

UC Davis

UC Davis Previously Published Works

Title

Cross-ancestry GWAS meta-analysis identifies six breast cancer loci in African and European ancestry women

Permalink

<https://escholarship.org/uc/item/5t975267>

Journal

Nature Communications, 12(1)

ISSN

2041-1723

Authors

Adedokun, Babatunde

Du, Zhaohui

Gao, Guimin

et al.

Publication Date

2021

DOI



10.1038/s41467-021-24327-x

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Cross-ancestry GWAS meta-analysis identifies six breast cancer loci in African and European ancestry women

Babatunde Adedokun¹, Zhaohui Du², Guimin Gao ³, Thomas U. Ahearn ⁴, Kathryn L. Lunetta ⁵, Gary Zirpoli⁶, Jonine Figueroa ⁷, Esther M. John⁸, Leslie Bernstein ⁹, Wei Zheng ¹⁰, Jennifer J. Hu¹¹, Regina G. Ziegler ⁴, Sarah Nyante ¹², Elisa V. Bandera¹³, Sue A. Ingles², Michael F. Press ¹⁴, Sandra L. Deming-Halverson¹⁰, Jorge L. Rodriguez-Gil ¹⁵, Song Yao ¹⁶, Temidayo O. Ogundiran¹⁷, Oladosu Ojengbede¹⁸, William Blot¹⁰, Melissa A. Troester¹⁹, Katherine L. Nathanson ²⁰, Anselm Hennis^{21,22}, Barbara Nemesure²², Stefan Ambbs ²³, Peter N. Fiorica³, Lara E. Sucheston-Campbell²⁴, Jeannette T. Bensen¹⁹, Lawrence H. Kushi ²⁵, Gabriela Torres-Mejia²⁶, Donglei Hu ²⁷, Laura Fejerman ²⁷, Manjeet K. Bolla²⁸, Joe Dennis ²⁸, Alison M. Dunning ²⁹, Douglas F. Easton ^{28,29}, Kyriaki Michailidou³⁰, Paul D. P. Pharoah ^{28,29}, Qin Wang²⁸, Dale P. Sandler ³¹, Jack A. Taylor ³¹, Katie M. O'Brien³¹, Cari M. Kitahara³², Adeyinka G. Falusi³³, Chinedum Babalola ³⁴, Joel Yarney³⁵, Baffour Awuah³⁶, Beatrice Addai-Wiafe³⁷, The GBHS Study Team, Stephen J. Chanock ⁴, Andrew F. Olshan¹⁹, Christine B. Ambrosone ¹⁶, David V. Conti², Elad Ziv²⁷, Olufunmilayo I. Olopade ¹, Montserrat Garcia-Closas ⁴, Julie R. Palmer⁶, Christopher A. Haiman ²✉ & Dezheng Huo ^{1,3}✉

Our study describes breast cancer risk loci using a cross-ancestry GWAS approach. We first identify variants that are associated with breast cancer at $P < 0.05$ from African ancestry GWAS meta-analysis (9241 cases and 10193 controls), then meta-analyze with European ancestry GWAS data (122977 cases and 105974 controls) from the Breast Cancer Association Consortium. The approach identifies four loci for overall breast cancer risk [1p13.3, 5q31.1, 15q24 (two independent signals), and 15q26.3] and two loci for estrogen receptor-negative disease (1q41 and 7q11.23) at genome-wide significance. Four of the index single nucleotide polymorphisms (SNPs) lie within introns of genes (*KCNK2*, *C5orf56*, *SCAMP2*, and *SIN3A*) and the other index SNPs are located close to *GSTM4*, *AMPD2*, *CASTOR2*, and *RP11-168G16.2*. Here we present risk loci with consistent direction of associations in African and European descendants. The study suggests that replication across multiple ancestry populations can help improve the understanding of breast cancer genetics and identify causal variants.

Breast cancer is the most common cancer in women worldwide and accounted for 2.1 million new cases and 627,000 deaths in 2018¹. Studies have shown a significant contribution of genetic factors to breast cancer risk^{2,3}, yet the landscape of this contribution has not been fully elucidated. Mutations in high- and moderate-penetrance genes confer relatively high risks of breast cancer but are rare in the population and account for <5–10% of cases⁴. Genome-wide association studies (GWAS) have been successful in identifying common low-penetrance genetic variation and approximately 200 risk loci have now been identified^{5–7}. The risk loci so far identified have provided clues to elucidating breast cancer tumorigenesis through previously unknown mechanisms. Additionally, when combined into risk scores, these polymorphisms can be used for breast cancer risk prediction⁸.

Despite the usefulness of GWAS, the majority of the GWAS studies have been performed among European ancestry populations^{9–13}, it is unclear whether the same genetic risk factors are also important in other populations, which may limit the applicability of the findings to other groups¹⁴. The earliest GWAS conducted in African ancestry populations identified genetic variants at 5p15.33 (TERT-CLPTMIL) associated with estrogen receptor (ER) negative breast cancer¹⁵. A larger analysis of African ancestry individuals which included several consortia identified a SNP at 3q26.21 also associated with ER-negative breast cancer¹⁶. Some common susceptibility loci are shared across populations, and the shared disease-associated variants are more likely to be causal^{16,9,14}.

Here we present, using a cross-ancestry GWAS approach in 248,000 women, genetic risk variants at 1p13.3, 5q31.1, 15q24, and 15q26.3 for overall breast cancer, and at 1q41 and 7q11.23 for ER-negative disease. The consistency of the directions of the risk for these loci in African and European samples increases the likelihood of their being causal variants.

Results

We discovered six loci containing seven SNPs significantly associated with breast cancer at $P < 5 \times 10^{-8}$ on cross-ancestry meta-analysis, with odds ratios (OR) ranging from 0.95 to 1.05 (Tables 1, 2; Supplementary Figs. 1, 2). Five SNPs were associated with overall breast cancer risk (rs17024628 at 1p13.3, rs2522057 at 5q31.1, rs1869959 at 15q24.1, rs60381548 at 15q24.2, rs181337095 at 15q26.3) and two were associated with ER-negative breast cancer (rs67931591 at 1q41 and rs1637365 at 7q11.2). The two SNPs at the 15q24 region were about 582 kb apart and independently associated with breast cancer risk. Four SNPs were within genes (rs67931591 in *KCNK2*, rs2522057 in *C5orf56*, rs1869959 in *SCAMP2*, and rs60381548 in *SIN3A*) and the others were in intergenic regions. The direction of the associations was consistent for the pooled African and European estimates. The estimates for overall and ER-negative breast cancer were generally consistent across the five contributing studies of African ancestry participants (Supplementary Table 2) and the BCAC European datasets (Supplementary Table 3).

Conditional analysis revealed three additional independent signals significant at $p < 10^{-4}$ at the 1p13.3 locus (rs116363925, rs114351980, and 1:109969874:C:T), two independent signals at 15q24 (rs113939578, rs12917507), and one each at 5q31.1 (5:132149322:G:GGCCGCC) and 15q26.3 (rs117793215) for overall breast cancer risk. Another independent SNP at 1q41 that was associated with ER-negative breast cancer was rs5780828 (Table 3).

Concerning pleiotropy, none of the SNPs identified above have been reported in previous GWAS associations at genome-wide

Table 1 Novel breast cancer risk loci identified by cross-ancestry meta-analysis of African and European populations.

SNP	Chr	Position	Test	Other	Locus	Within gene	African-specific meta-analysis			European-specific meta-analysis			Combined African and European meta-analysis ^a				
							TAF	OR (95% CI)	P value	TAF	OR (95% CI)	P value	TAF	OR (95% CI)	P value		
Overall																	
rs17024629	1	110,179,756	T	C	1p13.3	No	0.13	0.88 (0.83–0.95)	5.2E–04	0.16	0.96 (0.94–0.98)	1.2E–06	0.16	0.95 (0.94–0.97)	3.0E–08		
rs67931591	1	215,330,292	G	GCTGAGG- CAGGAGA	1q41	KCNK2	0.28	0.95 (0.90–1.00)	0.034	0.68	0.98 (0.96–0.99)	3.9E–04	0.66	0.97 (0.96–0.99)	7.4E–05		
rs2522057	5	131,801,947	C	G	5q31.1	C5orf56	0.86	0.92 (0.86–0.98)	0.0084	0.59	0.97 (0.96–0.98)	9.3E–08	0.60	0.97 (0.95–0.98)	1.1E–08		
rs1637365	7	74,359,358	T	C	7q11.23	No	0.62	1.06 (1.01–1.12)	0.024	0.28	1.04 (1.02–1.05)	3.3E–06	0.31	1.04 (1.02–1.05)	3.6E–07		
rs1869959	15	75,147,332	A	C	15q24.1	SCAMP2	0.40	0.95 (0.91–1.00)	0.043	0.30	0.97 (0.95–0.98)	3.6E–07	0.30	0.96 (0.95–0.98)	4.6E–08		
rs60381548	15	75,728,474	CA	C	15q24.2	SIN3A	0.50	0.93 (0.89–0.97)	0.0016	0.25	0.96 (0.95–0.98)	4.0E–07	0.27	0.96 (0.95–0.97)	6.6E–09		
rs181337095	15	100,907,094	A	G	15q26.3	No	0.69	1.06 (1.01–1.12)	0.017	0.87	1.05 (1.03–1.07)	3.4E–07	0.84	1.05 (1.04–1.07)	1.8E–08		
ER negative																	
rs17024629	1	110,179,756	T	C	1p13.3	No	0.13	0.83 (0.74–0.92)	0.00064	0.16	0.96 (0.93–0.99)	0.020	0.16	0.95 (0.93–0.98)	0.0014		
rs67931591	1	215,330,292	G	GCTGAGG- CAGGAGA	1q41	KCNK2	0.29	0.92 (0.85–0.99)	0.024	0.68	0.94 (0.92–0.96)	4.6E–07	0.65	0.94 (0.92–0.96)	4.3E–08		
rs2522057	5	131,801,947	C	G	5q31.1	C5orf56	0.86	0.93 (0.85–1.03)	0.18	0.59	0.99 (0.96–1.01)	0.22	0.60	0.98 (0.96–1.01)	0.14		
rs1637365	7	74,359,358	T	C	7q11.23	No	0.61	1.15 (1.06–1.25)	0.00069	0.28	1.07 (1.04–1.10)	9.0E–07	0.32	1.08 (1.05–1.11)	1.0E–08		
rs1869959	15	75,147,332	A	C	15q24.1	SCAMP2	0.41	0.96 (0.89–1.03)	0.22	0.30	0.97 (0.94–0.98)	0.0050	0.31	0.96 (0.94–0.99)	0.0022		
rs60381548	15	75,728,474	CA	C	15q24.2	SIN3A	0.51	0.99 (0.92–1.06)	0.78	0.25	0.95 (0.92–0.97)	8.6E–05	0.28	0.95 (0.93–0.98)	1.7E–04		
rs181337095	15	100,907,094	A	G	15q26.3	No	0.69	1.07 (0.99–1.16)	0.067	0.87	1.06 (1.02–1.10)	0.0016	0.83	1.06 (1.03–1.10)	2.9E–04		

TAF, Test allele frequency, SNP single nucleotide polymorphism, OR odds ratio, CI confidence intervals, ^atest for heterogeneity across studies was statistically significant only for rs1637365 and ER-negative breast cancer (P-for-heterogeneity = 0.025).

Table 2 Association analysis of novel SNPs in cross-ancestry combined meta-analysis by estrogen receptor status.

SNP	Chr	Position	Test	Other	Locus	Nearest genes	ER positive			ER negative			P for heterogeneity ^a
							TAF	OR (95% CI)	P value	TAF	OR (95% CI)	P value	
rs17024629	1	110,179,756	T	C	1p13.3	GSTM4, AMPD2, GSTM2, GSTM1, GNAT2, MIR197, GNA3, KCNK2, KCTD3, CENPF	0.16	0.95 (0.93-0.97)	1.1E-06	0.16	0.95 (0.93-0.98)	0.0014	0.96
rs67931591	1	215,330,292	G	GCTGAGG-CAGGAGA	1q41	C5orf56, IRF1, SLC22A5, IL5, RAD50	0.66	0.98 (0.96-0.99)	0.0020	0.65	0.94 (0.92-0.96)	4.3E-08	0.003
rs2522057	5	131,801,947	C	G	5q31.1	GTF2IRD2, STAG3L2, PMSZP5, WBSCR16	0.60	0.96 (0.95-0.98)	2.9E-07	0.60	0.98 (0.96-1.01)	0.14	0.011
rs1637365	7	74,359,358	T	C	7q11.23	CSK, CYP11A2, ULK3, MPI, SCAMP2, CRLX3, ARID3B, PTPN9	0.30	1.02 (1.00-1.04)	0.026	0.32	1.08 (1.05-1.11)	1.0E-08	3.9E-04
rs1869959	15	75,147,332	A	C	15q24.1	PTPN9, SCAMP5, C15orf39, COMMD4, SIN3A, SNUPN, ADAMTS17, PCSK6	0.30	0.96 (0.94-0.97)	6.0E-08	0.31	0.96 (0.94-0.99)	0.0022	0.64
rs60381548	15	75,728,474	CA	C	15q24.2		0.27	0.96 (0.95-0.98)	1.1E-05	0.28	0.95 (0.93-0.98)	1.7E-04	0.47
rs181337095	15	100,907,094	A	G	15q26.3		0.85	1.05 (1.03-1.08)	1.1E-05	0.83	1.06 (1.03-1.10)	2.9E-04	0.60

TAF Test allele frequency, SNP single nucleotide polymorphism, OR odds ratio, CI confidence intervals, ^ap for heterogeneity between ER-positive and ER-negative tumors.

significance with cancers. Associations with mosquito bite size and asthma had been reported for rs2522057 and SNPs in LD with this lead SNP. For the 15q24 region, associations with cardiovascular phenotypes have been previously reported for rs1869959 while body height, glomerular filtration rate, and type 2 diabetes have been associated with rs60381548 and SNPs highly correlated with this lead SNP (Supplementary Table 4).

The eQTL analysis of breast tumors revealed significant associations in four loci: 1p13.3, 5q31, 15q24.1, and 15q24.2 (Supplementary Table 5A). There were significant associations ($P < 10^{-6}$) between the protective allele of rs17024629 (T allele) at 1q13.3 and increased expression of *GSTM1*, *GSTM2*, and *GSTM4*, which are located 19 kb, 31 kb, and 51 kb downstream of the SNP. At 5q31, the top SNP rs2522057, located 15 kb downstream of *IRF1*, was most significantly associated with the gene's expression levels. At 15q24.1, rs1869959, located 35 kb upstream of *MPI* and 12 kb upstream of *ULK3*, was significantly associated with the expression of these two genes. The other top SNP at the 15q24 locus, rs60381548, located intron of *SIN3A*, 30 kb downstream of *PTPN9*, 162 kb downstream of *SNUPN*, and 212 kb upstream of *SNX33* was correlated with all four genes. The 1q41 locus revealed a significant association between rs67931591 and *PTPN14*. The SNP at 7q11.23 was significantly correlated with *STAG3L2*, a pseudogene. Previous published report on normal breast tissues from the GTEx revealed associations between rs2522057 and the *SLC22A5* gene, and between rs1869959 and the *ULK3* gene (Supplementary Table 5B).

Functional annotation analyses pointed out relationships with genomic functional biofeatures for rs2522057, rs17024629, rs1869959, and rs60381548 or SNPs in strong LD with these top SNPs in breast tissue-originated cell lines (Supplementary Tables 6, 7A, 7B). Active enhancer and promoter states were found for SNPs in strong LD with rs2522057 (rs2188962, rs4705950, rs4705950, rs72797306, rs11741255) using the 25-state chromatin model. Additional associations were found with histone modifications. These included: H3K4me1 and H3K27ac enhancer peaks for: rs2522057 and other SNPs in strong LD (rs2188962, rs17622378, rs12521868, rs146604341, rs11951091, rs6866614, rs4705950, rs72797303, rs2706396, rs2522052, rs2706403, rs2706336, rs72797306, rs2248116, rs11741255); those in strong LD with rs17024629 (rs538388, rs560674, rs568686, rs669426, rs3850616, rs17024628); a SNP in strong LD with rs1869959 (rs7180432); and for the top SNP rs60381548. H3K4me3 and H3K9ac promoter peaks were found for: rs2522057 and other SNPs in strong LD (rs12515180, rs11951091, rs72797306); SNPs in LD with rs17024629 (rs538388, rs669426, rs3850616); rs1869959 and other SNPs in strong LD (rs4886613, rs936230).

We evaluated the consistency of the association of the identified loci in Latinos, and found the effect and direction of the association were consistent in 8 out of 11 evaluated variants (Supplementary Table 8). However, none of these consistent variants was statistically significant at $p < 0.05$ in the Latino study of 2385 cases and 6416 controls.

Discussion

We found seven variants associated with breast cancer risk among women of African ancestry that may contribute to better prediction of breast cancer risk and provide further insights into mechanisms of breast cancer carcinogenesis. Although the discovery of the loci is largely driven by effects in European ancestry populations, observation of risk loci in multiple ancestral populations lends credence to the chances of those variants being causal. We designed our current approach of cross-ancestry meta-analysis to uncover genetic variants shared across ancestry.

Table 3 Conditional regression analysis of top SNPs and others in the loci.

Locus	Variants	Position	Test/other alleles	Marginal analysis		Conditional on lead SNP		Conditional on all other independent SNPs at locus ^a	
				OR (95% CI)	P value	OR (95% CI)	P value	OR (95% CI)	P value
Overall									
1p13.3	rs17024629	110,179,756	T/C	0.95 (0.94-0.97)	3.0E-08			0.96 (0.94-0.97)	1.07E-07
1p13.3	rs11633925	109,926,599	T/G	0.94 (0.91-0.97)	1.3E-04			0.93 (0.90-0.96)	1.46E-06
1p13.3	rs114351980	110,219,028	C/T	1.08 (1.04-1.12)	7.3E-05	0.94 (0.91-0.97)	1.4E-04	1.08 (1.04-1.12)	1.80E-05
1p13.3	rs10969874C:T	109,969,874	T/C	1.04 (1.02-1.06)	2.7E-04	1.08 (1.03-1.11)	8.2E-05	1.05 (1.02-1.07)	3.41E-05
15q24.2	rs60381548	75,728,474	C/A	0.96 (0.95-0.97)	6.6E-09	1.04 (1.02-1.06)	1.8E-04	0.97 (0.96-0.99)	8.62E-09
15q24.1	rs1869959	75,147,332	A/C	0.96 (0.95-0.98)	4.6E-08			0.96 (0.95-0.97)	1.57E-09
15q24.2	rs113939578 ^b	75,479,704	T/C	0.96 (0.94-0.98)	8.0E-05	0.96 (0.94-0.98)	1.9E-05	0.96 (0.94-0.98)	9.13E-06
15q24.2	rs12917507 ^b	75,953,903	T/G	1.02 (1.01-1.03)	0.0030	1.02 (1.01-1.03)	1.3E-04	1.02 (1.01-1.04)	6.51E-05
15q26.3	rs181337095	100,907,094	A/G	1.05 (1.04-1.07)	1.8E-08			1.05 (1.04-1.08)	1.26E-08
15q26.3	rs117793215	100,535,681	T/C	0.93 (0.89-0.97)	1.6E-04	0.93 (0.89-0.96)	7.7E-05	0.93 (0.89-0.96)	7.71E-05
5q31.1	rs2522057	131,801,947	C/G	0.97 (0.95-0.98)	1.1E-08			0.96 (0.95-0.97)	2.39E-10
5q31.1	rs132149322;G;GGCCCGCCGCC	132,149,322	GGCCCGCCGCC/G	1.03 (1.01-1.04)	7.4E-04	1.03 (1.02-1.05)	2.1E-05	1.04 (1.02-1.05)	2.13E-05
ER negative									
1q41	rs67931591	215,330,292	G/GCTGAGGCAGGAGA	0.94 (0.92-0.96)	4.3E-08	0.94 (0.92-0.96)	1.5E-06	0.94 (0.93-0.96)	1.59E-12
1q41	rs5780828	215,416,434	T/A	0.96 (0.93-0.98)	9.7E-05			0.94 (0.92-0.96)	1.46E-06

SNP Single nucleotide polymorphism, OR odds ratio, CI confidence intervals.

^aAll independent SNPs (Joint ORs) at each locus were in the same model.

^bThe conditional ORs are conditioned on rs60381548 and rs1869959.

The SNPs identified in this study lie in regions that are close to genes that have been previously implicated in cancer. Interestingly we found three variants located within the introns of genes. One of the variants, rs67931591 was found in *KCNK2* (also known as *TREK1*), which encodes the protein potassium channel subfamily K member 2, a member of the two-pore-domain background potassium channel family. Potassium channels are known to play a role in cancer and studies using TCGA data have shown associations with DNA methylation in the *KCNK* genes and triple negative breast cancer. Additionally, overexpression of *KCNK5*, *KCNK9*, and *KCNK12* and under-expression of *KCNK6* and *KCNK15* were associated with triple negative breast cancer¹⁷. Other studies investigated expression of *KCNK2* gene as potential prognostic markers. For example, Innamaa et al.¹⁸ found increased *KCNK2* expression in human ovaries and a role in cell proliferation and apoptosis for *KCNK2* modulators in ovarian cancer cell lines. Li et al.¹⁹ found differential expression of *KCNK2*, *KCNK15* and *KCNK17* in liver cancer cells compared to healthy tissue. *KCNK2* has also been reported in amplified regions in a genome-wide scan of chromosomal alterations in esophageal squamous cell carcinoma²⁰.

We found two independent SNPs at the 15q24 locus at about 582 kb apart (rs1869959 at 15q24.1 in the *SCAMP2* intron and rs60381548 at 15q24.2 in the *SIN3A* gene). The *SIN3A* gene was associated with rs60381548 in the eQTL analysis of breast tumor in the present study. Switch-independent 3 family A (*SIN3A*) is a transcriptional regulator, that along with its paralogs and corepressors play important roles in normal breast development, cancer and metastasis²¹⁻²³. Furthermore, *SIN3A* mediates STAT3 transcriptional repressor activity²⁴ and along with genes involved in histone modification such as *HDAC* and Lysine specific demethylase (*LSD*), inhibits several cancer genes including *CASP7*, *TGFB2*, *CDKN1A*, *HIF1A*, *TERT* and *MDM2*²⁵. Studies have shown key roles for *SIN3A* in breast cancer including sensitivity to chemotherapy²⁵ and breast cancer progression^{26,27}.

The other SNP at 15q24, rs1869959, is located in the intron of the *SCAMP2* gene that codes for secretory carrier associated membrane protein 2 that functions as carriers to the cell surface in post-golgi recycling pathways²⁸. The recent GTEx project pilot study found significant associations between the SNP and *SCAMP2* in esophageal mucosa, *ULK3* in breast mammary tissue, adipose, whole blood, and lung tissue^{29,30}. We also found that rs1869959 was associated the expression of *ULK3* in breast tumor. *ULK3* is a serine threonine kinase that activates *GLI2*, a key component of the Hedgehog signaling pathway, and implicated in many cancers^{31,32}.

Similarly, the *C5orf56* gene harboring the rs2522057 SNP returned no interesting associations with cancer. However, nearby genes in the 5q31 locus included *RAD50*, that codes for a DNA repair protein, a part of the MRE11-RAD50-NBS1 complex³³. Other nearby genes include *SLC22A5* solute carrier family 22 member 5 encoding the OCTN2 (organic cation transporter protein), and *IRF1* that encodes interferon regulatory factor 1. *SLC22A5* is an estrogen-dependent gene whose expression is associated with ER status in breast cancer cell lines and tissue specimens³⁴. Significantly decreased levels of *SLC22A5* have been reported in colorectal cancer tissues compared to normal tissues in eQTL studies³⁵. Moreover, eQTL studies report associations between rs2522057 and gene expression in several tissues including breast mammary tissue, lymphocytes, esophageal mucosa, lung, skeletal muscle, skin, thyroid and whole blood^{29,30}. We found an association between rs2522057 and *IRF1* expression levels in the eQTL analysis of breast cancer in this study lending support to the likelihood of involvement of the *IRF1* gene in the mechanism of the SNP on breast cancer carcinogenesis.

Additionally, IRF1 has been shown to have tumor suppressor functions in breast cancer through its inhibition of NF- κ B³⁶ and CASP8 activation and induction of apoptosis³⁷.

The majority of GWAS-identified SNPs were located in non-coding regions of the genome, and three loci in the present study were found in intergenic regions. The closest gene to rs17024629 is *AMPD2* (high adenosine monophosphate deaminase 2) and has recently been shown to predict worse outcomes in undifferentiated pleomorphic sarcoma³⁸. Earlier studies³⁹ found high expression levels of *AMPD2* in hepatocellular carcinoma, though the levels did not differ substantially from those in the non-tumorous organ. It is noteworthy that our eQTL analysis did not find a significant association with *AMPD2* expression. The carcinogen metabolism genes, *GSTM1*, *GSTM2*, and *GSTM4* are also located in this region and our eQTL analysis of breast tumor revealed highly significant associations between rs17024629 and these genes. The *GSTM1* null genotype has been associated with several cancers including cancers of the colorectum, oral cavity, lung, cervix, and stomach^{40–47}. In eQTL studies, *GSTM4* was significantly associated with gene expression in several tissues including the aorta, lungs, tibia nerve and whole blood^{29,30}.

The rs1637365 SNP at the 7q11.23 locus is near the *CASTOR2* gene (cytosolic arginine sensor for mTORC1 protein, also known as *GATSL1*, GATS-like protein 1). The CASTOR proteins are arginine sensors that function as negative regulators of the TORC1 signaling pathway, an often dysregulated pathway in human cancer, through the GATOR complex, inhibiting mTORC1^{48,49}. The rs181337095 SNP is located 6 kb 5' of RP11-168G16.2, an antisense DNA.

A potential limitation of this study is the different genotyping platforms used by the different consortia. However, stringent QC measures pre- and post-imputation were carried out. Additionally, the meta-analysis did not reveal significant heterogeneity across studies. Secondly, the sample size for ER-negative breast cancer cases was relatively small, thus reducing the precision of the estimates and providing less power for detecting risk loci. The third limitation is related to the additional SNPs identified at the same loci with the index SNPs from the conditional regression analysis. The regression procedures were based on a liberal p value cutoff of 10^{-4} , and the chance that some of the identified SNPs could be spurious findings cannot be ruled out. Another noteworthy point is that identification of genetic variants in GWAS is just the first step of the discovery of true causal variants and genes associated with breast cancer. Further studies are needed, including in vitro and in vivo functional studies to elucidate the mechanisms by which identified putative causal variants are acting and identify the targeted genes. Finally, although the direction and strength of the associations were consistent between African and European populations, and mostly consistent with Latino populations, we could not find statistically significant replication of the identified variants, which are likely due to the modest sample sizes of the Latino study.

Our study found six loci that could provide further insights into pathways for breast cancer carcinogenesis. The genetic variants that shared across ancestry populations makes them possible causal variants. Functional studies on these loci are desirable to identify causal variants and elucidate the mechanisms of breast cancer carcinogenesis. In addition, future studies can evaluate these variants for breast cancer risk prediction, particularly in African ancestry populations.

Methods

Study population. Data for this study were obtained from four consortia of African ancestry populations (ROOT, AMBER, AABC, and BCAC-African

ancestry)¹⁶ and the Ghana Breast Health Study (GBHS)^{50,51}, with a combined sample size of 19434 participants including 9241 cases and 10193 controls (Supplementary Table 1). Estimates from these studies were meta-analysed to generate pooled African ancestry estimates of breast cancer risk. Additionally, we used summary estimates (odds ratios, ORs) of breast cancer from European ancestry BCAC datasets (GWAS, iCOGs and OncoArray) with a combined sample size of 228,951 (122,977 cases and 105,974 controls)⁶.

Genotyping and quality control. Genotyping and quality control (QC) procedures have been described in detail for the three consortia¹⁶ and the BCAC European ancestry data⁶. The AABC was genotyped using the Illumina Human 1M-Duo BeadChip. After QC, a total of 3007 cases (1518 ER-positive, 987 ER-negative) and 2720 controls remained in the analysis⁵². Genotyping in the ROOT consortium was done using Illumina HumanOmni 2.5-8v1 array and 1657 cases (374 ER-positive, 403 ER-negative) and 2029 controls passed QC. In the BCAC-African ancestry consortium, genotyping was done using the Illumina OncoArray (260K GWAS backbone) and after removing overlapped samples between OncoArray with AABC, AMBER and ROOT and samples failed in QC, a total of 2271 cases (1130 ER-positive, 613 ER-negative) and 1406 controls remained for analysis. The Illumina MEGA array was used for genotyping in the AMBER consortium, and 1407 cases (952 ER-positive, 385 ER-negative) and 2408 controls remained in analysis passed QC. In the GBHS, Illumina Global Screening Array was used for genotyping, and 899 cases (296 ER-positive, 277 ER-negative) and 1630 controls were included in analysis. Imputation for all studies was done using the cosmopolitan reference panel in the 1000 Genomes Project (Phase 3 release).

In addition, we examined the association between the identified SNPs of interest and breast cancer risk in a GWAS of Latinos (2385 cases and 6416 controls). Details of the genotyping, QC and data analysis have been published⁵³.

Data analysis

GWAS. In the ROOT and AABC GWAS studies, genotyped SNPs were analyzed and imputed with imputation score >0.3 and minor allele frequency >0.01 to account for uncertainty in imputation. Unconditional logistic regression was used to examine the association of each SNP and breast cancer risk adjusting for age, study site and eigenvectors from Principal Components Analysis (PCA). In the ROOT GWAS, the first four eigenvectors were used to control for population stratification as only the first 4 eigenvectors were associated with case status. The AABC GWAS adjusted for the first 10 eigenvectors from the PCA. OR and 95% confidence intervals (CI) were calculated from the multivariable logistic regressions. All tests of statistical significance were two sided. Using similar methods, separate analyses were conducted to compare ER-positive and ER-negative breast cancers with controls. The AMBER consortium estimated ORs and P values using unconditional logistic regression, adjusting for 10-year age group, sample type (saliva, blood, other), study (Black Women's Health Study (BWHS) versus others) and PCs that associated with breast cancer at $P < 0.1$. The GBHS estimated per-allele ORs and 95% CI for each SNP on allele counts (dosages) using unconditional logistic regression adjusting for the first ten principal components, self-reported ethnicity and age. In the Oncoarray African ancestry samples, a total of 27 million SNPs with MAF $\geq 0.1\%$ and imputation quality score ≥ 0.3 were included in the analysis. PCs were estimated using EIGENSTRAT. ORs and P value of each SNP were estimated using unconditional logistic regression, adjusting for age, study (Women of African Ancestry Breast Cancer Study—WAABCS versus other) and the first ten PCs.

The BCAC European study used a two-stage imputation approach, using SHAPEIT2 for phasing and IMPUTE version 2 for imputation. The first ten principal components and country were adjusted for in the logistic regression, and per-allele ORs and standard errors were computed⁶.

Meta-analysis. Regression coefficient estimates from the five contributing African ancestry studies were combined in a fixed effects meta-analysis using METAL⁵⁴. Variants associated with breast cancer at $P < 0.05$ from the African ancestry meta-analysis were then combined in another fixed effects meta-analysis with the coefficients from the BCAC European ancestry data. Heterogeneity in both meta-analyses was assessed using the I^2 statistic. SNPs that were significant genome-wide ($P < 5 \times 10^{-8}$) in the cross-ancestry meta-analysis, and >500 kb away from the 180 loci known to be associated with breast cancer risk were identified^{5,6}. Conditional analysis below confirmed the identified loci. All analyses were done separately for ER-positive, ER-negative, and overall breast cancer risk.

Regression analysis conditional on index SNPs. In order to identify independent SNPs in the identified loci, conditional analysis was done in each of the regions, including all variants in the flanking ± 500 kb region of the lead SNP. The 15q24 region had two SNPs about 582 kb apart that were both genome-wide significant (see results for details). Hence, all variants in the region extending from 500 kb upstream of the proximal SNP and 500 kb downstream of the other SNP were included in the conditional analysis for this region. We used the GCTA software with the $-COJO$ option⁵⁵, that utilizes summary statistics and population-specific linkage disequilibrium (LD) from 1000 Genomes Project, for the computation of

conditional beta coefficients. SNPs significant at $p < 10^{-4}$ after adjusting for lead SNP were considered as independent signals. The $p < 10^{-4}$ cutoff was derived by applying a factor of 3000 (the ratio of the 3 billion base pairs genome-wide to the 1 million base pairs in each region in the conditional analysis) to the GWAS significance of 5×10^{-8} . This procedure was repeated until no additional independent signals were significant. In addition to the conditional analysis involving the lead SNP and one other candidate SNP, we also determined joint ORs including all independent loci in the same model. Separate analyses were done for African and European ancestry data, and the estimates from the conditional analysis were combined in a meta-analysis.

Functional annotation. The functional annotations of the SNPs were determined using HaploReg v4.1⁵⁶. Using data from ENCODE⁵⁷ and the Roadmap Epigenomics Consortium⁵⁸, we examined the chromatin states including core 15-state model and 25-state model using 12 imputed marks, H3K4me1 and H3K27ac (enhancers), and H3K4me3 and H3K9ac (promoters) for each identified SNP and other SNPs in strong LD with these lead SNPs (>0.8). We also assessed evolutionary conserved regions, DNase hypersensitivity sites, and variant effect on regulatory motifs, proteins bound and eQTL hits from previous studies.

eQTL analysis. We carried out a cis-eQTL analysis to understand possible target genes in the six loci. All genes within ± 1 MB around each index SNP were evaluated and gene expression in breast tumors from TCGA breast cancer patients (African ancestry, $n = 164$ and European ancestry, $n = 778$) were used in the analysis. A linear regression model estimated additive effects for each SNP, adjusting for age, ancestry, copy number variation, batch effect, and molecular subtype. Separate analyses were done for African and European ancestry samples and the estimates were meta-analysed to obtain overall estimates. Bonferroni significance levels were applied to determine statistical significance. We also checked associations between the identified loci and gene expression in several tissues, including normal breast, that had been published from previous eQTL analyses on the Haploreg website.

Allelic pleiotropy. We assessed the GWAS catalog (www.ebi.ac.uk) for previously reported associations for the identified lead SNPs and all other SNPs in LD with $r^2 > 0.4$ and phenotypes.

Ethical approval. Informed consent was obtained from all subjects included in the analysis. The relevant ethical review boards at all participating institutions approved study protocols.

Data availability

The genotype datasets used in this study are publicly available via dbGaP (<https://www.ncbi.nlm.nih.gov/gap/>) including AABC under accession code [phs000851.v1.p1](https://www.ncbi.nlm.nih.gov/gap/acc.cgi?acc=phs000851.v1.p1), ONCO under accession code [phs001265.v1.p1](https://www.ncbi.nlm.nih.gov/gap/acc.cgi?acc=phs001265.v1.p1), AMBER under accession code [phs000669.v1.p1](https://www.ncbi.nlm.nih.gov/gap/acc.cgi?acc=phs000669.v1.p1), ROOT under accession code [phs000383.v1.p1](https://www.ncbi.nlm.nih.gov/gap/acc.cgi?acc=phs000383.v1.p1), and GBHS under accession code [phs002387.v1.p1](https://www.ncbi.nlm.nih.gov/gap/acc.cgi?acc=phs002387.v1.p1). Data for TCGA is available via <https://portal.gdc.cancer.gov/>. The remaining data are available within the Article, Supplementary Information or Source Data file.

Code availability

The codes and summary statistics of the top 10000+ variants are stored on github (<https://github.com/gmgaous/Cross-ancestry-GWAS-for-breast-cancer>). The remaining data are available within the Article, Supplementary Information or available from the authors upon request.

Received: 23 July 2020; Accepted: 2 June 2021;

Published online: 07 July 2021

References

- Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424 (2018).
- Lichtenstein, P. et al. Environmental and heritable factors in the causation of cancer—analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000).
- Mucci, L. A. et al. Familial risk and heritability of cancer among twins in nordic countries. *JAMA* **315**, 68–76 (2016).
- Couch, F. J., Nathanson, K. L. & Offit, K. Two decades after BRCA: setting paradigms in personalized cancer care and prevention. *Science* **343**, 1466–1470 (2014).
- Lilyquist, J., Ruddy, K. J., Vachon, C. M. & Couch, F. J. Common genetic variation and breast cancer risk—past, present, and future. *Cancer Epidemiol. Biomark. Prev.* **27**, 380–394 (2018).
- Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
- Shu, X. et al. Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. *Nat. Commun.* **11**, 1217 (2020).
- Mavaddat, N. et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am. J. Hum. Genet.* **104**, 21–34 (2019).
- Milne, R. L. et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet.* **49**, 1767–1778 (2017).
- Michailidou, K. et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47**, 373–380 (2015).
- Fletcher, O. et al. Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. *J. Natl Cancer Inst.* **103**, 425–435 (2011).
- Turnbull, C. et al. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* **42**, 504–507 (2010).
- Ghoussaini, M. et al. Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat. Genet.* **44**, 312–318 (2012).
- Park, S. L., Cheng, I. & Haiman, C. A. Genome-wide association studies of cancer in diverse populations. *Cancer Epidemiol. Biomark. Prev.* **27**, 405–417 (2018).
- Haiman, C. A. et al. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat. Genet.* **43**, 1210–1214 (2011).
- Huo, D. et al. Genome-wide association studies in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer. *Hum. Mol. Genet.* **25**, 4835–4846 (2016).
- Dookeran, K. A., Zhang, W., Stayner, L. & Argos, M. Associations of two-pore domain potassium channels and triple negative breast cancer subtype in The Cancer Genome Atlas: systematic evaluation of gene expression and methylation. *BMC Res. Notes* **10**, 475 (2017).
- Innamaa, A. et al. Expression and effects of modulation of the K2P potassium channels TREK-1 (KCNK2) and TREK-2 (KCNK10) in the normal human ovary and epithelial ovarian cancer. *Clin. Transl. Oncol.* **15**, 910–918 (2013).
- Li, W. C. et al. KCNK levels are prognostic and diagnostic markers for hepatocellular carcinoma. *Aging (Albany NY)* **11**, 8169–8182 (2019).
- Chattopadhyay, I. et al. Genome-wide analysis of chromosomal alterations in patients with esophageal squamous cell carcinoma exposed to tobacco and betel quid from high-risk area in India. *Mutat. Res.* **696**, 130–138 (2010).
- Lewis, M. J. et al. SIN3A and SIN3B differentially regulate breast cancer metastasis. *Oncotarget* **7**, 78713–78725 (2016).
- Bansal, N., David, G., Farias, E. & Waxman, S. Emerging Roles of Epigenetic Regulator Sin3 in Cancer. *Adv. Cancer Res.* **130**, 113–135 (2016).
- Ellison-Zelski, S. J. & Alarid, E. T. Maximum growth and survival of estrogen receptor-alpha positive breast cancer cells requires the Sin3A transcriptional repressor. *Mol. Cancer* **9**, 263 (2010).
- Gambi, G. et al. The transcriptional regulator Sin3A contributes to the oncogenic potential of STAT3. *Cancer Res.* **79**, 3076–3087 (2019).
- Yang, Y. et al. LSD1 coordinates with the SIN3A/HDAC complex and maintains sensitivity to chemotherapy in breast cancer. *J. Mol. Cell Biol.* **10**, 285–301 (2018).
- Watanabe, K. et al. A novel somatic mutation of SIN3A detected in breast cancer by whole-exome sequencing enhances cell proliferation through ERalpha expression. *Sci. Rep.* **8**, 16000 (2018).
- Li, W. et al. The FOXN3-NEAT1-SIN3A repressor complex promotes progression of hormonally responsive breast cancer. *J. Clin. Investig.* **127**, 3421–3440 (2017).
- Singleton, D. R., Wu, T. T. & Castle, J. D. Three mammalian SCAMPs (secretory carrier membrane proteins) are highly related products of distinct genes having similar subcellular distributions. *J. Cell Sci.* **110**, 2099–2107 (1997).
- Consortium, G. T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
- Westra, H. J. et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
- Goruppi, S. et al. The ULK3 kinase is critical for convergent control of cancer-associated fibroblast activation by CSL and GLI. *Cell Rep.* **20**, 2468–2479 (2017).
- Maloverjan, A., Piirsoo, M., Michelson, P., Kogerman, P. & Osterlund, T. Identification of a novel serine/threonine kinase ULK3 as a positive regulator of Hedgehog pathway. *Exp. Cell Res.* **316**, 627–637 (2010).
- Zhong, Q. et al. Association of BRCA1 with the hRad50-hMre11-p95 complex and the DNA damage response. *Science* **285**, 747–750 (1999).

34. Wang, C., Uray, I. P., Mazumdar, A., Mayer, J. A. & Brown, P. H. SLC22A5/OCTN2 expression in breast cancer is induced by estrogen via a novel intronic estrogen-response element (ERE). *Breast Cancer Res. Treat.* **134**, 101–115 (2012).
35. Zou, D. et al. Integrative expression quantitative trait locus-based analysis of colorectal cancer identified a functional polymorphism regulating SLC22A5 expression. *Eur. J. Cancer* **93**, 1–9 (2018).
36. Armstrong, M. J. et al. IRF-1 inhibits NF-kappaB activity, suppresses TRAF2 and cIAP1 and induces breast cancer cell specific growth inhibition. *Cancer Biol. Ther.* **16**, 1029–1041 (2015).
37. Bouker, K. B. et al. Interferon regulatory factor-1 (IRF-1) exhibits tumor suppressor activities in breast cancer associated with caspase activation and induction of apoptosis. *Carcinogenesis* **26**, 1527–1535 (2005).
38. Orth, M. F. et al. Functional genomics identifies AMPD2 as a new prognostic marker for undifferentiated pleomorphic sarcoma. *Int. J. Cancer* **144**, 859–867 (2019).
39. Szydłowska, M. & Roszkowska, A. Expression patterns of AMP-deaminase isozymes in human hepatocellular carcinoma (HCC). *Mol. Cell Biochem.* **318**, 1–5 (2008).
40. Li, J. et al. Breast cancer genetic risk profile is differentially associated with interval and screen-detected breast cancers. *Ann. Oncol.* **26**, 517–522 (2015).
41. Liu, X. et al. Meta-analysis of GSTM1 null genotype and lung cancer risk in Asians. *Med. Sci. Monit.* **20**, 1239–1245 (2014).
42. Cai, X., Yang, L., Chen, H. & Wang, C. An updated meta-analysis of the association between GSTM1 polymorphism and colorectal cancer in Asians. *Tumour Biol.* **35**, 949–953 (2014).
43. Zhao, Y. et al. GSTM1 polymorphism and lung cancer risk among East Asian populations: a meta-analysis. *Tumour Biol.* **35**, 6493–6500 (2014).
44. Zhang, X. L. & Cui, Y. H. GSTM1 null genotype and gastric cancer risk in the Chinese population: an updated meta-analysis and review. *Onco Targets Ther.* **8**, 969–975 (2015).
45. Yang, H. et al. The association of GSTM1 deletion polymorphism with lung cancer risk in Chinese population: evidence from an updated meta-analysis. *Sci. Rep.* **5**, 9392 (2015).
46. Gu, J. et al. GSTM1 null genotype is associated with increased risk of gastric cancer in both ever-smokers and non-smokers: a meta-analysis of case-control studies. *Tumour Biol.* **35**, 3439–3445 (2014).
47. Economopoulos, K. P., Choussein, S., Vlahos, N. F. & Sergentanis, T. N. GSTM1 polymorphism, GSTT1 polymorphism, and cervical cancer risk: a meta-analysis. *Int. J. Gynecol. Cancer* **20**, 1576–1580 (2010).
48. Kim, L. C., Cook, R. S. & Chen, J. mTORC1 and mTORC2 in cancer and the tumor microenvironment. *Oncogene* **36**, 2191–2201 (2017).
49. Chantranupong, L. et al. The CASTOR Proteins Are Arginine Sensors for the mTORC1 Pathway. *Cell* **165**, 153–164 (2016).
50. Nyante, S. J. et al. Recruiting population controls for case-control studies in sub-Saharan Africa: the Ghana Breast Health Study. *PLoS ONE* **14**, e0215347 (2019).
51. Brinton, L. A. et al. Design considerations for identifying breast cancer risk factors in a population-based study in Africa. *Int. J. Cancer* **140**, 2667–2677 (2017).
52. Feng, Y. et al. A comprehensive examination of breast cancer risk loci in African American women. *Hum. Mol. Genet.* **23**, 5518–5526 (2014).
53. Hoffman, J. et al. Identification of novel common breast cancer risk variants at the 6q25 locus among Latinas. *Breast Cancer Res.* **21**, 3 (2019).
54. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genome-wide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
55. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
56. Ward, L. D. & Kellis, M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* **44**, D877–D881 (2016).
57. Encode Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
58. Roadmap Epigenomics C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

Acknowledgements

The ROOT Consortium was supported by National Cancer Institute grants R01-CA142996, R01-CA89085, R01-CA228198, and P20-CA233307. D.H. and G.G. were also partially supported by the National Cancer Institute (R03-CA227357 and R01-CA242929). D.H., O.I.O., and K.L.N. were also partially supported by Breast Cancer Research Foundation (BCRF-20-071). K.L.N. was also supported by Bassett Center for BRCA. AABC was supported by a Department of Defense Breast Cancer Research Program Era of Hope Scholar Award to C.A.H. [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, K05 CA136967); WCHS (U.S. Army Medical Research and Materiel 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-CA077305 and United States Army Medical Research Program grant DAMD-17-96-6071); NC-BCFR (National Institutes of Health grant U01-CA069417); 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). 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. MP was supported by Breast Cancer Research Foundation, Tower Cancer Research Foundation, and a gift from Dr. Richard Balch. AMBER was supported by the National Cancer Institute grants P01-CA151135, R01-CA098663, R01-CA058420, UM1-CA164974, R01-CA100598, P50-CA58223, and the University Cancer Research Fund of North Carolina. JRP was supported by the Susan G. Komen Foundation and the Karin Grunebaum Foundation. Pathology data were obtained from numerous state cancer registries (Arizona, California, Colorado, Connecticut, Delaware, District of Columbia, Florida, Georgia, Hawaii, Illinois, Indiana, Kentucky, Louisiana, Maryland, Massachusetts, Michigan, New Jersey, New York, North Carolina, Oklahoma, Pennsylvania, South Carolina, Tennessee, Texas, Virginia). The results reported do not necessarily represent their views or the views of the National Institutes of Health. The study protocol for the BWHS was approved by the Boston University Medical Center IRB and the IRBs of participating cancer registries, as required. BCAC is funded by Cancer Research UK [C1287/A16563, C1287/A10118], the European Union's Horizon 2020 Research and Innovation Program (grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), and by the European Community's Seventh Framework Program under grant agreement number 223175 (grant number HEALTH-F2-2009-223175) (COGS). The EU Horizon 2020 Research and Innovation Program funding source had no role in study design, data collection, data analysis, data interpretation or writing of the report. The Sister Study was funded by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES044005). GBHS authors acknowledge the research contributions of the Cancer Genomics Research Laboratory for their expertise, execution, and support of this research in the areas of project planning, wet laboratory processing of specimens, and bioinformatics analysis of generated data. This project has been funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, under NCI Contract No. 75N910D00024. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Author contributions

Study design: D.H., C.A.H.; Data analysis: B.A., Z.D., and G.G.; Data interpretation: B.A., Z.D., G.G., M.G., J.R.P., C.A.H., and D.H.; Writing of the paper: B.A., D.H.; Review of the paper: B.A., Z.D., G.G., T.A., K.L.L., G.Z., J.F., E.M.J., L.B., W.Z., J.J.H., R.G.Z., S.N., E.V.B., S.A.I., M.F.P., S.L.D., J.L.R., S.Y., T.O.O., O.O., W.B., M.A.T., K.L.N., A.H., B.N., S.A., P.N.F., L.E.S., J.T.B., L.H.K., G.T., D.H., L.F., M.K.B., J.D., A.M.D., D.F.E., K.M., P.D.P.P., Q.W., D.P.S., J.A.T., K.M.O., C.M.K., A.G.F., C.B., J.Y., B.A., B.A.W. S.J.C., A.F.O., C.B.A., D.V.C., E.Z., O.I.O., M.G., J.R.P., C.A.H., and D.H.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-021-24327-x>.

Correspondence and requests for materials should be addressed to C.A.H. or D.H.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021

¹Center for Clinical Cancer Genetics and Global Health, Department of Medicine, University of Chicago, Chicago, IL, USA. ²Department of Preventative Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. ³Department of Public Health Sciences, University of Chicago, Chicago, IL, USA. ⁴Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. ⁵Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. ⁶Slone Epidemiology Center, Boston University, Boston, MA, USA. ⁷Usher Institute and CRUK Edinburgh Centre, University of Edinburgh, Edinburgh, UK. ⁸Departments of Epidemiology & Population Health and of Medicine (Oncology) and Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA. ⁹Biomarkers of Early Detection and Prevention, Department of Population Sciences, Beckman Research Institute, City of Hope Comprehensive Cancer Center, Duarte, CA, USA. ¹⁰Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville, TN, USA. ¹¹Department of Public Health Sciences, University of Miami, Miami, FL, USA. ¹²Department of Radiology, University of North Carolina, Chapel Hill, NC, USA. ¹³Cancer Prevention and Control Program, Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA. ¹⁴Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. ¹⁵Genomics, Development and Disease Section, Genetic Disease Research Branch, National Human Genome Research Institute, NIH, Bethesda, MD, USA. ¹⁶Department of Cancer Prevention and Control, Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA. ¹⁷Department of Surgery, College of Medicine, University of Ibadan, Ibadan, Nigeria. ¹⁸Center for Population and Reproductive Health, College of Medicine, University of Ibadan, Ibadan, Nigeria. ¹⁹Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA. ²⁰Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. ²¹University of the West Indies, Bridgetown, Barbados. ²²Department of Family, Population and Preventive Medicine, Stony Brook University, Stony Brook, NY, USA. ²³Laboratory of Human Carcinogenesis, National Cancer Institute, Bethesda, MD, USA. ²⁴Department of Veterinary Biosciences, College of Veterinary Medicine, The Ohio State University, Columbus, OH, USA. ²⁵Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA. ²⁶Center for Population Health Research, Instituto Nacional de Salud Publica, Cuernavaca, Mexico. ²⁷Department of Medicine, University of California San Francisco, San Francisco, CA, USA. ²⁸Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. ²⁹Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK. ³⁰Biostatistics Unit, The Cyprus Institute of Neurology & Genetics, Nicosia, Cyprus. ³¹Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA. ³²Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. ³³Institute for Advanced Medical Research and Training, College of Medicine, University of Ibadan, Ibadan, Oyo, Nigeria. ³⁴Department of Pharmaceutical Chemistry, University of Ibadan, Ibadan, Oyo, Nigeria. ³⁵Korle Bu Teaching Hospital, Accra, Ghana. ³⁶Komfo Anokye Teaching Hospital, Kumasi, Ghana. ³⁷Peace and Love Hospital, Kumasi, Ghana. *A list of authors and their affiliations appears at the end of the paper. ✉email: haiman@usc.edu; dhuo@health.bsd.uchicago.edu

The GBHS Study Team

Thomas U. Ahearn⁴, Jonine Figueroa⁷, Joel Yarney³⁵, Baffour Awuah³⁶, Beatrice Addai Wiafe³⁷ & Montserrat Garcia-Closas⁴

A full list of members and their affiliations appears in the Supplementary Information.