A comprehensive examination of breast cancer risk loci in African American women

Ye Feng[1], Daniel O. Stram[1], Suhn Kyong Rhie[1], Robert C. Millikan[3,†], Christine B. Ambrosone[4], Esther M. John[5,6], Leslie Bernstein[7], Wei Zheng[8], Andrew F. Olshan[3], Jennifer J. Hu[9], Regina G. Ziegler[10], Sarah Nyante[11], Elisa V. Bandera[12], Sue A. Ingles[1], Michael F. Press[2], Sandra L. Deming[8], Jorge L. Rodriguez-Gil[9], Julie R. Palmer[13], Olufunmilayo I. Olopade[14], Dezheng Huo[15], Clement A. Adebamowo[16], Temidayo Ogundiran[17], Gary K. Chen[1], Alex Stram[1], Karen Park[1], Kristin A. Rand[1], Stephen J. Chanock[10], Loic Le Marchand[18], Laurence N. Kolonel[18], David V. Conti[1], Douglas Easton[19], Brian E. Henderson[1] and Christopher A. Haiman[1,*]

[1]Department of Preventive Medicine, Keck School of Medicine and Norris Comprehensive Cancer Center 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]Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY, USA, [5]Cancer Prevention Institute of California, Fremont, CA, USA, [6]Division of Epidemiology, Department of Health Research & Policy, and Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA, [7]Division of Cancer Etiology, Department of Population Science, Beckman Research Institute, City of Hope, Duarte, 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]Sylvester Comprehensive Cancer Center and Department of Epidemiology and Public Health, University of Miami Miller School of Medicine, Miami, FL, USA, [10]Epidemiology and Biostatistics Program, Division of Cancer Epidemiology and Genetics and [11]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA, [12]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA, [13]Slone Epidemiology Center at Boston University, Boston, MA, USA, [14]Center for Clinical Cancer Genetics & Global Health, Department of Medicine and, [15]Department of Health Studies, University of Chicago, Chicago, IL, USA, [16]Department of Epidemiology & Preventive Medicine, University of Maryland, Baltimore, MD, USA, [17]Department of Surgery, College of Medicine, University of Ibadan, Ibadan, Nigeria, [18]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA and [19]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

Genome-wide association studies have identified 73 breast cancer risk variants mainly in European populations. Given considerable differences in linkage disequilibrium structure between populations of European and African ancestry, the known risk variants may not be informative for risk in African ancestry populations. In a previous fine-mapping investigation of 19 breast cancer loci, we were able to identify SNPs in four regions that better captured risk associations in African American women. In this study of breast cancer in African American women (3016 cases, 2745 controls), we tested an additional 54 novel breast cancer risk variants. Thirty-eight variants (70%) were found to have an association with breast cancer in the same direction as previously reported, with eight (15%) replicating at *P* < 0.05. Through fine-mapping, in three regions (*1q32, 3p24,*

---

*To whom correspondence should be addressed at: USC Norris Comprehensive Cancer Center, Harlyne Norris Research Tower, 1450 Biggy Street, Room 1504, Los Angeles, CA 90033, USA. Tel: +1 3234427755; Fax: +1 3234427749; Email: haiman@usc.edu
†In Memorium.

10q25), we identified variants that better captured associations with overall breast cancer or estrogen receptor positive disease. We also observed suggestive associations with variants (at $P < 5 \times 10^{-6}$) in three separate regions (*6q25, 14q13, 22q12*) that may represent novel risk variants. Directional consistency of association observed for ∼65–70% of currently known genetic variants for breast cancer in women of African ancestry implies a shared functional common variant at most loci. To validate and enhance the spectrum of alleles that define associations at the known breast cancer risk loci, as well as genome-wide, will require even larger collaborative efforts in women of African ancestry.

INTRODUCTION

Genome-wide association studies (GWAS) have identified >70 risk variants for breast cancer (1–15). A large fraction of these discoveries have recently come from the COGS consortium which included follow-up testing of GWAS findings in ∼46 000 cases and ∼42 000 controls and revealed 41 loci for overall breast cancer (12) and four loci associated with estrogen receptor negative (ER−) but not ER positive (ER+) disease (14). Most of the >70 variants that are associated with breast cancer risk were found initially in women of European ancestry. Exceptions include a small number of variants located at *6q25* found in Asians (6,15) and *5p15,* which was identified in a multi-ethnic GWAS meta-analysis that included women of African ancestry in the discovery stage (10). A clear limitation of GWAS in non-European populations is sample size, and continued pooling of GWAS data and large-scale replication testing will be needed to reveal variants that may be unique to or are of particular importance in specific populations. At the same time, comprehensive testing of common genetic variation at known risk loci in multiple racial and ethnic populations will be required to understand the contribution of the locus to risk globally.

Population history has influenced recombination patterns, linkage disequilibrium (LD) structure and the number and frequency of polymorphic alleles between diverse populations. Thus, in the context of exploring genetic variation at known risk loci, a risk variant (i.e. 'index signal') found in European populations might not serve as a surrogate of (or 'tag') the biologically relevant risk variant in African ancestry populations. In addition, the complete spectrum of possible biologically meaningful genetic variation may not be examined if fine-mapping is limited to the population in which the signal was originally detected. We previously developed an analytic framework for fine-mapping of common variation at GWAS risk loci which we applied to testing of an initial set of 19 breast cancer susceptibility regions in an attempt to search for genetic markers that are the most informative for breast cancer risk in women of African ancestry (Materials and Methods) (16). We identified markers in four regions (*2q35, 5q11, 10q26* and *19p13*) that better capture the association with breast cancer risk in African Americans in comparison to the original index signal and thus are likely to be better markers of the biologically functional alleles in this population. We also identified 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 the present study, we have applied this analytical strategy to examine an additional 54 risk variants for breast cancer in 3016 cases and 2745 controls that are part of a breast cancer GWAS in African American women (16). In addition to testing the index signals, we conducted fine-mapping across each locus in search of risk variants that better define breast cancer risk in African Americans as well as secondary signals that are uncorrelated with the index signal and may define novel risk alleles. We also combine these new results with those from our previous report of the 19 loci, and summarize the evidence across all 73 loci.

RESULTS

For the 54 variants included in the analysis (38 genotyped and 16 imputed), the risk allele frequencies ranged from 0.003 for rs11571833 (*13q13*) to 0.98 for rs1353747 (*5q11*); 47 variants were appreciably common in African Americans with risk allele frequencies >0.1 (Supplementary Material, Table S1). Thirty-six of the 54 index variants (67%) showed positive associations (OR > 1) with overall breast cancer risk that were directionally consistent with the initial report of these variants, with seven nominally statistically significant at $P < 0.05$. Of the 54 variants (48 previously reported to be associated with overall breast cancer and six reported to be specifically associated with ER− disease), statistical power to detect a nominally statistically significant association was >80% for only two variants (Supplementary Material, Table S2).

Figure 1 shows the associations of all 73 variants with breast cancer risk in African Americans, which includes these 54 new variants as well as the 19 variants reported in our previous study (1–15). Of the 73 variants, 47 (64%) were positively associated with breast cancer risk in African American women. For 11 variants, the 95% confidence intervals (CI) reported from the previous studies excluded the ORs estimated in African Americans and for only eight variants were the 95% CI non-overlapping.

In analyses by ER status, 34 of the 54 (63%) variants were positively associated with ER+ breast cancer, with six significant at $P < 0.05$. Thirty-one variants (57%) were positively associated with ER− breast cancer (Seven at $P < 0.05$) (Supplementary Material, Table S3). In the case-only analysis, five variants showed a statistically significantly different association with breast cancer risk by ER status: rs10759243/*9q31* and rs13329835/*16q23*, which were more strongly associated with ER+ disease and rs10069690/*5p15*, rs1432679/*5q33* and rs2284378/*20q11* which were more strongly associated with ER− disease. These associations in ER subgroups were consistent with previous reports of these loci (Supplementary Material, Table S3) (10,12,13,17). Of the seven variants reported to be specifically associated with ER− breast cancer (rs6678914/*1q32*, rs4245739/*1q32*,
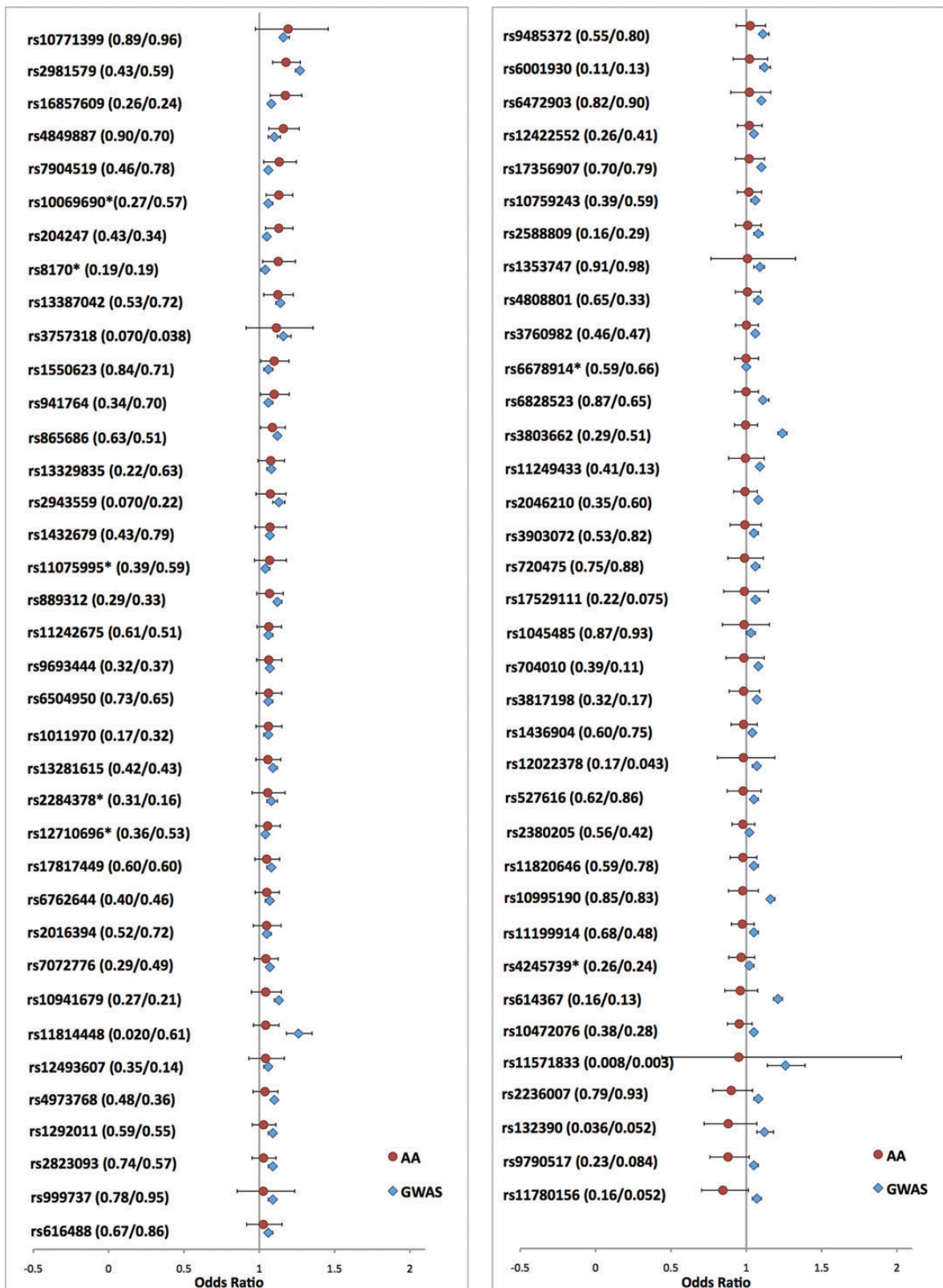
Figure 1. Effect estimates of overall breast cancer risk for all 73 known risk variants in GWAS-discovery and African-ancestry populations. Red circles represent the per-allele ORs estimated in women of African ancestry (AA). Blue diamonds represent the per-allele ORs reported in the initial GWAS. The horizontal lines represent 95% confidence limits. Asterisks represent SNPs that were reported for ER − disease. For each tested allele, frequencies in GWAS-discovery and African-ancestry populations are provided in parentheses. SNPs are sorted based on their ORs in AA. Detailed information for each SNP is provided in Supplementary Material, Table S1.

rs12710696/*2p24*, rs10069690/*5p15*, rs11075995/*16q12*, rs8170/*19p13*, rs2284378/*20q11* (7,10,13,14), we have previously reported positive associations for all seven variants, two of which were significant at $P < 0.05$ (rs10069690 on *5p15* and rs2284378 on *20q11*) (10,13,14,16). However, statistical power was >80% to detect the associations for only ER− variants rs10069690 on *5p15*, which this study contributed to identifying, and rs8170 on *19p13* (988 ER− cases and all controls; Supplementary Material, Table S2).

In addition to statistical power, the failure to replicate associations with the index variants implies that the particular risk variant found in GWAS in European or Asian populations might not be adequately correlated with the biologically relevant allele in African Americans. In an attempt to identify a better genetic marker of the biologically relevant allele in African Americans, we tested all genotyped and imputed SNPs (in the 1000 Genomes Project) that were correlated ($r^2 > 0.4$) with the index variant in European ancestry populations (see Materials and Methods for details of fine-mapping).

In three of the 54 regions (*1q32, 3p24, 10q25*), we found associations with variants that might better define risk in African Americans. The index variant on *1q32* (rs4245739) has been reported for ER− (OR = 1.14, $P = 3.9 \times 10^{-13}$) but not ER+ breast cancer (OR = 0.99, $P = 0.7$) (14). However, in this region, we observed suggestive evidence of a signal for ER+ breast cancer with a large cluster of alleles that are correlated with the index variant in European ancestry populations, the most significant of which was rs4951385 (OR = 1.17, $P = 1.2 \times 10^{-3}$). Variant rs4951385 is located 64.7 kb from the index SNP (rs4245739) in the 32nd intron of the *PIK3C2B* gene and is highly correlated with rs4245739 in European, but not African ancestry populations (EUR: $r^2 = 0.90$; AFR: $r^2 = 0.11$) (Supplementary Material, Table S4).

At *3q24*, the index SNP (rs12493607) was positively associated with overall breast cancer as well as ER+ and ER− disease in African Americans (Supplementary Material, Tables S1 and S3). Through fine-mapping, variant rs13086588 was detected to be more strongly associated with ER+ breast cancer (ER+: OR = 1.20, $P = 3.0 \times 10^{-4}$; ER−: OR = 1.04, $P = 0.54$; $p_{het} = 0.04$), which is consistent with this locus being more strongly associated with ER+ than ER− disease ($p_{het} = 0.02$) (12). Variant rs13086588 is located in the second intron of *TGFBR2*, and is strongly correlated with rs12493607 in Europeans but not in African ancestry populations (EUR: $r^2 = 0.76$; AFR: $r^2 = 0.08$; Supplementary Material, Table S4).

At *10q25*, the index variant (rs7904519) was significantly associated with the risk of overall breast cancer in African Americans (OR = 1.13, $P = 0.01$). Fine-mapping of this region revealed variant rs7919152 that is correlated with the index variant (rs7904519: EUR: $r^2 = 0.83$; AFR: $r^2 = 0.51$) and may be better capturing risk of overall breast cancer in this region (OR = 1.16, $P = 9.9 \times 10^{-4}$; Supplementary Material, Table S4).

In search of novel secondary signals at each risk locus, we tested associations of all SNPs within 250 kb surrounding each index variant with risk of overall breast cancer as well as ER+ and ER− disease (see Materials and Methods for details). In three of the 54 regions (*6q25, 14q13, 22q12*), we detected evidence of an independent signal at $P < 5 \times 10^{-6}$ (all SNPs uncorrelated $r^2 < 0.05$ with the index variants at $P \le 10^{-5}$ are shown in Supplementary Material, Table S5). At *6q25,* an intergenic variant, rs9390664, was found to be significantly associated with overall breast cancer risk (OR = 1.39, $P = 4.4 \times 10^{-7}$). This variant is located 33.9 kb from the index SNP (rs9485372) and is not correlated with rs9485372 in either European or African populations ($r^2 < 0.05$). At *14q13*, the index variant rs2236007 was reported to be more significantly associated with ER+ than ER− breast cancer in Europeans (OR = 1.10 versus 1.04, $p_{het} = 0.02$) (12). We observed an association with rs17104923, located 6.4 kb from rs2236007 in the 4th intron of the *PAX9* gene, which was also more strongly associated with ER+ breast cancer (Overall: OR = 1.28, $P = 1.1 \times 10^{-3}$; ER+: OR = 1.62, $P = 1.6 \times 10^{-6}$; ER−: OR = 1.13, $P = 0.27$; $p_{het} = 0.019$; Supplementary Material, Table S5; Fig. 2). Variant rs17104923 is not correlated with rs2236007 in either European or African populations ($r^2 < 0.01$). At *22q12,* the association with the index variant (rs132390) in Europeans was found to be stronger for ER+ disease (ER+: OR = 1.13, $P = 4.2 \times 10^{-5}$; ER−: OR = 1.08, $P = 0.11$). A secondary signal, rs67157227, located 100.4 kb from the index SNP (rs132390) in the 4th intron of the *KREMEN1* gene, was also identified to be significantly associated with ER+ breast cancer (Overall: OR = 1.24, $P = 5.5 \times 10^{-5}$; ER+: OR = 1.36, $P = 4.4 \times 10^{-6}$; ER−: OR = 1.12, $P = 0.13$; $p_{het} = 0.031$), suggesting that variation at this locus may also be more associated with ER+ disease in African Americans.

In attempt to confirm these findings, we tested the three significant secondary signals in an independent sample of 1657 breast cancer cases and 2028 controls of African ancestry (see Materials and Methods for details of this sample). Only one variant (rs17104923/*14q13*) was significantly associated with breast cancer risk (OR = 1.20, $P = 0.036$; Supplementary Material, Table S6). The association was stronger with ER+ than ER− disease ($n = 403$ ER+ cases: OR = 1.32, $P = 0.057$; $n = 374$ ER− cases: OR = 1.18, $P = 0.24$), which is consistent with the initial results in our study.

We also estimated the cumulative effects of all 73 breast cancer risk variants using risk score modeling. The risk per allele was 1.04 (95% CI: 1.03–1.05, $P = 1.6 \times 10^{-11}$) for overall breast cancer, 1.04 (95% CI: 1.02–1.05, $P = 2.0 \times 10^{-7}$) for ER+ disease and 1.03 (95% CI: 1.02–1.05, $P = 1.1 \times 10^{-4}$) for ER− disease. Compared with those in the lowest quintile, individuals in the top quintile of the risk allele distribution were at 1.78 ($P = 1.1 \times 10^{-10}$), 1.67 ($P = 1.8 \times 10^{-6}$) and 1.70 ($P = 4.4 \times 10^{-5}$) -fold greater risk of overall breast cancer, ER+ and ER− disease, respectively (Table 1).

DISCUSSION

In this study of breast cancer in African American women, we tested 54 recently identified variants with the vast majority identified through large-scale testing in the COGS consortium in European-ancestry populations (1–15). We observed 38 variants that were associated with overall or ER− breast cancer in African Americans in a direction consistent with that reported previously. The 54 variants tested in this study were previously reported to have an average odds ratio of 1.09, with only 13 (24%) having ORs >1.10. This is in contrast to the initial set of 19 breast cancer risk variants discovered through GWAS,
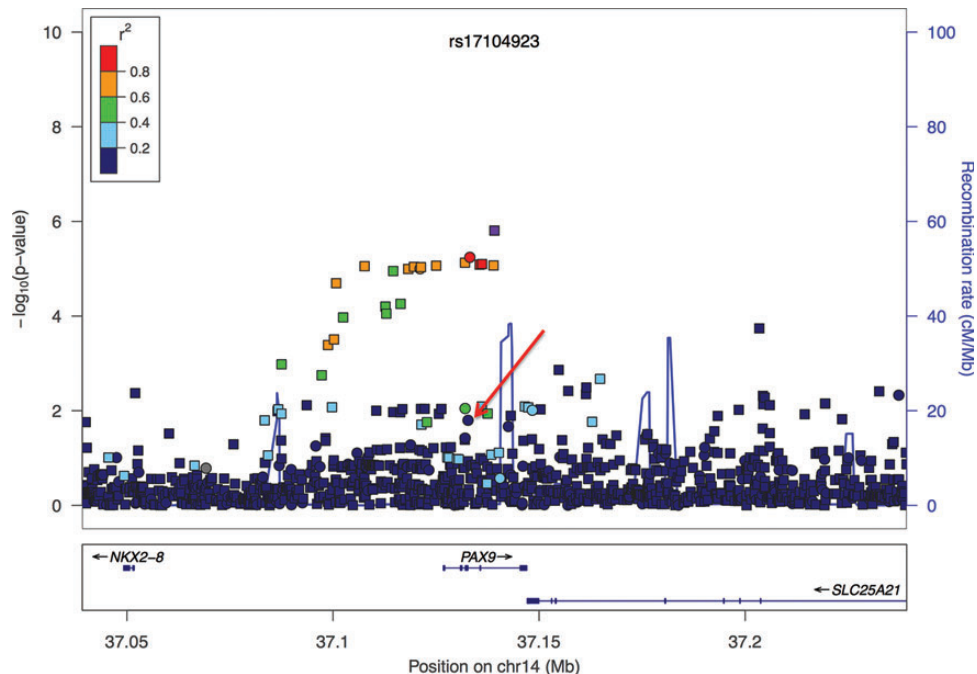
Figure 2. Regional plot of the secondary signal (rs17104923) on *14q13*. The chromosomal position (based on GRCh37) of SNPs on *14q13* against $-\log_{10}$ *P*-values for ER+ disease is shown. Genotyped SNPs are represented by circles. Imputed SNPs are represented by squares. The secondary signal rs17104923 is plotted by a purple square. The red arrow denotes the GWAS index variant rs2236007. SNPs surrounding the top SNPs are colored to indicate the LD structure using pairwise r^2 in reference to rs17104923 from the May 2012 AFR panel of 1000 Genomes. The plots were generated using LocusZoom ([18]).

which had larger effect sizes, with nine (47%) variants having ORs >1.10, and an average OR of 1.12. Thus, in general, these new variants had smaller effect sizes, implying a weaker biological influence on breast cancer ([1–15]). In our study in African Americans, statistical power was ≥80% to detect a nominally statistically significant association for eight (42%) of the 19 variants examined initially ([16]), while for only two (4%) of these additional 54 variants did we have ≥80% power to detect the odds ratios reported in the initial studies (Supplementary Material, Table S2).

Despite small effect sizes leading to limited power, failure of replication may also result from different LD structure between populations and more distinct markers of the index signal to represent the same biological signal in diverse populations. Using our stringent criteria, in only three (6%) of the 54 recently identified breast cancer susceptibility regions did we identify variants that might better define associations with overall breast cancer, ER+ or ER− disease in African Americans. The failure to enhance signals in these regions might also be attributed to limited statistical power. In utilizing the locus-specific α levels, statistical power was ≥80% to detect associations for only five of the 73 regions (Supplementary Material, Table S7). As described earlier, in the initial GWAS, these newly identified breast cancer risk variants had smaller odds ratios than the initial 19 GWAS identified risk variants. Given the observed diminishing effect sizes noted for the more recently identified GWAS variants, even larger sample size is needed to detect associations in non-European ancestry populations.

In three of the 54 regions, we observed significant associations ($P < 5 \times 10^{-6}$) with variants that were uncorrelated with the index SNPs, representing putative novel independent risk

signals. At *14q13*, the association with rs17104923 was stronger for ER+ breast cancer, with supportive evidence provided in the replication sample ($P = 0.04$ for overall breast cancer and $P = 0.06$ for ER+ disease). Variant rs17104923 is located in the 4th intron of the gene *PAX9* (paired box 9). In addition to it being a risk locus for breast cancer ([12]), the chromosome region containing *PAX9* on *14q13* has also been shown to be both amplified and deleted in lung cancer ([19]). Both the index variant at this locus and this putative novel signal appear to be more strongly associated with ER+ breast cancer, which provides further support for genetic determinants of breast cancer subtypes. At *6q25*, the intergenic variant rs9390664 is in close proximity to a number of genes, including *TAB2* (TGF-beta activated kinase 1/MAP3K7 binding protein 2), *SUMO4* (small ubiquitin-like modifier 4) and *UST* (uronyl-2-sulfotransferase). At *22q12*, the variant rs67157227 is located in the 4th intron of *KREMEN1* which is a component of a membrane complex that modulates canonical WNT signaling through lipoprotein receptor-related protein 6 (LRP6) ([20]). According to data harvested from the ENCODE project ([21]), only one of the suggestive secondary signals (rs17104923/*14q13*) was found to be located in proximity to a weak DNaseI signal, which marks for a nucleosome depleted region, in a breast cancer cell line (MCF7, ER+ cell line). Further support for the associations with these variants is needed as neither was found to be statistically significantly associated with risk in the replication sample.

Among the 73 known risk loci, 49 (67%) showed an association with overall breast cancer or ER− disease in the same direction as previously reported, with 12 (18%) showing directionally consistent and nominally statistically significant associations in African Americans. The directional consistency

Table 1. Associations with risk scores comprising 73 breast cancer risk variants in African Americans by ER status

	All cases versus controls	ER+ cases versus controls	ER− cases versus controls	p_{het}^a
Average number of risk alleles in controls (range)	71.3 (55.0–86.4)			
Per allele OR (95% CI)[b]	1.04 (1.03–1.05)	1.04 (1.02–1.05)	1.03 (1.02–1.05)	
P_{trend}^c	1.6×10^{-11}	2.0×10^{-7}	1.1×10^{-4}	0.36
n cases/*n* controls	3016/2745	1520/2745	988/2745	
Risk quintiles[d]				
Q1				
n cases/*n* controls	515/637	272/637	157/637	
OR (95% CI)	1.00 (reference)	1.00 (reference)	1.00 (reference)	
P-value	–	–	–	
Q2				
n cases/*n* controls	578/572	300/572	192/572	
OR (95% CI)	1.24 (1.05–1.47)	1.22 (0.99–1.50)	1.31 (1.02–1.69)	
P-value	0.014	0.063	0.036	
Q3				
n cases/*n* controls	628/527	306/527	218/527	
OR (95% CI)	1.46 (1.23–1.73)	1.36 (1.10–1.68)	1.58 (1.23–2.03)	
P-value	1.9×10^{-5}	4.1×10^{-3}	3.6×10^{-4}	
Q4				
n cases/*n* controls	615/538	302/538	201/538	
OR (95% CI)	1.44 (1.21–1.71)	1.34 (1.09–1.66)	1.49 (1.16–1.92)	
P-value	3.4×10^{-5}	6.3×10^{-3}	2.0×10^{-3}	
Q5				
n cases/*n* controls	680/471	340/471	220/471	
OR (95% CI)	1.78 (1.49–2.12)	1.67 (1.36–2.07)	1.70 (1.32–2.19)	
P-value	1.1×10^{-10}	1.8×10^{-6}	4.4×10^{-5}	

[a]*P*-value for case-only analysis (ER+ versus ER−).
[b]Odds ratio per allele based on analysis adjusted for age, study and the first 10 eigenvectors.
[c]*P*-value based on 1-degree-of-freedom Wald χ^2 trend test.
[d]Cut points based on the distribution of risk scores in controls.

noted implies a shared functional common variant at most loci. Long *et al.* (22) evaluated 67 breast cancer susceptibility loci in a study with 1231 African American cases and 2069 controls. Seven SNPs showed directionally consistent and significant associations with overall breast cancer, four of which were replicated in our African American sample. Through fine-mapping conducted in this study and in our previous study (16), we noted suggestive evidence in several regions with variants that may better characterize the association with breast cancer risk in African American women. As is currently ongoing for most phenotypes, combining GWAS data from large numbers of studies via meta-analyses followed by large-scale replication testing will continue to reveal variants with diminishing effect sizes. Additional studies in African ancestry populations and combining genetic data through large collaborative efforts will be needed in order to more fully understand the contribution to risk of the established breast cancer loci, especially for ER− disease, which disproportionally affects populations of African ancestry.

MATERIALS AND METHODS

Samples

The data in this study are from a GWAS of breast cancer in African American women which includes nine epidemiological studies of breast cancer, comprising a total of 3153 cases and 2831 controls (cases/controls): the Multiethnic Cohort study (MEC) (23), 734/1003; The Los Angeles component of The Women's Contraceptive and Reproductive Experiences (CARE) Study (24), 380/224; The Women's Circle of Health Study (WCHS) (25), 272/240; The San Francisco Bay Area Breast Cancer Study (SFBCS) (26), 172/231; The Northern California site of the Breast Cancer Family Registry (NC-BCFR) (27), 440/53; The Carolina Breast Cancer Study (CBCS) (28), 656/608; The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial Cohort (PLCO) (29), 64/133; The Nashville Breast Health Study (NBHS) (30), 310/186; and The Wake Forest University Breast Cancer Study (WFBC) (31), 125/153. Detailed information about the design of each study has been published previously (16,32). Sample size and selected characteristics for these studies are summarized in Supplementary Material, Table S8.

The replication sample included six studies of African ancestry and a total of 1657 cases and 2028 controls (cases/controls): the Nigerian Breast Cancer Study (NBCS) (33,34), 711/623; The Barbados National Cancer Study (BNCS) (35), 92/229; The Racial Variability in Genotypic Determinants of Breast Cancer Risk Study (RVGBC), 145/257; The Baltimore Breast Cancer Study (BBCS), 95/102; The Chicago Cancer Prone Study (CCPS), 394/387 and The Southern Community Cohort (SCCS) (36), 220/430. Detailed information about the design of each study is described in Zheng *et al.* (37).

Genotyping and quality control

Genotyping for the African American sample in this study was conducted using the Illumina Human1M-Duo BeadChip as described in Chen *et al.* (16). The average sample call rate was

99.8%. To confirm imputation (discussed subsequently), genotyping of the three significant secondary signals (rs9390664/*6q25,* rs17104923/*14q13* and rs67157227/*22q12*) was performed in 377 individuals from the MEC African American sample. Two variants (rs17104923/*14q13* and rs67157227/*22q12*) could be genotyped and had consistent genotypes with that from imputation, with an r^2 of 0.99 and 0.86, respectively.

Statistical analysis

In order to generate a data set suitable for fine-mapping, we performed genome-wide imputation using IMPUTE2 (38) to a cosmopolitan panel of all 1000 Genomes Project subjects (March 2012 release). Imputed SNPs with $r^2 > 0.8$ (defined as the observed variance divided by the expected variance) were used in the fine-mapping analyses. For the 54 index variants analyzed in this study, 16 were imputed and imputation quality scores were >0.8 for 14. Variants rs11571833/*13q13* and rs132390/*22q12* were imputed with scores 0.69 and 0.72, respectively, and both SNPs had small minor allele frequencies in the AFR population of the 1000 Genomes Project (rs11571833/*13q13*: MAF = 0.0060; rs132390/*22q12*: MAF = 0.059).

For each typed and imputed SNP, odds ratios (OR) and 95% CIs were estimated using unconditional logistic regression adjusting for age (at diagnosis for cases and age at the reference date for controls), study, and the first 10 eigenvectors from a principal components analysis (39). For each SNP, we tested for allele dosage effects using a 1-degree-of-freedom Wald χ^2 trend test.

To characterize alleles that might better represent the biologically functional variant, we searched and tested LD proxies among the genotyped and imputed SNPs that are correlated ($r^2 \geq 0.4$) with the index SNP (within 250 kb or larger if the index signal was contained within an LD block) in the GWAS discovery population (European ancestry). Two regions, *5p15* and *20q11* were excluded from locus fine-mapping as our African American sample was involved in the discovery of these loci (10,13). Locus-specific alpha levels were utilized, which accounts for multiple testing of correlated markers when searching for a stronger marker of the index signal in an African population (Supplementary Material, Table S7). It is calculated by 0.05/the number of tag SNPs in the African population (1000 Genomes, AFR) that capture ($r^2 \geq 0.8$) all SNPs correlated with the index signal in the European population (1000 Genomes, EUR). To reduce false-positive signals for all regions, we required the *P*-value of all the better markers to be less than 0.01. In an attempt to eliminate minor fluctuations in *P*-values for correlated SNPs, we also required the *P*-value to decrease by more than one order of magnitude compared with the association with the index signal. 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 European ancestry population.

We also looked for novel independent associations, focusing on the genotyped and imputed SNPs that were uncorrelated with the index signal in European ancestry populations ($r^2 < 0.4$). Here, we applied a significance criterion of $\alpha = 5 \times 10^{-6}$ for defining novel associations as significant in each region, which is an extension of the empirically determined Bonferroni correction used in Chen *et al.* (16) and is an approximation of the total number of tests to capture (at $r^2 \geq 0.8$) all common risk

alleles across the 73 risk regions in the African American population. These procedures were applied to the analysis of overall breast cancer as well as in hypothesis-generating analyses stratified by ER status.

To evaluate the combined effects of these risk markers, we modeled the cumulative genetic risk of breast cancer using the 73 reported risk variants in African Americans. We summed the number of risk alleles for each individual and estimated the odds ratio 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 applied this risk score to overall breast cancer, as well as ER+ and ER− disease. Missing values for ungenotyped markers were replaced with mean allele counts in the whole population.

Replication testing

The replication sample was genotyped with the Illumina 2.5 M array as described in Zheng *et al*. (37). For the three variants tested in this paper (rs9390664/*6q25,* rs17104923/*14q13* and rs67157227/*22q12*), only rs17104923 at *14q13* was genotyped (call rate of 99.8%). Variants rs9390664/*6q25* and rs67157227/*22q12* were imputed with scores of 0.95 and 0.97, respectively. Details of the imputation strategy used in the replication sample were described in Zheng *et al*. (37).

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

Conflict of Interest statement. None declared.

FUNDING

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. Cox, A., Dunning, A.M., Garcia-Closas, M., Balasubramanian, S., Reed, M.W., Pooley, K.A., Scollen, S., Baynes, C., Ponder, B.A., Chanock, S. *et al.* (2007) A common coding variant in CASP8 is associated with breast cancer risk. *Nat. Genet.*, **39**, 352–358.

3. 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.

4. 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.

5. 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.

6. 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.

7. 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.

8. Turnbull, C., Ahmed, S., 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. 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.

10. 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.

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

12. Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K. *et al.* (2013) Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.*, **45**, 353–361. 361e351–352.

13. Siddiq, A., Couch, F.J., Chen, G.K., Lindstrom, S., Eccles, D., Millikan, R.C., Michailidou, K., Stram, D.O., Beckmann, L., Rhie, S.K. *et al.* (2012) A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Hum. Mol. Genet.*, **21**, 5373–5384.

14. Garcia-Closas, M., Couch, F.J., Lindstrom, S., Michailidou, K., Schmidt, M.K., Brook, M.N., Orr, N., Rhie, S.K., Riboli, E., Feigelson, H.S. *et al.* (2013) Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat. Genet.*, **45**, 392–398. 398e391–392.

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

16. Chen, F., Chen, G.K., Millikan, R.C., John, E.M., Ambrosone, C.B., Bernstein, L., Zheng, W., Hu, J.J., Ziegler, R.G., Deming, S.L. *et al.* (2011) Fine-mapping of breast cancer susceptibility loci characterizes genetic risk in African Americans. *Hum. Mol. Genet.*, **20**, 4491–4503.

17. Palmer, J.R., Ruiz-Narvaez, E.A., Rotimi, C.N., Cupples, L.A., Cozier, Y.C., Adams-Campbell, L.L. and Rosenberg, L. (2013) Genetic susceptibility loci for subtypes of breast cancer in an African American population. *Cancer Epidemiol. Biomarkers Prev.*, **22**, 127–134.

18. 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.

19. Harris, T., Pan, Q., Sironi, J., Lutz, D., Tian, J., Sapkar, J., Perez-Soler, R., Keller, S. and Locker, J. (2011) Both gene amplification and allelic loss occur at 14q13.3 in lung cancer. *Clin. Cancer Res.*, **17**, 690–699.

20. Mao, B., Wu, W., Davidson, G., Marhold, J., Li, M., Mechler, B.M., Delius, H., Hoppe, D., Stannek, P., Walter, C. *et al.* (2002) Kremen proteins are Dickkopf receptors that regulate Wnt/beta-catenin signalling. *Nature*, **417**, 664–667.

21. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B. *et al.* (2012) The accessible chromatin landscape of the human genome. *Nature*, **489**, 75–82.

22. Long, J., Zhang, B., Signorello, L.B., Cai, Q., Deming-Halverson, S., Shrubsole, M.J., Sanderson, M., Dennis, J., Michailiou, K., Easton, D.F. *et al.* (2013) Evaluating genome-wide association study-identified breast cancer risk variants in African-American women. *PLoS One*, **8**, e58350.

23. 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.

24. 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.

25. 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.

26. 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.

27. 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.

28. 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.

29. 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.

30. 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.

31. 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.

32. Chen, F., Chen, G.K., Stram, D.O., Millikan, R.C., Ambrosone, C.B., John, E.M., Bernstein, L., Zheng, W., Palmer, J.R., Hu, J.J. *et al.* (2012) A genome-wide association study of breast cancer in women of African ancestry. *Hum. Genet.*, **132**, 39–48.

33. Huo, D., Adebamowo, C.A., Ogundiran, T.O., Akang, E.E., Campbell, O., Adenipekun, A., Cummings, S., Fackenthal, J., Ademuyiwa, F., Ahsan, H. *et al.* (2008) Parity and breastfeeding are protective against breast cancer in Nigerian women. *Br. J Cancer*, **98**, 992–996.

34. Huo, D., Kim, H.J., Adebamowo, C.A., Ogundiran, T.O., Akang, E.E., Campbell, O., Adenipekun, A., Niu, Q., Sveen, L., Fackenthal, J.D. *et al.* (2008) Genetic polymorphisms in uridine diphospho-glucuronosyltransferase 1A1 and breast cancer risk in Africans. *Breast Cancer Res. Treat.*, **110**, 367–376.

35. International HapMap, C., Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

36. Signorello, L.B., Hargreaves, M.K., Steinwandel, M.D., Zheng, W., Cai, Q., Schlundt, D.G., Buchowski, M.S., Arnold, C.W., McLaughlin, J.K. and Blot, W.J. (2005) Southern community cohort study: establishing a cohort to investigate health disparities. *J. Natl Med. Assoc.*, **97**, 972–979.

37. Zheng, Y., Ogundiran, T.O., Falusi, A.G., Nathanson, K.L., John, E.M., Hennis, A.J., Ambs, S., Domchek, S.M., Rebbeck, T.R., Simon, M.S. *et al.* (2013) Fine mapping of breast cancer genome-wide association studies loci in women of African ancestry identifies novel susceptibility markers. *Carcinogenesis*, **34**, 1520–1528.

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

39. 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.