Association between Ancestry-Specific 6q25 Variants and Breast Cancer Subtypes in Peruvian Women

Valentina A. Zavala[1], Sandro Casavilca-Zambrano[2], Jeannie Navarro-Vásquez[2], Carlos A. Castañeda[2], Guillermo Valencia[2], Zaida Morante[2], Monica Calderón[2], Julio E. Abugattas[2], Henry Gómez[2], Hugo A. Fuentes[2], Ruddy Liendo-Picoaga[2], Jose M. Cotrina[2], Claudia Monge[2], Silvia P. Neciosup[2], Scott Huntsman[3], Donglei Hu[3], Sixto E. Sánchez[4], Michelle A. Williams[5], Angel Núñez-Marrero[6], Lenin Godoy[6], Aaron Hechmer[7], Adam B. Olshen[7,8], Julie Dutil[6], Elad Ziv[3], Jovanny Zabaleta[9], Bizu Gelaye[5,10], Jule Vásquez[2], Marco Gálvez-Nino[2], Daniel Enriquez-Vera[11], Tatiana Vidaurre[2], and Laura Fejerman[1,12]

ABSTRACT

Background: Breast cancer incidence in the United States is lower in Hispanic/Latina (H/L) compared with African American/Black or Non-Hispanic White women. An Indigenous American breast cancer–protective germline variant (rs140068132) has been reported near the estrogen receptor 1 gene. This study tests the association of rs140068132 and other polymorphisms in the 6q25 region with subtype-specific breast cancer risk in H/Ls of high Indigenous American ancestry.

Methods: Genotypes were obtained for 5,094 Peruvian women with (1,755) and without (3,337) breast cancer. Associations between genotype and overall and subtype-specific risk for the protective variant were tested using logistic regression models and conditional analyses, including other risk-associated polymorphisms in the region.

Results: We replicated the reported association between rs140068132 and breast cancer risk overall [odds ratio (OR), 0.53; 95% confidence interval (CI), 0.47–0.59], as well as the

lower odds of developing hormone receptor negative (HR⁻) versus HR⁺ disease (OR, 0.77; 95% CI, 0.61–0.97). Models, including HER2, showed further heterogeneity with reduced odds for HR⁺HER2⁺ (OR, 0.68; 95% CI, 0.51–0.92), HR⁻HER2⁺ (OR, 0.63; 95% CI, 0.44–0.90) and HR⁻HER2⁻ (OR, 0.77; 95% CI, 0.56–1.05) compared with HR⁺HER2⁻. Inclusion of other risk-associated variants did not change these observations.

Conclusions: The rs140068132 polymorphism is associated with decreased risk of breast cancer in Peruvians and is more protective against HR⁻ and HER2⁺ diseases independently of other breast cancer-associated variants in the 6q25 region.

Impact: These results could inform functional analyses to understand the mechanism by which rs140068132-G reduces risk of breast cancer development in a subtype-specific manner. They also illustrate the importance of including diverse individuals in genetic studies.

Introduction

Breast cancer is the most common cancer and the second leading cause of cancer-related death among women in the United States (US; ref. 1). Analyses stratified by race and ethnicity categories have revealed differences in breast cancer incidence and subtype distribution among diverse populations (2). These differences are the result of structural, environmental, and genetic factors (3–6).

The genetic background of Latin American populations is shaped by the population history of the Americas and colonization (7–10), which

led to vast geographical variation in the average contribution of the different continental ancestry components (7, 10). Estimates of average European ancestry in individuals from Latin American countries vary from 84% in Uruguay (11), to 18% in Peru (12), with intermediate values in countries such as Mexico or Chile (10, 12). A complementary range of variation in Indigenous American ancestry has been described, with countries like Peru and Bolivia having average Indigenous American ancestry estimates of approximately 80% (10, 12, 13). The African ancestry component in Latin America varies from up to approximately 77% in the Caribbean (14, 15) to less than 5%

[1]Department of Public Health Sciences, University of California Davis, Davis, California. [2]Instituto Nacional de Enfermedades Neoplasicas, Lima, Peru. [3]Division of General Internal Medicine, Department of Medicine, University of California, San Francisco, San Francisco, California. [4]Universidad Peruana de Ciencias Aplicadas, Lima, Peru and Asociación Civil Proyectos en Salud (PROESA), Lima, Peru. [5]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts. [6]Department of Biochemistry, Cancer Biology Division, Ponce Research Institute, Ponce Health Sciences University, Ponce, Puerto Rico. [7]Helen Diller Comprehensive Cancer Center, University of California San Francisco, San Francisco, California. [8]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California. [9]Department of Pediatrics and Stanley S. Scott Cancer Center LSUHSC, New Orleans, Louisiana. [10]Department of Psychiatry, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts. [11]Universidad Privada San Juan Bautista, Lima, Peru. [12]UC Davis Comprehensive Cancer Center, University of California Davis, Davis, California.

in countries such as Peru or Argentina (10). Peru is among the countries with the largest contribution of Indigenous American ancestry in Latin America making it an ideal population for the identification of Indigenous American-specific trait/disease-associated genetic variants.

The first breast cancer genome-wide association study (GWAS) carried out in Hispanics/Latinas (H/L) from the US and Mexico identified a single-nucleotide polymorphism (SNP), rs140068132A>G, that was correlated with Indigenous American ancestry and associated with lower odds of developing breast cancer (16). This variant was observed at relatively high frequency in Latin American populations, with frequencies varying from 5% in Puerto Rico to 23% in Peru (17). In addition to being protective against breast cancer generally, it showed lower odds of estrogen receptor-negative (ER$^-$) disease compared with ER$^+$ disease (16).

The rs140068132 SNP is located within an intergenic enhancer close to the ER gene (*ESR1*) at 6q25, shown to be a transcription factor–binding site (16). This region harbors multiple polymorphisms associated with breast cancer risk in populations of European and Asian ancestry (18–25), which have been associated with subtype-specific effects (18, 19, 22, 26) and have been defined as expression quantitative trait loci (eQTL) for *ESR1* and other genes at 6q25 (24, 27). Among Europeans, five SNPs (rs3757322, rs9397437, rs851984, rs9918437, and rs2747652) define five independent loci associated with breast cancer risk (18). In Asian populations, rs12662670 and rs2046210 were extensively reported as susceptibility risk variants for breast cancer, as well as in populations of European descent (22, 24, 28). Although some of the index risk SNPs reported for other population groups present comparable frequencies in Latin American populations, they did not show strong associations in previous studies of H/L patients (16, 29) likely due to the different patterns of linkage disequilibrium (LD) between these populations. The second GWAS in H/Ls identified two independent loci defined by rs3778609 [odds ratio (OR), 0.76; 95% confidence interval (CI), 0.69–0.83] and rs851980 (OR, 1.28; 95% CI, 1.18–1.35) that are not in LD with rs140068132 (29). A meta-analysis of European and Asian populations confirmed the associations though they presented an attenuated effect (29).

This study aimed to test the extent to which rs140068132 is associated with breast cancer subtype–specific risk beyond ER status in the Peruvian Genetics and Genomics of Breast Cancer Study (PEGEN-BC). In addition, we evaluated previously reported SNPs in the region, their subtype-specific association, and their relationship with rs140068132.

Materials and Methods

Study participants

The PEGEN-BC study

As of March 2022, we have recruited 2,156 participants from the Instituto Nacional de Enfermedades Neoplásicas (INEN) in Lima, Peru, the largest cancer hospital in the country. Details about the study have been previously described (30). Briefly, women were invited to participate if they had a diagnosis of invasive breast cancer in the year 2010 or later, and were between 21 and 79 years of age when diagnosed. Approximately 70% of the patients invited to participate in PEGEN-BC provided their written informed consent and were included in the study. Demographic and clinical data were abstracted from electronic medical records (30). A blood sample was drawn by a certified phlebotomist at the INEN central laboratory. The current analysis includes a subset of 1,755 patients with available genotype data. Patient

and tumor characteristics of individuals in the PEGEN-BC study are similar to those previously described for the overall INEN breast cancer population (31). This study was approved by the INEN and the University of California Davis Institutional Review Boards.

The pregnancy outcomes, maternal and infant cohort study

This study recruited 3,347 women aged 18 or older who attended prenatal care clinics at the Instituto Nacional Materno Perinatal (INMP) in Lima, Peru (32). The pregnancy outcomes, maternal and infant cohort study (PrOMIS) participants were included in the present study as convenience controls to be compared with the PEGEN-BC study participants in association analyses focused on the 6q25 region. The average allele frequencies in this large non–cancer-focused study are expected to provide an estimate of allele frequencies for the general population of Lima, Peru. The study was approved by the institutional review boards of the INMP and the Office of Human Research Administration, Harvard T.H. Chan School of Public Health (Boston, MA). All participants provided written informed consent.

Tumoral tissue samples and subtype classification

Tumoral tissues were obtained from core biopsy or freshly resected invasive breast cancers pre-treatment that were formalin-fixed and paraffin-embedded following standard protocols at INEN. Tumor subtypes were defined using immunohistochemical (IHC) markers by a certified pathologist at INEN. Hormone receptor (HR) status was defined by ER and progesterone receptor (PR) expression. HR positivity was defined at 1% or more cells showing staining for these markers. HER2 positivity was defined as 3+ staining by IHC or by gene amplification detected by FISH following a borderline IHC result. These markers were used to classify tumors as HR$^+$HER2$^-$, HR$^+$HER2$^+$, HR$^-$HER2$^+$ and HR$^-$HER2$^-$ (30). IHC marker information was incomplete for 6% of the patients and therefore analyses by tumor subtype included 1,654 participants.

Genotyping, quality control, and imputation

The study focuses on selected SNPs that have been previously associated with breast cancer risk in the 6q25 locus; however, for estimation of continental ancestry and principal component analyses (PCA), genome-wide genotype data were used, as described below. Genotype data were obtained with the Affymetrix Precision Medicine Research Array for the PEGEN-BC study participants (30) and the Illumina Multi-Ethnic Global Array for PrOMIS study participants (32). Quality control (QC) of the genome-wide genotyped data was performed in PLINK v.1.9 (33) on each dataset separately. First, markers from sex chromosomes were excluded. SNPs with more than 2% missingness, that deviated from Hardy–Weinberg equilibrium at a P value of $<5\times10^{-5}$, or with a minor allele frequency below 5% were removed. Individuals missing more than 5% of genotype information were excluded. Genetically related pairs of individuals (>12.5% relatedness) were identified using KING v2.2.5 (34) and removed (75 individuals from PEGEN and 105 from PrOMIS). After these first QC steps, we identified 55,322 overlapping SNPs (excluding palindromic SNPs) between the PEGEN-BC dataset (1,755 participants), and the PrOMIS dataset (3,337 participants) that were used for ancestry estimation and PCAs. The average genotype call rate for the rs140068132 polymorphism was 0.99 in cases and in controls.

We imputed missing genotypes for the 6q25 region for each dataset separately using the Michigan Imputation Server (35), including individuals from the 1,000 Genomes project phase III as the reference panel. We filtered out low-quality–imputed variants ($r^2 < 0.3$). We

extracted imputed genotypes for 22 previously reported associated SNPs in the 6q25 region (Supplementary Table S1) to add to the analyses testing the association between rs140068132 and tumor subtype in the PEGEN-BC case only ($n = 1,654$ for which tumor subtype data were available) and case/control analyses (1,755 cases, 3,337 controls). These variants were extracted from published genome-wide association and fine mapping studies with and without functional analyses (16, 18–20, 23–25, 29, 36–39). Among the 22 variants in the 6q25 region that have been reported to be associated with breast cancer risk in other populations, we selected those that met the following conditions: (i) They had a P value of <0.15 in a case–control regression analysis adjusted by rs140068132; (ii) the OR in the case–control association analysis showed the same direction as reported in the literature, and (iii) they were in low LD with rs140068132. For SNPs in high LD with each other, we selected the one that presented the strongest evidence of functional activity in the literature and/or best score in the RegulomeDB (40). This process led to the selection of 5 SNPs: rs851984, rs9918437, rs3778609, rs2228480 and rs3798758. Among cases, the imputation quality (r^2) of rs851984, rs9918437, rs3778609, rs2228480 and rs3798758 was 0.95, 0.70, 0.99, 0.60 and 0.92, respectively. Among controls, the r^2 for rs851984, rs9918437, rs3778609, and rs3798758 was 0.99, 0.98, 0.99, 0.82, respectively. Among controls rs2228480 was genotyped, with an average genotype call rate of 0.99.

Identification of structure in the distribution of continental genetic variation

We conducted PCA on unrelated individuals to capture components of genetic structure to be used as covariates in case–control and case-only analyses and to confirm that variation between cases and controls was not driven by differences in the genotyping array. Genotyped data were pruned using PLINK v.1.9 (ref. 33; window size = 50, number of variants = 5, variance inflation factor threshold = 2) and merged with data from the 1,000 Genomes Project: (17) Admixed Americans (Peru, Colombia, Mexico and Puerto Rico), Europeans (Americans with Northern and Western European Ancestry, Italy, Spain, Finland, Scotland), East Asians (China, Japan, Vietnam), and African populations (Nigeria, Kenya, Gambia, Sierra Leone). The PEGEN-BC study includes a large proportion of patients with >98% Indigenous American ancestry, as previously reported (30), and therefore provides a source of non-admixed reference samples for this component. The PCA was performed on the merged dataset with 30,875 variants. T-distributed stochastic neighbor–embedding (Rt-SNE) analysis was performed in R 3.6.0 (41) with the Rtsne Package (42) using the first 30 principal components (PCs) of variation and perplexity = 30. Individual global genetic ancestry was estimated using ADMIXTURE (refs. 43, 44; unsupervised, $k = 4$) on this same dataset.

Data availability

Genotype data for rs140068132, the 22 additional SNPs in the 6q25 region, and global genetic ancestry estimates are available upon request.

Statistical analysis
Descriptive analyses of relevant clinical characteristics, genetic ancestry and rs140068132 genotypes

Differences in characteristics between tumor subtypes were tested by means of one-way ANOVA for normally distributed continuous variables, Kruskal–Wallis tests for non-normally distributed continuous variables, and χ^2 tests for categorical variables.

Case–control association analysis

Single-SNP case–control association analysis was performed using binomial logistic regression models, including the first 10 PCs of genetic variation as covariates (1,755 cases and 3,337 controls). We did not include age at diagnosis/recruitment as a covariate in the case–control analysis due to the limited overlap in age distribution between the PEGEN-BC study cases (mean 52 ± 11.0 years) and the PrOMIS controls (28 ± 6.3). Adding age at diagnosis as a covariate in previous case–control analyses did not have an effect on estimated coefficients or P values for the rs140068132 SNP (16). A conditional analysis was performed by including previously associated SNPs in the model to test the independent effect of rs140068132 from the former set of polymorphisms. We complemented this analysis with haplotype association analyses.

Association between rs140068132 and tumor subtype

Multinomial and binomial multivariate logistic regression analyses were performed in the PEGEN-BC study samples ($N = 1,654$) and controls ($N = 3,337$), including the first 10 PCs of genetic variation as covariates and controls as the reference group. When cases only were evaluated, age at diagnosis was included as covariate and the HR$^+$HER2$^-$ subtype was defined as the reference group. An additional model, including the region of residence as a covariate, was tested to account for residual confounding due to potential regional variation in allele frequency within the Indigenous American component of ancestry and variation in subtype distribution between INEN patients from different regions. A case-only conditional regression model was performed to test the independent association of rs140068132 and tumor subtype by including the previously reported breast cancer–associated SNPs that remained significant in the conditional case–control regression analysis (rs851984G>A, rs9918437G>T and rs3778609C>T). Analyses were adjusted by age at diagnosis and the first 10 PCs of genetic variation.

Haplotype association analysis

The four SNPs that remained associated with breast cancer risk in the conditional case–control model (rs140068132, rs851984, rs2228480, and rs3798758) were included in a haplotype analysis to estimate the magnitude of combined effects. Individual haplotypes defined by rs140068132, rs851984, rs2228480, rs3798758 for case–control association analysis were obtained using the haplo.stats R package (45). This package uses an optimized expectation–maximization algorithm to determine the posterior probabilities of pairs of haplotypes, conditional on unphased genotype data (45). We extracted individual haplotypes filtering out participants whose haplotype designation had a posterior probability <70% and filtered out haplotypes with a frequency <5%. For the haplotype-specific case–control analysis, 8 haplotypes and 1,132 cases and 2,076 controls were tested independently, using genotype 0 (cero copies of the haplotype) as reference. Alleles within each haplotype are represented with numbers within square brackets. To facilitate interpretation, we coded the risk allele as 1 for all SNPs and the non-risk allele as 0. Therefore, rs140068132 and rs2228480 minor alleles are denoted with a 0 due to their protective effect. The order of variants within each haplotype is as follows: rs140068132, rs851984, rs2228480, rs3798758. For case–control estimations, binomial logistic regression models were conducted to test the haplotype-specific ORs for each haplotype modeled as a continuous variable (0, 1, and 2 for absence, presence of one and presence of two copies of a specific haplotype). All analyses were adjusted by the first 10 PCs.

P values <=0.05 were considered statistically significant. All analyses were conducted in R v.3.6.0 (41).

Results

Characteristics of study participants

Clinical characteristics and genetic ancestry proportions for PEGEN-BC study patients, overall (*n* = 1,755) and by subtype (*n* = 1,654) are presented in **Table 1**. The average genetic ancestry proportions for the PEGEN-BC participants were 77% Indigenous American, 18% European, 4% African, and 1% East Asian. The average Indigenous American, European, African, and East Asian ancestry components for women without breast cancer were 81%, 14%, 3%, and 2%, respectively (**Table 1**; **Fig. 1A**). PCAs are consistent with these estimates, showing that a subset of individuals in the Peruvian cohort defines the Indigenous American cluster (**Fig. 1B** and **C**). In addition, visualization through a t-distributed stochastic neighbor–embedding (t-SNE) model shows that cases and controls have similar population genetic structure, including subcontinental clustering of individuals born in the Amazonian region (Supplementary Fig. S1).

Half of the patients (53%) were diagnosed with HR^+HER2^-, 19% with HR^+HER2^+, 13% with HR^-HER2^+, and 15% with HR^-HER2^- tumors. The mean proportion of Indigenous American ancestry was higher among women with the HR^- and $HER2^+$ subtypes ($P = 0.009$) compared with other subtypes. The average age at diagnosis was 53 years (± 11) and differed by tumor subtype (**Table 1**), suggesting an average older age at diagnosis for women with HR^+HER2^- tumors. There were suggestive differences in the distribution of tumor subtype by region, with a higher proportion of patients residing in the Amazonian region diagnosed with HR^-HER2^- disease and a higher proportion of patients from the Mountainous region diagnosed with HR^-HER2^+ disease. However, these differences were not statistically significant. The distribution of clinical characteristics and genetic ancestry by ER, PR, and HER2 status separately are consistent with the results observed by the subtypes defined by these three markers (Supplementary Table S2).

Table 1. Characteristics of the PEGEN-BC study participants by tumor subtype.

	PrOMIS Controls	PEGEN-BC patients					
		Overall	HR^+ $HER2^-$	HR^+ $HER2^+$	HR^- $HER2^+$	HR^- $HER2^-$	*P*
n (%)	3,337 (100)	1,755 (100)	877 (53)	313 (19)	211 (13)	253 (15)	
Age at diagnosis, y (mean, SD)	28.2 (6.3)	53 (11.4)	53 (11)	51 (11)	52 (11)	51 (12)	0.001
		Reproductive variables					
Age at menarche (mean, SD)		12.91 (1.71)	12.90 (1.76)	12.92 (1.69)	13.09 (1.54)	12.92 (1.68)	0.555
Age at first full-term pregnancy (mean, SD)		23.19 (5.66)	23.50 (5.83)	23.09 (5.27)	23.10 (6.13)	22.58 (5.37)	0.194
Number of pregnancies (mean, SD)		2.81 (1.67)	2.76 (1.64)	2.83 (1.73)	2.90 (1.75)	2.86 (1.60)	0.682
Breastfeeding history (yes; %)		1534 (89.1)	678 (96.3)	248 (96.5)	182 (96.3)	224 (96.6)	0.998
		Demographic variables					
Genetic ancestry (mean, SD)[a]							
Indigenous American	0.81 (0.13)	0.77 (0.17)	0.76 (0.17)	0.77 (0.17)	0.80 (0.14)	0.78 (0.16)	0.009
European	0.14 (0.10)	0.18 (0.12)	0.18 (0.13)	0.17 (0.12)	0.16 (0.11)	0.17 (0.12)	0.017
African	0.03 (0.05)	0.04 (0.08)	0.04 (0.08)	0.04 (0.09)	0.04 (0.06)	0.04 (0.07)	0.639
East Asian	0.02 (0.03)	0.01 (0.06)	0.02 (0.08)	0.01 (0.04)	0.01 (0.03)	0.01 (0.03)	0.093
Region of Residence, *n* (%)							
Amazonas		104 (5.9)	50 (5.7)	16 (5.1)	9 (4.3)	19 (7.5)	0.248
Coast		1381 (78.9)	704 (80.5)	243 (77.6)	161 (76.3)	191 (75.5)	
Mountains		266 (15.2)	121 (13.8)	54 (17.3)	41 (19.4)	43 (17.0)	
Tumor characteristics							
Histopathologic variety, *n* (%)							
Ductal		1,546 (89.6)	755 (86.5)	290 (92.9)	197 (94.7)	234 (93.2)	<0.001
Lobular		135 (7.8)	95 (10.9)	17 (5.4)	8 (3.8)	9 (3.6)	
Other		44 (2.6)	23 (2.6)	5 (1.6)	3 (1.4)	8 (3.2)	
Grade, *n* (%)							
Low		69 (4.1)	55 (6.4)	6 (1.9)	0 (0.0)	4 (1.6)	<0.001
Medium		719 (42.2)	503 (58.6)	111 (35.9)	31 (15.1)	36 (14.5)	
High		915 (53.7)	301 (35.0)	192 (62.1)	174 (84.9)	209 (83.9)	
Clinical stage, *n* (%)							
I		114 (7.0)	61 (7.0)	16 (5.2)	7 (3.3)	22 (8.8)	<0.001
II		790 (48.4)	455 (52.5)	127 (41.0)	67 (31.8)	96 (38.2)	
III		729 (44.6)	303 (35.0)	150 (48.4)	125 (59.2)	123 (49.0)	
IV		93 (5.7)	47 (5.4)	17 (5.5)	12 (5.7)	10 (4.0)	
rs140068132 genotype, *n* (%)							
AA	1,900 (56.6)	1,288 (73.4)	618 (70.5)	242 (77.3)	166 (78.7)	195 (77.1)	0.001
AG	1,239 (37.1)	438 (25.0)	249 (28.4)	70 (22.4)	43 (20.4)	50 (19.8)	
GG	203 (6.1)	22 (1.3)	10 (1.1)	1 (0.3)	2 (0.9)	8 (3.2)	

Note: Tumor subtypes are based on estrogen receptor (ER), progesterone receptor (PR) and HER2 status. ER, PR, and HER2 status are missing for 101 individuals (5.6%).
Abbreviation: MAF, minor allele frequency.
[a]*P* values for comparison of genetic ancestry proportions between cases and controls are <0.05 for all ancestry components.
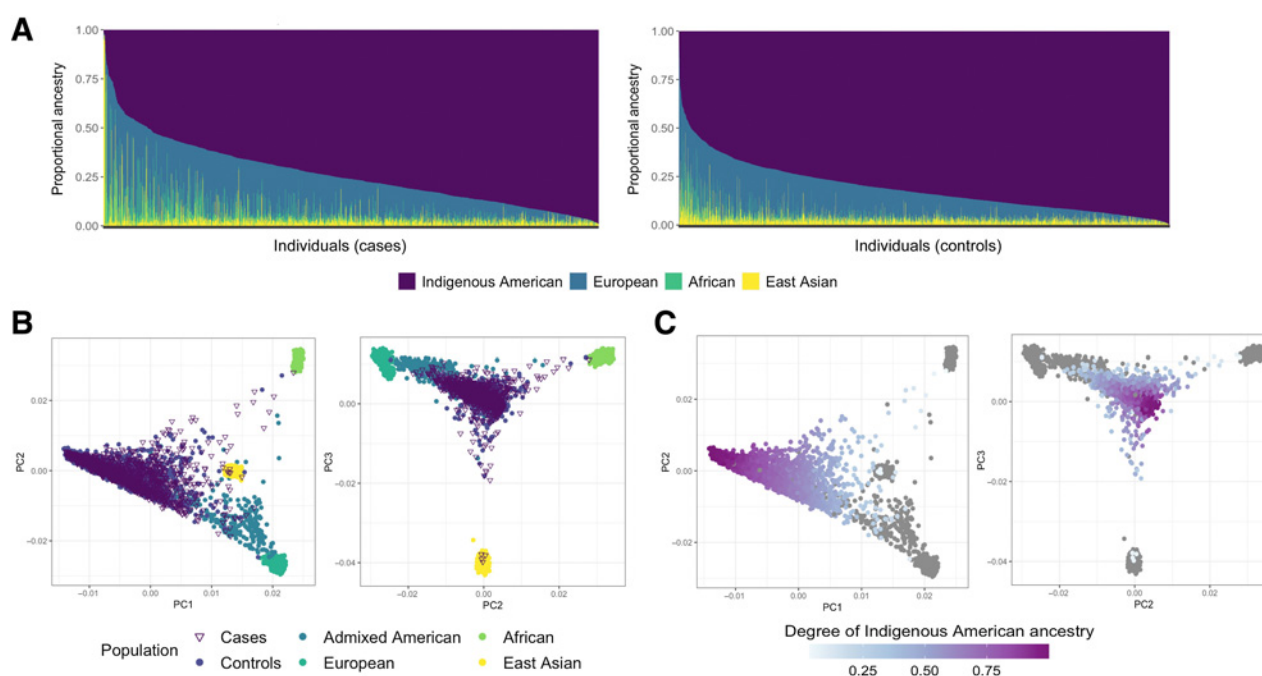
Figure 1.
Population genetic structure of Peruvian breast cancer cases and controls. **A,** ADMIXTURE continental ancestry estimates obtained in unsupervised analysis, assuming $K = 4$ for 1,755 cases and 3,337 controls. **B,** Principal component analysis (PCA), including cases, controls, and 1,000 Genomes Project individuals. The first three principal components are shown. **C,** Same PCA plots showing the degree of Indigenous American ancestry proportion among Peruvian women.

rs140068132 subtype-specific association

The frequency of the rs140068132-G allele among patients with breast cancer was lower than that of controls (14% vs. 25%; respectively, $P = 2.2 \times 10^{-16}$). In a logistic regression model, we replicated the previously reported association of the G allele with lower odds of developing breast cancer (refs. 29, 46; OR, 0.53; 95% CI, 0.47–0.59), and lower odds for ER$^-$ subtypes (OR, 0.48; 95% CI, 0.39–0.57) compared with ER$^+$ tumors (OR, 0.56, 95% CI, 0.49–0.64). A case-only analysis testing the association between rs140068132 and ER status supports the observed heterogeneity by ER status (ER$^+$ vs. ER$^-$ OR, 0.77, 95% CI, 0.60–0.99; **Table 2**).

The G allele was more common in patients with HR$^+$HER2$^-$ tumors (15%), and less so in patients with other subtypes (~11% in HR$^+$HER2$^+$ and HR$^-$HER2$^+$, and 13% in HR$^-$HER2$^-$). The multinomial logistic regression model, including controls as a reference group, confirmed that there is a protective effect for all tumor subtypes (**Table 2**). The G allele was associated with reduced odds of HR$^+$HER2$^+$ and HR$^-$HER2$^+$ tumors in a case-only comparison, showing statically significant differences between each subtype compared with HR$^+$HER2$^-$ (**Table 2**). Results did not change after including the region of residence as a covariate (Supplementary Table S3).

Independence of rs140068132 from other 6q25 previously associated variants

Previous studies in European, African, and Asian populations have reported variants in the 6q25 region associated with breast cancer risk with a differential magnitude of association by subtype (ref. 18; Supplementary Table S1). We tested whether there is a combined effect between rs140068132 and any of these previously reported variants on the observed association with tumor subtypes among the PEGEN-BC study patients. We selected SNPs from the 6q25 locus that are not in LD with rs140068132 for a conditional case–control association analysis (Supplementary Fig. S2). The selection was based on the concordance between the direction and magnitude of the association in the Peruvian case–control analysis and the previously reported ORs, and on their potential functional effect. We selected five candidate polymorphisms: rs851984G>A, rs9918437G>T, rs2228480G>A, rs3798758C>A, and rs3778609C>T. Despite the statistically significant association between each variant and breast cancer risk in the Peruvian study (Supplementary Table S1), only the rs851984, rs2228480, rs3798758, and rs140068132 SNPs remained associated with breast cancer risk in a conditional model (Supplementary Table S4), indicating the independent effect of these latter variants on breast cancer risk. To supplement the results of the conditional model, we tested the combined effect of the significant variants through a haplotype-specific analysis (Supplementary Table S5).

Conditional multinomial logistic regression analyses, including previously associated risk SNPs in the 6q25 region, did not show attenuation of the association between rs140068132 and tumor subtype (**Table 3**). However, rs9918437 showed a statistically significant association with HR$^-$HER2$^+$ subtype (OR, 1.64; 95% CI, 1.10–2.44; $P = 0.015$), independent of rs140068132 (OR, 1.57; 95% CI, 1.05–2.35, $P = 0.027$; **Table 3**).

Discussion

Breast cancer incidence in self-identified H/Ls is lower than in non-Hispanic White women (1), and the discovery of the rs140068132-protective SNP in the first GWAS, including a relatively large number of H/L participants, may be one of the multiple

Table 2. Regression analyses testing the association between rs140068132 and breast cancer risk overall and by tumor subtype.

Model	Comparison groups	OR (95% CI)	P
	Controls as reference (N = 3,337)		
Adjusted by PCs 1–10	All cases (n = 1,755)	0.53 (0.47–0.59)	1.4×10^{-26}
	ER$^+$ (N = 1,154)	0.56 (0.49–0.64)	7.1×10^{-18}
	ER$^-$ (N = 500)	0.48 (0.39–0.57)	2.7×10^{-15}
Adjusted by PCs 1–10	HR$^+$HER2$^-$	0.61 (0.53–0.71)	5.8×10^{-11}
	HR$^+$HER2$^+$	0.42 (0.33–0.55)	6.1×10^{-11}
	HR$^-$HER2$^+$	0.39 (0.28–0.53)	4.5×10^{-9}
	HR$^-$HER2$^-$	0.48 (0.37–0.64)	1.8×10^{-7}
	Cases only (N = 1,654)		
Adjusted by age + PCs 1–10	HR$^+$ vs. HR$^-$ (Ref = HR$^+$)[a]	0.77 (0.60–0.99)	0.042
Adjusted by age + PCs 1–10	HER2$^-$ vs. HER2$^+$ (Ref = HER2$^-$)	0.70 (0.55–0.89)	0.004
	By Subtype. HR$^+$HER2$^-$ as reference		
Adjusted by age + PCs 1–10	HR$^+$HER2$^+$	0.68 (0.51–0.92)	0.011
	HR$^-$HER2$^+$	0.63 (0.44–0.90)	0.012
	HR$^-$HER2$^-$	0.77 (0.56–1.05)	0.096

Note: Odds ratios reported per-allele. Overall breast cancer risk analyses were adjusted by the first 10 principal components (PCs) of genetic variation. Case-only subtype analyses were adjusted by age at diagnosis and the first 10 PCs of genetic variation. The rs140068132-A allele was used as reference.
[a]2% (N = 36) of samples are ER$^-$PR$^+$.

factors (e.g., alcohol intake and reproductive history) contributing to this trend (16). In previous studies, we demonstrated that rs140068132 showed heterogeneous association by ER status (16, 29), with lower odds for ER$^-$ compared with ER$^+$ diseases among G-allele carriers (16, 29). In the present study, we replicated this observation in a highly Indigenous American population and further tested the subtype-specific association of rs140068132 showing that this SNP is more protective for HER2$^+$ diseases compared with HR$^+$/HER2$^-$ subtype in patients with breast cancer from Peru. We evaluated the association of this SNP together with other known risk variants at the locus showing that the association between rs140068132 and breast cancer risk is independent of other risk-associated SNPs in the region. The direction of the associations of these additional SNPs is consistent with those previously reported in other studies (Supplementary Table S1). We also found that one of the previously associated SNPs, rs9918437, shows a subtype-specific association in the Peruvian samples, with the rs9918437-T variant conferring increased odds of HR$^-$HER2$^+$ disease. A previous study reported a higher magnitude of effect for ER$^-$ compared with ER$^+$ tumors (OR, 1.18, 95% CI, 1.11–1.27, compared with OR, 1.08, 95% CI, 1.04–1.13, respec-

tively). However, no significant association was reported for the HR$^-$HER2$^+$ subtype (18). These results confirm that multiple SNPs within the 6q25 region contribute independently to subtype-specific breast cancer risk.

The protective effect of rs140068132 for the development of HER2$^+$ disease suggests that the rs140068132 SNP might not only be involved in the regulation of ESR1 expression but might also affect ERBB2 expression (the gene that codes for HER2). Evidence showed that ER inhibits HER2 expression at the transcriptional level (47), and therefore, the fine-tuning regulation of ESR1 expression by rs140068132 or other variants within this locus might have a direct effect on HER2 levels. Other indirect mechanisms involve the modification of ER-related pathways (48). It is also possible that the inverse association with HER2 expression is due to molecular mechanisms that increase the chances of developing low proliferation cancer cell lineages that do not depend on HER2-associated pathways.

Previous studies have reported SNPs showing subtype-specific associations in breast cancer (18, 19, 22, 26, 49). For some of them, the functional interpretation and the link with the molecular events that explain the particular subtypes has been straightforward because

Table 3. Conditional multinomial logistic regression analyses testing the association between rs140068132 and breast tumor subtypes (case-only, n = 1,654; HR$^+$HER2$^-$ as reference), including previously reported breast cancer risk associated variants: rs851984, rs9918437, and rs3778609.

	HR$^-$HER2$^+$		HR$^+$HER2$^+$		HR$^-$HER2$^-$	
	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P
Single SNP model						
rs140068132	0.63 (0.44–0.90)	0.0114	0.68 (0.51–0.92)	0.0124	0.77 (0.56–1.05)	0.0965
rs851984	1.23 (0.97–1.57)	0.0838	1.16 (0.94–1.42)	0.1568	1.10 (0.88–1.38)	0.3843
rs9918437	1.64 (1.10–2.44)	0.0151	1.23 (0.85–1.79)	0.2678	1.20 (0.80–1.81)	0.3756
rs3778609	0.84 (0.65–1.08)	0.1659	1.04 (0.84–1.27)	0.7384	0.95 (0.76–1.19)	0.6568
rs2228480	0.85 (0.65–1.11)	0.2441	1.04 (0.84–1.30)	0.7089	1.15 (0.91–1.46)	0.2322
rs3798758	1.20 (0.93–1.56)	0.1672	1.01 (0.80–1.27)	0.9411	1.01 (0.79–1.31)	0.9095
Multi SNP models						
rs140068132 (+5 other SNP)	0.66 (0.45–0.96)	0.0318	0.66 (0.48–0.90)	0.0084	0.76 (0.54–1.06)	0.1019
rs9918437 (+rs140068132)	1.57 (1.05–2.35)	0.0268	1.20 (0.83–1.74)	0.3409	1.18 (0.78–1.77)	0.4330

Note: All analyses were adjusted by age at diagnosis and the first 10 principal components of genetic variation.

they are located within known breast cancer–associated genes of high or moderate penetrance (50). However, most SNPs identified in GWAS are located in non-coding regions (51), such as the case of rs140068132, and are associated with relatively modest ORs, making the functional interpretation challenging. Future *in vitro* studies will allow testing the effect of rs140068132 on *ESR1* gene expression in different tissue contexts, representing the heterogeneity of breast tumor subtypes.

Our study has some limitations. This study includes both case–case (subtype-specific) analyses as well as comparisons between cases and controls where the controls were part of the PrOMIS Study. These women are demographically similar to the women in the PEGEN-BC study (both studies recruited participants in public hospitals in Lima, Peru). However, given the nature of the two studies, the age distributions are different. We believe that this difference is unlikely to influence the results of our analyses and, if anything, the bias would be toward mitigation of allele frequency differences between breast cancer cases and controls assuming that some of the PrOMIS participants might develop breast cancer in the future. In addition, average ancestry differences between cases and controls are in the expected direction and of the same magnitude as what has been reported in previous studies of breast cancer in H/Ls from the US and Mexico (3). This provides additional reassurance about the adequacy of combining the PEGEN-BC and PrOMIS studies.

Another limitation was the relatively small number of overlapping SNPs between the arrays used in the PEGEN-BC and PrOMIS studies. Imputation quality was affected by the starting set and therefore full coverage of the 6q25 region was not possible to achieve due to the low imputation quality of some of the SNPs, limiting our analyses to previously reported variants. We plan to resolve this issue in the future to provide a further in-depth characterization of variation within the 6q25 region in Peruvian women.

In summary, the rs140068132 SNP is associated with breast cancer risk in a subtype-specific manner, independently of nearby variants previously described in other populations. The molecular mechanisms leading to the association between this protective variant and the least aggressive HR^+HER2^- breast cancer subtype needs further investigation. Ongoing eQTL analyses focused on rs140068132 in the context of the bidirectional molecular crosstalk between ER and HER2 pathways will contribute to understanding the effect of this variant on the etiology of breast cancer.

Authors' Disclosures

A. Hechmer reports grants from UCSF during the conduct of the study. E. Ziv reports grants from California Institute to Advance Precision Medicine and NIH/NCI during the conduct of the study. No disclosures were reported by the other authors.

Authors' Contributions

V.A. Zavala: Data curation, formal analysis, investigation, visualization, methodology, writing–original draft. **S. Casavilca-Zambrano:** Resources, writing–review and editing. **J. Navarro Vásquez:** Investigation, writing–review and editing. **C.A. Castañeda**: Investigation, writing–review and editing. **G. Valencia:** Investigation, writing–review and editing. **Z. Morante:** Investigation, writing–review and editing. **M. Calderón:** Investigation, writing–review and editing. **J.E. Abugattas:** Resources, investigation, writing–review and editing. **H. Gomez:** Investigation, writing–review and editing. **H.A. Fuentes:** Investigation, writing–review and editing. **R. Liendo-Picoaga:** Investigation, writing–review and editing. **J.M. Cotrina:** Investigation, writing–review and editing. **C. Monge:** Investigation, writing–review and editing. **S.P. Neciosup:** Investigation, writing–review and editing. **S. Huntsman:** Software, validation, writing–review and editing. **D. Hu:** Software, investigation, writing–review and editing. **S.E. Sanchez:** Resources, investigation, writing–review and editing. **M.A. Williams:** Resources, investigation, writing–review and editing. **A. Nunez-Marrero:** Investigation, writing–review and editing. **L. Godoy:** Investigation, writing–review and editing. **A. Hechmer:** Software, investigation, writing–review and editing. **A.B. Olshen:** Software, investigation, writing–review and editing. **J. Dutil:** Resources, validation, writing–review and editing. **E. Ziv:** Investigation, writing–review and editing. **J. Zabaleta:** Investigation, writing–review and editing. **B. Gelaye:** Resources, investigation, writing–review and editing. **J. Vasquez:** Investigation, writing–review and editing. **M. Galvez-Nino:** Investigation, writing–review and editing. **D. Enriquez-Vera:** Investigation, writing–review and editing. **T. Vidaurre:** Resources, writing–review and editing. **L. Fejerman:** Conceptualization, resources, formal analysis, supervision, funding acquisition, investigation, methodology, writing–original draft, project administration, writing–review and editing.

Note

Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

References

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2021. CA Cancer J Clin 2021;71:7–33.
2. Acheampong T, Kehm RD, Terry MB, Argov EL, Tehranifar P. Incidence trends of breast cancer molecular subtypes by age and race/ethnicity in the US From 2010 to 2016. JAMA Netw Open 2020;3:e2013226.
3. Fejerman L, John EM, Huntsman S, Beckman K, Choudhry S, Perez-Stable E, et al. Genetic ancestry and risk of breast cancer among U.S. Latinas. Cancer Res 2008;68:9723–8.
4. Fejerman L, Chen GK, Eng C, Huntsman S, Hu D, Williams A, et al. Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas. Hum Mol Genet 2012;21:1907–17.
5. Ziv E, John EM, Choudhry S, Kho J, Lorizio W, Perez-Stable EJ, et al. Genetic ancestry and risk factors for breast cancer among Latinas in the San Francisco Bay Area. Cancer Epidemiol Biomarkers Prev 2006;15:1878–85.
6. Zavala VA, Bracci PM, Carethers JM, Carvajal-Carmona L, Coggins NB, Cruz-Correa MR, et al. Cancer health disparities in racial/ethnic minorities in the United States. Br J Cancer 2021;124:315–32.
7. Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, et al. Genomic insights into the ancestry and demographic history of South America. PLOS Genet 2015;11:e1005602.
8. Chacón-Duque J-C, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuña-Alonzo V, Barquera R, et al. Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. Nat Commun 2018;9:1–13.

9. Posth C, Nakatsuka N, Lazaridis I, Skoglund P, Mallick S, Lamnidis TC, et al. Reconstructing the deep population history of central and South America. Cell 2018;175:1185–97.

10. Adhikari K, Chacón-Duque JC, Mendoza-Revilla J, Fuentes-Guajardo M, Ruiz-Linares A. The genetic diversity of the Americas. Annu Rev Genomics Hum Genet 2017;18:277–96.

11. Hidalgo PC, Bengochea M, Abilleira D, Cabrera A, Alvarez I. Genetic admixture estimate in the uruguayan population based on the loci LDLR, GYPA, HBGG, GC, and D7S8. Int J Hum Genet 2005;5:217–22.

12. Norris ET, Wang L, Conley AB, Rishishwar L, Mariño-Ramírez L, Valderrama-Aguirre A, et al. Genetic ancestry, admixture and health determinants in Latin America. BMC Genomics 2018;19:861.

13. Heinz T, Álvarez-Iglesias V, Pardo-Seco J, Taboada-Echalar P, Gómez-Carballa A, Torres-Balanza A, et al. Ancestry analysis reveals a predominant Native American component with moderate European admixture in Bolivians. Forensic Sci Int Genet 2013;7:537–42.

14. Murray T, Beaty TH, Mathias RA, Rafaels N, Grant AV, Faruque MU, et al. African and non-African admixture components in African Americans and an African Caribbean population. Genet Epidemiol 2010;34:561–8.

15. Moreno-Estrada A, Gravel S, Zakharia F, Mccauley JL, Byrnes JK, Gignoux CR, et al. Reconstructing the population genetic history of the Caribbean. PLoS Genet 2013;9:e1003925.

16. Fejerman L, Ahmadiyeh N, Hu D, Huntsman S, Beckman KB, Caswell JL, et al. Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25. Nat Commun 2014;5:5260.

17. 1000 Genomes Project Consortium, Campbell CL, Scheller, C, Horn H, Kidd JM, Doddapaneni H, et al. A global reference for human genetic variation. Nature 2015;526:68–74.

18. Dunning AM, Michailidou K, Kuchenbaecker KB, Thompson D, French JD, Beesley J, et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1, and CCDC170. Nat Genet 2016;48:374–86.

19. Cai Q, Wen W, Qu S, Li G, Egan KM, Chen K, et al. Replication and functional genomic analyses of the breast cancer susceptibility locus at 6q25.1 generalize its importance in women of Chinese, Japanese, and European ancestry. Cancer Res 2011;71:1344–55.

20. Fletcher O, Johnson N, Orr N, Hosking FJ, Gibson LJ, Walker K, et al. Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. J Natl Cancer Inst 2011;103:425–35.

21. Kim H-C, Lee J-Y, Sung H, Choi J-Y, Park SK, Lee K-M, et al. A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. Breast Cancer Res 2012;14:R56.

22. Hein R, Maranian M, Hopper JL, Kapuscinski MK, Southey MC, Park DJ, et al. Comparison of 6q25 breast cancer hits from Asian and European genome wide association studies in the Breast Cancer Association consortium (BCAC). PLoS ONE 2012;7:e42380.

23. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 2010;42:504–7.

24. Stacey SN, Sulem P, Zanon C, Gudjonsson SA, Thorleifsson G, Helgason A, et al. Ancestry-shift refinement mapping of the C6orf97-ESR1 breast cancer susceptibility locus. PLoS Genet 2010;6:1–12.

25. Zheng W, Long J, Gao Y-T, Li C, Zheng Y, Xiang Y-B, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 2009;41:324–8.

26. Zhang H, Ahearn TU, Lecarpentier J, Barnes D, Beesley J, Qi G, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nat Genet 2020;52:572–81.

27. Bailey SD, Desai K, Kron KJ, Mazrooei P, Sinnott-Armstrong NA, Treloar AE, et al. Noncoding somatic and inherited single-nucleotide variants converge to promote ESR1 expression in breast cancer. Nat Genet 2016; 48:1260–6.

28. Jin TF, Zhang WT, Zhou ZF. The 6q25.1 rs2046210 polymorphism is associated with an elevated susceptibility to breast cancer: a meta-analysis of 261,703 subjects. Mol. Genet. Genomic Med. 2019;7:e553.

29. Hoffman J, Fejerman L, Hu D, Huntsman S, Li M, John EM, et al. Identification of novel common breast cancer risk variants at the 6q25 locus among Latinas. Breast Cancer Res 2019;21:3.

30. Marker KM, Zavala VA, Vidaurre T, Lott PC, Vásquez JN, Casavilca-Zambrano S, et al. Human epidermal growth factor receptor 2–positive breast cancer is associated with indigenous American Ancestry in Latin American women. Cancer Res 2020;80:1893–901.

31. Vallejos CS, Gómez HL, Cruz WR, Pinto JA, Dyer RR, Velarde R, et al. Breast cancer classification according to immunohistochemistry markers: subtypes and association with clinicopathologic variables in a peruvian hospital database. Clin Breast Cancer 2010;10:294–300.

32. Gelaye B, Zhong Q-Y, Basu A, Levey EJ, Rondon MB, Sanchez S, et al. Trauma and traumatic stress in a sample of pregnant women. Psychiatry Res 2017;257:506–13.

33. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.

34. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics 2010; 26:2867–73.

35. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet 2016;48:1284–7.

36. Wang Y, He Y, Qin Z, Jiang Y, Jin G, Ma H, et al. Evaluation of functional genetic variants at 6q25.1 and risk of breast cancer in a Chinese population. Breast Cancer Res 2014;16:422.

37. Sun Y, Ye C, Guo X, Wen W, Long J, Gao Y-T, et al. Evaluation of potential regulatory function of breast cancer risk locus at 6q25.1. Carcinogenesis 2015;37: 163–8.

38. Long J, Cai Q, Sung H, Shi J, Zhang B, Choi J-Y, et al. Genome-wide association study in East Asians identifies novel susceptibility loci for breast cancer. PLoS Genet 2012;8:e1002532.

39. Long J, Zhang B, Signorello LB, Cai Q, Deming-Halverson S, Shrubsole MJ, et al. Evaluating genome-wide association study-identified breast cancer risk variants in African-American women. PLoS ONE 2013;8:e58350.

40. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 2012;22:1790–7.

41. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/.

42. Krijthe JH. Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. R package version 0.16. Available from: https://github.com/jkrijthe/Rtsne.

43. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19:1655–64.

44. Zhou H, Alexander D, Lange K. A quasi-Newton acceleration for high-dimensional optimization algorithms. Stat. Comput. 2011;21: 261–73.

45. Sinnwell JP, Schaid DJ. _haplo.stats: statistical analysis of haplotypes with traits and covariates when linkage phase is ambiguous. R package version 1.8.9. Availabe from: https://CRAN.R-project.org/package=haplo.stats.

46. Fejerman L, Hu D, Huntsman S, John EM, Stern MC, Haiman CA, et al. Genetic ancestry and risk of mortality among U.S. latinas with breast cancer. Cancer Res 2013;73:7243–53.

47. Newman SP, Bates NP, Vernimmen D, Parker MG, Hurst HC. Cofactor competition between the ligand-bound oestrogen receptor and an intron 1 enhancer leads to oestrogen repression of ERBB2 expression in breast cancer. Oncogene 2000;19:490–7.

48. Arpino G, Wiechmann L, Osborne CK, Schiff R. Crosstalk between the estrogen receptor and the HER tyrosine kinase receptor family: molecular mechanism and clinical implications for endocrine therapy resistance. Endocr Rev 2008;29:217–33.

49. Lilyquist J, Ruddy KJ, Vachon CM, Couch FJ. Common genetic variation and breast cancer risk—past, present, and future. Cancer Epidemiol Biomarkers Prev 2018;27:380–94.

50. Wang S, Zhang K, Tang L, Yang Y, Wang H, Zhou Z, et al. Association between single-nucleotide polymorphisms in breast cancer susceptibility genes and clinicopathological characteristics. Clin. Epidemiol. 2021;13: 103–12.

51. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature 2017;550:204–13.