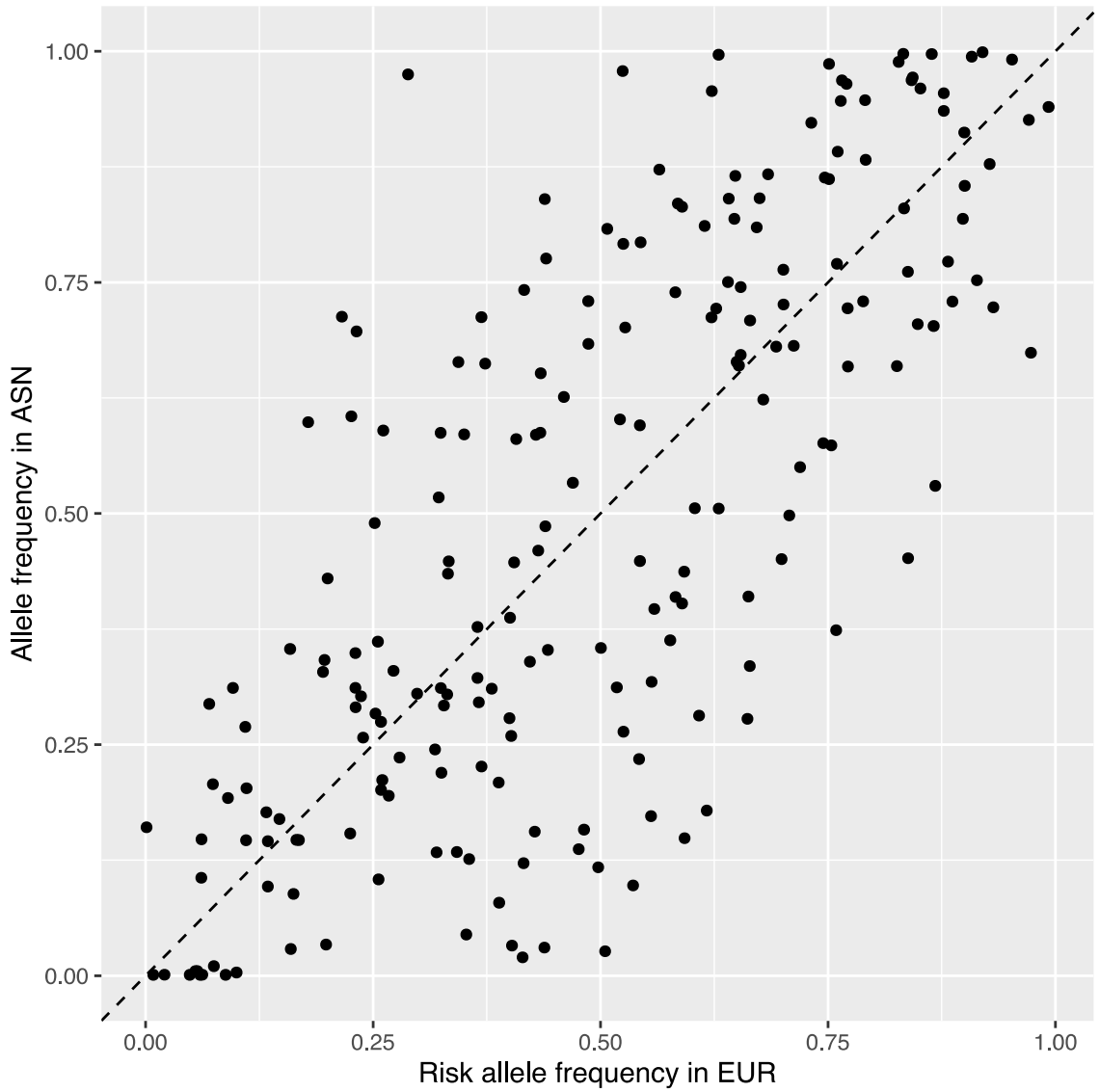

Figure S2. Comparison of allele frequency in Asian- and European-ancestry women for lead variants at risk loci identified by cross-ancestry meta-analysis. The counted allele was the allele in association with an increased risk of breast cancer in European-ancestry women. The black dashed line is the diagonal line.

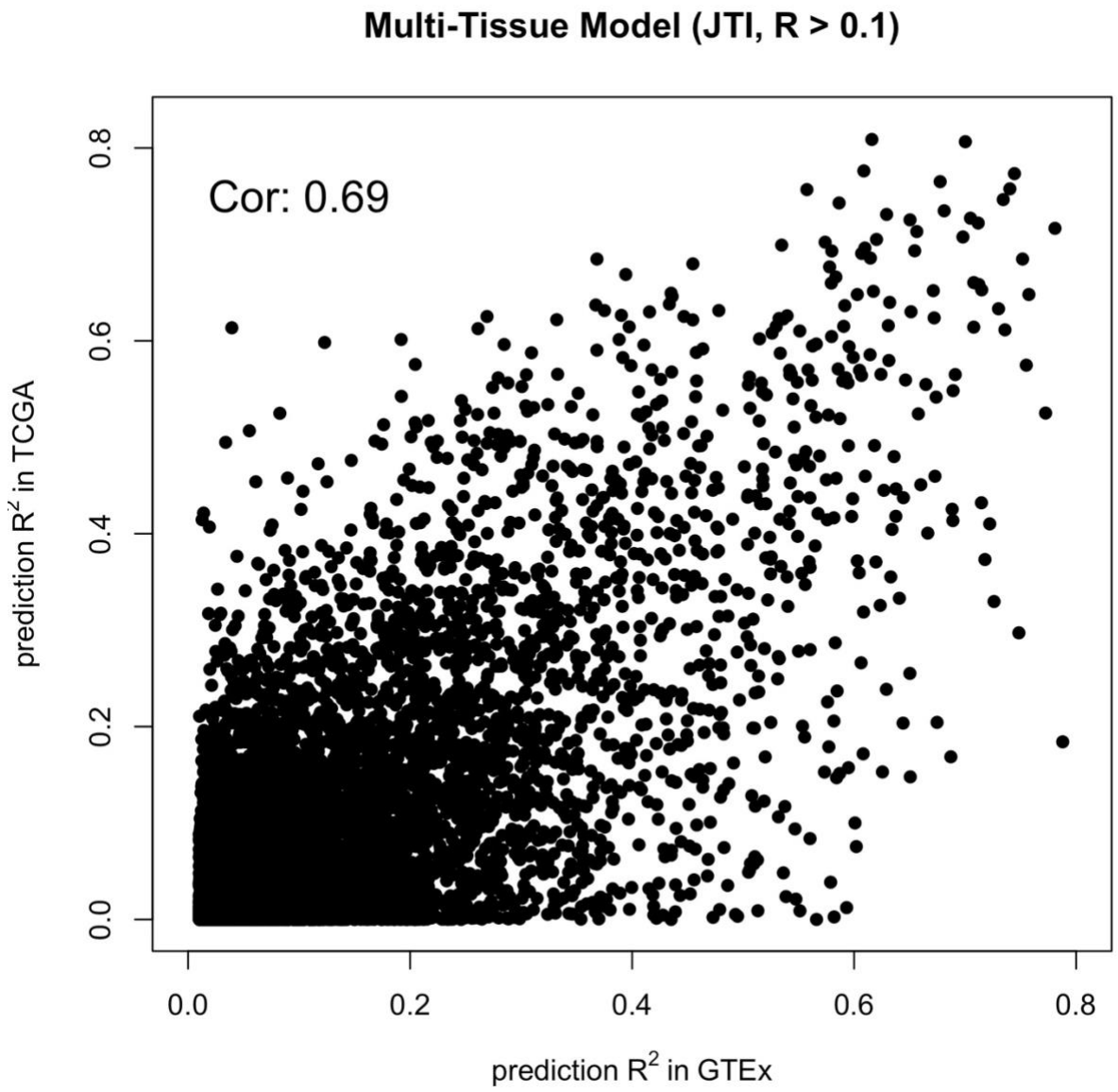

Figure S3. Performance of expression prediction model in GTEx and TCGA data for genes with over 10% correlation in GTEx data. The x axis represents the prediction performance (R^2) in the GTEx dataset (n = 115) and the y axis represents the prediction performance in the TCGA dataset (n = 86). Each dot represents the expression prediction model for one gene. There is a trend that genes with high prediction performance in the GTEx data also have high prediction performance in the TCGA (Pearson's correlation coefficient: 0.69).

Legends for Supplemental Tables

Table S1. Studies included in the cross-ancestry meta-analysis.

Table S2. Lead variants at risk loci for risk of overall breast cancer identified by meta-analyses.

Table S3. Lead variants at risk loci for risk of ER-positive breast cancer identified by meta-analyses.

Table S4. Lead variants at risk loci for risk of ER-negative breast cancer identified by meta-analyses.

Table S5. Results for the association of breast cancer risk with 17 novel risk loci in women from ABCC and BCAC

Table S6. Associations of novel risk variants for overall breast cancer risk from analyses using meta-regression.

Table S7. Associations by ER status for lead variants at risk loci identified by cross-ancestry meta-analyses.

Table S8. Associations with breast cancer risk for previously reported index SNPs not located at loci identified by our cross-ancestry meta-analysis.

Table S9. Independent association signals at novel breast cancer risk loci identified by conditional analysis in women of Asian and European ancestry.

Table S10. Samples by tissue type used in cross-tissue model building.

Table S11. Genes associated with breast cancer risk at the Bonferroni-corrected signficance level.

Table S12. Associations with breast cancer by ER status for genes identified at the Bonferroni-corrected significant level.

Table S13. Ancestry-specific associations with breast cancer risk for genes identified at the Bonferroni-corrected significant level.

Table S14. TWAS fine-mapping results for significant genes.

Table S15. Colocalization analysis results for TWAS significant genes from COLOC.

Table S16. Putative target protein-coding genes at risk loci for breast cancer risk.

Table S17. Pathway analyses for protein-coding genes associated with breast cancer.

Table S18. Summary of findings from genome- and transcriptome-wide association analyses with overall breast cancer and ER subtypes.

Supplemental Methods

I. Description of Study Populations

1. Description of Studies of the Asia Breast Cancer Consortium (ABCC)

1.1 Shanghai Breast Cancer Genetics Study (SBCGS)

The Chinese participants were drawn from Shanghai Breast Cancer Genetics Study (SBCGS), which consists of the Shanghai Breast Cancer Study (SBCS), Shanghai Breast Cancer Survival Study (SBCSS), Shanghai Endometrial Cancer Study (SECS, contributed control data only), and the Shanghai Women's Health Study (SWHS), four large population-based studies in urban Shanghai. All participants provided written informed consent prior to interview, and institutional review boards of all institutes in both China and the United States approved the study.

The SBCGS contributed samples to both ABCC and the BCAC Asian samples. Samples overlapped between ABCC and BCAC were only kept in the ABCC.

1.1.1 Shanghai Breast Cancer Study (SBCS)

The SBCS is a two-phase (SBCS-I and SBCS-II) population-based case-control study that recruited incident patients with breast cancer and controls in urban Shanghai, China.[1,2] The first phase (SBCS-I) recruited 1,602 eligible breast cancer cases and 1,724 eligible controls, from August 1996 to March 1998. Cases were recruited by a rapid case-ascertainment system and the population-based Shanghai Cancer Registry, and controls were randomly selected from the general population using the Shanghai Resident Registry. There were 1,459 cases (91.1%) and 1,556 controls (90.3%) who completed in-person interviews. Blood samples (10 ml from each woman) were obtained who completed the in-person interview (1,193 (82%) cases and 1,310 (84%) controls). A sample of exfoliated buccal cells was obtained using cotton swabs from virtually all study participants who did not provide a blood sample. The second phase (SBCS-II) recruited subjects between April 2002 and February 2005 using a protocol similar to the one used in the initial phase. Similar to the SBCS-I subjects, the majority of newly-recruited cases (n=1,932, 97.1%) and controls (n=1,857, 93.4%) provided a blood sample or an exfoliated buccal cell sample to the study. The modified mouthwash method initially reported by Lum A *et al*. was used.[3] Eligibility criteria for study participation were identical for SBCS-I and SBCS-II except age. The age ranged from 25 to 65 years for SBCS-I, and from 25 to 70 years in SBCS-II.

1.1.2 Shanghai Breast Cancer Survival Study (SBCSS)

The SBCSS included 6,303 breast cancer cases ascertained via the population-based Shanghai Cancer Registry between April 2002 and December 2006.[1] Information on known breast cancer risk factors as well as anthropometrics was collected by in-person interviews using a protocol and questionnaire similar to that used in the SBCS. Buccal cell samples were collected from 96% of study participants using the modified mouthwash method. There were 1,469 breast cancer patients participated in both SBCS-II and SBCSS due to the time overlap in the participant recruitment period.

1.1.3 Shanghai Endometrial Cancer Study (SECS)

The SECS is a population-based, case-control study of endometrial cancer conducted between January 1997 and December 2003 using a protocol similar to the SBCS, and only the community controls from the SECS were included in the present study.[1] Eligible cases were identified through the population-based Shanghai Cancer Registry and controls were randomly selected from the general population of Shanghai using the Shanghai Resident Registry and were age frequency matched to cases. Detailed information was collected by in-person interviews and anthropometrics measurements were taken. A total of 1,039 controls provided a blood sample or buccal cell sample using the mouthwash method, and these women were included in SBCGS.

1.1.4 Shanghai Women's Health Study (SWHS)

The SWHS is a population-based cohort study which recruited approximately 75,000 adult women from urban Shanghai between 1997 and 2000.[4] A total of 56,831 subjects, 75.8% of those who completed baseline survey through an in-person interview, donated a blood sample. An exfoliated buccal cell sample was collected from an additional 8,934 (49.3%) of the 18,111 subjects who did not provide a blood sample at baseline. Genomic DNA was available for about 88% of cohort members. Cancer cases were identified via record linkage with the population-based cancer registry and data collected at the Vital Statistic Unit, followed by home visits or telephone calls if necessary to confirm the diagnoses. Cancer diagnoses were verified by a review of medical records obtained from the diagnosing hospital.

Participants in SBCGS have been genotyped by Affymetrix Genome-Wide Human SNP Array 6.0, the Asian ExomeChip, and the Multi-Ethnic Global Array (MEGA). Similar genotyping and QC procedures have been described previously.[1,5] After imputation with the 1000 Genomes Project Phase 3 and QC exclusions, the final dataset included 2,511 cases and 2,135 controls for 11.1 million markers for the Affy6 dataset, 1,563 cases and 2,396 controls for 2.95 million markers for the ExomeChip dataset, and 1,794 cases and 2,059 controls for 14.1 million markers for the MEGA dataset.

1.2 Hwasun Cancer Epidemiology Study-Breast (HCES-Br)

The Hwasun Cancer Epidemiology Study (HCES-Br) is a hospital-based case-control study to identify factors of the cancer development and clinical progression in a Korean population.[6,7] The study included 3,387 female breast cancer cases diagnosed between April 2004 and February 2013 at Chonnam National University Hwasun Hospital, a cancer specified hospital in Jeollanam-do province, South Korea. Patients with secondary or recurrent tumor were excluded. Controls were 3,186 women who were randomly selected from among women with no previous cancer diagnosis at enrollment in the Namwon Study and the Dong-gu study, ongoing community-based cohort studies in South Korea.[8] Genomic DNA was extracted from their peripheral blood. Demographics data and conventional factors of breast cancer were collected by structured questionnaire and review of medical records. All cases and control subjects provided the informed consent to participate in the study and Institutional Review Board of Chonnam National University Hwasun Hospital approved this study. In the HCES-Br, there were 274 cases and 273 controls genotyped by MEGA and imputed with the 1000 Genomes Project Phase 3 data as reference.

1.3 Korea Precision Oncology Program (KPOP) - Breast Cancer

The KPOP – Breast Cancer study is a study to investigate genetic mutation/variants distribution of hereditary breast/ovarian cancer and risk stratification for women with or without family history of breast cancer. In addition, the risk factors of breast cancer were studied in women, stratified by family history of breast cancer. All cases had a histologically confirmed diagnosis of invasive breast cancer or ductal carcinoma in situ. The breast cancer cases were recruited from breast cancer center and genetic counseling clinic, National Cancer Center in Korea between 2013 and 2018. The controls were recruited from health screening examinees from National Cancer Center between 2013 and 2016 and they were women free of any cancer. After obtaining informed consent, cases and controls were asked to complete questionnaire on reproductive factors, lifestyle factors, and family history of cancer and provided blood samples. After separating plasma, serum, and whole blood, samples were stored at -70°C until assayed. Overall, 1904 breast cancer cases and 1195 controls were recruited. In KPOP, there were 963 cases and 921 controls were successfully genotyped by MEGA and imputed with the 1000 Genomes Project Phase 3 data as reference.

1.4 The Biobank Japan Project (BBJ2)

The BioBank Japan Project recruited around 200,000 patients with 47 diseases in Japan and collaboratively collected DNA and serum samples (https://biobankjp.org/english/index.html).[9,10] There were a total of 5,552 breast cancer patients and 89,731 female controls registered in Biobank Japan. Control samples were from population-based prospective cohorts and samples without related diagnoses. Samples were genotyped using the Illumina HumanOmniExpressExome BeadChip or a combination of the Illumina HumanOmniExpress and HumanExome BeadChips, and imputed with the 1000 Genomes Project Phase 3 data as reference.[11]

1.5 Seoul Breast Cancer Study (SeBCS):

The SeBCS is a hospital-based case-control study conducted in two teaching hospitals in Seoul.[12,13] Between 2001 and 2007, there were 2,342 patients with primary breast cancer recruited in the study. Information on known breast cancer risk factors and anthropometrics were collected by in-person interviews using a protocol and questionnaire. Medical charts were reviewed to verify clinical information. Eligible controls were derived from a large urban cohort included in the Korea Genome Epidemiology Study (KoGES), which was an ongoing cohort study that has sought to understand the causes and risk factors of disease in South Korea. A total of 2,052 controls were recruited between May 2006 and December 2007. They were frequency-matched to cases on the case's age at diagnosis in five-year intervals. Using a structured questionnaire and a protocol similar to the SeBCS, trained interviewers collected the demographic characteristics of the controls, their family histories with regard to breast cancer in first-degree relatives, reproductive and menstrual factors, and life-style habits. Samples were genotyped using Affymetrix 6.0 array. After quality control and imputation by the 1000 Genomes Project Phase 3, the final data set included 2,246 cases and 2,052 controls.[14]

In addition to AABC, the SeBCS also contributed samples to BCAC Asian dataset.

2. BCAC Asian samples

The studies included in the BCAC that contributed individual-level data to the Asian-specific meta-analysis were listed as Study, Location and BCAC project(s): ACP, Thailand, Oncoarray and iCOGS; CBCS, Canada, Oncoarray; HERPACC, Japan, Oncoarray and iCOGS; HKHBCFR, Hong Kong, Oncoarray; KOHBRA, Korea, Oncoarray; LAABC, USA, iCOGS; MYBRCA, Malaysia, Oncoarray and iCOGS; NC-BCFR, USA, Oncoarray; NGOBCS, Japan, Oncoarray; SBCGS, China, Oncoarray and iCOGS; SeBCS, Korea, Oncoarray and iCOGS; SGBCC, Singapore, Oncoarray and iCOGS; TWBCS, Taiwan, Oncoarray and iCOGS.

2.1 Asia Cancer Program (ACP):
The ACP is a hospital-based case-control study conducted in Thailand. Breast cancer cases were recruited between 1999-2000, and 2008-present at The National Cancer Institute (Central region), The Prince Songkla University Research Centre (South region), The HRH Princess Maha Chakri Sirindhorn Medical Centre (MSMC)-Srinakarinviroj University (Eastern region), Khon-Kaen University Cancer Centre (North-Eastern region). Women who were less than 71 years of age and underwent biopsy were eligible to participate in the study. All cases were pathologically diagnosed with breast cancer. Women resided in the same geographic area, younger than 71 years old, and reported no prior history of cancer were recruited as controls. In total, 944 invasive cases and 1,382 controls were included in the BCAC Asian dataset.

2.2 Canadian Breast Cancer Study (CBCS)
The CBCS is a population-based case-control study conducted in Canada.[15–18] Incident cases diagnosed between 2005 and 2009 were recruited from two areas, Vancouver, British Columbia and Kingston, Ontario. The cases were ascertained either from the population cancer registry (Vancouver, British Columbia) or participants of the Hotel Dieu Breast Assessment Program (Kingston, Ontario). Cancer-free controls were recruited through the Screening Mammography Program of British Columbia or the Hotel Dieu Breast Assessment Program in Kingston, Ontario. Controls were frequency matched by 5-year age groups.

2.3 Hospital-based Epidemiologic Research Program at Aichi Cancer Center (HERPACC)
The participants were recruited from a hospital-based case-control study conducted in Aichi, Japan.[19] All incident breast cancer cases were newly diagnosed within 1 year from the first visit to the Aichi Cancer Center between 2001 and 2013. Controls were selected from pool of non-cancer patients who firstly visited Aichi Cancer Center between 2001 and 2011. Subjects with previous cancer history were excluded.

2.4 Hong Kong Hereditary Breast Cancer (HKHBCFR)
Genetic screening of high-risk breast cancer patients was approached for the study enrollment from all hospitals in Hong Kong, China between 2006 and 2014.[20–22] Controls were selected from pool of non-cancer patients who visited hospitals in Hong Kong during the same period of recruitment as cases.

2.5 Korean Hereditary Breast Cancer (KOHBRA)

The KOHBRA study is an ongoing cohort study since 2007 to examine high risk groups for hereditary breast cancer such as female breast cancer patients with a family history, ovarian cancer, or other coincidental cancers, male breast cancer patients, and family members of breast cancer patients with *BRCA1/2* mutation. Final dataset included selected 1,397 female cancer patients without *BRCA1/2* mutation among KOHBRA subjects recruited in 2007-2009.[23]

2.6 Los Angeles County Asian-American Breast Cancer Case-Control Study (LAABC)

The LAABC is a population-based case-control study of incident breast cancer among Asian American women in Los Angeles County. Breast cancer cases were ascertained through the Los Angeles Cancer Surveillance Program. The included women were identified as Chinese, Japanese or Filipino women (aged 25-74 years) with a histologically confirmed primary breast cancer diagnosed between 1996 and 2006.[24–26] Controls were recruited from the same neighborhood as where cancer cases resided at the time of diagnosis. Cases and controls were frequency-matched on specific Asian ethnicities and 5-year age groups.

2.7 Malaysian Breast Cancer Genetic Study (MYBRCA)

Prevalent or incident breast cancer cases identified at the Breast Cancer Clinic in University Malaya Medical Centre from January 2003 to July 2014 and Subang Jaya Medical Centre from September 2012 to September 2014.[27] Controls are cancer-free individuals (37-74 years) selected from women attending mammographic screening at the same hospitals.

2.8 Northern California Breast Cancer Family Registry (NC-BCFR)

Incident breast cancer cases included women aged <65 years diagnosed from 1995-2009, identified through the SEER cancer registry of the Greater San Francisco Bay Area. All cases with indicators of increased genetic risk were eligible to enroll (diagnosed at age <35 years, personal history of ovarian or childhood cancer, bilateral breast cancer with 1st diagnosis at age <50, family history of breast or ovarian cancer in first-degree relatives).[28,29] Cases not meeting these criteria were randomly sampled (2.5% of non-Hispanic whites, 32% of other race/ethnicities). Incident cases also included men aged <80 years diagnosed from 1995-1998. Controls were those unaffected family members enrolled from 1995-2011 or unaffected unrelated subjects identified through random digit dialing conducted from 1999-2000 in the San Francisco Bay Area. Controls were frequency matched to cases diagnosed from 1995-1998 on 5-year age group and race/ethnicity, at a ratio of 1 control per 2 cases. Only women were included in the current analysis.

2.9 Nagano Breast Cancer Study (NGOBCS)

The Nagano Breast Cancer Study is a multicenter, hospital-based case-control study which was conducted from May 2001 to September 2005 at four hospitals in Nagano Prefecture, Japan.[30,31] Cases were admitted to the four hospitals during the survey period, and were a consecutive series of women aged 20-74 years with newly diagnosed, histologically confirmed invasive breast cancer. Among the 412 eligible patients, 405 (98%) agreed to participate. Controls were selected from medical checkup examinees in two of the hospitals who were confirmed having no cancer, with one control matched for each case by age (within three years) and residential area during the study period. Only one declined to participate among potential control subjects. Written informed consent was obtained from 405 matched pairs. Since two controls refused to provide blood samples, the analysis was restricted to 403 matched pairs. Participants completed a self-

administered questionnaire, which included questions on demographic characteristics, anthropometric factors, smoking habits, family history of cancer, physical activity, medical history, and menstrual and reproductive history. Dietary habits were investigated using a 136-item semi-quantitative food-frequency questionnaire, which was developed and validated in the Japanese population. The ER status of the patient's breast cancer tissue was obtained from medical records. Hormone receptor positivity values were determined either as specified by the laboratory that performed the assay, in accordance with the laboratory's written interpretation thereof, or both. The study protocol was approved by the Institutional Review Board of the National Cancer Center (Tokyo, Japan).

2.10 Singapore Breast Cancer Cohort (SGBCC)

The SGBCC is an open cohort with a recruitment target of 16,000 patients diagnosed with either breast carcinoma in situ or invasive breast cancer. Details of the study design has been published elsewere.[32] Briefly, recruitment started in 2010. All breast cancer patients who are at least 21 years of age at diagnosis, who are citizens or permanent residents of Singapore and who are attending any of the seven tertiary hospitals are invited to participate in SGBCC. Cases are a mixture of prevalent and incident cases. Three main ethnic groups are represented, namely, Chinese, Malays and Indians. Controls matched by age and ethnicity were selected from the Multi-ethnic Cohort (Phase 2, part of the Singapore Population Health Studies (SPHS).[33] Exclusion criteria for controls included a medical history of cancer, acute myocardial infarction or stroke, or major psychiatric morbidity including schizophrenia, psychotic depression, and advanced Alzheimer's disease.

2.11 Taiwanese Breast Cancer Study (TWBCS)

The study is a part of an ongoing collaborative study with a focus on understanding the cause of breast cancer among Taiwanese.[34,35] Breast cancer patients were recruited from those who were diagnosed and treated at the Tri-Service General Hospital or the Changhua Christian Hospital between March 2002 and August 2005. The controls were randomly selected from women who attended the same hospitals for a comprehensive health examination during the same period. If any evidence of breast cancer, precancerous lesions of breast or other cancers was found, the subject was excluded from the control group. Epidemiologic data were collected from the participants via a structured questionnaire by research nurses. Blood biospecimen was also collected. All the participants provided their informed consent before the data and sample collection.

3. BCAC European samples

Summary statistics data of European descendants from studies involved in the BCAC OncoArray, iCOGS, and GWAS projects were obtained and utilized in the cross-ancestry meta-analysis. Among 82 studies from the BCAC, the OncoArray dataset included 80,125 female cases with breast cancer and 58,383 female controls of European ancestry, and the Collaborative Oncological Gene-environment Study (iCOGS) included 38,349 breast cancer cases and 37,818 controls.[36] In addition, summary statistics from 11 other breast cancer genome-wide association

studies were also used in the meta-analysis with a combined sample of 14,910 cases and 17,588 controls. The genotyping data were imputed by IMPUTE version 2[37] with the 1000 Genomes Project Phase 3 as the reference panel.

II. Supplemental Statistical Analyses

Fine-mapping. We investigated the ancestral heterogeneity of the lead variants at risk loci. However, lead variants are not necessarily the causal variants, and the observed heterogeneity may be related to the different linkage disequilibrium (LD) pattern across populations. Therefore, we performed fine-mapping analyses to construct the 95% credible sets for the lead variants, and further investigated the ancestral heterogeneity of all variants in the credible sets. Fine-mapping analysis was performed using SuSiE[38]. Samples from 1000 Genome Project Phase 3 (EAS and EUR) were used as LD reference. An ancestry-specific LD matrix was used for risk loci identified by ancestry-specific analyses. For risk loci identified by cross-ancestry analyses, a cross-ancestry LD matrix was constructed by combining ancestry-specific LD matrices using weights of population sample sizes.

Gene prediction model building. We used whole genome sequencing (WGS) data in blood samples and RNA sequencing (RNA-seq) data from the Genotype-Tissue Expression Project (GTEx, version 8) to build prediction models for genes expressed in normal breast tissue. All genotyping and expression data were downloaded from dbGap (Accession Number: phs000424.v8.p2).

We kept samples from European-ancestry women with both expression and genotyping data (N =115). The following genetic variants were used to build genetic prediction models: 1) MAF ≥0.05, and 2) Hardy-Weinberg equilibrium $P \geq 10^{-4}$, and 3) call rate ≥ 95%, and 4) non A/T, C/G bi-allelic, and 5) available in BCAC. Finally, a total of 4,853,854 variants were kept for gene expression prediction model building.

There were 32 tissues with both RNA-Seq and WGS data available with sample size >50, and these 32 tissues were kept for model building. Detailed sample sizes by each tissue type were shown in Supplementary Table 10. Within each tissue type, we kept genes with a median expression level (transcript per million, TPM) >0 across samples for each tissue, and the expression level was log2 transformed. Then we performed quantile normalization to bring the expression profile of each sample to the same scale and performed inverse quantile normalization for each gene to the same scale. Then the expression levels were adjusted for age, the top three principal components (PCs) and the top probabilistic estimation of expression residuals (PEER) factors[39] to correct for batch effects and experimental confounders. After adjusting all these covariates, we performed another inverse quantile normalization for the residuals after PEER adjustment of each gene.

We built genetic models to predict gene expression levels in normal breast tissue using the joint-tissue imputation (JTI) approach, which borrows information across transcriptomes of different tissues to improve prediction performance.[40] Besides breast tissue, data from all 31 other tissues were borrowed in the JTI approach to leverage shared genetic regulation and improve prediction performance in a tissue-dependent manner. Gene expression levels were predicted using genetic variants within a flanking +/- 500kb from the respective gene boundaries. Five-fold cross-validation was used to validate the models internally. Genes with a model prediction R >0.1 (≥10% correlation between predicted and observed gene expression) were included for association analyses.

To evaluate the performance of prediction models, we further performed an external validation using 86 tumor-adjacent normal breast tissue samples from European-ancestry female breast cancer patients in the Cancer Genome Atlas (TCGA). Expression data were processed and normalized in similar approach for GTEx data as described above. We calculated the Spearman's correlation between the prediction performance (R^2) in GTEx and TCGA.

Association analyses of predicted gene expression with breast cancer risk. Based on the weight matrix from the prediction models and the summary statistics from meta-analysis of GWAS, we evaluated the association between genetically predicted gene expression and breast cancer risk using the method from the S-PrediXcan tool[41]. The details of the formula used in this method are

$$Z_g \approx \sum_{l \in Model_g} w_{lg} \frac{\widehat{\sigma_l}}{\widehat{\sigma_g}} \frac{\widehat{\beta_l}}{se(\widehat{\beta_l})}$$

In brief, the Z-score was used to estimate the association between predicted gene expression and breast cancer risk. In this formula, w_{lg} is the weight of variant l for predicting the expression of gene g. $\hat{\beta}_l$and $se(\hat{\beta}_l)$ are the association regression coefficient and its standard error for variant l in GWAS, and $\hat{\sigma}_l$ and $\hat{\sigma}_g$ are the estimated variances of variant l and the predicted expression of gene g, respectively. For this study, we estimated the correlations between variants included in the prediction models.

TWAS fine-mapping analyses. We performed TWAS fine-mapping for all genomic regions that contain one or more TWAS-identified risk genes using FOCUS (Fine-mapping Of CaUsal

gene Sets, v0.6.10)[42]. Regions were defined using the correlation matrix of predicted effects on gene expression around TWAS-identified genes. A posterior inclusion probability (PIP) was assigned to each gene for being possibly causal in each TWAS uncovered association signal. Based on the PIP of each gene and a null model, whereby no gene in the region is causal for the TWAS signal, a gene set for each region in which the sum of PIPs for all the genes was greater than or equal to 90% probability ($\sum_{i=1}^{k} nPIP \geq 90\%$) was defined as a credible gene set.

Colocalization analyses. COLOC were conducted to assess the probability that molecular traits as estimated by eQTL and physiological traits as estimated by GWAS share the same causal variant[43]. For each TWAS-identified risk gene, we only estimated variants with both gene-variant paired eQTL results from GTEx and GWAS association statistics (effect size estimate, standard error, and *P* value) and reached association p value less than 0.5. We obtained reference information such as MAF, sample size, and case-to-control proportions (in case of binary traits) for each variant. We defined a gene as having evidence of co-localization when gene-based posterior probability of co-localization PP[4] > 0.5.

Pathway analyses. Protein-coding genes identified by our TWAS were located at 46 GWAS-identified risk loci and seven novel risk loci. If there were multiple TWAS-identified genes at the same locus, genes which were included in the fine-mapping credible set or supported by colocalization analyses were selected for pathway analyses. At 150 additional GWAS-identified loci without protein-coding genes identified by our TWAS, previously reported putative target genes[44] or nearby protein-coding genes were selected for pathway analyses. A total of 221 putative genes for breast cancer were included for pathway analyses (Table S16). The WEB-

based Gene Set Analysis Toolkit (WebGestalt) was used to perform for KEGG pathways and gene ontology terms enrichment analyses[45,46].

Acknowledgements

The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agents. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This research was supported in part by the US National Institutes of Health grants R01CA235553, R01CA202981, R01CA124558, R01CA148667, R01CA158473, R01CA064277, R37CA070867, and UM1CA182910 (to W.Z.); R01CA118229 and R01CA092585 (to X.-O.S.); R01CA122756 (to Q.C.); and R01CA137013 (to J. Long), Department of Defense Idea Awards BC011118 (to X.-O.S.) and BC050791 (to Q.C.), and Ingram and Anne Potter Wilson Professorship and Research Reward funds (to W.Z.). Sample preparation and genotyping assays at Vanderbilt were conducted at the Survey and Biospecimen Shared Resources and Vanderbilt Microarray Shared Resource, which are supported in part by the Vanderbilt-Ingram Cancer Center (P30CA068485). Data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University. 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). KOHBRA/KOGES was supported by a grant from the National R&D Program for Cancer Control, Ministry for Health, Welfare and Family Affairs, Republic of Korea (#1020350). Studies conducted among Asian women include (Principal Investigator, grant support): the Shanghai Breast Cancer Study (W.Z. and X.-O.S., R01CA064277), the Shanghai Women's Health Study (W.Z., R37CA070867 and UM1CA182910), the Shanghai Breast Cancer Survival Study (X.-O. S., R01CA118229), the Shanghai Endometrial Cancer Study (X.-O.S., R01CA092585, controls only), the Seoul Breast Cancer Study [D.K., BRL (Basic Research Laboratory) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2012-0000347)], the BioBank Japan Project (S.-K.L., the Ministry of Education, Culture, Sports, Sciences and Technology from the Japanese Government); the Hwasun Cancer Epidemiology Study-Breast (S.-S.K., the Biobank of Chonnam National University Hwasun Hospital, a member of the Korea Biobank Network, # 07SA2014020), the Nagano Breast Cancer Study (M.I., National Cancer Center Research and Development Fund), the Hospital-based Epidemiologic Research Program at Aichi Cancer Center [Grant-in-Aid for Scientific Research on Priority Areas of Cancer (No. 17015018) from the Japanese Ministry of Education, Culture, Sports, Science and Technology and the "Practical Research for Innovative Cancer Control (15ck0106177h0001)" from the Japan Agency for Medical Research and development, AMED (K. Matsuo), and Cancer Bio Bank Aichi; the Asia Cancer Program (K. Muir and A.L., the NIHR Manchester Biomedical Research Centre and by the ICEP and CRUK, # C18281/A19169); the Canadian Breast Cancer Study (K.A. and J. Spinelli, the Canadian Cancer Society, # 313404); the Los Angeles County Asian-American Breast Cancer Case-Control Study (A.H.W., the California Breast Cancer Research Program [1RB-0287, 3PB-0102, 5PB-0018, 10PB- 0098]. Incident breast cancer cases were collected by the USC Cancer Surveillance Program (CSP) which is supported under subcontract by the California Department of Health. The CSP is also part of the National Cancer Institute's Division of Cancer Prevention and Control Surveillance, Epidemiology, and End Results Program, under contract number N01CN25403); the Malaysian Breast Cancer Genetic Study (S.-H.T., the Malaysian Ministry of Higher Education [UM.C/HlR/MOHE/06] and Cancer Research Malaysia. MYMAMMO is supported by research grants from Yayasan Sime Darby LPGA Tournament and Malaysian

Ministry of Higher Education [RP046B- 15HTM]); the Northern California Breast Cancer Family Registry (E.M.J., the National Cancer Institute [USA, UM1 CA164920]. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR.); the Singapore Breast Cancer Cohort (M.H., the NUS start-up Grant, National University Cancer Institute Singapore [NCIS] Centre Grant and the NMRC Clinician Scientist Award. Additional controls were recruited by the Singapore Consortium of Cohort Studies-Multi-ethnic cohort [SCCSMEC], which was funded by the Biomedical Research Council, grant number: 05/1/21/ 19/425); and the Taiwanese Breast Cancer Study (C.-Y.S., the Taiwan Biobank project of the Institute of Biomedical Sciences, Academia Sinica, Taiwan). Studies conducted among European-ancestry women Genotyping of the OncoArray was principally funded by three sources: the PERSPECTIVE project, funded from 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 the Quebec Breast Cancer Foundation; the NCI Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative and Discovery, Biology and Risk of Inherited Variants in Breast Cancer (DRIVE) project [NIH Grants U19 CA148065, X01HG007492]; and Cancer Research UK [C1287/A10118, C1287/A16563]. The BCAC is funded by Cancer Research UK [C1287/A16563], the European Community's Seventh Framework Programme under grant agreement 223175 [HEALTH-F2- 2009-223175] (COGS).

We also acknowledge the contribution from the following individuals to the SGBCC: Swee Ho Lim [1,2], Ern Yu Tan [3], Benita Kiat Tee Tan [2,4,5], Su-Ming Tan [6], Veronique Kiak Mien Tan [2,4,5], Ching Wan Chan [7], Siau-Wei Tang[7], Celene Wei Qi Ng[7], Geok Hoon Lim[1], Jinnie Siyan Pang[1], Jung Ah Lee[1], Patrick Mun Yew Chan[3], Juliana Chen[3], Sarah Qinghui Lu[3], Yirong Sim [2,4], Wei Sean Yong [2,4,5], Preetha Madhukumar [2,4,5], Fuh Yong Wong [8], Joanne Yuen Yie Ngeow [9,10], Tira Jing Ying Tan [9], Wai Peng Lee [6], Chi Wei Mok [6], Chin Mui Seah [6], Linda Tan [11], E Shyong Tai [11,12], Peh Joo Ho [13], Alexis Jiaying Khng [13]

[1] Breast Department, KK Women's and Children's Hospital, Singapore
[2] SingHealth Duke-NUS Breast Centre, Singapore
[3] Department of General Surgery, Tan Tock Seng Hospital, Singapore
[4] Division of Surgical Oncology, National Cancer Centre Singapore, Singapore
[5] Department of General Surgery, Singapore General Hospital, Singapore
[6] Division of Breast Surgery, Department of General Surgery, Changi General Hospital, Singapore
[7] National University Health System, Department of Surgery, Singapore, Singapore
[8] Division of Radiation Oncology, National Cancer Centre Singapore, Singapore
[9] Division of Medical Oncology, National Cancer Centre Singapore, Singapore
[10] Cancer Genetics Service, National Cancer Centre Singapore, Singapore
[11] Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore
[12] Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore
[13] Genome Institute of Singapore, Human Genetics, Singapore, Singapore

Supplemental references:

1. Zheng W, Long J, Gao YT, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet*. 2009;41(3):324-328.

2. Gao YT, Shu XO, Dai Q, et al. Association of menstrual and reproductive factors with breast cancer risk: Results from the Shanghai breast cancer study. *International Journal of Cancer*. 2000;87(2):295-300.

3. Lum A, Le Marchand L. A simple mouthwash method for obtaining genomic DNA in molecular epidemiological studies. *Cancer Epidemiol Biomarkers Prev*. 1998;7(8):719-724.

4. Zheng W, Chow WH, Yang G, et al. The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. *Am J Epidemiol*. 2005;162(11):1123-1131.

5. Zhang Y, Long J, Lu W, et al. Rare coding variants and breast cancer risk: evaluation of susceptibility Loci identified in genome-wide association studies. *Cancer Epidemiol Biomarkers Prev*. 2014;23(4):622-628.

6. Song HR, Shin MH, Kim HN, et al. Sex-specific differences in the association between ABO genotype and gastric cancer risk in a Korean population. *Gastric Cancer*. 2013;16(2):254-260.

7. Shim HJ, Lee R, Shin MH, Kim HN, Kweon SS. Association between the TCF7L2 polymorphism and colorectal cancer does not differ by diabetes and obesity statuses. *Cancer Epidemiol*. 2016;45:108-111.

8. Kweon SS, Shin MH, Jeong SK, et al. Cohort Profile: The Namwon Study and the Dong-gu Study. *Int J Epidemiol*. 2014;43(2):558-567.

9. Hirata M, Kamatani Y, Nagai A, et al. Cross-sectional analysis of BioBank Japan clinical data: A large cohort of 200,000 patients with 47 common diseases. *J Epidemiol*. 2017;27(3S):S9-S21.

10. Nagai A, Hirata M, Kamatani Y, et al. Overview of the BioBank Japan Project: Study design and profile. *J Epidemiol*. 2017;27(3S):S2-S8.

11. Ishigaki K, Akiyama M, Kanai M, et al. Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat Genet*. 2020;52(7):669-679.

12. Han S, Lee KM, Choi JY, et al. CASP8 polymorphisms, estrogen and progesterone receptor status, and breast cancer risk. *Breast Cancer Res Treat*. 2008;110(2):387-393.

13. Cho YS, Go MJ, Kim YJ, et al. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet*. 2009;41(5):527-534.

14. Han MR, Long J, Choi JY, et al. Genome-wide association study in East Asians identifies two novel breast cancer susceptibility loci. *Hum Mol Genet*. 2016;25(15):3361-3371.

15. Grundy A, Schuetz JM, Lai AS, et al. Shift work, circadian gene variants and risk of breast cancer. *Cancer Epidemiol*. 2013;37(5):606-612.

16. Kobayashi LC, Janssen I, Richardson H, Lai AS, Spinelli JJ, Aronson KJ. Moderate-to-vigorous intensity physical activity across the life course and risk of pre- and post-menopausal breast cancer. *Breast Cancer Res Treat*. 2013;139(3):851-861.

17. Grundy A, Richardson H, Burstyn I, et al. Increased risk of breast cancer associated with long-term shift work in Canada. *Occup Environ Med*. 2013;70(12):831-838.

18. Kobayashi LC, Janssen I, Richardson H, Lai AS, Spinelli JJ, Aronson KJ. A case-control study of lifetime light intensity physical activity and breast cancer risk. *Cancer Causes Control*. 2014;25(1):133-140.

19. Kawase T, Matsuo K, Suzuki T, et al. FGFR2 intronic polymorphisms interact with reproductive risk factors of breast cancer: results of a case control study in Japan. *Int J Cancer*. 2009;125(8):1946-1952.

20. Kwong A, Ng EKO, Law FBF, et al. Novel BRCA1 and BRCA2 genomic rearrangements in Southern Chinese breast/ovarian cancer patients. *Breast Cancer Res Treat*. 2012;136(3):931-933.

21. Kwong A, Ng EKO, Wong CLP, et al. Identification of BRCA1/2 founder mutations in Southern Chinese breast cancer patients using gene sequencing and high resolution DNA melting analysis. *PLoS One*. 2012;7(9):e43994.

22. Kwong A, Shin VY, Au CH, et al. Detection of Germline Mutation in Hereditary Breast and/or Ovarian Cancers by Next-Generation Sequencing on a Four-Gene Panel. *J Mol Diagn*. 2016;18(4):580-594.

23. Han SA, Park SK, Ahn SH, et al. The Korean Hereditary Breast Cancer (KOHBRA) study: protocols and interim report. *Clin Oncol (R Coll Radiol)*. 2011;23(7):434-441.

24. Wu AH, Yu MC, Tseng CC, Stanczyk FZ, Pike MC. Dietary patterns and breast cancer risk in Asian American women. *Am J Clin Nutr*. 2009;89(4):1145-1154.

25. Wu AH, McKean-Cowdin R, Tseng CC. Birth weight and other prenatal factors and risk of breast cancer in Asian-Americans. *Breast Cancer Res Treat*. 2011;130(3):917-925.

26. Wu AH, Vigen C, Butler LM, Tseng CC. Metabolic conditions and breast cancer risk among Los Angeles County Filipina Americans compared with Chinese and Japanese Americans. *Int J Cancer*. 2017;141(12):2450-2461.

27. Phuah SY, Looi LM, Hassan N, et al. Triple-negative breast cancer and PTEN (phosphatase and tensin homologue) loss are predictors of BRCA1 germline mutations in women with

early-onset and familial breast cancer, but not in women with isolated late-onset breast cancer. *Breast Cancer Res*. 2012;14(6):R142.

28. John EM, Hopper JL, Beck JC, et al. The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res*. 2004;6(4):R375-389.

29. Terry MB, Phillips KA, Daly MB, et al. Cohort Profile: The Breast Cancer Prospective Family Study Cohort (ProF-SC). *Int J Epidemiol*. 2016;45(3):683-692.

30. Itoh H, Iwasaki M, Hanaoka T, et al. Serum organochlorines and breast cancer risk in Japanese women: a case-control study. *Cancer Causes Control*. 2009;20(5):567-580.

31. Shimada N, Iwasaki M, Kasuga Y, et al. Genetic polymorphisms in estrogen metabolism and breast cancer risk in case-control studies in Japanese, Japanese Brazilians and non-Japanese Brazilians. *J Hum Genet*. 2009;54(4):209-215.

32. Ho PJ, Yeoh YS, Miao H, et al. Cohort profile: The Singapore Breast Cancer Cohort (SGBCC), a multi-center breast cancer cohort for evaluation of phenotypic risk factors and genetic markers. *PLOS ONE*. 2021;16(4):e0250102.

33. Tan KHX, Tan LWL, Sim X, et al. Cohort Profile: The Singapore Multi-Ethnic Cohort (MEC) study. *Int J Epidemiol*. 2018;47(3):699-699j.

34. Hsu HM, Wang HC, Chen ST, Hsu GC, Shen CY, Yu JC. Breast cancer risk is associated with the genes encoding the DNA double-strand break repair Mre11/Rad50/Nbs1 complex. *Cancer Epidemiol Biomarkers Prev*. 2007;16(10):2024-2032.

35. Ding SL, Yu JC, Chen ST, et al. Genetic variants of BLM interact with RAD51 to increase breast cancer susceptibility. *Carcinogenesis*. 2009;30(1):43-49.

36. Zhang H, Ahearn TU, Lecarpentier J, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat Genet*. 2020;52(6):572-581.

37. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 2009;5(6):e1000529.

38. Wang G, Sarkar A, Carbonetto P, Stephens M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2020;82(5):1273-1300.

39. Stegle O, Parts L, Durbin R, Winn J. A Bayesian Framework to Account for Complex Non-Genetic Factors in Gene Expression Levels Greatly Increases Power in eQTL Studies. *PLOS Computational Biology*. 2010;6(5):e1000770.

40. Zhou D, Jiang Y, Zhong X, Cox NJ, Liu C, Gamazon ER. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nat Genet.* 2020;52(11):1239-1246.

41. Barbeira AN, Dickinson SP, Bonazzola R, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun*. 2018;9(1):1825.

42. Mancuso N, Freund MK, Johnson R, et al. Probabilistic fine-mapping of transcriptome-wide association studies. *Nat Genet.* 2019;51(4):675-682.

43. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet.* 2014;10(5):e1004383.

44. Fachal L, Aschard H, Beesley J, et al. Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nature Genetics*. 2020;52(1):56-73.

45. Liao Y, Wang J, Jaehnig EJ, Shi Z, Zhang B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res*. 2019;47(W1):W199-W205.

46. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*. 2005;33(suppl_2):W741-W748.