Genome- and transcriptome-wide association studies in African-ancestry women uncover new insight into breast cancer genetics and improve risk prediction

By

Guochong Jia

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPY

in

Epidemiology

August 12, 2022

Nashville, Tennessee

Approved:

Wei Zheng, PhD, MD

Jirong Long, PhD

Tuya Pal, MD

Maureen Sanderson, PhD

Ran Tao, PhD

ACKNOWLEDGEMENTS

This research was made possible by the support of the funding of African-American Breast Cancer Genetic Study (R01 CA202981) and Anne Potter Wilson Chair fund. I would like to thank all project investigators of the participating studies in the AABCG consortium, and thank all African-ancestry female participants in the studies.

I sincerely appreciate my dissertation committee: Drs. Wei Zheng, Jirong Long, Tuya Pal, Maureen Sanderson, and Ran Tao. Your expertise has helped me a lot in elevating this work. Particularly, I would like to express the deepest thank you to my mentor, Dr. Wei Zheng. Your passion and thoughtfulness as a professional epidemiologist inspire me and I appreciate having you as an example in my PhD training stage.

Thank you to all my lab members. It is enjoyable to work in our research group because of the help and encouragement from each other. Especially, many thanks to the valuable help from my senior group members Lang Wu, Xiang Shu, Yaohua Yang, Ying Liu, and Jie Ping.

Many thanks to the PhD program at Vanderbilt University. Thank you to my academic advisor, Alicia Beeghly-Fadiel, and our program manager, Melissa Krasnove.

Thank you to all my friends. Best wishes for your bright future.

Many thanks to my parents. I am forever grateful for your unconditional love, encourage, and support to me. I love you with all my heart.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2SISTER	the Two Sister Study
AABC	the African American Breast Cancer consortium
AABCG	African American Breast Cancer Genetic
AMBER	African American Breast Cancer Epidemiology and Risk consortium
AUC	area under the receiver operating characteristic curve
BBCS	Baltimore Breast Cancer Study
BEST	Black Women: Etiology and Survival of Triple-negative Breast Cancers Study
BioVU	the Vanderbilt Biobank
BNCS	Barbados National Cancer Study
BWHS	Black Women's Health Study
CARE	the Los Angeles component of the Women's Contraceptive and Reproductive Experiences Study
CBCS	Carolina Breast Cancer Study
CCPS	Chicago Cancer Prone Study
eQTL	expression quantitative trait loci
ER-neg	estrogen receptor negative
ER-pos	estrogen receptor positive
FFPE	formalin-fixed paraffin-embedded
FPKM	fragments per kilobase of transcripts per million mapped reads
GBHS	Ghana Breast Health Study
GWAS	genome-wide association studies

iCOGS	Collaborative Oncological Gene-environment Study
LD	linkage disequilibrium
MAF	minor allele frequency
MDABCS	M.D. Anderson Breast Cancer Study
MEC	Multiethnic Cohort Study
MEGA	Multi-Ethnic Genotyping Array
NBCS	Nigerian Breast Cancer Study
NBHS	Nashville Breast Health Study
NC-BCFR	Northern California Breast Cancer Family Registry
NYUWHS	New York University Women’s Health Study
OncoArray	Genetic Associations and Mechanisms in Oncology OncoArray consortium
OR	odds ratio
PC	principal component
PEER	probabilistic estimation of expression residuals
PLCO	the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial
PRS	polygenic risk scores
QC	quality control Quality control
ROOT	the GWAS of Breast Cancer in the African Diaspora consortium
RVGBC	Racial Variability in Genotypic Determinants of Breast Cancer Risk Study
SCCS	Southern Community Cohort Study
SFBCS	San Francisco Bay Area Breast Cancer Study
SISTER	the Sister Study
SNV	single nucleotide variant

STSBHS	Southern Tri-State Breast Health Study
TCGA	the Cancer Genome Atlas
TNBC	triple-native breast cancer
TopMed	Trans-Omics for Precision Medicine
TWAS	transcriptome-wide association study
USC	University of South California
USRT	the United States Radiologic Technologists cohort
WAABCS	Women of African Ancestry Breast Cancer Study
WashU	Washington University
WCHS	Women’s Circle of Health Study
WFBC	Wake Forest University Breast Cancer Study
WGS	whole genome sequencing

I. SPECIFIC AIMS

Breast cancer is the most common cancer among females in the U.S., with 287,850 estimated new cases in 2022.[1] Females of African ancestry have a 41% higher age-adjusted mortality of breast cancer than females of European ancestry,[2] and they are also more likely to develop triple negative breast cancer, a more aggressive subtype, compared with European-ancestry females.[3] Previous studies have shown that genetic factors play an important role in the etiology of breast cancer, but reasons for the racial disparities remain unclear.[4]

To date, genome-wide association studies (GWAS) of breast cancer have been predominantly conducted among women of Asian and European ancestry,[5] and most identified risk variants cannot be directly replicated among women of African ancestry. A few GWAS have been conducted in African-ancestry women, but the results were limited due to small sample size.[6,7] Several polygenic risk scores (PRSs) have been constructed using common risk variants for breast cancer,[8,9] but the performance in African-ancestry women was much worse than in European-ancestry women. A well-powered genetic study of breast cancer is in need for women of African ancestry.

The African American Breast Cancer Genetic (AABCG) consortium includes genetic data from over 18,000 cases and 22,000 controls of African ancestry from over 20 studies conducted in the U.S. and Africa. The following aims were derived to better understand the genetics of breast cancer among women of African ancestry:

Aim 1: Identify genetic risk loci for breast cancer risk through a genome-wide association study in women of African ancestry.

Aim 2: Build a polygenic risk score for risk prediction among women of African ancestry

Aim 3: Identify predisposition genes for breast cancer among women of African ancestry by a transcriptome-wide association study.

This study is the largest genetic association study of breast cancer ever conducted in women of African ancestry. The purpose of this genetic study is to identify risk variants which could be used for risk assessment to identify high-risk individuals and genes which could inform the etiology of breast cancer.

II. BACKGROUND

Racial disparity of breast cancer

Breast cancer is the most common cancer among women in the United States. There are 287,850 estimated new cases in 2022.[1] The lifetime risk of developing breast cancer is 12.9% based on statistics from 2016 to 2018.[1] Women of African ancestry have a lifetime breast cancer risk of 11.6%, which is slightly lower than the lifetime risk in women of European ancestry (13.6%).[2] However, women of African ancestry have a higher breast cancer incidence rate than European descendants before age 40 years.[10,11] Women of African ancestry are also more likely to develop triple-negative breast cancer, a more aggressive subtype, compared with European descendants.[3] From 2012 to 2016, the incidence rate of triple-negative breast cancer in women of African ancestry was about twice as high as the incidence rate in women of European ancestry.[10] In addition, women of African ancestry have a 41% higher age-adjusted mortality of breast cancer than women of European ancestry.[2] Although the disparity of breast cancer can be partially explained by socioeconomic factors,[12,13] genetic components can also contribute to the disparity.[4]

Genetic factors in breast cancer

A genetic architecture has been proposed for complex diseases like cancer, which classifies genetic risk variants into three groups by their allele frequency and effect size: rare variants with large effect sizes causing Mendelian diseases, low-frequency variants with intermediate effect, and common variants with small effect sizes.[14]

The most common cause of hereditary breast cancer is mutations in genes *BRCA1* and *BRCA2*, which were identified in 1990s.[15–17] Deleterious mutations in *BRCA1* and *BRCA2* accounts for about 5% of all breast cancers.[18] Deleterious mutations in *TP53, PTEN, PALB2, CDH1,* and *STK11* are also classified as high-penetrance mutations.[19] Over the years, more genes with moderate-penetrance mutations have been identified for risk of breast cancer, including *ATM*, *CHEK2*, *NF1*, *RAD51C*, *RAD51D* and *BARD1*. In general, moderate-penetrance mutations are associated with a two to four times elevated risk of breast cancer.[20,21]

Two recent large case-control studies evaluated the associations of these susceptibility genes with risk of breast cancer. Significant associations were found for variants in genes *BRCA1*, *BRCA2*, *PALB2*, *BARD1*, *RAD51C*, *RAD51D*, *ATM*, and *CHEK2* in both studies, and *MSH6* and *CDH1* in one of the studies (Table 1).[22–24] Significant associations were not observed for *TP53* and *PTEN* mainly due to a low frequency of mutations.

Table 1. Summary of established high- or moderate-penetrance genes for breast cancer.

Genes	**Locus**	**Gene function**
BRCA1	17q21.31	Part of a complex that repairs double-strand breaks in DNA[25]
BRCA2	13q13.1	Interacts with the recombinase RAD51 in repair of double-strand breaks[26,27]
TP53	17p13.1	A key role in many cellular pathways controlling cell proliferation, cell survival, and genomic integrity[28]
PALB2	16p12.2	A molecular scaffold for BRCA2 and BRCA1 to form a complex that repairs double-strand breaks in DNA[29]
BARD1	2q35	A protein interacts with the N-terminal region of BRCA1 as a complex[30]
RAD51C	17q22	Essential for homologous recombination repair and repair of DNA[31]
RAD51D	17q12	Essential for homologous recombination repair and repair of DNA[32,33]
ATM	11q22.3	A serine/threonine protein kinase with a central role in the repair of DNA double-strand breaks[34]
CHEK2	22q12.1	A serine-threonine kinase in cell cycle regulation including DNA repair[35]
CDH1	16q22.1	A calcium-dependent cell–cell adhesion glycoprotein for cell invasion suppression[36]

MSH6	2p16.3	Envolve in DNA mismatch repair[37]
PTEN	10q23.31	A phosphatase to negatively regulates the PI3K/Akt/mTOR cell survival signaling pathway[38]
STK11	19p13.3	An upstream activator of AMPK/PAR1-related kinases, regulates cell polarity[39]
NF1	17q11.2	Coded protein, neurofibromin, negatively regulates RAS/MAPK pathway[40]

However, these high- and moderate-penetrance mutations explain only a small fraction of cancer events due to their low prevalence in the general population. For example, carriers of deleterious mutations in the *BRCA1* gene are estimated to have a 60-70% lifetime risk of developing breast cancer,[41,42] but only 0.24% of women in the general population carry *BRCA1* pathogenetic mutations.[43] The *CHEK2* mutation 1100delC is associated with about a 2-fold increased risk of breast cancer, while its frequency in the general population is 0.71%.[44]

Since 2007, GWAS have been used to identify common genetic variants in relation to breast cancer. High-throughput genomic technologies are used to scan the entire genome and conduct the association analyses of common variants and risk of disease. To date, common variants associated with risk of breast cancer have been identified at over 200 loci.[8,45,46] In our previous study combining data from 160,500 cases and 226,196 controls of Asian and European ancestry, we identified 222 genetic risk loci associated with breast cancer at genome-wide significance ($P <5.00\times10^{-8}$) and 22 additional known risk loci at nominal significance ($P <0.05$).[47] Unlike the rare deleterious coding variants in the high- or moderate-penetrance genes, the GWAS-identified single nucleotide variants (SNVs) are common variants (typically minor allele frequency >1%), and most of them are located at non-coding regions.[8,20]

Although the risk associated with each variant is small, individuals who carry multiple risk variants can be at a considerably elevated breast cancer risk. Breast cancer risk is consistent with the polygenic susceptibility model, with more common genetic variants each having a small to moderate multiplicative effect on cancer risk.[48,49] PRSs have been constructed by combining these common risk variants to identify individuals at a high genetic risk of breast cancer. In 2019, Mavaddat et al. constructed a PRS for breast cancer using 313 variants selected by stepwise forward regression among women of European ancestry, and the PRS had an area under the receiver operating characteristic curve (AUC) of 0.630 in validation.[9] This 313-variant PRS and subsequent PRSs modified from it have been widely used in further studies and showed a good performance in women of European and Asian ancestry.[46,50,51] Recently, a PRS constructed using genome-wide variants showed an AUC of 0.635 in women of Asian ancestry.[52] In addition, a PRS using 180 variants at known risk loci showed an AUC of 0.63 in U.S. Latinas and Latin American women.[53]

Under-representation of African-ancestry populations in genetic studies

Although over 200 risk loci have been identified by GWAS for risk of breast cancer, the GWAS have been predominantly conducted among women of Asian and European ancestry.[5,54] Women of African ancestry are under-represented in epidemiologic studies. To date, the largest GWAS ever conducted among women of European ancestry included 133,384 breast cancer cases and 113,789 controls.[8] In contrast, the largest GWAS previously conducted among women of African ancestry only included 6,657 breast cancer cases and 7,713 controls.[6] In addition, the genomes among populations of African ancestry has a higher diversity and smaller linkage disequilibrium (LD) blocks compared to the genomes among populations of European and Asian ancestry.[55] The reported index variants can be correlated with the causal variants in the

European-ancestry genome but not in the African-ancestry genome. Therefore, among the reported variants at risk loci identified from women of European ancestry, less than 20% have been directly replicated in women of African ancestry.[56–58]

The PRSs built from common risk variants in women of European ancestry also had a limited performance in risk prediction in women of African ancestry. A recent study evaluated the performance of the widely used 313-variant PRS. The AUC of the PRS in women of African ancestry was 0.571,[59] much lower than the previously reported AUC of 0.630 among European-ancestry, Asian-ancestry, or Hispanic women.[9,52,53]

Therefore, a well-powered genetic study of breast cancer is in need for women of African ancestry.

III. FIRST AIM: IDENTIFY GENETIC RISK LOCI FOR BREAST CANCER RISK THROUGH A GENOME-WIDE ASSOCIATION STUDY IN WOMEN OF AFRICAN ANCESTRY.

Overview

GWAS have identified common variants at over 200 susceptibility loci for breast cancer,[45,46,51] but previous studies were predominately conducted among women of European ancestry.[5,54] Less than 20% of the reported variants can be replicated in women of African ancestry because of the different genetic architectures and the small sample size of previous studies for African-ancestry women.[56–58]

I conducted a large-scale GWAS for breast cancer among women of African ancestry. The goal of this study was to identify novel risk loci and risk variants more specific for women of African ancestry at previously known risk loci.

Methods

Study populations

In this study, I used genetic data from the African American Breast Cancer Genetic (AABCG) consortium. The AABCG is a consortium for a genetic study of breast cancer for women of African ancestry, including genetic data of samples from over 20 studies conducted in the U.S. and Africa. Detailed descriptions of participating studies are included in the Appendix 1: Description of participating studies. Briefly, the genetic data consisted of three main components: whole genome sequencing data, newly generated genotyping data using the Multi-Ethnic Genotyping Array (MEGA), and genotyping data from existing studies or consortia

(Table 2). The same participants or first-degree relatives (Pi-HAT estimate >0.45) can be genotyped by different arrays. Of them, I kept samples genotyped by an array of a higher density. In total, there were 18,044 cases and 22,187 controls of African ancestry for association analyses. Information of immunohistochemistry markers was available for most breast cancer cases, including estrogen receptor (ER)-positive cases (n =9308), ER-negative cases (n =4,927), and triple-negative breast cancer (TNBC) (n = 2,862).

Genotyping and quality control

Except for MEGA genotyping data, all other sequencing or genotyping and quality control procedures have been described previously (Appendix 1).[6,59–62] The MEGA contains more than 2 million variants before imputation, with an excellent genomic coverage of common variants across multi-racial populations. Samples were genotyped by MEGA in Vanderbilt University Medical Center, Roswell Park Comprehensive Cancer Center, and University of Southern California, and the MEGA samples were therefore categorized into three datasets based on the institution. Within each dataset, quality control (QC) procedure included: samples were excluded if they (i) were not genetically female; (ii) had a call rate <95%; (iii) had a low proportion of African ancestry (<5%) using Admixture,[63] using 1000 Genome samples as reference; (iv) had a close relationship with a Pi-HAT estimate >0.45 (one of the pair was excluded); (v) had a high heterozygosity (which indicated contaminated samples). Variants were excluded if they had (i) a call rate <95%; (ii) a P value $<10^{-6}$ in the Hardy-Weinberg equilibrium test among the controls with African-ancestry population; (iii) a consistent rate <98% across duplicated QC samples; (iv) inconsistent alleles from 1000 Genome data. After quality control, all genotyping data were imputed using the Trans-Omics for Precision Medicine (TOPMed) as reference panel. Compared with the 1000 Genomes Project Phase 3 reference panel, the TOPMed panel had a better performance of imputation for populations of African ancestry, with a 2.3-fold increase in the number of well-imputed rare variants (minor allele frequency <0.5%) and 11% improvement in average imputation quality.[64] Three datasets of MEGA genotyping samples were imputed separately, and other genotyping samples by the same array were imputed together. In total, there were two whole-genome sequencing datasets, three imputation datasets of MEGA genotyping samples, and eight imputation datasets of other genotyping samples (Table 2).

Table 2. Sample sizes of studies contributing to the genome-wide association analysis.

Dataset [a]	**Study**	**Case**	**Control**	**Case by subtype [b]**		
				ER-pos	**ER-neg**	**TNBC**
Whole genome sequencing data						
WGS_AABCG	NBHS	91	16	29	62	32
	SCCS	321	376	172	147	77
	STSBHS	421	0	175	246	171
	GBHS	293	147	112	113	69
	MEC	211	119	126	73	5
WGS-2	SCCS	71	1,639	23	0	0
Subtotal		**1,408**	**2,297**	**637**	**641**	**354**
MEGA genotyping data						
Genotyped in Vanderbilt	NBHS [c]	138	147	60	0	0
	SCCS	708	678	281	104	50
	STSBHS [c]	692	683	565	77	36
	MDABCS [c]	1,294	1,222	700	309	220
	CCPS	366	279	242	103	69
	NBCS	695	376	56	162	82
	NC-BCFR [c]	185	213	106	53	35
	NYUWHS	72	58	33	11	5
Genotyped in Roswell Park	WCHS	1,326	851	891	368	235
	BWHS	1,282	1,879	752	334	191
Genotyped in USC	MEC	1,194	914	823	264	162
Subtotal		**7,952**	**7,300**	**4,509**	**1,785**	**1,085**
Genotyping data from existing studies or consortia						
AMBER	BWHS	307	2,098	207	63	42
	CBCS	602	1	393	185	136
	WCHS	472	243	334	135	88
ROOT	SCCS	126	323	78	31	18
	NBCS	702	602	91	134	78
	CCPS	365	376	161	130	84
	RVGBC	143	254	27	25	0
	BBCS	94	102	44	44	0
	BNCS	92	227	0	0	0
AABC	NBHS	255	161	130	35	20
	NC-BCFR	383	48	226	128	40
	CARE	254	204	126	84	31
	CBCS	614	570	263	308	193
	MEC	578	888	332	144	45
	PLCO	54	112	14	6	2
	SFBCS	157	210	87	49	2
	WCHS	63	21	36	26	21
	WFBC	113	138	62	41	19

Table 2. Continued

GBHS	GBHS	660	1,496	227	225	111
BCAC OncoArray	NBHS	51	53	13	12	8
	CBCS	855	46	494	288	215
	NC-BCFR	69	0	34	25	16
	MEC	605	607	419	160	91
	PLCO	24	68	12	2	2
	2SISTER	42	0	27	14	12
	SISTER	130	163	77	24	18
	USRT	26	38	0	0	0
	WAABCS	308	292	19	65	47
BioVU	BioVU	118	2,600	0	0	0
BEST	BEST [d]	359	356	223	117	84
BCAC iCOGS	NBHS	19	42	6	1	0
	SCCS	44	251	0	0	0
Subtotal		**8,684**	**12,590**	**4,162**	**2,501**	**1,423**
Total		**18,044**	**22,187**	**9,308**	**4,927**	**2,862**

Abbreviations: 2SISTER, the Two Sister Study; AABC, the African American Breast Cancer consortium; AMBER, African American Breast Cancer Epidemiology and Risk consortium; BBCS, Baltimore Breast Cancer Study; BEST, Black Women: Etiology and Survival of Triple-negative Breast Cancers Study; BioVU, the Vanderbilt Biobank; BNCS, Barbados National Cancer Study; BWHS, Black Women's Health Study; CARE, the Los Angeles component of the Women's Contraceptive and Reproductive Experiences Study; CBCS, Carolina Breast Cancer Study; CCPS, Chicago Cancer Prone Study; ER-neg, estrogen receptor negative; ER-pos, estrogen receptor positive; GBHS, Ghana Breast Health Study; iCOGS, Collaborative Oncological Gene-environment Study; MDABCS, M.D. Anderson Breast Cancer Study; MEC, Multiethnic Cohort Study; MEGA, Multi-Ethnic Genotyping Array; NBCS, Nigerian Breast Cancer Study; NBHS, Nashville Breast Health Study; NC-BCFR, Northern California Breast Cancer Family Registry; NYUWHS, New York University Women's Health Study; OncoArray, Genetic Associations and Mechanisms in Oncology OncoArray consortium; PLCO, the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial; ROOT, the GWAS of Breast Cancer in the African Diaspora consortium; RVGBC, Racial Variability in Genotypic Determinants of Breast Cancer Risk Study; SCCS, Southern Community Cohort Study; SFBCS, San Francisco Bay Area Breast Cancer Study; SISTER, the Sister Study; STSBHS, Southern Tri-State Breast Health Study; TNBC, triple-native breast cancer; USC, University of South California; USRT, the United States Radiologic Technologists cohort; WAABCS, Women of African Ancestry Breast Cancer Study; WCHS, Women's Circle of Health Study;
WFBC, Wake Forest University Breast Cancer Study; WGS, whole genome sequencing.

a. Samples with the same sequencing platform or genotyping array were pooled as one dataset. MEGA genotyping samples were pooled by the genotyping institution; b. Studies with subtype cases less than 10 were not included in subtype analyses; c. Controls matched from SCCS; d. Controls matched from BioVU.

Variants with a minor allele frequency (MAF) greater than 0.01 and an imputation quality score (r^2) greater than 0.3 were included in the association test. Due to a small number of cases, variants with a MAF less than 0.05 in the dataset WGS-2 and iCOGS were excluded. In addition, given that the cases in BEST were matched using controls in BioVU with different genotyping arrays, I excluded variants with a MAF less than 0.05 or imputation quality score (r^2) less than 0.95 in the BEST dataset.

Statistical analysis

Logistic regression analysis was performed within each dataset to estimate a per-allele odds ratio (OR) for each variant using PLINK2.0.[65] The list of datasets has been shown in Table 2. Principal components (PCs) were estimated using the genotyped variants within each dataset. Covariates included age (<45 years, 45-54 years, 55-69 years, ≥70 years), study, and top five PCs. In some datasets, studies with imbalanced cases and controls were combined in the analyses (MEC, GBHS, versus other studies in the dataset WGS_AABCG; BWHS versus other studies in AMBER; WAABCS versus other studies in OncoArray). QQ-plots and sample size-adjusted λ_{1000} [45,66] were used to check the genomic inflation and confirm the adjustment for top five PCs. Subtype analyses for ER-positive, ER-negative, and triple-negative breast cancer were conducted with the same approach in datasets except for WGS-2, BioVU and iCOGS. Studies with a number of subtype cases less than 10 in Table 2 were excluded from the analyses. I performed fixed-effects inverse-variance-weighted meta-analyses using METAL,[67] with a genome-wide significance level of $P < 5\times10^{-8}$. Some variants were not available in all contributing studies, and they were excluded from meta-analyses if available in less than half of the total cases. Heterogeneity across datasets was assessed by the Cochran's Q statistic and I^2.

For each locus identified at genome-wide significance, I conducted conditional analyses to identify additional independent signals located flanking ± 500kb from the sentinel variant using GCTA-COJO.[68] Variants with an imputation quality score (r^2) greater than 0.3 from MEGA genotyping data from Roswell Park Comprehensive Cancer Center were used as LD reference panel (N =5,338). Since the conditional analyses were restricted to local regions of the risk loci identified at genome-wide significance, I used 1×10^{-4} as significance level for independent association signal (adjusting for around 500 comparisons in each locus). In each iteration, the variant with the lowest conditional $P <1\times10^{-4}$ was considered an independent signal at that locus, and it was subsequently adjusted, along with the sentinel variant in later iterations. This process was repeated until there were no variants with a conditional $P <1\times10^{-4}$.

Statistical Power

Given a lifetime probability of developing breast cancer of 11.6% among African-ancestry women,[2] and a genome-wide significance level of 5×10^{-8}, I calculated the power to detect effect sizes of an OR from 1.0 to 1.2, for a MAF of 0.05, 0.15, 0.25, 0.35, and 0.45 (Figure 1). There was a power greater than 0.8 to detect a risk variant with a per-allele OR higher than 1.05 and a MAF greater than 0.15 among the combined dataset. For a variant with a MAF of 0.05, there was a power greater than 0.8 to detect a per-allele OR higher than 1.10. All power calculations were determined using Quanto version 1.2.4.[69]

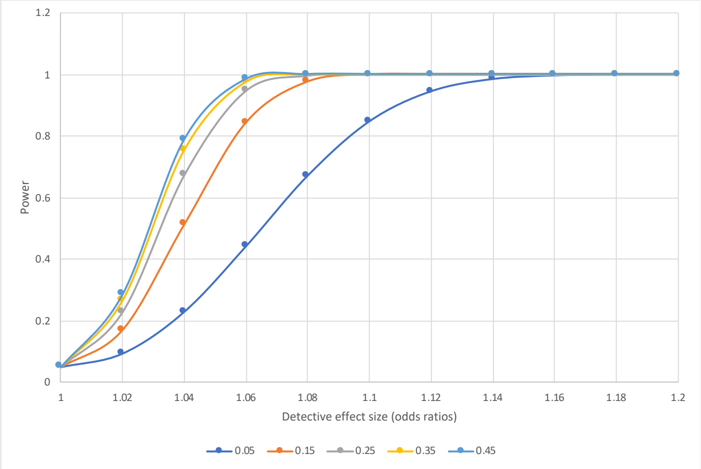

Figure 1. Power curves for genome-wide association analyses

Results

Using a fixed effects meta-analysis of GWAS from 18,044 cases and 22,187 controls of African ancestry, I identified 99 common variants at eight loci in association with overall breast cancer risk at genome-wide significance level (*P* $<5\times10^{-8}$). Sentinel variants at risk loci are shown in Table 3 and the Manhattan plot is shown as Figure 2. No obvious genomic inflation was observed (sample size-adjusted λ_{1000} ranged from 1.005 to 1.053 in the GWAS results). Although all the loci have been previously reported among women of European ancestry, sentinel variants at 4q24, 6q25.1, 14q13.3, and 18q12.1 are not in LD with the previously reported index SNVs in European-ancestry populations (Table 4). All sentinel variants showed a similar or larger effect size than the risk estimates reported in European-ancestry populations (Table 4). In particular, rs61751053 at 4q24 is a missense variant of gene *ARHGEF38*, with an OR of 1.48 (95% confidence interval [CI]: 1.30, 1.70). Evidence of heterogeneity across contributing studies was observed only for rs4784227 and rs56069439, which are both in high LD with the reported index SNVs in European-ancestry populations.

Analyses by subtype identified 55 variants at seven loci for ER-positive breast cancer, 67 variants at three loci for ER-negative breast cancer, and 85 variants at three loci for triple-negative breast cancer at genome-wide significance. Three loci for ER-positive, one locus for ER-negative, and one locus for triple-negative were not identified for overall breast cancer (Table 3). Of them, two risk loci (18q11.2 and 2q14.2 for ER-positive and ER-negative, respectively) were novel risk loci, with sentinel variants located at least 1Mb away from any of the previous GWAS-identified risk variants for breast cancer. The sentinel variant rs76664032 at the novel locus for ER-negative breast cancer was also associated with TNBC, with a higher OR of 1.30 (1.20, 1.42) and a *P* of 3.69×10^{-10}.

Of all the 12 risk loci identified at genome-wide significance, eight loci showed a significant different association by ER status (*P* <0.05 in heterogeneity test), including five loci with a stronger association with ER-positive breast cancer and three loci with a stronger association with ER-negative breast cancer (Table 5). All of the three loci with a stronger association with ER-negative breast cancer also showed a higher risk estimate in association with TNBC, with a similar or smaller *P* value.

For each locus identified at genome-wide significance, I performed conditional analyses for variants located within 500kb of the lead variant to identify potential secondary association signals. I found ten independent association signals (conditional *P* $<1.0\times10^{-4}$) at seven loci: 2q14.2, 2q35, 5p15.33, 6q25.1, 10q26.13, 14q13.3, and 16q12.1 (Table 6). Of them, the independent variants rs2736098 at 5p15.33 and rs57456888 at 16q12.1 reached genome-wide significance after adjusting for the sentinel variants. Without adjustment of the sentinel variants, rs2736098 and rs57456888 had a *P* of 1.59×10^{-7} and 8.18×10^{-8}, respectively.

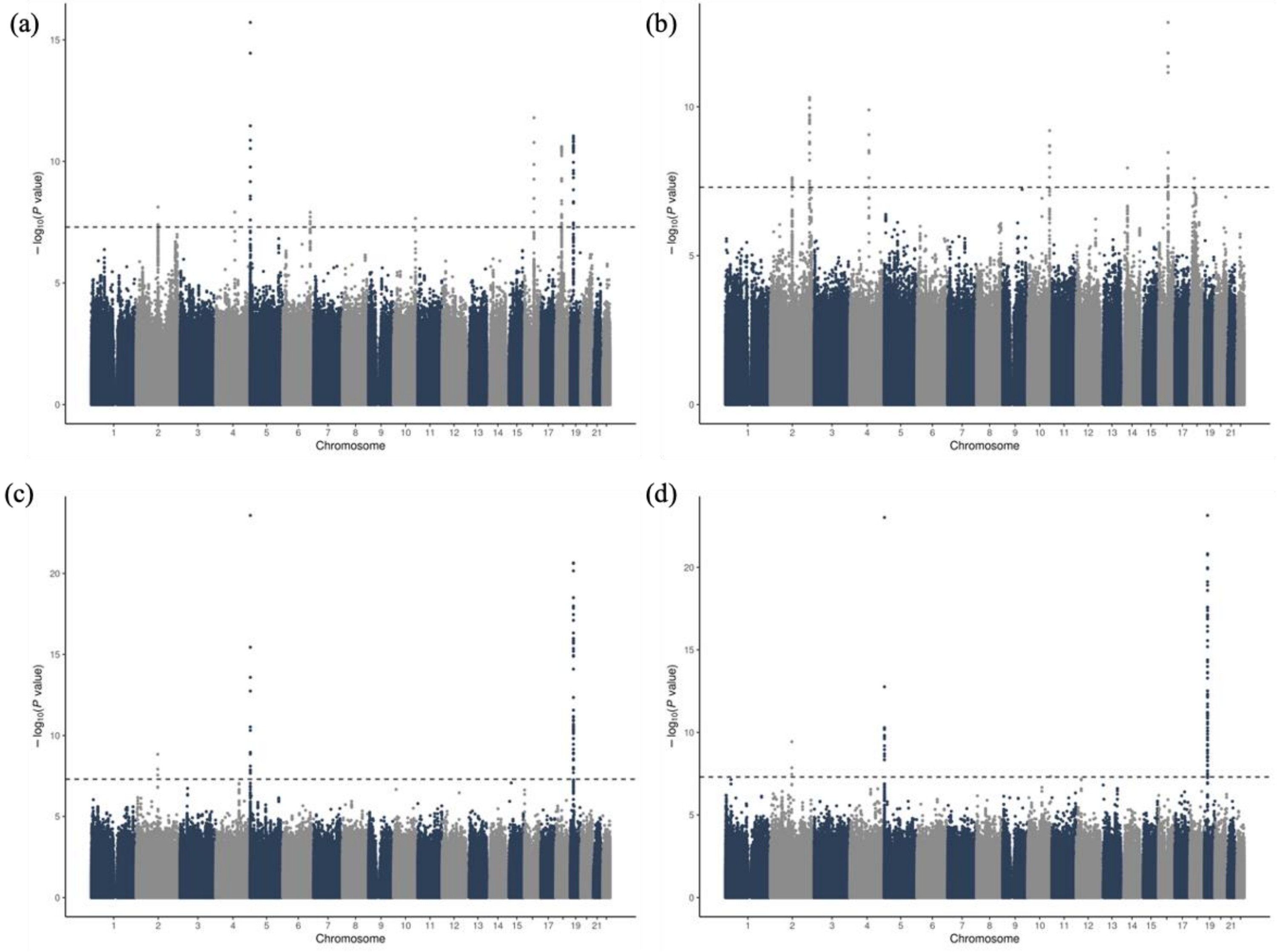

Figure 2. Genome-wide association with overall breast cancer and subtypes. (a) Overall breast cancer, (b) ER-positive breast cancer, (c) ER-negative breast cancer, (d) triple-negative breast cancer. Associations were estimated in each dataset and meta-analyzed with fixed effects model. The dashed line is genome-wide significance of 5×10^{-8}.

Table 3. Risk loci identified in women of African ancestry at genome-wide significance level, $P < 5\times10^{-8}$.

Variants	**Loci**	**Position (bd38)**	**Nearby gene**	**Gene region**	**Allele**[a]	**EAF**	**OR (95% CI)**	***P***	**I^2, %**	***P*_het**
Overall										
rs6750813	2q14.2	120501624	*INHBB*	Intergenic	C/T	0.83	1.13 (1.09, 1.18)	7.51×10^{-9}	20.3	0.24
rs61751053	4q24	105613442	*ARHGEF38*	Missense	T/C	0.01	1.48 (1.30, 1.70)	1.22×10^{-8}	0	0.64
rs10069690	5p15.33	1279675	*TERT*	Intron	T/C	0.61	1.14 (1.11, 1.18)	1.92×10^{-16}	0	0.98
rs35240111	6q25.1	151729388	*ESR1*	Intron	C/G	0.34	1.10 (1.06, 1.13)	1.25×10^{-8}	0	0.47
rs17542768	10q26.13	121578300	*FGFR2*	Intron	A/G	0.96	1.24 (1.15, 1.34)	2.22×10^{-8}	0.3	0.43
rs4784227	16q12.1	52565276	*TOX3*	Intergenic	T/C	0.07	1.25 (1.18, 1.34)	1.60×10^{-12}	51.6	0.02
rs16963205	18q12.1	32350613	*FAM59A*	Intron	T/C	0.76	1.13 (1.09, 1.17)	2.46×10^{-11}	33.8	0.11
rs56069439	19p13.11	17283116	*ANKLE1*	Intron	A/C	0.24	1.13 (1.09, 1.17)	9.07×10^{-12}	43.7	0.05
ER-positive										
rs2372943	2q35	217039053	*IGFBP5*	Intergenic	G/A	0.86	1.21 (1.14, 1.28)	4.87×10^{-11}	0	0.48
rs4575439	14q13.3	36682738	*SLC25A21*	Intron	G/A	0.59	1.12 (1.08, 1.17)	1.13×10^{-8}	0	0.82
rs10853615	18q11.2	24058545	*TTC39C*	Intron	A/C	0.20	1.15 (1.09, 1.21)	2.53×10^{-8}	17	0.29
ER-negative										
rs76664032	2q14.2	118823485	*RP11-19E11.1*	10kb from 3'	A/G	0.81	1.22 (1.15, 1.30)	1.44×10^{-9}	20.1	0.26
TNBC										
rs76664032	2q14.2	118823485	*RP11-19E11.1*	10kb from 3'	A/G	0.81	1.30 (1.20, 1.42)	3.69×10^{-10}	46.6	0.06

Abbreviations: AFR, African-ancestry populations; CI, confidence interval; EAF, effect allele frequency; ER, estrogen receptor; EUR, European-ancestry populations; OR, odds ratio; P_het, p value for heterogeneity; TNBC, triple-negative breast cancer.

a. Effect allele/other allele.

Table 4. Associations of known risk loci with breast cancer risk by population ancestry.

Loci	AFR[a]				EUR[b]				LD in EUR (r^2)	LD in AFR (r^2)
	Variants	Allele[c]	EAF	OR (95% CI)	Variants	Allele[c]	EAF	OR (95% CI)		
Overall										
2q14.2	rs6750813	C/T	0.83	1.13 (1.09, 1.18)	rs4849887	C/T	0.90	1.10 (1.08, 1.12)	<0.01	0.36
4q24	rs61751053	T/C	0.01	1.48 (1.30, 1.70)	rs9790517	T/C	0.23	1.04 (1.03, 1.06)	NA[d]	<0.01
5p15.33	rs10069690	T/C	0.61	1.14 (1.11, 1.18)	rs10069690	T/C	0.26	1.06 (1.05, 1.08)	1	1
6q25.1	rs35240111	C/G	0.34	1.10 (1.06, 1.13)	rs3757322	G/T	0.32	1.09 (1.08, 1.10)	0.02	0.01
10q26.13	rs17542768	A/G	0.96	1.24 (1.15, 1.34)	rs2981578	C/T	0.47	1.23 (1.22, 1.25)	0.15	0.51
16q12.1	rs4784227	T/C	0.07	1.25 (1.18, 1.34)	rs4784227	T/C	0.24	1.24 (1.23, 1.25)	1	1
18q12.1	rs16963205	T/C	0.76	1.13 (1.09, 1.17)	rs117618124	T/C	0.95	1.11 (1.08, 1.14)	0.82	NA[e]
19p13.11	rs56069439	A/C	0.24	1.13 (1.09, 1.17)	rs67397200	G/C	0.30	1.04 (1.02, 1.05)	0.99	0.62
ER-positive										
2q35	rs2372943	G/A	0.86	1.21 (1.14, 1.28)	rs4442975	G/T	0.50	1.14 (1.13, 1.15)	0.24	0.33
14q13.3	rs4575439	G/A	0.59	1.12 (1.08, 1.17)	rs2236007	G/A	0.79	1.07 (1.06, 1.09)	0.01	<0.01

Abbreviations: AFR, African-ancestry populations; ASN, Asian-ancestry populations; CI, confidence interval; EAF, effect allele frequency; ER, estrogen receptor; EUR, European-ancestry populations; NA, not applicable; LD, linkage disequilibrium; OR, odds ratio.

a. Risk estimates for sentinel variants identified in African-ancestry women in this aim; b. Risk estimates for index variants previously reported in European- and Asian-ancestry women; c. Effect allele/other allele; d. rs61751053 is monomorphic in European-ancestry populations; e. rs117618124 is monomorphic in African-ancestry populations

Table 5. Risk estimates at risk loci at genome-wide significance level by subtype among African-ancestry women.

Variants	**Loci**	**Allele** [a]	**EAF**	**ER-Positive**		**ER-Negative**		**TNBC**		***P* for ER heterogeneity**
				OR (95% CI)	***P***	**OR (95% CI)**	***P***	**OR (95% CI)**	***P***	
rs76664032 [b]	2q14.2	A/G	0.81	1.03 (0.98, 1.08)	0.31	1.22 (1.15, 1.30)	1.44×10^{-9}	1.30 (1.20, 1.42)	3.69×10^{-10}	2.99×10^{-5}
rs6750813	2q14.2	C/T	0.83	1.12 (1.06, 1.19)	1.58×10^{-5}	1.17 (1.10, 1.25)	3.47×10^{-6}	1.15 (1.06, 1.25)	7.82×10^{-4}	0.34
rs2372943	2q35	G/A	0.86	1.20 (1.14, 1.28)	4.87×10^{-11}	1.02 (0.95, 1.10)	0.53	1.01 (0.93, 1.10)	0.76	2.21×10^{-4}
rs61751053	4q24	T/C	0.01	1.69 (1.44, 1.98)	1.27×10^{-10}	1.34 (1.08, 1.66)	6.73×10^{-3}	1.16 (0.87, 1.54)	0.31	0.09
rs10069690	5p15.33	T/C	0.61	1.08 (1.04, 1.13)	7.58×10^{-5}	1.30 (1.23, 1.36)	2.63×10^{-24}	1.38 (1.30, 1.47)	9.70×10^{-24}	2.62×10^{-8}
rs35240111	6q25.1	C/G	0.34	1.10 (1.06, 1.15)	2.70×10^{-6}	1.09 (1.04, 1.15)	6.62×10^{-4}	1.07 (1.01, 1.14)	0.03	0.79
rs17542768	10q26.13	A/G	0.96	1.30 (1.18, 1.43)	7.14×10^{-8}	1.10 (0.98, 1.24)	0.10	1.04 (0.90, 1.21)	0.55	0.03
rs4575439	14q13.3	G/A	0.59	1.12 (1.08, 1.16)	1.13×10^{-8}	1.01 (0.96, 1.06)	0.66	0.99 (0.93, 1.05)	0.85	1.21×10^{-3}
rs4784227	16q12.1	T/C	0.07	1.32 (1.23, 1.43)	1.46×10^{-13}	1.13 (1.03, 1.25)	0.01	1.14 (1.00, 1.29)	0.05	0.01
rs10853615 [b]	18q11.2	A/C	0.20	1.15 (1.09, 1.21)	2.53×10^{-8}	0.99 (0.93, 1.06)	0.77	0.98 (0.91, 1.06)	0.67	2.92×10^{-4}
rs16963205	18q12.1	T/C	0.76	1.13 (1.08, 1.18)	1.00×10^{-7}	1.14 (1.07, 1.20)	1.24×10^{-5}	1.13 (1.05, 1.22)	6.60×10^{-4}	0.93
rs56069439	19p13.11	A/C	0.24	1.07 (1.03, 1.12)	1.92×10^{-3}	1.27 (1.21, 1.34)	3.49×10^{-18}	1.35 (1.26, 1.44)	4.22×10^{-18}	1.87×10^{-6}

Abbreviations: CI, confidence interval; EAF, effect allele frequency; ER, estrogen receptor; OR, odds ratio; TNBC, triple-negative breast cancer.
a. Effect allele/other allele; b. sentinel variants at novel risk loci.

Table 6. Independent signals identified by conditional analyses.

Variants	**Loci**	**Allele [a]**	**EAF**	**Adjusted Variant**	**OR (95% CI)**	**Conditional *P***
Overall						
rs10211615	2q14.2	G/A	0.60	rs6750813	1.07 (1.04, 1.11)	1.05×10^{-5}
rs2736098	5p15.33	C/T	0.90	rs10069690	1.16 (1.10, 1.22)	2.49×10^{-8}
rs6889886	5p15.33	C/T	0.80	rs10069690	1.10 (1.05, 1.14)	8.27×10^{-6}
rs3778610	6q25.1	A/G	0.51	rs35240111	1.07 (1.04, 1.10)	3.83×10^{-5}
rs2813569	6q25.1	A/G	0.60	rs35240111	1.07 (1.03, 1.10)	7.89×10^{-5}
rs79394706	10q26.13	G/A	0.04	rs17542768	1.19 (1.09, 1.29)	4.19×10^{-5}
rs57456888	16q12.1	G/A	0.20	rs4784227	1.13 (1.09, 1.18)	3.72×10^{-10}
ER-positive						
rs16856925	2q35	A/G	0.86	rs2372943	1.16 (1.10, 1.23)	1.78×10^{-7}
rs13008330	2q35	T/C	0.02	rs2372943	1.32 (1.16, 1.51)	3.09×10^{-5}
rs17104874	14q13.3	T/C	0.94	rs4575439	1.20 (1.10, 1.31)	2.01×10^{-5}

Abbreviations: CI, confidence interval; EAF, effect allele frequency; OR, odds ratio.
a. Effect allele/other allele;

Discussion

In this largest GWAS for breast cancer ever conducted in women of African ancestry, I identified 12 risk loci for breast cancer risk at genome-wide significance, including two novel risk loci for ER-positive and ER-negative breast cancer. Sentinel variants specific for women of African ancestry were also identified at previously known loci.

The sentinel variant rs76664032 at the novel locus for ER-negative and triple-negative breast cancer is located 10kb from the 3' of a long non-coding RNA gene *RP11-19E11.1*. Previous studies have reported that *RP11-19E11.1* is up-regulated in basal-like breast cancer and it functions as an *E2F1* target gene for cell proliferation.[70,71] This supported my finding that this risk locus showed a higher risk estimate with TNBC than ER-negative breast cancer. The rs76664032 showed a risk estimate of an OR of 1.30. Given that rs76664032 is a common variant with a MAF of 0.19, it may partially explain the high incidence of TNBC among African-ancestry women.

The sentinel variant rs10853615 at the other novel locus for ER-positive breast cancer is located at the intron of gene *TTC39C*. Although no previous studies have found association between gene *TTC39C* and breast cancer, a previous study has reported that *TTC39C* is up-regulated in cell lines with loss-of-function mutations of *STK11*, a major tumor suppressor gene in lung cancers.[72]

In this aim, ten previously known risk loci were identified at genome-wide significance. However, only five loci of them had sentinel variants as the same as or in LD with previous index variants reported in women of European ancestry. This supports that only a small proportion of index variants reported from European-ancestry populations can be directly replicated in African-ancestry populations due to the difference between their genetic architectures. The risk estimates at most risk loci showed a similar or larger effect size in African-ancestry women than those reported in European-ancestry women. A potential reason is that risk loci with higher risk estimates are more likely to be detected given that the sample size of African-ancestry GWAS is still relatively small than previous European-ancestry GWAS. Further studies are warranted to explore the underlying mechanism for these risk loci.

Among the nearby genes of the sentinel variants, *INHBB*, *ARHGEF38*, *TERT*, *ESR1*, *FGFR2*, *TOX3*, *ANKLE1*, *IGFBP5*, and *SLC25A21* have been identified as likely target genes by fine-mapping in women of European ancestry.[73] The sentinel variant rs61751053 is a missense variant of the gene *ARHGEF38*, which explains the high risk estimate of an OR of 1.48. This finding supports that *ARHGEF38* is a target gene for risk of breast cancer.

This is the largest GWAS for breast cancer ever conducted for women of African ancestry including 18,044 cases and 22,187 controls. However, the sample size was still relatively smaller compared with previous GWAS in women of European ancestry, and only ten known risk loci reached genome-wide significance. My work in the aim 2 investigated the risk estimates at all known loci in women of African ancestry.

In summary, I identified 12 risk loci for breast cancer risk at genome-wide significance among women of African ancestry, including two novel risk loci for ER-positive and ER-negative breast cancer. At known loci, sentinel variants were more specific for women of African ancestry.

IV. SECOND AIM: BUILD A POLYGENIC RISK SCORE FOR RISK PREDICTION AMONG WOMEN OF AFRICAN ANCESTRY

Overview

GWAS have identified common variants at over 200 susceptibility loci for breast cancer,[45,46,51] and common variants have been used to construct a PRS to identify individuals at a high genetic risk among the general populations.[74] In 2019, Mavaddat et al. selected 313 variants by stepwise forward regression among women of European ancestry, and constructed a PRS for breast cancer with an AUC of 0.630 in validation.[9] This 313-variant PRS and subsequent PRSs modified from it have been widely used in further studies and showed a good performance in women of European and Asian ancestry.[46,50,51] In 2020, Zhang et al. constructed a 330-variant subtype-specific PRS by adding 17 novel identified variants, and the per-standard deviation (SD) ORs were 1.83 and 1.65, with AUC of 0.661 and 0.636 for luminal A-like and triple-negative subtypes, respectively.[51]

However, among African-ancestry women, there have been no PRSs for breast cancer showing a performance of risk prediction as good as in European- or Asian-ancestry populations. The widely-used 313-variant PRS was built with selected variants and trained weights in women of European ancestry. In 2021, Du et al. evaluated the performance of the 313-variant PRS in women of African ancestry. The per-SD OR was 1.27 and the AUC was 0.571, which was much lower than the AUC of 0.630 in women of European ancestry.[59] A recent study developed a joint PRS combining the 313-variant PRS and a PRS built in a training dataset of African-ancestry women, and got an AUC of 0.581 in African-ancestry women.[75]

In this aim, I selected variants in association with breast cancer at known loci, constructed a PRS and evaluated the performance in a testing set. Ten previously known risk loci were identified at genome-wide significance in the first aim, but more known loci did not reach the genome-wide significance level, primarily due to the relatively small sample size of African-ancestry women. The purpose of this aim was to evaluate the known risk loci and select the variants in association with breast cancer among African-ancestry women but not reaching genome-wide significance. In our previous study, we identified common variants at 244 risk loci for breast cancer combining data from 160,500 cases and 226,196 controls of Asian and European ancestry.[47] Although the index variants may not be the causal variants, they are likely to be in high LD with the causal variants. Instead of exploring variant at whole genome, I focused on the risk estimates of previously reported index variants and their highly correlated variants, and therefore, reduced the multiple testing burden.

Methods

Study population

The study population in the aim 1 was divided into a training set for PRS construction and a testing set for performance evaluation. To avoid the heterogeneity across studies, all incident breast cancer cases (n =765) from Southern Community Cohort Study (SCCS) were selected into the testing set. Detailed descriptions of SCCS are included in the Appendix 1: Description of participating studies. The cases were sequenced or genotyped by Illunima HiSeq X Ten and BGISEQ-500 (n=263), Illumina HiSeq X and NovaSeq (n=38), Multi-Ethnic Genotyping Array (n=341), and Illumina HumanOmni2.5 Array (n=123). Controls were frequently matched on platform/array and stratum of year of birth (≤1935, 1936-1940, 1941-1945, 1946-1950, 1951-

1955, 1956-1960, 1961-1965, >1965). A total of 106 cases sequenced by Illunima HiSeq X Ten and BGISEQ-500 could not be matched with controls by the same sequencing platform, and therefore matched with controls sequenced by Illumina HiSeq X and NovaSeq. In total, 765 incident cases and 765 matched controls from SCCS were selected as a testing set. The study population in the aim 1 excluding the samples in the testing set was the training set.

Risk estimates at known risk loci

I acquired the list of 11,737 variants which are in high LD (r^2 >0.7) with the 244 index variants in European-ancestry samples from the 1000 genomes project phase 3 (East Asian samples used for two index variants exclusively found in Asian populations). Then I evaluated the risk estimates of the index variants and their highly correlated variants in the results of the meta-analyses in aim 1.

Variants selection for PRS

Association analyses were performed for overall breast cancer, ER-positive and ER-negative breast cancer with a similar approach as described in the aim 1 using samples in the training set. Sentinel variants at all risk loci at genome-wide significance in aim 1 were selected for PRS construction. At other known risk loci, variants were selected as the following steps: (1) I kept the index variants or highly correlated variants (r^2 >0.7) showing an association with breast cancer at a P <0.10 in the same direction as reported in women of European and Asian ancestry; (2) The index variant or a highly correlated insertion/deletion variant was selected in priority if it had a P <0.05; (3) The variant with the lowest P was selected if the index variant at the locus had a P >0.05 or greater than two orders of magnitude of the lowest P (lowest P×100); (4) If no variants had a P <0.05 at the known locus, I additionally selected the index variant or a variant in high LD (r^2 >0.9) with a P <0.10 in the same association direction as previously reported. If the selected variant was unavailable or had an imputation score (r^2) <0.8 in the testing dataset, a proxy variant would be selected.

The PRSs for ER-positive and ER-negative breast cancer were constructed using a subset of the variants selected for the PRS for overall breast cancer. For each ER-subtype, variants with associations at P >0.10 were excluded from the PRS for the breast cancer subtype.

To compare the performance of risk prediction, a PRS_{EUR} was also calculated using the reported 313 variants and reported weights from women of European ancestry. Variants were excluded if they had a MAF less than 0.01 in African-ancestry populations or with an imputation quality score (r^2) less than 0.8 in the testing set.

Statistical analysis

For each individual in the testing set, PRSs were calculated as:

$$PRS = \sum \beta_i SNV_i.$$

The weight βs used for PRS_{AFR}, $\text{PRS}_{\text{AFR\_ER+}}$, $\text{PRS}_{\text{AFR\_ER-}}$ were the beta coefficients estimated in association analyses for overall breast cancer, ER-positive and ER-negative breast cancer in the training set, respectively. The weight βs used for PRS_{EUR}, $\text{PRS}_{\text{EUR\_ER+}}$, $\text{PRS}_{\text{EUR\_ER-}}$ were the weights previously reported by Mavaddat et al.[9]

In the testing set, PRSs were categorized by percentile in controls (<20%, 20-40%, 40-60%, 60-80%, 80-90%, 90-95%, and ≥95%). Logistic regression was performed to estimate the OR per SD of continuous PRS and the OR for each PRS category (40-60% group used as reference) adjusted for the first five PCs. The AUC was estimated for continuous PRS using the Stata command *comproc* to adjust for the first five PCs. It fitted a linear regression of the PRS distribution on the five PCs derived from genotyped variants, and used standardized residuals based on this fitted linear model to estimate the AUC for cases and controls.[76]

The cumulative absolute risk of breast cancer was calculated for each category of $\mathrm{PRS}_{\mathrm{AFR}}$ using age-specific breast cancer incidence, breast cancer mortality, and all-cause mortality.[50,77,78] The age-specific breast cancer incidence and mortality rates were acquired from Surveillance, Epidemiology and End Results program (2015-2019) and age-specific all-cause mortality rates were acquired from National Center for Health Statistics, Centers for Disease Control and Prevention (2015-2019).

Results

First, I investigated the risk estimates at all known risk loci using the results of meta-analyses in the aim 1. Among the 11,737 highly correlated variants with 244 index variants (r^2>0.7 in European-ancestry populations), a total of 10,661 variants (including 212 index variants) had a MAF greater than 0.01 and were available in the meta-analyses. Among the 212 index variants, 56 variants showed an association with overall breast cancer risk in the same direction with a P <0.05, including 10 variants with a P $<1.0\times10^{-4}$. In addition, 42 index variants with a P >0.05 had a highly correlated variant showing an association in the same direction as reported in

European-ancestry populations with a *P* <0.05. In total, association in the same direction was observed at 98 known risk loci at nominal significance level (*P* <0.05). Additionally, 17 index variants or highly correlated variants (r^2 >0.9) showed an association in the same direction at a *P* >0.05 and <0.10. These 115 index variants or highly correlated variants are shown in Appendix 2: Known risk loci replicated in African-ancestry women.

Then, to avoid using samples twice, results of association analyses conducted in the training set were used to select variants to construct the PRS. Besides the 12 sentinel variants at risk loci at genome-wide significance, I selected 82 index variants or highly correlated variants (r^2>0.7) with a *P* <0.05, and 21 index variants or highly correlated variants (r^2>0.9) with a *P* >0.05 and <0.10, in association with breast cancer in the same direction as previously reported. Of them, six variants were unavailable or with an imputation score less than 0.8 in the testing set, and no variants were available as a proxy. In final, a total of 109 variants were selected to build the PRS for overall breast cancer in African-ancestry women, including 87 variants for ER-positive breast cancer and 38 variants for ER-negative breast cancer. The selected variants are shown in Appendix 3: Associations of selected variants used for risk score in African-ancestry women. Among the previously reported 313 variants by Mavaddat et al., 248 variants had a MAF greater than 0.01 with an imputation score greater than 0.8, and 236 variants were available in the testing set (Appendix 4: List of reported variants used for risk score with weights from European-ancestry women).

The PRS_{AFR} had a normal distribution in both the cases and controls of the testing samples, and the variance in cases was smaller than that in controls. The PRS_{EUR} had a skewed distribution in controls and a normal distribution in cases, and the variance in cases was smaller than that in controls (Figure 3).

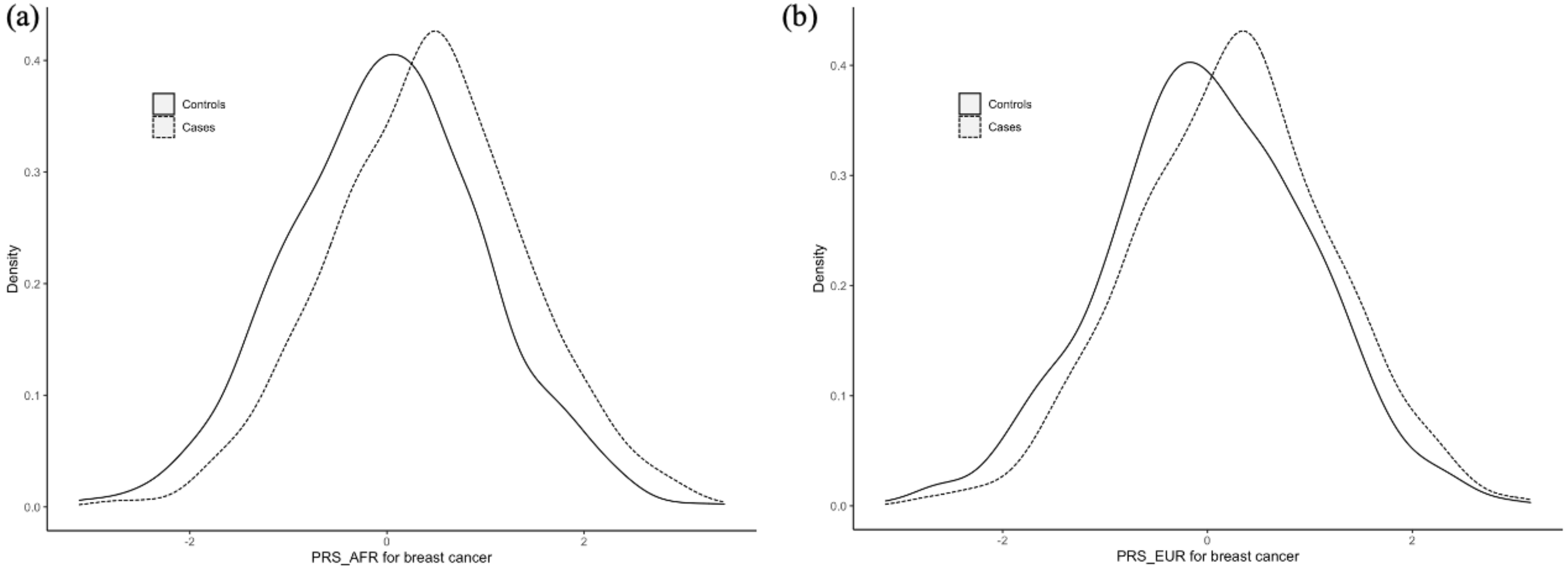

Figure 3. Distribution of PRS in testing samples. (a) The PRS_{AFR} was calculated using 109 selected variants with weights from association analyses in training samples; (b) The PRS_{EUR} was calculated using 236 reported variants with reported weight derived from European-ancestry women. Both PRSs were divided by the standard deviation in controls and centered at the mean in controls.

Table 7. Performance of polygenic risk scores (PRSs) in women of African ancestry

PRS and AUC [a]	No. of controls	PRS_{AFR}			PRS_{EUR}		
		No. of cases	**OR (95% CI)**	***P***	**No. of cases**	**OR (95% CI)**	***P***
<20%	153	79	0.65 (0.45, 0.93)	0.02	104	0.75 (0.54, 1.06)	0.10
20-40%	153	119	0.96 (0.68, 1.35)	0.81	121	0.88 (0.63, 1.22)	0.44
40-60%	153	125	1.00 (Reference)	NA	138	1.00 (Reference)	NA
60-80%	153	190	1.54 (1.12, 2.12)	0.01	210	1.52 (1.12, 2.08)	0.01
80-90%	76	99	1.67 (1.13, 2.45)	0.01	83	1.21 (0.82, 1.78)	0.33
90-95%	38	74	2.48 (1.56, 3.92)	1.10×10^{-4}	49	1.43 (0.88, 2.32)	0.15
>95%	39	79	2.57 (1.63, 4.04)	4.65×10^{-5}	60	1.71 (1.07, 2.71)	0.02
Per SD increase	765	765	1.54 (1.38, 1.71)	2.06×10^{-15}	765	1.28 (1.16, 1.42)	1.78×10^{-6}
AUC			0.620 (0.591, 0.648)			0.572 (0.542, 0.601)	

Abbreviations: AUC, area under the receiver operating characteristic curve; CI, confidence interval; NA, not applicable; OR, odds ratio; PRS, polygenic risk score; SD, standard deviation.

a. The AUC was estimated using the Stata command *comproc* to adjust for the first five PCs. It fitted a linear regression of the PRS distribution on the five PCs among controls, and used standardized residuals based on this fitted linear model to estimate the AUC for cases and controls.

Both PRSAFR and PRSEUR were associated with overall breast cancer risk, but the PRSAFR had a better performance of risk prediction than the PRSEUR (Table 7). Women in the top 5% of the PRSAFR had a 2.57-fold risk than those at average genetic risk. The PRSAFR had an OR per SD of 1.54 (1.38, 1.71), while the PRSEUR had an OR per SD of 1.28 (1.16, 1.42). The AUC of the PRSAFR was 0.620 (0.591, 0.648), higher than the AUC of 0.572 (0.542, 0.601) of the PRSEUR.

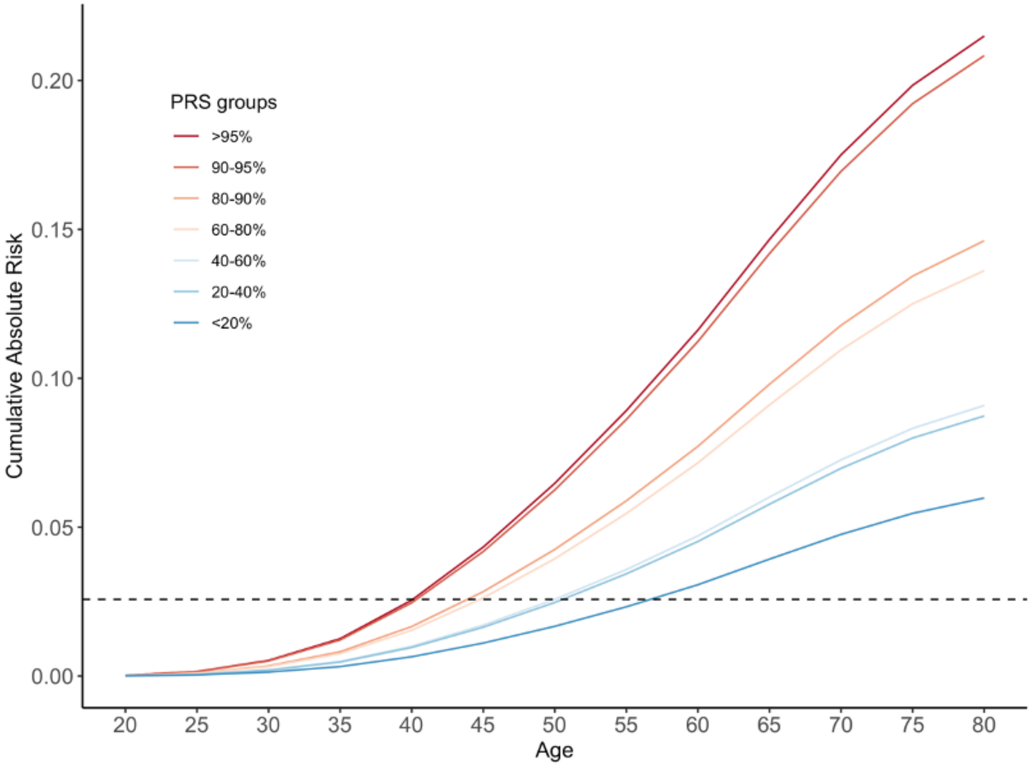

Figure 4. Cumulative absolute risk of breast cancer by polygenic risk score (PRS) category in women of African-ancestry. The horizontal line shows the cumulative risk at age of 50 years for women at average PRS (40-60%).

The cumulative absolute risk was estimated by age for each category of PRS_{AFR}. African-ancestry women in the top 10% of the PRS_{AFR} had a higher cumulative risk than women at average PRS. At age of 40 years, women in the top 10% of the PRS had the cumulative risk level which was reached by age of 50 years among women at average PRS (Figure 4).

For ER-positive breast cancer, the PRS_{AFR_ER+} showed an OR per SD of 1.47 (1.28, 1.768) with an AUC of 0.608 (0.571, 0.645), better than the performance of PRS_{EUR_ER+} (Table 8). Similar pattern was also observed for ER-negative breast cancer. The estimated ORs and AUCs for ER-negative breast cancer had a wider confidence interval than those for ER-positive breast cancer because of a smaller sample size of the testing set.

Table 8. Association of subtype-specific polygenic risk scores (PRSs) in women of African ancestry.

	ER-positive		**ER-negative**	
	PRS_{AFR_ER+}	**PRS_{EUR_ER+}**	**PRS_{AFR_ER-}**	**PRS_{EUR_ER-}**
Per SD OR (95% CI)	1.47 (1.28, 1.68)	1.37 (1.20, 1.57)	1.71 (1.40, 2.09)	1.38 (1.13, 1.69)
AUC	0.608 (0.571, 0.645)	0.589 (0.552, 0.626)	0.641 (0.591, 0.692)	0.588 (0.530, 0.645)

Abbreviation: AUC, area under the receiver operating characteristic curve; CI, confidence interval; ER, estrogen receptor; OR, odds ratio; PRS, polygenic risk score; SD, standard deviation.

Discussion

In this aim, I constructed a PRS for women of African ancestry with an AUC of 0.620 for risk prediction. This is the first study building a PRS for women of African ancestry with a performance as good as previous PRSs in European- or Asian-ancestry populations.

Compared with PRS_{EUR}, the PRS_{AFR} had a better performance for risk prediction for women of African ancestry. The 313 variants in the PRS_{EUR} were selected with weights trained in women of European ancestry. Due to the difference in genetic architecture, some variants were not in LD with the causal variants or had a low MAF in African-ancestry women, so they cannot present the signals at the risk loci.

Current screening guideline for breast cancer by U.S. Preventive Services Task Force (USPSTF) recommends screening mammography for women aged 50 to 74 years with average risk.[79] In this study, women in the top 10% of the PRS aged 40 years had the same cumulative risk of breast cancer as women aged 50 years with an average PRS. My findings support that PRS can be used to identify individuals at a high genetic risk to start screening at an early age.

There might be some limitations in this study. The testing samples were sequenced or genotyped using different platforms or arrays. To avoid confounding from the difference of genotyping arrays, variants were excluded from analyses if they had a MAF lower than 0.01, an imputation quality score lower than 0.8 in genotyping samples, or unavailable in sequencing samples. A sensitivity analysis was additionally performed adjusted for the genotyping platform, and the platform was not associated with the outcome. Using stringent inclusion criteria, 236 variants among the 313 reported variants were included to build the PRS_{EUR}. A previous study evaluated the 313-variant PRS in women of African-ancestry, including 311 variants with a MAF greater than 0.001 and imputation score higher than 0.3.[59] They reported an AUC of 0.571, which was close to the AUC of 0.572 estimated in this study. This supports that the performance of PRS_{EUR} was not underestimated due to fewer included variants in this aim.

V. THIRD AIM: IDENTIFY PREDISPOSITION GENES FOR BREAST CANCER AMONG WOMEN OF AFRICAN ANCESTRY BY A TRANSCRIPTOME-WIDE ASSOCIATION STUDY

Overview

Most risk variants identified by GWAS are located in non-coding regions of the genome, and their effects are likely to be involved in the gene regulation.[80] An expression quantitative trait loci **(**eQTL) refers to a genomic locus that explains a fraction of the genetic variance of a gene expression. The level of gene expression is usually measured by the transcript abundance. In an eQTL analysis, association tests are performed between the sentinel variants at susceptibility loci and local or distant genes. Conventionally, *cis*-eQTL analysis focuses on genes residing ±500 kb flanking the sentinel variant, and *trans*-eQTL focuses on genes located further downstream or upstream or on a different chromosome. The *cis*-eQTL analysis is commonly used to identify potential target genes of the risk variants.

While eQTL analyses examine the association with genetic risk variants and expression levels of potential target genes, it is also crucial to explore the association between gene expression levels and the trait. However, the sample size is usually limited for study populations with measurement of gene expression levels, due to the specimen availability and cost. In 2015, transcriptome-wide association study (TWAS) was developed to systematically investigate the association of gene expression with disease risk.[81,82] First, it uses genetic variants to construct prediction models for gene expression levels using samples with both genotype data and gene expression data. Then it applies the models to a large-scale GWAS population, and performs association tests between the genetically predicted gene expression and the risk of disease. The individual-level GWAS data are not necessary in this step. It only requires the summary association statistics from the large-scale GWAS.[82]

Different from GWAS, TWAS is a gene-based analysis. Some genetic variants may have a small effect size which is difficult to detect in an individual variant-based GWAS. In TWAS, the effects of multiple risk variants at the same locus can be aggregated into a single test at the gene level, increasing the study power to identify the risk locus.[83] A previous TWAS for breast cancer conducted among 122,977 cases and 105,974 controls of European ancestry found significant association with breast cancer for 48 genes of 8,597 genes evaluated, and 14 genes were located at loci which had not been previously reported for breast cancer.[83] In our previous study, we conducted a TWAS using GWAS data from 160,500 cases and 226,196 controls of European and Asian ancestry, and identified 137 genes in association with risk of breast cancer, including 14 genes at 13 loci are located at least 1Mb away from any of the previous GWAS-identified risk variants for breast cancer.[47]

Due to the lack of large-scale GWAS, no TWAS has been conducted among women of African ancestry yet. As a gene-based analysis, I expected the TWAS to identify additional risk loci for breast cancer in women of African ancestry. I also examined the susceptibility genes which were previously identified among women of European and Asian ancestry.

Methods

Study population

The same study population were used as the aim 1. Results of the meta-analyses of GWAS for overall, ER-positive, ER-negative, and triple-negative breast cancer in the aim 1 were used in the association analyses for TWAS.

Gene expression profiling and data processing

Transcriptome and high-density genotyping data of breast tumor tissue from SCCS (n =237) and the Cancer Genome Atlas (TCGA) (n =163) were used to build gene expression prediction models. SCCS samples were formalin-fixed paraffin-embedded (FFPE) breast tumor tissue samples, and TCGA samples were fresh frozen breast tumor tissue samples. Details of the profiling and mapping of the RNA sequencing data in SCCS samples have been described previously.[84] The HTSeq-FPKM data were downloaded for TCGA breast cancer samples using R package *TCGAbiolinks*,[85] and clinical information was acquired from the TCGA Pan-Cancer Clinical Data Resource.[86] The RNA expression levels of SCCS and TCGA samples were annotated using the GENCODE v22.[87]

The FPKM data for SCCS and TCGA samples were processed by excluding genes expressed in less than half of samples (median FPKM =0), log2 transformation, quantile normalization for each sample, rank-based inverse normalization for each gene, adjusted for probabilistic estimation of expression residuals (PEER) factors[88] (n =30), age, TNM stage, first five principal components, and copy number alternation (only for TCGA samples). TCGA samples were excluded if they had a missing value for age (n =1), unknown TNM stage (n =3), or unavailable data for copy number variation (n =2). One SCCS sample was also excluded with a missing value of age. A total of 43 SCCS samples had unknown TNM stage information, and they were categorized as a separate level of "unknown TNM stage". In total, 236 SCCS samples and 157 TCGA samples were used to build gene prediction models. In this study, protein-coding genes (16,001 in SCCS, 16,810 in TCGA, and 15,760 shared in both studies) and long intergenic non-coding RNA (lincRNA) (2,400 in SCCS, 2,950 in TCGA, and 1,977 shared in both studies) were included for analyses. Pseudogenes were not included due to potential concern of inaccurate calling.[89]

Genotyping and imputation

In this study, tumor samples collected from women of African ancestry were identified using the germline genotype data by projecting samples on the first two major principal components of four 1000 Genome populations (CEU, YRI, CHB, and JPT).[90] TCGA samples were genotyped by Affymetrix Genome-Wide Human SNP Array 6.0. SCCS samples were genotyped by three platform or arrays: Illumina HumanOmni2.5-Quad (n =90), Multi-Ethnic Genotyping Array chip (n=70), and Illunima HiSeq X Ten and BGISEQ-500 (n=77). Details of genotyping and quality control have been described in the aim 1. Genotyped variants shared across different platforms or arrays were kept. After quality control, TCGA and SCCS samples were imputed separately using TopMed as reference panel. Variants with a MAF greater than 0.05 and an imputation quality score (r^2) greater than 0.8 were included for model building. In final, there were 8,053,287 variants in SCCS and 8,083,361 variants in TCGA, with 7,477,713 shared variants.

Expression quantitative loci (eQTL) analysis

I performed *cis*-eQTL analyses to identify target genes of the sentinel variants at 12 loci identified in the aim 1 using gene expression data in SCCS and TCGA samples separately. A linear regression model was used to check associations between sentinel variants and genes located less than 500kb away from the sentinel variants.

Colocalization analysis

Colocalization tests were conducted between the variant-disease associations from GWAS meta-analyses and the variant-gene associations from eQTL analyses. At each locus, *cis*-eQTL analysis was performed for variants within 10kb from the sentinel variant in association with genes located within 500kb from the sentinel variant. Then the eQTL results were integrated with the GWAS meta-analyses results from the aim 1 using summary data-based Mendelian randomization (SMR) for a colocalization test.[91] For each significant association by SMR tests (P <0.05), a test on heterogeneity in dependent instruments (HEIDI test)[91] was conducted to check whether the gene expression and trait had independent causal variants instead of a shared causal variant.

Prediction model building

Gene expression prediction models were built using the elastic net method as implemented in the *glmnet* R package, with α =0.5, within SCCS and TCGA samples separately. For each gene, the prediction model was built selecting genetic variants located ±500kb from the gene boundary. Five-fold cross-validation was used to validate the models internally. Genes with a model prediction R^2 greater than 0.01 either in SCCS or in TCGA were included for association analyses.

Association analyses of predicted gene expression with breast cancer risk

Association analyses were conducted between predicted gene expression and breast cancer risk with S-PrediXcan tool,[92] using the GWAS summary statistics in the aim 1 for overall, ER-positive, ER-negative, and triple-negative breast cancer. The formula used was

$$Z_g \approx \sum_{l \in Model_g} w_{lg} \frac{\widehat{\sigma_l}}{\widehat{\sigma_g}} \frac{\widehat{\beta_l}}{se(\widehat{\beta_l})}$$

The Z-score was used to estimate the association between predicted gene expression and breast cancer risk. In this formula, w_{lg} is the weight of variant l for predicting the expression of gene g. $\hat{\beta}_l$and $se(\hat{\beta}_l)$ are the association beta coefficient and its standard error for variant l in GWAS, and $\hat{\sigma}_l$ and $\hat{\sigma}_g$ are the estimated variances of variant l and the predicted expression of gene g, respectively. In this study, I estimated the correlations between variants included in the prediction models.

Association analyses were performed for all genes with prediction models built from SCCS or TCGA samples with a R^2 greater than 0.01. Results of the genes available in both models were combined using the aggregated Cauchy association test (ACAT).[93] Bonferroni-corrected *P* values were calculated for multiple testing. In addition to the transcriptome-wide association analyses, I examined the association for 208 protein-coding genes and 18 lincRNA genes which were reported as target genes in previous TWAS.[47,83,94–96]

Results

By eQTL analyses, I identified two potential target genes in association with the sentinel variants in the same direction with a P <0.05 in both SCCS and TCGA (Table 9). Both sentinel variants were located at previously known risk loci. The association for gene *EPB41L5* at 2q14.2 was confirmed by colocalization analyses in both SCCS and TCGA, but the association for gene *ANKLE1* was only confirmed in SCCS.

Then I built gene prediction models using SCCS and TCGA samples and conducted transcriptome-wide association analyses. A total of 2,961 and 4,392 genes had prediction models with a R^2 greater than 0.01 using SCCS and TCGA samples, respectively. Among them, 846 genes were shared in SCCS and TCGA models. The Pearson correlation was 0.53 between the prediction performance (R^2) for these 846 genes in SCCS and TCGA. In total, 6,507 genes were evaluated in the association analyses. Among them, no genes showed an association with breast cancer risk at the Bonferroni-correction level of P <7.68×10^{-6}.

Of the 226 previously TWAS-reported protein-coding or lincRNA genes, 94 genes had prediction models with a R^2 greater than 0.01 and were therefore evaluated in the association analyses. Only one gene, *MAN2C1*, reached the Bonferroni-correction level of 5.32×10^{-4} (0.05/94). The expression level of *MAN2C1* was inversely associated with the risk of overall breast cancer, with a P of 5.26×10^{-4}. Nine more genes showed a nominal significant association with breast cancer risk in the same direction as previously reported (P <0.05), including two genes identified by association analyses for breast cancer subtypes (Table 10).

Among the 6,507 genes evaluated in the association analyses, 828 genes were located at GWAS-identified risk loci. Of them, 61 genes showed an association with breast cancer risk at a *P* <0.05. Only one gene, *GAREM1*, reached the Bonferroni-correction level of 6.04×10^{-5} (0.05/828). The expression level of *GAREM1* at 18q12.1 showed an inverse association with breast cancer risk with a *P* of 3.45×10^{-5}.

Table 9. Target genes identified by expression quantitative loci analyses.

Variants	Loci	Gene	Alleles [a]	SCCS				TCGA			
				EAF	Beta	SE	*P*	EAF	Beta	SE	*P*
rs6750813	2q14.2	*EPB41L5*	T/C	0.18	0.18	0.06	0.01	0.15	0.38	0.15	0.01
rs56069439	19p13.11	*ANKLE1*	C/A	0.76	-0.18	0.08	0.02	0.80	-0.35	0.12	0.005

Abbreviations: EAF, effect allele frequency; SCCS, Southern Community Cohort Study; SE, standard error; TCGA, the Cancer Genome Atlas.

a. Effect allele/other allele;

Table 10. Previously reported target genes in association with breast cancer risk, *P* <0.05.

Loci	Gene [a]	Model	Z score	*P*	R^2 [b]
Overall					
3p21.31	*GMPPB*	SCCS	2.27	0.02	0.01
11p15.5	*TNNT3*	TCGA	2.07	0.04	0.01
11q13.1	*SNX32*	TCGA	2.13	0.03	0.08
12p11.22	*CCDC91*	SCCS	2.34	0.02	0.02
14q32.33	*SIVA1*	SCCS	-2.11	0.03	0.05
15q24.2	*MAN2C1*	SCCS	-3.47	5.26E-04	0.09
16q12.2	*CES1*	TCGA	2.05	0.04	0.01
20q11.22	*CPNE1*	TCGA	-2.12	0.03	0.14
ER-positive					
1p36.13	*KLHDC7A*	SCCS/TCGA	-2.45/-1.05	0.03	0.09/0.05
TNBC					
2p23.3	*CENPO*	TCGA	-2.32	0.02	0.08

Abbreviations: SCCS, Southern Community Cohort Study; TCGA, the Cancer Genome Atlas.

a. All genes are protein coding genes; b. Performance of gene prediction models.

Discussion

In this aim, a TWAS was conducted among African-ancestry women. Although no genes reached Bonferroni-corrected significance in the transcriptome-wide analyses, my findings provided supports for some previously TWAS-reported genes and genes at GWAS-identified risk loci.

Among previously TWAS-identified genes, ten genes showed a nominal significant association in the same direction as previously reported, including gene *MAN2C1* at Bonferroni-corrected significance. Eight genes were located at GWAS-identified risk loci (*TNNT3*, *SNX32*, *CCDC91*, *SIVA1*, *MAN2C1*, *CPNE1*, *KLHDC7A*, and *CENPO*). Genes *CCDC91* and *KLHDC7A* have been reported as likely target genes by fine-mapping.[73] I found that the expression of *MAN2C1* gene was associated with a reduced risk of breast cancer in African-ancestry women, as reported by several previous TWAS for European- or Asian-ancestry women.[47,83,96] This finding is supported by a previous study showing that the protein encoded by *MAN2C1* binds PTEN and inhibits its phosphatase activity.[97]

Among genes located at GWAS-identified risk loci, I found 61 genes showed an association with breast cancer risk at a P <0.05. Gene *GAREM1* at 18q12.1 showed an inverse association with breast cancer risk with a P of 3.45×10^{-5}. The *GAREM1* encodes an adaptor protein which contributes to the epidermal growth factor (EGF) receptor-mediated signaling pathway, and it has been shown to contribute to cellular transformation.[98]

Our eQTL analyses identified potential target genes *EPB41L5* at 2q14.2 and *ANKLE1* at 19p13.11. The *EPB41L5* has been reported to be expressed at high levels in malignant breast cancer cells and involved with tumor invasion and metastasis.[99] The *ANKLE1* has been identified as the target gene in a previous eQTL study using adjacent normal breast tissue samples from European-ancestry women.[100]

In this aim, no genes were identified by TWAS in association with breast cancer at the Bonferroni-corrected threshold. This was mainly because of the relatively small sample size of the African-ancestry GWAS compared with previous European-ancestry GWAS. As I showed in the aim 1, the African-ancestry GWAS only identified ten risk loci at genome-wide significance among the 244 risk loci previously identified by European-ancestry GWAS. In addition, the gene prediction models were built using breast tumor samples due to limited available normal breast tissue samples collected from African-ancestry women. The performance of prediction models could also be limited because of the difference in gene expression between FFPE samples from SCCS and fresh frozen samples from TCGA.

VI. CONCLUSION AND FUTURE DIRECTIONS

Summary

In this study, I performed a large-scale GWAS for breast cancer among women of African ancestry, identified novel risk loci and evaluated risk estimates at all previously known loci. A PRS was constructed to identify African-ancestry women at a high genetic risk of breast cancer. Incorporating gene prediction models built from transcriptome data, a TWAS was performed to identify target genes for breast cancer in African-ancestry women.

In the first aim, I conducted GWAS for breast cancer among 18,044 cases and 22,187 controls of African ancestry. Samples were pooled into 13 datasets based on the genotyping platform and institution. GWAS was performed within each dataset, adjusted for age, study, and top five PCs. QQ-plots were checked and no observe genomic inflation was observed. Then a fixed effects meta-analysis was conducted to combine GWAS results from all datasets. I identified 12 risk loci for breast cancer risk at genome-wide significance, including two novel risk loci for ER-positive and ER-negative breast cancer. In addition, ten independent risk variants at seven risk loci were identified by conditional analyses.

Previous GWAS for breast cancer, predominately conducted among European-ancestry women, have identified common variants associated with risk of breast cancer at over 200 loci. Of them, only ten risk loci reached the genome-wide significance among African-ancestry women in the aim 1. In the second aim, I investigated the risk estimates for African-ancestry women at all previously reported risk loci for breast cancer. At each locus, I checked the previously reported index variant and variants in high LD in European-ancestry populations. Among the index variants, 56 variants showed an association with overall breast cancer risk in the same direction with a *P* <0.05, and 42 additional index variants had a highly correlated variant showing an association in the same direction with a *P* <0.05. In total, association in the same direction was observed at 98 known risk loci at nominal significance level (*P* <0.05). In addition, 17 more risk loci showed an association in the same direction at *P* <0.10, the nominal significance level for the one-side test.

Then I divided the study population into a training set to build a PRS and a testing set to evaluate the performance. Besides the 12 sentinel variants at risk loci at genome-wide significance in the aim 1, 102 variants were selected at known loci based on the results from association analyses conducted in the training set. Excluding five variants not available in the testing set, a PRS was constructed using 109 common variants. In the testing set, the PRS showed a dramatically improved performance in risk prediction compared with the reported 313-variant PRS for European-ancestry women.

In the third aim, I performed eQTL analyses and identified two potential target genes. I also built gene prediction models using gene expression data from SCCS and TCGA, and conducted a TWAS using the GWAS results in the aim 1. No genes were identified at the Bonferroni-corrected significance, mainly due to the relatively small sample size of the African-ancestry GWAS compared with previous European-ancestry GWAS. Among the genes reported by previous TWAS, ten genes were identified with a nominal significant association in the same direction as previously reported, including gene *MAN2C1* reaching the Bonferroni-corrected significance.

Implications

In this study, I performed the largest GWAS for breast cancer ever conducted among women of African ancestry, combining data from 18,044 cases and 22,187 controls of African ancestry. Two novel risk loci were identified for breast cancer in African-ancestry women, including one risk locus in association with ER-negative and triple-negative breast cancer. The sentinel variant at this locus had an OR of 1.30 and an MAF of 0.19, which may partially explain the high incidence of triple-negative breast cancer in African-ancestry women. Sentinel variants at multiple risk loci were not in LD with previously reported index variants, which indicates that I identified risk variants more specific for African-ancestry women. As the largest study to systematically evaluate associations of common variants across African-ancestry genome with breast cancer risk, the GWAS summary results derived from this study can benefit future studies on breast cancer among women of African ancestry.

The PRS constructed in this study had a good performance of risk prediction with an AUC of 0.620. This is the first time that a PRS built for African-ancestry women shows a performance as good as previous PRSs for European- or Asian-ancestry women. The cumulative absolute risk was calculated by age for each PRS category. Women in the top decile of the PRS aged 40 years reached the same cumulative risk level as women aged 50 years with an average PRS. Given that the current screening guideline by USPSTF recommends screening mammography for women aged 50 to 74 years with average risk, the PRS can be used to identify high-risk individuals to start screening earlier.

In the TWAS, although no genes were identified at Bonferroni-correction for transcriptome-wide analysis, I identified ten genes associated with breast cancer in the same direction as previously reported in European- or Asian-ancestry women. Future studies are warranted for these ten genes to explore underlying functions for breast cancer in African-ancestry women.

Future directions

In a future study, I will perform multi-ancestry meta-analyses and fine-mapping analyses combining the GWAS results from women of African, Asian, and European ancestry. The multi-ancestry meta-analyses will have a larger power to detected the risk loci shared across ancestry because of the increased sample size. The difference in genetic architecture also helps to identify the causal variants. Several previous studies have conducted multi-ancestry meta-analyses, combining Asian and European ancestry or African and European ancestry, and identified multiple novel risk loci for breast cancer.[46,47,62] I expect that the multi-ancestry meta-analysis for breast cancer will identified novel risk loci and reduce the size of credible sets for causal variants.

In the GWAS conducted in this study, I adjusted principal components for global ancestry. Compared with populations of European or Asian ancestry, populations of African ancestry have a higher degree of admixture. Future studies are warranted to take account of local ancestry in GWAS. It may increase the power and identify additional risk loci in African-ancestry women.

In this study, the PRS was constructed using variants selected from reported index variants and high correlated variants. In future studies, PRSs can be built by selecting variants from the whole genome using methods like LD-informed pruning and P value thresholding (PT),[101] LDpred2,[102] and PRS-CS.[103] GWAS results from women of European and Asian ancestry can be also integrated in PRS construction. For example, a recent published method, PRS-CSx, is a Bayesian polygenic prediction method that integrates GWAS summary statistics from populations of different ancestries.[104] Functional annotation can also be integrated to PRS construction using method like PolyPred+.[105] This may further improve the performance of the PRS for women of African ancestry.

A future TWAS with an increased power will be conducted using results from meta-analyses combining the GWAS from women of African, Asian, and European ancestry. Furthermore, prediction models will be constructed using newly generated RNA-sequencing data from normal breast tissue collected from African-ancestry women. The prediction performance can be better than the models built from breast tumor tissues in this study. I expect more target genes to be identified for African-ancestry women in the future TWAS.

VII. REFERENCES

1. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians*. 2022;72(1):7-33. doi:10.3322/caac.21708

2. Giaquinto AN, Miller KD, Tossas KY, Winn RA, Jemal A, Siegel RL. Cancer statistics for African American/Black People 2022. *CA: A Cancer Journal for Clinicians*. n/a(n/a). doi:10.3322/caac.21718

3. Iqbal J, Ginsburg O, Rochon PA, Sun P, Narod SA. Differences in Breast Cancer Stage at Diagnosis and Cancer-Specific Survival by Race and Ethnicity in the United States. *JAMA*. 2015;313(2):165-173. doi:10.1001/jama.2014.17322

4. Newman LA, Kaljee LM. Health Disparities and Triple-Negative Breast Cancer in African American Women: A Review. *JAMA Surg*. 2017;152(5):485-493. doi:10.1001/jamasurg.2017.0005

5. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019;51(4):584-591. doi:10.1038/s41588-019-0379-x

6. Huo D, Feng Y, Haddad S, et al. Genome-wide association studies in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer. *Hum Mol Genet*. 2016;25(21):4835-4846. doi:10.1093/hmg/ddw305

7. Chen F, Chen GK, Stram DO, et al. A genome-wide association study of breast cancer in women of African ancestry. *Hum Genet*. 2013;132(1):39-48. doi:10.1007/s00439-012-1214-y

8. Zhang H, Ahearn TU, Lecarpentier J, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat Genet*. 2020;52(6):572-581. doi:10.1038/s41588-020-0609-2

9. Mavaddat N, Michailidou K, Dennis J, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet*. 2019;104(1):21-34. doi:10.1016/j.ajhg.2018.11.002

10. DeSantis CE, Ma J, Gaudet MM, et al. Breast cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*. 2019;69(6):438-451. doi:10.3322/caac.21583

11. Brinton LA, Sherman ME, Carreon JD, Anderson WF. Recent Trends in Breast Cancer Among Younger Women in the United States. *JNCI: Journal of the National Cancer Institute*. 2008;100(22):1643-1648. doi:10.1093/jnci/djn344

12. Cho B, Han Y, Lian M, et al. Evaluation of Racial/Ethnic Differences in Treatment and Mortality Among Women With Triple-Negative Breast Cancer. *JAMA Oncology*. 2021;7(7):1016-1023. doi:10.1001/jamaoncol.2021.1254

13. Yedjou CG, Sims JN, Miele L, et al. Health and Racial Disparity in Breast Cancer. In: Ahmad A, ed. *Breast Cancer Metastasis and Drug Resistance: Challenges and Progress*. Advances in Experimental Medicine and Biology. Springer International Publishing; 2019:31-49. doi:10.1007/978-3-030-20301-6_3

14. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-753. doi:10.1038/nature08494

15. Hall JM, Lee MK, Newman B, et al. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*. 1990;250(4988):1684-1689. doi:10.1126/science.2270482

16. Wooster R, Bignell G, Lancaster J, et al. Identification of the breast cancer susceptibility gene BRCA2. *Nature*. 1995;378(6559):789-792. doi:10.1038/378789a0

17. Wooster R, Neuhausen SL, Mangion J, et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science*. 1994;265(5181):2088-2090. doi:10.1126/science.8091231

18. van der Groep P, van der Wall E, van Diest PJ. Pathology of hereditary breast cancer. *Cell Oncol (Dordr)*. 2011;34(2):71-88. doi:10.1007/s13402-011-0010-3

19. Shiovitz S, Korde LA. Genetics of breast cancer: a topic in evolution. *Annals of Oncology*. 2015;26(7):1291-1299. doi:10.1093/annonc/mdv022

20. Easton DF, Pharoah PDP, Antoniou AC, et al. Gene-Panel Sequencing and the Prediction of Breast-Cancer Risk. *New England Journal of Medicine*. 2015;372(23):2243-2257. doi:10.1056/NEJMsr1501341

21. Nielsen FC, van Overeem Hansen T, Sørensen CS. Hereditary breast and ovarian cancer: new genes in confined pathways. *Nature Reviews Cancer*. 2016;16(9):599-612. doi:10.1038/nrc.2016.72

22. Hu C, Hart SN, Gnanaolivu R, et al. A Population-Based Study of Genes Previously Implicated in Breast Cancer. *N Engl J Med*. 2021;384(5):440-451. doi:10.1056/NEJMoa2005936

23. Breast Cancer Association Consortium, Dorling L, Carvalho S, et al. Breast Cancer Risk Genes - Association Analysis in More than 113,000 Women. *N Engl J Med*. 2021;384(5):428-439. doi:10.1056/NEJMoa1913948

24. Narod SA. Which Genes for Hereditary Breast Cancer? *N Engl J Med*. 2021;384(5):471-473. doi:10.1056/NEJMe2035083

25. Foulkes WD, Shuen AY. In brief: BRCA1 and BRCA2. *J Pathol*. 2013;230(4):347-349. doi:10.1002/path.4205

26. Jensen RB, Carreira A, Kowalczykowski SC. Purified human BRCA2 stimulates RAD51-mediated recombination. *Nature*. 2010;467(7316):678-683. doi:10.1038/nature09399

27. Wang CX, Jimenez-Sainz J, Jensen RB, Mazin AV. The Post-Synaptic Function of Brca2. *Sci Rep*. 2019;9(1):4554. doi:10.1038/s41598-019-41054-y

28. Olivier M, Langerød A, Carrieri P, et al. The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer. *Clin Cancer Res*. 2006;12(4):1157-1167. doi:10.1158/1078-0432.CCR-05-1029

29. Buisson R, Masson JY. PALB2 self-interaction controls homologous recombination. *Nucleic Acids Res*. 2012;40(20):10312-10323. doi:10.1093/nar/gks807

30. Tarsounas M, Sung P. The antitumorigenic roles of BRCA1–BARD1 in DNA repair and replication. *Nat Rev Mol Cell Biol*. 2020;21(5):284-299. doi:10.1038/s41580-020-0218-z

31. Meindl A, Hellebrand H, Wiek C, et al. Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. *Nat Genet*. 2010;42(5):410-414. doi:10.1038/ng.569

32. Loveday C, Turnbull C, Ruark E, et al. Germline RAD51C mutations confer susceptibility to ovarian cancer. *Nat Genet*. 2012;44(5):475-476. doi:10.1038/ng.2224

33. Loveday C, Turnbull C, Ramsay E, et al. Germline mutations in RAD51D confer susceptibility to ovarian cancer. *Nat Genet*. 2011;43(9):879-882. doi:10.1038/ng.893

34. Ahmed M, Rahman N. ATM and breast cancer susceptibility. *Oncogene*. 2006;25(43):5906-5911. doi:10.1038/sj.onc.1209873

35. Nevanlinna H, Bartek J. The CHEK2 gene and inherited breast cancer susceptibility. *Oncogene*. 2006;25(43):5912-5919. doi:10.1038/sj.onc.1209877

36. Schrader KA, Masciari S, Boyd N, et al. Germline mutations in CDH1 are infrequent in women with early-onset or familial lobular breast cancers. *J Med Genet*. 2011;48(1):64-68. doi:10.1136/jmg.2010.079814

37. Zhan H, Mo F, Xu Q, et al. An integrative pan-cancer analysis reveals the oncogenic role of mutS homolog 6 (MSH6) in human tumors. *Aging (Albany NY)*. 2021;13(23):25271-25290. doi:10.18632/aging.203745

38. Bubien V, Bonnet F, Brouste V, et al. High cumulative risks of cancer in patients with PTEN hamartoma tumour syndrome. *J Med Genet*. 2013;50(4):255-263. doi:10.1136/jmedgenet-2012-101339

39. Baas AF, Smit L, Clevers H. LKB1 tumor suppressor protein: PARtaker in cell polarity. *Trends Cell Biol*. 2004;14(6):312-319. doi:10.1016/j.tcb.2004.04.001

40. Kang E, Kim YM, Seo GH, et al. Phenotype categorization of neurofibromatosis type I and correlation to NF1 mutation types. *J Hum Genet*. 2020;65(2):79-89. doi:10.1038/s10038-019-0695-0

41. Mersch J, Jackson M, Park M, et al. Cancers Associated with BRCA1 and BRCA2 Mutations other than Breast and Ovarian. *Cancer*. 2015;121(2):269-275. doi:10.1002/cncr.29041

42. Cancer Statistics Review, 1975-2014 - SEER Statistics. SEER. Accessed June 20, 2019. https://seer.cancer.gov/archive/csr/1975_2014/#citation

43. Whittemore AS, Gong G, John EM, et al. Prevalence of BRCA1 mutation carriers among U.S. non-Hispanic Whites. *Cancer Epidemiol Biomarkers Prev*. 2004;13(12):2078-2083.

44. Meijers-Heijboer H, van den Ouweland A, Klijn J, et al. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet*. 2002;31(1):55-59. doi:10.1038/ng879

45. Michailidou K, Lindström S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*. 2017;551(7678):92-94. doi:10.1038/nature24284

46. Shu X, Long J, Cai Q, et al. Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. *Nat Commun*. 2020;11(1):1217. doi:10.1038/s41467-020-15046-w

47. Jia G, Ping J, Shu X, et al. Genome- and transcriptome-wide association studies of 386,000 Asian and European-ancestry women provide new insights into breast cancer genetics [In Review]. *The American Journal of Human Genetics*.

48. Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. *Nat Genet*. 2008;40(1):17-22. doi:10.1038/ng.2007.53

49. Zhang B, Beeghly-Fadiel A, Long J, Zheng W. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *The Lancet Oncology*. 2011;12(5):477-488. doi:10.1016/S1470-2045(11)70076-6

50. Jia G, Lu Y, Wen W, et al. Evaluating the Utility of Polygenic Risk Scores in Identifying High-Risk Individuals for Eight Common Cancers. *JNCI Cancer Spectr*. 2020;4(3):pkaa021. doi:10.1093/jncics/pkaa021

51. Zhang H, Ahearn TU, Lecarpentier J, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat Genet*. 2020;52(6):572-581. doi:10.1038/s41588-020-0609-2

52. Ho WK, Tai MC, Dennis J, et al. Polygenic risk scores for prediction of breast cancer risk in Asian populations. *Genetics in Medicine*. 2022;24(3):586-600. doi:10.1016/j.gim.2021.11.008

53. Shieh Y, Fejerman L, Lott PC, et al. A Polygenic Risk Score for Breast Cancer in US Latinas and Latin American Women. *JNCI: Journal of the National Cancer Institute*. 2020;112(6):590-598. doi:10.1093/jnci/djz174

54. Popejoy AB, Fullerton SM. Genomics is failing on diversity. *Nature*. 2016;538(7624):161-164. doi:10.1038/538161a

55. Martin AR, Teferra S, Möller M, Hoal EG, Daly MJ. The critical needs and challenges for genetic architecture studies in Africa. *Current Opinion in Genetics & Development*. 2018;53:113-120. doi:10.1016/j.gde.2018.08.005

56. Zheng Y, Ogundiran TO, Falusi AG, et al. Fine mapping of breast cancer genome-wide association studies loci in women of African ancestry identifies novel susceptibility markers. *Carcinogenesis*. 2013;34(7):1520-1528. doi:10.1093/carcin/bgt090

57. Feng Y, Stram DO, Rhie SK, et al. A comprehensive examination of breast cancer risk loci in African American women. *Hum Mol Genet*. 2014;23(20):5518-5526. doi:10.1093/hmg/ddu252

58. Huo D, Zheng Y, Ogundiran TO, et al. Evaluation of 19 susceptibility loci of breast cancer in women of African ancestry. *Carcinogenesis*. 2012;33(4):835-840. doi:10.1093/carcin/bgs093

59. Du Z, Gao G, Adedokun B, et al. Evaluating Polygenic Risk Scores for Breast Cancer in Women of African Ancestry. *JNCI: Journal of the National Cancer Institute*. 2021;113(9):1168-1176. doi:10.1093/jnci/djab050

60. Chen Z, Guo X, Long J, et al. Discovery of structural deletions in breast cancer predisposition genes using whole genome sequencing data from > 2000 women of African-ancestry. *Hum Genet*. 2021;140(10):1449-1457. doi:10.1007/s00439-021-02342-8

61. Kasimatis KR, Abraham A, Ralph PL, Kern AD, Capra JA, Phillips PC. Evaluating human autosomal loci for sexually antagonistic viability selection in two large biobanks. *Genetics*. 2021;217(1):1-10. doi:10.1093/genetics/iyaa015

62. Adedokun B, Du Z, Gao G, et al. Cross-ancestry GWAS meta-analysis identifies six breast cancer loci in African and European ancestry women. *Nat Commun*. 2021;12(1):4198. doi:10.1038/s41467-021-24327-x

63. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19(9):1655-1664. doi:10.1101/gr.094052.109

64. Kowalski MH, Qian H, Hou Z, et al. Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLOS Genetics*. 2019;15(12):e1008500. doi:10.1371/journal.pgen.1008500

65. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4:7. doi:10.1186/s13742-015-0047-8

66. Devlin B, Roeder K. Genomic Control for Association Studies. *Biometrics*. 1999;55(4):997-1004. doi:10.1111/j.0006-341X.1999.00997.x

67. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26(17):2190-2191. doi:10.1093/bioinformatics/btq340

68. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. *The American Journal of Human Genetics*. 2011;88(1):76-82. doi:10.1016/j.ajhg.2010.11.011

69. Gauderman WJ. Sample Size Requirements for Association Studies of Gene-Gene Interaction. *American Journal of Epidemiology*. 2002;155(5):478-484. doi:10.1093/aje/155.5.478

70. Giro-Perafita A, Luo L, Khodadadi-Jamayran A, et al. LncRNA RP11-19E11 is an E2F1 target required for proliferation and survival of basal breast cancer. *NPJ Breast Cancer*. 2020;6:1. doi:10.1038/s41523-019-0144-4

71. Han YJ, Boatman SM, Zhang J, et al. LncRNA BLAT1 is Upregulated in Basal-like Breast Cancer through Epigenetic Modifications. *Sci Rep*. 2018;8(1):15572. doi:10.1038/s41598-018-33629-y

72. Park C, Lee Y, Je S, et al. Overexpression and Selective Anticancer Efficacy of ENO3 in STK11 Mutant Lung Cancers. *Mol Cells*. 2019;42(11):804-809. doi:10.14348/molcells.2019.0099

73. Fachal L, Aschard H, Beesley J, et al. Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nature Genetics*. 2020;52(1):56-73. doi:10.1038/s41588-019-0537-1

74. Mavaddat N, Pharoah PDP, Michailidou K, et al. Prediction of Breast Cancer Risk Based on Profiling With Common Genetic Variants. *J Natl Cancer Inst*. 2015;107(5). doi:10.1093/jnci/djv036

75. Gao G, Zhao F, Ahearn TU, et al. Polygenic Risk Scores for Prediction of Breast Cancer Risk in Women of African Ancestry: a Cross-Ancestry Approach. *Human Molecular Genetics*. Published online May 12, 2022:ddac102. doi:10.1093/hmg/ddac102

76. Pepe MS, Longton G, Janes H. Estimation and Comparison of Receiver Operating Characteristic Curves. *The Stata Journal*. 2009;9(1):1-16. doi:10.1177/1536867X0900900101

77. Zheng W, Wen W, Gao YT, et al. Genetic and Clinical Predictors for Breast Cancer Risk Assessment and Stratification Among Chinese Women. *J Natl Cancer Inst*. 2010;102(13):972-981. doi:10.1093/jnci/djq170

78. Wen W, Shu X ou, Guo X, et al. Prediction of breast cancer risk based on common genetic variants in women of East Asian ancestry. *Breast Cancer Research*. 2016;18(1):124. doi:10.1186/s13058-016-0786-1

79. Siu AL. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann Intern Med*. 2016;164(4):279-296. doi:10.7326/M15-2886

80. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci*. 2013;368(1620):20120362. doi:10.1098/rstb.2012.0362

81. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet*. 2015;47(9):1091-1098. doi:10.1038/ng.3367

82. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet*. 2016;48(3):245-252. doi:10.1038/ng.3506

83. Wu L, Shi W, Long J, et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nature Genetics*. 2018;50(7):968-978. doi:10.1038/s41588-018-0132-x

84. Ping J, Guo X, Ye F, et al. Differences in gene-expression profiles in breast cancer between African and European-ancestry women. *Carcinogenesis*. 2020;41(7):887-893. doi:10.1093/carcin/bgaa035

85. Colaprico A, Silva TC, Olsen C, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*. 2016;44(8):e71. doi:10.1093/nar/gkv1507

86. Liu J, Lichtenberg T, Hoadley KA, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*. 2018;173(2):400-416.e11. doi:10.1016/j.cell.2018.02.052

87. Frankish A, Diekhans M, Ferreira AM, et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res*. 2019;47(D1):D766-D773. doi:10.1093/nar/gky955

88. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat Protoc*. 2012;7(3):500-507. doi:10.1038/nprot.2011.457

89. Guo X, Lin M, Rockowitz S, Lachman HM, Zheng D. Characterization of Human Pseudogene-Derived Non-Coding RNAs for Functional Potential. *PLOS ONE*. 2014;9(4):e93972. doi:10.1371/journal.pone.0093972

90. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

91. Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*. 2016;48(5):481-487. doi:10.1038/ng.3538

92. Barbeira AN, Dickinson SP, Bonazzola R, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat Commun*. 2018;9(1):1825. doi:10.1038/s41467-018-03621-1

93. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *The American Journal of Human Genetics*. 2019;104(3):410-421. doi:10.1016/j.ajhg.2019.01.002

94. Ferreira MA, Gamazon ER, Al-Ejeh F, et al. Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. *Nat Commun*. 2019;10(1):1741. doi:10.1038/s41467-018-08053-5

95. Wen W, Chen Z, Bao J, et al. Genetic variations of DNA bindings of FOXA1 and co-factors in breast cancer susceptibility. *Nat Commun*. 2021;12(1):5318. doi:10.1038/s41467-021-25670-9

96. Feng H, Gusev A, Pasaniuc B, et al. Transcriptome-wide association study of breast cancer risk by estrogen-receptor status. *Genet Epidemiol*. 2020;44(5):442-468. doi:10.1002/gepi.22288

97. He L, Fan C, Kapoor A, et al. α-Mannosidase 2C1 attenuates PTEN function in prostate cancer cells. *Nat Commun*. 2011;2(1):307. doi:10.1038/ncomms1309

98. Tashiro K, Tsunematsu T, Okubo H, et al. GAREM, a novel adaptor protein for growth factor receptor-bound protein 2, contributes to cellular transformation through the activation of extracellular signal-regulated kinase signaling. *J Biol Chem*. 2009;284(30):20206-20214. doi:10.1074/jbc.M109.021139

99. Hashimoto A, Hashimoto S, Sugino H, et al. ZEB1 induces EPB41L5 in the cancer mesenchymal program that drives ARF6-based invasion, metastasis and drug resistance. *Oncogenesis*. 2016;5(9):e259-e259. doi:10.1038/oncsis.2016.60

100. Lawrenson K, Kar S, McCue K, et al. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast-ovarian cancer susceptibility locus. *Nat Commun*. 2016;7:12675. doi:10.1038/ncomms12675

101. Common polygenic variation contributes to risk of schizophrenia that overlaps with bipolar disorder. *Nature*. 2009;460(7256):748-752. doi:10.1038/nature08185

102. Privé F, Arbel J, Vilhjálmsson BJ. LDpred2: better, faster, stronger. *Bioinformatics*. 2020;36(22-23):5424-5431. doi:10.1093/bioinformatics/btaa1029

103. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun*. 2019;10(1):1776. doi:10.1038/s41467-019-09718-5

104. Ruan Y, Lin YF, Feng YCA, et al. Improving polygenic prediction in ancestrally diverse populations. *Nat Genet*. Published online May 5, 2022:1-8. doi:10.1038/s41588-022-01054-7

105. Weissbrod O, Kanai M, Shi H, et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. *Nat Genet*. 2022;54(4):450-458. doi:10.1038/s41588-022-01036-9

VIII. APPENDICES

Appendix 1: Description of participating studies

The African American Breast Cancer Genetic (AABCG) consortium consists of whole genome sequencing data, newly generated genotyping data using the Multi-Ethnic Genotyping Array (MEGA), and genotyping data from existing studies/consortia. Relatives or the same participants (a close relationship with a Pi-HAT estimate >0.45) can be genotyped by different arrays. Of them, the samples genotyped by an array of a higher density were kept.

1. Whole genome sequencing data

The whole genome sequencing data included 1,337 breast cancer cases and 658 cancer-free controls from five studies: SCCS, NBHS, STSBHS, GBHS, and MEC. The whole-genome sequencing was performed using the Illunima HiSeq X Ten and BGISEQ-500 platforms. Details on sequencing library construction and data processing have been previously published.[1] In brief, all of the sequencing samples reached the average sequencing depth with at least 30X. The sequencing reads were aligned to the human reference genome (GRCh38) using the Burrows–Wheeler Aligner (BWA) program (version 0.79a). The mapped reads were further processed by removing the duplicated reads using MarkDuplicates from the Picard tool and recalibrating the base quality scores using BaseRecalibrator.

1.1 Southern Community Cohort Study (SCCS)

The SCCS is a prospective cohort study focused on the recruitment of a low-income, predominantly African-American population from a 12-state area of southeastern U.S.[2,3] Approximately 86,000 study participants aged 40–79 years were recruited between 2002 and 2009. About 32,500 of the SCCS participants were African American women. An in-person interview was completed at enrollment. Participants were asked to donate a 20-ml blood sample, and a buccal cell or saliva specimen was accepted if the subject did not wish to donate blood. To obtain follow-up data on cancer development, procedures for data linkage, processing, and quality control were established with the 12-state cancer registries covering the SCCS catchment area (Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, and West Virginia). In SCCS, samples from 321 cases and 376 controls were sequenced by Illunima HiSeq X Ten and BGISEQ-500 platforms. In addition, 71 cases and 1,639 controls from SCCS were sequenced at >20x coverage on the Illumina HiSeq X and NovaSeq platforms with paired-end 150 bp reads, which were called as WGS-2 in this study.

1.2 Nashville Breast Health Study (NBHS)

The NBHS is a population-based case-control study conducted in the Nashville metropolitan area.[4] Participants were recruited between 2001 and 2011. Eligible cases were women newly diagnosed with primary breast cancer between 25 and 75 years of age and with no prior history of cancer other than nonmelanoma skin cancer. Breast cancer cases (n =2,694) were identified through the Tennessee State Cancer Registry and five major hospitals in Nashville that provide medical care for breast cancer patients. Controls (n =2,384) were identified via random digit dialing of households in the same geographic area as the cases. Saliva samples were collected as a source of genomic DNA for genetic studies of breast cancer. In NBHS, samples from 91 cases and 16 controls were sequenced by BGISEQ-500 platform.

1.3 Southern Tri-State Breast Health Study (STSBHS)

The STSBHS is a population-based case-only study conducted in Tennessee, South Carolina, and Georgia.[5] It recruited women with incident invasive breast cancer diagnosed between ages 25 and 75 years from 2012 to 2018, without any prior cancer history other than non-melanoma skin cancer. Breast cancer cases were identified through the Tennessee Cancer Registry, the South Carolina Central Cancer Registry, and the Georgia Comprehensive Cancer Registry. In STSBHS, samples from 421 cases were sequenced by BGISEQ-500 platform.

1.4 Multiethnic Cohort Study (MEC)

The MEC is a prospective cohort study conducted in Hawaii and the Los Angeles area and included 215,251 study participants recruited between 1993 and 1996.[6] African American study participants were recruited from the Los Angeles area. Data collection at baseline included a detailed, self-administered questionnaire that obtained information on basic demographic variables and several lifestyle and medical variables that have been associated with cancer risk. Incident cancer cases were identified through two state-wide Surveillance, Epidemiology, and End Results (SEER) registries: the Hawaii Tumor Registry and the California State Cancer Registry. Blood samples were collected from a substantial portion of the cohort. In MEC, samples from 211 cases and 119 controls were sequenced by Illunima HiSeq X Ten.

1.5 Ghana Breast Health Study (GBHS)

The GBHS is a population-based case-control study conducted in Accra and Kumasi, Ghana, between 2013 and 2015.[7,8] Breast cancer cases were identified from women recommended for biopsy of a breast lesion suspicious for malignancy at one of the three major cancer treatment hospitals in Ghana (Korle Bu Teaching Hospital in Accra, and Komfo Anokye Teaching Hospital, and Peace and Love Hospital in Kumasi); or women presenting for treatment of pathologically-confirmed breast cancer at Korle Bu, Komfo Anokye, or Peace and Love Hospitals within one year of diagnosis. Controls were frequency matched to cases by age and district of residence. In GBHS, samples from 293 cases and 147 controls were sequenced by BGISEQ-500 platform.

2. Newly generated genotyping data using Multi-Ethnic Genotyping Array

In the African American Breast Cancer Genetic (AABCG) consortium, we genotyped samples from multiple studies using the Multi-Ethnic Genotyping Array (MEGA). The MEGA chip contains over 2 million variants with an excellent genomic coverage of common variants across multi-racial populations. Samples were genotyped in Vanderbilt University Medical Center (SCCS, NBHS, STSBHS, MDABCS, CCPS, NBCS, NC-BCFR, NYUWHS), Roswell Park Comprehensive Cancer Center (BWHS, WCHS), and University of Southern California (MEC). Samples overlapped between whole genome sequencing data and genotyping data were only kept in the whole genome sequencing dataset.

Quality control (QC) procedure includes: samples were excluded if they (i) were not genetically female; (ii) had a call rate <95%; (iii) had a low proportion of African ancestry (<25%) using Admixture,[9] using 1000 Genome samples as reference; (iv) had a close relationship with a Pi-HAT estimate >0.45 (one of the pair was excluded). Variants were excluded if they had (i) a call rate <95%; (ii) a P value $<10^{-6}$ in the Hardy-Weinberg equilibrium test among the controls; (iii) a consistent rate <98% across duplicated QC samples; (iv) inconsistent alleles from 1000 Genome data.

2.1 Southern Community Cohort Study (SCCS)

The SCCS has been described above. After quality control, there were 708 cases and 678 controls from SCCS newly genotyped by MEGA. In addition, 2,265 controls from SCCS were selected for cases from NBHS, STSBHS, MDABCS, and NC-BCFR. The controls were frequently matched on age and state of residence (only for NBHS and STSBHS).

2.2 Nashville Breast Health Study (NBHS)

The NBHS has been described above. After quality control, new MEGA genotyping data included 138 cases from NBHS and 147 controls matched from SCCS.

2.3 Southern Tri-State Breast Health Study (STSBHS)

The STSBHS has been described above. After quality control, new MEGA genotyping data included 692 cases from STSBHS and 683 controls matched from SCCS.

2.4 M.D. Anderson Breast Cancer Study (MDABCS)

All breast cancer cases in MDABCS are newly registered, histologically confirmed breast cancer patients at MD Anderson Cancer Center.[10] Basic demographic and epidemiological information including smoking, alcohol, education, and family history data were collected as part of institutional patient history database. Clinical data were abstracted from electronic medical records by clinical coding specialists. DNA were extracted from residual blood samples and banked in the institutional Blood Specimen Research Resource. In the MEGA genotyping data, controls from SCCS were frequently matched on age. After quality control, new MEGA genotyping data included 1,294 cases from MDABCS and 1,222 controls matched from SCCS.

2.5 Chicago Cancer Prone Study (CCPS)

The CCPS is a hospital-based case-control study designed to investigate the genetics of young-onset breast cancer. Cases with histologically confirmed breast cancer were enrolled through the Cancer Risk Clinic at the University of Chicago. Young-onset cases and African Americans were oversampled. Controls were gender- and age-matched with cases and enrolled from patients who visited the same hospital and were willing to donate blood for genetic studies. After quality control, new MEGA genotyping data included 366 cases and 279 controls from CCPS.

2.6 Nigerian Breast Cancer Study (NBCS)

The NBCS is an ongoing case-control study of breast cancer in Ibadan, Nigeria initiated in 1998.[11,12] Breast cancer cases were 20 years or older, ascertained at the University College Hospital, Ibadan, which is the oldest tertiary hospital in Nigerian with a catchment population of approximate three million. Controls were recruited from a randomly selected community in one of the communities adjoining the hospital. The majority of the study subjects were Yoruba and

Yoruba is one of the populations selected by the International HapMap Project to represent African continent. After quality control, new MEGA genotyping data included 695 cases and 376 controls from NBCS.

2.7 Northern California Breast Cancer Family Registry (NC-BCFR)

Incident breast cancer cases included women aged <65 years, identified through the SEER cancer registry of the Greater San Francisco Bay Area (diagnoses 1995-2009) and the Sacramento region (diagnoses 2005-2006).[13,14] All cases with indicators of inherited breast cancer were included. Among cases aged 35-64 years without such indicators, cases from racial and ethnic minority populations were oversampled. Controls were identified through random-digit dialing and frequency matched to cases diagnosed from 1995-1998 on 5-year age group and race/ethnicity, at a ratio of one control per two cases. In the MEGA genotyping data, controls from SCCS were selected by frequently matched on age. After quality control, new MEGA genotyping data included 185 cases from NC-BCFR and 213 controls matched from SCCS.

2.8 New York University Women's Health Study (NYUWHS)

The NYUWHS is a cohort study which enrolled 14,274 women aged 34 to 65 years attending Guttman Breast Diagnostic Institute in New York City for yearly screening from 1985 to 1991.[15,16] Self-administered questionnaires were used to collect demographic, medical, anthropometric, reproductive, and dietary. Non-fasting peripheral venous blood was drawn prior to breast examination and serum samples were stored at -80°C for subsequent biochemical analyses. Up until 1991, women who returned to the clinic for annual breast cancer screening were asked to donate blood at each of their visits. Cases were breast cancer patients arising from in the cohort, and controls were women selected from the same cohort who were not diagnosed with breast cancer and matched to cases on age and follow up time. After quality control, new MEGA genotyping data included 72 cases and 58 controls from NYUWHS.

2.9 Women's Circle of Health Study (WCHS)

The WCHS is a case-control study established in 2003 in the New York City metropolitan areas, and beginning in 2006, from 10 counties in New Jersey.[17] Eligible cases included women who were diagnosed with invasive breast cancer between 20 and 75 years of age and self-identified as European-American or African-American. Controls were initially identified through random digit dialing and were matched to cases by self-reported race and 5-year age categories. From 2009-2012, controls were recruited through community events, particularly through churches.[18] After quality control, new MEGA genotyping data included 1,326 cases and 851 controls from WCHS.

2.10 Black Women's Health Study (BWHS)

The BWHS is a prospective cohort study which recruited approximately 59,000 African American women, aged 21-60 years, from all regions of the United States since 1995.[19] Participants were enrolled by completing a postal health questionnaire and were followed by mail questionnaires every two years. DNA samples were obtained from BWHS participants (26,800 women) by the mouthwash-swish method with all samples stored in freezers at −80°C. After quality control, new MEGA genotyping data included 1,282 cases and 1,879 controls from BWHS.

2.11 Multiethnic Cohort Study (MEC)

The MEC has been described above. In the MEGA genotyping data, cases were from MEC and controls were from African American Eye Disease Study (AFEDS). The AFEDS is a population-based cohort study conducted from April 2014 to April 2018 including 6,347 African American adults, 40 years of age and older residing in 32 census tracts in and around Inglewood, CA within Los Angeles County. The participation rate for eligible residents who completed the clinical examination and home interview was 80% (6,347 of 7,957 eligible). While the study was conducted to fill the gaps in our understanding of vision health in African American adults, selection was independent of eye health. Biological samples included a blood draw and/or a saliva sample depending on the selection by the participant after informed consent was completed. Data on demographic and behavioral characteristics, medical and ocular history, insurance status and access to care were collected, and a comprehensive eye examination was conducted. Detailed methods have been published elsewhere.[20] A random sample of female participants from the cohort were selected for inclusion in the AABCGS genetic study based on number of budgeted tests (N=969). Women with a self-reported history of breast cancer were excluded. The mean age of the genotyping set was 57.8 years (SD=10.0) with a minimum age of 40 and a maximum age of 83 years. Of the 969 participants, 4.3% had less than a high school education, 17.4% had a high school education, 38.5% had some college education, and 37% had a college degree or higher (2.8% not reported). After quality control, the MEGA genotyping data included 1,194 cases from MEC and 914 controls from AFEDS.

3. African American Breast Cancer Epidemiology and Risk (AMBER) consortium

The AMBER is a collaboration of four studies: the Women's Circle of Health Study (WCHS), the Black Women's Health Study (BWHS), the Carolina Breast Cancer Study (CBCS), and the Multiethnic Cohort Study (MEC) funded by the National Cancer Institute. In the AMBER phase 2, a total of 4,224 study samples were genotyped by MEGA chip and a set of custom variants selected from breast cancer candidate loci at the Center for Inherited Disease Research at Johns Hopkins University.

3.1 Carolina Breast Cancer Study (CBCS)

The CBCS is a population-based case-control study conducted in 24 counties of central and eastern North Carolina.[21] From 1993 to 2001, it recruited women aged between 20 and 74 years and diagnosed with invasive breast cancer. African-American women and women aged less than 50 years were oversampled. Cases were identified by rapid case ascertainment system in cooperation with the North Carolina Central Cancer Registry. Controls were selected from the North Carolina Division of Motor Vehicle (for women younger than 65 years) and United States Health Care Financing Administration (for women aged 65 and older). Controls were approximately frequency matched to cases by age and race. Blood samples were collected from participants with consent. After quality control, we included 602 cases and 1 control of African ancestry from CBCS in the AMBER phase2.

3.2 Women's Circle of Health Study (WCHS)

The WCHS has been described above. After quality control, we included 472 cases and 243 controls of African ancestry from WCHS in the AMBER phase2.

3.3 Black Women's Health Study (BWHS)

The BWHS has been described above. After quality control, we included 307 cases and 2,098 controls of African ancestry from BWHS in the AMBER phase2.

4. The GWAS of Breast Cancer in the African Diaspora (ROOT) consortium

The ROOT consortium consists of samples from NBCS, BNCS, RVGBC, CCPS, BBCS, and SCCS. The samples in ROOT consortium were genotyped using the Illumina HumanOmni2.5-8v1 array.

4.1 Baltimore Breast Cancer Study (BBCS)

The BBCS is a case control study of breast cancer designed to identify and characterize markers of disease aggressiveness and poor outcome. From 1993 to 2003, incident breast cancer cases and controls were recruited from six hospitals in the greater Baltimore area, including the University of Maryland Medical Center, the Baltimore Veterans Affairs Medical Center, Union Memorial Hospital, Mercy Medical Center, and the Sinai Hospital. Controls were frequency matched to cases by race and age. After quality control, we included 94 cases and 102 controls of African ancestry from BBCS in the ROOT consortium.

4.2 Barbados National Cancer Study (BNCS)

The BNCS is a population-based case-control study of incident breast and prostate cancer in the predominantly African population of Barbados, West Indies.[22] Breast cancer cases were histologically confirmed incident cases identified through the only pathology department on the island, located at the Queen Elizabeth Hospital, between July 2002 and March 2006. Controls were selected from a national database provided by the Barbados Statistical Services Department, and were frequency matched to breast cancer cases at a 2:1 ratio and by 5-year age groups. Blood samples were collected from participants. After quality control, we included 92 cases and 227 controls of African ancestry from BNCS in the ROOT consortium.

4.3 Racial Variability in Genotypic Determinants of Breast Cancer Risk Study (RVGBC)

RVGBC is a hospital-based case-control study conducted in Philadelphia and Detroit metropolitan areas from 1999 to 2003. Breast cancer cases were identified in the University of Pennsylvania Health System and Karmanos Cancer Institute. Local advertisement was also put to recruit breast cancer cases living in the Philadelphia and Detroit area. Controls were recruited in the same way as cases except that they did not have breast cancer. Patients with breast cancer had to be diagnosed within 18 months of recruitment and have invasive ductal cancer. The study over-sampled women diagnosed with breast cancer under age of 40 years. After quality control, we included 143 cases and 254 controls of African ancestry from BNCS in the ROOT consortium.

4.4 Chicago Cancer Prone Study (CCPS)

The CCPS has been described above. After quality control, we included 365 cases and 376 controls of African ancestry from CCPS in the ROOT consortium.

4.5 Nigerian Breast Cancer Study (NBCS)

The NBCS has been described above. After quality control, we included 702 cases and 602 controls of African ancestry from NBCS in the ROOT consortium.

4.6 Southern Community Cohort Study (SCCS)

The SCCS has been described above. After quality control, we included 126 cases and 323 controls of African ancestry from SCCS in the ROOT consortium.

5. The African American Breast Cancer (AABC) consortium

The AABC consortium consists of samples from NBHS, NC-BCFR, CARE, CBCS, MEC, PLCO, SFBCS, WCHS, and WFBC. The samples in AABC consortium were genotyped using IlluminaHuman1M-Duo BeadChip.

5.1 The Los Angeles component of the Women's Contraceptive and Reproductive Experiences Study (CARE)

The Women's CARE Study is a large multi-center population-based case-control study sponsored by the National Institute of Child Health and Human Development (NICHD).[23] It was designed to examine the effects of oral contraceptive use on invasive breast cancer risk. Cases diagnosed with breast cancer between 34 and 64 years of age were recruited in five U.S. locations (Atlanta, Detroit, Los Angeles, Philadelphia, and Seattle). Cases in Los Angeles County were diagnosed from July 1, 1994 through April 30, 1998, and controls were sampled by random-digit dialing from the same population and time period. After quality control, we included 254 cases and 204 controls of African ancestry from CARE in the AABC consortium.

5.2 The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)

The PLCO is a multicenter, two-armed, randomized trial designed to evaluate the screening efficacy for prostate, lung, colorectal and ovarian cancer.[24] It recruited approximately 155,000 men and women, aged 55-74 years, from 1993 to 2001. After quality control, we included 24 cases and 68 controls of African ancestry from PLCO in the AABC consortium.

5.3 San Francisco Bay Area Breast Cancer Study (SFBCS)

The SFBCS is a population-based case control study of invasive breast cancer in Hispanic, African American and non-Hispanic White women in the San Francisco Bay Area.[25] From 1995 to 2003, women aged 35–79 years, diagnosed with a first primary invasive breast cancer were identified through the California population-based Greater Bay Area Cancer Registry. Population controls were identified through random digit dialing. After quality control, we included 157 cases and 210 controls of African ancestry from SFBCS in the AABC consortium.

5.4 Wake Forest University Breast Cancer Study (WFBC)

The WFBC is a clinic-based case-control study at Wake Forest University Health Sciences from 1998 to 2008.[26,27] Incident breast cancer cases were recruited at the Wake Forest University Breast Care Center. Controls were recruited from the patient population receiving routine mammography at the Outpatient Radiology-Breast Screening Center. Blood samples (20 ml) were collected from all study subjects. After quality control, we included 113 cases and 138 controls of African ancestry from WGBC in the AABC consortium.

5.5 Nashville Breast Health Study (NBHS)

The NBHS has been described above. After quality control, we included 255 cases and 161 controls of African ancestry from NBHS in the AABC consortium.

5.6 Northern California Breast Cancer Family Registry (NC-BCFR)

The NC-BCFR has been described above. After quality control, we included 383 cases and 48 controls of African ancestry from NC-BCFR in the AABC consortium.

5.7 Carolina Breast Cancer Study (CBCS)

The CBCS has been described above. After quality control, we included 614 cases and 570 controls of African ancestry from CBCS in the AABC consortium.

5.8 Multiethnic Cohort Study (MEC)

The MEC has been described above. After quality control, we included 578 cases and 888 controls of African ancestry from MEC in the AABC consortium.

5.9 Women's Circle of Health Study (WCHS)

The WCHS has been described above. After quality control, we included 63 cases and 21 controls of African ancestry from WCHS in the AABC consortium.

6. Ghana Breast Health Study (GBHS)

The GBHS has been described above. The GBHS samples were genotyped using Infinium Global Screening Array-24. After quality control, we included 660 cases and 1,496 controls of African ancestry from GBHS genotyping data.

7. Genetic Associations and Mechanisms in Oncology (GAME-ON) OncoArray consortium

The GAME-ON OncoArray consortium consists of samples from NBHS, CBCS, NC-BCFR, MEC, PLCO, 2SISTER, SISTER, USRT, and WAABCS. The samples in GAME-ON OncoArray consortium were genotyped using the Infinium OncoArray-500k BeadChip.

7.1 The Sister Study (SISTER)

The Sister Study is a prospective cohort study designed address genetic and environmental risk factors for breast cancer by the National Institute of Environmental Health Sciences. From 2003 through 2009, 50,884 U.S. and Puerto Rican women were recruited through a national multimedia campaign and network of recruitment volunteers, breast cancer professionals, and advocates. Participants were women aged 35 to 74 years and had a sister diagnosed with breast cancer.[28] At enrollment, participants completed baseline questionnaires on medical and family history, lifestyle factors, and demographics. Blood samples were collected during a home visit by trained phlebotomists and shipped overnight to the Sister Study laboratory where they were processed to obtain serum and stored at −80°C. After quality control, we included 130 cases and 163 controls of African ancestry from SISTER in the OncoArray consortium.

7.2 The Two Sister Study (2SISTER)

The Two Sister Study is a family-based retrospective study developed from the Sister Study. The Two Sister Study recruited the case sisters in the Sister Study who were diagnosed within 4 years and had been younger than age 50 years at diagnosis.[29] After quality control, we included 42 cases of African ancestry from 2SISTER in the OncoArray consortium.

7.3 The United States Radiologic Technologists (USRT) cohort

The USRT is a cohort study for cancer incidence and mortality which recruited approximately 140,00 U.S. radiologic technologists who were certified for at least two years between 1926 and 1982.[30] Breast cancer cases were confirmed based on pathology or medical records. After quality control, we included 26 cases and 38 controls of African ancestry from USRT in the OncoArray consortium.

7.4 Women of African Ancestry Breast Cancer Study (WAABCS)

The WAABCS is a hospital-based case-control study originally started in Nigeria in 1998 and was expanded to Uganda and Cameroon in 2011 with the same questionnaires and protocol.[31,32] After quality control, we included 308 cases and 292 controls of African ancestry from WAABCS in the OncoArray consortium.

7.5 Nashville Breast Health Study (NBHS)

The NBHS has been described above. After quality control, we included 51 cases and 53 controls of African ancestry from NBHS in the OncoArray consortium.

7.6 Carolina Breast Cancer Study (CBCS)

The CBCS has been described above. After quality control, we included 614 cases and 570 controls of African ancestry from CBCS in the OncoArray consortium.

7.7 Northern California Breast Cancer Family Registry (NC-BCFR)

The NC-BCFR has been described above. After quality control, we included 69 cases of African ancestry from NC-BCFR in the OncoArray consortium

7.8 Multiethnic Cohort Study (MEC)

The MEC has been described above. After quality control, we included 605 cases and 607 controls of African ancestry from MEC in the OncoArray consortium.

7.9 The Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)

The PLCO has been described above. After quality control, we included 24 cases and 68 controls of African ancestry from PLCO in the OncoArray consortium.

8. The Vanderbilt Biobank (BioVU)

The DNA biobank at Vanderbilt University consists of DNA extracted from blood collected during routine clinical testing and linked de-identified medical records.[33,34] Samples from more than 90,000 individuals were genotyped using the Illumina MEGA-Ex chip. Breast cancer cases were identified from the electronic medical record systems. In this study, we only kept adult participants of African ancestry. After quality control, we included 118 cases and 2,600 controls of African ancestry from BioVU. In addition, 356 controls of African ancestry from BioVU were selected and frequently matched on with cases from BEST study.

9. Black Women: Etiology and Survival of Triple-negative Breast Cancers (BEST) Study

The BEST is a case-only study which recruited African-ancestry women who were diagnosed with invasive breast cancer at age ≤50 years between 2009 and 2012 and lived in Florida at the time of their diagnosis.[35] Breast cancer cases were identified through the Florida Cancer Registry. Participants provided a saliva sample through mail for DNA extraction and *BRCA* testing.

The BEST samples were genotyped using the Infinium OncoArray-500k BeadChip. Controls from BioVU were selected to match cases from BEST by age. Given that different genotyping arrays were used for BEST and BioVU samples, we only kept genotyped variants shared by both arrays. Other criteria of quality control were the same as MEGA genotyping samples. After quality control, we included 359 cases from BEST and 356 controls matched from BioVU.

10. Data from Collaborative Oncological Gene-environment Study (iCOGS)

NBHS and SCCS (have been described above) contributed some samples in the iCOGS. Samples were genotyped using the Illumina iSelect Genotyping Array. After quality control and excluding the overlapped samples, we included 19 cases and 42 controls from NBHS, 44 cases and 251 controls from SCCS in iCOGS.

References for Appendix 1

1. Chen Z, Guo X, Long J, et al. Discovery of structural deletions in breast cancer predisposition genes using whole genome sequencing data from > 2000 women of African-ancestry. *Hum Genet*. 2021;140(10):1449-1457. doi:10.1007/s00439-021-02342-8

2. Signorello LB, Hargreaves MK, Steinwandel MD, et al. Southern community cohort study: establishing a cohort to investigate health disparities. *J Natl Med Assoc*. 2005;97(7):972-979.

3. Signorello LB, Hargreaves MK, Blot WJ. The Southern Community Cohort Study: Investigating Health Disparities. *J Health Care Poor Underserved*. 2010;21(1):26-37. doi:10.1353/hpu.0.0245

4. Cui Y, Deming-Halverson SL, Beeghly-Fadiel A, et al. Interactions of Hormone Replacement Therapy, Body Weight, and Bilateral Oophorectomy in Breast Cancer Risk. *Clinical Cancer Research*. 2014;20(5):1169-1178. doi:10.1158/1078-0432.CCR-13-2094

5. Sanderson M, Pal T, Beeghly-Fadiel A, et al. A Pooled Case-only Analysis of Reproductive Risk Factors and Breast Cancer Subtype Among Black Women in the Southeastern United States. *Cancer Epidemiol Biomarkers Prev*. 2021;30(7):1416-1423. doi:10.1158/1055-9965.EPI-20-1784

6. Kolonel LN, Altshuler D, Henderson BE. The multiethnic cohort study: exploring genes, lifestyle and cancer risk. *Nat Rev Cancer*. 2004;4(7):519-527. doi:10.1038/nrc1389

7. Brinton LA, Awuah B, Nat Clegg-Lamptey J, et al. Design considerations for identifying breast cancer risk factors in a population-based study in Africa. *International Journal of Cancer*. 2017;140(12):2667-2677. doi:10.1002/ijc.30688

8. Nyante SJ, Biritwum R, Figueroa J, et al. Recruiting population controls for case-control studies in sub-Saharan Africa: The Ghana Breast Health Study. *PLoS One*. 2019;14(4):e0215347. doi:10.1371/journal.pone.0215347

9. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 2009;19(9):1655-1664. doi:10.1101/gr.094052.109

10. Wu X, Hildebrandt MA, Ye Y, et al. Cohort Profile: The MD Anderson Cancer Patients and Survivors Cohort (MDA-CPSC). *International Journal of Epidemiology*. 2016;45(3):713-713f. doi:10.1093/ije/dyv317

11. Huo D, Kim HJ, Adebamowo CA, et al. Genetic polymorphisms in uridine diphospho-glucuronosyltransferase 1A1 and breast cancer risk in Africans. *Breast Cancer Res Treat*. 2008;110(2):367-376. doi:10.1007/s10549-007-9720-7

12. Huo D, Adebamowo CA, Ogundiran TO, et al. Parity and breastfeeding are protective against breast cancer in Nigerian women. *Br J Cancer*. 2008;98(5):992-996. doi:10.1038/sj.bjc.6604275

13. John EM, Hopper JL, Beck JC, et al. The Breast Cancer Family Registry: an infrastructure for cooperative multinational, interdisciplinary and translational studies of the genetic epidemiology of breast cancer. *Breast Cancer Res*. 2004;6(4):R375-389. doi:10.1186/bcr801

14. John EM, Sangaramoorthy M, Koo J, Whittemore AS, West DW. Enrollment and biospecimen collection in a multiethnic family cohort: the Northern California site of the Breast Cancer Family Registry. *Cancer Causes Control*. 2019;30(4):395-408. doi:10.1007/s10552-019-01154-6

15. Toniolo PG, Pasternack BS, Shore RE, et al. Endogenous hormones and breast cancer: a prospective cohort study. *Breast Cancer Res Treat*. 1991;18 Suppl 1:S23-26. doi:10.1007/BF02633522

16. Zeleniuch-Jacquotte A, Afanasyeva Y, Kaaks R, et al. Premenopausal serum androgens and breast cancer risk: a nested case-control study. *Breast Cancer Res*. 2012;14(1):1-12. doi:10.1186/bcr3117

17. Ambrosone CB, Ciupak GL, Bandera EV, et al. 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;2009:871250. doi:10.1155/2009/871250

18. Bandera EV, Chandran U, Zirpoli G, McCann SE, Ciupak G, Ambrosone CB. Rethinking sources of representative controls for the conduct of case–control studies in minority populations. *BMC Med Res Methodol*. 2013;13(1):71. doi:10.1186/1471-2288-13-71

19. Palmer JR, Ruiz-Narvaez EA, Rotimi CN, et al. Genetic Susceptibility Loci for Subtypes of Breast Cancer in an African American Population. *Cancer Epidemiology, Biomarkers & Prevention*. 2013;22(1):127-134. doi:10.1158/1055-9965.EPI-12-0769

20. McKean-Cowdin R, Fairbrother-Crisp A, Torres M, et al. The African American Eye Disease Study: Design and Methods. *Ophthalmic Epidemiology*. 2018;25(4):306-314. doi:10.1080/09286586.2018.1454965

21. Newman B, Moorman PG, Millikan R, et al. The Carolina Breast Cancer Study: integrating population-based epidemiology and molecular biology. *Breast Cancer Res Treat*. 1995;35(1):51-60. doi:10.1007/BF00694745

22. Nemesure B, Wu SY, Hambleton IR, Leske MC, Hennis AJ. Risk factors for breast cancer in a black population—The Barbados National Cancer Study. *International Journal of Cancer*. 2009;124(1):174-179. doi:10.1002/ijc.23827

23. Marchbanks PA, Mcdonald JA, Wilson HG, et al. The NICHD Women's Contraceptive and Reproductive Experiences Study: Methods and Operational Results. *Annals of Epidemiology*. 2002;12(4):213-221. doi:10.1016/S1047-2797(01)00274-5

24. Prorok PC, Andriole GL, Bresalier RS, et al. Design of the prostate, lung, colorectal and ovarian (PLCO) cancer screening trial. *Controlled Clinical Trials*. 2000;21(6, Supplement 1):273S-309S. doi:10.1016/S0197-2456(00)00098-2

25. John EM, Schwartz GG, Koo J, Wang W, Ingles SA. Sun Exposure, Vitamin D Receptor Gene Polymorphisms, and Breast Cancer Risk in a Multiethnic Population. *American Journal of Epidemiology*. 2007;166(12):1409-1419. doi:10.1093/aje/kwm259

26. Smith TR, Levine EA, Freimanis RI, et al. Polygenic model of DNA repair genetic polymorphisms in human breast cancer risk. *Carcinogenesis*. 2008;29(11):2132-2138. doi:10.1093/carcin/bgn193

27. Smith TR, Levine EA, Perrier ND, et al. DNA-Repair Genetic Polymorphisms and Breast Cancer Risk. *Cancer Epidemiology, Biomarkers & Prevention*. 2003;12(11):1200-1204.

28. White AJ, Nichols HB, Bradshaw PT, Sandler DP. Overall and central adiposity and breast cancer risk in the sister study. *Cancer*. 2015;121(20):3700-3708. doi:10.1002/cncr.29552

29. Fei C, DeRoo LA, Sandler DP, Weinberg CR. Fertility Drugs and Young-Onset Breast Cancer: Results From the Two Sister Study. *JNCI: Journal of the National Cancer Institute*. 2012;104(13):1021-1027. doi:10.1093/jnci/djs255

30. Bhatti P, Struewing JP, Alexander BH, et al. Polymorphisms in DNA repair genes, ionizing radiation exposure and risk of breast cancer in U.S. Radiologic technologists. *International Journal of Cancer*. 2008;122(1):177-182. doi:10.1002/ijc.23066

31. Adebamowo CA, Ogundiran TO, Adenipekun AA, et al. Obesity and Height in Urban Nigerian Women with Breast Cancer. *Annals of Epidemiology*. 2003;13(6):455-461. doi:10.1016/S1047-2797(02)00426-X

32. Hou N, Ndom P, Jombwe J, et al. An Epidemiologic Investigation of Physical Activity and Breast Cancer Risk in Africa. *Cancer Epidemiology, Biomarkers & Prevention*. 2014;23(12):2748-2756. doi:10.1158/1055-9965.EPI-14-0675

33. Roden DM, Pulley JM, Basford MA, et al. Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clinical Pharmacology & Therapeutics*. 2008;84(3):362-369. doi:https://doi.org/10.1038/clpt.2008.89

34. Kasimatis KR, Abraham A, Ralph PL, Kern AD, Capra JA, Phillips PC. Evaluating human autosomal loci for sexually antagonistic viability selection in two large biobanks. *Genetics*. 2021;217(1):1-10. doi:10.1093/genetics/iyaa015

35. Pal T, Bonner D, Cragun D, et al. A high frequency of BRCA mutations in young black women with breast cancer residing in Florida. *Cancer*. 2015;121(23):4173-4180. doi:10.1002/cncr.29645

Appendix 2: Known risk loci replicated in African-ancestry women.

Variant	CHR	Position (bd38)	Allele[a]	EAF	Beta	SE	P	Index variant	LD in EUR (r^2)
rs2506885	1	10520994	A/T	0.89	0.06	0.03	0.03	rs2506885	1
rs34167614	1	17401576	A/G	0.12	-0.06	0.02	0.02	rs6586541	0.90
rs7515948	1	41613375	T/G	0.40	-0.05	0.02	1.38×10^{-3}	rs10749837	0.97
rs2066319	1	50994608	A/C	0.59	0.03	0.02	0.05	rs11588271	0.96
rs12118297	1	87313534	T/G	0.19	-0.04	0.02	0.05	rs12118297	1
rs56747346	1	155551443	A/G	0.15	0.05	0.02	0.05	rs12091730	0.75
rs6427303	1	156181635	T/G	0.43	-0.03	0.02	0.08	rs11264454	0.96
rs67931591	1	215156949	G/GCTGAGGCAGGAGA	0.29	-0.04	0.02	0.03	rs67931591	1
rs12075072	1	217029330	T/C	0.74	0.04	0.02	0.01	rs11117758	0.99
rs11684853	2	19111157	T/G	0.41	-0.05	0.02	3.68×10^{-3}	rs11684853	1
rs10637593	2	25176684	CTG/C	0.41	0.04	0.02	0.02	rs10637593	1
rs56158184	2	28973050	T/C	0.91	0.06	0.03	0.02	rs56158184	1
rs727477	2	40299988	T/G	0.41	0.04	0.02	0.01	rs727477	1
rs1430782	2	67647946	A/C	0.67	-0.03	0.02	0.06	rs9712235	0.99
rs6757464	2	69183301	C/G	0.85	0.05	0.02	0.03	rs62134416	0.73
rs3833441	2	111168154	CTTATGTT/C	0.79	-0.04	0.02	0.04	rs73954922	0.97
rs12711947	2	120486696	T/C	0.30	-0.10	0.02	4.16×10^{-8}	rs12711947	1
rs2010610	2	173346180	C/G	0.29	-0.04	0.02	0.03	rs2010610	1
rs3769821	2	201258707	T/C	0.41	-0.05	0.02	3.67×10^{-3}	rs3769821	1
rs4442975	2	217056046	T/G	0.33	-0.09	0.02	2.75×10^{-7}	rs4442975	1
rs16857609	2	217431785	T/C	0.25	0.06	0.02	5.82×10^{-4}	rs16857609	1
rs4266007	2	226370592	A/G	0.45	-0.07	0.02	2.32×10^{-6}	rs12479355	0.80
rs6762558	3	4700567	A/G	0.70	-0.08	0.02	5.00×10^{-6}	rs6762558	1
rs9868094	3	16730246	A/G	0.32	0.04	0.02	0.01	rs9868094	1
rs1352944	3	27332610	A/C	0.10	-0.06	0.03	0.04	rs552647	0.97
rs35263707	3	30642605	A/G	0.32	0.04	0.02	0.02	rs12493607	0.92

Appendix 2. Continued

rs6787229	3	46847697	A/G	0.25	-0.05	0.02	0.01	rs559989662	0.98
rs55917937	3	63909153	A/C	0.56	-0.03	0.02	0.03	rs73117066	0.72
rs11714337	3	71533370	A/G	0.11	-0.06	0.03	0.03	rs6805189	0.75
rs9833888	3	100004736	T/G	0.09	0.08	0.03	0.01	rs9833888	1
rs75942495	3	150760090	A/G	0.18	-0.04	0.02	0.04	rs73006998	0.92
rs75575928	3	172560023	T/C	0.16	0.05	0.02	0.02	rs62282635	0.93
rs16342	4	1983005	T/TAACA	0.80	-0.04	0.02	0.08	rs495367	0.92
rs28713645	4	174923380	T/G	0.53	0.03	0.02	0.07	rs7697216	1
rs6863730	5	324888	T/G	0.45	0.04	0.02	0.01	rs62641919	0.94
rs2853669	5	1295234	A/G	0.88	0.13	0.02	2.57×10^{-7}	rs2853669	1
rs4702131	5	16233510	T/C	0.79	0.05	0.02	0.01	rs4702131	1
rs4866905	5	44555765	T/C	0.38	0.03	0.02	0.05	rs7710996	0.99
rs930395	5	44822356	A/G	0.19	0.04	0.02	0.03	rs10941679	0.76
rs59957907	5	56731413	A/G	0.82	-0.07	0.02	4.97×10^{-4}	rs59957907	1
rs35712350	5	91383702	T/C	0.45	-0.04	0.02	0.01	rs1895449	0.97
rs2522057	5	132466255	C/G	0.86	-0.06	0.02	0.01	rs2522057	1
rs1432679	5	158817075	T/C	0.19	-0.10	0.02	5.74×10^{-7}	rs1432679	1
rs58242049	5	170144460	A/G	0.18	0.05	0.02	0.01	rs56234354	0.97
rs405447	6	13712867	A/G	0.32	0.04	0.02	0.01	rs405447	1
rs10484439	6	26309680	A/G	0.13	-0.05	0.02	0.02	rs71557345	0.74
rs12207986	6	80384570	A/G	0.10	0.05	0.03	0.09	rs1836962	0.98
rs9383590	6	151632630	T/C	0.97	-0.19	0.05	3.63×10^{-5}	rs60954078	0.89
rs910416	6	152111767	T/C	0.54	0.06	0.02	3.60×10^{-4}	rs910416	1
rs2516603	6	167973374	C/G	0.98	-0.11	0.06	0.09	rs3778663	1
rs6940159	6	170017397	T/C	0.41	-0.03	0.02	0.06	rs6940159	1
rs17167582	7	13850781	T/C	0.44	0.03	0.02	0.03	rs17167576	0.71
rs10273849	7	92107807	T/C	0.47	-0.03	0.02	0.03	rs2075881	1
rs35818859	7	94573736	A/T	0.92	-0.05	0.03	0.10	rs2188648	0.90

Appendix 2. Continued

rs68056147	7	130989722	A/G	0.11	0.10	0.03	6.82×10^{-4}	rs68056147	1
rs11977670	7	140242504	A/G	0.27	0.05	0.02	0.01	rs11977670	1
rs310291	8	23801998	A/G	0.57	0.04	0.02	0.03	rs310291	1
rs13256025	8	25974262	T/C	0.07	0.08	0.03	0.01	rs13256025	1
rs7463114	8	29649578	T/C	0.39	0.05	0.02	1.28×10^{-3}	rs7463114	1
rs7816345	8	36988591	T/C	0.40	-0.03	0.02	0.03	rs75772194	0.99
rs72658071	8	75393550	A/T	0.94	-0.14	0.03	4.54×10^{-5}	rs72658071	1
rs2860518	8	100225892	T/G	0.46	0.03	0.02	0.08	rs2849506	0.94
rs13277568	8	115667320	A/G	0.22	0.04	0.02	0.07	rs13277568	1
rs58847541	8	123597926	A/G	0.30	0.04	0.02	0.01	rs58847541	1
rs10096351	8	127359926	A/G	0.38	-0.04	0.02	0.01	rs10096351	1
rs1121948	8	128152810	A/G	0.76	-0.05	0.02	0.01	rs1121948	1
rs9942894	9	104152624	A/G	0.24	-0.05	0.02	0.01	rs10820600	0.99
rs60037937	9	107541527	T/TAA	0.17	0.05	0.02	0.02	rs60037937	1
rs7862747	9	108130619	A/C	0.77	0.04	0.02	0.04	rs7862747	1
rs4455975	9	126620920	A/G	0.21	0.03	0.02	0.09	rs10760444	0.99
rs68088353	10	9048153	T/C	0.88	-0.06	0.02	0.01	rs35781392	0.82
rs6482189	10	21600209	A/G	0.59	0.03	0.02	0.04	rs7098100	0.74
rs10995187	10	62513267	A/G	0.07	-0.06	0.03	0.08	rs10995201	0.94
rs704010	10	79081391	T/C	0.09	0.06	0.03	0.04	rs704010	1
rs12250948	10	113368732	T/C	0.75	0.06	0.02	9.05×10^{-4}	rs12250948	1
rs9420318	10	121333668	A/G	0.52	-0.04	0.02	0.01	rs9420318	1
rs2981579	10	121577821	A/G	0.61	0.07	0.02	5.82×10^{-6}	rs2981579	1
rs588321	11	1875727	C/G	0.20	-0.04	0.02	0.04	rs588321	1
rs10838267	11	44347342	A/G	0.14	0.05	0.02	0.03	rs4755816	0.94
rs78540526	11	69516650	T/C	0.01	0.16	0.08	0.03	rs78540526	1
rs228606	11	108217120	T/G	0.16	-0.05	0.02	0.04	rs199504893	0.98
rs145400227	11	116960788	CAGTAAA/C	0.11	0.06	0.03	0.01	rs36028244	0.85

Appendix 2. Continued

rs10894076	11	129593165	T/C	0.82	0.05	0.02	0.02	rs10894076	1
rs7297051	12	28021884	T/C	0.13	-0.06	0.02	0.01	rs7297051	1
rs17356907	12	95633983	A/G	0.81	0.06	0.02	4.68×10^{-3}	rs17356907	1
rs1292011	12	115398717	A/G	0.55	0.03	0.02	0.09	rs1292011	1
rs61962260	13	50508668	A/G	0.05	-0.06	0.04	0.09	rs2286657	0.99
rs12883049	14	36662856	A/G	0.81	0.07	0.02	6.88×10^{-4}	rs12881240	0.84
rs371902365	14	68209961	CTTT/C	0.77	-0.04	0.02	0.08	rs1744947	0.99
rs11628293	14	68570603	A/G	0.95	0.08	0.04	0.05	rs10483813	0.88
rs4983544	14	104747641	T/G	0.13	-0.05	0.02	0.04	rs4983544	1
rs12900028	15	74373686	C/G	0.13	0.05	0.02	0.04	rs1484216	0.73
rs6938	15	74843920	C/G	0.83	-0.08	0.02	7.29×10^{-4}	rs1869959	0.78
rs4486847	15	75436190	C/G	0.45	0.07	0.02	2.66×10^{-5}	rs8035987	0.93
rs2290203	15	90968837	A/G	0.42	-0.06	0.02	4.24×10^{-4}	rs2290203	1
rs4784227	16	52565276	T/C	0.07	0.23	0.03	1.60×10^{-12}	rs4784227	1
rs7190396	16	53788590	T/G	0.53	0.08	0.02	1.83×10^{-6}	rs62048402	0.93
rs3893264	16	54649890	T/C	0.82	-0.05	0.02	0.02	rs16953806	0.81
rs76535198	16	71858595	A/C	0.91	0.08	0.03	0.01	rs76535198	1
rs8056731	16	80615933	A/G	0.44	-0.06	0.02	5.47×10^{-4}	rs12446424	0.87
rs9898886	17	55172318	A/G	0.63	0.03	0.02	0.05	rs244373	0.81
rs521667	18	26758515	A/C	0.16	-0.05	0.02	0.02	rs527616	0.94
rs2307561	18	26923542	A/AAGTGTT	0.25	-0.04	0.02	0.03	rs2307561	1
rs72931898	18	32401563	A/G	0.15	-0.09	0.02	3.49×10^{-5}	rs72931898	1
rs9954058	18	44831838	C/G	0.31	-0.05	0.02	3.15×10^{-3}	rs9954058	1
rs4609972	19	17280108	C/G	0.42	-0.09	0.02	4.55×10^{-9}	rs4609972	1
rs76194	19	18519770	T/G	0.24	0.03	0.02	0.07	rs8105994	0.96
rs56681946	19	43778879	T/C	0.77	-0.06	0.02	2.36×10^{-3}	rs56681946	1
rs1319363	20	33954070	A/T	0.82	-0.06	0.02	4.73×10^{-3}	rs2284378	0.95
rs6020465	20	50346469	A/G	0.71	-0.04	0.02	0.02	rs6012911	0.99

Appendix 2. Continued

rs531874019	20	53677557	G/GT	0.90	-0.07	0.03	0.01	rs531874019	1
rs11088317	21	15201802	T/C	0.06	-0.09	0.04	0.01	rs2403907	0.90
rs12628403	22	38962032	A/C	0.98	-0.14	0.06	0.01	rs12628403	1
rs73167066	22	40477714	T/C	0.98	-0.15	0.06	0.02	rs66987842	0.80

Abbreviations: CHR, chromosome; EAF, effect allele frequency; EUR, European-ancestry populations; LD, linkage disequilibrium; SE, standard error.

a. Effect allele/other allele.

Appendix 3: Associations of selected variants used for risk score in African-ancestry women.

Variant	CHR	Position (bd38)	Allele[a]	EAF	Beta	Se	P	Index variant	LD in EUR (r^2)
rs2506885[b]	1	10520994	A/T	0.89	0.05	0.03	0.04	rs2506885	1
rs12375[c]	1	10536284	T/C	0.10	-0.06	0.03	0.02	rs2506885	0.70
rs34167614	1	17401576	A/G	0.12	-0.06	0.03	0.02	rs6586541	0.90
rs7515948	1	41613375	T/G	0.40	-0.05	0.02	1.99E-03	rs10749837	0.97
rs1707302	1	46135245	A/G	0.39	-0.03	0.02	0.08	rs1707302	1
rs2066319	1	50994608	A/C	0.59	0.03	0.02	0.03	rs11588271	0.96
rs11264372	1	155443139	A/G	0.84	-0.05	0.02	0.04	rs12091730	0.87
rs6427303	1	156181635	T/G	0.43	-0.03	0.02	0.07	rs11264454	0.96
rs67931591	1	215156949	G/GCTGAGGCAGGAGA	0.29	-0.03	0.02	0.05	rs67931591	1
rs12075072	1	217029330	T/C	0.74	0.04	0.02	0.02	rs11117758	0.99
rs11684853	2	19111157	T/G	0.41	-0.04	0.02	0.01	rs11684853	1
rs10637593	2	25176684	CTG/C	0.41	0.03	0.02	0.04	rs10637593	1
rs56158184	2	28973050	T/C	0.91	0.05	0.03	0.07	rs56158184	1
rs727477	2	40299988	T/G	0.41	0.04	0.02	0.03	rs727477	1
rs1430782	2	67647946	A/C	0.67	-0.03	0.02	0.04	rs9712235	0.99
rs6757464	2	69183301	C/G	0.85	0.05	0.02	0.05	rs62134416	0.73
rs3833441	2	111168154	CTTATGTT/C	0.79	-0.03	0.02	0.09	rs73954922	0.97
rs76664032	2	118823485	A/G	0.81	0.07	0.02	3.62E-04	Sentinel in aim 1	NA
rs6750813	2	120501624	T/C	0.17	-0.12	0.02	7.51E-09	Sentinel in aim 1	NA
rs2010610	2	173346180	C/G	0.29	-0.05	0.02	0.01	rs2010610	1
rs3769821	2	201258707	T/C	0.42	-0.05	0.02	1.63E-03	rs3769821	1
rs2372943	2	217039053	A/G	0.14	-0.12	0.02	3.31E-07	Sentinel in aim 1	NA
rs16857609	2	217431785	T/C	0.25	0.06	0.02	8.61E-04	rs16857609	1
rs4266007	2	226370592	A/G	0.45	-0.07	0.02	1.03E-05	rs12479355	0.80
rs6762558	3	4700567	A/G	0.70	-0.07	0.02	4.26E-05	rs6762558	1
rs9868094	3	16730246	A/G	0.32	0.05	0.02	0.01	rs9868094	1

Appendix 3. Continued

rs1352944	3	27332610	A/C	0.10	-0.06	0.03	0.05	rs552647	0.97
rs35263707	3	30642605	A/G	0.32	0.04	0.02	0.01	rs12493607	0.92
rs6787229	3	46847697	A/G	0.25	-0.04	0.02	0.02	rs559989662	0.98
rs9833888	3	100004736	T/G	0.09	0.08	0.03	4.22E-03	rs9833888	1
rs75942495	3	150760090	A/G	0.18	-0.04	0.02	0.04	rs73006998	0.92
rs75575928	3	172560023	T/C	0.16	0.05	0.02	0.03	rs62282635	0.93
rs16342[d]	4	1983005	T/TAACA	0.80	-0.04	0.02	0.08	rs495367	0.92
rs61751053	4	105613442	T/C	0.01	0.40	0.07	1.22E-08	Sentinel in aim 1	NA
rs6826366	4	174924959	A/G	0.47	-0.03	0.02	0.09	rs7697216	1
rs6863730	5	324888	T/G	0.45	0.03	0.02	0.05	rs62641919	0.94
rs10069690	5	1279675	T/C	0.61	0.13	0.02	1.92E-16	Sentinel in aim 1	NA
rs4702131	5	16233510	T/C	0.79	0.05	0.02	0.02	rs4702131	1
rs4866905	5	44555765	T/C	0.38	0.03	0.02	0.04	rs7710996	0.99
rs930395	5	44822356	A/G	0.19	0.05	0.02	0.03	rs10941679	0.76
rs59957907	5	56731413	A/G	0.82	-0.07	0.02	5.54E-04	rs59957907	1
rs35712350	5	91383702	T/C	0.45	-0.05	0.02	4.84E-03	rs1895449	0.97
rs2522057	5	132466255	C/G	0.86	-0.06	0.02	0.01	rs2522057	1
rs1432679	5	158817075	T/C	0.19	-0.10	0.02	8.82E-07	rs1432679	1
rs58242049	5	170144460	A/G	0.18	0.05	0.02	0.01	rs56234354	0.97
rs405447	6	13712867	A/G	0.32	0.05	0.02	4.58E-03	rs405447	1
rs10484439	6	26309680	A/G	0.13	-0.05	0.02	0.02	rs71557345	0.74
rs12207986	6	80384570	A/G	0.10	0.05	0.03	0.07	rs1836962	0.98
rs35240111	6	151729388	C/G	0.34	0.09	0.02	1.25E-08	Sentinel in aim 1	NA
rs910416	6	152111767	T/C	0.54	0.06	0.02	1.73E-04	rs910416	1
rs2516603	6	167973374	C/G	0.98	-0.11	0.07	0.08	rs3778663	1
rs6940159	6	170017397	T/C	0.41	-0.03	0.02	0.08	rs6940159	1
rs17167582	7	13850781	T/C	0.44	0.04	0.02	0.02	rs17167576	0.71
rs10273849	7	92107807	T/C	0.47	-0.03	0.02	0.04	rs2075881	1

Appendix 3. Continued

rs35818859[d]	7	94573736	A/T	0.92	-0.05	0.03	0.09	rs2188648	0.90
rs68056147[d]	7	130989722	A/G	0.11	0.11	0.03	1.12E-04	rs68056147	1
rs11977670	7	140242504	A/G	0.27	0.05	0.02	0.01	rs11977670	1
rs7822515	8	206504	A/G	0.18	0.04	0.02	0.04	rs116426014	0.81
rs310291	8	23801998	A/G	0.57	0.03	0.02	0.03	rs310291	1
rs13256025	8	25974262	T/C	0.07	0.07	0.03	0.02	rs13256025	1
rs7463114	8	29649578	T/C	0.39	0.05	0.02	2.02E-03	rs7463114	1
rs7816345	8	36988591	T/C	0.40	-0.04	0.02	0.01	rs75772194	0.99
rs72658071	8	75393550	A/T	0.94	-0.13	0.03	1.07E-04	rs72658071	1
rs2860518	8	100225892	T/G	0.46	0.03	0.02	0.08	rs2849506	0.94
rs58847541	8	123597926	A/G	0.30	0.04	0.02	0.01	rs58847541	1
rs13279803	8	123738281	T/C	0.30	0.03	0.02	0.09	rs13279803	1
rs10096351	8	127359926	A/G	0.38	-0.04	0.02	0.01	rs10096351	1
rs1121948	8	128152810	A/G	0.77	-0.05	0.02	0.01	rs1121948	1
rs9942894	9	104152624	A/G	0.24	-0.04	0.02	0.02	rs10820600	0.99
rs60037937	9	107541527	T/TAA	0.17	0.05	0.02	0.01	rs60037937	1
rs7862747	9	108130619	A/C	0.77	0.04	0.02	0.02	rs7862747	1
rs68088353	10	9048153	T/C	0.88	-0.06	0.02	0.01	rs35781392	0.82
rs6482189	10	21600209	A/G	0.59	0.03	0.02	0.04	rs7098100	0.74
rs10995187	10	62513267	A/G	0.07	-0.06	0.03	0.06	rs10995201	0.94
rs704010	10	79081391	T/C	0.09	0.05	0.03	0.07	rs704010	1
rs12250948	10	113368732	T/C	0.75	0.05	0.02	4.87E-03	rs12250948	1
rs9420318	10	121333668	A/G	0.53	-0.04	0.02	0.01	rs9420318	1
rs17542768	10	121578300	A/G	0.96	0.22	0.04	2.22E-08	Sentinel in aim 1	NA
rs588321	11	1875727	C/G	0.20	-0.04	0.02	0.07	rs588321	1
rs10838267	11	44347342	A/G	0.14	0.06	0.02	0.02	rs4755816	0.94
rs78540526	11	69516650	T/C	0.01	0.14	0.08	0.08	rs78540526	1
rs228606	11	108217120	T/G	0.16	-0.04	0.02	0.05	rs199504893	0.98

Appendix 3. Continued

rs145400227	11	116960788	CAGTAAA/C	0.11	0.05	0.03	0.04	rs36028244	0.85
rs10894076	11	129593165	T/C	0.82	0.05	0.02	0.01	rs10894076	1
rs7297051[d]	12	28021884	T/C	0.13	-0.06	0.02	0.02	rs7297051	1
rs17356907	12	95633983	A/G	0.81	0.05	0.02	0.01	rs17356907	1
rs1292011	12	115398717	A/G	0.55	0.03	0.02	0.05	rs1292011	1
rs9530172	13	73240055	A/G	0.07	0.05	0.03	0.10	rs17181761	1.00
rs4575439	14	36682738	A/G	0.41	-0.07	0.02	1.35E-05	Sentinel in aim 1	NA
rs739874	14	68507693	A/G	0.95	0.09	0.04	0.02	rs10483813	0.72
rs72699870	14	92650006	T/C	0.80	0.04	0.02	0.05	rs78440108	0.82
rs8011461	14	104751584	A/G	0.17	-0.05	0.02	0.02	rs4983544	0.98
rs12900028	15	74373686	C/G	0.14	0.05	0.02	0.04	rs1484216	0.73
rs6938	15	74843920	C/G	0.83	-0.07	0.02	1.50E-03	rs1869959	0.78
rs4486847	15	75436190	C/G	0.45	0.07	0.02	1.19E-05	rs8035987	0.93
rs2290203	15	90968837	A/G	0.42	-0.06	0.02	4.03E-04	rs2290203	1
rs4784227	16	52565276	T/C	0.07	0.22	0.03	2.75E-11	Sentinel in aim 1	NA
rs7190396	16	53788590	T/G	0.53	0.07	0.02	7.95E-06	rs62048402	0.93
rs3893264	16	54649890	T/C	0.82	-0.05	0.02	0.02	rs16953806	0.81
rs76535198	16	71858595	A/C	0.91	0.07	0.03	0.02	rs76535198	1
rs8056731	16	80615933	A/G	0.44	-0.05	0.02	1.79E-03	rs12446424	0.87
rs551011992[d]	17	45851263	G/GCACA	0.13	-0.06	0.03	0.03	rs572771346	0.83
rs35051208	17	54975207	T/C	0.08	-0.08	0.03	0.01	rs244373	0.75
rs10853615	18	24058545	A/C	0.19	0.08	0.02	1.95E-04	Sentinel in aim 1	NA
rs521667	18	26758515	A/C	0.16	-0.05	0.02	0.03	rs527616	0.94
rs2307561	18	26923542	A/AAGTGTT	0.25	-0.04	0.02	0.03	rs2307561	1
rs16963205	18	32350613	T/C	0.76	0.12	0.02	4.39E-11	Sentinel in aim 1	NA
rs9954058	18	44831838	C/G	0.31	-0.05	0.02	2.83E-03	rs9954058	1
rs56069439	19	17283116	A/C	0.24	0.12	0.02	9.07E-12	Sentinel in aim 1	NA
rs56681946	19	43778879	T/C	0.77	-0.06	0.02	2.73E-03	rs56681946	1

Appendix 3. Continued

rs1319363	20	33954070	A/T	0.82	-0.05	0.02	0.01	rs2284378	0.95
rs6020465	20	50346469	A/G	0.71	-0.04	0.02	0.02	rs6012911	0.99
rs531874019	20	53677557	G/GT	0.90	-0.06	0.03	0.04	rs531874019	1
rs11088317	21	15201802	T/C	0.06	-0.08	0.04	0.03	rs2403907	0.90
rs12628403[d]	22	38962032	A/C	0.98	-0.15	0.06	0.01	rs12628403	1
rs66987842	22	40508703	CT/C	0.88	-0.05	0.02	0.05	rs66987842	1

Abbreviations: CHR, chromosome; EAF, effect allele frequency; EUR, European-ancestry populations; LD, linkage disequilibrium; SE, standard error.
a. Effect allele/other allele; b. Unavailable in testing set and rs12375 was used as proxy; c. Proxy for rs2506885; d. Unavailable in testing set.

Appendix 4: List of reported variants used for risk score with weights from European-ancestry women.

Variant[a]	CHR	Position (bd38)	Allele[b]	Beta[c]	Included[d]
1_7917076_G_A	1	7857016	A/G	-0.04	Yes
1_10566215_A_G	1	10506158	G/A	-0.06	Yes
1_18807339_T_C	1	18480845	C/T	-0.06	Yes
1_41380440_C_T	1	40914768	T/C	0.04	Yes
1_41389220_T_C	1	40923548	C/T	0.16	Yes
1_46670206_TC_T	1	46204534	T/TC	0.04	Yes
1_51467096_CT_C	1	51001424	C/CT	0.04	No
1_88156923_G_A	1	87691240	A/G	0.05	Yes
1_88428199_C_A	1	87962516	A/C	-0.04	Yes
1_100880328_A_T	1	100414772	T/A	0.04	Yes
1_110198129_CAAA_C	1	109655507	C/CAAA	0.05	No
1_114445880_G_A	1	113903258	A/G	0.06	Yes
1_118141492_A_C	1	117598870	C/A	0.05	Yes
1_120257110_T_C	1	119714487	C/T	0.04	Yes
1_121280613_A_G	1	121538815	G/A	0.09	Yes
1_121287994_A_G	1	121546196	G/A	-0.07	No
1_145604302_C_CT	1	145830809	CT/C	-0.04	No
1_149906413_T_C	1	149934520	C/T	0.05	Yes
1_155556971_G_A	1	155587180	A/G	0.05	Yes
1_168171052_CA_C	1	168201814	C/CA	-0.07	No
1_172328767_T_TA	1	172359627	TA/T	-0.04	No
1_201437832_C_T	1	201468704	T/C	0.09	Yes
1_202184600_C_T	1	202215472	T/C	-0.01	Yes
1_203770448_T_A	1	203801320	A/T	0.05	Yes
1_204502514_T_TTCTGAAACAGGG	1	204533386	TTCTGAAACAGGG/T	-0.03	No
1_208076291_G_A	1	207902946	A/G	-0.04	Yes
1_217053815_T_G	1	216880473	G/T	0.04	Yes
1_217220574_G_A	1	217047232	A/G	-0.04	Yes
1_220671050_C_T	1	220497708	T/C	0.04	Yes
1_242034263_A_G	1	241870961	G/A	0.14	Yes
2_10138983_T_C	2	9998855	C/T	0.06	No
2_19315675_T_A	2	19115914	A/T	-0.03	Yes
2_25129473_A_G	2	24906604	G/A	-0.04	Yes

Appendix 4. Continued

2_29179452_G_C	2	28956586	C/G	-0.01	Yes
2_29615233_T_C	2	29392367	C/T	-0.04	Yes
2_39699510_C_CT	2	39472369	CT/C	-0.04	No
2_70172587_G_A	2	69945455	A/G	-0.04	Yes
2_88358825_G_C	2	88059306	C/G	0.05	Yes
2_121058254_A_G	2	120300678	G/A	-0.03	Yes
2_121089731_T_C	2	120332155	C/T	-0.04	Yes
2_121159205_G_A	2	120401629	A/G	-0.04	Yes
2_121246568_T_C	2	120488992	C/T	0.10	Yes
2_172974566_C_G	2	172109838	G/C	-0.05	Yes
2_174212910_A_G	2	173348182	G/A	0.06	Yes
2_192381934_C_T	2	191517208	T/C	0.03	Yes
2_202204741_T_C	2	201340018	C/T	-0.05	Yes
2_217920769_G_T	2	217056046	T/G	-0.13	Yes
2_217955896_GA_G	2	217091173	G/GA	-0.20	No
2_218292158_C_G	2	217427435	G/C	-0.08	Yes
2_218714845_G_A	2	217850122	A/G	-0.04	Yes
2_241388857_C_A	2	240449440	A/C	-0.12	No
3_4742251_A_G	3	4700567	G/A	0.06	Yes
3_27353716_C_A	3	27312225	A/C	0.07	Yes
3_27388664_C_G	3	27347173	G/C	0.05	Yes
3_29294845_C_T	3	29253354	T/C	-0.13	Yes
3_30684907_C_T	3	30643415	T/C	0.06	Yes
3_46888198_T_C	3	46846708	C/T	-0.08	Yes
3_49709912_C_CT	3	49672479	CT/C	-0.04	No
3_55970777_A_AT	3	55936749	AT/A	-0.12	Yes
3_59373745_C_T	3	59388019	T/C	-0.04	Yes
3_63887449_T_TTG	3	63901773	TTG/T	0.06	No
3_71620370_T_G	3	71571219	G/T	-0.04	Yes
3_87037543_A_G	3	86988393	G/A	-0.07	Yes
3_99403877_G_A	3	99685033	A/G	-0.04	Yes
3_141112859_CTT_C	3	141394017	C/CTT	0.06	No
3_172285237_G_A	3	172567447	A/G	0.04	Yes
3_189774456_C_T	3	190056667	T/C	-0.05	Yes

Appendix 4. Continued

4_38784633_G_T	4	38783012	T/G	0.05	Yes
4_84370124_TAA_TA	4	83448971	TA/TAA	-0.05	No
4_89240476_G_A	4	88319324	A/G	0.04	Yes
4_92594859_TTCTTTC_T	4	91673708	T/TTCTTTC	-0.04	No
4_106069013_G_T	4	105147856	T/G	0.05	No
4_126752992_A_AAT	4	125831837	AAT/A	-0.04	No
4_143467195_C_T	4	142546042	T/C	-0.06	Yes
4_151218296_CATATTT_C	4	150297144	C/CATATTT	0.04	Yes
4_175842495_G_A	4	174921344	A/G	-0.09	Yes
4_175847436_C_A	4	174926285	A/C	0.03	No
4_187503758_A_T	4	186582604	T/A	0.04	No
5_345109_T_C	5	344994	C/T	0.08	No
5_1279790_C_T	5	1279675	T/C	0.06	Yes
5_1296255_A_AG	5	1296140	AG/A	-0.05	Yes
5_1353077_T_C	5	1352962	C/T	0.16	Yes
5_2777029_G_A	5	2776915	A/G	0.04	Yes
5_16231194_G_C	5	16231085	C/G	-0.04	Yes
5_32579616_TCA_T	5	32579510	T/TCA	0.04	Yes
5_44508264_G_GT	5	44508162	GT/G	-0.12	No
5_44619502_A_G	5	44619400	G/A	-0.11	Yes
5_44649944_C_T	5	44649842	T/C	0.05	Yes
5_44706498_A_G	5	44706396	G/A	0.05	Yes
5_44853593_G_C	5	44853491	C/G	-0.03	Yes
5_52679539_C_CA	5	53383709	CA/C	0.06	No
5_55662540_C_CT	5	56366713	CT/C	-0.05	No
5_55965167_C_T	5	56669340	T/C	0.04	Yes
5_56023083_T_G	5	56727256	G/T	0.14	Yes
5_56042972_C_T	5	56747145	T/C	0.09	Yes
5_56045081_T_C	5	56749254	C/T	-0.06	Yes
5_58241712_C_T	5	58945885	T/C	-0.04	No
5_71965007_G_A	5	72669180	A/G	-0.04	No
5_73234583_T_C	5	73938758	C/T	-0.04	Yes
5_77155397_GT_G	5	77859573	G/GT	-0.04	Yes
5_79180995_G_GA	5	79885172	GA/G	0.03	No

Appendix 4. Continued

5_81512947_TA_T	5	82217128	T/TA	-0.06	Yes
5_90789470_G_A	5	91493653	A/G	-0.06	Yes
5_104300273_G_T	5	104964572	T/G	-0.05	Yes
5_122478676_C_A	5	123142982	A/C	-0.04	Yes
5_122705244_C_T	5	123369550	T/C	0.09	Yes
5_131640536_A_G	5	132304843	G/A	0.04	Yes
5_132407058_C_T	5	133071366	T/C	-0.04	Yes
5_158244083_C_T	5	158817075	T/C	-0.07	Yes
5_169591460_T_C	5	170164456	C/T	0.04	Yes
5_173358154_G_A	5	173931151	A/G	0.04	Yes
5_176134882_T_C	5	176707881	C/T	0.04	No
6_13713366_G_C	6	13713134	C/G	-0.06	Yes
6_16399557_C_T	6	16399326	T/C	-0.04	Yes
6_18783140_G_A	6	18782909	A/G	0.03	Yes
6_20537845_CA_C	6	20537614	C/CA	-0.04	No
6_21923810_T_C	6	21923579	C/T	-0.03	Yes
6_27425644_G_C	6	27457865	C/G	-0.07	Yes
6_43227141_G_A	6	43259403	A/G	-0.06	Yes
6_82263549_AAT_A	6	81553832	A/AAT	0.05	No
6_85912194_CAA_C	6	85202476	C/CAA	0.08	Yes
6_87803819_T_C	6	87094101	C/T	0.04	No
6_130341728_C_CT	6	130020583	CT/C	0.05	Yes
6_149595505_T_C	6	149274369	C/T	-0.05	Yes
6_151949806_A_C	6	151628671	C/A	0.07	Yes
6_151955914_A_G	6	151634779	G/A	0.14	No
6_152022664_CAAAAAAA_C	6	151701529	C/CAAAAAAA	0.01	No
6_152023191_G_A	6	151702056	A/G	0.06	Yes
6_152055978_A_T	6	151734843	T/A	0.07	Yes
6_152432902_C_T	6	152111767	T/C	0.06	Yes
6_169006947_C_G	6	168606267	G/C	-0.03	Yes
6_170332621_T_C	6	170017397	C/T	0.04	Yes
7_21940960_A_G	7	21901342	G/A	-0.05	Yes
7_25569548_C_T	7	25529928	T/C	-0.05	Yes
7_28869017_G_A	7	28829400	A/G	-0.06	Yes

Appendix 4. Continued

7_55192256_A_C	7	55124563	C/A	-0.03	No
7_91459189_A_ATT	7	91829875	ATT/A	0.05	No
7_94113799_T_C	7	94484487	C/T	0.04	Yes
7_98005235_G_A	7	98375923	A/G	-0.05	Yes
7_99948655_T_G	7	100351032	G/T	0.04	No
7_101552440_G_A	7	101909160	A/G	-0.06	No
7_102481842_T_C	7	102841395	C/T	0.04	Yes
7_130656911_C_T	7	130972152	T/C	-0.05	Yes
7_130674481_G_A	7	130989722	A/G	0.04	No
7_139943702_CT_C	7	140243902	C/CT	0.06	No
7_144048902_G_T	7	144351809	T/G	-0.06	No
8_170692_T_C	8	220692	C/T	0.05	Yes
8_17787610_CT_C	8	17930101	C/CT	-0.04	No
8_23447496_A_G	8	23589983	G/A	-0.04	Yes
8_23663653_C_A	8	23806140	A/C	0.03	Yes
8_29509616_A_C	8	29652100	C/A	-0.06	Yes
8_36858483_A_G	8	37000965	G/A	-0.08	Yes
8_76230943_A_G	8	75318708	G/A	0.08	Yes
8_76333056_C_T	8	75420821	T/C	0.11	Yes
8_76378165_G_T	8	75465930	T/G	-0.04	Yes
8_102483100_T_C	8	101470872	C/T	0.06	Yes
8_106358620_A_T	8	105346392	T/A	-0.07	Yes
8_117209548_A_G	8	116197325	G/A	-0.04	Yes
8_120862186_A_G	8	119849946	G/A	0.05	Yes
8_124563705_T_C	8	123551465	C/T	0.05	Yes
8_124571581_G_A	8	123559341	A/G	0.03	Yes
8_124739913_T_G	8	123727673	G/T	0.05	Yes
8_128213561_C_CA	8	127201316	CA/C	-0.04	Yes
8_128370949_C_G	8	127358703	G/C	0.06	Yes
8_128372172_A_G	8	127359926	G/A	0.06	Yes
8_129199566_G_A	8	128187320	A/G	0.06	Yes
8_143669254_A_G	8	142587893	G/A	-0.03	Yes
9_6880263_A_G	9	6880263	G/A	0.03	Yes
9_21964882_CAAAA_C	9	21964883	C/CAAAA	0.06	No

Appendix 4. Continued

9_22041998_C_G	9	22041999	G/C	0.03	Yes
9_36928288_T_C	9	36928291	C/T	0.02	Yes
9_87782211_T_C	9	85167296	C/T	0.04	Yes
9_98362587_T_C	9	95600305	C/T	0.06	Yes
9_110303808_TAA_T	9	107541527	T/TAA	0.08	Yes
9_110837073_A_G	9	108074792	G/A	0.12	Yes
9_110837176_C_T	9	108074895	T/C	0.07	Yes
9_110849525_G_T	9	108087244	T/G	0.02	Yes
9_110885479_C_T	9	108123199	T/C	0.09	Yes
9_119313486_A_G	9	116551207	G/A	-0.05	Yes
9_129424719_A_G	9	126662440	G/A	-0.04	Yes
9_136146597_C_T	9	133271182	T/C	0.04	Yes
10_5794652_A_G	10	5752689	G/A	0.05	Yes
10_13892298_G_A	10	13850298	A/G	0.04	Yes
10_22032942_A_G	10	21744013	G/A	-0.06	Yes
10_22477776_ACC_A	10	22188847	A/ACC	0.17	No
10_22861490_A_C	10	22572561	C/A	0.09	No
10_38523626_C_A	10	38234698	A/C	0.04	No
10_64299890_A_G	10	62540131	G/A	-0.13	Yes
10_64819996_G_T	10	63060236	T/G	0.05	Yes
10_71335574_C_T	10	69575818	T/C	-0.04	No
10_80851257_G_T	10	79091500	T/G	-0.08	Yes
10_80886726_A_G	10	79126969	G/A	0.08	Yes
10_95292187_CAA_C	10	93532430	C/CAA	-0.05	No
10_114777670_C_T	10	113017911	T/C	0.05	Yes
10_115128491_T_C	10	113368732	C/T	-0.06	Yes
10_123095209_G_A	10	121335695	A/G	-0.05	No
10_123340107_A_G	10	121580593	G/A	0.15	Yes
10_123340431_GC_G	10	121580917	G/GC	-0.24	Yes
10_123349324_A_T	10	121589810	T/A	-0.26	Yes
11_433617_T_C	11	433617	C/T	-0.04	No
11_803017_A_G	11	803017	G/A	0.05	Yes
11_1895708_C_A	11	1874478	A/C	-0.08	Yes
11_18664241_T_G	11	18642694	G/T	0.05	Yes

Appendix 4. Continued

11_42844441_C_T	11	42822891	T/C	-0.03	Yes
11_44368892_G_A	11	44347342	A/G	0.04	Yes
11_46318032_C_G	11	46296481	G/C	-0.07	Yes
11_65553492_C_A	11	65786021	A/C	0.04	Yes
11_65572431_G_A	11	65804960	A/G	-0.03	Yes
11_69328130_A_T	11	69513362	T/A	-0.04	Yes
11_69330983_G_A	11	69516215	A/G	0.10	Yes
11_69331418_C_T	11	69516650	T/C	0.18	Yes
11_103614438_T_G	11	103743710	G/T	0.01	Yes
11_108267402_C_CA	11	108396675	CA/C	0.00	No
11_111696440_T_C	11	111825716	C/T	-0.04	Yes
11_116727936_A_T	11	116857220	T/A	-0.04	Yes
11_122966626_A_G	11	123095918	G/A	-0.04	Yes
11_129243417_T_G	11	129373522	G/T	-0.05	Yes
11_129461016_A_G	11	129591121	G/A	0.05	Yes
12_293626_A_G	12	184460	G/A	0.04	Yes
12_14413931_G_C	12	14260997	C/G	0.05	Yes
12_28149568_C_T	12	27996635	T/C	-0.06	Yes
12_28174817_C_T	12	28021884	T/C	-0.09	Yes
12_28347382_C_T	12	28194449	T/C	-0.05	Yes
12_29140260_G_A	12	28987327	A/G	0.06	No
12_57146069_T_G	12	56752285	G/T	-0.06	No
12_70798355_A_T	12	70404575	T/A	0.05	Yes
12_83064195_G_GA	12	82670416	GA/G	0.07	Yes
12_85004551_C_T	12	84610772	T/C	0.03	Yes
12_96027759_A_G	12	95633983	G/A	-0.09	Yes
12_103097887_C_T	12	102704109	T/C	0.05	Yes
12_111600134_G_T	12	111162330	T/G	-0.04	Yes
12_115108136_T_C	12	114670331	C/T	0.05	Yes
12_115796577_A_G	12	115358772	G/A	-0.04	Yes
12_115835836_T_C	12	115398031	C/T	-0.08	Yes
12_120832146_C_T	12	120394343	T/C	0.05	Yes
13_32839990_G_A	13	32265853	A/G	0.04	Yes
13_32972626_A_T	13	32398489	T/A	0.27	Yes

Appendix 4. Continued

13_43501356_A_G	13	42927220	G/A	0.05	Yes
13_73806982_T_C	13	73232845	C/T	0.03	Yes
13_73960952_A_G	13	73386815	G/A	0.04	Yes
14_37128564_C_A	14	36659359	A/C	-0.07	Yes
14_37228504_C_T	14	36759299	T/C	0.04	Yes
14_68660428_T_C	14	68193711	C/T	-0.05	Yes
14_68979835_T_C	14	68513118	C/T	-0.09	Yes
14_91751788_TC_T	14	91285444	T/TC	0.04	Yes
14_91841069_A_G	14	91374725	G/A	0.05	Yes
14_93070286_C_T	14	92603941	T/C	-0.06	Yes
14_105213978_T_G	14	104747641	G/T	0.04	Yes
15_46680811_C_A	15	46388613	A/C	-0.20	Yes
15_50694306_A_G	15	50402109	G/A	-0.04	Yes
15_66630569_G_A	15	66338231	A/G	-0.04	Yes
15_67457698_A_G	15	67165360	G/A	0.08	Yes
15_75750383_T_C	15	75458042	C/T	-0.04	Yes
15_91512267_G_T	15	90969037	T/G	-0.06	Yes
15_100905819_A_C	15	100365614	C/A	-0.06	Yes
16_4008542_CAAAAA_C	16	3958541	C/CAAAAA	-0.03	No
16_4106788_C_A	16	4056787	A/C	-0.03	Yes
16_6963972_C_G	16	6913971	G/C	0.04	Yes
16_10706580_G_A	16	10612723	A/G	-0.07	Yes
16_23007047_G_T	16	22995726	T/G	0.12	Yes
16_52538825_C_A	16	52504913	A/C	0.11	Yes
16_52599188_C_T	16	52565276	T/C	0.11	Yes
16_53809123_C_T	16	53775211	T/C	-0.07	Yes
16_53861139_C_T	16	53827227	T/C	-0.03	Yes
16_53861592_G_A	16	53827680	A/G	-0.03	Yes
16_54682064_G_A	16	54648152	A/G	0.05	Yes
16_80648296_A_G	16	80614399	G/A	0.08	Yes
16_85145977_T_C	16	85112371	C/T	-0.02	No
16_87086492_T_C	16	87052886	C/T	-0.05	Yes
17_29168077_G_T	17	30841059	T/G	-0.06	No
17_39251123_T_C	17	41094871	C/T	0.08	Yes

Appendix 4. Continued

17_40127060_T_C	17	41975042	C/T	0.02	Yes
17_40485239_G_T	17	42333221	T/G	-0.06	Yes
17_40744470_G_A	17	42592452	A/G	0.20	Yes
17_43212339_C_CT	17	45134972	CT/C	0.04	No
17_44283858_G_A	17	46206492	A/G	-0.05	No
17_53209774_A_C	17	55132413	C/A	-0.08	No
17_77781725_A_G	17	79807926	G/A	-0.04	Yes
18_11696613_C_T	18	11696614	T/C	-0.04	No
18_20634253_C_T	18	23054290	T/C	-0.04	Yes
18_24125857_T_C	18	26545893	C/T	0.03	Yes
18_24337424_C_G	18	26757460	G/C	0.05	Yes
18_24518050_AT_A	18	26938086	A/AT	-0.06	Yes
18_25407513_C_G	18	27827549	G/C	0.04	Yes
18_29981526_G_A	18	32401563	A/G	-0.11	Yes
18_42411803_G_C	18	44831838	C/G	-0.09	Yes
18_42888797_T_C	18	45308832	C/T	-0.05	Yes
19_13249921_G_T	19	13139107	T/G	0.10	Yes
19_17393925_C_A	19	17283116	A/C	0.04	Yes
19_18569492_C_T	19	18458682	T/C	-0.07	Yes
19_19517054_C_CGGGCG	19	19406245	CGGGCG/C	0.04	No
19_44283031_T_C	19	43778879	C/T	0.06	Yes
19_46166073_T_C	19	45662815	C/T	-0.04	Yes
19_55816678_C_T	19	55305310	T/C	-0.04	Yes
20_5948227_G_A	20	5967581	A/G	0.08	Yes
20_11379842_T_C	20	11399194	C/T	0.08	No
20_41613706_C_G	20	42985066	G/C	0.03	Yes
20_52296849_G_A	20	53680310	A/G	0.04	Yes
21_16364756_T_G	21	14992435	G/T	0.06	Yes
21_16566350_A_G	21	15194030	G/A	0.06	Yes
21_16574455_C_A	21	15202135	A/C	-0.07	Yes
21_47762932_G_A	21	46343018	A/G	0.09	Yes
22_19766137_C_T	22	19778614	T/C	-0.04	Yes
22_29121087_A_G	22	28725099	G/A	0.18	No
22_29135543_G_A	22	28739555	A/G	0.07	Yes

Appendix 4. Continued

22_29203724_C_T	22	28807736	T/C	0.14	Yes
22_29551872_A_G	22	29155884	G/A	-0.17	No
22_38583315_AAAAG_AAAAGAAAG	22	38187308	AAAAGAAAG/AAAAG	-0.05	No
22_39343916_T_A	22	38947911	A/T	0.04	Yes
22_40904707_CT_C	22	40508703	C/CT	0.11	Yes
22_43433100_C_T	22	43037094	T/C	-0.06	Yes
22_45319953_G_A	22	44924073	A/G	-0.01	No
22_46283297_G_A	22	45887417	A/G	0.07	Yes

Abbreviations: CHR, chromosome; EAF, effect allele frequency; EUR, European-ancestry populations; LD, linkage disequilibrium; SE, standard error.
a. Variant names coded using chromosome, position bd37, and alleles; b. Effect allele/other allele; c. Beta were reported for effect alleles by Mavaddat et al. in 2019; d. Of the 313 reported variants, 236 variants were used in the aim 2.