Research Article

Cancer Epidemiology, Biomarkers & Prevention

Germline Variation and Breast Cancer Incidence: A Gene-Based Association Study and Whole-Genome Prediction of Early-Onset Breast Cancer

Molly Scannell Bryan[1,2], Maria Argos[2], Irene L. Andrulis[3], John L. Hopper[4], Jenny Chang-Claude[5,6], Kathleen E. Malone[7], Esther M. John[8,9], Marilie D. Gammon[10], Mary B. Daly[11], Mary Beth Terry[12], Saundra S. Buys[13], Dezheng Huo[1], Olofunmilayo I. Olopade[1], Jeanine M. Genkinger[12], Alice S. Whittemore[14], Farzana Jasmine[1], Muhammad G. Kibriya[1], Lin S. Chen[1], and Habibul Ahsan[1]

Abstract

Background: Although germline genetics influences breast cancer incidence, published research only explains approximately half of the expected association. Moreover, the accuracy of prediction models remains low. For women who develop breast cancer early, the genetic architecture is less established.

Methods: To identify loci associated with early-onset breast cancer, gene-based tests were carried out using exome array data from 3,479 women with breast cancer diagnosed before age 50 and 973 age-matched controls. Replication was undertaken in a population that developed breast cancer at all ages of onset.

Results: Three gene regions were associated with breast cancer incidence: *FGFR2* ($P = 1.23 \times 10^{-5}$; replication $P < 1.00 \times 10^{-6}$), *NEK10* ($P = 3.57 \times 10^{-4}$; replication $P < 1.00 \times 10^{-6}$), and *SIVA1* ($P = 5.49 \times 10^{-4}$; replication $P < 1.00 \times 10^{-6}$). Of the 151 gene regions reported in previous literature, 19 (12.5%) showed evidence of association ($P < 0.05$) with the risk of early-onset breast cancer in the early-onset population. To predict incidence, whole-genome prediction was implemented on a subset of 3,076 participants who were additionally genotyped on a genome wide array. The whole-genome prediction outperformed a polygenic risk score [AUC, 0.636; 95% confidence interval (CI), 0.614–0.659 compared with 0.601; 95% CI, 0.578–0.623], and when combined with known epidemiologic risk factors, the AUC rose to 0.662 (95% CI, 0.640–0.684).

Conclusions: This research supports a role for variation within *FGFR2* and *NEK10* in breast cancer incidence, and suggests *SIVA1* as a novel risk locus.

Impact: This analysis supports a shared genetic etiology between women with early- and late-onset breast cancer, and suggests whole-genome data can improve risk assessment. *Cancer Epidemiol Biomarkers Prev; 27(9); 1057–64. ©2018 AACR.*

Introduction

Germline genetic variation is an established risk factor for breast cancer. The identified risk loci have implicated molecular processes involved in oncogenesis, and suggested targets for preventative efforts (1, 2). However, the risk loci that have been identified only contribute about half of the total expected risk due to genetics (3, 4). For women who are diagnosed with breast cancer before age 50 (one in five of American diagnoses; ref. 5) less is known, and it remains unclear the extent to which the genetic influences on the risk of late onset disease are also risk factors in younger women.

Much of the recent research into the genetic determinants of breast cancer has focused on single-variant tests common in genome-wide association studies (GWAS), which often incorporate common genetic variation and variants that can be reliably imputed from them (4, 6, 7). While GWAS have been successful in establishing breast cancer as a complex disease

[1]Department of Public Health Sciences, University of Chicago, Chicago, Illinois. [2]University of Illinois at Chicago, Chicago, Illinois. [3]Lunefeld-Tanenbaum Research Institute, Sinai Health System and Department of Molecular Genetics, University of Toronto, Toronto, Canada. [4]University of Melbourne, Parkville, Victoria, Australia. [5]Deutsches Krebsforschungszentrum in der Helmholtz-Gemeinshaft, Heidelberg, Germany. [6]University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [7]Fred Hutchinson Cancer Research Center, Seattle, Washington. [8]Cancer Prevention Institute of California, Fremont, California. [9]Stanford Cancer Institute, Stanford, California. [10]University of North Carolina at Chapel Hill, Chapel Hill, North Carolina. [11]Fox Chase Cancer Center, Philadelphia, Pennsylvania. [12]Columbia University, New York, New York. [13]University of Utah Salt Lake City, Salt Lake City, Utah. [14]Stanford University, Stanford, California.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

Prior presentations: Portions of this manuscript have been included in the doctoral thesis of M. Scannell Bryan at the University of Chicago (under embargo until 2019), and posters presented at the Society for Epidemiologic Research June 2017 meeting and the American Society of Preventative Oncology meeting March 2017.

Corresponding Author: Molly Scannell Bryan, University of Illinois at Chicago, 1747 W. Roosevelt Rd., Chicago, IL 60608-1264. Phone: 312-355-2328; Fax: 312-996-2703; E-mail: scannemo@uic.edu

doi: 10.1158/1055-9965.EPI-17-1185

with many genetic influences, the GWAS design does not incorporate evidence from variants that are not in strong linkage disequilibrium with those measured on genome-wide arrays. In contrast, exome arrays allow for a more comprehensive interrogation of variation near coding portions of the genome. When combined with set-based tests such as SKAT-O (8), the data provided by exome arrays can detect genes in which multiple variants are associated with breast cancer even when their single associations do not meet genome-wide significance thresholds, and can also incorporate evidence from variants that are too rare to be tested individually (9). Gene-based tests that have focused on exonic variation have only been implemented in three published studies of breast cancer (10–12). Of these, none targeted women with early-onset disease, and all excluded variants from their analysis that were common or not annotated as "functional" (e.g., restricted to nonsynonymous variants), possibly limiting their findings.

In addition to identifying individual risk loci, it is also of clinical interest to incorporate genetic information into predictions of breast cancer, as a strong prediction model could better identify high-risk women who may want to choose more aggressive screening or chemoprevention, and also low-risk women who may benefit from knowledge of their lower risk. The most common family of prediction models that incorporate genetic variation are polygenic risk scores (PRS), which only include variants that meet significance thresholds (13). In contrast, whole-genome linear mixed model (LMM) prediction methods include all measured genetic variants (14). These methods have been shown to have more predictive power than PRSs for complex traits (4, 6, 7), but have not yet been applied to breast cancer incidence.

In this study, the genetic architecture of breast cancer is described in a population of women who were under the age of 50 at the time of their diagnosis. Gene-based tests were implemented to identify risk loci, and identified loci were examined for evidence of replication in a larger cohort of women who were diagnosed at all ages. In addition, a whole-genome LMM prediction of breast cancer incidence was undertaken.

Materials and Methods

Participants

Primary. The 4,914 study participants (3,876 cases and 1,038 controls) were sampled from five ongoing studies designed to assess the risk factors associated with early-onset breast cancer. Details of recruitment and data collection are found in the supplementary text. All participants were women of European ancestry, and participants were excluded if they were found to have any known pathogenic mutations in the *BRCA1* or *BRCA2* genes. Almost all of the participants (98%) were younger than 50 at the time of their diagnosis or enrollment. Informed written consent was obtained from all participants, and the research was carried out under the supervision of multiple institutional review boards, as detailed in the Supplementary Data.

Replication. Gene regions found to be suggestive in the primary study population (details below) were examined for the evidence of replication in the summary statistics provided by the participants of the 2017 meta-analysis published by Michailidou and colleagues (137,045 breast cancer cases diagnosed at all ages of onset and 119,078 controls; ref. 4). Details of the study population and analysis methods that were used to calculate these associations can be found in the Michailidou and colleagues' article (4) and the associations were retrieved from the study website in March of 2018 (15). Three percent of the cases and 0.8% of the controls in the replication were also participants of the primary study.

Genotyping

Details of the genotyping, imputation, and quality control of the primary study participants are found in the Supplementary Data. Briefly, germline DNA from the 4,914 participants was genotyped on the Illumina HumanExome array. If the variant passed variant-level quality control, was polymorphic in the study population, and was 1 kb from the transcription start or end site of a gene, it was included in the gene-based analyses (125,388 polymorphic variants in 16,813 genes). Women were excluded for genotyping rate less than 95%, sex mismatch, heterozygosity greater than 4 standard deviations from the mean, principal component outliers (first or second principal components constructed from common variants greater than 6 standard deviations from the mean), and estimated genetic relatedness closer than 0.4, resulting in 4,452 women available for the gene-based SKAT-O analyses (3,479 cases and 973 controls).

A subset (n = 3,374; 2,340 cases and 1,034 controls, demographics and geographic information in Supplementary Tables S1 and S2) was additionally genotyped using the Illumina 610-Quad or Cyto12 v2 BeadChips (555,259 variants; 3,310,148 after imputation to the 1,000 genomes phase III release; ref. 16). The same quality control steps described above were applied. For the prediction analysis, these genome-wide array variants were combined with the exome array variants using PLINK (17, 18). Fewer than 100 of the variants that were interrogated on both platforms were discordant. In those cases, the variant called by the exome array was used. The prediction analyses were carried out on the variants that could be annotated by ANNOVAR (ref. 19; 3,415,850 variants; 3,076 participants; 2,109 cases and 967 controls available for the prediction analysis).

Statistical analysis

Statistical approach for gene-based tests. To identify gene regions in which variation was associated with early-onset breast cancer incidence, gene-based logistic regressions were implemented on the 3,479 cases and 973 controls genotyped on the exome array. The analysis was carried out using the R software SKAT-O package (20, 21). Each variant was weighted by the combined annotation dependent depletion (CADD; ref. 22) score of the predicted deleteriousness of the minor allele. These weights allow for functional information about the variant to be incorporated, and no variants in the gene regions were excluded. To counter the potential for confounding between genetic ancestry and risk, EIGENSTRAT (23) was used to construct two sets of six principal components, one set from common variants [variants with minor allele frequency (MAF) above the threshold of $(\frac{1}{2n})^{\frac{1}{2}} = 0.0106$ following Wu and colleagues; ref. 24] and one set from rare variants (principal components 1 and 2 of both common and rare variants are plotted in Supplementary Fig. S1). The first principal component constructed from common variants and the second principal component constructed from rare variants were associated with

case status in a logistic regression, and were subsequently included as covariates. The threshold for single-study genome-wide significance was set at 3.0×10^{-6} (0.05, Bonferroni corrected for 16,813 genes). Gene-level summary statistics of this analysis are available in dbGaP, accession phs001589.v1.p1.

To compare the results of the gene-based tests in the primary study population to the evidence generated from common variants measured in the replication data, we converted the individual-variant summary statistics from the Michailidou meta-analysis into gene-based tests using the VEGAS method (9) for the 20 genes with the smallest *P* value in the primary early-onset analysis. The meta-analysis variants were included if they were within 1 kb of the transcription start or end site as annotated by ANNOVAR. Genes for which the VEGAS analysis gave a *P* value less than the Bonferroni-corrected threshold ($0.05/20 = 2.5 \times 10^{-3}$) were considered to be associated with breast cancer. Because of the overlap between the meta-analysis consortium and the primary study population (3.2% of the cases and 0.3% of the controls), the test statistics were not formally combined.

To determine whether the results were sensitive to the weighting method, the analyses were repeated twice: using weights that were a beta (1, 25) transformation of the MAF (as suggested by the SKAT authors; ref. 8), and using equal weights. The distribution of the per-variant beta weights is contrasted to the distribution of the per-variant CADD weights in Supplementary Fig. S2.

To establish whether any identified genes were driven by novel risk loci, single nucleotide variants (SNV) were selected for analysis from the NHGRI-EBI GWAS catalog (2, 25) if the SNV fell within any genes that were identified by the primary analysis (GWAS catalog $P < 5 \times 10^{-8}$). These genes were then reanalyzed in the early-onset primary study population, controlling for the effects of the GWAS catalog SNVs using the prepCondScores and skatOMeta functions of the skatMeta R package (20).

Finally, to examine whether there was evidence for a shared genetic etiology between early-onset breast cancer and all ages of onset represented by previously identified risk loci, in the primary study population of early-onset cases, we examined the 151 gene regions in which variation was associated with breast cancer of any age in the GWAS catalog as of March 2018 (GWAS catalog reported $P < 5 \times 10^{-8}$, gene regions of these variants listed in Supplementary Data; 14 additional gene regions were listed in the GWAS catalog that were not assayed on the exome chip).

Statistical approach for whole-genome prediction. The whole-genome LMM prediction was carried out using the genetic data from the combined whole-exome and whole-genome arrays. This analysis was carried out on the subset of participants for whom the full set of epidemiologic risk factors was available (3,076 women; 2,109 cases; and 967 controls; demographics in Supplementary Table S1). The genetic relatedness matrices were created by the GCTA software (26), and the prediction was implemented using the R package "omicKriging." (14) Using 10-fold cross validation, a probability of disease was estimated, which was used to compute an area under the receiver operating characteristic curve (AUC). This was repeated 200 times; the reported AUC is the mean of the 200 calculated AUCs and the reported 95% confidence intervals (CI) are the 2.5th and 97.5th percentiles of the replications.

The performance of Kriging prediction improves if multiple genetic relatedness matrices are constructed with each matrix containing variants with similar association strengths (14). For this reason, seven matrices were constructed. The first matrix contained the 1,137 SNVs that were either listed in the GWAS catalog as associated with breast cancer risk, or in linkage disequilibrium with them ($r^2 > 0.2$ in the 1,000 Genomes CEU reference panel; ref. 27) The other six matrices were constructed to reflect the hypothesis that variants of the same MAF and predicted functionality may have similar effect sizes: three matrices were constructed that partitioned rare variants (MAF below 0.0127) into three functional categories as defined by ANNOVAR: (1) intergenic, (2) gene regions, but not predicted to cause an amino acid change, and (3) predicted to cause an amino acid change. The final three matrices were constructed from common variants that were classified into each of these functional annotations. The weights of each of these seven matrices were empirically varied to obtain an optimal AUC, with the optimal weights and number of variants in each GRM noted in Table 2.

To compare the effectiveness of the LMM prediction with nongenetic risk models, the predictive power of nongenetic epidemiologic risk factors was calculated. The risk factors included were age (although the age match between cases and controls limited its typically strong predictive ability), socioeconomic status as captured through education and marital status, smoking status, hormonal contraceptive use, number of pregnancies, age at menarche, menopausal status, and race/ethnicity, as measured by the two principal components described above.

To compare the whole-genome prediction to a polygenic risk score, a PRS was calculated using the 155 SNVs that were measured in either the exome array or the GWAS array that also had at least one published effect allele in the GWAS catalog with $P < 5 \times 10^{-8}$. Statistics reported as odds ratios in the GWAS catalog were converted to beta estimates. If a variant was listed multiple times in the GWAS catalog by the same research group, the most recent effect size was used. If a variant was listed multiple times in the GWAS catalog by different research groups, the average of the effect sizes was used. Three variants were excluded due to conflicting directions of effect reported by differing studies.

Using the GCTA software (26, 28), the proportion of variation in breast cancer risk that was associated with genotypic variation was estimated (or "chip heritability") using the data provided by both arrays. This estimation used the single genetic relatedness matrix, and estimated a prevalence of 12%.

Results

Gene-based tests

A summary of the results of the gene-based tests is shown in Fig. 1, and the 20 genes with the smallest *P* values are presented in Table 1. The primary analysis alone identified no gene regions in which variants were associated with early-onset breast cancer incidence, but when combined with the evidence from the replication sample, three gene regions were identified that contained variants associated with breast cancer incidence: *FGFR2* ($P = 1.23 \times 10^{-5}$; replication $P < 1.00 \times 10^{-6}$), *NEK10* ($P = 3.57 \times 10^{-4}$; replication $P < 1.00 \times 10^{-6}$), and *SIVA1*

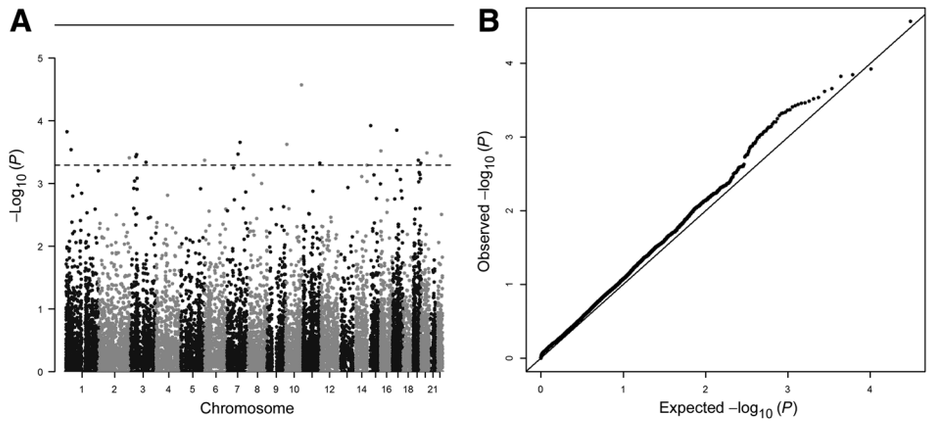

Figure 1.
Gene-based genome-wide associations with early-onset breast cancer incidence. **A,** Manhattan plot of the gene-based logistic regressions of breast cancer incidence in 3,479 early-onset breast cancer cases and 973 age-matched controls. The solid line represents a Bonferroni-corrected *P* value threshold for significance ($P = 3.0 \times 10^{-6}$), not met by any genes. The dashed line represents the *P* value of the twentieth most significant gene. **B,** Quantile–quantile plot of the *P* values of the same analysis.

($P = 5.49 \times 10^{-4}$; replication $P < 1.00 \times 10^{-6}$). The individual variants in the identified genes, along with their weights and annotations, are shown in Supplementary Tables S3–S5. These results were not sensitive to whether the replication data included only women of European descent, or whether it included both women of European and East Asian descent.

The results were sensitive to the weighting method. The substitution of beta weights for CADD weights did not identify any of the three genes highlighted by the CADD-weighted primary analysis. The analysis with beta weights also provided an extremely inflated estimate ($P = 1.1 \times 10^{-19}$) for the gene *MSGN1*, a signal that was entirely driven by a variant that was observed in two heterozygous controls and never in cases. In contrast, the analysis with equal weights produced similar results to those in the analysis with CADD weights.

In two of three risk loci (*SIVA1* and *NEK10*), the variation driving the signal had not been identified previously. As of March 2018, no variants within *SIVA1* were listed in the GWAS catalog. In *NEK10*, the GWAS catalog lists three variants as associated with breast cancer, but these variants were neither measured by the exome array nor in high linkage disequilibrium ($r^2 > 0.8$ in European populations) with any measured variant. In contrast, the GWAS catalog lists seven variants within *FGFR2* as associated with breast cancer, four of which were measured on the exome array. After controlling for these variants, the association within *FGFR2* was no longer statistically significant ($P = 7.57 \times 10^{-2}$).

Of the gene regions identified previously as associated with breast cancer, many were also associated with early-onset cases. In the analysis of the early-onset cases, we found evidence in 19 of 151 gene regions (12.6%) that variation in that region was associated ($P < 0.05$) with early-onset breast cancer: *ARHGAP27*, *ATG10*, *CASC16*, *EBF1*, *EMBP1*, *FGFR2*, *KANSL1*, *LSP1*, *MAP3K1*, *MAPT*, *MKL1*, *NEK10*, *PRC1-AS1*, *RALY*, *SLC4A7*, *TNS1*, *TRAK2*, *WNT3*, and *ZNF45*. The gene-based *P* values for all 151 regions are presented in Supplementary Table S6.

Table 1. Gene-based association test results and corresponding *P* value in replication sample

		Primary analysis			Replication analysis	
CHR	**Gene**	**SNVs**	**MAC**	**SKAT *P***	**SNVs**	**VEGAS *P***
17	*AP2B1*	2	4,261	5.37×10^{-4}	414	9.41×10^{-1}
10	*CELF2*	2	11	2.39×10^{-4}	254	6.06×10^{-2}
3	*DALRD3*	4	2,080	2.90×10^{-4}	5	6.62×10^{-1}
10	*FGFR2*[a]	8	16,039	1.23×10^{-5}	234	1.00×10^{-6}
1	*FNDC7*	14	8,321	4.16×10^{-4}	75	2.33×10^{-1}
1	*GJA9*	10	4,455	3.35×10^{-4}	77	5.86×10^{-1}
19	*HSPBP1*	2	16	4.73×10^{-4}	44	5.74×10^{-1}
7	*LANCL2*	4	3,046	3.26×10^{-5}	190	9.47×10^{-1}
3	*NEK10*[a]	11	10,381	3.57×10^{-4}	598	1.00×10^{-6}
3	*PLSCR4*	9	8,084	7.53×10^{-5}	248	2.15×10^{-1}
16	*RAB26*	5	26	1.73×10^{-4}	10	2.33×10^{-2}
1	*RHBDL2*	6	4,134	2.95×10^{-4}	230	6.43×10^{-1}
14	*SIVA1*[a]	4	1,744	5.49×10^{-4}	28	1.00×10^{-6}
17	*SLFN14*	10	12,264	9.44×10^{-5}	57	5.50×10^{-1}
15	*SNURF*	3	6	1.19×10^{-4}	43	6.95×10^{-1}
14	*SYNE2*	104	19,281	3.18×10^{-4}	328	2.89×10^{-2}
3	*UPK1B*	8	388	3.71×10^{-4}	124	6.42×10^{-1}
7	*WBSCR17/GALNT17*	10	9,519	2.40×10^{-4}	91	6.95×10^{-1}
20	*WFDC11*	1	19	3.26×10^{-4}	73	2.91×10^{-2}
19	*ZNF665*	6	8,052	3.84×10^{-4}	132	3.09×10^{-1}

NOTE: SNVs is the count of single nucleotide variants included in test.
Abbreviation: MAC, minor allele count.
[a]Highlights genes with replication *P* values less than the Bonferroni-corrected level of 2.5×10^{-3}. Because of the simulation design of VEGAS, the minimum *P* value that can be calculated is 10^{-6}.

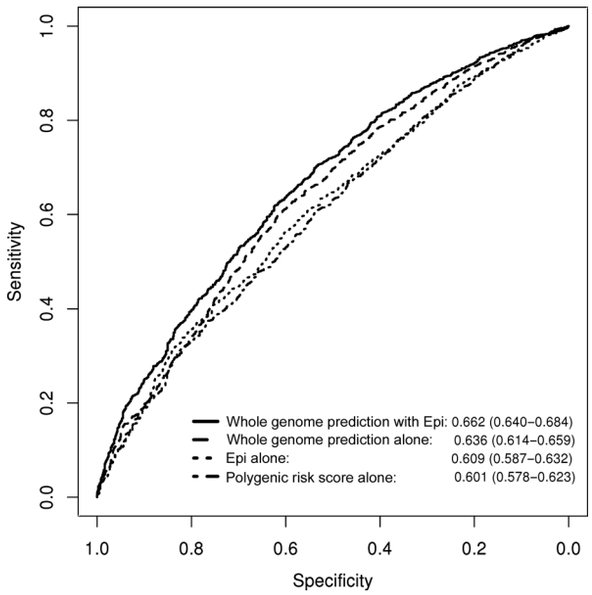

Figure 2.

Comparison of prediction models. The receiver operator characteristic curves for four linear prediction models of breast cancer incidence for the 3,076 women (2,109 cases and 967 controls) with information on epidemiologic risk factors. Epidemiologic risk factors alone produce the short dotted line, the whole-genome LMM produces the long dashed line. The model that combines both epidemiologic and whole-genome information produces the solid line. For comparison, the receiver operator characteristic curve of the polygenic risk score with the 155 SNVs that are listed in the NHGRI-EBI GWAS catalog is also plotted with the short–long dashed line.

Whole-genome prediction

The nongenetic epidemiologic risk factors alone were able to predict breast cancer risk in the primary study participants with an AUC of 0.609 (95% CI, 0.587–0.632; Fig. 2, short dashed line). Removing marital status as a predictor did not substantively change the predictive power of the nongenetic epidemiologic risk factor model (AUC without marital status: 0.6026).

The ability to predict breast cancer diagnosis was improved by the addition of the genetic data. The LLM whole-genome prediction produced an AUC of 0.636 (95% CI, 0.614–0.659; Fig. 2, long dashed line) using only genetic data. The optimal weights (Table 2) were not unique, as multiple other combinations of weights produced substantively similar prediction metrics, but consistently, any nonzero weight on the two GRMs that contained rare variants that do not cause amino acid changes reduced the discrimination of the model. In contrast, the predictive power of the 155-SNV polygenic risk score (AUC 0.601; 95% CI from 2,000 bootstrap replications: 0.578–0.623; Fig. 2, short-long dashed line) was lower than the whole-genome prediction model.

Combining both the epidemiologic risk factors and the LMM whole genome prediction produced an AUC of 0.662 (95% CI, 0.640–0.684; Fig. 2, solid line).

The heritability estimate that incorporated the variants measured on the combined genome wide and exome array was 0.45, with standard error 0.09.

Table 2. Summary of the parameters of the seven genetic relatedness matrices used in the whole-genome prediction model

Genetic relatedness matrix	Variants in matrix	Weight given
NHGRI-EBI variants	1,137	0.200
Not NHGRI-EBI, common protein damaging	16,283	0.104
Not NHGRI-EBI, common other gene region	1,402,269	0.112
Not NHGRI-EBI, common intergenic	1,793,952	0.240
Not NHGRI-EBI, rare protein damaging	83,423	0.334
Not NHGRI-EBI, rare other gene region	58,879	0.000
Not NHGRI-EBI, rare intergenic	59,908	0.000

Discussion

This study finds three genes in which germline variation is associated breast cancer incidence: *FGFR2*, *NEK10*, and *SIVA1*. These genes were identified through analysis of a population that developed breast cancer before age 50, and replicated in a population of all ages of onset, suggesting that variation within these genes is associated with risk across the age spectrum. The association at *FGFR2* replicates past findings, but the variants driving the association at *NEK10* have not previously been implicated in breast cancer, and *SIVA1* is a novel risk locus. Of note, in the Michailidou meta-analysis, the lead SNV in *SIVA1* (rs8006310) was associated with a *P* value of $5.897 \cdot 10^{-8}$ (15), on the threshold of single-study genome-wide significance. All three loci are involved in oncogenic cellular processes: *FGFR2* is part of the known cancer pathway of *PI3K-AKT* (29, 30), *NEK10* is involved in cell-cycle control (31), and *SIVA1* regulates the apoptotic pathway and metastasis (32, 33). The EMBL-EBI gene expression atlas (34) supports a functional mechanism linking these genes and breast cancer. All three are expressed in normal breast tissue, and FGFR2 and NEK10 are reported as underexpressed in breast tumors when compared with normal breast tissue. In addition, *NEK10* is part of a gene-rich region on chromosome 3 that also includes *SLC4A7*, which has previously been implicated in breast cancer risk (7, 35), and, as our recent work suggests, may be related to progesterone receptor status of the tumor, (36) suggesting that future research in the *NEK10* region may further elucidate the mechanisms of breast cancer initiation.

The whole-genome LMM prediction model produces an AUC of 0.636, which is a modest improvement over the discriminative accuracy of the PRS in the same sample, and is more discriminating than recently published PRSs of breast cancer (37, 38). The AUC rose further to 0.662 when combined with epidemiologic risk factors for breast cancer. While verification in an independent study is necessary, this is consistent with other evidence (39) that suggests whole-genome LMM prediction models may better predict risk than a PRS even in diseases, such as breast cancer, where more than 100 risk variants have been identified.

These investigations represent only the fourth study of breast cancer risk that has incorporated either sequencing or exome chip design to better capture variation in exomes directly, and the first such one in which the primary study population was comprised of women with early-onset breast cancer. Each of the previous studies differed in their methodology from this study, by restricting to rare or putative functional variants (10–12), or the use of a burden test, rather than an omnibus gene-based test (11). We included all variants, and weighted them using CADD

scores. This implemented the often suggested (22, 40–42), but infrequently practiced recommendation to include functional information in weighting. The results suggest that CADD score weighting (which includes MAF as an input) may improve upon weighting methods that weight by transformations of MAF alone (such as beta weights), as MAF-only methods may be overly sensitive to rare variants in smaller genes. The present analysis also suggests that the omnibus SKAT-O test is a more appropriate method to elucidate the genetic architecture of breast cancer than the burden test alone, as the mixing parameter needed to mix the burden and SKAT test for the identified genes varied from zero (*FGFR2*) to one (*NEK10*).

This study was limited by the replication data, which did not measure rare variants directly, and also included a small percentage of the participants from the primary analysis. A fully independent replication dataset would have been preferred, and one that was comprised of women younger than 50 would have allowed the results to be interpreted as focusing solely on the genetic architecture of early-onset breast cancer. As no loci were identified as associated with early-onset breast cancer in the primary analysis alone (possibly due to sample size limitations) this analysis is unable to make a direct contrast between the genetic architectures of early- and late-onset breast cancer. Nevertheless, the findings provide novel insight into the genetic etiologies shared by women of all ages, through the identification of the three genes highlighted in this analysis and the 19 loci identified through the NHGRI-EBI catalog that were also associated in the early-onset population. It is likely that additional shared risk loci exist because the exome array does not fully capture the common variation that was queried by previous analyses, and the sample size was modest in this current study.

The findings can provide insight into the design of future studies of the genetic etiology of early-onset breast cancer incidence. The LMM prediction suggests that additional risk loci exist beyond those that have already been discovered, as there was still predictive power apart from the variants that had been identified by previous research. In addition, the heritability estimate of this work approaches that found through family-based studies (which were primarily carried out without restricting age of onset; refs. 3, 43). Taken together, this suggests that the variation measured by arrays can capture the influence of most risk loci, even although their effect sizes may be too small to meet significance thresholds. The prediction model implicates two classes of variants as the most likely location of these yet-to-be identified variants: common variants (MAF >1%) of all functional categories, and rare variants that result in a nonsynonymous mutation. However, when these gene-based results are combined with our earlier single-variant analysis on the GWAS array in this same population (6), it suggests that these additional risk loci may be difficult to identify with a population-based study. To have not been identified in either the previous GWAS or this gene-based analysis, the undiscovered risk loci are likely to be off modest size (and require a large sample size to detect) or not in high linkage disequilibrium with SNVs on arrays, and may need sequencing to identify. If undiscovered risk alleles with a large effect size exist, they are likely to be quite rare, and family-based studies would likely be better powered to identify them than population-based studies of unrelated individuals (44).

This study examined women without known pathogenic mutations in *BRCA1* and *BRCA2*, which are more prevalent in early-onset breast cancer cases for women of European ancestry (45). Given recent work that suggests germline variants increase risk of breast cancer in the presence of *BRCA1* and *BRCA2* mutations (46, 47), the role of gene-by-*BRCA* interactions in early-onset breast cancer is worthy of future study.

In conclusion, this study identifies three loci, *FGFR2*, *NEK10*, and *SIVA1*, which are associated with breast cancer incidence in women of European descent, and demonstrates the utility of whole-genome prediction methods, including those that incorporate known epidemiologic risk factors, in determining risk of early-onset breast cancer.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Disclaimer

The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR.

Authors' Contributions

Conception and design: J.L. Hopper, J. Chang-Claude, H. Ahsan
Development of methodology: J.L. Hopper
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): I.L. Andrulis, J.L. Hopper, J. Chang-Claude, K.E. Malone, E.M. John, M.D. Gammon, M.B. Daly, M.B. Terry, S.S. Buys, D. Huo, O.I. Olopade, J.M. Genkinger, A.S. Whittemore, F. Jasmine, M.G. Kibriya, L.S. Chen, H. Ahsan
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): M. Scannell Bryan, J.L. Hopper, M.G. Kibriya, L.S. Chen, H. Ahsan
Writing, review, and/or revision of the manuscript: M. Scannell Bryan, M. Argos, I.L. Andrulis, J.L. Hopper, J. Chang-Claude, K.E. Malone, E.M. John, M.D. Gammon, M.B. Daly, M.B. Terry, S.S. Buys, D. Huo, O.I. Olopade, J.M. Genkinger, H. Ahsan
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): M. Scannell Bryan, J.L. Hopper, M.B. Daly, F. Jasmine, M.G. Kibriya, H. Ahsan
Study supervision: J.L. Hopper, M.B. Terry, M.G. Kibriya, L.S. Chen, H. Ahsan

Acknowledgments

This work was supported by the NIH grant numbers to support the BCFR: RC1CA145506 and U01CA122171 (to H. Ahsan and M. Argos); RC1CA145506, R01CA094069, and UM1CA164920 (to I.L. Andrulis); and UO1CA66572 and U19CA148065. L.S. Chen was supported through R01 GM108711. M. Scannell Bryan was supported by NIH grant numbers R25-CA057699 2T32 and CA057699–26. L.S. Chen was supported by R01GM108711. Samples from the CPIC were processed and distributed by the Coriell Institute for Medical Research. The studies included in the meta-analysis were supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the 'Ministère de l'Économie, de la Science et de l'Innovation du Québec' through Genome Québec and grant PSR-SIIRI-701, the NIH (U19 CA148065 and X01HG007492), Cancer Research UK (C1287/A10118, C1287/A16563, and C1287/A10710), and The European Union (HEALTH-F2-2009-223175 and H2020 633784 and 634935). All studies and funders for the meta-analysis are listed in Michailidou and colleagues (2017).

The authors would like to thank Regina M. Santella of Columbia University.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received January 3, 2018; revised April 3, 2018; accepted June 8, 2018; published first June 13, 2018.

References

1. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome wide association studies. Nat Genet 2006;38:209–13.
2. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome wide association loci for human diseases and traits. Proc Natl Acad Sci USA 2009;106:9362–7.
3. Bahcall O. Common variation and heritability estimates for breast, ovarian and prostate cancers [Internet]. Nat Genet 2013. Available from: tp://www.nature.com/icogs/primer/common-variation-and-heritability-estimates-for-breast-ovarian-and-prostate-cancers/.
4. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature 2017;551:92.
5. DeSantis CE, Fedewa SA, Goding Sauer A, Kramer JL, Smith RA, Jemal A. Breast cancer statistics, 2015: convergence of incidence rates between black and white women. CA Cancer J Clin 2016;66:31–42.
6. Ahsan H, Halpern J, Kibriya MG, Pierce BL, Tong L, Gamazon E, et al. A genome wide association study of early-onset breast cancer identifies PFKM as a novel breast cancer gene and supports a common genetic spectrum for breast cancer at any age. Cancer Epidemiol Biomarkers Prev 2014;23:658–69.
7. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet 2013;45:353–61.
8. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case–control whole-exome sequencing studies. Am J Hum Genet 2012;91:224–37.
9. Mishra A, Macgregor S. VEGAS2: software for more flexible gene-based testing. Twin Res Hum Genet 2015;18:86–91.
10. Haddad SA, Ruiz-Narváez EA, Haiman CA, Sucheston-Campbell LE, Bensen JT, Zhu Q, et al. An exome-wide analysis of low frequency and rare variants in relation to risk of breast cancer in African American women: the AMBER Consortium. Carcinogenesis 2016;37:870–7.
11. Haiman CA, Han Y, Feng Y, Xia L, Hsu C, Sheng X, et al. Genome wide testing of putative functional exonic variants in relationship with breast and prostate cancer risk in a multiethnic population. PLoS Genet 2013;9:e1003419.
12. Zhou W, Jiang Y, Zhu M, Hang D, Chen J, Zhou J, et al. Low-frequency nonsynonymous variants in FKBPL and ARPC1B genes are associated with breast cancer risk in Chinese women. Mol Carcinog 2017;56:774–80.
13. Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet 2013;9:e1003348.
14. Wheeler HE, Aquino-Michaels K, Gamazon ER, Trubetskoy VV, Dolan ME, Huang RS, et al. Poly-omic prediction of complex traits: omicKriging. Genet Epidemiol 2014;38:402–15.
15. Genome wide Association Studies, iCOGS and OncoArray Summary Results [Internet]; 2018.. Available from: http://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/gwas-icogs-and-oncoarray-summary-results/.
16. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015;526:68–74.
17. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.
18. Purcell S. PLINK v1.07; 2009. Available from: http://pngu.mgh.harvard.edu/purcell/plink/.
19. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc 2015;10:1556–66.
20. Voorman A, Brody J, Lumley T. skatMeta: efficient meta analysis for the SKAT test [Internet]; 2013. Available from: https://cran.r-project.org/web/packages/skatMeta/index.html.
21. R Core Team. R: a language and environment for statistical computing, version3.2.2; 2017. Available from: http://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=2144573.
22. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46:310–5.
23. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome wide association studies. Nat Genet 2006;38:904–9.
24. Wu B, Pankow JS, Guan W. Sequence kernel association analysis of rare variant set based on the marginal regression model for binary traits. Genet Epidemiol 2015;39:399–405.
25. Hindorff LA, MacArther J, Morales J, Junkins HA, Hall PN, Klemm AK, et al. GWAS Catalog [Internet]. NHGRI-EBI Cat. Publ. Genome wide Assoc. Stud; 2016. Available from: http://www.ebi.ac.uk/gwas/search?query=FGFR2.
26. Yang J, Lee SH, Goddard ME, Visscher PM. Genome wide complex trait analysis (GCTA): methods, data analyses, and interpretations. Methods Mol Biol 2013;1019:215–36.
27. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, Bakker PIW de. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 2008;24:2938–9.
28. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet 2010;42:565–9.
29. Dufour C, Guenou H, Kaabeche K, Bouvard D, Sanjay A, Marie PJ. FGFR2-Cbl interaction in lipid rafts triggers attenuation of PI3K/Akt signaling and osteoblast survival. Bone 2008;42:1032–9.
30. Ruiz-Narváez EA, Haddad SA, Lunetta KL, Yao S, Bensen JT, Sucheston-Campbell LE, et al. Gene-based analysis of the fibroblast growth factor receptor signaling pathway in relation to breast cancer in African American women: the AMBER consortium. Breast Cancer Res Treat 2016;155:355–63.
31. Moniz LS, Stambolic V. Nek10 mediates G_2–M cell cycle arrest and MEK autoactivation in response to UV irradiation. Mol Cell Biol 2011;31:30–42.
32. Li N, Jiang P, Du W, Wu Z, Li C, Qiao M, et al. Siva1 suppresses epithelial–mesenchymal transition and metastasis of tumor cells by inhibiting stathmin and stabilizing microtubules. Proc Natl Acad Sci USA 2011;108:12851–6.
33. Wang X, Zha M, Zhao X, Jiang P, Du W, Tam AYH, et al. Siva1 inhibits p53 function by acting as an ARF E3 ubiquitin ligase. Nat Commun 2013;4:1551.
34. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, et al. Gene expression atlas at the European bioinformatics institute. Nucleic Acids Res 2010;38:D690–8.
35. Fletcher O, Johnson N, Orr N, Hosking FJ, Gibson LJ, Walker K, et al. Novel breast cancer susceptibility locus at 9q31.2: results of a genome wide association study. J Natl Cancer Inst 2011;103:425–35.
36. Scannell Bryan M, Argos M, Andrulis IL, Hopper JL, Chang-Claude J, Malone K, et al. Limited influence of germline genetic variation on all-cause mortality in women with early onset breast cancer: evidence from gene-based tests, single-marker regression, and whole-genome prediction. Breast Cancer Res Treat 2017;164:707–17.
37. Mavaddat N, Pharoah PDP, Michailidou K, Tyrer J, Brook MN, Bolla MK, et al. Prediction of breast cancer risk based on profiling with common genetic variants. J Natl Cancer Inst 2015;107:djv036.
38. Shieh Y, Hu D, Ma L, Huntsman S, Gard CC, Leung JWT, et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. Breast Cancer Res Treat 2016;159:513–25.
39. Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, et al. Beyond missing heritability: prediction of complex traits. PLoS Genet 2011;7:e1002051.
40. Roeder K, Devlin B, Wasserman L. Improving power in genome wide association studies: weights tip the scale. Genet Epidemiol 2007;31:741–7.
41. Sveinbjornsson G, Albrechtsen A, Zink F, Gudjonsson SA, Oddson A, Másson G, et al. Weighting sequence variants based on their annotation increases power of whole-genome association studies. Nat Genet 2016;48:314–7.
42. He Z, Zhang D, Renton AE, Li B, Zhao L, Wang GT, et al. The Rare-Variant generalized disequilibrium test for association analysis of nuclear and extended pedigrees with application to Alzheimer disease WGS data. Am J Hum Genet. 100:193–204.
43. Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, et al. Familial risk and heritability of cancer among twins in nordic countries. JAMA 2016;315:68–76.

44. Shen J, Liao Y, Hopper JL, Goldberg M, Santella RM, Terry MB. Dependence of cancer risk from environmental exposures on underlying genetic susceptibility: an illustration with polycyclic aromatic hydrocarbons and breast cancer. Br J Cancer 2017;116:1229–33.
45. Peto J, Collins N, Barfoot R, Seal S, Warren W, Rahman N, et al. Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. J Natl Cancer Inst 1999;91:943–9.
46. Turnbull C, Seal S, Renwick A, Warren-Perry M, Hughes D, Elliott A, et al. Gene-gene interactions in breast cancer susceptibility. Hum Mol Genet 2012;21:958–62.
47. Foo TK, Tischkowitz M, Simhadri S, Boshari T, Zayed N, Burke KA, et al. Compromised BRCA1–PALB2 interaction is associated with breast cancer risk. Oncogene 2017;36:4161–70.