Identification of Novel Common Breast Cancer Risk Variants in Latinas at the 6q25 Locus

Joshua Hoffman[1]
Laura Fejerman[2]
Donglei Hu[3]
Scott Huntsman[4]
Min Li[5]
Esther John[6]
Gabriela Torres-Mejia[7]
Larry Kushi[8]
Yuan Chun Ding[9]
Jeffrey Weitzel[10]
Susan L. Neuhausen[11]
Paul Lott[12]
COLUMBUS Consortium
Magdalena Echeverry[13]
Luis Carvajal-Carmona[14]
Esteban Burchard[15]
Celeste Eng[16]
Wei Zheng[17]
Jirong Long[18]
Olufunmilayo Olopade[19]
Dezheng Huo[20]
Christopher Haiman[21]
Elad Ziv[22],[#]

1. jdh0210@gmail.com Department of Epidemiology and Biostatistics, Institute of Human Genetics, University of California, San Francisco, San Francisco, CA
2. laura.fejerman@ucsf.edu Division of General Internal Medicine, Department of Medicine, Institute of Human Genetics, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA
3. donglei.hu@ucsf.edu Division of General Internal Medicine, Department of Medicine, University of California, San Francisco, San Francisco, CA
4. scott.huntsman@ucsf.edu Division of General Internal Medicine, Department of Medicine, University of California, San Francisco, San Francisco, CA
5. min.li2@ucsf.edu Division of General Internal Medicine, Department of Medicine, University of California, San Francisco, San Francisco, CA
6. esther.john@cpic.org Cancer Prevention Institute of California, Fremont, CA, Stanford Cancer Institute, Stanford University, Palo Alto, CA
7. gtorres@insp.mx National Institute of Public Health, Cuernavaca, Mexico
8. larry.kushi@kp.org Division of Research, Kaiser Permanente, Northern California
9. ycding@coh.org Beckman Research Institute of City of Hope, Department of Population Sciences, City of Hope, Duarte, CA

10. JWeitzel@coh.org City of Hope National Medical Center, Clinical Cancer Genetics, Duarte CA

11. sneuhausen@coh.org Beckman Research Institute of City of Hope, Department of Population Sciences, City of Hope, Duarte, CA

12. pclott@ucdavis.org Department of Biochemistry and Molecular Medicine, University of California, Davis, Davis, CA

13. echeverrydepolanco@hotmail.com University of Tolima, Ibague, Colombia

14. lgcarvajal@ucdavis.edu Department of Biochemistry and Molecular Medicine, University of California, Davis, Davis, CA

15. esteban.burchard@ucsf.edu Department of Biophamaceutical Sciences, Lung Biology Center, Department of Medicine, Institute for Human Genetics, University of California, San Francisco, San Francisco, CA

16. celeste.eng@ucsf.edu Department of Biophamaceutical Sciences, Lung Biology Center, Department of Medicine, Institute for Human Genetics, University of California, San Francisco, San Francisco, CA

17. jirong.long@Vanderbilt.Edu Department of Epidemiology, Vanderbilt University Medical Center, TN

18. wei.zheng@Vanderbilt.Edu Department of Epidemiology, Vanderbilt University Medical Center, TN

19. folopade@medicine.bsd.uchicago.edu Department of Public Health Sciences, University of Chicago School of Medicine, Chicago, IL

20. dhuo@health.bsd.uchicago.edu Department of Medicine, Section of Oncology, University of Chicago School of Medicine, Chicago, IL

21. Christopher.Haiman@med.usc.edu Department of Preventive Medicine, Norris Comprehensive Cancer Center, Keck School of Medicine, University of Southern California, Los Angeles, CA

22. Elad.Ziv@ucsf.edu Division of General Internal Medicine, Department of Medicine, Institute of Human Genetics, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA

Corresponding author:
Elad Ziv
Division of General Internal Medicine,
Department of Medicine, UCSF
Box 0320
San Francisco, CA 94143
Elad.ziv@ucsf.edu
Fax: 415-353-7981

Abstract:

Background: Breast cancer is a partially heritable trait and over 180 common genetic variants have been associated with breast cancer in genome wide association studies (GWAS). We have previously performed breast cancer GWAS in Latinas and identified a strongly protective single nucleotide polymorphism (SNP) at 6q25 with the protective minor allele originating from Indigenous American ancestry. Here we report on additional GWAS and replication in Latinas.

Methods: We performed GWAS in 2385 cases and 7342 controls who were either U.S. Latinas or Mexican women. We replicated 2412 cases and 1620 controls of U.S Latina, Mexican, and Colombian women. In addition, we replicated the top novel variants in study of African American and African women and in one study of Chinese women. In each dataset we used logistic regression models to test the association between SNPs and breast cancer risk and corrected for genetic ancestry using either principal components or genetic ancestry inferred from ancestry informative markers using a model based approach.

Results: We identified 3 SNPs (p=1.9×10^{-8} - 2.8×10^{-8}) at 6q25 locus not in linkage disequilibrium (LD) with variants previously reported at this locus. These SNPs were in high LD with each other, with the top SNP, rs3778609, associated with breast cancer with an odds ratio (OR) and 95% confidence interval (95% CI) of 0.75 (0.68-0.83). In a replication in women of Latin American origin, we also observed a consistent effect (OR: 0.88; 95% CI: 0.78-0.99; p=0.037). Since the minor allele was common in East Asians and African American but not European ancestry populations, we replicated in a meta-analysis of those populations and also observed a consistent effect (OR 0.94; 95% CI: 0.91 – 0.97; p=0.013).

Conclusion: The effect size of this variant is relatively large compared to other common variants associated with breast cancer and adds to evidence about the importance of the 6q25 locus for breast cancer susceptibility. Our finding also highlights the utility of performing additional searches for genetic variants for breast cancer in non-European populations.

Introduction:

Breast cancer is a partially heritable disease. Mutations in several high penetrance genes including *BRCA1*[1, 2], *BRCA2*[3] and others[4] are associated with high risk of breast cancer among carriers and explain a fraction of the heritability. Genome-wide association studies have identified over 180 common single nucleotide polymorphisms (SNPs) associated with risk of breast cancer [5-20]. The majority of these SNPs were identified in European ancestry and East Asian ancestry populations, although some unique SNPs have been identified in African American populations[21] and in Latina populations[22, 23].

Several GWAS studies have identified SNPs at 6q25 that are associated with breast cancer risk[13, 18, 20, 23-27] and mammographic density[23, 27-30]. The initial report identified a SNP in the intergenic region between *ESR1* and *CDCC170* in an East Asian population[24]. The locus was then confirmed in other populations and several additional variants were identified[11, 18, 25, 26, 31]. More recently, a fine-mapping and functional approach at this locus identified five distinct common variants associated with risk of different subtypes of breast cancer[27].

Hispanic/Latino populations are the second largest ethnic group in the U.S.[32] and yet have been understudied in genome wide association studies[33]. Latinos are a population of mixed ancestry with European, Indigenous American and African ancestral contributions[34-37]. Since there are no large studies of breast cancer in Indigenous American populations, studies in Latinos may identify novel variants associated breast cancer unique to or substantially more common in this population. We have previously used an admixture mapping approach to search for breast cancer susceptibility loci in Latinas and identified a large region at 6q25 where Indigenous American ancestry was associated with decreased risk of breast cancer[22]. Subsequently, we identified a SNP (rs140068132) that was common (minor allele frequency ~0.1) only in Latinas with Indigenous American ancestry and was associated with substantially lower risk of breast cancer, particularly estrogen receptor (ER) negative breast cancer and with

lower mammographic density[23]. However, the variant we identified did not completely explain the risk associated with locus specific ancestry at 6q25 in Latinas, suggesting that other variants may account for this risk. We set out to fine-map and identify additional variants at 6q25 associated with breast cancer risk among Latinas.

Methods:

Populations:

San Francisco Bay Area Breast Cancer Study (SFBCS): The SFBCS is a population-based multiethnic case–control study of breast cancer. Cases aged 35–79 years diagnosed with invasive breast cancer from 1995 to 2002 were identified through the Greater Bay Area Cancer Registry. Controls were identified by random-digit dialing and matched on 5-year age groups. Blood collection was initiated in 1999. For this study, we focused only on cases and matched controls who self-identified as Latina or Hispanic and included 351 cases and 579 controls. Samples from this study were used as part of the initial discovery set.

GALA1: GALA1 is a family-based study (including children with asthma and their parents) of pediatric asthma in Latino Americans. We included 112 females of self-reported Mexican origin from the GALA1 study to our set of population controls. The individuals are between 11 and 42 years of age (85% are older than 20 years). Samples from this study were used as part of the initial discovery set.

Breast Cancer Family Registry (BCFR): The BCFR is an international, National Cancer Institute (NCI)-funded family study that has recruited and followed over 13,000 breast cancer families and breast cancer cases with strong likelihood of genetic contribution to disease45. The present study includes samples from the population-based Northern California site of the BCFR. Cases aged 18–64 years diagnosed from 1995 to 2007 were ascertained through the Greater Bay Area Cancer Registry. Cases with indicators of

increased genetic susceptibility (diagnosis at the age of <35 years, bilateral breast cancer with the first diagnosis at the age of <50 years, a personal history of ovarian or childhood cancer and a family history of breast or ovarian cancer in first-degree relatives) were oversampled. Cases not meeting these criteria were randomly sampled46. Population controls were identified through random-digit dialing and frequency-matched on 5-year age groups to cases diagnosed from 1995 to 1998. We included 641 cases and 61 controls who self-identified as Latina or Hispanic from this study. Samples from this study were used as part of the initial discovery set.

Multiethnic Cohort (MEC): The MEC is a large prospective cohort study in California (mainly Los Angeles County) and Hawaii. The breast cancer study is a nested case–control study including women with invasive breast cancer diagnosed at the age of >45 years and controls matched on age (within 5 years) and self-identified ethnicity47. For the current study, we used data and genetic data from 546 Latina women with breast cancer and 558 matched Latina controls. We also included an additional 1,941 controls who self-identified as Hispanic/Latino from this study (935 of these controls are men) selected as part of a GWAS of type 2 diabetes[38]. Samples from this study were used as part of the initial discovery set.

Research Project on Genes Environment and Health (RPGEH): The RPGEH is a large cohort study of over 100,000 men and women of all racial/ethnic groups who are members of the Kaiser Permanente Health Plan (additional recruitment criteria?). This analysis focuses only on women who are of self-reported Latina/Hispanic ethnicity (N=3801). We included both incident and prevalent cases (total N=225) in our analyses. We identified 44 women who were also included the SFBCS. The genetic data from these participants were included as part of the RPGEH since we considered the Affymetrix Lat array as a more comprehensive array. Samples from this study were used as part of the initial discovery set.

Cancer de Mama (CAMA) Study: This study is a population-based case–control study of breast cancer conducted in Mexico City, Monterrey and Veracruz. Cases aged 35–69

years diagnosed between 2005 and 2007 were recruited from 11 hospitals (three to five in each region). Controls were recruited based on membership in the same health plan as the cases and are frequency-matched on 5-year age groups. For the current study, we used data and DNA samples from 1008 women with breast cancer and 1,063 controls. Of these 698 cases and 599 controls were genotyped with Oncoarray and included in the discovery. An additional, 310 cases and 464 controls were included as part of the replication dataset. A subset of the samples from this study were used as part of the initial discovery set and another subset were used as part of the replication.

Colombian Study of Environmental and Heritable Causes of Breast Cancer (COLUMBUS): COLUMBUS is a population-based case–control study of breast cancer conducted in four cities: Bogota, Ibague and Neiva, from the Central Colombian Andes region, and Pasto, from the Colombian South. Incident cases with invasive breast cancer aged 18–75 years have been recruited in two population registries and two large cancer hospitals. Recruitment started in 2011. Cancer-free controls were recruited through the same institutions and were matched on education, socioeconomic status and local origin using a genealogical interview. In the current study, we used data from 954 cases and 769 controls for the replication study.

Hereditary Cancer Registry of City of Hope (HCRCOH): (Southern California; PI Jeffrey Weitzel). Latina breast cancer cases are part of the HCRCOH through the Clinical Cancer Genetics Community Research Network (CCGCRN). The CCGCRN includes cancer center and community-based clinics that provide genetic counseling to individuals with a personal or family history of cancer [39]. All patients are invited to participate in the HCRCOH at the time of consultation (>90% participation). Starting in May 1998 and continuing to present, female breast cancer cases with self-reported Latino origin were seen for GCRA, enrolled in the Registry and underwent *BRCA1/2* testing after informed consent. In the current study we genotyped 1148 cases. The 347 unaffected female Latina controls were from Southern

California and were invited to participate at community health fairs, flyers, and at City of Hope. These samples were used as part of the replication study.

African American Breast Cancer GWAS (AABC): The GWAS includes African American participants from 9 epidemiological studies of breast cancer, comprising a total of 3,153 cases and 2,831 controls (cases/controls: The Multiethnic Cohort Study (MEC), 734/1,003; The Los Angeles component of The Women's Contraceptive and Reproductive Experiences (CARE) Study, 380/224; The Women's Circle of Health Study (WCHS), 272/240; The San Francisco Bay Area Breast Cancer Study (SFBCS), 172/231; The Northern California Breast Cancer Family Registry (NC-BCFR), 440/53;The Carolina Breast Cancer Study (CBCS), 656/608; The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) Cohort, 64/133; The Nashville Breast Health Study (NBHS), 310/186; and, The Wake Forest University Breast Cancer Study (WFBC), 125/153). Additional details can be found in [21, 40]. These samples were used as part of the replication study.

The ROOT consortium included six studies and a total of 1,657 cases and 2,029 controls of African ancestry: The Nigerian Breast Cancer Study (NBCS), 711/624; The Barbados National Cancer Study (BNCS), 92/229; The Racial Variability in Genotypic Determinants of Breast Cancer Risk Study (RVGBC), 145/257; The Baltimore Breast Cancer Study (BBCS), 95/102; The Chicago Cancer Prone Study (CCPS), 394/387; and The Southern Community Cohort (SCCS), 220/430. Additional details can be found in [21]. These samples were used as part of the replication study.

Shanghai Breast Cancer Study: The SBCS is a population-based, case-control study conducted in urban Shanghai. Subject recruitment in the initial phase of the SBCS (SBCS-I) was conducted between August 1996 and March 1998. The second phase (SBCS-II) of recruitment occurred between April 2002 and February 2005. Breast cancer cases were identified through the population-based Shanghai Cancer Registry and supplemented by a rapid case-ascertainment system. Controls were randomly selected using the Shanghai Resident

Registry. Approximately 3500 cases and 3500 controls were recruited in the study. A subset of these including with GWAS data including 2731 cases and 2135 controls were used as part o the replication study.

Genotyping:

Genome wide association: The SFBCS, NC-BCFR, and GALA samples were all genotyped with Affymetrix 6.0 arrays at UCSF. The MEC samples were genotyped with Illumina 660 array at USC (546 Latina women with breast cancer and 558 matched Latina controls) and an additional 1941 controls were typed on an Illumina 2.5M array at the Broad Institute (Cambridge, MA). The RPGEH samples were typed on an Affymetrix LAT array at UCSF. The CAMA samples were typed on an Ilumina Oncoarray at the Quebec Genome Center. The COLUMBUS samples were typed on an Affymetrix Biobank Array. Genotyping in the AABC consortium was conducted using the IlluminaHuman1M-Duo BeadChip. Genotyping in the ROOT consortium was conducted using the Illumina HumanOmni2.5-8v1 array at Johns Hopkins University Center for Inherited Disease Research. The Shanghai Breast Cancer Study samples were typed on on an Affymetrix Genome-Wide Human SNP Array 6.0. After quality control exclusions, the final data set included 2731 cases and 2135 controls for 668 499 markers.

Replication genotyping: The CAMA samples which were not included in the GWAS and the CCGRN samples, were genotyped using Taqman probes for rs3778609 . The CAMA samples included 106 ancestry informative markers from genotyped on a Sequenome platform as previously described[41]. CCGRN samples included 100 ancestry informative markers that were included as part of a sequencing project. The sequence data were aligned to Hg37 using Burrows-Wheeler Alignment and genotype calls were made using Haplotypecaller which is part of the GATK platform.

Analysis:

Genotyping Quality Control and Imputation: Samples with >5% missing genotypes were removed from each dataset. We dropped variants with >5% missing data from each dataset. Since excess homozygosity is more common in populations with substructure, particularly with ancestry informative markers, we did not use deviation from Hardy-Weinberg equilibrium as a criteria for excluding markers. All datasets were entered mapped to Hg19. Each dataset was then phased using SHAPEIT and imputed using the Haplotype Reference Consortium (HRC) with Minimac3 [45]. For the MEC datasets which included both 660K and 2.5M arrays, we used the overlapping SNPs (N=192,795) and imputed from those since we found that if we imputed them separately and then analyzed them together we got a large number of false positives. Each of the remaining GWAS datasets was submitted to the HRC server individually for imputation. Only variants with imputation quality scores of R2>0.5 were selected for additional analysis.

Genotype imputation for the ROOT consortium was conducted using the *IMPUTE2* software [42] with the 1000 Genomes Project phase I cosmopolitan variant set as the reference panel (October 2011 release). Genotype imputation in AABC was conducted using *IMPUTE2* software [42] to a cosmopolitan panel of all 1000 Genome Project subjects (March 2012 release). Variants with imputation score >0.3 were included in the analysis.

The Shangai Breast Cancer Study GWAS data were phased with Minimac2 and imputed with SHAPEIT using 1000 Genomes Project Phase 3. Only SNPs with an MAF ≥ 0.01 and high imputation quality (RSQR ≥ 0.5) in three GWAS in the analyses.

We used KING[43] to identify relative pairs either within the RPGEH cohort or between the RPGEH and SFBCS and/or NC-BCFR and performed the same analysis within the MEC and the CAMA study. We identified pairs of individuals with kinship coefficient >0.2 and dropped one from each of these pairs. If a relative pair included a case and control then we excluded the control. If a relative pair includes two cases or two controls we randomly dropped

one of them. We dropped 127 individuals to eliminate all closely related individuals from the combined RPGEH, SFBCS and NC-BCFR.

Empirical Assessment of Imputation accuracy: We genotyped rs3778609, the top novel SNP, in the CAMA study in samples that also had GWAS data and checked the concordance between genotyped and imputed results. We found excellent concordance between the imputed and genotyped data with 1361/1369 (99.4%) concordance between the genotyped and imputed datasets.

Genetic Ancestry Inference We implemented principal component analysis to assess genetic ancestry in each of the discovery datasets in unrelated individuals. To do so, we first LD pruned typed SNPs with $r^2 > 0.2$ in PLINK. With the remaining data, we determined the principal components (PC) using EIGENSTRAT[44] within smartpca. For the replication datasets, we used ancestry informative markers and used the program ADMIXTURE[45] to calculate genetic ancestry, assuming a 3 population model with ancestry from African, European and Native American populations.

Association Testing: We performed single variant association testing using logistic regression models and adjusting for PC's 1-10 in PLINK[46].   For the replication datasets we entered ancestry into the model as covariates. We also performed association testing separately for estrogen receptor (ER)-positive and ER-negative breast cancer using this approach. In each analysis we also included study as a covariate.

To calculate linkage disequilibrium (LD), we calculated R^2 in the controls in our dataset using PLINK. We then performed conditional analyses by entering the most significant SNP in the model as a covariate in addition to PC's 1-10.

Power: Based on the sample size for discovery (2396 cases and 7468 controls) we had ~80% power to detect an odds ratio of 1.25, 1.355 and 1.475 with allele frequencies of 0.4, 0.2, and 0.1 respectively.

Results:

Individual Association Analyses: We conducted a meta-analysis across four GWAS discovery studies (Table 1) and identified 28 variants with genome-wide significant p-values at the 6q25 susceptibility region(Supplementary Table 1). No additional genome wide significant SNPs were identified. The top variants in the region included rs140068132 and rs147157845 which are in near perfect LD (r^2=0.96) and which we have previously reported as genome-wide significant in this population[23]. Of the 28 top variants, 25 were in strong linkage disequilibrium (r^2>0.4) with rs140068132, and 3 were in low LD (r^2<0.2). These SNPs, rs3778609, rs7771984 and rs6914438 have a minor allele frequency of 0.19 and are in near perfect LD (r^2=0.99). The minor allele of these SNPs are associated with lower risk of breast cancer, and the odds ratio (OR) for rs3778609 was 0.75 (95% CI: 0.68 – 0.93, p=1.9×10^{-8}; Table 2). These SNPs are also independent (r^2<0.2) of previously reported SNPs at this locus (Supplementary Table 2). Another SNP, rs851983, was associated with a near genome-wide significant level of association (MAF=0.35; OR: 1.24, 95% CI: 1.25-1.34, p= 5.6×10^{-8} Table 2). However, this SNP is in strong LD with SNPs that were previously reported (Supplementary Table 2).

Table 1: Discovery and Replication Samples Used

Study	Genotyping Platform	Cases	Controls
Discovery: Latinas			
BCFR/SFBACS	Affy 6.0	942	699
RPGEH	Affy Axiom	225	3574
MEC	Illumina 1M, 2.5M	520	2470
CAMA	Illumina Oncoarray	698	599
Total		**2385**	**7342**
Replication: Latinas			
COLUMBUS	Affy Axiom	954	769
CCGRN	Taqman	1148	387
CAMA	Taqman	310	464
Total Latina replication		**2412**	**1620**
Replication: African American			
AABC	Illumina 1M	3153	2831
ROOT	Illumina 2.5M	1657	2029
Total African ancestry replication		4810	4860
Replication: East Asian ancestry			
Shanghai Breast Cancer Study	Affymetrix 6.0	2731	2135

We performed conditional analyses by entering rs140068132 and other top SNPs at this locus in joint models. We found that rs3778609, rs7771984 and rs6914438 all remained nominally significant in joint models adjusting for rs140068132 (Table 2), although the adjusted odds ratios for these 3 SNPs were attenuated (OR~0.81-0.84; p<=0.001). We also found that rs851983 remained nominally significant in joint models with rs140068132 with mild attenuation. When we included 3 SNPs that best represent each of the signals from each set of associated variants (rs140068132, rs3778609 and rs851983) in the same model all of the SNPs remained nominally significant with minimal attenuation of the odds ratios (compared to models including just pairs of variants; Table 2).

Table 2: Representative SNPs and association statistics from each of 4 different SNP clusters/regions that are genome wide significant.

SNP/Risk allele	Position (BP, Hg19)	Odds Ratio (95% CI)	P value	Conditional OR* p value	Joint OR** P value
rs140068132-G	6:151954834	0.58 (0.51-0.67)	1.2×10^{-13}		0.66, 5.1×10^{-8}
rs3778609-T	6:152133187	0.75 (0.68-0.83)	1.9×10^{-8}	0.83, 0.0009	0.87, 0.007
rs7771894-T	6:152145916	0.76 (0.68-0.83)	2.3×10^{-8}	0.84, 0.001	
rs6914438-T	6:152129588	0.73 (0.65-0.81)	4.3×10^{-8}	0.81, 0.001	
rs851983-G	6: 152024415	1.24 (1.15-1.34)	5.6×10^{-8}	1.17, 0.0007	1.15, 0.001

*Conditional on rs140068132
 ** Joint model with rs140068132, rs3778609 and rs851983

Technical Validation and Replication: We used data from the portion of the CAMA study that did not have GWAS data, The COLUMBUS study and the CCGRN to replicate the association with rs3778609. We found a consistent direction in all 3 studies and a nominally significant association in a meta-analysis of the 3 studies (n= cases; n=controls; OR=0.88, 95% CI: 0.78-0.99, p=0.037, Table 3).

Table 3: Replication of rs3778609 in other Latina datasets

Study	Odds Ratio (95% CI)	P value
COLUMBUS	0.87(0.73 -1.04)	0.119
CCGRN	0.89(0.70 -1.14)	0.375
CAMA (excluding GWAS)	0.88(0.69 -1.13)	0.314
Meta-analysis	**0.88(0.78 -0.99)**	**0.037**

Association of Previously Identified SNPs at 6q25: We examined previously reported SNPs in our discovery dataset (Supplementary Table 3). Only rs851984 was significant in our study. However, nearly all of the others were directionally consistent and the 95% confidence intervals overlapped with the results from previous studies.

Association with Estrogen receptor subtypes: We analyzed the association for each of the top SNPs separately and jointly by ER-status. As we have previously reported the minor (low risk) allele of rs140068132 is associated with a significantly lower odds ratio for ER-negative than for ER-positive breast cancer.   We also found a significantly stronger effect size for ER-negative breast cancer for rs3778609. The effect size for rs851983 is also greater for ER-negative breast cancer; however, there is no significant difference between ER-negative and ER-positive breast cancer for this SNP.

Table 4: Association by Estrogen Receptor Status

	ER-positive		ER-negative		P value for ER-positive vs. ER-negative
SNP	Odds Ratio (95% CI)	P value	Odds Ratio (95% CI)	P value	
rs140068132-G	0.57 (0.46-0.72)	1.0×10^{-6}	0.35 (0.23-0.54)	2.2×10^{-6}	0.033
rs3778609-T	0.75 (0.65-0.86)	6.7×10^{-5}	0.59 (0.46-0.75)	2.7×10^{-5}	0.037
rs851983-G	1.22 (1.10-1.36)	2.4×10^{-4}	1.36 (1.15-1.62)	4.6×10^{-4}	0.33

Replication in non-Latinas: The cluster of new SNPs we identified in this locus represented by rs37786109 are common in East Asians (minor allele frequency 0.27) and African (minor allele frequency 0.3) ancestry populations (Supplementary table 4), but not in European populations (minor allele frequency 0.019 in 1000 Genomes). Therefore, we

evaluated the association with this SNP in several additional populations in patients of these ancestries including the African American Breast Cancer study, the ROOT study, and the Shanghai Breast Cancer Study. We found a consistent effect for of these studies with a significant effect in a meta-analysis of rs3778609, rs7771894 and rs6914438 (Supplementary table 4). We also examined rs3778610 which is in strong LD with these SNPs and had near genome-wide significant associations in our discovery sample (Supplementary Table 1). The strongest association in non-Latina populations was with rs3778610 which was particularly stronger in African ancestry populations (Supplementary table 4).

Discussion:

We have previously reported on a SNP at 6q25 associated with a minor allele that is unique to Indigenous American populations and associated with decreased risk of breast cancer [23]. Here, we investigate this locus in greater depth in an expanded sample size of Latina breast cancer cases and controls, the largest sample size of this population analyzed for breast cancer risk to date. We have identified several SNPs that are genome-wide significant and associated with breast cancer at this locus independently of other SNPs at this region previously reported by us and those previously reported by other groups at this locus. These SNPs are located in the region of ~152.13 – 152.14 MB (Hg19). Replication in African American and Asian samples supports the association with this SNP. In addition, we have also shown that these novel variants at this locus have significantly stronger effect sizes on ER-negative breast cancer. Prior studies have also demonstrated a stronger effect size with ER-negative breast cancer for most variants at the 6q25 locus, consistent with our data [18, 27].

Prior studies in other populations have reported a series of independent SNPs affecting breast cancer risk [11, 18, 24-27, 31]. The variants previously reported in other populations are not significant in our studies but most are directionally consistent. A combined fine-mapping and functional study of SNPs mapped in other populations at this locus found that they affect

expression of *ESR1*, *RMND1* and *CCDC170[27]*. Since the new variants we report are common only in non-European ancestry populations, there is limited data to explore the potential effects of these variants on gene expression.

Our study is limited by sample size. Therefore, it is possible that we have missed other variants at this locus. In fact, several previously reported variants have odds ratios with point estimates that are close to the previously reported results, but have 95% confidence intervals that include 1, as would be expected with insufficient power.   The effect size we observed in the replication dataset is substantially lower than in the discovery dataset, likely due to winner's curse. However, even if we take the replication odds ratio (0.88) as the closest to the true effect size of these SNPs, this is still a relatively large effect for a common variant. It is likely that there are other variants that have not yet been identified in European GWAS due to low allele frequency and that could be identified in Latinas where they are more common. Larger studies of Latina women are needed to identify these variants.

Conclusion

Our study demonstrates additional unique associations with variants at 6q25 and breast cancer risk. This further highlights the important contribution of this locus to breast cancer susceptibility, particularly ER-negative breast cancer susceptibility. Additional fine- mapping and functional studies are needed to elucidate all of the causal variants in our population. However, the variants we identified in this study can be useful to add to the increasing pool of common variants coming from GWAS and will be particularly useful to risk stratify women of Latin American ancestry for breast cancer risk.

List of abbreviations:

GWAS: Genome wide association study
SNP: Single nucleotide polymorphism
OR: odds ratio
CI: confidence interval
PCA: Principal components analysis
PC: Principal component
ER: Estrogen receptor

Declarations:

Ethics approval and consent to participate: The following Institutional Review Boards have approved the collection of different datasets which have contributed to this study: University of California, San Francisco, the Cancer Prevention Institute of California, University of Southern California, Kaiser Permanente Division of Research Northern California, University of Chicago, Vanderbilt University, National Institute of Public Health, Mexico

Consent for publication: Not applicable

Availability of data and material: The following GWAS datasets have been submitted to dbGAP: SFBACS (phs000912.v1.p1), MEC Latina (phs000517.v2.p1), AABC (phs000851.v1.p1), ROOT (phs000383.v1.p1), Shanghai Breast Cancer Study (phs000799.v1.p1). The NC-BCFR was determined by the IRB not to be eligible for submission to dbGAP based on the consent. Summary statistics are available for the 6q25 locus from the NC-BCFR, the Kaiser RPGEH and the COLUMBUS study are available from the corresponding author by request.

Competing interests: The authors declare no competing interests.

References:

1. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC: **Linkage of early-onset familial breast cancer to chromosome 17q21**. *Science* 1990, **250**(4988):1684-1689.

2. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W *et al*: **A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1**. *Science* 1994, **266**(5182):66-71.

3. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G: **Identification of the breast cancer susceptibility gene BRCA2**. *Nature* 1995, **378**(6559):789-792.

4. Walsh T, King MC: **Ten genes for inherited breast cancer**. *Cancer Cell* 2007, **11**(2):103-105.

5. Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R *et al*: **Genome-wide association study identifies novel breast cancer susceptibility loci**. *Nature* 2007, **447**(7148):1087-1093.

6. Hunter DJ, Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A *et al*: **A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer**. *Nat Genet* 2007, **39**(7):870-874.

7. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A *et al*: **Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer**. *Nat Genet* 2007, **39**(7):865-869.

8. Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, Morrison J, Maranian M, Pooley KA, Luben R *et al*: **Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2**. *Nat Genet* 2009, **41**(5):585-590.

9. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K *et al*: **A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1)**. *Nat Genet* 2009, **41**(5):579-584.

10. Long J, Cai Q, Shu XO, Qu S, Li C, Zheng Y, Gu K, Wang W, Xiang YB, Cheng J *et al*: **Identification of a functional genetic variant at 16q12.1 for breast cancer risk: results from the Asia Breast Cancer Consortium**. *PLoS Genet* 2010, **6**(6):e1001002.

11. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, Seal S, Ghoussaini M, Hines S, Healey CS *et al*: **Genome-wide association study identifies five new breast cancer susceptibility loci**. *Nat Genet* 2010, **42**(6):504-507.

12. Wang X, Pankratz VS, Fredericksen Z, Tarrell R, Karaus M, McGuffog L, Pharoah PD, Ponder BA, Dunning AM, Peock S *et al*: **Common variants associated with breast cancer in genome-wide association studies are modifiers of breast cancer risk in BRCA1 and BRCA2 mutation carriers**. *Hum Mol Genet* 2010, **19**(14):2886-2897.

13. Antoniou AC, Kartsonaki C, Sinilnikova OM, Soucy P, McGuffog L, Healey S, Lee A, Peterlongo P, Manoukian S, Peissel B *et al*: **Common alleles at 6q25.1 and 1p11.2 are associated with breast cancer risk for BRCA1 and BRCA2 mutation carriers**. *Hum Mol Genet* 2011, **20**(16):3304-3321.

14. Cai Q, Long J, Lu W, Qu S, Wen W, Kang D, Lee JY, Chen K, Shen H, Shen CY *et al*: **Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium**. *Hum Mol Genet* 2011, **20**(24):4991-4999.

15. Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, Millikan RC, Wang X, Ademuyiwa F, Ahmed S, Ambrosone CB *et al*: **A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer**. *Nat Genet* 2011, **43**(12):1210-1214.

16. Couch FJ, Gaudet MM, Antoniou AC, Ramus SJ, Kuchenbaecker KB, Soucy P, Beesley J, Chen X, Wang X, Kirchhoff T *et al*: **Common variants at the 19p13.1 and ZNF365 loci are associated with ER subtypes of breast cancer and ovarian cancer risk in BRCA1 and BRCA2 mutation carriers**. *Cancer Epidemiol Biomarkers Prev* 2012, **21**(4):645-657.

17. Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, Dicks E, Dennis J, Wang Q, Humphreys MK, Luccarini C *et al*: **Genome-wide association analysis identifies three new breast cancer susceptibility loci**. *Nat Genet* 2012, **44**(3):312-318.

18. Siddiq A, Couch FJ, Chen GK, Lindstrom S, Eccles D, Millikan RC, Michailidou K, Stram DO, Beckmann L, Rhie SK *et al*: **A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11**. *Hum Mol Genet* 2012, **21**(24):5373-5384.

19. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, Schmidt MK, Chang-Claude J, Bojesen SE, Bolla MK *et al*: **Large-scale genotyping identifies 41 new loci associated with breast cancer risk**. *Nat Genet* 2013, **45**(4):353-361, 361e351-352.

20. Zheng W, Zhang B, Cai Q, Sung H, Michailidou K, Shi J, Choi JY, Long J, Dennis J, Humphreys MK *et al*: **Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls**. *Hum Mol Genet* 2013, **22**(12):2539-2550.

21. Huo D, Feng Y, Haddad S, Zheng Y, Yao S, Han YJ, Ogundiran TO, Adebamowo C, Ojengbede O, Falusi AG *et al*: **Genome-wide association studies in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer**. *Hum Mol Genet* 2016, **25**(21):4835-4846.

22. Fejerman L, Chen GK, Eng C, Huntsman S, Hu D, Williams A, Pasaniuc B, John EM, Via M, Gignoux C *et al*: **Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas**. *Hum Mol Genet* 2012, **21**(8):1907-1917.

23. Fejerman L, Ahmadiyeh N, Hu D, Huntsman S, Beckman KB, Caswell JL, Tsung K, John EM, Torres-Mejia G, Carvajal-Carmona L *et al*: **Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25**. *Nat Commun* 2014, **5**:5260.

24. Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL *et al*: **Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1**. *Nat Genet* 2009, **41**(3):324-328.

25. Cai Q, Wen W, Qu S, Li G, Egan KM, Chen K, Deming SL, Shen H, Shen CY, Gammon MD *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**(4):1344-1355.

26. Hein R, Maranian M, Hopper JL, Kapuscinski MK, Southey MC, Park DJ, Schmidt MK, Broeks A, Hogervorst FB, Bueno-de-Mesquita HB *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**(8):e42380.

27. Dunning AM, Michailidou K, Kuchenbaecker KB, Thompson D, French JD, Beesley J, Healey CS, Kar S, Pooley KA, Lopez-Knowles E *et al*: **Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170**. *Nat Genet* 2016, **48**(4):374-386.

28. Brand JS, Li J, Humphreys K, Karlsson R, Eriksson M, Ivansson E, Hall P, Czene K: **Identification of two novel mammographic density loci at 6Q25.1**. *Breast Cancer Res* 2015, **17**:75.

29. Ellingjord-Dale M, Grotmol T, Lee E, Van Den Berg DJ, Hofvind S, Couto E, Sovio U, Dos-Santos-Silva I, Ursin G: **Breast cancer susceptibility variants and mammographic density phenotypes in norwegian postmenopausal women**. *Cancer Epidemiol Biomarkers Prev* 2014, **23**(9):1752-1763.

30. Mariapun S, Ho WK, Kang PC, Li J, Lindstrom S, Yip CH, Teo SH: **Variants in 6q25.1 Are Associated with Mammographic Density in Malaysian Chinese Women**. *Cancer Epidemiol Biomarkers Prev* 2016, **25**(2):327-333.

31. Long J, Cai Q, Sung H, Shi J, Zhang B, Choi JY, Wen W, Delahanty RJ, Lu W, Gao YT *et al*: **Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer**. *PLoS Genet* 2012, **8**(2):e1002532.

32. Gonzalez Burchard E, Borrell LN, Choudhry S, Naqvi M, Tsai HJ, Rodriguez-Santana JR, Chapela R, Rogers SD, Mei R, Rodriguez-Cintron W *et al*: **Latino populations: a unique opportunity for the study of race, genetics, and social environment in epidemiological research**. *Am J Public Health* 2005, **95**(12):2161-2168.

33. Park SL, Cheng I, Haiman CA: **Genome-wide association studies of cancer in diverse populations**. *Cancer Epidemiol Biomarkers Prev* 2017.

34. Lisker R, Ramirez E, Briceno RP, Granados J, Babinsky V: **Gene frequencies and admixture estimates in four Mexican urban centers**. *Hum Biol* 1990, **62**(6):791-801.

35. Hanis CL, Hewett-Emmett D, Bertin TK, Schull WJ: **Origins of U.S. Hispanics. Implications for diabetes**. *Diabetes Care* 1991, **14**(7):618-627.

36. Williams RC, Knowler WC, Pettitt DJ, Long JC, Rokala DA, Polesky HF, Hackenberg RA, Steinberg AG, Bennett PH: **The magnitude and origin of European-American admixture in the Gila River Indian Community of Arizona: a union of genetics and demography**. *Am J Hum Genet* 1992, **51**(1):101-110.

37. Bertoni B, Budowle B, Sans M, Barton SA, Chakraborty R: **Admixture in Hispanics: distribution of ancestral population contributions in the Continental United States**. *Hum Biol* 2003, **75**(1):1-11.

38. Consortium STD, Williams AL, Jacobs SB, Moreno-Macias H, Huerta-Chagoya A, Churchhouse C, Marquez-Luna C, Garcia-Ortiz H, Gomez-Vazquez MJ, Burtt NP *et al*: **Sequence variants in SLC16A11 are a common risk factor for type 2 diabetes in Mexico**. *Nature* 2014, **506**(7486):97-101.

39. MacDonald DJ, Blazer KR, Weitzel JN: **Extending comprehensive cancer center expertise in clinical cancer genetics and genomics to diverse communities: the power of partnership**. *Journal of the National Comprehensive Cancer Network : JNCCN* 2010, **8**(5):615-624.

40. Feng Y, Stram DO, Rhie SK, Millikan RC, Ambrosone CB, John EM, Bernstein L, Zheng W, Olshan AF, Hu JJ *et al*: **A comprehensive examination of breast cancer risk loci in African American women**. *Hum Mol Genet* 2014, **23**(20):5518-5526.

41. Fejerman L, Romieu I, John EM, Lazcano-Ponce E, Huntsman S, Beckman KB, Perez-Stable EJ, Gonzalez Burchard E, Ziv E, Torres-Mejia G: **European ancestry is positively associated with breast cancer risk in Mexican women**. *Cancer Epidemiol Biomarkers Prev* 2010, **19**(4):1074-1082.

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

43. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM: **Robust relationship inference in genome-wide association studies**. *Bioinformatics* 2010, **26**(22):2867-2873.

44.   Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nat Genet* 2006, **38**(8):904-909.

45.   Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals**. *Genome Res* 2009, **19**(9):1655-1664.

46.   Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ *et al*: **PLINK: a tool set for whole-genome association and population-based linkage analyses**. *Am J Hum Genet* 2007, **81**(3):559-575.