A Large-Scale, Exome-Wide Association Study of Han Chinese Women Identifies Three Novel Loci Predisposing to Breast Cancer

Bo Zhang1,2, Men-Yun Chen2, Yu-Jun Shen3, Xian-Bo Zhuo3, Ping Gao4, Fu-Sheng Zhou3, Bo Liang3, Jun Zu3, Qin Zhang4, Sufyan Suleman6, Yi-Hui Xu2, Min-Gui Xu2, Jin-Kai Xu2, Chen-Cheng Liu2, Nikolaos Giannareas4, Ji-Han Xia4, Yuan Zhao3, Zhong-Lian Huang1, Zhen Yang1, Huai-Dong Cheng1, Na Li1, Yan-Yan Hong1, Wei Li1, Min-Jun Zhang1, Ke-Da Yu5, Guoliang Li6, Meng-Hong Sun5, Zhen-Dong Chen1, Gong-Hong Wei4, and Zhi-Min Shao5

Abstract

Genome-wide association studies have identified more than 90 susceptibility loci for breast cancer. However, the missing heritability is evident, and the contributions of coding variants to breast cancer susceptibility have not yet been systematically evaluated. Here, we present a large-scale whole-exome association study for breast cancer consisting of 24,162 individuals (10,055 cases and 14,107 controls). In addition to replicating known susceptibility loci (e.g., ERB1, FGFR2, and TOX3), we identify two novel nonsense variants in C21orf58 (rs13047478, Pmet = 4.52 × 10^{-8}) and ZNF526 (rs3810151, Pmeta = 7.60 × 10^{-8}) and one new non-coding variant at 7q21.11 (P < 5 × 10^{-8}). C21orf58 and ZNF526 possessed functional roles in the control of breast cancer cell growth, and the two coding variants were found to be eQTL for several nearby genes. rs13047478 was significantly (P < 5.00 × 10^{-8}) associated with the expression of genes MCM3AP and YBEY in breast mammary tissues. rs3810151 was found to be significantly associated with the expression of genes PAFAH1B3 (P = 8.39 × 10^{-8}) and CNFN (P = 3.77 × 10^{-8}) in human blood samples. C21orf58 and ZNF526, together with these eQTL genes, were differentially expressed in breast tumors versus normal breast. Our study reveals additional loci and novel genes for genetic predisposition to breast cancer and highlights a polygenic basis of disease development.

Significance: Large-scale genetic screening identifies novel missense variants and a noncoding variant as predisposing factors for breast cancer. Cancer Res; 78(11); 3087–97. ©2018 AACR.

Introduction

Breast cancer is the most common type of cancer and the leading cause of cancer-related deaths in women worldwide (1). Morbidity and mortality associated with breast cancer have increased rapidly in China (2). Although the precise mechanisms underlying this heterogeneous disease have not been fully elucidated, increasing evidence indicated that common genetic variants may contribute to the heritable risk of breast cancer (3). Our understanding of the genetic architecture of breast cancer has been rapidly increased through genome-wide association studies (GWAS), which have identified numerous breast cancer risk-associated variants within more than 90 susceptibility loci (www.genome.gov/gwastudies). However, these variants only explain a small proportion of the genetic variation in breast cancer. The "missing heritability" in breast cancer is evident (4, 5). Furthermore, most of the previously identified variants related to breast cancer susceptibility are located in noncoding genomic regions (6) and thus provide few clues to the functional mechanisms through which these variants affect susceptibility to the disease. Analysis of coding variation could provide more direct biological and functional interpretation for etiology of disease. To assess the role of coding variants with high penetrance that were poorly covered in conventional GWAS may contribute to identifying the "missing heritability" in polygenic disorders (7–9).

Recent technological advances in high-throughput sequencing (10) have provided an opportunity to resequence multiple genetic regions. Such studies have generated compelling evidence that coding variants contribute to the mechanisms of breast cancer (4) and other complex disorders (11–16). Recently, studies employing new exome chips have demonstrated that such chips can be used to comprehensively identify coding variants for several complex traits. Because of the relatively high cost of high-throughput sequencing, exome chips provide a cost-effective approach...
method for the investigation of coding variants. In this study, we sought to identify novel genetic loci predisposing to breast cancer using exome chips in a Chinese population.

**Patients and Methods**

**Study samples**

We implemented a two-stage case–control design in this study. The subjects, consisting of 10,055 cases and 14,107 healthy controls, were enrolled through a collaborative consortium in China (Table 1). All cases were diagnosed by at least two pathologists, and their clinical information was collected through a comprehensive clinical check-up by professional investigators. In addition, demographic information was collected from all participants through a structured questionnaire. All of the healthy controls were clinically determined to be without breast cancer, a family history of breast cancer (and including first- and second-degree relatives). All cases and controls were female. All samples were self-reported Han Chinese. Written, informed consent was given by all participants. The study was approved by the institutional ethics committee of each hospital and was conducted according to the Declaration of Helsinki principles.

**Exome array and genotyping**

In this study, we used custom Illumina Human Exome Asian BeadChip (Exome_Asian Array). The platform includes 242,102 markers focused on putative functional coding variants from >12,000 exome and genome sequences representing multiple ethnicities and complex traits in addition to 30,642 Chinese population–specific coding variants, identified by whole-exome sequencing performed in 676 controls by our group (17). The details of the SNP content and selection strategies are described on the exome array design webpage (http://genome.sph.umich.edu/wiki/Exome_Chip_Design).

In this study, a cohort including 16,066 samples (8,031 cases and 8,035 controls) was genotyped using the Exome_Asian array. The genotyping was conducted at the State Key Lab Incubation and 8,035 controls) was genotyped using the Exome_Asian array. The genotyping was conducted at the State Key Lab Incubation Base of Dermatology, Ministry of National Science and Technology (Anhui Medical University, Heifei, Anhui, China). The genotype calling and the clustering of study sample genotypes were performed using Illumina’s GenTrain (version 1.0) clustering algorithm in Genome Studio (version 2011.1).

**Quality controls**

We excluded samples with genotyping call rates <98% in the first stage. Then, we examined potential genetic relatedness based on pairwise identity by state (IBS) for all the successfully genotyped samples using PLINK 1.07 software (18). On the identification of a first- or second-degree relative pair, we removed one of the two related individuals (the sample with the lower call rate was removed). We defined close relatives as those for whom the estimated genome-wide identity-by-descent proportion of alleles shared was > 0.10. In total, 452 samples (388 cases and 64 controls) were removed due to sample duplication and genetic relatedness. The remaining samples were subsequently assessed for population outliers and stratification using a PCA-based approach (19). For all PCA, all HLA SNPs on chr.6 and SNPs on nonautosomes were removed (Supplementary Fig. S1). Furthermore, we excluded SNPs with a call rate < 99%, a minor allele frequency (MAF) < 0.01, and/or a significant deviation from Hardy–Weinberg equilibrium (HWE) in the controls (P < 10⁻⁵) during each stage. For quality control, 272,744 variants (Exome_Asian Array) were included. After quality control, the genotype data of 33,347 autosomal variants in 8,031 cases and 8,035 controls were included for further analysis.

**SNP selection and genotyping for replication**

To replicate the association results of the Exome_Asian array, we further analyzed the 60 top variants in an additional 8,096 samples (2,024 cases and 6,072 controls; Supplementary Table S1) using the Sequenom MassARRAY system (Sequenom, Inc.) and Multiplex SnapShot technology (Applied Biosystems, Inc.; six of these SNPs failing for primer design). All of these selected SNPs met the following quality criteria: (i) the MAF was higher than 1% in both the cases and controls; (ii) HWE in the controls was P ≥ 0.01 and the HWE in the cases was P > 10⁻⁵; (iii) in each locus, one or two of the most significant SNPs were selected for validation.

**Statistical analyses**

Single-variant association analyses were performed to test for disease–SNP associations, assuming an additive allelic effect and using logistic regression in each stage. The Cochran–Armitage trend test was conducted in these two-stage samples. We performed heterogeneity tests (I² and P values of the Q statistics) between the two groups using the Breslow–Day test (20), and the extent of heterogeneity was assessed using the I² index (21). To improve the statistical power, we combined the association results in two stages using meta-analysis. The fixed effect model (Mantel–Haenszel test; ref. 22) was applied when I² was less than 30%. Otherwise, the random effect model (DerSimonian–Laird; ref. 23) was implemented.

**Cell culture**

The experiments were performed using MCF7, MDA-MB-231, and T47D breast cancer cell lines that were originally purchased from ATCC in 2011, and regularly tested for Mycoplasma. Only Mycoplasma-negative cells were used for experimentation. All the cell lines are usually used for 4 to 9 passages from the initial expansion and frozen down. MCF7 cells were grown in RPMI 1640 medium (Sigma), and MDA-MB-231 was grown in low-glucose DMEM (21885025, Invitrogen). All mediums used for cell culture were supplemented with 10% FBS and 1% penicillin and streptomycin (Sigma). All cells were grown at 37°C with 9.5% air and 5% CO₂.
siRNA transfection

siRNA used in this experiment were purchased from Qiagen and can be found in Supplementary Table S2. The siRNA knockdown assay was performed as described previously (24). In brief, 50%-60% confluent MDA-MB-231 cells were seeded in 6-well plates. Twelve hours later, we performed transfection with siRNA following the instructions of HiPerfect Transfection Reagent (301705, Qiagen). The cells were collected for RNA purification in 48 hours.

Quantitative real-time PCR

PureLink RNA Mini Kit (12183018A, Invitrogen) was used to isolate RNA from cells. The DNA was removed by RNase-Free DNase (79254, QIAGEN). High-Capacity cDNA Reverse Transcription Kit (A368614, Applied Biosystems) was applied to synthesize cDNA from RNA. Quantitative RT-PCR reactions were performed by using the SYBR Select Master Mix (4472908, Applied Biosystems). We selected two high specificity primers for each target. Primer sequences used in this experiment can be found in Supplementary Table S3. For the determination of mRNA levels of each genes, three replications of each gene were performed and the data were normalized against an endogenous ACTB (β-actin) control.

Plasmids, gene and SNP region cloning, and site-directed mutagenesis

The cDNA of human C21orf58 or ZNF526 gene was amplified from a human cDNA library and cloned into pcDNA3.1-V5 vector. Primer sequences are shown in Supplementary Table S3. Site-directed mutagenesis was also made to obtain the G allele at the rs3810151 site of ZNF526 cDNA and the A allele at the rs13047478 site of C21orf58 cDNA in the pcDNA3.1-V5 constructs.

For cloning SNP and promoter regions, the pGL3 basic and pGL3 or pGL4 promoter vectors (Promega) were used, both vectors encode luciferase-reporter gene luc2 (Photinus pyralis). Experimental inserts of 787 bp for rs13047478 (chr21: 47734341-47735127, GRCh37/hg19) and 751 bp for rs3810151 (chr19: 42728444-42729194, GRCh37/hg19) were amplified from the VCaP genomic DNA using the cloning primers listed in Supplementary Table S3. The inserts were cloned upstream of luciferase gene into pGL3 promoter vectors, the insert sequences and alleles of SNPs rs13047478 and rs3810151 were confirmed by sequencing. The allele determined were as rs13047478-A and rs3810151-A. For the measurement of the allele-specific enhancer activity, the determined alleles were mutated to rs13047478-G and rs3810151-G by using site-directed mutagenesis primers (Supplementary Table S3). To eliminate the possibility of enhancer activity from regions other than SNP-containing regions, two control regions from C21orf58 gene were selected and cloned into pGL3 promoter vector. The fragment I is an intergenic region between YBEY and C21orf58, fragment II is a random intronic region of gene C21orf58.

To test and validate the allele-specific impact of rs13047478 on gene-specific promoter regions, MCM3AP and YBEY promoter regions were cloned downstream of SNP regions and upstream of luciferase gene into pGL3 basic vector using the primers listed in (Supplementary Table S3). In addition, promoter regions of MCM3AP and YBEY were also cloned into pGL3 basic vector to run as control along with the empty pGL3 basic vector. All cloning inserts were confirmed by sequencing. In all reporter assays, a Renilla luciferase reporter vector pGL75 (Promega) was used as an internal control to compare transfection efficiency.

Transient transfection

For plasmid transfection (pcDNA3.1-V5-C21orf58/ZNF526) on 6-well culture plates, 1.5 × 10^6 breast cancer MCF7 cells per well were applied. Transient transfections were applied using Lipofectamine 3000 Transfection Reagent (Thermo Fisher Scientific) following the manufacturer’s instructions. After 48 hours, cells were harvest for protein blot analysis.

Western blot analysis

Cell lysate was prepared in lysis buffer (600 mmol/L NaCl, 1% Triton X-100 in PBS). Protein samples were denatured in 1× SDS loading buffer (Thermo Fisher Scientific) and 100 mmol/L DTT, separated by 10% SDS-PAGE and blotted onto 0.45-μm polyvinylidene difluoride transfer membrane (Immobilon-P, Millipore) with a Semi-Dry transfer cell (Trans-Blot SD, Bio-Rad). Membranes were blocked with 5% nonfat milk (Cell Signaling Technology) in TBST (50 mmol/L Tris-HCl, pH 7.5, 150 mmol/L NaCl, 0.05% Tween-20) and then exposed to antibodies (1:5,000 dilution) targeting V5 tag (monoclonal antibody, HRP, R961-25, Invitrogen) and actin (ab20272, abcam). Membranes were developed with SuperSignal West Femto Maximum Sensitivity Substrate (34095, Thermo Scientific). Membranes were imaged with a LAS-3000 Luminescent Image Analyzer (Fujifilm).

Cell viability and proliferation assays

HiPerFect Transfection Reagent (301705, Qiagen) was used for reverse transfection of control siRNA (1027281), C21orf58 siRNA (SI04208631, SI04282838, SI04314142 and SI04367083), and ZNF526 siRNA (SI00775593, SI04130966, SI04205859, and SI04230422) into MDA-MB-231 cells. The detailed protocol was as described previously (25). Briefly, 6 μmol/L siRNA was diluted in 18 μmol/L optiMEM, and then 1.5 μL HiPerFect Transfection Reagent was added and mixed for each well of 96-well plate. After 10-minute incubation, cells (2.5 × 10^4 per well) were added. We added XTT (11465015001, Roche) reagent at the time point of 3 days and 5 days, measured the absorbance at 450 nm according to the manufacturer’s instructions. Data were collected from five replicate wells and analyzed by the two-tailed t test to determine the significances between different target cells at each time point.

Transfection and enhancer reporter assay

For enhancer reporter assay, the cells were cultured in white 96-well plate with 100-μL suspension of MCF7 at 4 × 10^5 cells/mL per well. Cells were reverse transfection according to the manufacturer’s protocol with 100 ng of experimental and control plasmids, and 4 ng of pGL75 as an internal control/well using X-tremeGENE HP DNA Transfection Reagent (Roche) and incubated at 37°C. After 48 hours, the luciferase activity was measured using the Dual-Glo Luciferase Assay System (Promega) by following the manufacturer’s protocol. In all transfection reactions, three replicates were made, which were compared with control plasmid for enhancer activity determination.

Differential gene expression analysis of the clinical breast cancer datasets

The differential gene expression analyses were performed to identify which transcripts/genes from the breast cancer tumor samples were being produced at a significantly higher or lower
level than that in the healthy tissues. Our study involved a batch of datasets from Oncomine (26) and TCGA cohorts (27). Genes with missing expression value in more than 50% of total samples were not taken for analysis. Mann–Whitney U test was employed to investigate the differential gene expression between the tumor samples and normal samples. Kruskal–Wallis H test was also applied in certain datasets where more than two groups were available. Figures were produced with R (28).

Survival analysis
To investigate the association of the expression of certain genes and the overall survival or biochemical relapse rate in breast cancer, we analyzed a batch of datasets from Oncomine (26) and TCGA (27). Gene expression values from Oncomine are microarray-based while RNA-seq–based in TCGA.

The Kaplan–Meier survival function was applied in the survival analyses. The idea is to define the probability of surviving to a certain time period. The survival probability or biochemical event–free probability at any particular time period is calculated by the formula shown below:

In our analyses, we did not take into account the genes whose expression data were missing in more than 50% of total samples. We used the average gene expression value as a measurement to stratify tumor samples into two groups. The strategy for stratification is illustrated as follows:

Higher expression: Expression of gene X in subject i > mean (gene X in all subjects k)
Lower expression: Expression of gene X in subject i < mean (gene X in all subjects k)

Where, i ∈ [1, k]

Compared with the sample mean, subjects with higher gene expression value were defined as higher expression, while lower expression subjects were classified as lower expression category. In other words, subjects with positive/negative scores indicate higher/lower gene expression compared with the average expression. We then performed the Kaplan–Meier function based on the stratification, and figures were plotted using R package ‘survplot’ (28) with modifications fitting to own needs.

Ethics approval and consent to participate
All participants have given written and informed consent. The study was approved by the institutional ethics committee of each hospital and was conducted according to the Declaration of Helsinki principles.

Results
Overview of exome-wide association analyses
To identify novel loci conferring susceptibility to breast cancer, we conducted a large-scale exome-wide association study using a two-stage case–control design (10,055 cases and 14,107 controls; Table 1) in a Han Chinese population by using the Illumina Human Exome_Asian Array (Illumina, Inc.), Sequenom MassArray system (Sequenom, Inc.), and Multiplex Snapshot technology (Applied Biosystems, Inc.). In the discovery stage, 272,744 markers were genotyped in 8,031 cases and 8,035 controls using the Exome_Asian Array (Fig. 1). After quality control and principal component analysis (Supplementary Methods), 33,347 non-MHC single-nucleotide polymorphisms (SNP) were identified and selected for further analysis. Quantile–quantile (QQ) plots and Manhattan plots were generated using the Cochran–Armitage test for trend (Fig. 2; Supplementary Fig. S1–S2). A clear deviation from the expected null distribution was observed in the QQ plot (Supplementary Fig. S2).

Using the genome-wide results from the discovery stage, we first investigated the evidence for the previously reported GWAS loci in breast cancer. To date, 125 breast cancer susceptibility SNPs within 98 loci have been discovered at genome-wide significance ($P < 5 \times 10^{-8}$) (Supplementary Table S1). In this study, twenty-three of these SNPs were directly covered by the Exome_Asian array and passed our quality control. Significant associations were observed for 15 known breast cancer risk variants within 10 loci such as $CCDC170$ (rs3734805, $P = 7.10 \times 10^{-7}$), $ESR1$ (rs2046210, $P = 9.08 \times 10^{-7}$), $TOX3$ (rs4784227, $P = 3.19 \times 10^{-7}$), $TNRC9$ (rs3803662, $P = 4.99 \times 10^{-11}$), and $FGFR2$ (rs2981579, $P = 7.80 \times 10^{-8}$) and rs1219648, $P = 2.69 \times 10^{-7}$; Supplementary Table S5; Supplementary Fig. S3). Our data also showed nominal association for another three previously reported SNPs within three loci ($ANKLE1$ rs3263956, $FGF10$ rs4415084, and $PTPN22$ rs11552449; $P < 0.05$ each). For the 23 SNPs, most of them also showed effect in the same direction as the previously reported studies, although some variants displayed no significant associations with breast cancer risk in our study cohort. Together, these analyses not only confirmed the associations of these 15 SNPs within 10 independent reported loci with breast cancer in Chinese population, but also ensure the good quality of the genotype data obtained from the Exome_Asian Array for our downstream analyses.

Discovery of new susceptibility loci for breast cancer
To identify true genetic factors and novel susceptibility loci for breast cancer, we next selected the top 60 SNPs with Pvalues of less than $10^{-6}$ (Supplementary Table S1) for the stage 2 of replication study (six of them failing for primer design). These selected SNPs were further genotyped in an independent replication cohort including 2,024 cases and 6,072 controls. We evaluated these SNPs passed in the replication stage for achieved nominal association evidence without Bonferroni correction, similar to the research methods adopted in the previous studies (29–31). In the process of quality control stage, we also evaluated association heterogeneity for these SNPs in the discovery and the replication studies. After quality control at the replication stage, 14 SNPs at 14 loci were considered as associated with breast cancer ($1.93 \times 10^{-22} < P < 4.95 \times 10^{-12}$; Table 2; Supplementary Tables S1 and S6). Meta-analysis of the SNPs in the discovery (stage 1) and replication (stage 2) studies identified two new missense variants, including $ZNF526$ (rs3810151, $P_{meta} = 7.60 \times 10^{-8}$), and $C21orf58$ (rs13047478, $P_{meta} = 4.52 \times 10^{-8}$). In addition, a new noncoding variant were identified at $7q21.11$ (rs7807771, $P_{meta} = 6.71 \times 10^{-8}$; Table 2; Supplementary Figs. S4 and S5). Notably, none of these three SNPs exhibited any significant heterogeneity in the discovery and the replication studies. For the three novel SNPs, we analyzed the Linkage disequilibrium (LD) patterns using the genotyping data (only SNPs with MAF > 0.01) from our Illumina Human Exome_Asian Array data in the Haploview (Supplementary Fig. S6; ref. 32).

Functional analyses of $C21orf58$ and $ZNF526$ in breast cancer
In the meta-analysis of this exome-wide association study, we discovered two novel coding variants associated with breast cancer, rs13047478 within the exon of $C21orf58$ (chromosome 21 open reading frame 58) at $21q22.3$ and rs3810151 in the zinc
finger protein 526 gene (ZNF526) at 19q13.2. The role of C21orf58 and ZNF526 at these two loci in breast cancer development remains totally unknown. Using the data of genome-wide CRISPR-Cas9-based loss-of-function screens in 33 cancer cell lines for the identification of genes that are essential for cell growth and survival (33), we found that C21orf58 and ZNF526 were important for the survival of breast cancer cells (Fig. 3A and B; Supplementary Fig. S7), suggesting that these genes possess unknown function in the control of breast cancer cell growth. Consistent with this, our cell proliferation assays showed that breast cancer cells harboring short interfering RNA (siRNA) against C21orf58 or ZNF526 showed markedly reduced cell growth and viability compared to cells harboring control siRNA (Fig. 3C and D; Supplementary Fig. S8). Furthermore, using the Oncomine analysis tool (26), we compared the mRNA expression levels of C21orf58 and ZNF526 in the Finak breast cancer dataset (34), and found that both genes were highly expressed in invasive breast carcinoma compared to normal breast (Fig. 3E). The analysis of several additional large-scale clinical datasets (27, 35, 36) showed that C21orf58 and ZNF526 were greatly upregulated in breast cancer in comparison with normal breast tissues (Fig. 3F and G; Supplementary Fig. S9A). Furthermore, high mRNA levels of C21orf58 or ZNF526 showed marginal associations with poor survival in multiple independent cohorts of breast cancer patients (Supplementary Figs. S9B–S9D; refs. 37–39). Together, our analyses reveal a previously unknown role of C21orf58 and ZNF526 in breast carcinogenesis.

Functional annotation of the variants at the three novel breast cancer susceptibility loci

The SNP rs13047478 at 21q22.3 is located within the exon of C21orf58, which results in an amino acid change of proline to serine. C21orf58 is an uncharacterized gene, and its role in breast cancer remains completely unknown. Here, we provide experimental and clinical evidence of a potential role for C21orf58 in breast cell growth and carcinogenesis (Fig. 3; Supplementary Figs. S7–S9). We next cloned C21orf58 with different alleles of rs13047478 into mammalian expression vector pcDNA3.1-V5 and examined the effect of rs13047478 on C21orf58 expression by transient transfection in the breast cancer cell line MCF7 (see Materials and Methods). We found that the expression levels of C21orf58 with the A allele of rs13047478 are approximately...
1.5-fold higher than that of \( C21orf58 \) with the G allele (Supplementary Fig. S10A). In contrast, we observed no impact of the rs13047478-associated amino acid change on the function of \( C21orf58 \) in the growth control of breast cancer cell line (Supplementary Fig. S10B).

Despite the fact that rs13047478 in \( C21orf58 \) is a coding variant functioning on amino acid substitution, by querying a large collection of ChIP-seq datasets (40), we observed the binding of multiple transcription factors and epigenetic features at rs13047478 region (Supplementary Fig. S11), suggesting this SNP-containing genomic region may be a possible exonic transcriptional regulatory element with impact on gene expression (41). Consistent with this observation, our enhancer luciferase reporter assay showed that rs13047478 region may possesses enhancer activity (Fig. 4A). We also mapped rs13047478 within several transcription factor DNA-binding positional weight matrix (PWM) derived from HaploReg database (Supplementary Table S7; refs. 42, 43) but found no direct impact of rs13047478 on PWMs of the transcription factors enriched at rs13047478 region (Supplementary Fig. S11).

We next performed the eQTL analysis using the Gene-Tissue Expression (GTEx) database (44) and unexpectedly revealed a significant association of rs13047478 with several genes across many types of human tissues and cells. In particular, we found that rs13047478 was in the eQTLs for the genes \( MCM3AP \) \((P = 4.90 \times 10^{-10})\) and \( YBEY \) \((P = 3.00 \times 10^{-11})\) in normal breast tissues (Fig. 4B and C). Chromatin looping data (45) indicated direct physical interactions among rs13047478/C21orf58, \( MCM3AP \) and \( YBEY \) in the breast cancer cell line MCF7 (Fig. 4D). To directly test the effect of the rs13047478-containing enhancer in \( MCM3AP \) or \( YBEY \) regulation, we inserted rs13047478-centered DNA fragment upstream of the \( MCM3AP \) or \( YBEY \) promoter in a pGL3-Basic vector and performed luciferase reporter assays in MCF7 cells (Fig. 4E and F). The results showed that, compared with the A allele of rs13047478, the G allele indicated a lower activity on the basal \( MCM3AP \) promoter and higher activity on the \( YBEY \) promoter, consistent with the eQTL results, showing a significant association of the G allele of rs13047478 with decreased mRNA levels of \( MCM3AP \) and elevated expression of \( YBEY \) in breast mammary tissues (Fig. 4B and C). Collectively, these analyses suggest the causal effect of rs13047478 on the expression of \( MCM3AP \) and \( YBEY \).

A recent study showed that the expression of \( MCM3AP \) was significantly decreased in human breast tumors (46). In addition, \( MCM3AP \) can serve as an independent predictor and its lower expression was associated with poor prognosis of patients with breast cancer. Mammary gland–specific \( MCM3AP \) knockout mice showed severe impairment of mammary gland development during pregnancy and were more likely to develop mammary gland tumors (46). Moreover, tumor formation also occurred in female mice with \( MCM3AP \) heterozygosity. In addition, \( MCM3AP \) plays a significant role in the suppression of DNA damage caused by estrogen in human breast cancer cell lines (46). Together, these results indicated that the \( MCM3AP \) is associated with breast cancer resistance. Consistent with these observations, we found that \( MCM3AP \) was greatly downregulated in breast cancer compared with normal breast samples in a cohort of over 2000 breast cancer patients (Supplementary Fig. S12A; ref. 34). Furthermore, we observed that lower expression of \( MCM3AP \) showed a strong association with decreased metastasis-free survival in a collection of 195 breast tumors (Supplementary Fig. S12B; ref. 47). Together, these results suggest a causal role for \( MCM3AP \) in breast cancer.

Although there were no suggested links of the other rs13047478 eQTL gene \( YBEY \) to breast cancer, we observed that \( YBEY \) is likely to be essential for the growth and survival of ER-positive breast cancer cells (Supplementary Fig. S13A and S13B). Furthermore, we found that \( YBEY \) was greatly upregulated in breast cancers in comparison with adjacent normal breast tissues (Supplementary Fig. S13C; ref. 34). Notably, higher expression of \( YBEY \) and lower mRNA levels of \( MCM3AP \) in breast cancer are also consistent with stronger activity of \( YBEY \) promoter and weaker activity of \( MCM3AP \) promoter, respectively, observed in the luciferase reporter experiments in MCF7 cells (Fig. 4E and F). Altogether, these analyses raised potential roles of the rs13047478
eQTL genes MCM3AP and YBEY in breast cancer, and indicated a likely function of rs13047478 as an exonic enhancer variant. Further work will be needed to address whether rs13047478 could tag regulatory genetic variants within active transcription factor–binding sites and regulatory enhancers in regulating the expression of MCM3AP and YBEY conferring breast cancer susceptibility.

The SNP rs3810151 at 19q13.2 is located within ZNF526 gene, and a missense variant where the minor allele G results in a valine to alanine amino acid change in the ZNF526. This single amino acid change also showed a slight effect on ZNF526 expression (Supplementary Fig. S10C). ZNF526 is a member of zinc finger protein family and its exact function is unclear. Interestingly, many zinc finger proteins have been reported to be associated with breast cancer, such as ZNF365 (25, 48, 49) and ZNF545 (50). Here, we show that ZNF526 may be essential for the survival of breast cancer cells, and a striking upregulation of ZNF526 in breast cancer samples compared with normal, indicating a potential function of ZNF526 in breast tumorigenesis (Fig. 3; Supplementary Figs. S7–S9).

Similar to rs13047478, we observed that the rs3810151-containing region in ZNF526 was also featured as a cis-regulatory element with several transcription factor binding, active chromatin marks, and enhancer activity (Supplementary Fig. S14A and S14B; ref. 40) that may impact gene regulation. Regulatory motif analysis of the DNA sequence surrounding rs3810151 showed that rs13047478 may alter PWMs of several transcription factors (Supplementary Table S7; refs. 42, 43) but not for ESR1, FOXA1, MYC, and so on occupied at rs3810151 region (Supplementary Fig. S14A; ref. 40). We next performed the eQTL analysis to examine whether the variant rs3810151 correlate with expression of nearby genes, using a publicly available database (51). This analysis revealed a significant cis-association of rs3810151 with the expression of two genes: PAFAH1B3 ($P = 8.39 \times 10^{-5}$) and CNFN ($P = 3.77 \times 10^{-4}$; Supplementary Fig. S14C). Consistent with this, a long-range chromatin interaction between rs3810151/ZNF526 and PAFAH1B3 was observed in the MCF7 breast cancer cells using ChiP–PET method (Supplementary Fig. S15; ref. 45). PAFAH1B3 has been previously identified as a key metabolic driver of breast cancer pathogenicity, which is upregulated in primary human breast tumors, and correlated with poor prognosis (52). Metabolomic profiling suggests that PAFAH1B3 inactivation attenuates cancer pathogenicity through enhancing tumor-suppressing signaling lipids (52). Consistent with this putative oncogenic role of PAFAH1B3 in breast cancer, our analysis of multiple large-scale breast cancer datasets revealed that PAFAH1B3 was highly expressed in breast tumor samples and significantly associated with poor prognosis of the patients with breast cancer (Supplementary Fig. S16; refs. 27, 35, 36, 38, 47, 53). We also observed a significant upregulation of the other rs3810151 eQTL gene CNFN in breast tumor tissues (Supplementary Fig. S17A and S17B; refs. 27, 35), and increased risk for relapse in breast cancer patients harboring the tumors with high mRNA levels of CNFN expression (Supplementary Fig. S17C; ref. 54).

In addition to the two novel regulatory exonic SNPs, we identified a new noncoding variant rs7807771 (7q21.11). Regulatory motif analysis indicated that rs7807771 may impact several transcription factor DNA-bind PWMs (Supplementary Table S7; refs. 42, 43), whereas observed no enrichment of transcription factor binding and active chromatin marks at this region based on
Figure 3.
The rs13047478-associated gene C21orf58 and the rs3810151-associated ZNF526 show potential effects on breast cancer cell survival and display high expression in breast cancer tissues. A and B, Genome-wide loss-of-function screening of the genes that are essential for the survival of the ER-negative breast cancer cell line CAL120 (A) and the ER-positive breast cancer cell line T47D (B). Lower ATARiS values indicate elevated dependency of the cells on given genes. BRD4, CCND1, and MYC are known to be important for breast cancer cell growth and survival (29). Note that the essentiality is strikingly higher for C21orf58 in comparison with CCND1 (A), and ZNF526 with MYC (B). C and D, Cell proliferation was measured at 5 days (C) and 3 days (D), respectively, by XTT colorimetric assay (absorbance at 450 nm (OD450); mean ± SD of five technical replicates. **, P < 0.01; ***, P < 0.001; ****, P < 0.0001, two-tailed Student t test. E, The expression levels of two genes in cancerous and normal tissues of the patients with breast cancer. Oncomine analysis of the Finak data set (33) of 53 invasive breast tumors and 6 normal samples shows the expression of both genes to be deregulated (P < 10^-5) in breast cancer. Colors indicate z-score normalized to depict relative values in each row. F and G, C21orf58 (F) or ZNF526 (G) mRNA expression was significantly upregulated in human breast cancers. The horizontal lines represent the median values. The P values were calculated using Mann–Whitney U tests. The analyses are based on the datasets from Curtis and colleagues (34) and TCGA at the Oncomine database.
Figure 4.
Association of rs13047478 genotype with the expression of MCM3AP and YBEY, and the chromatin interaction between rs13047478/C21orf58 and the two eQTL genes. 

A, Luciferase reporter assays showing an enhancer activity of rs13047478-centered genomic region in comparison with the control of pGL3-promoter vector. Note that the randomly selected control region 1 and 2 show no enhancer activity. NS, not significant. 

B and C, The expression quantitative trait locus analysis of the breast cancer risk-associated SNP rs13047478. The expression quantitative trait locus analyses indicate that rs13047478 is associated with the expression of two genes MCM3AP (B) and YBEY (C), respectively, in breast tissue samples. Note that the risk allele G of rs13047478 is associated with decreased mRNA levels of MCM3AP and increased expression of YBEY. The analyses are based on the source data of the GTEx project (43). 

D, The chromatin interactions among C21orf58 (rs13047478), MCM3AP, and YBEY were defined by ChIA-PET experiments in the breast cancer cell line MCF-7 (44). 

E and F, Luciferase reporter assays indicate increased enhancer activity with the A allele of rs13047478 relative to the G allele for the MCM3AP promoter (E) and reversely for the YBEY promoter. In A, E, and F, error bars show mean ± SD (n = 4 technical replicates). The P values were evaluated using two-tailed Student t tests. ***, P < 0.001; ****, P < 0.0001.
the query of a large collection of ChIP-seq data [40]. rs7807771 is located 332-kb upstream of SEMA3D. SEMA3D is a member of the class-3 semaphorin family, and could inhibit tumor development through affecting the expression of appropriate semaphorin receptors [55]. The expression SEMA3D is increased in pancreatic ductal adenocarcinoma (PDA) tumors. Mouse PDA cells in which SEMA3D was knocked down exhibited decreased invasive and metastatic potential in culture and in mice [56]. Consistent with its oncogenic property, we showed that SEMA3D was highly upregulated in breast cancer tissues (Supplementary Fig. S18; ref. 35).

Protein–protein interaction and pathway enrichment analysis

Finally, we evaluated the connectivity at the protein–protein interaction (PPI) level for the genes at 98 previously reported GWAS loci in breast cancer (Supplementary Table S4) and the three novel loci discovered in this study (Table 2). Using the search tool to retrieve the interacting genes/proteins (STRING database; ref. 57), we observed a significant PPI enrichment (P value: 0, hypergeometric test; Supplementary Fig. S19) for these genes, suggesting at least partially functional and biological connections among them. The most significantly overrepresented pathways are related to the mammary gland epithelium development (Supplementary Table S8; P = 2.62 × 10⁻⁶).

Discussion

In this study, to the best of our knowledge, we performed the first comprehensive assessment of coding variation using the exome array for breast cancer in Han Chinese women. This analysis led to the identification of two novel missense variants within two uncharacterized genes (ZNF526 and C21orf58) in breast cancer, and a new noncoding variant at 7q21.11. These loci have not been discovered in previous GWAS and other genetic association studies in breast cancer. We demonstrated that ZNF526 and C21orf58 played roles in breast cancer cell growth and disease progression. We unexpectedly found that the two missense variants function as regulatory coding SNPs in the eQTLs with several genes including MCM3AP, YBEY, PAFAH1B3, and CNFN that are potentially important for breast cancer. Our findings suggest that genetic variants and genes at these loci contribute to the development of breast cancer. Our work highlights polygenic contributions to the pathogenesis of breast cancer and identifies additional susceptibility loci for breast cancer. We acknowledge that, although the overall evidence reaches genome-wide significance for the three newly identified breast cancer risk loci, the sample size and thus the power of the validation study are rather limited. Further studies in large independent samples will be needed to replicate these findings. In addition, future studies involving fine-mapping of targeted regions and deep functional studies are required to delineate the molecular mechanisms underlying the risk variants identified in this study.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: B. Zhang, Y.-J. Sheng, H. Cheng, Y.-Y. Hong, K.-D. Yu, G.-H. Wei


Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): B. Zhang, M.-Y. Chen, X.-B. Zhao, Q. Zhang, S. Suleman, J.-K. Xu, C.-C. Liu, Z. Yang, Y.-Y. Hong, G. Li, Z.-D. Chen, G.-H. Wei

Writing, review, and/or revision of the manuscript: B. Zhang, Y.-J. Sheng, S. Suleman, H. Cheng, Y.-Y. Hong, K.-D. Yu, G.-H. Wei

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): B. Zhang, Y. Zhao, Y.-Y. Hong, K.-D. Yu, M.-H. Sun, G.-H. Wei

Study supervision: B. Zhang, H. Cheng, Y.-Y. Hong, G.-H. Wei, Z.-M. Shao

Acknowledgments

We thank the State Key Laboratory Incubation Base of Dermatology, Ministry of National Science and Technology (Hefei, China) for providing research platform. This work was supported by the Young Program of the National Natural Science Foundation of China (81301771; awarded to B. Zhang), the Academy of Finland (284618 and 279760), University of Oulu Strategic Funds, and Jane & Aatos Erkko Foundation grants awarded (to G.-H. Wei).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received June 14, 2017; revised October 25, 2017; accepted March 20, 2018; published first March 23, 2018.

Published OnlineFirst March 23, 2018; DOI: 10.1158/0008-5472.CAN-17-1721


A Large-Scale, Exome-Wide Association Study of Han Chinese Women Identifies Three Novel Loci Predisposing to Breast Cancer

Bo Zhang, Men-Yun Chen, Yu-Jun Shen, et al.


Updated version
Access the most recent version of this article at:
doi:10.1158/0008-5472.CAN-17-1721

Supplementary Material
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2018/03/23/0008-5472.CAN-17-1721.DC1

Cited articles
This article cites 56 articles, 7 of which you can access for free at:
http://cancerres.aacrjournals.org/content/78/11/3087.full#ref-list-1

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, use this link
http://cancerres.aacrjournals.org/content/78/11/3087.
Click on "Request Permissions" which will take you to the Copyright Clearance Center's (CCC) Rightslink site.