AD_____

Award Number: W81XWH-08-1-0383

TITLE: A Genome-wide Breast Cancer Scan in African Americans

PRINCIPAL INVESTIGATOR: Christopher A. Haiman, Sc.D.

CONTRACTING ORGANIZATION: University of Southern California, Los Angeles, CA 90089

REPORT DATE: June 2012

TYPE OF REPORT: Revised Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
June 2012	Revised Final	01June 2008 - 31 May 2012

4. TITLE AND SUBTITLE
A Genome-wide Breast Cancer Scan in African Americans

5a. CONTRACT NUMBER
.

5b. GRANT NUMBER
W81XWH-08-1-0383

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Christopher A. Haiman

E-Mail: Haiman@usc.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Southern California
Los Angeles, California 90089-9235

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The focus of this proposal was to discover susceptibility loci for overall and estrogen-receptor (ER) negative breast cancer that are particularly important for women of African ancestry. Over the past four years, we conducted the first genome-wide association study (GWAS) of breast cancer in African American women. For this effort, we were successful in establishing a consortium of breast cancer case-control studies with DNA available for genomic analysis. The GWAS, and subsequent replication genotyping of strong signals from stage 1, did not reveal any novel locus for overall breast cancer in this population. However, working with ongoing GWAS of ER-negative disease in European ancestry populations, we revealed two novel risk loci for ER-negative disease that are particularly important for women of African ancestry. We estimate that one of these loci may explain 20% of the greater risk of ER-negative disease subtypes, including triple negative disease, in women of African ancestry compared to women of other ancestries. The GWAS data have also been utilized to fine-map the more than 70 known breast cancer risk loci which has revealed generalizability of the known risk variants found in European and information that we believe will inform risk stratification in this population.

15. SUBJECT TERMS
Breast Cancer, Genome-wide Association Study, African Americans

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	19	USAMRMC

Table of Contents

Introduction

Genome-wide association studies (GWAS) of breast cancer have been completed among populations of European ancestry, and several regions have been identified that appear to contribute susceptibility to this cancer. Recent data suggests that not all risk alleles for common cancers will be revealed however by studies limited to Whites of European ancestry, and that similar efforts in other racial and ethnic populations will be needed to identify the full spectrum of common risk alleles that contribute to disease risk in the population. To identify genetic risk alleles for breast cancer risk among African American women we have performed a well-powered whole-genome association scan. For this project we have established a collaborative network of investigators whose careers have been dedicated to studying breast cancer in minority populations who have contributed samples and covariates from each of their respective studies to identify genetic variants that contribute to risk of breast cancer in this minority population. We have completed a GWAS of >1.1 SNPs in >3000 African American breast cancer cases and >2,700 controls. With these data we have validated and improved upon markers of risk at the known breast cancer risk regions that better characterize their contribution to breast cancer risk in women of African ancestry. In collaboration with GWAS in populations of European ancestry we have also revealed novel risk loci for breast cancer including regions that contribute to risk for estrogen receptor (ER)-negative breast cancer.

Body

The Specific Aim of this application was to identify genetic risk alleles for breast cancer among African American women by performing a well-powered whole-genome association scan. Here we describe the major research accomplishments associated with each task outlined in the approved Statement of Work as well as additional novel findings and scientific contributions that have emanated from this work.

Task 1: To genotype 1,000,000 single nucleotide polymorphisms (SNPs) using the Illumina Infinium 1M technology in 1,000 invasive African American breast cancer cases and 1,000 African American controls.

With the costs of genotyping decreasing we were able to genotype >3,000 cases and >2,800 controls. These samples were selected from the studies participating in this effort (Table 1).

Table 1. African American Breast Cancer Studies.

Study	Full Name	Cases	Controls
MEC	Multiethnic Cohort	734	1027
BWHS	Black Women's Health Study	825	1170
CARE	The Los Angeles component of The Women's Contraceptive and Reproductive Experiences Study	380	224
CBCS	The Carolina Breast Cancer Study	656	608
NBHS/ SCCS	The Nashville Breast Health Study/ Southern Community Cohort	1242	1002
WAABCO	Pennsylvania, Nigeria, Barbados, Baltimore, Chicago	1281	1148
PLCO	Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial	64	133
SFBCS/ NC-BCFR	The San Francisco Bay Area Breast Cancer Study / Northern California Breast Cancer Family Registry	612	284
WCHS	The Women's Circle of Health Study	272	240
WFBC	Wake Forest University Breast Cancer Study	125	153
WHI	The Women's Health Initiative	316	316
WISE	The Women's Insights and Shared Experiences Study	145	367
TOTAL		**6661**	**6622**

Specific details about genotyping, quality control, statistical analysis and results of the GWAS are described in detail below as well as in our recent publication [1] (Chen et al, Human Genetics, 2012, see Appendix). Tables and Figures that support this work are provided in the attached manuscript.

Genotyping in stage 1 was conducted using the Illumina Human1M-Duo BeadChip. Of the 5,984 samples from these studies (3,153 cases and 2,831 controls), we attempted genotyping of 5,932, removing samples (n = 52) with DNA concentrations <20 ng/ul. Following genotyping, we removed samples based on the following exclusion criteria: 1) unexpected replicates (≥98.9% genetically identical) that we were able to confirm through discussions with study investigators (only one of each replicate was removed, n = 15); 2) unknown replicates that we were not able to confirm (pair or triplicate removed, n = 14); 3) samples with call rates <95% after a second genotyping attempt (n = 100); 4) samples with ≤ 5% African ancestry (n = 36) (discussed below); and 5) samples with <15% mean heterozygosity of SNPs on the X chromosome and/or similar mean allele intensities of SNPs on the X and Y chromosomes (n = 6) as these are likely to be males.

We removed SNPs with <95% call rate (n = 21,732) or minor allele frequencies (MAFs) <1% (n = 80,193). To assess genotyping reproducibility we included 138 known replicate samples; the average concordance rate was 99.95% (>99.93% for all pairs). We also eliminated SNPs with genotyping concordance rates <98% based on the replicates (n = 11,701). The final analysis dataset included 1,043,036 SNPs genotyped on 3,016 cases and 2,745 controls, with an average SNP call rate of 99.7% and average sample call rate of 99.8%. Hardy-Weinberg equilibrium (HWE) was not used as a criterion for removing SNPs; none of the SNPs selected for replication deviated from HWE in controls in each study (based on a cut-off of $p<0.001$).

In stage 1, we utilized STRUCTURE [2] to infer percent African ancestry on an individual level. A total of 2,546 ancestry-informative SNPs from the Illumina array were selected based on low inter-marker correlation and ability to differentiate between samples of African and European descent. In evaluating the distribution of the fraction of African ancestry across the stage 1 populations, statistically significant differences (ANOVA $p<10^{-16}$) were noted. We also applied principal components analysis (PCA) [3] to estimate axes of variation among the 5,761 individuals using the same 2,546 ancestry informative markers. The first eigenvector accounted for 10.1% of the variation between subjects, and subsequent eigenvectors accounted for no more than 0.5%. Using input genotypes from the HapMap populations, CEU (CEPH Utah), YRI (Yoruba), and JPT (Japanese), we determined that the first eigenvector captures clearly differentiates Europeans (CEU) and West Africans (YRI) in the HapMap samples.

In Stage 1, we observed no evidence of inflation of the test statistic ($\lambda = 1.01$) for the 1,043,036 genotyped and 2,067,098 imputed SNPs analyzed in stage 1, and no excess of very small p-values beyond what was expected. We observed no SNP to be associated with disease status at a genome-wide level of significance ($p<5\times10^{-8}$) in stage 1. The most statistically significant association was noted with SNP rs7610073 located in intron 2 of the gene *GRM7* (metabotropic glutamate receptor 7) on chromosome 3p26 (risk allele frequency 0.64; OR per allele = 1.22; $p = 7.4\times10^{-7}$). A second signal was also noted ~486 kb upstream of *GRM7* (rs10510333: risk allele frequency = 0.18; OR per allele = 1.24; $p = 8.2\times10^{-6}$). The associations with these 2 markers were independent and remained statistically significant when both were included in the same model (p-values of 8.3×10^{-7} and 9.3×10^{-6}, respectively).

<u>Task 2</u>: We will perform follow-up genotyping of a minimum of 13,800 SNPs using an Illumina Infinium iSELECT custom SNP array in 2,000 African American breast cancer cases and 2,000 African American controls. The actual number of SNPs to be examined in stage 2 will depend on the per chip/sample genotyping cost when stage 2 genotyping will be conducted. Fewer

SNPs were genotyped in Stage 2 because a substantially larger number of samples were genotyped in Stage 1.

In Stage 2, we genotyped 66 SNPs with association p-values less than 2×10^{-4} (from Stage 1) for replication testing in the stage 2 studies (>3,000 cases and >3,000 controls). None of these SNPs replicated with stage 2-wide significance of <0.0008 (0.05/66), but 2 replicated with a p-value <0.05 and an OR in the same direction as that observed in stage 1. Combining results from stages 1 and 2, no SNP achieved genome-wide significance. The smallest combined p-values were noted for the two SNPs that replicated in stage 2: rs4322600 located ~100 kb upstream of the gene *GALC* (galactosylceramidase) on chromosome 14q31 (risk allele frequency = 0.78, OR per allele = 1.18, $p = 4.3\times10^{-6}$) and rs10510333 located ~486 kb upstream of *GRM7* on chromosome 3p26 (risk allele frequency = 0.18, OR per allele = 1.15, $p = 1.5\times10^{-5}$). We found no strong statistical evidence that the associations with these two loci differ by ER status (p-values for heterogeneity in case-only testing: rs10510333: p=0.67; rs4322600: p=0.85).

<u>Task 3</u>: Case-only analyses will be performed using the combined data from stages 1 + 2 to assess potential heterogeneity of allelic effects by disease phenotype (e.g. ER- and/or aggressive tumors) using a model for exposure as a function of genotype only for the data from the cases.

With only 1,000 ER-negative cases included in Stage 1, **in years 3 and 4** we reached out to other ongoing GWAS of ER-negative disease in other populations. These efforts to find loci for ER-negative disease are described below and in a number of manuscripts (Haiman et al, Nature Genetics, 2012 and Siddiq et al, Human Molecular Genetics, in press; see Appendix). [4]

Chromosome 5p15

To search for genetic risk factors for ER-negative breast cancer phenotypes, we initially combined results the African American GWAS [AABC: 3,016 cases (1,004 with ER-negative disease) and 2,745 controls] with results from a GWAS of triple negative breast cancer in women of European ancestry (TNBCC: 1,562 cases and 3,399 controls) (Haiman et al, Nature Genetics, 2012, see Appendix). ***This work took place in years 3 and 4 of the project period.*** In TNBCC, cases were genotyped with the Illumina 660W array. Genotypes of TNBCC cases were compared with GWAS data for publicly available controls. Both studies imputed genotypes for common SNPs in Phase 2 HapMap populations (release 21) and a total of 3,154,485 SNPs, genotyped and imputed were analyzed in stage 1 of the meta-analysis.

We observed little evidence of inflation in the test statistics in AABC (λ=1.01), TNBCC (λ=1.04) or in the meta-analysis of the two GWAS (λ=1.02). In the combined results, only SNP rs10069690 (NCBI36/hg18, chr5:1,332,790) located in intron 4 of the *TERT* gene at chromosome 5p15 displayed a genome-wide significant association with ER negative breast cancer (AABC: OR per allele=1.32, $p=1.3\text{x}10^{-6}$; TNBCC: OR=1.25, $p=1.2\text{x}10^{-3}$; combined OR =1.29, $p=1.0\text{x}10^{-8}$). To further confirm the association at 5p15, we genotyped SNP rs10069690 in women of European ancestry, which included 8,365 cases (1,359 ER negatives) and 10,935 controls from the NCI Breast and Prostate Cancer Cohort Consortium (BPC3) and 6,182 cases (933 ER negatives) and 5,966 controls from Studies of Epidemiology and Risk Factors in Cancer Heredity (SEARCH). Evidence for replication was observed for rs10069690 and ER negative breast cancer in both studies (BPC3: OR=1.09, p=0.072; SEARCH: OR=1.21, $p=6.9\text{x}10^{-4}$).

In an analysis of ER positive cases, rs10069690 was only weakly associated with risk in African Americans (AABC: 1,558 ER positive cases and 2,743 controls with genotype data: OR=1.08; p=0.10) and in women of European ancestry (BPC3: 4,890 ER positive cases and

10,397 controls, OR=1.04, p=0.19; SEARCH: 3,534 ER positive cases and 5,966 controls, OR=1.03, p=0.37; combined for all populations: OR=1.04, p=0.03, pHet = 0.69). The statistical power to detect an OR of 1.19 (observed for ER negative disease) for ER+ positive disease was >99% in the combined sample (9,982 cases and 19,106 controls) assuming the risk allele frequency of 0.26 in Europeans. This result suggests that the association with breast cancer might be specific for ER negative subtypes (P-value for case-only test of ER negative versus ER positive = $1.0x10^{-4}$).

We further stratified the cases by HER2 status to assess whether this region may be a risk locus for triple negative disease. In AABC, BPC3 and SEARCH the association with rs10069690 was greater for ER/PR/HER2 negative tumors than for ER/PR negative/HER2 positive tumors, and in combining all studies, including TNBCC, the association with rs10069690 was significantly greater for triple negative disease [3,706 ER/PR/HER2 negative cases and 19,728 controls with genotype data, OR=1.25, $p=8.6x10^{-10}$; 376 ER/PR negative/HER2 positive cases and 19,106 controls, OR=1,04, p=0.64, P-value for case-only test =0.011]. The association with rs10069690 was also observed to be significantly greater for ER negative and triple negative disease at younger ages (<50 years: ER negative, OR=1.32, $p=7.0x10^{-9}$; triple negative, OR=1.47, $p=2.4x10^{-9}$; P for interaction with age = 0.039 and $3.8x10^{-3}$, respectively). We found no significant association with rs1006960 among ER/PR positive cases when stratified by HER2 status [513 ER/PR/HER2 positive cases and 18,126 controls, OR=1.08, p=0.30; 2,808 ER/PR positive/HER negative cases and 18,126 controls, OR=1.03, p=0.30], which suggests the association may be limited to triple negative disease and not all HER2 negative tumors.

Chromosome 20q11

In order to identify genetic loci associated with risk of ER-negative breast cancer, we conducted a meta-analysis of three GWAS of ER-negative breast cancer, comprising 4,754 cases and 31,663 controls with further replication in an additional 11,209 cases (946 with ER-negative disease) and 16,057 controls (Siddiq et al, Human Molecular Genetics, in press; see Appendix). ***This work took place in year 4 of the study period.***

The meta-analysis included GWAS of ER-negative breast cancer (4,754 ER-negative cases and 31,663 controls) from the NCI Breast and Prostate Cancer Cohort Consortium (BPC3) (2,188 ER-negative cases and 25,519 controls of European ancestry), the Triple Negative Breast Cancer Consortium (TNBCC) (1,562 triple negative cases and 3,399 controls of European ancestry) and the African American Breast Cancer Consortium (AABC) (1,004 ER-negative cases and 2,745 controls). We observed little evidence of over-inflation in the test statistics ($\lambda \leq 1.04$ for each study; $\lambda=1.04$ for meta-analysis). A total of 86 SNPs were associated with ER-negative breast cancer at $P \leq 10^{-5}$. An in silico replication of the 86 SNPs was conducted using GWAS of European (BCAC combined), Latino (MEC-LAT, SFBCS/NC-BCFR) and Japanese (MEC-JPT) ancestry populations, totaling 11,209 breast cancer cases (946 with ER-negative disease) and 8,404 controls (Stage 2).

Combining results for ER-negative breast cancer from stages 1 and 2, variants in three regions showed genome-wide significance [20q11-rs2284378, T allele: odds ratio, OR=1.16, $P = 1.1x10^{-8}$ (PGC = $7.7x10^{-8}$; Table 1); 19p13-rs8100241, G allele: OR=1.14, $P=3.5x10^{-8}$; 6q25-rs9383938, T allele: OR=1.28, $P = 2.37 \times 10^{-10}$]. Variants at 6q25 have previously been associated with breast cancer risk [5], and variants at the 19p13 locus have been associated with ER-negative and triple negative breast cancer risk [6,7]. The rs2284378 variant at 20q11 is located in a region containing *RALY* (RNA binding protein, autoantigenic), *EIF2S2* (eukaryotic translation initiation factor 2, subunit 2 beta) and ~100kb upstream of *ASIP* (agouti signaling protein), and is in high linkage disequilibrium ($r^2=0.96$ and D'=1) with rs4911414, which has been associated with melanoma and basal cell carcinoma. [8-10] The T allele at rs2284378 was associated with an increased ER-negative breast cancer risk (OR>1) in all racial/ethnic

populations, except Japanese (OR=0.99). However this group had the smallest sample size. Furthermore, no significant evidence of heterogeneity was observed by race (P=0.28) or study (P=0.54). When the study was extended to include all available breast cancer cases (ER-positive and ER-negative) and controls from the participating GWAS, rs2284378 showed a weaker association with overall breast cancer (OR=1.08, $P=1.3x10^{-6}$ based on 17,868 cases and 43,744 controls) and no evidence for association with ER-positive disease (OR=1.01, P=0.67 based on 9,965 cases and 22,902 controls. A case-only analysis of ER-negative versus ER-positive breast cancer indicated a highly significant difference in ORs by ER status ($P=1.3x10^{-4}$). Furthermore, rs2284378 appeared more strongly associated with triple negative breast cancer (OR=1.16; $P=6.4x10^{-3}$), than ER-negative, PR-negative, HER2-positive breast cancer (OR=1.07, P=0.41), although these differences were not statistically significant (case-only P=0.44).

Next, we examined the associations between all candidate loci from stage 1 (n=86 SNPs) and overall breast cancer risk using all available breast cancer cases and controls from the studies in stages 1 and 2. We identified genome-wide statistically significant associations with variants at 6q25 (rs9383938, T allele: OR=1.20; $P=8.7x10^{-14}$), and a recently reported risk locus near the PTHLH gene at 12p11 [11] (rs1975930, T allele: OR=1.22; $P=1.4x10^{-13}$). In addition, we observed genome wide significant associations with multiple variants in a gene-desert located at 6q14. Allele C of rs17530068 at 6q14 was associated with increased risk for overall breast cancer risk (OR=1.12; $P=1.1x10^{-9}$; PGC $=9.4x10^{-9}$) and both ER-positive (OR=1.09; $P=1.5x10^{-5}$) and ER-negative (OR=1.16, $P=2.5x10^{-7}$) breast cancer. We observed no evidence of risk heterogeneity for rs17530068 by ER status (case-only analysis P=0.53); study (Phet=0.16); or race/ethnicity (Phet =0.30). Furthermore, rs17530068 appeared more strongly associated with ER-negative, PR-negative, HER2-positive breast cancer (OR=1.26, $P=8.0x10^{-3}$), than triple negative breast cancer (OR=1.12, P=0.07), although these differences were not statistically significant (case-only P=0.17).

Fine-mapping of Known Breast Cancer Risk Loci. This study does not fall under any of the Tasks specifically outlined in the Statement of Work however it is a logical extension of our work and makes good use of the dense SNP data genome-wide generated in Stage 1 of the scan. A manuscript describing these finding is provided in the Appendix (Haiman et al, Human Molecular Genetics, 2012). [12] ***This work started in year 3 and was completed in year 4 of the study period.***

We tested common genetic variation at the breast cancer risk loci identified in women of European and Asian descent in the stage 1 African American breast cancer sample to identify markers of risk that are relevant to this population. More specifically, we examined the index variants and conducted fine-mapping of the locus to both improve the current set of risk markers in African Americans as well as to identify new risk variants for breast cancer. We then applied this information to model breast cancer risk in African American women in attempt to characterize the spectrum of genetic risk in this population defined by common variants at the known risk loci.

We tested the 19 validated breast cancer risk variants (referred as "index variants") at 1p11, 2q35, 3p24, 5p12, 5q11, 6q25, 8q24, 9p21, 9q31, 10p15, 10q21, 10q22, 10q26, 11p15, 11q13, 14q24, 16q12, 17q23 and 19p13 in models adjusted for age, study, global ancestry (the first 10 eigenvectors) and local ancestry;[5,12-17] 17 SNPs were directly genotyped, while 2 were imputed using MACH ($r^2>0.98$). All 19 variants were common (≥0.05) in African Americans, with 11 variants being more common in Europeans than African Americans. In previous GWAS, the index signals had very modest odds ratios (1.05-1.29 per copy of the risk allele) and our sample size provided ≥70% statistical power to detect the reported effects for 12 of the 19 variants (at P<0.05). We observed positive associations with 11 of the 19 variants (OR >1) however only 4 were statistically significant (P<0.05 at 2q35, 9q31, 10q26 and 19p13). Of the 15 variants that

were not replicated at P<0.05, statistical power was <70% for only 7 of the variants. Although power was more limited, we also evaluated associations by estrogen receptor (ER) status as some risk variants have been found to be more strongly associated with ER-positive (ER+) or ER-negative (ER-) breast cancer. We observed positive associations with 12 variants (2 at P<0.05) for ER+ disease (n=1,520) and with 9 variants for ER- (3 at P<0.05; n=988). For only one variant did we observe statistically significant risk heterogeneity by ER status (rs13387042 at 2q35, P=0.013).

Aside from statistical power, the lack of a statistically significant association with an index variant (OR>1 and p<0.05) suggests that the particular variant revealed in the GWAS populations may not be adequately correlated with the biologically relevant allele in African Americans. In an attempt to identify a better genetic marker of risk in African Americans we conducted fine-mapping across all risk regions using genotyped SNPs on the Illumina 1M array and imputed SNPs to Phase 2 HapMap populations. Through fine-mapping we revealed markers in four regions that were more significantly associated with risk than the index signal (>1 order of magnitude change in the p-value) and are likely capturing the same signal (2q35, 5q11, 10q26 and 19p13). We also identified markers in four regions that are not correlated with the index signal in the GWAS populations (8q24, 10q22, 11q13 and 16q12) and may represent putative novel risk variants, with one being specific for ER+ disease (8q24). These regions are discussed below.

Risk variants that better define the index signal in African Americans

2q35

The index signal at 2q35 was statistically significantly associated with risk of overall breast cancer (rs13387042: OR=1.12, $P=7.5\times10^{-3}$) and ER+ disease (OR=1.22, $P=2.6\times10^{-4}$). However, we found stronger associations with two markers that are each modestly correlated with the index signal in CEU and YRI: rs13000023 with overall breast cancer (OR=1.20, $P=5.8\times10^{-4}$) and rs12998806: with ER+ disease (OR=1.39, $P=3.3\times10^{-6}$). The signal in this region appeared limited to ER+ breast cancer, which is consistent with the initial report of this risk locus. [15]

5q11

We found a positive non-significant association with the index signal at 5q11, which is located 79 kb centromeric of the *MAP3K1* gene (rs889312: OR=1.07, P=0.084). Fine-mapping revealed statistically significant associations with markers, rs16886165 for overall breast cancer (OR=1.15, $P=6.5\times10^{-4}$) and rs832529 for ER- disease (OR=1.22, $P=1.3\times10^{-3}$). These SNPs show greater correlation with the index signal in Europeans (CEU, r^2=0.40 and 0.46) than in Africans (YRI, r^2<0.01 and r^2=0.09), which suggests that they may be better markers of the biologically functional variant in African Americans.

10q26

Both the index signal, rs2981582 (OR=1.11, $P=8.6\times10^{-3}$), and rs2981578, that was identified previously through fine-mapping in African Americans (which some of these studies contributed to)[18], were statistically significantly associated with risk (OR=1.24, $P=1.7\times10^{-4}$). Variant rs2981578 was the most strongly associated marker in the region for overall breast cancer and for ER+ disease, which is consistent with previous reports of variation in this region being more strongly associated with ER+ breast cancer.[19] In fine-mapping the locus we observed a suggestive association with a correlated marker and ER- disease (rs2912774: OR=1.19, $P=2.1\times10^{-3}$) however the association was also noted with ER+ disease (OR=1.10, P=0.041) and is likely capturing the same signal as rs2981578.

19p13

19p13 was the first risk locus reported to harbor a variant that may be specific for ER- disease.[20] In African Americans, the index variant was statistically significantly associated with risk of overall breast cancer (rs2363956: OR=1.14, $P=8.0\times10^{-4}$), as well as ER+ (OR=1.12, P=0.016) and ER- disease (OR=1.14, P=0.01). The most significant association in the region for overall

breast cancer and ER+ disease was with rs3745185 ($P=3.7\times10^{-5}$ and $P=8.2\times10^{-4}$, respectively), which is likely to be capturing the same functional variant ($r^2=0.57$ in CEU and 0.19 in YRI). The most significant marker for ER- breast cancer was correlated with both rs2363956 and rs3745185 (rs11668840: OR=1.25, $P=5.1\times10^{-5}$).

Novel risk-associated markers at breast cancer susceptibility loci.

8q24

Given the importance of the 8q24 locus in cancer, we conducted association testing across the entire cancer risk region (126.0 Mb-130.0 Mb).[21,22] The index signal (rs13281615) was not statistically significantly associated with risk in African Americans, nor did we identify significant associations with correlated SNPs. However, we did detect a significant association with rs16902056 and ER+ breast cancer (risk allele frequency, 0.95; $P=6.7\times10^{-6}$; ER-: P=0.66). This SNP is located 78 kb centromeric of the index variant and is not correlated with the index variant ($r^2<0.01$ in CEU and $r^2=0.027$ in YRI). No statistically significant associations were observed with variants found previously in association with cancers of the bladder and ovary, or leukemia (rs9642880: OR=1.03, P=0.58; rs10088218: OR=1.02, P=0.62; rs2456449: OR=1.07, P=0.14). Of the known risk variants for prostate cancer we found a single nominally significant ($P<0.05$) association with the same risk allele of rs1016343 (P=0.015) which is located >260 kb centromeric of the breast cancer risk region and is not correlated with rs13281615 or rs16902056.

10q22

We observed no association with the index signal at 10q22 (rs704010) which is located in intron 1 of the gene *ZMIZ1*, or with any correlated markers. However, we did detect strong evidence of a second signal located 215 kb telomeric in intron 12 of the gene *ZMIZ1* (rs12355688: OR=1.24, $P=6.8\times10^{-6}$). This putative novel risk variant is not correlated with the index variant in the CEU or YRI populations ($r^2<0.01$).

11q13

No positive association was noted with the index variant at 11q13. However, we did detect evidence of a second independent signal (rs609275: OR=1.20, $P=1.0\times10^{-5}$), located 74 kb telomeric, and 53 kb centromeric of *CCND1.* The variant is monomorphic and uncorrelated with the index signal in the CEU population; and r^2 with the index signal in the YRI population is <0.01.

16q12

As in previous studies of African Americans we were not able to replicate the association signal defined by the index variant rs3803662.[23,24] A recent study of African Americans reported a suggestive association with SNP rs3104746, which is located 15 kb telomeric of rs3803662.[25] This SNP has a minor allele frequency of 0.04 in the HapMap CEU population, 0.19 in our African American controls, and is modestly correlated with rs3803662 in Africans ($r^2=0.31$ in YRI), but not in Europeans ($r^2=0.038$). Fine-mapping around this putative signal revealed a perfect proxy ($r^2=1$) for rs3104746, rs3112572, which is significantly associated with breast cancer risk in African Americans (OR=1.18, $P=3.9\times10^{-4}$) with the association noted to be stronger for ER+ breast cancer (OR=1.27, $P=3.1\times10^{-5}$).

For index SNPs found to be nominally associated with breast cancer risk, as well as risk-associated markers identified through fine-mapping, we also tested for associations by genotype. Results from the genotype-specific model were consistent with log-additive-associations. Risk variants at 2q35 and 8q24 were also found to have significantly stronger associations with ER+ breast cancer than ER- disease which is consistent with previous studies.[19]

We observed no statistically significant associations with common variation at 10 risk loci on 1p11, 3p24, 5p12, 6q25, 9p21, 10p15, 10q21, 11p15, 14q24 and 17q23.

Risk modeling

In this study we also estimated the cumulative effect of all breast cancer risk variants, and compared a summary risk score comprised of unweighted counts of all GWAS reported risk variants to a risk score that included variants we identified as being associated with risk in African Americans. Using the 19 index signals from GWAS, the risk per allele was 1.04 (95% CI, 1.02-1.06; $P=6.1\times10^{-5}$) and individuals in the top quintile of the risk allele distribution were at 1.4-fold greater risk ($P=7.4\times10^{-5}$) of breast cancer compared to those in the lowest quartile. As expected, the risk score was improved when utilizing the markers that we identified at the known risk loci as being more relevant to African Americans (8 alleles for overall breast cancer: 2q35, 5q11, 9q31, 10q22, 10q26, 11q13, 16q12 and 19p13; OR=1.18; 95% CI, 1.14-1.22; $P=2.8\times10^{-24}$), with risk for those in the top quartile being 2.2-times that observed in the lowest quintile ($P=3.6\times10^{-17}$). We observed an increase of 1.9 percentage points in the area under the curve (AUC) ($P=2.6\times10^{-6}$). This score was significantly associated with risk of both ER+ (OR=1.20, $P=1.7\times10^{-19}$) and ER- (OR=1.15, $P=2.8\times10^{-9}$) disease ($P_{het}=0.12$).

Future Work to Better Address the Topic: Additional Ongoing Efforts to Reveal Loci for Breast Cancer in Women of African Ancestry.

We are currently conducting additional meta-analyses and follow-up genotyping with new studies of breast cancer in African ancestry populations. In October of 2012, we will be meta-analyzing GWAS results from our AABC GWAS with a GWAS of breast cancer in Nigerian women (>1,000 cases and >1,000 controls). The 50,000 most significant associations from the meta-analysis will be included on a custom iSelect array to be genotyped by the AMBER breast cancer consortium (>3,000 cases and >3,000 controls). We expect findings from this work to reveal additional loci for overall breast cancer and ER-negative disease that are important for women of African ancestry. The custom array will also include SNP content (~80,000 SNPs), for fine-mapping of the ~80 known breast cancer risk loci in this population.

Key Research Accomplishments

- Established a consortia to study breast cancer among women of African ancestry
- Conducted the first genome-wide association study of breast cancer among African American women
- Ruled out common genetic variants with large effects as contributors to breast cancer risk in women of African ancestry
- Pulled together all existing GWAS of ER-negative breast cancer for meta-analysis
- Identified three susceptibility loci for breast cancer with two being specific for ER-negative breast cancer
- Identified a locus for ER-negative breast cancer that contributes to greater risk of ER-negative disease and triple negative disease in women of African ancestry
- Via fine-mapping we improved upon markers of risk at known susceptibility loci that better characterize their contribution to breast cancer risk in women of African ancestry

Reportable Outcomes and Studies that have Emanated from the GWAS of Breast Cancer in Women of African Ancestry.

Manuscripts (provided in Appendix):

A genome-wide association study of breast cancer in women of African ancestry.
Chen F, Chen GK, Stram DO, Millikan RC, Ambrosone CB, John EM, Bernstein L, Zheng W, Palmer JR, Hu JJ, Rebbeck TR, Ziegler RG, Nyante S, Bandera EV, Ingles SA, Press MF, Ruiz-Narvaez EA, Deming SL, Rodriguez-Gil JL, Demichele A, Chanock SJ, Blot W, Signorello L, Cai Q, Li G, Long J, Huo D, Zheng Y, Cox NJ, Olopade OI, Ogundiran TO, Adebamowo C, Nathanson KL, Domchek SM, Simon MS, Hennis A, Nemesure B, Wu SY, Leske MC, Ambs S, Hutter CM, Young A, Kooperberg C, Peters U, Rhie SK, Wan P, Sheng X, Pooler LC, Van Den Berg DJ, Le Marchand L, Kolonel LN, Henderson BE, **Haiman CA**.
Human Genetics 2012 Aug 25

A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer.
Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, Millikan RC, Wang X, Ademuyiwa F, Ahmed S, Ambrosone CB, Baglietto L, Balleine R, Bandera EV, Beckmann MW, Berg CD, Bernstein L, Blomqvist C, Blot WJ, Brauch H, Buring JE, Carey LA, Carpenter JE, Chang-Claude J, Chanock SJ, Chasman DI, Clarke CL, Cox A, Cross SS, Deming SL, Diasio RB, Dimopoulos AM, Driver WR, Dünnebier T, Durcan L, Eccles D, Edlund CK, Ekici AB, Fasching PA, Feigelson HS, Flesch-Janys D, Fostira F, Försti A, Fountzilas G, Gerty SM; Gene Environment Interaction and Breast Cancer in Germany (GENICA) Consortium, Giles GG, Godwin AK, Goodfellow P, Graham N, Greco D, Hamann U, Hankinson SE, Hartmann A, Hein R, Heinz J, Holbrook A, Hoover RN, Hu JJ, Hunter DJ, Ingles SA, Irwanto A, Ivanovich J, John EM, Johnson N, Jukkola-Vuorinen A, Kaaks R, Ko YD, Kolonel LN, Konstantopoulou I, Kosma VM, Kulkarni S, Lambrechts D, Lee AM, Marchand LL, Lesnick T, Liu J, Lindstrom S, Mannermaa A, Margolin S, Martin NG, Miron P, Montgomery GW, Nevanlinna H, Nickels S, Nyante S, Olswold C, Palmer J, Pathak H, Pectasides D, Perou CM, Peto J, Pharoah PD, Pooler LC, Press MF, Pylkäs K, Rebbeck TR, Rodriguez-Gil JL, Rosenberg L, Ross E, Rüdiger T, Silva Idos S, Sawyer E, Schmidt MK, Schulz-Wendtland R, Schumacher F, Severi G, Sheng X, Signorello LB, Sinn HP, Stevens KN, Southey MC, Tapper WJ, Tomlinson I, Hogervorst FB, Wauters E, Weaver J, Wildiers H, Winqvist R, Van Den Berg D, Wan P, Xia LY, Yannoukakos D, Zheng W, Ziegler RG, Siddiq A, Slager SL, Stram DO, Easton D, Kraft P, Henderson BE, Couch FJ.
Nature Genetics 2011 Oct 30;43(12):1210-4.

A meta-analysis of estrogen receptor negative breast cancer GWAS identifies two novel susceptibility loci at 6q14 and 20q11
Siddiq A, Couch FJ, Chen GK, Garcia-Closas M, …+.>100 co-authors…Ziv E, Easton DF, Nevanlinna H, Hunter DJ, Chanock SJ, Kraft P, **Haiman CA**, Vachon CM
Human Molecular Genetics (in press)

Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans.
Chen F, Chen GK, Millikan RC, John EM, Ambrosone CB, Bernstein L, Zheng W, Hu JJ, Ziegler RG, Deming SL, Bandera EV, Nyante S, Palmer JR, Rebbeck TR, Ingles SA, Press MF, Rodriguez-Gil JL, Chanock SJ, Le Marchand L, Kolonel LN, Henderson BE, Stram DO, **Haiman CA**.

Human Molecular Genetics 2011 Nov 15;20(22):4491-503

Abstracts/Posters:

"Towards Understanding Breast Cancer Susceptibility in Women of African Ancestry", Department of Defense Congressionally Directed Breast Cancer Research Program, Era of Hope, 2011.

"Genome-wide association studies identify novel ER-negative specific breast cancer risk loci", American Society of Human Genetics, 2012.

Presentations:

"A Genome-wide Association Study of Breast Cancer in African American women" American Association for Cancer Research, Denver, 2009.

"The Life and Times of Genome-wide Scans", American Statistical Association Conference of Radiation and Health, Annapolis, MD, 2010.

"Towards Understanding Breast Cancer Susceptibility in Women of African Ancestry", Department of Defense Congressionally Directed Breast Cancer Research Program, Era of Hope Conference, Orlando, FL, 2011.

"Genetic Studies of Cancer in Multiethnic Populations" Moffitt Cancer Center, Tampa Bay, FL, 2012.

Funding applied for based on work supported by this award:

Title: Epidemiology of Breast Cancer Subtypes in African American Women: A Consortium
Agency: NIH/NCI P01 (PI, C. Ambrosone, Roswell Park Cancer Center)

Title: Whole-genome sequencing of breast cancer in African American women
Agency: NIH/NCI R01 (Co-Principal Investigator)

In addition to the findings above, the GWAS data generated as part of this project have been utilized in a number of additional genetic studies in African ancestry populations. ***These projects took place in years 3 and 4 of the study period and some are still ongoing.***

Manuscripts:

Genome-wide meta-analyses of smoking behaviors in African Americans.
David SP, Hamidovic A, Chen GK, Bergen AW, Wessel J, Kasberger JL, Brown WM, Petruzella S, Thacker EL, Kim Y, Nalls MA, Tranah GJ, Sung YJ, Ambrosone CB, Arnett D, Bandera EV, Becker DM, Becker L, Berndt SI, Bernstein L, Blot WJ, Broeckel U, Buxbaum SG, Caporaso N, Casey G, Chanock SJ, Deming SL, Diver WR, Eaton CB, Evans DS, Evans MK, Fornage M, Franceschini N, Harris TB, Henderson BE, Hernandez DG, Hitsman B, Hu JJ, Hunt SC, Ingles SA, John EM, Kittles R, Kolb S, Kolonel LN, Le Marchand L, Liu Y, Lohman KK, McKnight B,

Millikan RC, Murphy A, Neslund-Dudas C, Nyante S, Press M, Psaty BM, Rao DC, Redline S, Rodriguez-Gil JL, Rybicki BA, Signorello LB, Singleton AB, Smoller J, Snively B, Spring B, Stanford JL, Strom SS, Swan GE, Taylor KD, Thun MJ, Wilson AF, Witte JS, Yamamura Y, Yanek LR, Yu K, Zheng W, Ziegler RG, Zonderman AB, Jorgenson E, **Haiman CA**, Furberg H.
Translational Psychiatry. 2012 May 22;2:e119.

Identification, replication, and fine-mapping of Loci associated with adult height in individuals of african ancestry.
N'Diaye A, Chen GK, Palmer CD, Ge B, Tayo B, Mathias RA, Ding J, Nalls MA, Adeyemo A, Adoue V, Ambrosone CB, Atwood L, Bandera EV, Becker LC, Berndt SI, Bernstein L, Blot WJ, Boerwinkle E, Britton A, Casey G, Chanock SJ, Demerath E, Deming SL, Diver WR, Fox C, Harris TB, Hernandez DG, Hu JJ, Ingles SA, John EM, Johnson C, Keating B, Kittles RA, Kolonel LN, Kritchevsky SB, Le Marchand L, Lohman K, Liu J, Millikan RC, Murphy A, Musani S, Neslund-Dudas C, North KE, Nyante S, Ogunniyi A, Ostrander EA, Papanicolaou G, Patel S, Pettaway CA, Press MF, Redline S, Rodriguez-Gil JL, Rotimi C, Rybicki BA, Salako B, Schreiner PJ, Signorello LB, Singleton AB, Stanford JL, Stram AH, Stram DO, Strom SS, Suktitipat B, Thun MJ, Witte JS, Yanek LR, Ziegler RG, Zheng W, Zhu X, Zmuda JM, Zonderman AB, Evans MK, Liu Y, Becker DM, Cooper RS, Pastinen T, Henderson BE, Hirschhorn JN, Lettre G, **Haiman CA**.
PLoS Genetics 2011 Oct;7(10):e1002298.

Detectable clonal mosaicism from birth to old age and its relationship to cancer.
Laurie CC, Laurie CA, Rice K, Doheny KF, Zelnick LR, McHugh CP, Ling H, Hetrick KN, Pugh EW, Amos C, Wei Q, Wang LE, Lee JE, Barnes KC, Hansel NN, Mathias R, Daley D, Beaty TH, Scott AF, Ruczinski I, Scharpf RB, Bierut LJ, Hartz SM, Landi MT, Freedman ND, Goldin LR, Ginsburg D, Li J, Desch KC, Strom SS, Blot WJ, Signorello LB, Ingles SA, Chanock SJ, Berndt SI, Le Marchand L, Henderson BE, Monroe KR, Heit JA, de Andrade M, Armasu SM, Regnier C, Lowe WL, Hayes MG, Marazita ML, Feingold E, Murray JC, Melbye M, Feenstra B, Kang JH, Wiggs JL, Jarvik GP, McDavid AN, Seshan VE, Mirel DB, Crenshaw A, Sharopova N, Wise A, Shen J, Crosslin DR, Levine DM, Zheng X, Udren JI, Bennett S, Nelson SC, Gogarten SM, Conomos MP, Heagerty P, Manolio T, Pasquale LR, **Haiman CA**, Caporaso N, Weir BS.
Nature Genetics 2012 May 6;44(6):642-50.

The landscape of recombination in African Americans.
Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL, Aldrich MC, Ambrosone CB, Amos C, Bandera EV, Berndt SI, Bernstein L, Blot WJ, Bock CH, Boerwinkle E, Cai Q, Caporaso N, Casey G, Cupples LA, Deming SL, Diver WR, Divers J, Fornage M, Gillanders EM, Glessner J, Harris CC, Hu JJ, Ingles SA, Isaacs W, John EM, Kao WH, Keating B, Kittles RA, Kolonel LN, Larkin E, Le Marchand L, McNeill LH, Millikan RC, Murphy A, Musani S, Neslund-Dudas C, Nyante S, Papanicolaou GJ, Press MF, Psaty BM, Reiner AP, Rich SS, Rodriguez-Gil JL, Rotter JI, Rybicki BA, Schwartz AG, Signorello LB, Spitz M, Strom SS, Thun MJ, Tucker MA, Wang Z, Wiencke JK, Witte JS, Wrensch M, Wu X, Yamamura Y, Zanetti KA, Zheng W, Ziegler RG, Zhu X, Redline S, Hirschhorn JN, Henderson BE, Taylor HA Jr, Price AL, Hakonarson H, Chanock SJ, **Haiman CA**, Wilson JG, Reich D, Myers SR.
Nature. 2011 Jul 20;476(7359):170-5.

Three Novel Loci Identified with Body Mass Index in a Genome-wide Association Study of 60,306 Men and Women of African Ancestry
Monda KL, Chen GK, Taylor KC, >200 co-authors, Loos R, North KE, **Haiman CA**
Nature Genetics (under review)

Abstracts/Posters

“Genome-wide Association Study of Body Mass Index in 33,525 African Americans Validates Four Loci Previously-Identified in Europeans and Reveals Two Novel Loci”, American Heart Association, 2011.

“Genome-wide Association Study of Body Mass Index in 40,016 African Americans Reveals Two Novel Loci and Validates Six Loci Previously-Identified in Europeans”, The Obesity Society, 2011.

“Identification, replication, and fine-mapping of loci associated with adult height in individuals of African ancestry”, American Society of Human Genetics, 2011.

“Heritability estimation for human height using GWAS-based studies in African-Americans”, American Society of Human Genetics, 2011.

“Three Novel Loci Identified with BMI in a Genome-wide Association Study of 47,098 Men and Women of African ancestry”, American Society of Human Genetics, 2012.

Conclusion

Genome-wide studies of common and rare genetic variation conducted in multiple populations will be required to reveal the complete spectrum of susceptibility alleles that contribute to risk of breast cancer globally. In a genome-wide scan of common genetic variation in >3,000 African American cases and >2,700 controls, followed by replication testing of the most significant associations ($p<10^{-4}$) in an independent set of >3,000 cases and >3,000 controls, we identified two suggestive associations with breast cancer risk that replicated in stage 2 at $p<0.05$ [chromosome 14q31 ($p = 4.3\times10^{-6}$) and 3p26 ($p = 1.5\times10^{-5}$)]; however, these associations did not reach the standard level of genome-wide significance. These regions have not been highlighted in previous GWAS conducted in other racial/ethnic populations and each association requires further validation in additional studies. A strength of the 2-stage GWAS we conducted is that it includes most existing case-control studies of breast cancer conducted in women of African ancestry. In this 2-stage design, we had 80% statistical power to identify a common risk variant (frequency of ≥ 10%) that conveys a risk per allele of 1.3 at genome-wide significance (p=5×10-8). Thus, we were able to rule out variants with large effects if they were among the top 0.007% in stage 1 (and thus taken to stage 2) and were adequately tagged by the common SNPs on the 1M array. However, we are likely to have missed some milder associations. In previous GWAS of breast cancer in European ancestry populations, most risk variants eventually identified were not among the most statistically significant in stage 1 and were only revealed through testing of large numbers of SNPs in additional replication stages. To identify novel risk loci for overall breast cancer in African ancestry populations will require continued collaborative efforts and investigators willing to test larger numbers of SNPs in their respective studies.

In our meta-analyses of GWAS for ER-negative breast cancer we identified three novel loci for breast cancer with two being specific for ER-negative disease. SNP rs17530068 at chromosome 6q14 was associated with overall breast cancer risk and showed no differential association depending on ER status. The association of SNP rs2284378 at 20q11, however, was stronger for ER-negative than ER-positive breast cancer. SNP rs10069690 at 5p15 also appeared to be more associated with ER-negative and triple negative disease. Identification of the variants directly responsible for the association will be required to fully address the extent to which these loci contribute to the greater incidence of ER-negative and triple negative tumors in women of African ancestry. However, it is notable that the risk allele frequency of rs10069690 is greater in African American women (frequency, 0.57) than in women of European ancestry (frequency, 0.26). If this variant is an equally good surrogate for the biologically functional allele in each population, then this locus may be responsible for a 15% (95% CI, 10-20%) increase in the incidence rate of ER negative or triple negative breast cancer in women of African compared to European ancestry. Larger studies with well-characterized tumor pathology information will be needed to determine if the associations we observed applies other breast cancer subtypes. Furthermore, our findings provide further support for the presence of genetic susceptibility to ER-negative breast cancer subtypes.

"So What?"

Identifying new loci associated with ER-negative and triple negative breast cancer will continue to provide insight into the biological mechanisms underlying this more aggressive form of breast cancer, and could result in improvements in risk prediction and treatment.

References

1. Chen, F. *et al.* A genome-wide association study of breast cancer in women of African ancestry. *Hum Genet* (2012).
2. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-59 (2000).
3. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904-9 (2006).
4. Haiman, C.A. *et al.* A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat Genet* **43**, 1210-4 (2011).
5. Zheng, W. *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet* **41**, 324-8 (2009).
6. Stevens, K.N. *et al.* 19p13.1 is a triple-negative-specific breast cancer susceptibility locus. *Cancer Res* **72**, 1795-803 (2012).
7. Antoniou, A.C. *et al.* A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat Genet* **42**, 885-92 (2010).
8. Maccioni, L. *et al.* Variants at chromosome 20 (ASIP locus) and melanoma risk. *Int J Cancer* (2012).
9. Lin, W. *et al.* ASIP genetic variants and the number of non-melanoma skin cancers. *Cancer Causes Control* **22**, 495-501 (2011).
10. Nan, H., Kraft, P., Hunter, D.J. & Han, J. Genetic variants in pigmentation genes, pigmentary phenotypes, and risk of skin cancer in Caucasians. *Int J Cancer* **125**, 909-17 (2009).
11. Ghoussaini, M. *et al.* Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat Genet* **44**, 312-8 (2012).
12. Chen, F. *et al.* Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans. *Hum Mol Genet* **20**, 4491-503 (2011).
13. Easton DF, P.K., Dunning AM, Pharoah PDP, Thompson D, Ballinger DG, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature advance online publication* **doi:10.1038/nature05887**(2007).
14. Hunter DJ, K.P., Jacobs KB, Cox DG, Yeager M, Hankinson, SE, et al. . A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer *Nature Genetics advance online publication* **doi:10.1038/ng2064**(2007).
15. Stacey, S.N. *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* **39**, 865-9 (2007).
16. Stacey, S.N. *et al.* Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* **40**, 703-6 (2008).
17. Thomas, G. *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* **41**, 579-84 (2009).
18. Udler, M.S. *et al.* FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Hum Mol Genet* **18**, 1692-703 (2009).
19. Garcia-Closas, M. *et al.* Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS Genet* **4**, e1000054 (2008).
20. Antoniou, A.C. *et al.* A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat Genet* **42**, 885-92.

21. Haiman, C.A. *et al.* A common genetic risk factor for colorectal and prostate cancer. *Nat Genet* **39**, 954-6 (2007).
22. Haiman, C.A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat Genet* **39**, 638-44 (2007).
23. Udler, M.S. *et al.* Fine scale mapping of the breast cancer 16q12 locus. *Hum Mol Genet* **19**, 2507-15.
24. Zheng, W. *et al.* Evaluation of 11 breast cancer susceptibility loci in African-American women. *Cancer Epidemiol Biomarkers Prev* **18**, 2761-4 (2009).
25. Ruiz-Narvaez, E.A. *et al.* Polymorphisms in the TOX3/LOC643714 locus and risk of breast cancer in African-American women. *Cancer Epidemiology Biomarkers & Prevention* **19**, 1320-7 (2010).

Appendices
See attached.

Hum Genet
DOI 10.1007/s00439-012-1214-y

ORIGINAL INVESTIGATION

A genome-wide association study of breast cancer in women of African ancestry

Fang Chen · Gary K. Chen · Daniel O. Stram · Robert C. Millikan · Christine B. Ambrosone · Esther M. John · Leslie Bernstein · Wei Zheng · Julie R. Palmer · Jennifer J. Hu · Tim R. Rebbeck · Regina G. Ziegler · Sarah Nyante · Elisa V. Bandera · Sue A. Ingles · Michael F. Press · Edward A. Ruiz-Narvaez · Sandra L. Deming · Jorge L. Rodriguez-Gil · Angela DeMichele · Stephen J. Chanock · William Blot · Lisa Signorello · Qiuyin Cai · Guoliang Li · Jirong Long · Dezheng Huo · Yonglan Zheng · Nancy J. Cox · Olufunmilayo I. Olopade · Temidayo O. Ogundiran · Clement Adebamowo · Katherine L. Nathanson · Susan M. Domchek · Michael S. Simon · Anselm Hennis · Barbara Nemesure · Suh-Yuh Wu · M. Cristina Leske · Stefan Ambs · Carolyn M. Hutter · Alicia Young · Charles Kooperberg · Ulrike Peters · Suhn K. Rhie · Peggy Wan · Xin Sheng · Loreall C. Pooler · David J. Van Den Berg · Loic Le Marchand · Laurence N. Kolonel · Brian E. Henderson · Christopher A. Haiman

Received: 23 April 2012 / Accepted: 31 July 2012

Abstract Genome-wide association studies (GWAS) in diverse populations are needed to reveal variants that are more common and/or limited to defined populations. We conducted a GWAS of breast cancer in women of African ancestry, with genotyping of >1,000,000 SNPs in 3,153 African American cases and 2,831 controls, and replication testing of the top 66 associations in an additional 3,607 breast cancer cases and 11,330 controls of African ancestry. Two of the 66 SNPs replicated ($p < 0.05$) in stage 2, which reached statistical significance levels of 10^{-6} and 10^{-5} in the stage 1 and 2 combined analysis (rs4322600 at chromosome 14q31: OR = 1.18, $p = 4.3 \times 10^{-6}$; rs10510 333 at chromosome 3p26: OR = 1.15, $p = 1.5 \times 10^{-5}$). These suggestive risk loci have not been identified in

F. Chen and G. K. Chen contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-012-1214-y) contains supplementary material, which is available to authorized users.

F. Chen · G. K. Chen · D. O. Stram · S. A. Ingles · S. K. Rhie · P. Wan · X. Sheng · L. C. Pooler · D. J. Van Den Berg · B. E. Henderson · C. A. Haiman
Department of Preventive Medicine, Keck School of Medicine, Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA
e-mail: fangchen@usc.edu

R. C. Millikan · S. Nyante
Department of Epidemiology, Gillings School of Global Public Health, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA

C. B. Ambrosone
Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY, USA

E. M. John
Cancer Prevention Institute of California, Fremont, CA, USA

E. M. John
Stanford University School of Medicine, Stanford Cancer Institute, Stanford, CA, USA

L. Bernstein
Division of Cancer Etiology, Department of Population Science, Beckman Research Institute, City of Hope, Duarte, CA, USA

W. Zheng · S. L. Deming · W. Blot · L. Signorello · Q. Cai · G. Li · J. Long
Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA

J. R. Palmer · E. A. Ruiz-Narvaez
Slone Epidemiology Center at Boston University, Boston, MA, USA

J. J. Hu · J. L. Rodriguez-Gil
Department of Epidemiology and Public Health, Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL, USA

T. R. Rebbeck · A. DeMichele · K. L. Nathanson · S. M. Domchek
University of Pennsylvania School of Medicine, Philadelphia, PA, USA

Published online: 25 August 2012

previous GWAS in other populations and will need to be examined in additional samples. Identification of novel risk variants for breast cancer in women of African ancestry will demand testing of a substantially larger set of markers from stage 1 in a larger replication sample.

Introduction

Genome-wide association studies (GWAS) of breast cancer have been conducted almost exclusively in populations of European ancestry, and have firmly established associations with a number of common susceptibility loci that contribute modest effects (relative risks $\leq$1.3) (Ahmed et al. 2009; Antoniou et al. 2010; Easton et al. 2007;Fletcher et al. 2011; Ghoussaini et al. 2012; Haiman et al. 2011b; Hunter et al. 2007; Kim et al. 2012; Long et al. 2012; Stacey et al. 2007, 2008; Thomas et al. 2009; Turnbull et al. 2010; Zheng et al. 2009b). These discoveries provide support for the polygenic model of breast cancer susceptibility (Pharoah et al. 2002), as well as clues as to important biological pathways involved in the pathogenesis of breast cancer. For example, the most strongly associated risk locus for breast cancer revealed through GWAS has been the region containing the fibroblast growth factor receptor 2 (*FGFR2*) at chromosome 10q26 (Easton et al. 2007; Hunter et al. 2007; Meyer et al. 2008). *FGFR2* is a member of the *FGFR* family of receptor tyrosine kinases (RTKs) which regulate cell proliferation, differentiation and apoptosis (Tenhagen et al. 2012). The risk variant on chromosome 14q24 is located in intron 12 of *RAD51B* which is a member of the RAD51 protein family. RAD51 proteins are essential for DNA repair by homologous recombination (Tarsounas et al. 2004), a DNA repair pathway with an established and important role in breast cancer development. A more recent study, which included African American subjects from the current study, revealed a risk marker at the telomerase reverse transcriptase (*TERT*) locus (Haiman et al. 2011b), a protein that controls telomere length and is also implicated in oncogenesis (Kim et al. 1994). Many of the risk variants identified by GWAS, however, are located in gene deserts, or near genes with roles in breast cancer etiology that are currently unknown.

The search for additional low penetrance alleles for breast cancer in specific racial/ethnic populations has revealed additional variants that are important globally or more common and/or limited to defined populations. For example, a GWAS conducted among Chinese women identified a novel risk locus for breast cancer near the gene for the estrogen receptor (ER) on chromosome 6 which had not been revealed in previous, well-powered GWAS in

R. G. Ziegler · S. J. Chanock
Epidemiology and Biostatistics Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, DC, USA

E. V. Bandera
The Cancer Institute of New Jersey, New Brunswick, NJ, USA

M. F. Press
Department of Pathology, Keck School of Medicine and Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA

D. Huo
Department of Health Studies, University of Chicago, Chicago, IL, USA

Y. Zheng · N. J. Cox · O. I. Olopade
Department of Medicine, University of Chicago, Chicago, IL, USA

T. O. Ogundiran
Department of Surgery, College of Medicine, University of Ibadan, Ibadan, Nigeria

C. Adebamowo
Department of Epidemiology and Preventive Medicine, University of Maryland, Baltimore, MD, USA

M. S. Simon
Department of Oncology, Karmanos Cancer Institute, Wayne State University, Detroit, MI, USA

A. Hennis
Chronic Disease Research Centre, Tropical Medicine Research Institute, University of the West Indies, Bridgetown, Barbados

A. Hennis · B. Nemesure · S.-Y. Wu · M. C. Leske
Department of Preventive Medicine, State University of New York at Stony Brook, Stony Brook, NY, USA

S. Ambs
Laboratory of Human Carcinogenesis, National Cancer Institute, Bethesda, MD, USA

C. M. Hutter · A. Young · C. Kooperberg · U. Peters
Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

D. J. Van Den Berg
Epigenome Center, Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA

L. Le Marchand · L. N. Kolonel
Epidemiology Program, Cancer Research Center, University of Hawaii, Honolulu, HI, USA

C. A. Haiman (✉)
USC Norris Comprehensive Cancer Center, Harlyne Norris Research Tower, 1450 Biggy Street, Room 1504, Los Angeles, CA 90033, USA
e-mail: haiman@usc.edu

populations of European ancestry (Zheng et al. 2009b). A GWAS of prostate cancer in men of African ancestry also identified a novel risk variant at 17q12 that is not observed in other populations (Haiman et al. 2011a). In search for risk variants for breast cancer that may be important to women of African ancestry, we analyzed >1 million common SNPs in 3,153 African American breast cancer cases and 2,831 African American controls, and examined the most statistically significant associations in a second stage of 3,607 cases and 11,330 controls of African ancestry.

Materials and methods

Study populations

Stage 1 of the GWAS included 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). Replication testing was conducted in an independent sample of 3,607 breast cancer cases and 11,330 controls from 9 additional studies of breast cancer in women of African ancestry (The Black Women's Health Study (BWHS), 826/1,167; The Women's Insights and Shared Experiences study (WISE), 174/458; NBHS/Southern Community Cohort (SCCS), 981/851; The Nigerian Breast Cancer Study (NBCS), 681/282; The Barbados National Cancer Study (BNCS), 93/244; The Racial Variability in Genotypic Determinants of Breast Cancer Risk Study (RVGBC), 151/272; The Baltimore Breast Cancer Study (BBCS), 117/111; The Chicago Cancer Prone Study (CCPS), 268/261; and, The Women's Health Initiative (WHI), 316/7,484).

Sample size and selected characteristics for these studies are summarized in Supplemental Tables 1 and 2 and detailed information about the design and organization of each study is provided in supporting information.

Genotyping and quality control

Genotyping in stage 1 was conducted using the Illumina Human1M-Duo BeadChip. Of the 5,984 samples from these studies (3,153 cases and 2,831 controls), we attempted genotyping of 5,932, removing samples ($n = 52$) with DNA concentrations <20 ng/ul. Following genotyping, we removed samples based on the following exclusion criteria: (1) unexpected replicates (≥98.9 % genetically identical) that we were able to confirm through discussions with study investigators (only one of each replicate was removed, $n = 15$); (2) unknown replicates that we were not able to confirm (pair or triplicate removed, $n = 14$); (3) samples with call rates <95 % after a second genotyping attempt ($n = 100$); (4) samples with ≤5 % African ancestry ($n = 36$) (discussed below); and (5) samples with <15 % mean heterozygosity of SNPs on the X chromosome and/or similar mean allele intensities of SNPs on the X and Y chromosomes ($n = 6$) as these are likely to be males.

We removed SNPs with <95 % call rate ($n = 21,732$) or minor allele frequencies (MAFs) <1 % ($n = 80,193$). To assess genotyping reproducibility, we included 138 known replicate samples; the average concordance rate was 99.95 % (>99.93 % for all pairs). We also eliminated SNPs with genotyping concordance rates <98 % based on the replicates ($n = 11,701$). The final analysis dataset included 1,043,036 SNPs genotyped on 3,016 cases and 2,745 controls, with an average SNP call rate of 99.7 % and average sample call rate of 99.8 %. Hardy–Weinberg equilibrium (HWE) was not used as a criterion for removing SNPs; none of the SNPs selected for replication deviated from HWE in controls in each study (based on a cut-off of $p < 0.001$).

We selected 66 SNPs with p values $<2 \times 10^{-4}$ in stage 1 for evaluation in the second stage. These SNPs were selected from 53 regions following linkage disequilibrium (LD) pruning of correlated SNPs. Two of these SNPs were located near a previously validated breast cancer risk locus [rs12355688 at 10q22, 241 kb downstream of rs704010, $r^2 = 0$ in both CEU and YRI populations from 1000 Genomes Project (March 2010 release) (Turnbull et al. 2010); and rs3745185 at 19p13, 10 kb downstream of rs2363956, $r^2 = 0.57$ and 0.19 in the CEU and YRI populations from 1000 Genomes Project (March 2010 release), respectively (Antoniou et al. 2010)]. Genotyping in the replication studies was performed using the Sequenom platform (BWHS), OpenArray (WISE and NBHS/SCCS), the Affymetrix 6.0 SNP array (WHI) (Hutter et al. 2011) and Illumina GoldenGate (all other studies) (see Supporting Information). Blinded duplicate samples (5–10 %) were included in the replication studies and concordance of these samples was ≥98 % in all studies. The number of SNPs that were genotyped successfully in each stage 2 study ranged from 51 to 63. The average call rate for all SNPs in stage 2 was 98.8 % (range for call rates of a SNP within study 71.4–100 %). Call rates by SNP and study are shown in Supplemental Table 3.

Estimation of African ancestry

In stage 1, we utilized STRUCTURE (Pritchard et al. 2000) to infer percent African ancestry on an individual level. A total of 2,546 ancestry-informative SNPs from the Illumina array were selected based on low inter-marker correlation and ability to differentiate between samples of African and European descent. In evaluating the distribution of the fraction of African ancestry across the stage 1 populations, statistically significant differences (ANOVA $p < 10^{-16}$) were noted (Supplemental Figure 1). We also applied principal components analysis (PCA) (Price et al. 2006) to estimate axes of variation among the 5,761 individuals using the same 2,546 ancestry informative markers. The first eigenvector accounted for 10.1 % of the variation between subjects, and subsequent eigenvectors accounted for not more than 0.5 %. Using input genotypes from the HapMap populations, CEU (CEPH Utah), YRI (Yoruba), and JPT (Japanese), we determined that the first eigenvector clearly differentiates between Europeans (CEU) and West Africans (YRI) in the HapMap samples (Supplemental Fig. 2).

Statistical analysis

We examined the observed versus the expected distribution of the Chi-squared test statistics using a 1-degree of freedom (df) trend test, comparing genotype counts for each SNP in cases versus controls. All tests of statistical significance were two-sided. To improve coverage, we augmented the set of SNPs tested for association through imputation using MACH (Li and Abecasis 2006). Phased haplotypes from the 120 CEU and 120 YRI founders in HapMap Phase 2 were used to infer genotypes of all Phase 2 SNPs that were not available on the Illumina 1M Duo or did not pass our quality control (QC) criteria. Odds ratios (OR) and 95 % confidence intervals (CI) for each SNP were estimated using unconditional logistic regression, adjusting for age, the first eigenvector and study. The SFBCS and NC-BCFR studies were conducted in the same San Francisco Bay Area population and were combined in all analyses.

In the replication studies, ORs and 95 % CIs for each SNP were estimated using unconditional logistic regression, adjusting for age, region within the WHI and estimated genetic ancestry. Ancestry information was available for all stage 2 studies except WISE (Supporting Information). Overall testing of single SNP associations was conducted via meta-analyses of results from the stage 1 and stage 2 studies.

We also conducted combined GWAS and admixture-based statistical tests to assess the contribution of local ancestry on the SNP associations. For each subject in our analysis, we inferred local ancestry, which defines the proportion of European and African ancestry at each genotyped and imputed SNP. To infer local ancestry in our GWAS panel of 5,761 African American women, we applied the program HAPMIX (Price et al. 2009). HAPMIX builds a Hidden Markov Model (HMM) using phased haplotype data that are representative of the two source populations assumed to be ancestral to the admixed (study) data. In this case, we provided the same HapMap dataset that was used for imputation (i.e., 240 CEU + YRI founder haplotypes per chromosome) as input. HAPMIX reports posterior probabilities for each subject at each SNP of carrying 0, 1 and 2 copies of a European allele.

Combined GWAS and admixture-based statistical tests were conducted to make inferences about regions of the genome that explain not only case–control differences in disease risk based on SNP associations, but also risk differences based on local genetic ancestry. We utilized the MIXSCORE program (Pasaniuc et al. 2011) which takes as input results from a GWAS scan and an admixture scan (specifically HAPMIX output), and computes several statistics that incorporate allele frequency information from both sources of evidence. The SUM score is a 2-df Chi-squared test that simultaneously tests for association (i.e., a case–control difference in allele frequency) and admixture evidence (i.e., a deviation from the genome-wide proportion of European ancestry). The MIX score also tests for both evidence of admixture and association, but assumes the odds ratios for admixture and association are equal, which is potentially more powerful when this assumption is true since it is a 1-df test.

Results

The stage 1 analysis included 3,016 cases and 2,745 controls among African American women from 9 epidemiological studies of breast cancer. The age of the cases and controls in stage 1 ranged from 22 to 87 years with the median ages being 55 and 58 years, respectively (Supplemental Table 1). The analysis of the most statistically significant associations from stage 1 was conducted in 3,533 cases and 11,046 controls from an additional 9 studies. The age of the cases and controls in stage 2 ranged from 18 to 92 years with the median ages being 50 and 53 years, respectively (Supplemental Table 2).

We observed no evidence of inflation of the test statistic ($\lambda = 1.01$) for the 1,043,036 genotyped and 2,067,098 imputed SNPs analyzed in stage 1, and no excess of very small p values beyond what was expected (Fig. 1). We observed no SNP to be associated with disease status at a genome-wide level of significance ($p < 5 \times 10^{-8}$) in stage 1 (Fig. 2). The most statistically significant association was

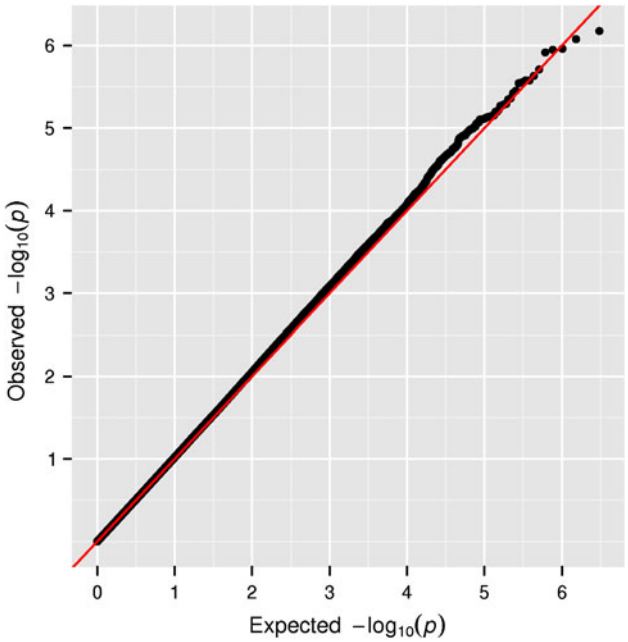

Fig. 1 The distribution of observed versus expected $-\log_{10} p$ values from stage 1 adjusted for age, study and the first principal component (PC1)

noted with SNP rs7610073 located in intron 2 of the gene *GRM7* (metabotropic glutamate receptor 7) on chromosome 3p26 (risk allele frequency 0.64; OR per allele 1.22; $p = 7.4 \times 10^{-7}$). A second signal was also noted ~486 kb upstream of *GRM7* (rs10510333: risk allele frequency 0.18; OR per allele 1.24; $p = 8.2 \times 10^{-6}$). The associations with these two markers were independent and remained statistically significant when both were included in the same model (p values of 8.3×10^{-7} and 9.3×10^{-6}, respectively). Shown in Table 1 are the genotyped SNPs with p values $<10^{-5}$ in stage 1, as well as SNPs that replicated in stage 2 (discussed below).

We selected 66 genotyped SNPs with association p values less than 2×10^{-4} for replication testing in the stage 2 studies. None of these SNPs replicated with stage 2-wide significance of <0.0008 (0.05/66), but two replicated with a p value <0.05 and an OR in the same direction as that observed in stage 1 (Table 1). Combining results from stages 1 and 2, no SNP achieved genome-wide significance. The smallest combined p values were noted for the two SNPs that replicated in stage 2: rs4322600 located ~100 kb upstream of the gene *GALC* (galactosylceramidase) on chromosome 14q31(risk allele frequency 0.78, OR per allele 1.18, $p = 4.3 \times 10^{-6}$) and rs10510333 located ~486 kb upstream of *GRM7* on chromosome 3p26 (risk allele frequency 0.18, OR per allele 1.15, $p = 1.5 \times 10^{-5}$) (Table 1). We found no strong statistical evidence that the associations with these two loci differ by ER status (p values for heterogeneity in case-only testing: rs10510333: $p = 0.67$; rs4322600: $p = 0.85$).

Using the MIXSCORE program, we simultaneously tested the null hypothesis of no association and admixture at each loci defined by the 66 most significant variants identified in Stage 1. SNP rs7610073, which had the largest

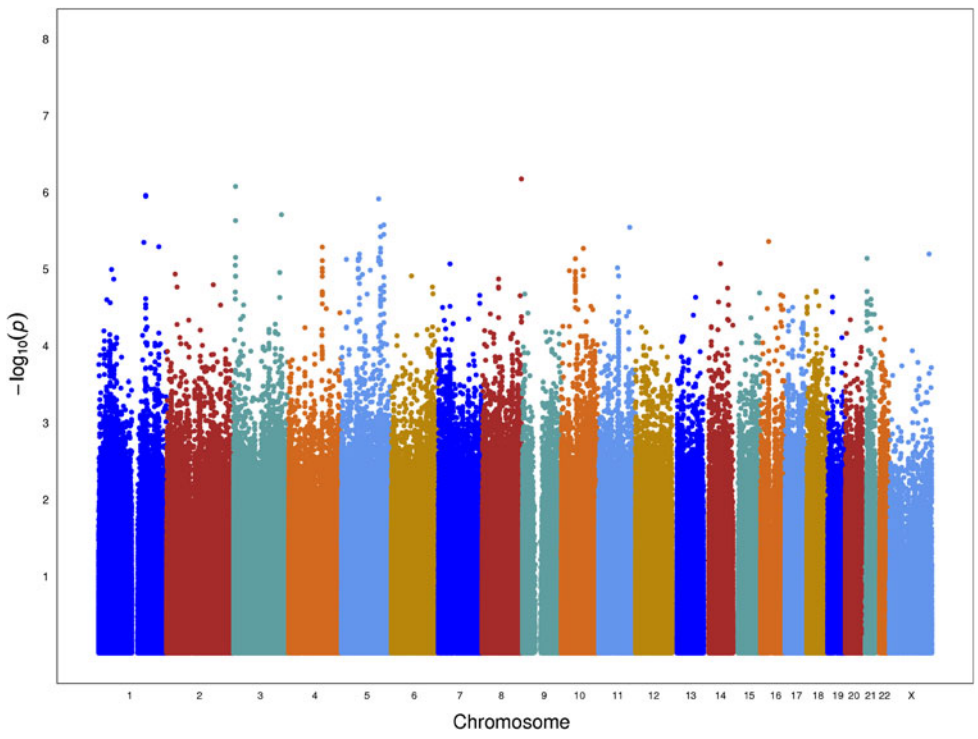

Fig. 2 A Manhattan plot showing the $-\log_{10} p$ values which test for case–control association to disease for genotyped and imputed SNPs by chromosome in stage 1

Table 1 SNPs with $p < 10^{-5}$ in stage 1 and SNPs that replicated at $p < 0.05$ in stage 2 of the African American breast cancer GWAS

SNP	Chr position[a]	Nearest genes	Risk allele	RAF[b]	Stage 1 3,016 cases, 2,745 controls		Stage 2 3,533 cases, 11,046 controls		Stage 1 + stage 2 6,549 cases, 13,791 controls	
					OR (95 % CI)	*p* value	OR (95 % CI)	*p* value	OR (95 % CI)	*p* value
rs7610073	3p26.1, 7275601	*GRM7*	A	0.64	1.22 (1.13–1.32)	7.4×10^{-7}	0.95 (0.89–1.12)	0.13	1.05 (1.00–1.11)	0.045
rs3861950	1q25.1, 171422915	*TNFSF4*	T	0.22	1.27 (1.15–1.39)	1.1×10^{-6}	0.98 (0.91–1.06)	0.63	1.09 (1.02–1.16)	6.4×10^{-3}
rs6880837	5q31.2, 135424568	*TGFBI*	C	0.69	1.23 (1.13–1.34)	1.8×10^{-6}	0.96 (0.89–1.03)	0.26	1.07 (1.01–1.13)	0.027
rs13074711	3q26.31, 173750497	*TNFSF10*	T	0.68	1.22 (1.13–1.33)	2.1×10^{-6}	1.01 (0.93–1.10)	0.78	1.11 (1.04–1.17)	5.7×10^{-4}
rs2085421	11q23.2, 113395918	*HTR3A,ZBTB16*	A	0.47	1.20 (1.11–1.29)	3.4×10^{-6}	0.99 (0.92–1.07)	0.86	1.09 (1.03–1.15)	1.7×10^{-3}
rs153170	5q31.3, 142257435	*ARHGAP26*	A	0.71	1.22 (1.12–1.33)	4.7×10^{-6}	1.00 (0.93–1.07)	0.94	1.09 (1.03–1.15)	4.0×10^{-3}
rs13172733	5q31.3, 142351873	*ARHGAP26*	A	0.69	1.22 (1.12–1.33)	4.8×10^{-6}	1.00 (0.91–1.08)	0.92	1.10 (1.04–1.17)	1.7×10^{-3}
rs12355688	10q22.3, 80725632	*ZMIZ1*	T	0.20	1.24 (1.13–1.36)	6.5×10^{-6}	1.03 (0.95–1.12)	0.43	1.12 (1.05–1.19)	3.8×10^{-4}
rs10510333	3p26.1, 6391779	*GRM7*	T	0.18	1.24 (1.13–1.36)	8.2×10^{-6}	1.08 (1.00–1.17)	0.048	1.15 (1.08–1.22)	1.5×10^{-5}
rs7727166	5q12.3, 65676926	*SFRS12*	T	0.16	1.25 (1.13–1.38)	9.4×10^{-6}	1.07 (0.98–1.17)	0.11	1.15 (1.07–1.22)	5.3×10^{-5}
rs4322600	14q31.3, 87365353	*GALC*	G	0.78	1.22 (1.11–1.34)	2.5×10^{-5}	1.12 (1.01–1.25)	0.036	1.18 (1.10–1.27)	4.3×10^{-6}

[a] NCBI build 36

[b] Risk allele frequency (RAF) in stage 1

[c] The stage 2 studies contributing to the results for each SNP are shown in Supplemental Table 3

MIX score of 24.5 ($p = 7.5 \times 10^{-7}$) also had the smallest *p* value in the first stage (Supplemental Table 4). The risk allele (the "A" allele for rs7610073) was not strongly differentiated (60 % in HapMap YRI vs. 81 % in HapMap CEU) and the MIX score *p* value was almost identical to the *p* value from our association scan. Association *p* values were generally stronger than the SUM or MIX score, so admixture did not make a substantive contribution in joint evidence of admixture and association for these 66 SNPs, as indicated in Supplemental Table 4. All together, these findings seem to indicate that the associations at the most significant loci in Stage 1 are not influenced by differences in local ancestry between cases and controls, meaning that any causal variants in these regions are not appreciably differentiated in frequency between cases and controls.

Discussion

Genome-wide studies of common and rare genetic variation conducted in multiple populations will be required to reveal the complete spectrum of susceptibility alleles that contribute to risk of breast cancer globally. In a genome-wide scan of common genetic variation in >3,000 African American cases and >2,700 controls, followed by replication testing of the most significant associations ($p < 2 \times 10^{-4}$) in an independent set of >3,500 cases and >11,000 controls, we identified two suggestive associations with breast cancer risk that replicated in stage 2 at $p < 0.05$ [chromosome 14q31 ($p = 4.3 \times 10^{-6}$) and 3p26 ($p = 1.5 \times 10^{-5}$)]; however, these associations did not reach the standard level of genome-wide significance. These regions have not been highlighted in previous GWAS conducted in other racial/ethnic populations and each association requires further validation in additional studies.

Populations of African ancestry have greater genetic diversity and lower levels of LD among chromosomal loci (Campbell and Tishkoff 2008; Reed and Tishkoff 2006). Because of LD patterns and allele frequencies that differ from non-African populations, GWAS results from European or Asian populations are not always replicable in populations of African ancestry (Chen et al. 2010; Huo et al. 2012; Hutter et al. 2011; Ruiz-Narvaez et al. 2010; Zheng et al. 2009a). Fine mapping of known breast cancer risk loci in populations of African ancestry has revealed risk-associated markers that are more relevant to African populations and contribute to modeling of genetic risk in this population (Chen et al. 2011; Ruiz-Narvaez et al. 2010; Udler et al. 2009). Large GWAS in populations of African ancestry, with proper control of population structure, will be required to discover additional disease susceptibility variants that better define the genetic profile of breast cancer in this population.

A strength of the present study is that it includes most existing case–control studies of breast cancer conducted in women of African ancestry. In this two-stage design, we had 80 % statistical power to identify a common risk variant (frequency of ≥10 %) that conveys a risk per allele of 1.3 at genome-wide significance ($p = 5 \times 10^{-8}$). Thus, we were able to rule out variants with large effects if they were among the top 0.007 % in stage 1 (and thus taken to stage 2) and were adequately tagged by the common SNPs on the 1 M array. However, we are likely to have missed some milder associations. In previous GWAS of breast cancer in European ancestry populations, most risk variants eventually identified were not among the most statistically significant in stage 1 and were only revealed through testing of large numbers of SNPs in additional replication stages. To identify novel risk loci for breast cancer in African ancestry populations will require continued collaborative efforts and investigators willing to test larger numbers of SNPs in their respective studies.

Our attempt to apply joint admixture and association mapping, using MIXSCORE, did not provide additional suggestive risk variants beyond those found using association methods alone. This suggests that the associations observed at the most significant regions in Stage 1 are not weakened by ancestry differences between cases and controls, and thus, the biologically functional alleles are unlikely to be highly differentiated in frequency between cases and controls. Because of the limited number of ER-negative cases in stage 1 ($n = 988$) and stage 2 ($n = 423$), the statistical power to look at subtypes with rate differences (e.g., ER-negative disease, more common in African American than European American women) was limited and not attempted for GWAS or admixture testing. However, in collaboration with GWAS of ER-negative breast cancer in European ancestry populations, which have substantially larger numbers of ER-negative cases, we have identified a novel locus for ER-negative breast cancer at 5p15 (*TERT*) (Haiman et al. 2011b). Genetic variation at this locus may contribute in part to the higher incidence of ER-negative disease subtypes in women of African ancestry (frequency of 0.56 in African Americans and frequency of 0.26 in Whites) (Haiman et al. 2011b). As for the analysis of overall breast cancer, larger studies of breast cancer in women of African ancestry will be needed to search for novel risk loci for ER-negative disease subtypes that are important for and may be limited to this population.

This study is the first genome-wide investigation of common genetic variation in relationship with breast cancer risk in women of African ancestry. The suggestive associations noted with risk variants at 14q31 and 3p26 require further validation in additional samples of African ancestry as well as in other populations. Identification of common risk variants for breast cancer in African ancestry populations will require testing a larger number of the most statistically significant SNPs from stage 1 in additional samples.

Acknowledgments This work was supported by a Department of Defense Breast Cancer Research Program Era of Hope Scholar Award to CAH [W81XWH-08-1-0383] and the Norris Foundation. Each of the participating studies was supported by the following grants: MEC (National Institutes of Health grants R01-CA63464 and R37-CA54281); CARE (National Institute for Child Health and Development grant NO1-HD-3-3175); WCHS (U.S. Army Medical Research and Material Command (USAMRMC) grant DAMD-17-01-0-0334, the National Institutes of Health grant R01-CA100598, and the Breast Cancer Research Foundation); SFBCS (National Institutes of Health grant R01-CA77305 and United States Army Medical Research Program grant DAMD17-96-6071); NC-BCFR (National Institutes of Health grant U01-CA69417); CBCS (National Institutes of Health Specialized Program of Research Excellence in Breast Cancer, grant number P50-CA58223, and Center for Environmental Health and Susceptibility National Institute of Environmental Health Sciences, National Institutes of Health, grant number P30-ES10126); PLCO (Intramural Research Program, National Cancer Institute, National Institutes of Health); NBHS (National Institutes of Health grant R01-CA100374); SCCS (National Institutes of Health grant R01-CA092447), WFBC (National Institutes of Health grant R01-CA73629); BWHS (National Institutes of Health grants R01-CA58420 and R01-CA98663) and WISE (National Institutes of Health grant P01-CA77596). OI Olopade and D Huo were supported by National Institutes of Health Specialized Program of Research Excellence in Breast Cancer, grant number P50-CA125183 and National Cancer Institute R01-CA141712. BBCS is supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research. The Breast Cancer Family Registry (BCFR) was supported by the National Cancer Institute, National Institutes of Health under RFA-CA-06-503 and through cooperative agreements with members of the Breast Cancer Family Registry and Principal Investigators. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the BCFR, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government or the BCFR. The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institute of Health, U.S. Department of Health and Human Services through contracts N01WH22110, 24152, 32100-2, 32105-6, 32108-9, 32111-13, 32115, 32118-32119,32122, 42107-26, 42129-32, and 44221. Funding for WHI SHARe genotyping was provided by NHLBI Contract N02-HL-64278. We thank the women who volunteered to participate in each study. We also thank Madhavi Eranti, Andrea Holbrook, Paul Poznaik, David Wong and Lucy Xia from the University of Southern California for their technical support. We would also like to acknowledge co-investigators from the WCHS study: Dana H. Bovbjerg (University of Pittsburgh), Lina Jandorf (Mount Sinai School of Medicine) and Gregory Ciupak, Warren Davis, Gary Zirpoli, Song Yao and Michelle Roberts from Roswell Park Cancer Institute.

Conflict of interest The authors declare that they have no conflict of interest.

Ethical statement The experiments done in this manuscript comply with the current laws of the country of USA.

References

Ahmed S, Thomas G, Ghoussaini M, Healey CS, Humphreys MK, Platte R, Morrison J, Maranian M, Pooley KA, Luben R, Eccles

D, Evans DG, Fletcher O, Johnson N, dos Santos Silva I, Peto J, Stratton MR, Rahman N, Jacobs K, Prentice R, Anderson GL, Rajkovic A, Curb JD, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver WR, Bojesen S, Nordestgaard BG, Flyger H, Dork T, Schurmann P, Hillemanns P, Karstens JH, Bogdanova NV, Antonenkova NN, Zalutsky IV, Bermisheva M, Fedorova S, Khusnutdinova E, Kang D, Yoo KY, Noh DY, Ahn SH, Devilee P, van Asperen CJ, Tollenaar RA, Seynaeve C, Garcia-Closas M, Lissowska J, Brinton L, Peplonska B, Nevanlinna H, Heikkinen T, Aittomaki K, Blomqvist C, Hopper JL, Southey MC, Smith L, Spurdle AB, Schmidt MK, Broeks A, van Hien RR, Cornelissen S, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Schmutzler RK, Burwinkel B, Bartram CR, Meindl A, Brauch H, Justenhoven C, Hamann U, Chang-Claude J, Hein R, Wang-Gohrke S, Lindblom A, Margolin S, Mannermaa A, Kosma VM, Kataja V, Olson JE, Wang X, Fredericksen Z, Giles GG, Severi G, Baglietto L, English DR, Hankinson SE, Cox DG, Kraft P, Vatten LJ, Hveem K, Kumle M et al (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 41:585–590. doi:10.1038/ng.354

Antoniou AC, Wang X, Fredericksen ZS, McGuffog L, Tarrell R, Sinilnikova OM, Healey S, Morrison J, Kartsonaki C, Lesnick T, Ghoussaini M, Barrowdale D, Peock S, Cook M, Oliver C, Frost D, Eccles D, Evans DG, Eeles R, Izatt L, Chu C, Douglas F, Paterson J, Stoppa-Lyonnet D, Houdayer C, Mazoyer S, Giraud S, Lasset C, Remenieras A, Caron O, Hardouin A, Berthet P, Hogervorst FB, Rookus MA, Jager A, van den Ouweland A, Hoogerbrugge N, van der Luijt RB, Meijers-Heijboer H, Gomez Garcia EB, Devilee P, Vreeswijk MP, Lubinski J, Jakubowska A, Gronwald J, Huzarski T, Byrski T, Gorski B, Cybulski C, Spurdle AB, Holland H, Goldgar DE, John EM, Hopper JL, Southey M, Buys SS, Daly MB, Terry MB, Schmutzler RK, Wappenschmidt B, Engel C, Meindl A, Preisler-Adams S, Arnold N, Niederacher D, Sutter C, Domchek SM, Nathanson KL, Rebbeck T, Blum JL, Piedmonte M, Rodriguez GC, Wakeley K, Boggess JF, Basil J, Blank SV, Friedman E, Kaufman B, Laitman Y, Milgrom R, Andrulis IL, Glendon G, Ozcelik H, Kirchhoff T, Vijai J, Gaudet MM, Altshuler D, Guiducci C, Loman N, Harbst K, Rantala J, Ehrencrona H, Gerdes AM, Thomassen M, Sunde L, Peterlongo P, Manoukian S, Bonanni B, Viel A, Radice P et al (2010) A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. Nat Genet 42:885–892. doi:10.1038/ng.669

Campbell MC, Tishkoff SA (2008) African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet 9:403–433. doi:10.1146/annurev.genom.9.081307.164258

Chen F, Stram DO, Le Marchand L, Monroe KR, Kolonel LN, Henderson BE, Haiman CA (2010) Caution in generalizing known genetic risk markers for breast cancer across all ethnic/racial populations. Eur J Hum Genet 19:243–245. doi:10.1038/ejhg.2010.185

Chen F, Chen GK, Millikan RC, John EM, Ambrosone CB, Bernstein L, Zheng W, Hu JJ, Ziegler RG, Deming SL, Bandera EV, Nyante S, Palmer JR, Rebbeck TR, Ingles SA, Press MF, Rodriguez-Gil JL, Chanock SJ, Le Marchand L, Kolonel LN, Henderson BE, Stram DO, Haiman CA (2011) Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans. Hum Mol Genet 20:4491–4503. doi:10.1093/hmg/ddr367

Easton DF, Pooley KA, Dunning AM, Pharoah PD, Thompson D, Ballinger DG, Struewing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, Johnson N, Seal S, Stratton MR, Rahman N, Chenevix-Trench G, Bojesen SE, Nordestgaard BG, Axelsson CK, Garcia-Closas M, Brinton L, Chanock S, Lissowska J, Peplonska B, Nevanlinna H, Fagerholm R, Eerola H, Kang D, Yoo KY, Noh DY, Ahn SH, Hunter DJ, Hankinson SE, Cox DG, Hall P, Wedren S, Liu J, Low YL, Bogdanova N, Schurmann P, Dork T, Tollenaar RA, Jacobi CE, Devilee P, Klijn JG, Sigurdson AJ, Doody MM, Alexander BH, Zhang J, Cox A, Brock IW, MacPherson G, Reed MW, Couch FJ, Goode EL, Olson JE, Meijers-Heijboer H, van den Ouweland A, Uitterlinden A, Rivadeneira F, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Hopper JL, McCredie M, Southey M, Giles GG, Schroen C, Justenhoven C, Brauch H, Hamann U, Ko YD, Spurdle AB, Beesley J, Chen X, Mannermaa A, Kosma VM, Kataja V, Hartikainen J, Day NE et al (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087–1093. doi:10.1038/nature05887

Fletcher O, Johnson N, Orr N, Hosking FJ, Gibson LJ, Walker K, Zelenika D, Gut I, Heath S, Palles C, Coupland B, Broderick P, Schoemaker M, Jones M, Williamson J, Chilcott-Burns S, Tomczyk K, Simpson G, Jacobs KB, Chanock SJ, Hunter DJ, Tomlinson IP, Swerdlow A, Ashworth A, Ross G, dos Santos Silva I, Lathrop M, Houlston RS, Peto J (2011) Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. J Natl Cancer Inst 103:425–435. doi:10.1093/jnci/djq563

Ghoussaini M, Fletcher O, Michailidou K, Turnbull C, Schmidt MK, Dicks E, Dennis J, Wang Q, Humphreys MK, Luccarini C, Baynes C, Conroy D, Maranian M, Ahmed S, Driver K, Johnson N, Orr N, dos Santos Silva I, Waisfisz Q, Meijers-Heijboer H, Uitterlinden AG, Rivadeneira F, Hall P, Czene K, Irwanto A, Liu J, Nevanlinna H, Aittomaki K, Blomqvist C, Meindl A, Schmutzler RK, Muller-Myhsok B, Lichtner P, Chang-Claude J, Hein R, Nickels S, Flesch-Janys D, Tsimiklis H, Makalic E, Schmidt D, Bui M, Hopper JL, Apicella C, Park DJ, Southey M, Hunter DJ, Chanock SJ, Broeks A, Verhoef S, Hogervorst FB, Fasching PA, Lux MP, Beckmann MW, Ekici AB, Sawyer E, Tomlinson I, Kerin M, Marme F, Schneeweiss A, Sohn C, Burwinkel B, Guenel P, Truong T, Cordina-Duverger E, Menegaux F, Bojesen SE, Nordestgaard BG, Nielsen SF, Flyger H, Milne RL, Alonso MR, Gonzalez-Neira A, Benitez J, Anton-Culver H, Ziogas A, Bernstein L, Dur CC, Brenner H, Muller H, Arndt V, Stegmaier C, Justenhoven C, Brauch H, Bruning T, Wang-Gohrke S, Eilber U, Dork T, Schurmann P, Bremer M, Hillemanns P, Bogdanova NV, Antonenkova NN, Rogov YI, Karstens JH, Bermisheva M, Prokofieva D, Khusnutdinova E, Lindblom A, Margolin S, Mannermaa A et al (2012) Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nat Genet 44:312–318. doi:10.1038/ng.1049

Haiman CA, Chen GK, Blot WJ, Strom SS, Berndt SI, Kittles RA, Rybicki BA, Isaacs WB, Ingles SA, Stanford JL, Diver WR, Witte JS, Hsing AW, Nemesure B, Rebbeck TR, Cooney KA, Xu J, Kibel AS, Hu JJ, John EM, Gueye SM, Watya S, Signorello LB, Hayes RB, Wang Z, Yeboah E, Tettey Y, Cai Q, Kolb S, Ostrander EA, Zeigler-Johnson C, Yamamura Y, Neslund-Dudas C, Haslag-Minoff J, Wu W, Thomas V, Allen GO, Murphy A, Chang BL, Zheng SL, Leske MC, Wu SY, Ray AM, Hennis AJ, Thun MJ, Carpten J, Casey G, Carter EN, Duarte ER, Xia LY, Sheng X, Wan P, Pooler LC, Cheng I, Monroe KR, Schumacher F, Le Marchand L, Kolonel LN, Chanock SJ, Berg DV, Stram DO, Henderson BE (2011a) Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. Nat Genet 43:570–573. doi:10.1038/ng.839

Haiman CA, Chen GK, Vachon CM, Canzian F, Dunning A, Millikan RC, Wang X, Ademuyiwa F, Ahmed S, Ambrosone CB, Baglietto L, Balleine R, Bandera EV, Beckmann MW, Berg CD, Bernstein L, Blomqvist C, Blot WJ, Brauch H, Buring JE, Carey LA, Carpenter JE, Chang-Claude J, Chanock SJ, Chasman DI, Clarke CL, Cox A, Cross SS, Deming SL, Diasio RB, Dimopoulos AM, Driver WR, Dunnebier T, Durcan L, Eccles D, Edlund CK, Ekici AB, Fasching PA, Feigelson HS, Flesch-Janys D, Fostira F, Forsti A, Fountzilas G, Gerty SM, Giles GG, Godwin AK, Goodfellow P, Graham N, Greco D, Hamann U, Hankinson SE, Hartmann A, Hein R, Heinz J, Holbrook A, Hoover RN, Hu JJ, Hunter DJ, Ingles SA, Irwanto A, Ivanovich J, John EM, Johnson N, Jukkola-Vuorinen A, Kaaks R, Ko YD, Kolonel LN, Konstantopoulou I, Kosma VM, Kulkarni S, Lambrechts D, Lee AM, Marchand LL, Lesnick T, Liu J, Lindstrom S, Mannermaa A, Margolin S, Martin NG, Miron P, Montgomery GW, Nevanlinna H, Nickels S, Nyante S, Olswold C, Palmer J, Pathak H, Pectasides D, Perou CM, Peto J, Pharoah PD, Pooler LC, Press MF, Pylkas K, Rebbeck TR, Rodriguez-Gil JL, Rosenberg L, Ross E, Rudiger T, Silva Idos S et al (2011b) A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. Nat Genet 43:1210–1214. doi:10.1038/ng.985ng.985

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

Huo D, Zheng Y, Ogundiran TO, Adebamowo C, Nathanson KL, Domchek SM, Rebbeck TR, Simon MS, John EM, Hennis A, Nemesure B, Wu SY, Leske MC, Ambs S, Niu Q, Zhang J, Cox NJ, Olopade OI (2012) Evaluation of 19 susceptibility loci of breast cancer in women of African ancestry. Carcinogenesis 33:835–840. doi:10.1093/carcin/bgs093

Hutter CM, Young AM, Ochs-Balcom HM, Carty CL, Wang T, Chen CT, Rohan TE, Kooperberg C, Peters U (2011) Replication of breast cancer GWAS susceptibility loci in the Women's Health Initiative African American SHARe Study. Cancer Epidemiol Biomarkers Prev 20:1950–1959. doi:10.1158/1055-9965.EPI-11-0524

Kim NW, Piatyszek MA, Prowse KR, Harley CB, West MD, Ho PL, Coviello GM, Wright WE, Weinrich SL, Shay JW (1994) Specific association of human telomerase activity with immortal cells and cancer. Science 266:2011–2015

Kim HC, Lee JY, Sung H, Choi JY, Park SK, Lee KM, Kim YJ, Go MJ, Li L, Cho YS, Park M, Kim DJ, Oh JH, Kim JW, Jeon JP, Jeon SY, Min H, Kim HM, Park J, Yoo KY, Noh DY, Ahn SH, Lee MH, Kim SW, Lee JW, Park BW, Park WY, Kim EH, Kim MK, Han W, Lee SA, Matsuo K, Shen CY, Wu PE, Hsiung CN, Kim HL, Han BG, Kang D (2012) 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 14:R56. doi:10.1186/bcr3158

Li Y, Abecasis GR (2006) Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet S79:2290

Long J, Cai Q, Sung H, Shi J, Zhang B, Choi JY, Wen W, Delahanty RJ, Lu W, Gao YT, Shen H, Park SK, Chen K, Shen CY, Ren Z, Haiman CA, Matsuo K, Kim MK, Khoo US, Iwasaki M, Zheng Y, Xiang YB, Gu K, Rothman N, Wang W, Hu Z, Liu Y, Yoo KY, Noh DY, Han BG, Lee MH, Zheng H, Zhang L, Wu PE, Shieh YL, Chan SY, Wang S, Xie X, Kim SW, Henderson BE, Le Marchand L, Ito H, Kasuga Y, Ahn SH, Kang HS, Chan KY, Iwata H, Tsugane S, Li C, Shu XO, Kang DH, Zheng W (2012) Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer. PLoS Genet 8:e1002532. doi:10.1371/journal.pgen.1002532PGENETICS-D-11-01537

Meyer KB, Maia AT, O'Reilly M, Teschendorff AE, Chin SF, Caldas C, Ponder BA (2008) Allele-specific up-regulation of FGFR2 increases susceptibility to breast cancer. PLoS Biol 6:e108. doi:10.1371/journal.pbio.0060108

Pasaniuc B, Zaitlen N, Lettre G, Chen G, Tandon A, Kao L, Ruczinski I, Fornage M, Siscovick D, Zhu X, Larkin E, Lange L (2011) Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a breast cancer consortium (under review)

Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA (2002) Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 31:33–36. doi:10.1038/ng853ng853

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–9. doi:10.1038/ng1847

Price AL, Tandon A, Patterson N, Barnes KC, Rafaels N, Ruczinski I, Beaty TH, Mathias R, Reich D, Myers S (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet 5:e1000519. doi:10.1371/journal.pgen.1000519

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959

Reed FA, Tishkoff SA (2006) African human diversity, origins and migrations. Curr Opin Genet Dev 16:597–605. doi:10.1016/j.gde.2006.10.008

Ruiz-Narvaez EA, Rosenberg L, Cozier YC, Cupples LA, Adams-Campbell LL, Palmer JR (2010) Polymorphisms in the TOX3/LOC643714 locus and risk of breast cancer in African-American women. Cancer Epidemiol Biomarkers Prev 19:1320–1327

Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, Masson G, Jakobsdottir M, Thorlacius S, Helgason A, Aben KK, Strobbe LJ, Albers-Akkers MT, Swinkels DW, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Godino J, Garcia-Prats MD, Polo E, Tres A, Mouy M, Saemundsdottir J, Backman VM, Gudmundsson L, Kristjansson K, Bergthorsson JT, Kostic J, Frigge ML, Geller F, Gudbjartsson D, Sigurdsson H, Jonsdottir T, Hrafnkelsson J, Johannsson J, Sveinsson T, Myrdal G, Grimsson HN, Jonsson T, von Holst S, Werelius B, Margolin S, Lindblom A, Mayordomo JI, Haiman CA, Kiemeney LA, Johannsson OT, Gulcher JR, Thorsteinsdottir U, Kong A, Stefansson K (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 39:865–869. doi:10.1038/ng2064

Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, Jonsson GF, Jakobsdottir M, Bergthorsson JT, Gudmundsson J, Aben KK, Strobbe LJ, Swinkels DW, van Engelenburg KC, Henderson BE, Kolonel LN, Le Marchand L, Millastre E, Andres R, Saez B, Lambea J, Godino J, Polo E, Tres A, Picelli S, Rantala J, Margolin S, Jonsson T, Sigurdsson H, Jonsdottir T, Hrafnkelsson J, Johannsson J, Sveinsson T, Myrdal G, Grimsson HN, Sveinsdottir SG, Alexiusdottir K, Saemundsdottir J, Sigurdsson A, Kostic J, Gudmundsson L, Kristjansson K, Masson G, Fackenthal JD, Adebamowo C, Ogundiran T, Olopade OI, Haiman CA, Lindblom A, Mayordomo JI, Kiemeney LA, Gulcher JR, Rafnar T, Thorsteinsdottir U, Johannsson OT, Kong A, Stefansson K (2008) Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 40:703–706. doi:10.1038/ng.131

Tarsounas M, Davies AA, West SC (2004) RAD51 localization and activation following DNA damage. Philos Trans R Soc Lond B Biol Sci 359:87–93. doi:10.1098/rstb.2003.1368

Tenhagen M, van Diest PJ, Ivanova IA, van der Wall E, van der Groep P (2012) Fibroblast growth factor receptors in breast

cancer: expression, downstream effects, and possible drug targets. Endocr Relat Cancer 19:R115–R129. doi:10.1530/ERC-12-0060

Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, Hankinson SE, Hutchinson A, Wang Z, Yu K, Chatterjee N, Garcia-Closas M, Gonzalez-Bosquet J, Prokunina-Olsson L, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Diver R, Prentice R, Jackson R, Kooperberg C, Chlebowski R, Lissowska J, Peplonska B, Brinton LA, Sigurdson A, Doody M, Bhatti P, Alexander BH, Buring J, Lee IM, Vatten LJ, Hveem K, Kumle M, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Chanock SJ, Hunter DJ (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet 41:579–584. doi:10.1038/ng.353

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

Udler MS, Meyer KB, Pooley KA, Karlins E, Struewing JP, Zhang J, Doody DR, MacArthur S, Tyrer J, Pharoah PD, Luben R, Bernstein L, Kolonel LN, Henderson BE, Le Marchand L, Ursin G, Press MF, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Kang D, Yoo KY, Noh DY, Ahn SH, Ponder BA, Haiman CA, Malone KE, Dunning AM, Ostrander EA, Easton DF (2009) FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. Hum Mol Genet 18:1692–1703. doi:10.1093/hmg/ddp078

Zheng W, Cai Q, Signorello LB, Long J, Hargreaves MK, Deming SL, Li G, Li C, Cui Y, Blot WJ (2009a) Evaluation of 11 breast cancer susceptibility loci in African–American women. Cancer Epidemiol Biomarkers Prev 18:2761–2764. doi:10.1158/1055-9965.EPI-09-0624

Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL, Gu K, Fair AM, Cai Q, Lu W, Shu XO (2009b) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 41:324–328. doi:ng.31810.1038/ng.318

nature genetics

A common variant at the *TERT-CLPTM1L* locus is associated with estrogen receptor–negative breast cancer

Estrogen receptor (ER)-negative breast cancer shows a higher incidence in women of African ancestry compared to women of European ancestry. In search of common risk alleles for ER-negative breast cancer, we combined genome-wide association study (GWAS) data from women of African ancestry (1,004 ER-negative cases and 2,745 controls) and European ancestry (1,718 ER-negative cases and 3,670 controls), with replication testing conducted in an additional 2,292 ER-negative cases and 16,901 controls of European ancestry. We identified a common risk variant for ER-negative breast cancer at the *TERT-CLPTM1L* locus on chromosome 5p15 (rs10069690: per-allele odds ratio (OR) = 1.18 per allele, $P = 1.0 \times 10^{-10}$). The variant was also significantly associated with triple-negative (ER-negative, progesterone receptor (PR)-negative and human epidermal growth factor-2 (HER2)-negative) breast cancer (OR = 1.25, $P = 1.1 \times 10^{-9}$), particularly in younger women (<50 years of age) (OR = 1.48, $P = 1.9 \times 10^{-9}$). Our results identify a genetic locus associated with estrogen receptor negative breast cancer subtypes in multiple populations.

Compared to women of European ancestry, women of African descent are more likely to be diagnosed with ER-negative breast cancer[1]. ER-negative tumors and triple-negative tumors are observed at even higher rates among African women currently residing in Africa[2], suggesting a genetic component to the high risk of ER-negative phenotypes in women of African descent. Similarly, ER-negative breast cancers and triple-negative breast cancers are also the predominant histological subtypes in women with germline mutations in *BRCA1* (ref. 3). The enrichment for ER-negative disease in this genetically predisposed population also suggests the existence of additional genetic factors that contribute to the risk of ER-negative disease. Support for the presence of these factors was recently provided by a GWAS of breast cancer in *BRCA1* mutation carriers, in which a common risk variant for ER-negative breast cancer on chromosome 19p13 was identified that also was significantly associated with ER-negative and triple-negative disease in the general population[4].

To search for genetic risk factors for ER-negative breast cancer phenotypes, we combined results from a GWAS of breast cancer in African-American women (African American Breast Cancer Consortium (AABC): 3,016 cases (1,004 with ER-negative disease) and 2,745 controls) with results from a GWAS of triple-negative breast cancer in women of European ancestry (Triple-Negative Breast Cancer Consortium (TNBCC): 1,718 cases and 3,670 controls). Genotyping in AABC was conducted with the Illumina Infinium 1M Duo. In TNBCC, cases were genotyped with the Illumina 660W array, a subset of cases from the Mammary Carcinoma Risk Factor Investigation (MARIE) component were genotyped using the Illumina CNV370 SNP array, and cases and controls from the Helsinki Breast Cancer Study (HEBCS) component were genotyped using the Illumina 550-Duo SNP array. Genotypes of TNBCC cases were compared with GWAS data for publicly available controls (Online Methods). Both studies imputed genotypes for common SNPs in phase 2 HapMap populations (release 21) (**Supplementary Table 1** and Online Methods). A total of 3,154,485 SNPs, genotyped and imputed, were analyzed in stage 1 of the meta-analysis.

We observed little evidence of inflation in the test statistics in AABC ($\lambda = 1.01$) or TNBCC ($\lambda = 1.04$) or in the meta-analysis of the two GWAS ($\lambda = 1.02$; **Supplementary Fig. 1**). In the combined results, only SNP rs10069690 (NCBI36/hg18, chr5:1,332,790) located in intron 4 of the *TERT* gene (encoding telomerase reverse transcriptase) at chromosome 5p15 showed a genome-wide significant association with ER-negative breast cancer (AABC: OR per allele = 1.32, $P = 1.3 \times 10^{-6}$; TNBCC: OR = 1.25, $P = 1.2 \times 10^{-3}$; combined OR = 1.29,

Table 1 Association of rs10069690 at 5p15 and ER-negative breast cancer risk

Stage	Consortium or study	Cases/controls[a]	RAF[b] T allele	Heterozygotes OR (95% CI)[c]	Homozygotes OR (95% CI)[c]	Per-allele OR (95% CI)[c]	*P* value (1-d.f.)[d]
1	AABC	1,002/2,743	0.57	1.32 (1.05–1.67)	1.74 (1.37–2.21)	1.32 (1.18–1.48)	1.3×10^{-6}
1	TNBCC	2,785/1,602	0.27	1.10 (0.97–1.26)	1.53 (1.21–1.95)	1.18 (1.07–1.30)	1.0×10^{-3}
2	BPC3	1,289/10,397	0.26	1.08 (0.96–1.22)	1.19 (0.95–1.49)	1.09 (0.99–1.19)	0.077
2	SEARCH	933/5,966	0.26	1.23 (1.06–1.43)	1.44 (1.10–1.89)	1.21 (1.09–1.36)	6.9×10^{-4}
Combined		6,009/20,708		1.15 (1.06–1.23)	1.46 (1.29–1.64)	1.18 (1.13–1.25)	1.0×10^{-10}

[a]Number of cases and controls with genotype data for rs10069690. All subjects were directly genotyped. [b]Risk allele frequency (RAF) in controls. [c]Adjusted for age, study and principal components in AABC. Adjusted for age and country in TNBCC. Adjusted for age, study and country (European Prospective Investigation into Cancer and Nutrition (EPIC) only) in BPC3. Adjusted for age in SEARCH. Combined results are from the meta-analysis. [d]*P* for trend (one degree of freedom (1-d.f.)).

A full list of authors and affiliations appears at the end of the paper.

Received 3 May; accepted 28 September; published online 30 October 2011; doi:10.1038/ng.985

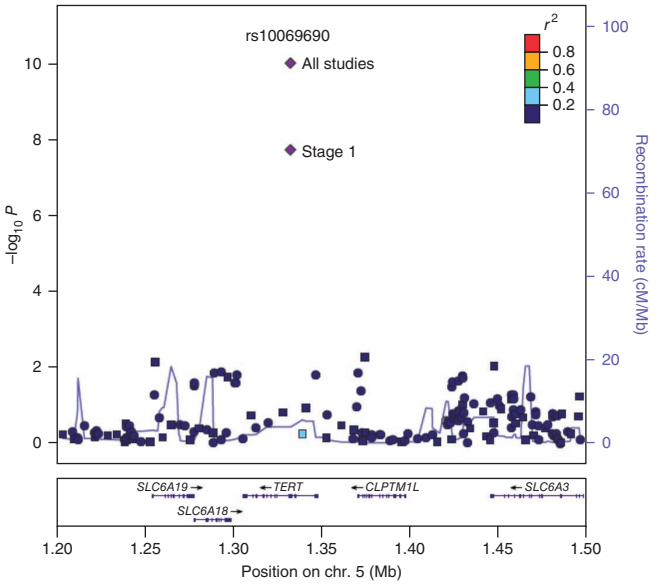

Figure 1 A regional plot of the $-\log_{10} P$ values for SNPs at the chromosome 5p15 risk locus from the meta-analysis of the AABC and TNBCC stage 1 studies. SNP rs10069690 is designated with the purple diamonds. The colors depict the strength of the correlation (r^2) between SNP rs10069690 and the SNPs tested in the region. The correlation is estimated using 1000 Genomes Project (1KGP) data for the HapMap CEU population (June 2010). Squares are SNPs that were genotyped in AABC and TNBCC. Circles are SNPs that were genotyped in one study and imputed in the other or imputed in both studies. The blue line indicates the recombination rates in centimorgans (cM) per megabase (Mb). Also shown are the SNP Build 36 coordinates and genes in the region.

$P = 1.0 \times 10^{-8}$). Whereas SNP rs10069690 was genotyped in AABC, it was imputed in TNBCC ($R^2 = 0.55$). To verify the imputed genotypes and the significance of the association in TNBCC, we re-genotyped rs10069690 in available DNA samples from 2,963 TNBCC cases and 1,632 study-specific TNBCC controls (Online Methods). Although the overlapping samples between the TNBCC GWAS and the re-genotyping study showed that the quality of imputation for rs10069690 in the GWAS was poor (Online Methods), the association with ER-negative breast cancer for rs10069690 remained statistically significant in the larger re-genotyped TNBCC sample (OR = 1.18, $P = 1.0 \times 10^{-3}$; **Table 1** and **Fig. 1**) and in the new combined results for AABC and the re-genotyped TNBCC sample (OR = 1.24, $P = 1.6 \times 10^{-8}$).

To further confirm the association at 5p15, we genotyped SNP rs10069690 in women of European ancestry, which included 8,365 cases (1,359 ER negative) and 10,935 controls from the US National Cancer Institute Breast and Prostate Cancer Cohort Consortium (BPC3) and 6,182 cases (933 ER negative) and 5,966 controls from Studies of Epidemiology and Risk Factors in Cancer Heredity (SEARCH). Evidence for replication was observed for rs10069690 and ER-negative breast cancer in both studies (BPC3: OR = 1.09, $P = 0.077$; SEARCH: OR = 1.21, $P = 6.9 \times 10^{-4}$; **Table 1**).

In combining the results across all studies (6,009 ER-negative cases and 20,708 controls with genotype data), rs10069690 was significantly associated with an increased risk of ER-negative breast cancer (OR = 1.18, 95% confidence interval (CI), 1.13–1.25; $P = 1.0 \times 10^{-10}$; **Table 1**). The risk for heterozygote and homozygote carriers was 1.15 (95% CI, 1.06–1.23) and 1.46 (95% CI, 1.29–1.64), respectively. We observed little evidence of heterogeneity for the reported association for this variant by study or country in AABC (test for heterogeneity, $p_{het} = 0.86$), TNBCC ($p_{het} = 0.85$) or BPC3 ($p_{het} = 0.37$; **Supplementary Table 2**).

In an analysis of ER-positive cases, rs10069690 was only weakly associated with risk in African Americans (AABC: 1,558 ER-positive cases and 2,743 controls with genotype data, OR = 1.08, $P = 0.10$) and in women of European ancestry (BPC3: 4,890 ER-positive cases and 10,397 controls, OR = 1.03, $P = 0.31$; SEARCH: 3,534 ER positive cases and 5,966 controls, OR = 1.03, $P = 0.37$; combined for all populations: OR = 1.04, $P = 0.06$, $p_{het} = 0.64$). The statistical power to detect an OR of 1.18 (observed for ER-negative disease) for ER-positive disease was >99% in the combined sample (9,982 cases and 19,106 controls), assuming the risk allele frequency of 0.26 in people of European decent. This result suggests that the association with breast cancer might be specific for ER-negative subtypes (P value for case-only test of ER negative versus ER positive = 1.7×10^{-4}).

We further stratified the cases by HER2 status to assess whether this region may be a risk locus for triple-negative disease. In AABC, BPC3 and SEARCH the association with rs10069690 was greater for triple-negative tumors than for ER-negative, PR-negative, HER2-positive tumors (**Table 2**), and, in combining all studies, including TNBCC, the association with rs10069690 was significantly greater for triple-negative disease (3,707 triple-negative cases and 19,728 controls with genotype data, OR = 1.25, $P = 1.1 \times 10^{-9}$; 376 ER-negative, PR-negative, HER2-positive cases and 18,126 controls, OR = 1.03, $P = 0.71$; P value for case-only test = 0.010). The association with rs10069690 was also observed to be significantly greater for ER-negative and triple-negative disease at younger ages (<50 years: ER negative,

Table 2 Association of rs10069690 at 5p15 stratified by HER2 status

Consortium or study	Subtype	Cases/controls[a]	Heterozygotes OR (95% CI)[b]	Homozygotes OR (95% CI)[b]	Per-allele OR (95% CI)[b]	P value (1-d.f.)[c]	Case-only P
AABC[d]	ER$^-$PR$^-$HER2$^-$	440/2,407	1.35 (0.97–1.89)	1.78 (1.27–2.49)	1.33 (1.14–1.55)	3.0×10^{-4}	0.19
	ER$^-$PR$^-$HER2$^+$	115/2,407	1.83 (0.99–3.40)	1.59 (0.82–3.05)	1.15 (0.86–1.52)	0.34	
TNBCC	ER$^-$PR$^-$HER2$^-$	2,785/1,602	1.10 (0.97–1.26)	1.53 (1.21–1.95)	1.18 (1.07–1.30)	1.0×10^{-3}	–
BPC3[e]	ER$^-$PR$^-$HER2$^-$	300/9,753	1.19 (0.93–1.52)	1.64 (1.10–2.46)	1.25 (1.04–1.49)	0.015	0.13
	ER$^-$PR$^-$HER2$^+$	198/9,753	0.99 (0.73–1.33)	0.95 (0.53–1.70)	0.98 (0.78–1.23)	0.87	
SEARCH	ER$^-$PR$^-$HER2$^-$	182/5,966	1.42 (1.03–1.95)	2.41 (1.47–3.95)	1.51 (1.20–1.89)	4.2×10^{-4}	0.058
	ER$^-$PR$^-$HER2$^+$	63/5,966	1.31 (0.79–2.16)	0.27 (0.04–1.95)	0.97 (0.64–1.46)	0.88	
Combined	ER$^-$PR$^-$HER2$^-$	3,707/19,728[f]	1.17 (1.06–1.30)	1.69 (1.43–1.99)	1.25 (1.16–1.34)	1.1×10^{-9}	0.010
	ER$^-$PR$^-$HER2$^+$	376/18,126	1.15 (0.91–1.46)	1.11 (0.73–1.70)	1.03 (0.88–1.21)	0.71	

[a]Number of cases and controls with genotype data for rs10069690. All subjects were directly genotyped. [b]Adjusted for age, study and principal components in AABC. Adjusted for age and country in TNBCC. Adjusted for age, study and country (EPIC only) in BPC3. Adjusted for age in SEARCH. Combined results are from the meta-analysis. [c]P for trend (1-d.f.). [d]Excludes San Francisco Bay Area Breast Cancer Study (SFBCS) and Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO), as HER2 data were not available. [e]Excludes WHS , as HER2 data were not available. [f]Includes TNBCC. Without TNBCC: 922 ER$^-$PR$^-$HER2$^-$ cases and 18,126 controls; OR per allele = 1.33 (1.20–1.48), $P = 6.3 \times 10^{-8}$; heterozygotes: OR = 1.29 (1.09–1.53); homozygotes: OR = 1.85 (1.47–2.33).

OR = 1.32, $P = 1.4 \times 10^{-8}$; triple negative, OR = 1.48, $P = 1.9 \times 10^{-9}$; P for interaction with age = 0.035 and 3.2×10^{-3}, respectively; **Supplementary Table 3**). We found no significant association with rs1006960 among ER- and PR-positive cases when stratified by HER2 status (513 triple-positive cases and 18,126 controls, OR = 1.09, $P = 0.21$; 2,808 ER-positive, PR-positive, HER2-negative cases and 18,126 controls, OR = 1.04, $P = 0.29$), which suggests the association may be limited to triple-negative disease and not all HER2-negative tumors.

Similar to 8q24 (refs. 5–7) and 11q13 (refs. 8–10), the *TERT-CLPTM1L* locus harbors multiple risk variants for different cancers (reviewed in ref. 11). SNP rs10069690 is modestly correlated ($r^2 = 0.13$–0.43 in 1000 Genomes Project populations of European and African ancestry, **Supplementary Fig. 2**) with variants found for serous ovarian cancer (rs7726159), glioma (rs2736100) and lung cancer (rs2736100, rs2735940)[12–14]. Aside from risk variant rs2853676 found for glioma[14], which we found to be associated with risk in TNBCC ($P = 0.014$, $r^2 = 0.05$ with rs10069690), none of the known risk variants identified for other cancers in the *TERT-CLPTM1L* region was significantly associated with breast cancer risk in TNBCC or AABC. Although rs7726159 was not tested in AABC or TNBCC (as it is not on the Illumina arrays or in HapMap), it is noteworthy that the first common risk variant identified for ER-negative breast cancer, at chromosome 19p13, is also associated with risk for serous ovarian cancer[15]. The *TERT* gene encodes the catalytic subunit of telomerase, which controls telomere length, a process linked with genomic instability and implicated in tumorigenesis. Sequencing of the coding exons of *TERT* in 96 African-American women (Online Methods) did not reveal a coding variant strongly correlated with rs10069690. The *TERT* locus may highlight another biological process common to the pathogenesis of ER-negative breast cancer subtypes and serous ovarian cancer that is also shared with other cancers.

Identification of the variant directly responsible for the association will be required to fully address the extent to which this locus contributes to the greater incidence of ER-negative and triple-negative tumors in women of African ancestry. However, it is notable that the risk allele frequency of rs10069690 is greater in African American women (frequency, 0.57) than in women of European ancestry (frequency, 0.26). If this variant is an equally good surrogate for the biologically functional allele in each population, then this locus may be responsible for a 15% (95% CI, 10–20%) higher incidence rate of ER-negative or triple-negative breast cancer in women of African compared to European ancestry (Online Methods). Larger studies with well-characterized tumor pathology information will be needed to determine whether the association we observed applies to all ER-negative disease or just the triple-negative subtype. Our findings provide further support for the presence of genetic susceptibility to ER-negative breast cancer subtypes and demonstrate the importance of discovery efforts in multiple populations.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This work was supported by a US Department of Defense Breast Cancer Research Program Era of Hope Scholar Award to C.A.H. (W81XWH-08-1-0383), the Norris Foundation, the Mayo Clinic College of Medicine, Komen Foundation for the Cure, the Breast Cancer Research Foundation and US National Institutes of Health grants CA128978, CA122340 and CA148065. Study specific acknowledgments are listed in the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

Conceived of and designed the experiments: C.A.H. and F.J.C. Performed the experiments and analyzed the data: C.A.H., L.C.P., D.V.D.B., X.S., G.K.C., A. Holbrook, P.W., F.C., D.O.S., X.W., T.L., C.O., K.N.S., A.M.L., L.Y.X., S.L.S. and C.M.V. Contributed reagents, materials, analysis tools or comments on the manuscript: C.A.H., C.M.V., A.D., R.C.M., X.W., F.A., S.A., C.B.A., L. Baglietto, R.B., E.V.B., M.W.B., C.D.B., L. Bernstein, C.B., W.J.B., H.B., J.E.B., L.A.C., J.E.C., J.C.-C., S.J.C., D.I.C., C.L.C., A.C., S.S.C., S.L.D., R.B.D., A.M.D., W.R.D., T.D., L.D., D.E., C.K.E., A.B.E., P.A.F., H.S.F., D.F.-J., F.F., A.F., G.F., S.M.G., G.G.G., A.K.G., P.G., N.G., D.G., U.H., S.E.H., A. Hartmann, R.H., J.H., R.N.H., J.J.H., D.J.H., S.A.I., A.I., J.I., E.M.J., N.J., A.J.-V., R.K., Y.-D.K., L.N.K., I.K., V.-M.K., S.K., D.L., A.M.L., L.L.M., T.L., J.L., S.L., A.M., S.M., N.G.M., P.M., G.W.M., H.N., S. Nickels, S. Nyante, C.O., J. Palmer, H.P., D.P., C.M.P., J. Peto, P.D.P.P., L.C.P., M.F.P., K.P., T.R.R., J.L.R.-G., L.R., E.R., T.R., I.d.S.S., E.S., M.K.S., R.S.-W., F.S., G.S., X.S., L.B.S., H.-P.S., K.N.S., M.C.S., W.J.T., I.T., F.B.L.H., E.W., J.W., H.W., R.W., D.Y., W.Z., R.G.Z., A.S., S.L.S., D.O.S., D.E., P.K., B.E.H. and F.J.C. Wrote the paper: C.A.H. and F.J.C.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Carey, L.A. *et al.* Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *J. Am. Med. Assoc.* **295**, 2492–2502 (2006).
2. Huo, D. *et al.* Population differences in breast cancer: survey in indigenous African women reveals over-representation of triple-negative breast cancer. *J. Clin. Oncol.* **27**, 4515–4521 (2009).
3. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. USA* **98**, 10869–10874 (2001).
4. Antoniou, A.C. *et al.* A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat. Genet.* **42**, 885–892 (2010).
5. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
6. Haiman, C.A. *et al.* A common genetic risk factor for colorectal and prostate cancer. *Nat. Genet.* **39**, 954–956 (2007).
7. Kiemeney, L.A. *et al.* Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat. Genet.* **40**, 1307–1312 (2008).
8. Purdue, M.P. *et al.* Genome-wide association study of renal cell carcinoma identifies two susceptibility loci on 2p21 and 11q13.3. *Nat. Genet.* **43**, 60–65 (2011).
9. Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.* **40**, 310–315 (2008).
10. Turnbull, C. *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* **42**, 504–507 (2010).
11. Baird, D.M. Variation at the TERT locus and predisposition for cancer. *Expert Rev. Mol. Med.* **12**, e16 (2010).
12. Johnatty, S.E. *et al.* Evaluation of candidate stromal epithelial cross-talk genes identifies association between risk of serous ovarian cancer and TERT, a cancer susceptibility "hot-spot". *PLoS Genet.* **6**, e1001016 (2010).
13. McKay, J.D. *et al.* Lung cancer susceptibility locus at 5p15.33. *Nat. Genet.* **40**, 1404–1406 (2008).
14. Shete, S. *et al.* Genome-wide association study identifies five susceptibility loci for glioma. *Nat. Genet.* **41**, 899–904 (2009).
15. Bolton, K.L. *et al.* Common variants at 19p13 are associated with susceptibility to ovarian cancer. *Nat. Genet.* **42**, 880–884 (2010).

Christopher A Haiman[1], Gary K Chen[1], Celine M Vachon[2], Federico Canzian[3], Alison Dunning[4], Robert C Millikan[5], Xianshu Wang[6], Foluso Ademuyiwa[7], Shahana Ahmed[4], Christine B Ambrosone[8], Laura Baglietto[9], Rosemary Balleine[10], Elisa V Bandera[11], Matthias W Beckmann[12], Christine D Berg[13], Leslie Bernstein[14], Carl Blomqvist[15], William J Blot[16,17], Hiltrud Brauch[18,19], Julie E Buring[20], Lisa A Carey[21], Jane E Carpenter[22], Jenny Chang-Claude[23], Stephen J Chanock[24], Daniel I Chasman[20], Christine L Clarke[22],

Angela Cox[25], Simon S Cross[26], Sandra L Deming[16], Robert B Diasio[27], Athanasios M Dimopoulos[28], W Ryan Driver[29], Thomas Dünnebier[30], Lorraine Durcan[31], Diana Eccles[31], Christopher K Edlund[1], Arif B Ekici[32], Peter A Fasching[12,33], Heather S Feigelson[34], Dieter Flesch-Janys[35], Florentia Fostira[36], Asta Försti[37,38], George Fountzilas[39], Susan M Gerty[31], The Gene Environment Interaction and Breast Cancer in Germany (GENICA) Consortium[40], Graham G Giles[9], Andrew K Godwin[41], Paul Goodfellow[42], Nikki Graham[31], Dario Greco[43], Ute Hamann[30], Susan E Hankinson[44,45], Arndt Hartmann[46], Rebecca Hein[23], Judith Heinz[35], Andrea Holbrook[1], Robert N Hoover[24], Jennifer J Hu[47], David J Hunter[45,48], Sue A Ingles[1], Astrid Irwanto[49], Jennifer Ivanovich[42], Esther M John[50,51], Nicola Johnson[52], Arja Jukkola-Vuorinen[53], Rudolf Kaaks[54], Yon-Dschun Ko[55], Laurence N Kolonel[56], Irene Konstantopoulou[36], Veli-Matti Kosma[57], Swati Kulkarni[58], Diether Lambrechts[59,60], Adam M Lee[27], Loïc Le Marchand[56], Timothy Lesnick[2], Jianjun Liu[49], Sara Lindstrom[45,48], Arto Mannermaa[61,62], Sara Margolin[63], Nicholas G Martin[64], Penelope Miron[65], Grant W Montgomery[64], Heli Nevanlinna[43], Stephan Nickels[23], Sarah Nyante[5], Curtis Olswold[2], Julie Palmer[66], Harsh Pathak[67], Dimitrios Pectasides[68], Charles M Perou[69], Julian Peto[70], Paul D P Pharoah[4], Loreall C Pooler[1], Michael F Press[71], Katri Pylkäs[72], Timothy R Rebbeck[73], Jorge L Rodriguez-Gil[47], Lynn Rosenberg[66], Eric Ross[74], Thomas Rüdiger[75], Isabel dos Santos Silva[70], Elinor Sawyer[76], Marjanka K Schmidt[77], Rüdiger Schulz-Wendtland[46], Fredrick Schumacher[1], Gianluca Severi[9], Xin Sheng[1], Lisa B Signorello[16,17], Hans-Peter Sinn[78], Kristen N Stevens[2], Melissa C Southey[79], William J Tapper[31], Ian Tomlinson[80], Frans B L Hogervorst[81], Els Wauters[59,60], JoEllen Weaver[67], Hans Wildiers[82], Robert Winqvist[72], David Van Den Berg[1], Peggy Wan[1], Lucy Y Xia[1], Drakoulis Yannoukakos[36], Wei Zheng[16], Regina G Ziegler[24], Afshan Siddiq[83], Susan L Slager[2], Daniel O Stram[1], Douglas Easton[4], Peter Kraft[45,48,84], Brian E Henderson[1] & Fergus J Couch[2,6]

[1]Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California, USA. [2]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA. [3]Genomic Epidemiology Group, DKFZ, Heidelberg, Germany. [4]Centre for Cancer Genetic Epidemiology, Strangeways Laboratory, Worts Causeway, Cambridge, UK. [5]Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina, USA. [6]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA. [7]Department of Medicine, Roswell Park Cancer Institute, Buffalo, New York, USA. [8]Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, New York, USA. [9]Cancer Epidemiology Centre, The Cancer Council Victoria & Centre for Molecular, Environmental, Genetic, and Analytic Epidemiology, The University of Melbourne, Victoria, Australia. [10]Department of Translational Oncology, Westmead Hospital, Western Sydney Local Health Network, Westmead, New South Wales, Australia. [11]The Cancer Institute of New Jersey, New Brunswick, New Jersey, USA. [12]Department of Gynecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Erlangen, Germany. [13]Division of Cancer Prevention, National Cancer Institute, US National Institutes of Health, Bethesda, Maryland, USA. [14]Division of Cancer Etiology, Department of Population Science, Beckman Research Institute, City of Hope, California, USA. [15]Department of Oncology, Helsinki University Central Hospital, Helsinki, Finland. [16]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center and Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. [17]International Epidemiology Institute, Rockville, Maryland, USA. [18]Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany. [19]University of Tübingen, Tübingen, Germany. [20]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA. [21]Department of Medicine, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina, USA. [22]Australian Breast Cancer Tissue Bank, University of Sydney at the Westmead Millennium Institute, Westmead, New South Wales, Australia. [23]Division of Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany. [24]Division of Cancer Epidemiology and Genetics, National Cancer Institute, US National Institutes of Health, Bethesda, Maryland, USA. [25]Institute for Cancer Studies, Department of Oncology, Faculty of Medicine, Dentistry & Health, University of Sheffield, Sheffield, UK. [26]Academic Unit of Pathology, Department of Neuroscience, Faculty of Medicine, Dentistry & Health, University of Sheffield, Sheffield, UK. [27]Department of Pharmacology, Mayo Clinic, Rochester, Minnesota, USA. [28]Department of Clinical Therapeutics, "Alexandra" Hospital, University of Athens School of Medicine, Athens, Greece. [29]Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA. [30]Molecular Genetics of Breast Cancer, DKFZ, Heidelberg, Germany. [31]Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton, UK. [32]Institute of Human Genetics, Friedrich-Alexander University of Erlangen-Nuremberg, Erlangen, Germany. [33]Department of Medicine, Division of Hematology and Oncology, David Geffen School of Medicine, University of California–Los Angeles, Los Angeles, California, USA. [34]Kaiser Permanente Colorado, Denver, Colorado, USA. [35]Institute for Medical Biometrics and Epidemiology, University Clinic Hamburg-Eppendorf, Hamburg, Germany. [36]Molecular Diagnostics Laboratory Institute of Radioisotopes and Radiodiagnostic Products, National Centre for Scientific Research "Demokritos", Athens, Greece. [37]Division of Molecular Genetic Epidemiology, DKFZ, Heidelberg, Germany. [38]Center for Primary Health Care Research, University of Lund, Malmö, Sweden. [39]Department of Medical Oncology, Aristotle University of Thessaloniki, Papageorgiou Hospital, Thessaloniki, Greece. [40]A full list of members is provided in the **Supplementary Note**. [41]Department of Pathology and Laboratory Medicine, Kansas University Medical Center, Lawrence, Kansas, USA. [42]Washington University School of Medicine, Barnes-Jewish Hospital and Siteman Cancer Center, St. Louis, Missouri, USA. [43]Department of Obstetrics and Gynecology, Helsinki University Central Hospital, Helsinki, Finland. [44]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [45]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [46]Institute of Pathology, University Hospital Erlangen, Friedrich-Alexander University of Erlangen-Nuremberg, Erlangen, Germany. [47]Sylvester Comprehensive Cancer Center and Department of Epidemiology and Public Health, University of Miami Miller School of Medicine, Miami, Florida, USA. [48]Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [49]Human Genetics Division, Genome Institute of Singapore, Singapore. [50]Cancer Prevention Institute of California, Fremont, California. [51]Stanford University School of Medicine and Stanford Cancer Center, Stanford, California, USA. [52]Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, London, UK. [53]Department of Oncology, Oulu University Hospital, University of Oulu, Oulu, Finland. [54]Division of Cancer Epidemiology, DKFZ, Heidelberg, Germany. [55]Department of Internal Medicine, Evangelische Kliniken Johanniter- und Waldkrankenhaus Bonn gGmbH, Bonn, Germany. [56]Epidemiology Program, Cancer Research Center, University of Hawaii, Honolulu, Hawaii, USA. [57]Department of Pathology, Imaging Centre, Kuopio University Hospital, Kuopio, Finland. [58]Department of Surgical Oncology, Roswell Park Cancer Institute, Buffalo, New York, USA. [59]Vesalius Research Center, Vlaams Instituut voor Biotechnologie, Leuven, Belgium. [60]Vesalius Research Center, University of Leuven, Leuven, Belgium. [61]Institute of Clinical Medicine, Department of Pathology, University of Eastern Finland Biocenter Kuopio, Kuopio, Finland. [62]Department of Pathology, Imaging Centre, Kuopio University Hospital, Kuopio, Finland. [63]Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden. [64]Queensland Institute of Medical Research (QIMR) Genome-Wide Association Study Collective, Brisbane, Queensland, Australia. [65]Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [66]Slone Epidemiology Center at Boston University, Boston, Massachusetts, USA. [67]Department of Medical Oncology, Fox Chase Cancer Center, Philadelphia,

Pennsylvania, USA. [68]Department of Internal Medicine, Oncology Section, "Hippokration" Hospital, Athens, Greece. [69]Departments of Genetics and Pathology, Lineberger Comprehensive Cancer Center, The University of North Carolina, Chapel Hill, North Carolina, USA. [70]Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK. [71]Department of Pathology, Keck School of Medicine and Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, California. [72]Laboratory of Cancer Genetics, Department of Clinical Genetics and Biocenter Oulu, University of Oulu, Oulu University Hospital, Oulu, Finland. [73]University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, USA. [74]Department of Biostatistics, Fox Chase Cancer Center, Philadelphia, Pennsylvania, USA. [75]Institute of Pathology, Städtisches Klinikum Karlsruhe, Karlsruhe, Germany. [76]National Institute for Health Research Comprehensive Biomedical Research Centre, Guy's & St. Thomas' National Health Service Foundation Trust, London, UK. [77]Division of Experimental Therapy and Molecular Pathology and Division of Epidemiology, Netherlands Cancer Institute–Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands. [78]Department of Pathology, University Hospital Heidelberg, Heidelberg, Germany. [79]Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Melbourne, Victoria, Australia. [80]Wellcome Trust Centre for Human Genetics and Oxford Biomedical Research Centre, University of Oxford, Oxford, UK. [81]Family Cancer Clinic, Netherlands Cancer Institut–Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands. [82]Multidisciplinary Breast Center, University Hospital Gasthuisberg, Leuven, Belgium. [83]Imperial College, London, UK. [84]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. Correspondence should be addressed to C.A.H. (haiman@usc.edu) or F.J.C. (couch.fergus@mayo.edu).

ONLINE METHODS

Study populations. Stage 1 included the studies of the AABC and the TNBCC. AABC includes 3,153 breast cancer cases (1,017 ER negative and 1,608 ER positive) and 2,831 controls from 9 studies (**Supplementary Table 1**). TNBCC is composed of 2,963 triple-negative breast cancer cases and 1,632 controls from 22 studies, GWAS genotype data from an additional 85 triple-negative breast cancer cases and 222 controls from HEBCS, and public GWAS genotype data from 3,448 controls from Cancer Genetic Markers of Susceptibility (CGEMS), Wellcome Trust Case-Control Consortium (WTCCC), KORA and QIMR (**Supplementary Table 1**). Replication studies include 8,365 breast cancer cases (1,359 ER negative and 5,255 ER positive) and 10,935 controls of the BPC3 and 6,182 breast cancer cases (933 ER negative and 3,434 ER positive) and 5,966 controls of the SEARCH. All participants in these studies have provided written informed consent for the research, and approval for the study was obtained from the ethics review boards at all the local institutions. A description of each participating study is provided in the **Supplementary Note**. Details regarding the measurement and collection of ER, PR and HER2 data for each study are provided in **Supplementary Table 4**.

Genotyping and quality control. Genotyping in AABC was conducted using the Illumina Human1M-Duo BeadChip. Of the 5,984 samples in the AABC Consortium (3,153 cases and 2,831 controls), we attempted genotyping of 5,932, removing samples ($n = 52$) with DNA concentrations <20 ng/μl. Following genotyping, we removed samples on the basis of the following exclusion criteria: (i) unknown replicates (≥98.9% genetically identical, $n = 29$); (ii) samples with call rates <95% after a second attempt ($n = 100$); (iii) samples with ≤5% African ancestry ($n = 36$) (discussed below); and (iv) samples with <15% mean heterozygosity of SNPs on the X chromosome and/or similar mean allele intensities of SNPs on the X and Y chromosomes ($n = 6$). In the analysis, we removed SNPs with <95% call rates ($n = 21,732$) or minor allele frequencies (MAFs) <1% ($n = 80,193$). The concordance rate for blinded duplicates was 99.95%. We also eliminated SNPs with genotyping concordance rates <98% based on the replicates ($n = 11,701$). The final analysis data set included 1,043,036 SNPs genotyped on 3,016 cases (988 ER negative, 1,520 ER positive and the remaining 508 cases with unknown ER status) and 2,745 controls, with an average SNP call rate of 99.7% and average sample call rate of 99.8%. The call rate for rs10069690 was very high in stage 1 (99.9%) and similar in cases (99.9%) and controls (99.9%). We also re-genotyped rs10069690 using TaqMan in 1,456 of the stage 1 samples; the concordance was 99.8%.

Genotyping for the TNBCC GWAS was conducted on 1,577 cases from ten studies (Australian Breast Cancer Tissue Bank (ABCTB), Bavarian Breast Cancer Cases and Controls (BBCC), Dana-Farber Cancer Institute, Fox Chase Cancer Center, GENICA, MARIE, Melbourne Collaborative Cohort Study (MCBCS), Prospective Study of Outcomes in Sporadic Versus Hereditary Breast Cancer (POSH), Sheffield Breast Cancer Study (SBCS)) using the Illumina 660-Quad SNP array. In addition, a set of MARIE cases ($n = 56$) were genotyped using the Illumina CNV370 SNP array. HEBCS cases ($n = 85$) were genotyped using the Illumina 550-Duo SNP array, bringing the total number of cases to 1,718. Population allele and genotype frequencies on healthy population controls ($n = 222$) genotyped on Illumina HumanHap 370CNV in the NordicDB, a Nordic pool and portal for genome-wide control data, were obtained from the Finnish Genome Center. GWAS data for public controls ($n = 3,448$) were generated using the following arrays: Illumina 660-Quad (QIMR), Illumina 550(v1) (CGEMS), Illumina 550 (KORA) and Illumina 1.2M (WTCCC). The combined total number of controls was 3,670. These GWAS data were independently evaluated by an iterative quality control process with the following exclusion criteria: MAF <0.01, call rate <95%, Hardy-Weinberg equilibrium (HWE) P value $< 1 \times 10^{-7}$ among controls and sample call rate <98%. In total, we excluded cases failing in the genotyping process ($n = 5$), previously unknown replicates ($n = 2$) and samples with call rates <98% ($n = 83$), samples that failed sex check ($n = 10$), cases identified as non–triple-negative breast cancer ($n = 20$) and related samples ($n = 27$). We removed SNPs with <95% call rates or MAF <5%. Because a number of our samples were genotyped at different locations, we removed SNPs if there was a difference of >0.10 between the study allele frequency and the median frequency across all studies. Eigensoft was used to evaluate confounding due to population stratification. We removed 101 subjects that did not cluster with the CEU HapMap phase 2 samples, resulting in 1,562 cases and 3,578 controls in the GWAS analyses.

Re-genotyping of rs10069690 on 2,963 TNBCC cases and 1,632 study-specific controls was conducted using a single multiplex on the iPLEX Mass Array platform (Sequenom). We removed 31 cases from MCCS that were part of the MCCS replication sample in BPC3. SNPs and samples evaluated on the iPLEX were excluded on the basis of the following criteria: SNP call rate was <97%, HWE P value < 0.001 among controls and sample call rate <95% (for the overall experiment). The final data set of 2,849 cases and 1,602 controls for rs10069690 had a SNP call rate >99% and HWE P value of 0.53 in controls. The concordance rate, on the basis of blinded duplicates, was 100%. The concordance of the imputed ($R^2 = 0.55$) versus the genotyped data was 70%.

Replication genotyping. In BPC3, genotyping of rs10069690 was performed by TaqMan in five laboratories (Cancer Prevention Study II Nutrition Cohort (CPS2) and Multiethnic Cohort (MEC) at the University of Southern California; the Nurses' Health Study (NHS) and the Women's Health Study (WHS) at Harvard University; EPIC at the German Cancer Research Center in Heidelberg; MCCS at Melbourne University and PLCO at the NCI Core Genotyping Facility). Genotyping in SEARCH was performed by TaqMan at Cambridge University. Genotype call rates were >92% in cases and controls, and concordance of blinded duplicates was ≥99.5% in all studies. The P value for HWE in controls was >0.01 in all studies except WHS ($P = 0.007$).

DNA sequencing. Bi-directional sequencing of the 15 coding exons of *TERT* was performed in 96 African-American women using the ABI 3730xl DNA Analyzer (Applied Biosystems). Sequencing purification was performed using DyeDX 96 columns (Qiagen) following their standard protocol, and PolyPhred was used for analyzing sequence traces (http://droog.gs.washington.edu/polyphred/). More than 95% of samples were sequenced for each exon except for exon 15 ($n = 74$) and 16 ($n = 86$). Exon 1 could not be sequenced, as well as 112bp (9%) of exon 2, because of high GC content.

Statistical analysis. In AABC, we tested for gene dosage effects through a one-degree-of-freedom likelihood ratio test in models adjusted for age, study and genetic ancestry eigenvectors 1–10. OR and 95% CI were estimated using unconditional logistic regression. In TNBCC, unconditional logistic regression was used to assess single SNP associations also assuming a log-additive model, adjusting for country and the first two principal components. For the analyses of the iPLEX genotyping data on rs10069690, unconditional logistic regression was used assuming a log-additive model and adjusting for age and country.

In both AABC and TNBCC, phased haplotype data from the founders of the CEU and YRI HapMap Phase 2 samples (build 21) were used to infer linkage disequilibrium patterns in order to impute untyped markers. For both studies, genome-wide imputation was carried out using the software MACH. Filtered from the analysis were SNPs with $R^2 < 0.3$.

We conducted a fixed-effect meta-analysis of AABC and TNBCC using the inverse variance weighted method. The number of SNPs available for meta-analysis from AABC and TNBCC was 3,055,415 and 2,134,490 respectively. The union of these two data sets (3,154,485 SNPs) was meta-analyzed using the program METAL.

SNP rs10069690 was analyzed in BPC3 and SEARCH using logistic regression controlling for age and study or country (BPC3 only). The meta-analysis of rs10069690 from AABC, TNBCC, BPC3 and SEARCH was conducted using the inverse variance weighted method. Testing for heterogeneity by study was evaluated using the Q statistic. Case-only analyses were performed to test for differences in the association by tumor subtypes.

We estimated the relative risk in African-ancestry women compared to women of European descent that could plausibly be attributable to the association with rs10069690. The calculation of the attributable racial/ethnic ratio (ARR) is $\mathrm{ARR} = \sum_{i=0}^{2} f_A OR^i / \sum_{i=0}^{2} f_E OR^i$, where $f_A(i)$ is the probability in the African American women of carrying $i = 0$, 1 or 2 copies of the risk variant and $f_E(i)$ is the same probability for European women. The per-allele OR is for triple-negative disease from the meta-analysis (1.25), and both a log linear model for risk and Hardy-Weinberg equilibrium for the alleles (in both populations) is assumed. A confidence interval for the ARR is calculated from the confidence interval for the OR in the meta-analysis.

Human Molecular Genetics, 2011, Vol. 20, No. 22 ***4491–4503***
doi:10.1093/hmg/ddr367
Advance Access published on August 18, 2011

Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans

Fang Chen[1], Gary K. Chen[1], Robert C. Millikan[3], Esther M. John[4,5], Christine B. Ambrosone[6], Leslie Bernstein[7], Wei Zheng[8], Jennifer J. Hu[9], Regina G. Ziegler[10], Sandra L. Deming[8], Elisa V. Bandera[11], Sarah Nyante[3], Julie R. Palmer[12], Timothy R. Rebbeck[13], Sue A. Ingles[1], Michael F. Press[2], Jorge L. Rodriguez-Gil[9], Stephen J. Chanock[10], Loïc Le Marchand[14], Laurence N. Kolonel[14], Brian E. Henderson[1], Daniel O. Stram[1] and Christopher A. Haiman[1,*]

[1]Department of Preventive Medicine and [2]Department of Pathology, Keck School of Medicine and Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA, [3]Department of Epidemiology, Gillings School of Global Public Health, and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA, [4]Northern California Cancer Center, Fremont, CA, USA, [5]Stanford University School of Medicine and Stanford Cancer Center, Stanford, CA, USA, [6]Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY, USA, [7]Division of Cancer Etiology, Department of Population Science, Beckman Research Institute, City of Hope, CA, USA, [8]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, and Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA, [9]Department of Epidemiology and Public Health, and Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL, USA, [10]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA, [11]The Cancer Institute of New Jersey, New Brunswick, NJ, USA, [12]Slone Epidemiology Center at Boston University, Boston, MA, USA, [13]University of Pennsylvania School of Medicine, Philadelphia, PA, USA and [14]Epidemiology Program, Cancer Research Center, University of Hawaii, Honolulu, HI, USA

Received May 4, 2011; Revised July 15, 2011; Accepted August 15, 2011

Genome-wide association studies (GWAS) have revealed 19 common genetic variants that are associated with breast cancer risk. Testing of the index signals found through GWAS and fine-mapping of each locus in diverse populations will be necessary for characterizing the role of these risk regions in contributing to inherited susceptibility. In this large study of breast cancer in African-American women (3016 cases and 2745 controls), we tested the 19 known risk variants identified by GWAS and replicated associations ($P < 0.05$) with only 4 variants. Through fine-mapping, we identified markers in four regions that better capture the association with breast cancer risk in African Americans as defined by the index signal (2q35, 5q11, 10q26 and 19p13). We also identified statistically significant associations with markers in four separate regions (8q24, 10q22, 11q13 and 16q12) that are independent of the index signals and may represent putative novel risk variants. In aggregate, the more informative markers found in the study enhance the association of these risk regions with breast cancer in African Americans [per allele odds ratio (OR) = 1.18, $P = 2.8 \times 10^{-24}$ versus OR = 1.04, $P = 6.1 \times 10^{-5}$]. In this detailed analysis of the known breast cancer risk loci, we have validated and improved upon markers of risk that better characterize their association with breast cancer in women of African ancestry.

*To whom correspondence should be addressed at: Harlyne Norris Research Tower, 1450 Biggy Street, Room 1504, Los Angeles, CA 90033, USA, Tel: +1 3234427755; Fax: +1 3234427749; E-mail: haiman@usc.edu

INTRODUCTION

Genome-wide association studies (GWAS) of breast cancer have identified at least 19 chromosomal regions that harbor common alleles that contribute to genetic susceptibility (1–10). These discoveries have allowed for improved understanding of genetic risk for this common cancer, although it is argued that many more markers will be needed to elucidate disease heritability, and in the clinical setting for disease prediction (11–13). Except for the breast cancer risk locus at 6q25 identified in a GWAS of Chinese women, the risk loci for breast cancer have been revealed in studies in women of European ancestry. We have recently shown in a multiethnic study that a summary score comprised of the index variants at many of these risk loci is statistically significantly associated with breast cancer risk in multiple populations [odds ratio (OR) per allele of >1.10], but not in African Americans (14). Similar studies in African-American women have also reported lack of replication with many of the reported index signals (15–17). Limited statistical power of these initial reports as well as variation in both allele frequency and patterns of linkage disequilibrium (LD) across populations may be contributing factors as to why the associations found in the GWAS populations may not be generalizable to African Americans. Association testing of the risk variants as well as fine-mapping in a sufficiently large sample of African Americans will be needed to identify and localize the subset of markers that best define risk of the functional allele(s) within known risk regions.

In the present study, we tested common genetic variation at the breast cancer risk loci identified in women of European and Asian descent in a large sample comprised of 3016 African-American breast cancer cases and 2745 controls to identify markers of risk that are relevant to this population. More specifically, we examined the index variants and conducted fine-mapping of the locus to both improve the current set of risk markers in African Americans as well as to identify new risk variants for breast cancer. We then applied this information to model breast cancer risk in African-American women in an attempt to characterize the spectrum of genetic risk in this population defined by common variants at the known risk loci.

RESULTS

The ages of cases and controls ranged from 22 to 87 years and 23 to 86 years, respectively, with cases and controls having similar mean ages (55 and 58 years, respectively; Supplementary Material, Table S1).

We tested 19 validated breast cancer risk variants (referred to as 'index variants' throughout the paper) at 1p11, 2q35, 3p24, 5p12, 5q11, 6q25, 8q24, 9p21, 9q31, 10p15, 10q21, 10q22, 10q26, 11p15, 11q13, 14q24, 16q12, 17q23 and 19p13 in models adjusted for age, study, global ancestry (the first 10 eigenvectors) and local ancestry (Table 1; Supplementary Material, Table S2) (1–10); 17 SNPs were directly genotyped, whereas 2 were imputed ($r^2 > 0.98$; see Materials and Methods). All 19 variants were common (≥0.05) in African Americans, with 11 variants being more common in Europeans than in African Americans (Table 1, Fig. 1). In previous GWAS, the index signals had modest ORs (1.05–1.29 per copy of the risk allele) and our sample size provided ≥70% statistical power to detect the reported effects for 12 of the 19 variants (at $P < 0.05$; Supplementary Material, Table S2).

We observed positive associations with 11 of the 19 variants (OR > 1); however, only 4 were statistically significant ($P < 0.05$ at 2q35, 9q31, 10q26 and 19p13; Table 1). Of the 15 variants that were not replicated at $P < 0.05$, statistical power was <70% for only 7 of the variants. Although power was more limited, we also evaluated associations by estrogen receptor (ER) status as some risk variants have been found to be more strongly associated with ER-positive (ER+) or ER-negative (ER−) breast cancer (2,18). We observed positive associations with 12 variants (2 at $P < 0.05$) for ER+ disease ($n = 1520$) and with 9 variants for ER− (3 at $P < 0.05$; $n = 988$) (Supplementary Material, Table S3). For only one variant did we observe statistically significant risk heterogeneity by ER status (rs13387042 at 2q35, $P = 0.013$) (Supplementary Material, Table S3).

Local ancestry was included in all models, as it was found to be associated with breast cancer risk in many regions (Supplementary Material, Table S4). We observed nominally significant associations between local ancestry and overall breast cancer, ER+ or ER− disease risk at 5 loci (5p12, 6q25, 8q24, 10p15, 10q26). The most statistically significant association was between European ancestry and ER+ breast cancer risk at 6q25 (OR per European allele chromosome = 1.19, $P = 6.2 \times 10^{-3}$). The inverse association observed between European ancestry and ER+ disease risk at 10q26 (OR per European chromosome = 0.85, $P = 0.011$) is consistent with previous reports of over-representation of African ancestry at this locus in many of these same cases (19,20).

Aside from statistical power, the lack of a statistically significant association with an index variant (OR > 1 and $P < 0.05$) suggests that the particular variant revealed in the GWAS populations may not be adequately correlated with the biologically relevant allele in African Americans. In an attempt to identify a better genetic marker of risk in African Americans, we conducted fine-mapping across all risk regions, using genotyped SNPs on the Illumina 1M array and imputed SNPs to Phase 2 HapMap populations (see Materials and Methods). If a marker associated with risk in African Americans represents the same signal as that reported in the initial GWAS, then it should be correlated to some degree with the index signal in the GWAS population. Using HapMap data for the populations in which the risk variant was identified [Utah residents with ancestry from northern and western Europe (CEU), or Han Chinese in Beijing, China (CHB)], we catalogued and tested all SNPs that were correlated ($r^2 \geq 0.2$) with the index signal (within 250 kb), applying an α_a of 3.2×10^{-3} which was estimated to be 0.05 divided by the average number of tags needed to capture ($r^2 \geq 0.8$) the common risk alleles correlated with the index allele in each region in the Yoruba HapMap population [in Ibadan, Nigeria (YRI); Supplementary Material, Table S5]. We also tested for novel independent associations, focusing on SNPs that were uncorrelated with the index signal in the initial GWAS populations. Here, we applied a Bonferroni correction for defining novel associations as statistically

Table 1. Associations with common variants at known breast cancer risk regions in African Americans

Chr., nearest genes	Index SNP from GWAS (3016 cases, 2745 controls)		Best marker in African Americans (3016 cases, 2745 controls)		
	Marker, position, alleles (risk/reference)	RAF in CEU/AA[a], OR (95% CI), P_{trend}	Marker, position, alleles (risk/reference)	RAF in CEU/AA[a], OR (95% CI), P_{trend} from stepwise analysis	r^2 with index in CEU/YRI[b]
1p11	rs11249433, 120982136, G/A	0.43/0.13, 1.01 (0.90–1.14), 0.84			
2q35	rs13387042, 217614077, A/G	0.56/0.72, 1.12 (1.03–1.21), 7.5×10^{-3}	rs13000023[c], 217632639, G/A	0.82/0.83, 1.20 (1.09–1.33), 5.8×10^{-4}	0.35/0.53
3p24, *NEK10*	rs4973768, 27391017, T/C	0.44/0.36, 1.04 (0.96–1.13), 0.32			
5p12, *MRPS30*	rs4415084, 44698272, T/C	0.38/0.63, 1.02 (0.95–1.11), 0.54			
5q11, *MAP3K1*	rs889312, 56067641, C/A	0.30/0.34, 1.07 (0.99–1.18), 0.084	rs16886165, 56058840, G/T	0.16/0.31, 1.15 (1.06–1.25), 6.5×10^{-4}	0.40/<0.01
6q25, *C6orf97*	rs2046210[c,d], 151990059, A/G	0.38/0.60, 1.00 (0.93–1.09), 0.88			
8q24	rs13281615, 128424800, G/A	0.45/0.43, 1.05 (0.97–1.13), 0.20			
9p21, *CDKN2B*	rs1011970, 22052134, T/G	0.17/0.33, 1.05 (0.97–1.14), 0.24			
9q31	rs865686, 109928199, T/G	0.61/0.52, 1.08 (1.01–1.17), 0.034			
10p15, *ANKRD16*	rs2380205, 5926740, C/T	0.52/0.42, 0.98 (0.91–1.06), 0.60			
10q21, *ZNF365*	rs10995190, 63948688, G/A	0.87/0.83, 0.97 (0.88–1.08), 0.57			
10q22, *ZMIZ1*	rs704010, 80511154, T/C	0.43/0.11, 0.99 (0.87–1.12), 0.83	rs12355688, 80725632, T/C	0.090/0.20, 1.24 (1.13–1.36), 6.8×10^{-6}	<0.01/<0.01
10q26, *FGFR2*	rs2981582, 123342307, A/G	0.46/0.46, 1.11 (1.03–1.19), 8.6×10^{-3}	rs2981578[c], 123330301, C/T	0.46/0.81, 1.24 (1.11–1.39), 1.7×10^{-4}	0.66/0.059
11p15, *LSP1*	rs3817198, 1865582, C/T	0.33/0.17, 0.98 (0.88–1.08), 0.63			
11q13	rs614367, 69037945, T/C	0.18/0.13, 0.96 (0.86–1.07), 0.45	rs609275[c], 69112096, C/T	1.00/0.59, 1.20 (1.11–1.30), 1.0×10^{-5}	NA/<0.01
14q24, *RAD51L1*	rs999737, 68104435, T/C	0.26/0.051, 0.98 (0.82–1.17), 0.80			
16q12, *TNRC9*	rs3803662, 51143842, A/G	0.25/0.51, 0.99 (0.92–1.08), 0.85	rs3112572, 51157948, A/G	0.020/0.20, 1.18 (1.08–1.30), 3.9×10^{-4}	0.038/0.31
17q23, *COX11*	rs6504950[c], 50411470, G/A	0.70/0.66, 1.05 (0.97–1.14), 0.19			
19p13, *ANKLE1*	rs2363956, 17255124, T/G	0.45/0.49, 1.14 (1.05–1.22), 8.0×10^{-4}	rs3745185, 17245267, G/A	0.52/0.75, 1.20 (1.10–1.32), 3.7×10^{-5}	0.57/0.19

SNP positions are based on NCBI build 36.
ORs are per allele odds ratios adjusted for age, study, the first 10 eigenvectors and local ancestry at each risk locus.
P_{trend} values are based on test of trend (1 d.f.).
[a]RAF, risk allele frequencies in the original GWAS population (HapMap CEU, or CHB for rs2046210) and AA (African American) controls in this study. Risk allele is the allele associated with increased risk in previous GWAS.
[b]Pairwise correlations (r^2) between the index signal and the best marker are from the CEU (CHB for rs2046210) and YRI populations in the 1000 Genomes Project (March 2010 release).
[c]Imputed SNPs.
[d]Index signal reported in Han Chinese. RAFs based on HapMap CHB and r^2 based on CHB in the 1000 Genomes Project (March 2010 release).

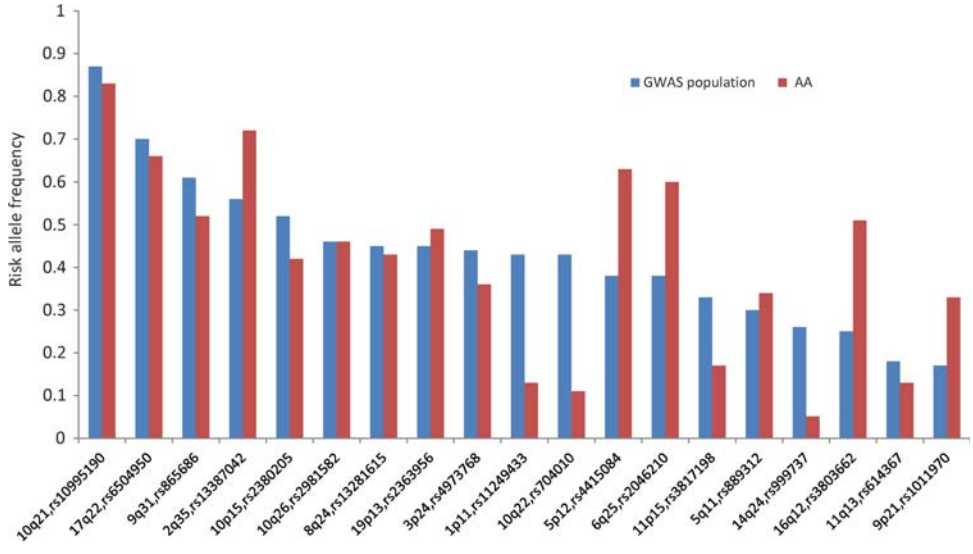

Figure 1. RAFs in Europeans and African Americans. The distribution of RAFs for the 19 index SNPs (from Table 1) in HapMap CEU (CHB for rs2046210) and African Americans (AA). The variants are sorted based on the RAF in the GWAS population.

significant in each region, with α_b estimated to be 0.05 divided by the total number of tags needed to capture ($r^2 \geq 0.8$) all common risk alleles in the 19 regions in the YRI population ($\alpha_b = 1.0 \times 10^{-5}$; similar to the genome-wide-type correction of 5×10^{-8}, which accounts for the number of tags needed to capture all common alleles in the genome; Supplementary Material, Table S5). For each region, stepwise logistic regression was used with SNPs kept in the final model based on α_a or α_b (results for each model are provided in Supplementary Material, Tables S6 and S7). These procedures were applied to all cases and controls as well as in hypothesis-generating analyses stratified by ER status.

At nine loci, we detected variants that were statistically significantly associated with breast cancer risk in African Americans. These regions include 9q31, where the sole marker of risk was the index signal (rs865686: OR = 1.08, $P = 0.034$; Table 1). In five of these nine regions, the index marker itself was not statistically significantly associated with disease risk. Through fine-mapping, we revealed markers in four regions that were more significantly associated with risk than the index signal (>1 order of magnitude change in the *P*-value) and are likely to capture the same signal (2q35, 5q11, 10q26 and 19p13). We also identified markers in four regions that are not correlated with the index signal in the GWAS populations (8q24, 10q22, 11q13 and 16q12) and may represent putative novel risk variants, with one being specific for ER+ disease (8q24) (Table 1, Fig. 2 and Supplementary Material, Table S8). These regions are discussed in what follows.

Risk variants that better define the index signal in African Americans

2q35. The index signal at 2q35 was statistically significantly associated with risk of overall breast cancer (rs13387042: OR = 1.12, $P = 7.5 \times 10^{-3}$; Table 1) and ER+ disease (OR = 1.22, $P = 2.6 \times 10^{-4}$; Supplementary Material, Table S3). However, we found stronger associations with two markers that are each modestly correlated with the index signal in CEU and YRI: rs13000023 with overall breast cancer (OR = 1.20, $P = 5.8 \times 10^{-4}$) and rs12998806 with ER+ disease (OR = 1.39, $P = 3.3 \times 10^{-6}$) (Table 1 and Supplementary Material, Table S8). As shown in Supplementary Material, Figure S1, the signal in this region appeared limited to ER+ breast cancer, which is consistent with the initial report of this risk locus (2) but not with subsequent large-scale replication efforts in European populations (21).

5q11. We found a positive non-significant association with the index signal at 5q11, which is located 79 kb centromeric of the *MAP3K1* gene (rs889312: OR = 1.07, $P = 0.084$; Table 1). Fine-mapping revealed statistically significant associations with markers, rs16886165 for overall breast cancer (OR = 1.15, $P = 6.5 \times 10^{-4}$) and rs832529 for ER− disease (OR = 1.22, $P = 1.3 \times 10^{-3}$; Table 1 and Supplementary Material, Table S8). These SNPs show greater correlation with the index signal in Europeans (CEU, $r^2 = 0.40$ and 0.46) than in Africans (YRI, $r^2 < 0.01$ and $r^2 = 0.09$), which suggests that they may be better markers of the biologically functional variant in African Americans (Table 1, Fig. 2).

10q26. Both the index signal, rs2981582 (OR = 1.11, $P = 8.6 \times 10^{-3}$; Table 1) and rs2981578, which was identified previously through fine-mapping in African Americans (which some of these studies contributed to) (22), were statistically significantly associated with risk (OR = 1.24, $P = 1.7 \times 10^{-4}$, Table 1). Variant rs2981578 was the most strongly associated marker in the region for overall breast cancer and for ER+ disease, which is consistent with previous reports of variation in this region being more strongly associated with ER+ breast cancer (Supplementary Material, Table S8) (18). In fine-mapping the locus, we observed a suggestive association with a correlated marker and ER− disease (rs2912774: OR = 1.19, $P = 2.1 \times 10^{-3}$; Supplementary Material, Table

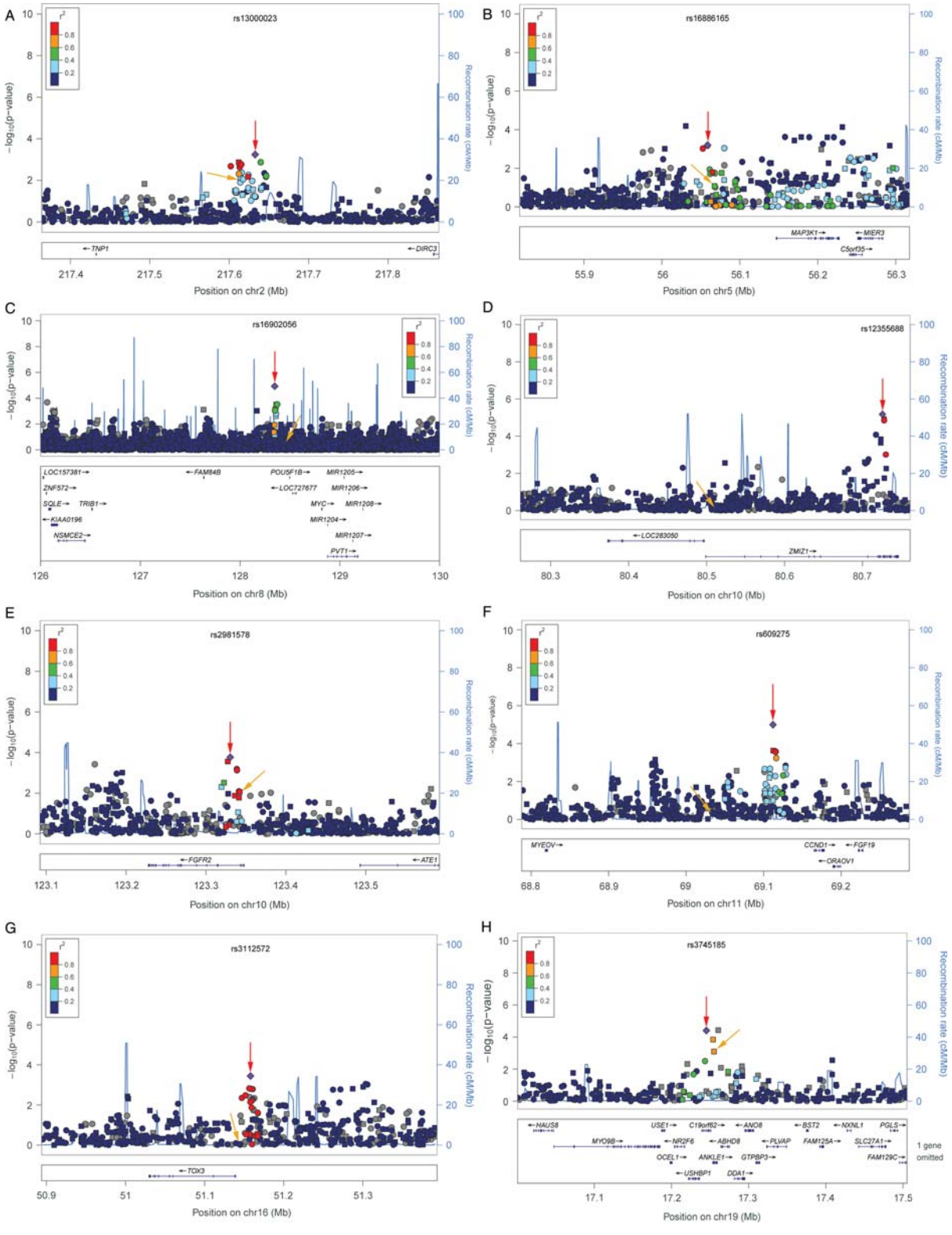

Figure 2. −Log *P* plots for common alleles at eight breast cancer risk loci in African Americans. −Log *P*-values for risk-associated alleles in African Americans from logistic regression models adjusted for age, study, global ancestry (the first 10 eigenvectors) and local ancestry. *P*-values are for overall breast cancer risk except for 8q24, which is for ER+ breast cancer. Pairwise correlations (r^2) in the HapMap CEU population are shown in relation to markers identified through fine-mapping in African Americans (diamond), except for 11q13, where r^2 is shown in HapMap YRI as the marker is monomorphic in CEU. Squares denote genotyped SNPs; circles, imputed SNPs. Gray squares and circles denote that r^2 cannot be estimated (not in HapMap or monomorphic in CEU). Red arrows denote markers identified in African Americans; yellow arrows, GWAS index variants. Each panel shows a −log *P* plot for common alleles for regions: (**A**) 2q35; (**B**) 5q11; (**C**) 8q24; (**D**) 10q22; (**E**)10q26; (**F**) 11q13; (**G**) 16q12; (**H**) 19p13. The plots were generated using LocusZoom (55).

S8); however, the association was also noted with ER+ disease (OR = 1.10, $P = 0.041$; Supplementary Material, Table S9) and is likely to capture the same signal as rs2981578.

19p13. 19p13 was the first risk locus reported to harbor a variant that may be specific for ER− disease (9). In African Americans, the index variant was statistically significantly associated with risk of overall breast cancer (rs2363956: OR = 1.14, $P = 8.0 \times 10^{-4}$), as well as ER+ (OR = 1.12, $P = 0.016$) and ER− disease (OR = 1.14, $P = 0.018$; Table 1 and Supplementary Material, Table S3). The most significant association in the region for overall breast cancer and ER+ disease was with rs3745185 ($P = 3.7 \times 10^{-5}$ and $P = 8.2 \times 10^{-4}$, respectively), which is likely to capture the same functional variant ($r^2 = 0.57$ in CEU and 0.19 in YRI; Table 1 and Supplementary Material, Table S8). The most significant marker for ER− breast cancer was correlated with both rs2363956 and rs3745185 (rs11668840: OR = 1.25, $P = 5.1 \times 10^{-5}$; Supplementary Material, Tables S8 and S10).

Novel risk-associated markers at breast cancer susceptibility loci

8q24. Given the importance of the 8q24 locus in cancer, we conducted association testing across the entire cancer risk region (126.0–130.0 Mb) (23–25). The index signal (rs13281615) was not statistically significantly associated with risk in African Americans (Table 1 and Supplementary Material, Table S3), nor did we identify significant associations with correlated SNPs. However, we did detect a significant association with rs16902056 and ER+ breast cancer [risk allele frequency (RAF) 0.95; $P = 6.7 \times 10^{-6}$; ER−: $P = 0.66$; Supplementary Material, Table S8]. This SNP is located 78 kb centromeric of the index variant and is not correlated with the index variant ($r^2 < 0.01$ in CEU and $r^2 = 0.027$ in YRI). No statistically significant associations were observed with variants found previously in association with cancers of the bladder and ovary, or leukemia (rs9642880: OR = 1.03, $P = 0.58$; rs10088218: OR = 1.02, $P = 0.62$; rs2456449: OR = 1.07, $P = 0.14$) (26–28). Of the known risk variants for prostate cancer (29–35), we found a single nominally significant ($P < 0.05$) association with the same risk allele of rs1016343 ($P = 0.015$) which is located >260 kb centromeric of the breast cancer risk region and is not correlated with rs13281615 or rs16902056.

10q22. We observed no association with the index signal at 10q22 (rs704010) which is located in intron 1 of the gene *ZMIZ1*, or with any correlated markers. However, we did detect strong evidence of a second signal located 215 kb telomeric in intron 12 of the gene *ZMIZ1* (rs12355688: OR = 1.24, $P = 6.8 \times 10^{-6}$). As is shown in Table 1 and Figure 2, this putative novel risk variant is not correlated with the index variant in the CEU or YRI populations ($r^2 < 0.01$).

11q13. No positive association was noted with the index variant at 11q13. However, we did detect evidence of a second independent signal (rs609275: OR = 1.20, $P = 1.0 \times 10^{-5}$), located 74 kb telomeric, and 53 kb centromeric of *CCND1*. The variant is monomorphic and uncorrelated with the index signal in the CEU population; and r^2 with the index signal in the YRI population is <0.01 (Table 1).

16q12. As in previous studies of African Americans, we were not able to replicate the association signal defined by the index variant rs3803662 (Table 1) (15,16). A recent study of African Americans reported a suggestive association with SNP rs3104746, which is located 15 kb telomeric of rs3803662 (16). This SNP has a minor allele frequency (MAF) of 0.04 in the HapMap CEU population, 0.19 in our African-American controls, and is modestly correlated with rs3803662 in Africans ($r^2 = 0.31$ in YRI), but not in Europeans ($r^2 = 0.038$; Supplementary Material, Table S10). Fine-mapping around this putative signal revealed a perfect proxy ($r^2 = 1$) for rs3104746, rs3112572, which is significantly associated with breast cancer risk in African Americans (OR = 1.18, $P = 3.9 \times 10^{-4}$), with the association noted to be stronger for ER+ breast cancer (OR = 1.27, $P = 3.1 \times 10^{-5}$; Table 1 and Supplementary Material, Table S8).

For index SNPs found to be nominally associated with breast cancer risk, as well as risk-associated markers identified through fine-mapping, we also tested for associations by genotype. Results from the genotype-specific model were consistent with log-additive associations (Supplementary Material, Tables S9 and S11). Risk variants at 2q35 and 8q24 were also found to have significantly stronger associations with ER+ breast cancer than ER− disease (Supplementary Material, Table S7), which is consistent with previous studies (2,18).

We observed no statistically significant associations with common variation at 10 risk loci on 1p11, 3p24, 5p12, 6q25, 9p21, 10p15, 10q21, 11p15, 14q24 and 17q23 (Supplementary Material, Fig. S2). We also could not replicate the association with the recently identified SNP rs9397435 at 6q25 that was found through fine-mapping in European, African and Asian population samples (17) ($P = 0.26$ for overall breast cancer, $P = 0.71$ for ER+ and $P = 0.36$ for ER− tumor subtypes). Neither could we replicate the association with SNP rs4784227 at 16q12, which was identified by a recent multistage GWAS in women of Asian ancestry (36) in our African-American sample ($P = 0.51$ overall, $P = 0.35$ and $P = 0.65$ for ER+ and ER− subtypes, respectively).

Risk modeling

We next estimated the cumulative effect of all breast cancer risk variants, and compared a summary risk score comprised of unweighted counts of all GWAS-reported risk variants with a risk score that included variants we identified as being associated with risk in African Americans (Table 2). Using the 19 index signals from GWAS (see Materials and Methods), the risk per allele was 1.04 [95% confidence interval (CI) 1.02–1.06; $P = 6.1 \times 10^{-5}$], and individuals in the top quintile of the risk allele distribution were at 1.4-fold greater risk ($P = 7.4 \times 10^{-5}$) of breast cancer compared with those in the lowest quintile (Table 2). As expected, the risk score was improved when utilizing the markers that we identified at the known risk loci as being more relevant to African Americans (eight markers for overall breast cancer: 2q35, 5q11, 9q31, 10q22, 10q26, 11q13, 16q12 and 19p13;

Table 2. The association of the total risk score with breast cancer risk in African Americans

	Index markers from GWAS (19 markers)	Risk-associated best markers in African Americans[a] (8 markers)		
Mean number of risk alleles in controls (range)	15.7 (6–25)	8.4 (3–14)		
Per allele OR (95% CI)	1.04 (1.02–1.06)	1.18 (1.14–1.22)		
P_{trend}	6.1×10^{-5}	2.8×10^{-24}		
			First-degree family history negative[b]	First-degree family history positive[b]
Subjects, *n* cases/*n* controls	3016/2745	3016/2745	2387/2349	554/303
Risk quintiles[c]				
Q1				
n cases/*n* controls	536/549	352/462	281/387	62/57
OR (95%CI)	1.00 (ref.)	1.00 (ref.)	1.00 (ref.)	1.58 (1.06–2.37)
P-value	—	—	—	0.025
Q2				
n cases/*n* controls	722/742	430/505	344/437	77/47
OR (95% CI)	0.99 (0.84–1.16)	1.17 (0.96–1.42)	1.15 (0.93–1.43)	2.18 (1.46–3.26)
P-value	0.88	0.11	0.18	1.5×10^{-4}
Q3				
n cases/*n* controls	435/382	632/625	503/549	115/53
OR (95%CI)	1.15 (0.96–1.39)	1.37 (1.14–1.64)	1.31 (1.07–1.60)	3.14 (2.17–4.53)
P-value	0.14	7.2×10^{-4}	8.0×10^{-3}	1.2×10^{-9}
Q4				
n cases/*n* controls	753/669	665/566	517/476	132/75
OR (95%CI)	1.16 (0.98–1.36)	1.56 (1.30–1.87)	1.51 (1.24–1.86)	2.52 (1.81–3.52)
P-value	0.080	2.3×10^{-6}	6.2×10^{-5}	4.0×10^{-8}
Q5				
n cases/*n* controls	570/403	937/587	742/500	168/71
OR (95%CI)	1.44 (1.20–1.72)	2.16 (1.80–2.58)	2.11 (1.73–2.56)	3.44 (2.47–4.77)
P-value	7.4×10^{-5}	3.6×10^{-17}	1.3×10^{-13}	9.9×10^{-14}

ORs are adjusted for age, study and the first 10 eigenvectors.
P_{trend} values are based on test of trend (1 d.f.).
[a]The most significant markers from the stepwise analysis for overall breast cancer in each region from Table 1.
[b]Information about first-degree family history of breast cancer is available on 97.5% of cases and 96.6% of controls.
[c]Based on distribution in controls (cut points for index markers aggregate: 13.3, 15, 16, 18; cut points for best markers aggregate: 7, 8, 9, 10).

OR = 1.18; 95% CI 1.14–1.22; $P = 2.8 \times 10^{-24}$), with risk for those in the top quartile being 2.2 times that observed in the lowest quintile ($P = 3.6 \times 10^{-17}$). This score was significantly associated with risk of both ER+ (OR = 1.20, $P = 1.7 \times 10^{-19}$) and ER− (OR = 1.15, $P = 2.8 \times 10^{-9}$) disease ($P_{het} = 0.12$) (Supplementary Material, Table S12).

Stratifying by first-degree family history of breast cancer differentiated risk further with those with a family history and in the top quintile of the risk score distribution (4% of the population) having a 3.4-fold greater risk ($P = 9.9 \times 10^{-14}$) compared with those without a family history and in the lowest quintile of the risk score (Table 2).

In hypothesis-generating analyses, we also developed risk scores for ER+ and ER− breast tumor subtypes, utilizing the most informative markers revealed through fine-mapping of each phenotype. These phenotype-specific scores were highly significant (ER+: OR = 1.30, $P = 6.0 \times 10^{-18}$; ER−: OR = 1.20, $P = 2.3 \times 10^{-10}$) with statistically significant heterogeneity noted when the scores were applied to the other subtype ($P_{het} = 1.7 \times 10^{-5}$ and 5.0×10^{-3} for ER+ and ER− scores, respectively) (Supplementary Material, Table S12).

DISCUSSION

In this large study of breast cancer in African-American women, we were able to replicate associations with 4 of the 19 index variants (at $P < 0.05$). Through fine-mapping, we observed that overall breast cancer risk was statistically significantly associated with markers in four regions which are likely to capture the GWAS-reported signal and to serve as better markers of the functional allele and risk in African Americans. We also detected putative novel associations that are independent of the index signals in three regions for overall breast cancer (10q22, 11q13 and 16q12) and in one region for ER+ disease (8q24). In 10 of the risk regions, however, we were not able to replicate the GWAS index signals, nor did we detect statistically significant associations of common SNPs with breast cancer risk at the levels of statistical significance we set for fine-mapping. The inability to replicate associations with the index signals despite adequate statistical power (>70% power for 12 of 19 variants) suggests that they are unlikely to be functional variants or capture the functional variants as efficiently in this population. Our ability to find associated markers in five regions where index signals were not significantly associated with risk also demonstrates the value of testing common variation at GWAS-identified risk loci in additional populations (14,16,17,22,37,38).

In four regions, we observed risk markers that are correlated with, and in the same LD block as the index markers in CEU (rs13000023 at 2q35, rs16886165 at 5q11, rs2981578 at 10q26 and rs3745185 at 19p13). It is likely that these risk markers capture the same signal as defined by the index markers

based on the r^2 values between these markers and the index markers ($\geq$0.35). We cannot rule out the possibility, though, that some of them may represent a second, independent signal in the same region.

In the four regions where we observed independent signals, the risk alleles (rs16902056 at 8q24, rs12355688 at 10q22, rs609275 at 11q13 and rs3112572 at 16q12) were uncorrelated with, and not in, the same LD block as the index variant in Europeans (CEU, $r^2 < 0.04$)) (distances from the index signal ranged from 14 kb at 16q12 to 215 kb at 10q22) (Supplementary Material, Fig. S3). Therefore, these variants are likely to pick up a novel signal independent of the index signal. However, because of different LD patterns in European and African ancestry populations, they may each mark the same functional variant, and if the functional variant is less common it may not be well captured by either common marker alone. At 10q22, both the index SNP and the novel variant are located within introns of the *ZMIZ1* gene. *ZMIZ1* encodes zinc finger MIZ-type containing 1, which regulates the activity of various transcription factors (39–41). At 11q13, rs609275 lies 74 kb telomeric of the index signal and in closer proximity to a number of candidate genes, including *CCND1* (encoding cyclin D1, a protein crucial for cell-cycle control), *ORAOV1* (encoding oral cancer overexpressed 1) and *FGF19* (encoding fibroblast growth factor 19). The association at 16q12 confirms the findings of a previous, smaller study of African Americans (16), and is consistent with a previous fine-mapping study suggesting that African Americans may harbor a separate causal variant in this region (42). Whether this variant is influencing the same genes/pathways as the index variant rs3803662 is not known; however, the stronger associations noted for both variants with ER+ disease (2,18) suggest that they may affect the same biological process.

Notably, at region 19p13, which was originally reported in association with ER− breast cancer (9), the index signal was statistically significantly associated with both ER+ and ER− subtypes in African Americans. In addition, we found a stronger marker in this region (rs3745185) for ER+ as well as overall breast cancer risk (Table 1 and Supplementary Material, Table S8). We also found stronger associations with ER+ than ER− disease for variants in many regions, including 2q35, 8q24, 10q26 and 16q12, which is consistent with previous reports (2,18). In the study, we also found strong signals for ER− disease in regions 5q11, 10q26 and 19p13. It is possible that these signals may explain some of the excess risk for ER− disease in African Americans, since these risk alleles have higher frequencies in this population than they do in European-ancestry populations. However, our understanding of their contribution to racial and ethnic differences in disease incidence will only be determined once the functional variants have been identified and tested across populations. Unfortunately, we were not able to assess associations with triple-negative (ER/PR/HER2-negative; PR, progesterone receptor; HER2, human epidermal growth factor receptor 2) breast cancer, since HER2 status was available for only a limited number of cases. However, in a large study of women of European ancestry which tested many of these same index variants, further stratification on tumor subtype using HER2 status was not additionally informative for ER/PR-negative breast cancer (43).

The observation of secondary signals at many loci, and associations of variants with different tumor subtypes that have not yet been reported in European-ancestry populations could indicate a different genetic architecture of breast cancer across populations. For example, the index signal at *TNRC9* does not replicate in African Americans, but there appears to be a second risk variant that is unique to this population. At *FGFR2*, which was originally reported to be associated with ER+ disease in women of European ancestry, we found a signal for ER− disease with a marker correlated with the index variant. Similarly, for chromosome 19p13, which was reported as an ER− locus, we observed an association with ER+ breast cancer. However, these findings and their implications require further validation.

We investigated local ancestry as a potential confounding factor in the analysis of each risk locus. At five loci, we observed nominally significant evidence of association between local ancestry and breast cancer risk, with the most statistically significant association observed at 6q25 between European ancestry and ER+ breast cancer risk. Although the association of local ancestry and breast cancer risk needs to be validated in additional large studies, the inability to identify a risk variant that is differentiated in frequency between populations of European and African ancestry implies that either the association with local ancestry at many regions is a false-positive signal and/or we have not tested an adequate surrogate of the functional alleles.

The majority of the variants identified by GWAS for common cancers are of low risk (relative risks $<$1.30) and in aggregate are not yet informative for risk prediction (11–13). Until the functional alleles at each susceptibility locus are identified and their effects are accurately estimated, modeling of the genetic risk will rely on markers that best capture risk for a given population. Many of the markers we identified at these risk loci appear to have stronger associations with breast cancer risk compared with the GWAS-identified variants in African-American women. The risk score for overall breast cancer was also equally efficient for ER+ and ER− tumors. However, our hypothesis-generating model suggests that identification of tumor subtype-specific variants will improve the fit of these models.

While this is the largest study of African Americans to date to investigate genetic risk at known breast cancer susceptibility loci, statistical power was still limited. We had only 35% power to detect an OR of 1.10 for a risk allele of 0.10 frequency which may account for our inability to replicate GWAS signals or risk-associated markers in 10 of the regions. While attempting to apply a strict threshold for declaring significance through fine-mapping, we did not take into account testing for multiple phenotypes (overall breast as well as ER+ and ER− disease). As a result, the α-levels used as selection criteria may be too liberal. However, our risk modeling focused on the variants revealed for overall breast cancer, whereas we consider the associations observed for markers identified for ER+ or ER− disease and used in the subtype-specific risk modeling as hypothesis-generating. Since all of the cases and controls used for fine-mapping/discovery were also included in the risk modeling,

the risk model is likely to over-estimate the level of association due to winner's curse. Instead of partitioning the sample into test and validation sets, we felt it was necessary to use all of the subjects in the association testing of known variants and in fine-mapping to increase the statistical power to detect associations in each region. Therefore, other studies with reasonable power in African Americans must be performed in the future to test the model presented.

In summary, through fine-mapping of the breast cancer susceptibility regions in a large sample of African-American women, we identified markers with enhanced association with breast cancer in this population. Validation and augmentation of this model are needed before risk modeling based on genetic variants of low risk can be implemented in the clinical setting.

MATERIALS AND METHODS

Ethics statement

The Institutional Review Board at the University of Southern California approved the study protocol.

Study populations

This study included 9 epidemiological studies of breast cancer among African-American women, which comprise a total of 3153 cases and 2831 controls. Sample size and selected characteristics for these studies are summarized in Supplementary Material, Table S1. What follows is a brief description of these studies.

The Multiethnic Cohort Study (MEC). The MEC is a prospective cohort study of 215 000 men and women in Hawaii and Los Angeles (44) between the ages of 45 and 75 years at baseline (1993–1996). Through 31 December 2007, a nested breast cancer case–control study in the MEC included 556 African-American cases (544 invasive and 12 *in situ*) and 1003 African-American controls. An additional 178 African-American breast cancer cases (ages: 50–84) diagnosed between 1 June 2006 and 31 December 2007 in Los Angeles County (but outside of the MEC) were included in the study.

The Los Angeles component of The Women's Contraceptive and Reproductive Experiences (CARE) Study. The CARE Study is a large multi-center, population-based case–control study that was designed to examine the effects of oral contraceptive use on invasive breast cancer risk among African-American women and white women aged 35–64 years in five US locations (45). Cases in Los Angeles County were diagnosed from 1 July 1994 through 30 April 1998, and controls were sampled by random-digit dialing (RDD) from the same population and time period; 380 African-American cases and 224 African-American controls were included in the study.

The Women's Circle of Health Study (WCHS). The WCHS is an ongoing case–control study of breast cancer among European women and African-American women in the New York City boroughs and in seven counties in New Jersey (46). Eligible cases included women with invasive breast cancer between 20 and 74 years of age; controls were identified through RDD. The WCHS contributed 272 invasive African-American cases and 240 African-American controls.

The San Francisco Bay Area Breast Cancer Study (SFBCS). The SFBCS is a population-based case–control study of invasive breast cancer in Hispanic, African-American and non-Hispanic white women conducted between 1995 and 2003 in the San Francisco Bay Area (47). African-American cases, aged 35–79 years, were diagnosed between 1 April 1995 and 30 April 1999, with controls identified through RDD. Included from this study were 172 invasive African-American cases and 231 African-American controls.

The Northern California Breast Cancer Family Registry (NC-BCFR). The NC-BCFR is a population-based family study conducted in the Greater San Francisco Bay Area, and one of six sites of the Breast Cancer Family Registry (BCFR) (48). African-American breast cancer cases in NC-BCFR were diagnosed after 1 January 1995 and between the ages of 18 and 64 years; population controls were identified through RDD. Genotyping was conducted for 440 invasive African-American cases and 53 African-American controls.

The Carolina Breast Cancer Study (CBCS). The CBCS is a population-based case–control study conducted between 1993 and 2001 in 24 counties of central and eastern North Carolina (49). Cases were identified by rapid case ascertainment system in cooperation with the North Carolina Central Cancer Registry, and controls were selected from the North Carolina Division of Motor Vehicle and United States Health Care Financing Administration beneficiary lists. Participants' ages ranged from 20 to 74 years. DNA samples were provided from 656 African-American cases with invasive breast cancer and 608 African-American controls.

The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO) Cohort. PLCO, coordinated by the US National Cancer Institute (NCI) in 10 US centers, enrolled approximately 155 000 men and women aged 55–74 years during 1993–2001 in a randomized, two-arm trial to evaluate the efficacy of screening for these four cancers (50). A total of 64 African-American invasive breast cancer cases and 133 African-American controls contributed to this study.

The Nashville Breast Health Study (NBHS). The NBHS is a population-based case–control study of incident breast cancer conducted in Tennessee (15). The study was initiated in 2001 to recruit patients with invasive breast cancer or ductal carcinoma *in situ*, and controls, recruited through RDD between the ages of 25 and 75 years. NBHS contributed 310 African-American cases (57 *in situ*) and 186 African-American controls.

Wake Forest University Breast Cancer Study (WFBC). African-American breast cancer cases and controls in WFBC were recruited at Wake Forest University Health Sciences

from November 1998 through December 2008 (51). Controls were recruited from the patient population receiving routine mammography at the Breast Screening and Diagnostic Center. Age range of participants was 30–86 years. WFBC contributed 125 cases (116 invasive and 9 *in situ*) and 153 controls to the analysis.

Genotyping and quality control

Genotyping in stage 1 was conducted using the Illumina Human1M-Duo BeadChip. Of the 5984 samples from these studies (3153 cases and 2831 controls), we attempted genotyping of 5932, removing samples ($n = 52$) with DNA concentrations <20 ng/μl. Following genotyping, we removed samples based on the following exclusion criteria: (i) unknown replicates ($\geq$98.9% genetically identical) that we were able to confirm (only one of each duplicate was removed, $n = 15$); (ii) unknown replicates that we were not able to confirm through discussions with study investigators (pair or triplicate removed, $n = 14$); (iii) samples with call rates $<$95% after a second attempt ($n = 100$); (iv) samples with $\leq$5% African ancestry ($n = 36$) (discussed in what follows); and (v) samples with $<$15% mean heterozygosity of SNPs on the X chromosome and/or similar mean allele intensities of SNPs on the X and Y chromosomes ($n = 6$) (these are likely to be males).

In the analysis, we removed SNPs with $<$95% call rates ($n = 21\,732$) or MAFs $<$1% ($n = 80\,193$). To assess genotyping reproducibility, we included 138 replicate samples; the average concordance rate was 99.95% ($>$99.93% for all pairs). We also eliminated SNPs with genotyping concordance rates $<$98% based on the replicates ($n = 11\,701$). The final analysis data set included 1 043 036 SNPs genotyped on 3016 cases (1520 ER+, 988 ER− and the remaining 508 cases with unknown ER status) and 2745 controls, with an average SNP call rate of 99.7% and average sample call rate of 99.8%.

Statistical analysis

Ancestry estimation. We used principal components analysis (52) to estimate global ancestry among the 5761 individuals, using 2546 ancestry informative markers. Eigenvector 1 was highly correlated ($\rho = 0.997$, $P < 1 \times 10^{-16}$) with percentage of European ancestry, estimated in HAPMIX (53), and accounted for 10.1% of the variation between subjects; subsequent eigenvectors accounted for no more than 0.5%. At each locus and for each participant, we also estimated local ancestry [i.e. the number of European chromosomes (continuous between 0 and 2) carried by the participant], using the HAPMIX program (53). To summarize local ancestry at each region, for each individual we averaged across all local ancestry estimates that were within the start and end points of the region (Supplementary Material, Table S5). To address the potential for confounding by genetic ancestry, we adjusted for both global and local ancestry in all analyses.

SNP imputation. In order to generate a data set suitable for fine-mapping, we carried out genome-wide imputation using the software MACH (54). Phased haplotype data from the founders of the CEU and YRI HapMap Phase 2 samples were used to infer LD patterns in order to impute ungenotyped markers. The r^2 metric, defined as the observed variance divided by the expected variance, provides a measure of the quality of the imputation at any SNP, and was used as a threshold in determining which SNPs to filter from analysis ($r^2 < 0.3$). Of the 1 539 328 common SNPs (MAF ≥ 0.05) in the YRI population in HapMap Phase 2, we could impute 1 392 294 (90%) with $r^2 \geq 0.8$. For all the imputed SNPs presented in Results and the tables reported herein, the average r^2 was 0.92 (estimated in MACH).

Association testing. For each typed and imputed SNP, ORs and 95% CIs were estimated using unconditional logistic regression adjusting for age at diagnosis (or age at the reference date for controls), study, the first 10 eigenvalues and local ancestry. For each SNP, we tested for allele dosage effects through a 1 d.f. Wald χ^2 trend test.

We fine-mapped each risk locus using the combined genotyped and imputed SNPs in search of (i) an SNP that is more associated with risk in African Americans than the index signal; and (ii) a novel signal that is independent of the index signal. As some risk loci have been found to be more strongly associated with breast cancer subtypes, we investigated three outcomes: (i) overall breast cancer, (ii) ER+ breast cancer, and (iii) ER− breast cancer, with the latter two being hypothesis-generating. These analyses included SNPs (genotyped and imputed) spanning 250 kb upstream and 250 kb downstream of each index signal. If the index signal was contained within an LD block (based on the D′ statistic) of $>$250 kb, then the region was extended to include the entire region of LD.

Stepwise regression was performed by region to select the most informative risk variants as discussed in what follows, in models adjusted for age, study, global ancestry (the first 10 eigenvectors) and local ancestry. In the stepwise regression, we preserved the original sample size by using the mean genotype of typed subjects in place of 'no-calls' for SNPs with $<$100% genotyping completion rate.

Within each known risk locus, it is expected that markers that are associated with risk in African Americans will be correlated with the index signal reported in Europeans. Thus, we identified and tested SNPs that are correlated ($r^2 > 0.2$) with the index signals in the GWAS populations (HapMap CEU or CHB for 6q25). For each region, we determined the number of tags needed to capture all the SNPs correlated with the index signal in the YRI population (Phase 2 HapMap). The average number of tags in each region was then used as the correction factor for Bonferroni correction. An α-level of 0.05 divided by average number of tags needed in each region was applied in the stepwise regression process. For all of the remaining markers that were not correlated with the index signal (in Europeans), we applied a more stringent α-level for defining statistical significance. In each risk region, we determined the number of tag SNPs needed to capture all common alleles (MAF > 0.05, with $r^2 > 0.8$) in the YRI HapMap population. The total number of tags across the 19 regions was then used as a correction factor, as they define the number of independent tests in each region. An α of 0.05 divided by the number of tags was

applied to assess statistical significance for any putative novel, independent signal in each region. For correlated SNPs that were selected to be better markers, we also assessed phase to ensure that the new risk allele is on the same haplotype as the GWAS-reported risk allele in the HapMap CEU population.

Risk modeling. We modeled the cumulative genetic risk of breast cancer using the risk variants reported in previous GWAS (total = 19). We compared the results with a model of the SNPs found to be significantly associated with risk in African Americans, which included SNPs identified from the stepwise procedures at all loci for overall breast cancer risk (presented in Table 1). More specifically, in each case we summed the number of risk alleles for each individual and estimated the OR per allele for this aggregate-unweighted allele count variable as an approximate risk score appropriate for unlinked variants with independent effects of approximately the same magnitude for each allele. We then applied this risk score to overall breast cancer as well as ER+/ER− breast cancer subtypes. We also constructed risk scores based on risk alleles for ER+ and ER− tumor subtypes separately, and, as hypothesis-generating, applied both risk scores to overall and ER+/ER− breast cancer subtypes.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

We thank the women who volunteered to participate in each study. We also thank Madhavi Eranti, Andrea Holbrook, Paul Poznaik, Loreall Pooler, Xin Sheng and David Wong from the University of Southern California for their technical support. We would also like to acknowledge co-investigators from the WCHS study: Dana H. Bovbjerg (University of Pittsburgh), Lina Jandorf (Mount Sinai School of Medicine) and Gregory Ciupak, Warren Davis, Gary Zirpoli, Song Yao and Michelle Roberts from Roswell Park Cancer Institute.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by a Department of Defense Breast Cancer Research Program Era of Hope Scholar Award to C.A.H. (W81XWH-08-1-0383), a National Institute of Health grant to C.A.H. (R01-CA132839), the Norris Foundation, and a grant from the California Breast Cancer Research Program to D.O.S. (15UB-8402). Each of the participating studies was supported by the following grants: MEC: by National Institutes of Health (R01-CA63464 and R37-CA54281); CARE: by National Institute for Child Health and Development (NO1-HD-3-3175), WCHS: by US Army Medical Research and Material Command (USAMRMC) (DAMD-17-01-0-0334); the National Institutes of Health (R01-CA100598); and the Breast Cancer Research Foundation; SFBCS: by National Institutes of Health (R01-CA77305) and United States Army Medical Research Program (DAMD17-96-6071); NC-BCFR: by National Institutes of Health (U01-CA69417); CBCS: by National Institutes of Health Specialized Program of Research Excellence in Breast Cancer (P50-CA58223) and Center for Environmental Health and Susceptibility, National Institute of Environmental Health Sciences, National Institutes of Health (P30-ES10126); PLCO: by Intramural Research Program, National Cancer Institute, National Institutes of Health; NBHS: by National Institutes of Health (R01-CA100374); WFBC: by National Institutes of Health (R01-CA73629). The Breast Cancer Family Registry (BCFR) was supported by the National Cancer Institute, National Institutes of Health under (RFA CA-95-011) and through cooperative agreements with members of the Breast Cancer Family Registry and Principal Investigators.

REFERENCES

1. Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., Luben, R. *et al.* (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087–1093.
2. Stacey, S.N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S.A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A. *et al.* (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.*, **39**, 865–869.
3. Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, M., Hankinson, S.E., Wacholder, S., Wang, Z., Welch, R., Huntchinson, A. *et al.* (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.*, **39**, 870–874.
4. Stacey, S.N., Manolescu, A., Sulem, P., Thorlacius, S., Gudjonsson, S.A., Jonsson, G.F., Jakobsdottir, M., Bergthorsson, J.T., Gudmundsson, J., Aben, K.K. *et al.* (2008) Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.*, **40**, 703–706.
5. Zheng, W., Long, J., Gao, Y.T., Li, C., Zheng, Y., Xiang, Y.B., Wen, W., Levy, S., Deming, S.L., Haines, J.L. *et al.* (2009) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.*, **41**, 324–328.
6. Thomas, G., Jacobs, K.B., Kraft, P., Yeager, M., Wacholder, S., Cox, D.G., Hankinson, S.E., Hutchinson, A., Wang, Z., Yu, K. *et al.* (2009) A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat. Genet.*, **41**, 579–584.
7. Ahmed, S., Thomas, G., Ghoussaini, M., Healey, C.S., Humphreys, M.K., Platte, R., Morrison, J., Maranian, M., Pooley, K.A., Luben, R. *et al.* (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat. Genet.*, **41**, 585–590.
8. Turnbull, C., Shahana, A., Morrison, J., Pernet, D., Renwick, A., Maranian, M., Seal, S., Ghoussaini, M., Hines, S., Healey, C.S. *et al.* (2010) Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.*, **42**, 504–507.
9. Antoniou, A.C., Wang, X., Fredericksen, Z.S., McGuffog, L., Tarrell, R., Sinilnikova, O.M., Healey, S., Morrison, J., Kartsonaki, C., Lesnick, T. *et al.* (2010) A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat. Genet.*, **42**, 885–892.
10. Fletcher, O., Johnson, N., Orr, N., Hosking, F.J., Gibson, L.J., Walker, K., Zelenika, D., Gut, I., Heath, S., Palles, C. *et al.* (2011) Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. *J. Natl Cancer Inst.*, **103**, 425–435.
11. Pepe, M.S. and Janes, H.E. (2008) Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J. Natl Cancer Inst.*, **100**, 978–979.

12. Gail, M.H. (2008) Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *J. Natl Cancer Inst.*, **100**, 1037–1041.
13. Pharoah, P.D., Antoniou, A., Bobrow, M., Zimmern, R.L., Easton, D.F. and Ponder, B.A. (2002) Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.*, **31**, 33–36.
14. Chen, F., Stram, D.O., Le Marchand, L., Monroe, K.R., Kolonel, L.N., Henderson, B.E. and Haiman, C.A. (2010) Caution in generalizing known genetic risk markers for breast cancer across all ethnic/racial populations. *Eur. J. Hum. Genet.*, **19**, 243–245.
15. Zheng, W., Cai, Q., Signorello, L.B., Long, J., Hargreaves, M.K., Deming, S.L., Li, G., Li, C., Cui, Y. and Blot, W.J. (2009) Evaluation of 11 breast cancer susceptibility loci in African-American women. *Cancer Epidemiol. Biomarkers Prev.*, **18**, 2761–2764.
16. Ruiz-Narvaez, E.A., Rosenberg, L., Cozier, Y.C., Cupples, L.A., Adams-Campbell, L.L. and Palmer, J.R. (2010) Polymorphisms in the TOX3/LOC643714 locus and risk of breast cancer in African-American women. *Cancer Epidemiol. Biomarkers Prev.*, **19**, 1320–1327.
17. Stacey, S.N., Sulem, P., Zanon, C., Gudjonsson, S.A., Thorleifsson, G., Helgason, A., Jonasdottir, A., Besenbacher, S., Kostic, J.P., Fackenthal, J.D. *et al.* (2010) Ancestry-shift refinement mapping of the C6orf97-ESR1 breast cancer susceptibility locus. *PLoS Genet.*, **6**, e1001029.
18. Garcia-Closas, M., Hall, P., Nevanlinna, H., Pooley, K., Morrison, J., Richesson, D.A., Bojesen, S.E., Nordestgaard, B.G., Axelsson, C.K., Arias, J.I. *et al.* (2008) Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics. *PLoS Genet.*, **4**, e1000054.
19. Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G., Tandon, A., Kao, L., Ruczinski, I., Fornage, M., Siscovick, D., Zhu, X. *et al.* (2011) Enhanced statistical tests for GWAS in admixed populations: assessment using African Americans from CARe and a breast cancer consortium. *PLoS Genet.*, **7**, e1001371.
20. Fejerman, L., Haiman, C.A., Reich, D., Tandon, A., Deo, R.C., John, E.M., Ingles, S.A., Ambrosone, C.B., Bovbjerg, D.H., Jandorf, L.H. *et al.* (2009) An admixture scan in 1484 African American women with breast cancer. *Cancer Epidemiol. Biomarkers Prev.*, **18**, 3110–3117.
21. Milne, R.L., Benitez, J., Nevanlinna, H., Heikkinen, T., Aittomaki, K., Blomqvist, C., Arias, J.I., Zamora, M.P., Burwinkel, B., Bartram, C.R. *et al.* (2009) Risk of estrogen receptor-positive and -negative breast cancer and single-nucleotide polymorphism 2q35-rs13387042. *J. Natl Cancer Inst.*, **101**, 1012–1018.
22. Udler, M.S., Meyer, K.B., Pooley, K.A., Karlins, E., Struewing, J.P., Zhang, J., Doody, D.R., MacArthur, S., Tyrer, J., Pharoah, P.D. *et al.* (2009) FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Hum. Mol. Genet.*, **18**, 1692–1703.
23. Jia, L., Landan, G., Pomerantz, M., Jaschek, R., Herman, P., Reich, D., Yan, C., Khalid, O., Kantoff, P., Oh, W. *et al.* (2009) Functional enhancers at the gene-poor 8q24 cancer-linked locus. *PLoS Genet.*, **5**, e1000597.
24. Freedman, M.L. (2006) Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl Acad. Sci. USA*, **103**, 14068–14073.
25. Ghoussaini, M., Song, H., Koessler, T., Al Olama, A.A., Kote-Jarai, Z., Driver, K.E., Pooley, K.A., Ramus, S.J., Kjaer, S.K., Hogdall, E. *et al.* (2008) Multiple loci with different cancer specificities within the 8q24 gene desert. *J. Natl Cancer Inst.*, **100**, 962–966.
26. Kiemeney, L.A., Thorlacius, S., Sulem, P., Geller, F., Aben, K.K., Stacey, S.N., Gudmundsson, J., Jakobsdottir, M., Bergthorsson, J.T., Sigurdsson, A. *et al.* (2008) Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat. Genet.*, **40**, 1307–1312.
27. Goode, E.L., Chenevix-Trench, G., Song, H., Ramus, S.J., Notaridou, M., Lawrenson, K., Widschwendter, M., Vierkant, R.A., Larson, M.C., Kjaer, S.K. *et al.* (2010) A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nat. Genet.*, **42**, 874–879.
28. Crowther-Swanepoel, D., Broderick, P., Di Bernardo, M.C., Dobbins, S.E., Torres, M., Mansouri, M., Ruiz-Ponte, C., Enjuanes, A., Rosenquist, R., Carracedo, A. *et al.* (2010) Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat. Genet.*, **42**, 132–136.
29. Salinas, C.A., Kwon, E., Carlson, C.S., Koopmeiners, J.S., Feng, Z., Karyadi, D.M., Ostrander, E.A. and Stanford, J.L. (2008) Multiple independent genetic variants in the 8q24 region are associated with prostate cancer risk. *Cancer Epidemiol. Biomarkers Prev.*, **17**, 1203–1213.
30. Gudmundsson, J., Sulem, P., Gudbjartsson, D.F., Blondal, T., Gylfason, A., Agnarsson, B.A., Benediktsdottir, K.R., Magnusdottir, D.N., Orlygsdottir, G., Jakobsdottir, M. *et al.* (2009) Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat. Genet.*, **41**, 1122–1126.
31. Gudmundsson, J., Sulem, P., Manolescu, A., Amundadottir, L.T., Gudbjartsson, D., Helgason, A., Rafnar, T., Bergthorsson, J.T., Agnarsson, B.A., Baker, A. *et al.* (2007) Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.*, **39**, 631–637.
32. Yeager, M., Orr, N., Hayes, R.B., Jacobs, K.B., Kraft, P., Wacholder, S., Minichiello, M.J., Fearnhead, P., Yu, K., Chatterjee, N. *et al.* (2007) Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.*, **39**, 645–649.
33. Yeager, M., Chatterjee, N., Ciampa, J., Jacobs, K.B., Gonzalez-Bosquet, J., Hayes, R.B., Kraft, P., Wacholder, S., Orr, N., Berndt, S. *et al.* (2009) Identification of a new prostate cancer susceptibility locus on chromosome 8q24. *Nat. Genet.*, **41**, 1055–1057.
34. Al Olama, A.A., Kote-Jarai, Z., Giles, G.G., Guy, M., Morrison, J., Severi, G., Leongamornlert, D.A., Tymrakiewicz, M., Jhavar, S., Saunders, E. *et al.* (2009) Multiple loci on 8q24 associated with prostate cancer susceptibility. *Nat. Genet.*, **41**, 1058–1060.
35. Haiman, C.A., Patterson, N., Freedman, M.L., Myers, S.R., Pike, M.C., Waliszewska, A., Neubauer, J., Tandon, A., Schirmer, C., McDonald, G.J. *et al.* (2007) Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.*, **39**, 638–644.
36. Long, J., Cai, Q., Shu, X.O., Qu, S., Li, C., Zheng, Y., Gu, K., Wang, W., Xiang, Y.B., Cheng, J. *et al.* (2010) Identification of a functional genetic variant at 16q12.1 for breast cancer risk: results from the Asia Breast Cancer Consortium. *PLoS Genet.*, **6**, e1001002.
37. Waters, K.M., Stram, D.O., Hassanein, M.T., Le Marchand, L., Wilkens, L.R., Maskarinec, G., Monroe, K.R., Kolonel, L.N., Altshuler, D., Henderson, B.E. *et al.* (2010) Consistent association of type 2 diabetes risk variants found in Europeans in diverse racial and ethnic groups. *PLoS Genet.*, **6**, e1001078.
38. Waters, K.M., Le Marchand, L., Kolonel, L.N., Monroe, K.R., Stram, D.O., Henderson, B.E. and Haiman, C.A. (2009) Generalizability of associations from prostate cancer genome-wide association studies in multiple populations. *Cancer Epidemiol. Biomarkers Prev.*, **18**, 1285–1289.
39. Sharma, M., Li, X., Wang, Y., Zarnegar, M., Huang, C.-Y., Palvimo, J.J., Lim, B. and Sun, Z. (2003) hZimp10 is an androgen receptor co-activator and forms a complex with SUMO-1 at replication foci. *EMBO J.*, **22**, 6101–6114.
40. Li, X., Thyssen, G., Beliakoff, J. and Sun, Z. (2006) The novel PIAS-like protein hZimp10 enhances Smad transcriptional activity. *J. Biol. Chem.*, **281**, 23748–23756.
41. Lee, J., Beliakoff, J. and Sun, Z. (2007) The novel PIAS-like protein hZimp10 is a transcriptional co-activator of the p53 tumor suppressor. *Nucleic Acids Res.*, **35**, 4523–4534.
42. Udler, M.S., Ahmed, S., Healey, C.S., Meyer, K., Struewing, J., Maranian, M., Kwon, E.M., Zhang, J., Tyrer, J., Karlins, E. *et al.* (2010) Fine scale mapping of the breast cancer 16q12 locus. *Hum. Mol. Genet.*, **19**, 2507–2515.
43. Broeks, A., Schmidt, M.K., Sherman, M.E., Couch, F.J., Hopper, J.L., Dite, G.S., Apicella, C., Smith, L.D., Hammet, F., Southey, M.C. *et al.* (2011) Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. *Hum. Mol. Genet.*, **20**, 3289–3303.
44. Kolonel, L.N., Henderson, B.E., Hankin, J.H., Nomura, A.M., Wilkens, L.R., Pike, M.C., Stram, D.O., Monroe, K.R., Earle, M.E. and Nagamine, F.S. (2000) A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.*, **151**, 346–357.
45. Marchbanks, P.A., McDonald, J.A., Wilson, H.G., Burnett, N.M., Daling, J.R., Bernstein, L., Malone, K.E., Strom, B.L., Norman, S.A., Weiss, L.K. *et al.* (2002) The NICHD Women's Contraceptive and Reproductive Experiences Study: methods and operational results. *Ann. Epidemiol.*, **12**, 213–221.
46. Ambrosone, C.B., Ciupak, G.L., Bandera, E.V., Jandorf, L., Bovbjerg, D.H., Zirpoli, G., Pawlish, K., Godbold, J., Furberg, H., Fatone, A. *et al.* (2009) Conducting molecular epidemiological research in the age of HIPAA: a multi-institutional case-control study of breast cancer in

African-American and European-American women. *J. Oncol.*, **2009**, 871250.
47. John, E.M., Schwartz, G.G., Koo, J., Wang, W. and Ingles, S.A. (2007) Sun exposure, vitamin D receptor gene polymorphisms, and breast cancer risk in a multiethnic population. *Am. J. Epidemiol.*, **166**, 1409–1419.
48. John, E.M., Hopper, J.L., Beck, J.C., Knight, J.A., Neuhausen, S.L., Senie, R.T., Ziogas, A., Andrulis, I.L., Anton-Culver, H., Boyd, N. *et al.* (2004) The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res.*, **6**, R375–R389.
49. Newman, B., Moorman, P.G., Millikan, R., Qaqish, B.F., Geradts, J., Aldrich, T.E. and Liu, E.T. (1995) The Carolina Breast Cancer Study: integrating population-based epidemiology and molecular biology. *Breast Cancer Res. Treat*, **35**, 51–60.
50. Prorok, P.C., Andriole, G.L., Bresalier, R.S., Buys, S.S., Chia, D., Crawford, E.D., Fogel, R., Gelmann, E.P., Gilbert, F., Hasson, M.A. *et al.* (2000) Design of the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial. *Control Clin. Trials*, **21**, 273S–309S.
51. Smith, T.R., Levine, E.A., Freimanis, R.I., Akman, S.A., Allen, G.O., Hoang, K.N., Liu-Mares, W. and Hu, J.J. (2008) Polygenic model of DNA repair genetic polymorphisms in human breast cancer risk. *Carcinogenesis*, **29**, 2132–2138.
52. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.
53. Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D. and Myers, S. (2009) Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet.*, **5**, e1000519.
54. Li, Y., Willer, C., Sanna, S. and Abecasis, G. (2009) Genotype imputation. *Annu. Rev. Genomics Hum. Genet.*, **10**, 387–406.
55. Pruim, R.J., Welch, R.P., Sanna, S., Teslovich, T.M., Chines, P.S., Gliedt, T.P., Boehnke, M., Abecasis, G.R. and Willer, C.J. (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, **26**, 2336–2337.

A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11

Afshan Siddiq[1]†, Fergus J. Couch[2,3]†, Gary K.Chen[4]†, Sara Lindström[5], Diana Eccles[6], Robert C. Millikan[7], Kyriaki Michailidou[8], Daniel O. Stram[4], Lars Beckmann[9], Suhn Kyong Rhie[4], Christine B. Ambrosone[10], Kristiina Aittomäki[11], Pilar Amiano[12], Carmel Apicella[13], Australian Breast Cancer Tissue Bank Investigators[14], Laura Baglietto[13,15], Elisa V. Bandera[16], Matthias W. Beckmann[17], Christine D. Berg[18], Leslie Bernstein[19], Carl Blomqvist[20], Hiltrud Brauch[21], Louise Brinton[22], Quang M. Bui[13], Julie E. Buring[23], Saundra S. Buys[24], Daniele Campa[25], Jane E. Carpenter[26], Daniel I. Chasman[27], Jenny Chang-Claude[28], Constance Chen[5], Françoise Clavel-Chapelon[29], Angela Cox[30] , Simon S. Cross[31], Kamila Czene[32], Sandra L. Deming[33], Robert B. Diasio[34], W. Ryan Diver[35], Alison M. Dunning[36], Lorraine Durcan[6], Arif B. Ekici[37], Peter A. Fasching[17,38], Familial Breast Cancer Study[39], Heather Spencer Feigelson[40], Laura Fejerman[41], Jonine D Figueroa[22], Olivia Fletcher[42], Dieter Flesch-Janys[43], Mia M. Gaudet[35], The GENICA Consortium[44], Susan M. Gerty[6], Jorge L. Rodriguez-Gil[45], Graham G. Giles[13,15], Carla H. van Gils[46], Andrew K. Godwin[47], Nikki Graham[6], Dario Greco[48], Per Hall[32], Susan E. Hankinson[23], Arndt Hartmann[49], Rebecca Hein[28,50], Judith Heinz[43], Robert N. Hoover[22], John L Hopper[13], Jennifer J. Hu[45], Scott Huntsman[51], Sue A. Ingles[4], Astrid Irwanto[52], Claudine Isaacs[53], Kevin B. Jacobs[22,54,55], Esther M. John[56], Christina Justenhoven[21], Rudolf Kaaks[28], Laurence N. Kolonel[57], Gerhard A. Coetzee[4,87], Mark Lathrop[58,59], Loic Le Marchand[57], Adam M. Lee[34], I-Min Lee[23], Timothy Lesnick[2], Peter Lichtner[60], Jianjun Liu[52], Eiliv Lund[61], Enes Makalic[13], Nicholas G. Martin[62], Catriona A McLean[63], Hanne Meijers-Heijboer[64], Alfons Meindl[65], Penelope Miron[66], Kristine R. Monroe[4], Grant W. Montgomery[62], Bertram Müller-Myhsok[67], Stefan Nickels[28], Sarah J. Nyante[22], Curtis Olswold[2], Kim Overvad[68], Domenico Palli[69], Daniel J Park[70], Julie R. Palmer[71], Harsh Pathak[47], Julian Peto[72], Paul Pharoah[36], Nazneen Rahman[39], Fernando Rivadeneira[73], Daniel F. Schmidt[13], Rita K Schmutzler[74], Susan Slager[2], Melissa C. Southey[70],

Kristen N. Stevens[2], Hans-Peter Sinn[75], Michael F. Press[76], Eric Ross[77], Elio Riboli[78], Paul M. Ridker[27], Fredrick R. Schumacher[4], Gianluca Severi[13,15], Isabel dos Santos Silva[72], Jennifer Stone[13], Malin Sund[79], William J. Tapper[6], Michael J. Thun[35], Ruth C. Travis[80], Clare Turnbull[39], Andre G. Uitterlinden[73], Quinten Waisfisz[64], Xianshu Wang[3], Zhaoming Wang[22,54], JoEllen Weaver[81], Rüdiger Schulz-Wendtland[82], Lynne R. Wilkens[57], David Van Den Berg[4], Wei Zheng[83], Regina G. Ziegler[22], Elad Ziv[51], Heli Nevanlinna[48], Douglas F. Easton[36], David J. Hunter[84,85], Brian E. Henderson[4], Stephen J. Chanock[22], Montserrat Garcia-Closas[86], Peter Kraft[5]†, Christopher A. Haiman[4]†, Celine M. Vachon[2]†*

† These authors contributed equally.

*To whom correspondence should be addressed: Celine M. Vachon, Mayo Clinic, 200 First St SW, Charlton 6-239, Rochester, MN 55905, (Tel): 507-284-9977, (Fax): 507-266-2478, E-mail: Vachon.Celine@mayo.edu or Christopher A. Haiman, Harlyne Norris Research Tower, 1450 Biggy Street, Room 1504, Los Angeles, CA 90033 USA, Email: Christopher.Haiman@med.usc.edu

[1] Department of Epidemiology and Biostatistics & Department of Genomics of Common Disease, School of Public Health, Imperial College London, United Kingdom

[2] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

[3] Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA

[4] Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California, USA

[5] Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, MA, USA

[6] Faculty of Medicine, University of Southampton, Southampton, UK

[7] Department of Epidemiology, Gillings School of Global Public Health, and Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC, USA

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

[9] Institute for Quality and Efficiency in Health Care, IQWiG, Cologne, Germany

[10] Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY, USA

[11] Department of Clinical Genetics, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland

[12] Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Madrid, Spain

[13] Centre for Molecular, Environmental, Genetic, and Analytic Epidemiology, Melbourne School of Population Health, The University of Melbourne, Australia

[14] ABCTB, University of Sydney, NSW, Australia

[15] Cancer Epidemiology Centre, The Cancer Council Victoria, Melbourne, Australia & Centre for Molecular, Environmental, Genetic, and Analytic Epidemiology, The University of Melbourne, Australia

[16] The Cancer Institute of New Jersey, New Brunswick, NJ, USA

[17] Friedrich-Alexander University Erlangen-Nuremberg , University Hospital Erlangen, University Breast Center Franconia, Department of Gynecology and Obstetrics, Erlangen, Germany

[18] Early Detection Research Group, Division of Cancer Prevention, National Cancer Institute, Rockville, Maryland, USA

[19] Division of Cancer Etiology, Department of Population Science, Beckman Research Institute, City of Hope, CA, USA

[20] Department of Oncology, Helsinki University Central Hospital, Helsinki, Finland

[21] Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, and University Tübingen, Germany

[22] Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA

[23] Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

[24]Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA

[25] Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany

[26] Australian Breast Cancer Tissue Bank, University of Sydney at the Westmead Millennium Institute, Westmead, NSW, Australia

[27] Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA

[28]Division of Cancer Epidemiology, German Cancer Research Center, Deutsches Krebsforschungszentrum, Heidelberg, Germany

[29] INSERM UMR 1018, Team 9: Nutrition, Hormones et Sante´ desfemmes, Centre de Recherche en Epide´miologie et Sante´ des Populations, Hoˆpital Paul Brousse, Villejuif, France

[30] Institute for Cancer Studies, Department of Oncology , Faculty of Medicine, Dentistry & Health, University of Sheffield, UK

[31] Academic Unit of Pathology, Department of Neuroscience, Faculty of Medicine, Dentistry & Health, University of Sheffield, UK

[32] Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 17177, Sweden

[33] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center and Vanderbilt-Ingram Cancer Center Vanderbilt University School of Medicine, Nashville, TN, USA

[34] Department of Pharmacology, Mayo Clinic, Rochester, MN, USA

[35] Epidemiology Research Program, American Cancer Society, Atlanta, GA, USA

[36] Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK

[37] Friedrich-Alexander University Erlangen-Nuremberg, Institute of Human Genetics, Erlangen, Germany

[38] University of California at Los Angeles, David Geffen School of Medicine, Department of Medicine, Division of hematology and Oncology, Los Angeles, CA, USA

[39] Section of Cancer Genetics, Institute of Cancer Research, Sutton, UK

[40] Institute for Health Research, Kaiser Permanente, Denver, CO, USA

[41] Division of General Internal Medicine and Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, California

[42] Breakthrough Breast Cancer Research Centre, Institute of Cancer Research, London, UK

[43] Department of Cancer Epidemiology/Clinical Cancer Registry University Cancer Center Hamburg (UCCH) and Department of Medical Biometrics and Epidemiology University Medical Center Hamburg-Eppendorf, Hamburg, Germany

[44] Gene Environment Interaction and Breast Cancer in Germany (GENICA): Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, and University Tübingen, Germany (HB, CJ); Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany (Ute Hamann); Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany (Yon-Dschun Ko, Christian Baisch); Institute of Pathology, Medical Faculty of the University of Bonn, Germany (Hans-Peter Fischer); Institute for Prevention and Occupational Medicine of the German Social Accident Insurance (IPA), Bochum, Germany (Thomas Bruening, Beate Pesch, Sylvia Rabstein), Institute and Outpatient Clinic of Occupational Medicine, Saarland University Medical Center and Saarland University Faculty of Medicine, Homburg, Germany (Volker Harth)

[45] Sylvester Comprehensive Cancer Center and Department of Epidemiology and Public Health, University of Miami Miller School of Medicine, Miami, FL, USA

[46] Julius Center, University Medical Center, Utrecht, The Netherlands

[47] Department of Pathology and Laboratory Medicine, Kansas University Medical Center, Kansas City, KS, USA

[48] Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland

[49] Friedrich-Alexander University Erlangen-Nuremberg, Institute of Pathology, University Hospital Erlangen, Erlangen, Germany

[50] PMV Research Group at the Department of Child and Adolescent Psychiatry and Psychotherapy, University of Cologne, Cologne, Germany

[51] Division of General Internal Medicine, Department of Medicine, Institute for Human Genetics and Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, USA

[52] Human Genetics Division, Genome Institute of Singapore, Singapore.

[53] Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC

[54] Core Genotyping Facility, SAIC-Frederick Inc., NCI-Frederick, Frederick, MD, USA

[55] Bioinformed Consulting Services, Gaithersburg, MD, USA

[56] Cancer Prevention Institute of California, Fremont, CA, USA, and Stanford University School of Medicine and Stanford Cancer Institute, Stanford, CA, USA

[57] Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA

[58] Centre National de Genotypage, Evry, France.

[59] Fondation Jean Dausset – CEPH, Paris, France.

[60] Institute of Human Genetics, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[61] Institute of Community Medicine University of Tromsø, Tromsø,Norway

[62] QIMR GWAS Collective, Queensland Institute of Medical Research, Brisbane, Australia

[63] The Alfred Hospital, Melbourne, Australia

[64] Department of Clinical Genetics, VU University Medical Center, section Oncogenetics, Amsterdam, The Netherlands

[65] Clinic of Gynaecology and Obstetrics, Division for Gynaecological Tumor-Genetics, Technische Universität München, München, Germany

[66] Dana Farber Cancer Institute, Boston, MA, USA

[67] Max Planck Institute of Psychiatry, Munich, Germany

[68] Department of Cardiology, Center for Cardiovascular Research,Aalborg Hospital, Aarhus University Hospital, Aalborg, Denmark

[69] Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Institute, ISPO, Florence, Italy

[70] Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Australia

[71] Slone Epidemiology Center at Boston University, Boston, MA, USA

[72] Non-communicable Disease Epidemiology Department, London School of Hygiene and Tropical Medicine, London, UK.

[73] Department of Internal Medicine and Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands

[74] Department of Obstetrics and Gynaecology, Division of Molecular Gynaeco-Oncology, University of Cologne, Germany

[75] Department of Pathology, University Hospital Heidelberg, Heidelberg, Germany

[76] Department of Pathology, Keck School of Medicine and Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA, USA

[77] Department of Biostatistics, Fox Chase Cancer Center, Philadelphia, PA, USA

[78] Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, United Kingdom

[79] Department of Surgery, Umea° University, Umea°, Sweden

[80] Cancer Epidemiology Unit, Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK

[81] Biosample Repository, Fox Chase Cancer Center, Philadelphia, PA, USA

[82] Friedrich-Alexander University Erlangen-Nuremberg, Institute of Diagnostic Radiology, University Hospital Erlangen, Erlangen, Germany

[83] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, and Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA

[84] Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA

[85] Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, MA, USA

[86] Section of Epidemiology and Genetics, Institute of Cancer Research, Sutton, United Kingdom

[87] Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, CA, 90089

ABSTRACT

Genome-wide association studies (GWAS) of breast cancer defined by hormone receptor status have revealed loci contributing to susceptibility of estrogen receptor (ER)-negative subtypes. To identify additional genetic variants for ER-negative breast cancer we conducted the largest meta-analysis of ER-negative disease to date, comprising 4,754 ER-negative cases and 31,663 controls from three GWAS: NCI Breast and Prostate Cancer Cohort Consortium (BPC3) (2,188 ER-negative cases; 25,519 controls of European ancestry), Triple Negative Breast Cancer Consortium (TNBCC) (1,562 triple negative cases; 3,399 controls of European ancestry) and African American Breast Cancer Consortium (AABC) (1,004 ER-negative cases; 2,745 controls). We performed *in silico* replication of 86 SNPs at $P \leq 1x10^{-5}$ in an additional 11,209 breast cancer cases (946 with ER-negative disease) and 16,057 controls of Japanese, Latino and European ancestry. We identified two novel loci for breast cancer at 20q11 and 6q14. SNP rs2284378 at 20q11 was associated with ER-negative breast cancer (combined two stage OR=1.16; $P= 1.1x10^{-8}$) but showed a weaker association with overall breast cancer (OR=1.08, $P=1.3x10^{-6}$) based on 17,869 cases and 43,745 controls and no association with ER-positive disease (OR=1.01, P=0.67) based on 9,965 cases and 22,902 controls. Similarly, rs17530068 at 6q14 was associated with breast cancer (OR=1.12; $P= 1.1x10^{-9}$), and with both ER-positive (OR=1.09; $P=1.5x10^{-5}$) and ER-negative (OR=1.16, $P=2.5x10^{-7}$) disease. We also confirmed three known loci associated with ER-negative (19p13) and both ER-negative and ER-positive breast cancer (6q25 and 12p11). Our results highlight the value of large-scale collaborative studies to identify novel breast cancer risk loci.

INTRODUCTION

Breast cancer is a heterogeneous disease and has multiple histological and molecular subtypes, likely with distinct etiologies. Tumors that lack expression of the estrogen receptor (ER) tend to have more aggressive disease, higher histological grade, and lower survival rates (1). ER-negative breast cancer is more common in women of African ancestry, accounting for as much as 40% of cases in African American women compared with 15-20% in women of European ancestry. The etiologic heterogeneity between breast cancer subtypes is supported by different associations with ER-positive versus ER-negative disease for many of the known breast cancer risk factors (such as reproductive factors and BMI)(2). Tumors in women with *BRCA1* mutations are predominantly ER-negative, while tumors in *BRCA2* mutation carriers are predominantly ER-positive(3). Furthermore, genome-wide association studies have identified multiple common genetic variants more strongly associated with ER-positive than ER-negative breast cancer(4). Through collaborative efforts, we recently identified risk loci on 5p15 and 19p13 that are associated specifically with ER-negative and triple negative (TN) (ER-negative, progresterone (PR)-negative and HER2-negative) breast cancer(5-7).

In order to identify genetic loci associated with risk of ER-negative breast cancer, we conducted a meta-analysis of three GWAS of ER-negative breast cancer, comprising 4,754 cases and 31,663 controls with further replication in an additional 11,209 cases (946 with ER-negative disease) and 16,057 controls.

RESULTS

The meta-analysis included GWAS of ER-negative breast cancer (4,754 ER-negative cases and 31,663 controls) from the NCI Breast and Prostate Cancer Cohort Consortium (BPC3) (2,188 ER-negative cases and 25,519 controls of European ancestry), the Triple Negative Breast Cancer Consortium (TNBCC) (1,562 triple negative cases and 3,399 controls of European ancestry) and the African American Breast Cancer Consortium (AABC) (1,004 ER-negative cases and 2,745 controls). **(Figure 1, Supplementary Table 1)**. We observed little evidence of over-inflation in the test statistics ($\lambda \leq 1.04$ for each study; $\lambda=1.04$ for meta-analysis) **(Supplementary Figure 1)**. A total of 86 SNPs were associated with ER-

negative breast cancer at $P \leq 10^{-5}$ **(Supplementary Table 2).** An *in silico* replication of the 86 SNPs was conducted using GWAS of European (BCAC combined), Latino (MEC-LAT, SFBCS/NC-BCFR) and Japanese (MEC-JPT) ancestry populations, totaling 11,209 breast cancer cases (946 with ER-negative disease) and 8,404 controls (Stage 2)**(Supplementary Table 1).**

Combining results for ER-negative breast cancer from stages 1 and 2, variants in three regions showed genome-wide significance [20q11-rs2284378, T allele: odds ratio, OR=1.16, $P = 1.1x10^{-8}$ (**Table 1**); 19p13-rs8100241, G allele: OR=1.14, $P=3.5x10^{-8}$; 6q25-rs9383938, T allele: OR=1.28, $P = 2.37 \times 10^{-10}$]. Variants at 6q25 have previously been associated with breast cancer risk(8), and variants at the 19p13 locus have been associated with ER-negative and TN breast cancer risk(5, 7). The rs2284378 variant at 20q11 is located in a region containing *RALY* (RNA binding protein, autoantigenic), *EIF2S2* (eukaryotic translation initiation factor 2, subunit 2 beta) and ~100kb upstream of *ASIP* (agouti signaling protein), and is in high linkage disequilibrium (r^2=0.96 and D'=1) with rs4911414, which has been associated with melanoma and basal cell carcinoma(9) (**Supplementary Figure 2).** The T allele at rs2284378 was associated with an increased ER-negative breast cancer risk (OR>1) in all racial/ethnic populations, except Japanese (OR=0.99) (**Table 1**). However this group had the smallest sample size. Furthermore, no significant evidence of heterogeneity was observed by race (*P*=0.28) or study (*P*=0.54) **(Table 1, Supplementary Table 3)**. When the study was extended to include all available breast cancer cases (ER-positive and ER-negative) and controls from the participating GWAS, rs2284378 showed a weaker association with overall breast cancer (OR=1.08, $P=1.3x10^{-6}$ based on 17,868 cases and 43,744 controls; **Table 1**) and no evidence for association with ER-positive disease (OR=1.01, *P*=0.67 based on 9,965 cases and 22,902 controls **(Supplementary Table 5)**. A case-only analysis of ER-negative versus ER-positive breast cancer indicated a highly significant difference in ORs by ER status ($P=1.3x10^{-4}$, **Supplementary Table 5**). Furthermore, rs2284378 appeared more strongly associated with triple negative (TN) breast cancer (OR=1.16; $P=6.4x10^{-3}$), than ER-negative, PR-negative, HER2-positive breast cancer (OR=1.07, *P*=0.41), although these differences were not statistically significant (case-only *P*=0.44) **(Supplementary Table 5)**.

Next, we examined the associations between all candidate loci from stage 1 (n=86 SNPs) and overall breast cancer risk using all available breast cancer cases and controls from the studies in stages 1 and 2 (**Figure 1**). We identified genome-wide statistically significant associations with variants at 6q25 (rs9383938, T allele: OR=1.20; P=8.7x10^{-14}), and a recently reported risk locus near the *PTHLH* gene at 12p11 (rs1975930, T allele: OR=1.22; P=1.4x10^{-13})(10). In addition, we observed genome wide significant associations with multiple variants in a gene-desert located at 6q14. Allele C of rs17530068 at 6q14 was associated with increased risk for overall breast cancer risk (OR=1.12; P=1.1x10^{-9}) (**Table 2, Supplementary Figure 3**, **Supplementary Table 4)** and both ER-positive (OR=1.09; P=1.5x10^{-5}) **(Supplementary Table 6)** and ER-negative (OR=1.16, P=2.5x10^{-7}) **(Table 2)** breast cancer. We observed no evidence of risk heterogeneity for rs17530068 by ER status (case-only analysis P=0.53) (**Supplementary Table 6**); study (P_{het}=0.16); or race/ethnicity (P_{het} =0.30) **(Table 2).** Furthermore, rs17530068 appeared more strongly associated with ER-negative, PR-negative, HER2-positive breast cancer (OR=1.26, P=8.0x10^{-3}), than TN breast cancer (OR=1.12, P=0.07), although these differences were not statistically significant (case-only P=0.17) **(Supplementary Table 6)**.

We also evaluated associations for 25 known breast cancer risk markers in European-ancestry women from our study (**Supplementary Table 7 and Supplementary Figure 4**). In our samples 8 of the 13 markers previously associated with both ER-negative and ER-positive disease or with ER-negative disease only (TERT and 19p13.1), were nominally significantly associated (P<0.05) with ER-negative disease. In contrast, none of the 10 markers previously associated with ER-positive disease only were associated with ER-negative disease. A risk score formed by summing the risk alleles at all 25 previously identified loci was significantly associated with ER-negative disease in our study (OR=1.06 (1.04-1.07); P=2.9 x10^{-14}). Risk scores for subsets of markers associated with ER-negative disease only (2 markers) or both ER-negative and ER-positive disease (11 markers) were also significantly associated with ER-negative disease (OR=1.22 (1.14-1.31), P=1.0 x10^{-8} and OR=1.08 (1.05-1.10), P=9.5 x10^{-12}, respectively). A risk score for the subset of loci previously associated with ER-positive disease only (10 markers) was not associated with risk of ER-negative disease (OR=1.02 (1.00-1.04), P=0.08). These score

results provide some confirmation of earlier results and an estimate of the effects of previously-identified breast cancer risk markers on risk of ER-negative disease.

DISCUSSION

We present results from the largest meta-analysis to date to specifically focus on ER-negative disease. We identify two novel loci for breast cancer: 20q11 associated with ER-negative and triple negative, but not ER-positive breast cancer, and 6q14 associated with both ER-positive and ER-negative breast cancer. In addition, we confirm three known regions previously associated with ER-negative (19p13) or ER-negative and ER-positive breast cancer (6q25 and 12p11). Correction for genomic control results in similar but attenuated findings for 20q11-rs2284378 (P_{GC}=2.4x10^{-8}) and 6q14-rs17530068 (P_{GC} =3.2x10^{-9}).

The novel association at 20q11 with ER-negative breast cancer spans the *ASIP, RALY* and *EIF2S2* genes. Agouti signaling protein (product of the *ASIP* gene) was first described to inhibit melanogenesis in human melanocytes in 1997(11). ASIP is a melanocortin 1 receptor (MC1R) ligand that antagonises the function of the transmembrane receptor(12). The variants we identified at 20q11 for breast cancer are highly correlated with variants previously associated with pigmentation traits as well as risk of both cutaneous melanoma and basal cell carcinoma(9), suggesting a possible biological link between these cancers. Further studies have confirmed the importance of the genetic variation spanning the *ASIP* locus, where a variant at 20q11 showed the strongest association with pigmentation and was implicated in a probable linkage disequilibrium (LD) with variants within an *ASIP* regulatory region(13). *EIF2S2* encodes eukaryotic translation initiation factor 2, subunit 2 beta, which is involved in early steps of protein synthesis by forming a ternary complex with GTP and initiator tRNA. The deletion of *Eif2s2* has been associated with suppression of testicular germ cell tumor incidence and recessive lethality in mice(14). The agouti-yellow (*AV*) deletion is a genetic modifier known to suppress testicular germ cell tumor susceptibility in mice and humans. The *AV* mutation deletes both *RALY* and *Eif2s2,* and induces the ectopic expression of *agouti,* all of which are potential testicular germ cell tumor-modifying variations

(14). Both *RALY* and *EIF2S2* are expressed in many tissues including mammary gland(15). SNP rs2284378 was not consistently associated with expression of *EIF2S2*, *RALY*, or *ASIP* in lymphocytes (11), adipocytes or skin cells(16)although there was marginal evidence for association between rs2284378 and *EIF2S2* expression in one study (16)(**Supplementary Table 8**). However, several SNPs in high linkage disequilibrium with SNP rs2284378 (r2>0.8) within a 1MB region were significantly associated with expression of nearby genes *EIF2S2* and *RALY*. Rs4911379 (r^2=0.96) is statistically significantly associated with *EIF2S2* expression in fibroblasts (*P*=3.6 x10^{-4}) (17)and SNPs rs761238 and rs761236 (r^2=0.85) are associated with *RALY* expression in lymphocytes (P=8.3x10^{-4})(16). An additional 13 SNPs (r^2>0.85) have been associated with expression of *RALY*, *GGTL3*, *DYNLRB1*, and *AK054906* in liver cells, monocytes and lymphoblastoid cell lines (**Supplementary Table 9**). In addition to expression, several enhancer as well as promoter regions defined by overlapping chromatin marks in human mammary epithelial cells were found at 20q11 (**Supplemental Figure 5**). SNPs in high LD with rs2284378 (r^2>0.7), such as rs4911395, rs4911396 and rs1007090, are located in the promoter region of RALY. SNPs rs6142101, rs6087557, and rs4911408 (r^2>0.7) are present in the promoter region of *EIF2S*2, and rs1054534, rs1555075, rs2268086, rs2268088, rs4911401, rs2284388, rs2284389 and rs932388 are located in predicted enhancer regions in introns of *RALY*. Thus, variants at 20q11 may influence expression of multiple genes in mammary epithelial cells, as has been seen in prostate cancer (18).

In contrast, rs17530068 at 6q14 is located in a gene desert with no evidence of an open/active regulatory region in human mammary epithelial cells (**Supplementary Figure 6**). The closest gene (~262kb), family with sequence similarity 46, member A (*FAM46A*/*C6orf37*), encodes a protein of unknown function. Five SNPs in this region in low linkage disequilibrium with SNP rs17530068 (r2<0.02) were associated with expression of *IBTK* in lymphoblastoid cell lines (**Supplementary Table 10).** Additional studies of both of these novel regions will be necessary to identify the underlying biologically relevant variant/s.

SNP rs17530068 at chromosome 6q14 was associated with overall breast cancer risk and showed no differential association depending on ER status. The association of SNP rs2284378 at 20q11, however, was stronger for ER-negative than ER-positive breast cancer. This finding underscores the importance of investigating genetic variants for specific subtypes of breast cancer, as this locus had not been previously identified in the many GWAS of breast cancer to date that did not focus on this specific breast cancer subtype. The etiology of ER-negative disease is largely unknown. Identifying new loci associated with ER-negative and TN breast cancer will continue to provide insight into the biological mechanisms underlying this more aggressive form of breast cancer, and could result in improvements in risk prediction and treatment.

MATERIALS AND METHODS

Study populations

Stage 1 included the studies of the NCI Breast and Prostate Cancer Cohort Consortium (BPC3), Triple Negative Breast Cancer Consortium (TNBCC) and African American Breast Cancer Consortium (AABC). The BPC3 study includes 2,188 ER-negative cases and 25,519 controls, AABC includes 3,153 cases (1,004 ER-negative) and 2,745 controls from 9 studies and TNBCC includes 1,562 cases and 3,399 controls from 15 studies **(Supplementary Table 1)**. Replication studies include 886 cases (84 ER-negative) and 830 controls from a GWAS of breast cancer in Japanese (MEC-JPT) women and 546 cases (112 ER-negative) and 558 controls from a GWAS of breast cancer in Latino (MEC-LAT) women in the Multiethnic Cohort (MEC), 992 (188 ER-negative) and 640 controls from the San Francisco Bay Area Breast Cancer Study (SFBCS) and the Northern California Breast Cancer Family Registry (NC-BCFR), and 8,785 (562 ER-negative) and 14,029 controls from eight combined GWAS of breast cancer from BCAC. All participants in these studies have provided written consent for the research and approval for the study was obtained from the ethical review board from all local institutions. A description of each participating study has been provided in supplementary material.

Stage 1 genotyping and quality control

Genotyping in AABC was conducted using the Illumina Human1M-Duo BeadChip. Of the 5,984 samples in the AABC Consortium (3,153 cases and 2,831 controls), we attempted genotyping of 5,932, removing samples (n=52) with DNA concentrations <20 ng/ul. Following genotyping, we removed samples based on the following exclusion criteria: 1) unknown replicates (≥98.9% genetically identical) that we were able to confirm, n=15); 2) unknown replicates pair or triplicate removed, n=14); 3) samples with call rates <95% after a second attempt (n=100); 4) samples with ≤ 5% African ancestry (n=36) (discussed below); and, 5) samples with <15% mean heterozygosity of SNPs in the X chromosome and/or similar mean allele intensities of SNPs on the X and Y chromosomes (n=6). In the analysis, we removed SNPs with <95% call rates (n=21,732) or minor allele frequencies (MAFs) <1% (n=80,193). The concordance rate for blinded duplicates was 99.95%. We also eliminated SNPs with genotyping concordance rates <98% based on the replicates (n=11,701). The final analysis dataset included 1,043,036 SNPs genotyped on 3,016 cases (988 ER-negative, 1520 ER-positive, and the remaining 508 cases with unknown ER status) and 2,745 controls, with an average SNP call rate of 99.7% and average sample call rate of 99.8%.

Genotyping for the TNBCC GWAS was conducted on 1,718 cases from 10 studies (ABCTB, BBCC, DFCI, FCCC, GENICA, MARIE, MCBCS, MCCS, POSH, SBCS) using the Illumina 660-Quad SNP array. In addition, a subset of MARIE cases (n=52) were genotyped using the Illumina CNV370 SNP array. HEBCS cases (n=85) were genotyped using the Illumina 550 SNP array and population allele and genotype frequencies on healthy population controls (n=222) were genotyped on Illumina 370 SNP array, and obtained from the NordicDB, a Nordic pool and portal for genome-wide control data(19) from the Finnish Genome Center. GWAS data for public controls (n=3,448) were generated using the following arrays: Illumina 660-Quad SNP array (QIMR), Illumina 550 SNP array (CGEMS), Illumina 550 SNP array (KORA), and Illumina 1.2M (WTCCC). These GWAS data were independently evaluated by an iterative QC process with the following exclusion criteria: minor allele frequency (MAF) <0.01, call rate <95%, HWE p-value $<1x10^{-7}$ among controls and sample call rate <98%. In total, we excluded previously unknown replicates (n=2) and samples with call rates <98% (n=83), samples that failed sex

check (n=10), cases identified as non-triple negative breast cancer (n=20) and related samples (n=27). We removed SNPs with <95% call rates or MAF <5%. Because a number of our samples were genotyped at different locations, we removed SNPs if there was a difference >0.10 between the study allele frequency and the median frequency across all studies. Eigensoft software which uses principle component analysis (PCA) was used to evaluate confounding due to population stratification. We removed 101 subjects that did not cluster with the CEU HapMap Phase 2 samples, and a further 179 controls were removed which overlapped with CGEMS/NHS controls in BPC3, resulting in 1,562 cases and 3,399 controls in the GWAS analyses.

BPC3 GWAS genotyping was conducted at three genotyping centers (NCI Core Genotyping Facility, USA; University of Southern California, USA; and Imperial College London, UK). Subjects from CPSII, EPIC, MEC, PLCO, and PBCS were genotyped using the Illumina Human 660k-Quad SNP array(Illumina, Inc), NHSI/NHSII and part of the PLCO study were genotyped previously using the Illumina Human 550 SNP array(Illumina, Inc) (20). SNPs were filtered and removed based on deviations from Hardy-Weinberg proportions in control subjects (p<10e-5), autosomal SNPs with MAF of less than 5% and completion rate less than 95%. Samples were excluded based on genotyping call rates less than 95% (n=195), samples with extreme heterozygosity were excluded from the analysis (n=35), sex discordance (n=3), unexpected duplicates and relatedness (n=6), Subjects with evidence of significant non-European ancestry and population structure were also excluded. Non-European ancestry was assessed utilizing a subset of unlinked, population informative SNPs (21). Individuals determined to have less than 80% European ancestry were excluded from future analyses (n=16). The average concordance rate of blinded duplicates was 99.95%. In order to resolve a more detailed population substructure, PCA was conducted using *struct.pca* module of GLU (http://code.google.com/p/glu-genetics/). PCA was only performed in subjects with over 80% European ancestry. Furthermore, 958 controls from NHS (CGEMS) were removed from BPC3 analyses due to overlap between TNBCC and BPC3 studies. The overall number of cases and controls after all exclusions which contributed to the stage 1 analysis were 1,998 cases and 2,305 controls.

The WHS cohort subjects in BPC3 were previously genotyped using the Human-Hap300 Duo-plus BeadChip (22). Among the final 23,294 individuals of verified European ancestry, genotypes for a total of 2,608,509 SNPs were imputed from the experimental genotypes and LD relationships implicit in the HapMap r. 22 CEU samples. WHS contributed 190 cases and 23,214 control subjects to stage 1. WHS was meta-analyzed with the remaining BPC3 studies contributing a total of 2,188 cases and 25,519 control subjects to stage 1 analysis.

SNP rs2284378 and rs17530068 were genotyped in all stage 1 studies.

Stage 2 genotyping and quality control

The San Francisco Bay Area Breast Cancer Study (SFBCS)(23)and the Northern California Breast Cancer Family Registry (NC-BCFR)(24) study samples were genotyped with the Affymetrix 6.0 array according to the manufacturer's instructions (https://www.affymetrix.com) in the laboratory of Esteban Gonzalez Burchard at UCSF. A total of 15 cases and 30 controls were excluded from the SFBCS and NC-BCFR sample set that had a genotyping call rate <95% or showed either known or cryptic relatedness. The final sample included in the analysis was 992 cases (188 ER-negative cases) and 640 controls. Imputation was conducted with the program BEAGLE, with all unrelated HapMap Phase II samples included as references (http://hapmap.ncbi.nlm.nih.gov).

GWAS of breast cancer in Latino (MEC-LAT) and Japanese (MEC-JPT) samples from the MEC were genotyped with the Illumina 660W array at USC. For MEC-LAT, we excluded 48 samples from the MEC that had a genotyping call rate of <95% and 34 that showed either known or cryptic relatedness. The final MEC-LAT sample included 546 (112 ER-negative) and 558 controls. With similar exclusions, the final MEC-JPT sample included 886 (84 ER-negative) and 830 controls.

The BCAC combined GWAS includes primary genotype data from eight breast cancer GWAS in populations of European ancestry (ABCFS, BBCS, , GC-HBOC, MARIE, HEBCS, SASBAC, UK2, DFBBCS). All studies were genotyped with various versions of Illumina arrays, except GC-HBOC which was performed with the Affymetrix 5.0 (cases) and 6.0 (controls) arrays. Standard QC was performed on all scans. Specifically, all individuals with low call rate (<95%), extreme high or low

heterozygosity ($P<10^{-5}$), and all individuals evaluated to be of non-European ancestry (>15% non-European component, by multidimensional scaling using the three Hapmap2 populations as a reference) were excluded. SNPs with call rate <95%; call rate <99% and MAF<5%, all SNPs with MAF<1%, and SNPs with genotype frequencies departing from Hardy-Weinberg equilibrium at $P<10^{-6}$ in controls or $P<10^{-12}$ in cases were also excluded. Data were imputed for ~2.6M SNPs for all scans using Mach v1.0 with HapMap version 2 CEU as a reference. BBCS and UK2 used the same control data (WTCCC2). These studies were imputed separately. For the combined analysis, the control set was divided randomly between the two studies, in proportion to the size of case series, to provide disjoint strata. Estimated per-allele ORs and standard errors were generated from the imputed genotypes using Probabel (25).

SNP rs2284378 and rs17530068 were genotyped in all stage 2 studies except SFBCS and NC-BCFR where they were imputed. Both SNPs were genotyped by TaqMan in 483 samples from these studies and genotype concordance versus imputed genotypes was 93.3% for rs2284378 and 94.9% for rs17530068.

Taqman gentoyping in BPC3 for SNP rs2284378 and SNP rs17530068

In BPC3, genotyping of SNP rs2284378 and rs17530068 was performed for all available breast cancer cases and controls by TaqMan in four laboratories (CPS-II and MEC at the University of Southern California; NHS and WHS at Harvard University; EPIC at the German Cancer Research Center in Heidelberg; and PLCO at the NCI/Core Genotyping Facility). All studies typed SNP rs17530068; however for SNP rs2284378, PLCO and CPS-II typed a proxy SNP rs6059651 (r^2 =1, D'=1). The concordance for the Taqman genotyping data with that generated from Illumina for stage 1 ER-negative cases and controls was 0.997 for rs17530068 and 0.986 for rs2284378 for CPS2, MEC, NHS, EPIC and PLCO. The genotype concordance versus imputed for WHS was 95% for rs2284378 and 97% for rs17530068

Statistical analysis

In AABC, we tested for gene dosage effects in models adjusted for age, study and eigenvectors 1-10. Odds ratios (OR) and 95% confidence intervals (95% CI) were estimated using unconditional logistic regression. In TNBCC, unconditional logistic regression was used to assess single SNP associations also assuming a log-additive model, adjusting for country and the first two principal components. In BPC3, unconditional logistic regression model was used to assess single SNP associations adjusting for age categories and the top 6 eigenvectors.

In both AABC and TNBCC, phased haplotype data from the founders of the CEU and YRI HapMap Phase 2 samples (build 21) were used to infer LD patterns in order to impute untyped markers. For BPC3, Hapmap Phase 2 (release 21) and Hapmap Phase 3 were used to impute untyped markers. For all studies, genome-wide imputation was carried out using the software MACH. Filtered from the analysis were SNPs with Rsq<0.3 and MAF <1%.

We conducted a fixed effect meta-analysis of AABC, TNBCC and BPC3 using the inverse variance weighted method. The number of SNPs available for meta-analysis from AABC, TNBCC and BPC3 in stage 1 were 3,055,415, 2,134,490 and 245,3207 respectively. The union of these three data sets was meta-analyzed using the program METAL. We conducted *in silico* replication of 86 SNP with p-values $\leq 10^{-5}$ in stage 1 in the stage 2 studies, and a meta-analysis of these SNPs from stage 1 and 2 for both ER- negative and overall breast cancer. P-values from our top two loci were corrected for genomic inflation (P_{GC}) using the lambda value from the overall meta-analysis. Testing for heterogeneity by study was evaluated using the Q-statistic. Case-only analyses were performed to test for differences in the association by tumor subtypes, study and race/ethnicity.

The association between risk scores of 25 previously-identified breast cancer risk alleles and risk of breast cancer in our samples was calculated using meta-regression, assuming the per-allele odds ratio was constant across the markers analyzed. This is equivalent to combining the summary log odds ratio estimates at independent loci using inverse-variance weighted meta-analysis. The overlap between subjects contributing to this study and those contributing to previous studies varied from marker to marker (e.g. the TNBCC contributed to the initial report on rs8170 (5) and the BPC3 and TNBCC

contributed to the initial report on the *TERT* locus (6).Thus, the results could be overestimates since some of the studies here contributed to the discovery of these 25 loci.

Functional analysis

Expression quantitative trait loci (eQTL) were assessed for all SNPs in the chromosome 6 and 20 loci using the GTEX database (http://www.ncbi.nlm.nih.gov/gtex/GTEX2/gtex.cgi), University of Chicago eQTL Browser (http://eqtl.uchicago.edu) and Genevar (http://www.sanger.ac.uk/resources/software/genevar/) (26)

In an attempt to identify functionality at the two novel breast cancer risk loci, we used the open-source R/Bioconductor package FunciSNP version 0.99(27), which systematically integrates the 1,000 Genomes Project SNP data (April 2012 data release) with chromatin features of interest. For each of the two novel breast cancer markers, we analyzed all SNPs with an r^2 value > 0.5 with each index SNP in the 1,000 Genomes Project EUR populations in a 1MB window around each index variant. We assessed whether these SNPs were co-located with 12 different chromatin features generated by next-generation sequencing technologies, which capture open chromatin regions, promoters, and enhancers genome-wide in human mammary epithelial cells (HMEC) as well as known DNaseI hypersensitive locations, FAIRE-seq peaks, and CTCF binding sites from more than 100 different cell types, which were collected in ENCODE data(28). We utilized the UCSC Genome Browser (http://genome.ucsc.edu/) to illustrate the correlated SNPs, which overlap chromatin features as well as chromatin feature tracks (**Supplemental Figures 5-6**).

FUNDING

AABC was supported by a Department of Defense Breast Cancer Research Program Era of Hope Scholar Award to CAH [W81XWH-08-1-0383] and the Norris Foundation. Each of the participating **AABC** studies was supported by the following grants: MEC (National Institutes of Health grants R01-CA63464 and R37-CA54281); CARE (National Institute for Child Health and Development grant NO1-HD-3-3175), WCHS (U.S. Army Medical Research and Material Command (USAMRMC) grant DAMD-17-01-0-0334, the National Institutes of Health grant R01-CA100598, and the Breast Cancer Research Foundation, SFBCS (National Institutes of Health grant R01-CA77305 and United States Army Medical Research Program grant DAMD17-96-6071), NC-BCFR (National Institutes of Health grant U01-CA69417), CBCS (National Institutes of Health Specialized Program of Research Excellence in Breast Cancer, grant number P50-CA58223, and Center for Environmental Health and Susceptibility, National Institute of Environmental Health Sciences, National Institutes of Health, grant number P30-ES10126), PLCO (Intramural Research Program, National Cancer Institute, National Institutes of Health), and NBHS (National Institutes of Health grant R01-CA100374),WFBC (National Institutes of Health grant R01-CA73629). TheNC-BCFR is one of 6 sites participating in The Breast Cancer Family Registry (BCFR) which was supported by the National Cancer Institute, National Institutes of Health under RFA CA-06-503 and through cooperative agreements with members of the Breast Cancer Family Registry and Principal Investigators. The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the BCFR, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government or the BCFR.

The **TNBCC** studies were supported by the following grants: MCBCS (National Institutes of Health Grants CA122340 and a Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA116201), and the Breast Cancer Research Foundation (BCRF); MARIE (Deutsche Krebshilfe e.V., grant number 70-2892-BR I, the Hamburg Cancer Society, the German Cancer Research Center (DKFZ) and the Federal Ministry of Education and Research (BMBF) Germany grant 01KH0402); GENICA (Federal Ministry of Education and Research (BMBF) Germany grants 01KW9975/5, 01KW9976/8,

01KW9977/0, 01KW0114, and the Robert Bosch Foundation Stuttgart, Germany;MCCS (Australian NHMRC grants 209057, 251553 and 504711 and infrastructure provided by the Cancer Council Victoria); SBCS (Breast Cancer Campaign (grant 2004Nov49 to AC), and by Yorkshire Cancer Research core funding); DFCI (DFCI Breast Cancer SPORE NIH P50 CA089393); POSH (Cancer Research UK); DEMOKRITOS (Hellenic Cooperative Oncology Group research grant (HR R_BG/04) and the Greek General Secretary for Research and Technology (GSRT) Program, Research Excellence II, funded at 75% by the European Union); BBCC (Dr. Mildred Scheel Stiftung of the Deutsche Krebshilfe e.V.); BBCS (Cancer Research UK and Breakthrough Breast Cancer and NHS funding to the NIHR biomedical Research Centre and the National Cancer Research Network (NCRN); LMBC (European Union Framework Programme 6 Project LSHC-CT-2003-503297 (the Cancerdegradome) and by the 'Stichting tegen Kanker' (232-2008); OBCS (Finnish Cancer Foundation, the Sigrid Juselius Foundation, the Academy of Finland, the University of Finland, and Oulu University Hospital); HEBCS (Helsinki University Central Hospital Research Fund, Academy of Finland (132473), the Finnish Cancer Society, The Nordic Cancer Union and the Sigrid Juselius Foundation); FCCC (U01CA69631, 5U01CA113916, the University of Kansas Cancer Center and the Kansas Bioscience Authority Eminent Scholar Program); RPCI (RPCI DataBank and BioRepository (DBBR), a Cancer Center Support Grant Shared Resource (P30 CA016056-32); SKKDKFZS (Deutsches Krebsforschungszentrum); BIGGS (National Institute for Health Research (NIHR) Comprehensive Biomedical Research Centre, Guy's & St. Thomas' NHS Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust); ABCTB (National Health and Medical Research Council of Australia, The Cancer Institute NSW and the National Breast Cancer Foundation); ABCS (Dutch Cancer Society grant number 2009-4363); KARBAC (The Stockholm Cancer Society).

BPC3 is supported by the US National Institutes of Health, National Cancer Institute under cooperative agreements U01-CA98233 (NHS, NHSII, WHS), U01-CA98710 (CPS2), U01-CA98216 (EPIC), U01-CA98758 (MEC) and Intramural Research Program of NIH/National Cancer Institute, Division of Cancer Epidemiology and Genetics (PLCO). The authors thank Drs. Christine Berg and

Philip Prorok, Division of Cancer Prevention, NCI, the screening center investigators and staff of the PLCO Cancer Screening Trial, Mr. Thomas Riley and staff at Information Management Services, Inc., and Ms. Barbara O'Brien and staff at Westat, Inc. for their contributions to the PLCO Cancer Screening Trial.

The WHS is supported by HL043851 and HL080467 from the National Heart, Lung, and Blood Institute and CA 047988 from the National Cancer Institute, the Donald W. Reynolds Foundation and the Fondation Leducq, with collaborative scientific support and funding for genotyping provided by Amgen.

The **UK2** GWAS was funded by Wellcome Trust and Cancer Research UK. The WTCCC was funded by the Wellcome Trust. **BCAC** is funded by CR-UK [C1287/A10118, C1287/A12014] and by the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS). Meetings of the BCAC have been funded by the European Union COST programme [BM0606]. The **ABCFS** study was supported by the National Health and Medical Research Council of Australia, the New South Wales Cancer Council, the Victorian Health Promotion Foundation (Australia), and the National Cancer Institute, National Institutes of Health under RFA-CA-06-503 and through cooperative agreements with members of the Breast Cancer Family Registry (BCFR) and the Principle Investigators. The University of Melbourne (U01 CA69638) contributed data to this study. The content of this manuscript does not necessarily reflect the views or the policies of the National Cancer Institute or any of the collaborating centers in the BCFR, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government or the BCFR. We extend our thanks to the many women and their families that generously participated in the Australian Breast Cancer Family Study and consented to us accessing their pathology material. JLH is a National Health and Medical Research Council Australia Fellow. MCS is a National Health and Medical Research Council Senior Research Fellow. JLH and MCS are both group leaders of the Victoria Breast Cancer Research Consortium. The **BBCS** is funded by Cancer Research UK and Breakthrough Breast Cancer and acknowledges NHS funding to the NIHR Biomedical Research Centre, and the National Cancer Research Network (NCRN). The BBCS GWAS received funding from The Institut National de Cancer. The

DFBBCS GWAS was funded by The Netherlands Organisation for Scientific Research (NWO) as part of a ZonMw/VIDI grant number 91756341. We thank Muriel Adank for selecting the samples and Margreet Ausems, Christi van Asperen, Senno Verhoef, and Rogier van Oldenburg for providing samples from their Clinical Genetic centers. The **GC-HBOC** was supported by Deutsche Krebshilfe [107054], the Dietmar-Hopp Foundation, the Helmholtz society and the German Cancer Research Centre (DKFZ). The GC-HBOC GWAS was supported by the German Cancer Aid (grant no. 107352). The **MARIE** study was supported by the Deutsche Krebshilfe e.V. [70-2892-BR I], the Hamburg Cancer Society, the German Cancer Research Center and the genotype work in part by the Federal Ministry of Education and Research (BMBF) Germany [01KH0402]. MARIE would like to thank Tracy Slanger and Elke Mutschelknauss for their valuable contributions, and S. Behrens, R. Birr, M.Celik, U. Eilber, B. Kaspereit, N. Knese and K. Smit for their excellent technical assistance. The **SASBAC** study was supported by funding from the Agency for Science, Technology and Research of Singapore (A*STAR), the US National Institute of Health (NIH) and the Susan G. Komen Breast Cancer Foundation. **CGEMS.** The Nurses' Health Studies are supported by NIH grants CA 65725, CA87969, CA49449, CA67262, CA50385 and 5UO1CA098233. The **HEBCS** study has been financially supported by the Helsinki University Central Hospital Research Fund, Academy of Finland (132473), the Finnish Cancer Society, The Nordic Cancer Union and the Sigrid Juselius Foundation. The population allele and genotype frequencies were obtained from the data source funded by the Nordic Center of Excellence in Disease Genetics based on samples regionally selected from Finland, Sweden and Denmark. We thank Drs. Kirsimari Aaltonen, Päivi Heikkilä and Tuomas Heikkinen and RN Hanna Jäntti and Irja Erkkilä for their help with the HEBCS data and samples.

The biofeature analysis was supported by NIH grant CA109147.

The breast cancer **GWAS in Japanese and Latinos** in the MEC (MEC-LAT and MEC-JPT) were supported by NIH grants CA132839, CA54281 and CA63464. Genotyping of the Latino breast cancer cases and controls from SFBCS and NC-BCFR was supported by NIH grant **CA120120**..

ACKNOWLEDGEMENTS

We thank the women who volunteered to participate in each study. We also thank Madhavi Eranti, Andrea Holbrook, Paul Poznaik, and David Wong from the University of Southern California for their technical support. We would also like to acknowledge co-investigators from the WCHS study: Dana H. Bovbjerg (University of Pittsburgh), Lina Jandorf (Mount Sinai School of Medicine) and Gregory Ciupak, Warren Davis, Gary Zirpoli, Song Yao and Michelle Roberts from Roswell Park Cancer Institute.

CONFLICT OF INTEREST

The authors have no conflicts of interest to declare.

REFERENCES

1 Parl, F.F., Schmidt, B.P., Dupont, W.D. and Wagner, R.K. (1984) Prognostic significance of estrogen receptor status in breast cancer in relation to tumor stage, axillary node metastasis, and histopathologic grading. *Cancer*, **54**, 2237-2242.

2 Yang, X.R., Chang-Claude, J., Goode, E.L., Couch, F.J., Nevanlinna, H., Milne, R.L., Gaudet, M., Schmidt, M.K., Broeks, A., Cox, A. *et al.* (2011) Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies. *J. Natl. Cancer. Inst.*, **103**, 250-263.

3 Milne, R.L. and Antoniou, A.C. (2011) Genetic modifiers of cancer risk for BRCA1 and BRCA2 mutation carriers. *Ann. Oncol.*, **22 Suppl 1**, i11-17.

4 Broeks, A., Schmidt, M.K., Sherman, M.E., Couch, F.J., Hopper, J.L., Dite, G.S., Apicella, C., Smith, L.D., Hammet, F., Southey, M.C. *et al.* (2011) Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. *Hum. Mol. Genet.*, **20**, 3289-3303.

5 Antoniou, A.C., Wang, X., Fredericksen, Z.S., McGuffog, L., Tarrell, R., Sinilnikova, O.M., Healey, S., Morrison, J., Kartsonaki, C., Lesnick, T. *et al.* (2010) A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat. Genet.*, **42**, 885-892.

6 Haiman, C.A., Chen, G.K., Vachon, C.M., Canzian, F., Dunning, A., Millikan, R.C., Wang, X., Ademuyiwa, F., Ahmed, S., Ambrosone, C.B. *et al.* (2011) A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat. Genet.*, **43**, 1210-1214.

7 Stevens, K.N., Vachon, C.M., Lee, A.M., Slager, S., Lesnick, T., Olswold, C., Fasching, P.A., Miron, P., Eccles, D., Carpenter, J.E. *et al.* (2011) Common breast cancer susceptibility loci are associated with triple-negative breast cancer. *Cancer Res.*, **71**, 6240-6249.

8 Zheng, W., Long, J., Gao, Y.T., Li, C., Zheng, Y., Xiang, Y.B., Wen, W., Levy, S., Deming, S.L., Haines, J.L. *et al.* (2009) Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.*, **41**, 324-328.

9 Gudbjartsson, D.F., Sulem, P., Stacey, S.N., Goldstein, A.M., Rafnar, T., Sigurgeirsson, B., Benediktsdottir, K.R., Thorisdottir, K., Ragnarsson, R., Sveinsdottir, S.G. *et al.* (2008) ASIP and TYR pigmentation variants associate with cutaneous melanoma and basal cell carcinoma. *Nat. Genet.*, **40**, 886-891.

10 Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C. *et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, **6**, e107.

11 Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E., Sekowska, M., Smith, G.D., Evans, D., Gutierrez-Arcelus, M. *et al.* (2012) Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.*, **8**, e1002639.

12 Scherer, D. and Kumar, R. (2010) Genetics of pigmentation in skin cancer--a review. *Mutat. Res.*, **705**, 141-153.

13 Barrett, J.H., Iles, M.M., Harland, M., Taylor, J.C., Aitken, J.F., Andresen, P.A., Akslen, L.A., Armstrong, B.K., Avril, M.F., Azizi, E. *et al.* (2011) Genome-wide association study identifies three new melanoma susceptibility loci. *Nat. Genet.*, **43**, 1108-1113.

14 Heaney, J.D., Michelson, M.V., Youngren, K.K., Lam, M.Y. and Nadeau, J.H. (2009) Deletion of eIF2beta suppresses testicular cancer incidence and causes recessive lethality in agouti-yellow mice. *Hum. Mol. Genet.*, **18**, 1395-1404.

15 Mosca, E., Alfieri, R., Merelli, I., Viti, F., Calabria, A. and Milanesi, L. (2010) A multilevel data integration resource for breast cancer study. *BMC Syst. Biol.*, **4**, 76.

16 Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K. *et al.* (2011) The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.*, **7**, e1002003.

17 Dimas, A.S., Deutsch, S., Stranger, B.E., Montgomery, S.B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M. *et al.* (2009) Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science*, **325**, 1246-1250.

18 Pomerantz, M.M., Shrestha, Y., Flavin, R.J., Regan, M.M., Penney, K.L., Mucci, L.A., Stampfer, M.J., Hunter, D.J., Chanock, S.J., Schafer, E.J. *et al.* (2010) Analysis of the 10q11 cancer risk locus implicates MSMB and NCOA4 in human prostate tumorigenesis. *PLoS Genet.*, **6**, e1001204.

19 Leu, M., Humphreys, K., Surakka, I., Rehnberg, E., Muilu, J., Rosenstrom, P., Almgren, P., Jaaskelainen, J., Lifton, R.P., Kyvik, K.O. *et al.* (2010) NordicDB: a Nordic pool and portal for genome-wide control data. *Eur. J. Hum. Genet.*, **18**, 1322-1326.

20 Hunter, D.J., Kraft, P., Jacobs, K.B., Cox, D.G., Yeager, M., Hankinson, S.E., Wacholder, S., Wang, Z., Welch, R., Hutchinson, A. *et al.* (2007) A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.*, **39**, 870-874.

21 Yu, K., Wang, Z., Li, Q., Wacholder, S., Hunter, D.J., Hoover, R.N., Chanock, S. and Thomas, G. (2008) Population substructure and control selection in genome-wide association studies. *PLoS One.*, **3**, e2551.

22 Ridker, P.M., Chasman, D.I., Zee, R.Y., Parker, A., Rose, L., Cook, N.R. and Buring, J.E. (2008) Rationale, design, and methodology of the Women's Genome Health Study: a genome-wide association study of more than 25,000 initially healthy american women. *Clin. Chem.*, **54**, 249-255.

23 John, E.M., Schwartz, G.G., Koo, J., Wang, W. and Ingles, S.A. (2007) Sun exposure, vitamin D receptor gene polymorphisms, and breast cancer risk in a multiethnic population. *Am. J. Epidemiol.*, **166**, 1409-1419.

24 John, E.M., Miron, A., Gong, G., Phipps, A.I., Felberg, A., Li, F.P., West, D.W. and Whittemore, A.S. (2007) Prevalence of pathogenic BRCA1 mutation carriers in 5 US racial/ethnic groups. *JAMA*, **298**, 2869-2876.

25 Aulchenko, Y.S., Struchalin, M.V. and van Duijn, C.M. (2010) ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics*, **11**, 134.

26 Yang, T.P., Beazley, C., Montgomery, S.B., Dimas, A.S., Gutierrez-Arcelus, M., Stranger, B.E., Deloukas, P. and Dermitzakis, E.T. (2010) Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. *Bioinformatics*, **26**, 2474-2476.

27 Coetzee, S.G., Rhie, S.K., Berman, B.P., Coetzee, G.A. and Noushmehr, H. (2012) FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res.*, June 22 [Epub ahead of print].

28 Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43-49.

LEGEND

Figure 1. Multi-stage study design.

Table 1. Association of SNP rs2284378 (T/C) at chromosome 20q11 and breast cancer risk by study and race/ethnicity

Consortium/ Study	Race/ Ethnicity	Case/ control[a]	RAF (T allele)[b]	OR (95% CI)[c]	*P*-value[d]	$P_{Het\text{-}study}$/ $P_{Het\text{-}race}$[e]
Stage 1 ER-negative cases versus controls						
BPC3	European	2,188/25,519	0.31	1.14 (1.05-1.24)	0.0028	
TNBCC	European	1,478/3,345	0.33	1.18(1.07-1.30)	0.0010	
AABC	African	1,004/2,744	0.16	1.19 (1.03-1.37)	0.020	
Stage 1		**4,670/31,608**		**1.16 (1.09-1.23)**	**6.5×10^{-7}**	**0.85/0.76**
Stage 2 ER-negative cases versus controls						
BCAC Combined GWAS	European	562/6410	0.35	1.10 (0.96-1.25)	0.17	
MEC-JPT	Japanese	84/830	0.26	0.99 (0.68-1.44)	0.95	
MEC-LAT	Latino	112/553	0.29	1.27 (0.94-1.71)	0.13	
SFBCS/NC-BCFR	Latino	188/611	0.29	1.45 (1.13-1.87)	0.004	
Stage 2 (ER-negative)		946/8,404		1.16 (1.04-1.29)	0.0048	**0.98/0.12**
Stage 1+2 (ER-negative)		**5,616/40,012**		**1.16 (1.10-1.22)**	**1.1×10^{-8}**	**0.54/0.28**
All breast cancer cases versus controls						
AABC	African	3,016/2,745	0.16	1.06 (0.95-1.17)	0.30	
BCAC Combined GWAS	European	8,785/10,142	0.35	1.04 (0.99-1.09)	0.11	
MEC-JPT	Japanese	886/830	0.26	1.08 (0.91-1.24)	0.46	
MEC-LAT	Latino	546/553	0.29	1.24 (1.03-1.48)	0.021	
SFBCS/NC-BCFR	Latino	970/611	0.29	1.23 (1.05-1.44)	0.011	
Stage 2 (all cases)		14,202/14,880		1.06 (1.02-1.10)	0.0025	**0.14/0.073**
Stage 1+2 (all cases)		**17,869/43,745**		**1.08 (1.05-1.12)**	**1.3×10^{-6}**	**0.056/0.19**

[a]Number of cases and controls with genotype data for rs2284378. [b]Risk Allele Frequency (RAF) in controls. [c]Adjusted for age, study and principal components in AABC. Adjusted for age and country in TNBCC. Adjusted for age categories and top 6 eigenvectors in BPC3. Adjusted for age and top 10 eigenvectors in MEC-JPT, MEC-LAT and SFBCS/NC-BCFR studies. Combined analysis (Stage1, Stage2 and Stage 1+2) are from the meta-analysis. [d]*P* for trend (1-d.f.). [e]*P* for heterogeneity by study and race/ethnicity, respectively.

Table 2. Association of SNP rs17530068 (C/T) at chromosome 6q14 and breast cancer risk by study and race/ethnicity

Consortium/ Study	Race/ Ethnicity	Case/ control[a]	RAF (C allele)[b]	OR (95% CI)[c]	*P*-value[d]	$P_{Het\text{-}study}$/ $P_{Het\text{-}race}$[e]
Stage 1 ER-negative cases versus controls						
BPC3	European	2,188/25,519	0.24	1.23 (1.12-1.35)	2.23×10^{-5}	
TNBCC	European	1,478/3,345	0.24	1.13 (1.02-1.26)	0.023	
AABC	African	1,004/2,745	0.07	1.07 (0.86-1.34)	0.54	
Stage 1		**4,670/31,609**		**1.17 (1.09-1.26)**	**3.5×10^{-6}**	**0.37/0.41**
Stage 2 ER-negative cases versus controls						
BCAC combined GWAS	European	562/6,410	0.22	1.09 (0.95-1.25)	0.24	
MEC-JPT	Japanese	84/830	0.19	1.16 (0.79-1.71)	0.45	
MEC-LAT	Latino	112/553	0.23	1.06 (0.75-1.50)	0.73	
SFBCS/NC-BCFR	Latino	188/611	0.22	1.40 (1.07-1.84)	0.014	
Stage 2 (ER-negative)		946/8,404		1.14 (1.02-1.28)	0.022	**0.41/0.52**
Stage 1+2 (ER-negative)		**5,616/40,013**		**1.16 (1.10-1.23)**	**2.5×10^{-7}**	**0.54/0.78**
All breast cancer cases versus controls						
AABC	African	3,016/2,745	0.07	1.04 (0.89-1.21)	0.63	
BCAC combined GWAS	European	8,785/10,142	0.22	1.08 (1.02-1.14)	0.0021	
MEC-JPT	Japanese	886/830	0.19	1.13 (0.96-1.34)	0.14	
MEC-LAT	Latino	546/553	0.23	1.21 (0.99-1.47)	0.056	
SFBCS/NC-BCFR	Latino	970/611	0.22	1.27 (1.07-1.51)	0.006	
Stage 2 (all cases)		14,203/14,881		1.10 (1.05-1.15)	1.8×10^{-5}	**0.31/0.20**
Stage 1+2 (all cases)		**17,869/43,745**		**1.12 (1.08-1.16)**	**1.1×10^{-9}**	**0.16/0.30**

[a]Number of cases and controls with genotype data for rs17530068. [b]Risk Allele Frequency (RAF) in controls. [c]Adjusted for age, study and principal components in AABC. Adjusted for age and country in TNBCC. Adjusted for age categories and top 6 eigenvectors in BPC3. Adjusted for age and top 10 eigenvectors in MEC-JPT, MEC-LAT and SFBCS/NC-BCFR studies. Combined analysis (Stage1, Stage2 and Stage 1+2) are from the meta-analysis. [d]*P* for trend. [e]*P* for heterogeneity by study and race/ethnicity, respectively.

ABBREVIATIONS

ER=Estrogen Receptor

PR=Progesterone Receptor

SNP=Single nucleotide polymorphism

GWAS=Genome-wide Association Study

OR=Odds Ratio

BPC3=NCI Breast and Prostate Cancer Cohort Consortium

TNBCC=Triple Negative Breast Cancer Consortium

AABC=African American Breast Cancer Consortium