HHS Public Access

Author manuscript

Int J Cancer. Author manuscript; available in PMC 2020 August 28.

Published in final edited form as:

Int J Cancer. 2020 April 15; 146(8): 2175–2181. doi:10.1002/ijc.32825.

Discovery of rare coding variants in *OGDHL* and *BRCA2* in relation to breast cancer risk in Chinese women

Xingyi Guo[1], **Jirong Long**[1], **Zhishan Chen**[1], **Xiao-ou Shu**[1], **Yong-Bing Xiang**[2], **Wanqing Wen**[1], **Chenjie Zeng**[1], **Yu-Tang Gao**[2], **Qiuyin Cai**[1], **Wei Zheng**[1]

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

[2]State Key Laboratory of Oncogene and Related Genes & Department of Epidemiology, Shanghai Cancer Institute, Renji Hospital, Shanghai Jiaotong University School of Medicine, Shanghai, China

Abstract

The missing heritability of breast cancer could be partially attributed to rare variants (MAF < 0.5%). To identify breast cancer-associated rare coding variants, we conducted whole-exome sequencing (~50×) in genomic DNA samples obtained from 831 breast cancer cases and 839 controls of Chinese females. Using burden tests for each gene that included rare missense or predicted deleterious variants, we identified 29 genes showing promising associations with breast cancer risk. We replicated the association for two genes, *OGDHL* and *BRCA*2, at a Bonferroni-corrected $p < 0.05$, by genotyping an independent set of samples from 1,628 breast cancer cases and 1,943 controls. The association for *OGDHL* was primarily driven by three predicted deleterious variants (p.Val827Met, p.Pro839Leu, p.Phe836Ser; $p < 0.01$ for all). For *BRCA*2, we characterized a total of 27 disruptive variants, including 18 nonsense, six frameshift and three splicing variants, whereas they were only detected in cases, but none of the controls. All of these variants were either very rare (AF < 0.1%) or not detected in >4,500 East Asian women from the genome Aggregation database (gnomAD), providing additional support to our findings. Our study revealed a potential novel gene and multiple disruptive variants of *BRCA*2 for breast cancer risk, which may identify high-risk women in Chinese populations.

Keywords

breast cancer; whole-exome sequencing; MEGA; *OGDHL*; *BRCA2*; rare coding variants

Introduction

Genetic factors contribute to the etiology of both sporadic and familial breast cancer.[1] To date, genome-wide association studies (GWAS) have identified more than 180 common

Correspondence to: Xingyi Guo, xingyi.guo@vumc.org.

Additional Supporting Information may be found in the online version of this article.

Conflict of interest

No potential conflicts of interest were disclosed.

genetic variants (minor allele frequency [MAF] > 5%) associated with breast cancer risk.[2–10] The association of these variants with breast cancer is generally weak and together, they explain only a small fraction of breast cancer heritability.[5,7,11] The missing heritability could be partially attributed to risk-associated rare coding variants (MAF < 0.5%), which typically have large effect sizes, as demonstrated in multiple hereditary breast cancer genes, such as *BRCA1*, *BRCA2*, *ATM*, *TP53*, *CHEK2*, *PALB2*, *CDH1*, *STK11*, *NF1* and *PTEN*.[12–25] Two recent case–control association studies conducted in European-ancestry populations discovered many new pathogenic coding variants by sequencing known cancer predisposition genes.[22,23] However, rare coding variants in known breast cancer susceptibility genes and other less well-characterized genes have not been adequately investigated in Asian populations.

In our study, we utilized data resources from the Shanghai Breast Cancer Genetic Study (SBCGS) to search for coding variants associated with breast cancer risk. In the discovery stage, we conducted whole-exome sequencing (WES; mean read depth with 50×) in DNA extracted from blood samples obtained from 831 breast cancer cases and 839 controls. We selected 29 genes that showed promising associations with breast cancer risk in the discovery stage for replication. Rare coding variants identified in WES in our study and other sequencing data sources for these genes were included as a custom content to the Multi-Ethnic Global Array (MEGA), which was used to genotype an independent set of samples from 1,628 breast cancer cases and 1,943 controls (Supporting Information Table S1).

Materials and Methods

Study populations

A total of 831 cases and 839 controls (for whole-exome sequencing), and a total of 1,628 breast cancer cases and 1,943 controls (for MEGA, in the replication stage) were drawn from participants of the Shanghai Breast Cancer Study (SBCS), and the Shanghai Women Health Study (SWHS). Detailed descriptions of these studies have been described in previous literature,[9] and brief summary statistics are presented in Supporting Information Data S1. Both cases and controls came from population-based studies conducted in urban Shanghai, China, including the Shanghai Breast Cancer Studies (SBCS-I and SBCS-II) and the Shanghai Women's Health Study (SWHS). The details of these studies have been previously described. Briefly, for the SBCS-I, participants were recruited between 1996 and 1998. Cancer diagnoses for all cases were reviewed and confirmed by two senior pathologists. Controls were randomly selected from the general population using the Shanghai Resident Registry, a population registry containing demographic information for all residents of urban Shanghai. The inclusion criteria for controls were identical to those for cases, with the exception of a breast cancer diagnosis. Using a protocol similar to that of the SBCS-I, the SBCS-II recruited 1,989 incident breast cancer cases and 1,989 community controls between 2002 and 2005. The age range was expanded from 25–65 years in SBCS-I to 25–70 years in SBCS-II. The SWHS study was a population-based cohort study conducted in Shanghai with baseline surveys conducted from 1996 to 2000. Information on breast cancer diagnoses was collected from standardized and structured interviews and

ascertained through the Shanghai Cancer Registry. Medical charts and pathology slides from diagnostic hospitals were reviewed to further verify the cancer diagnoses. The protocols for these three studies were approved by their relevant institutional review boards, and all participants provided written informed consent.

WES data analysis

We performed WES using the ILLumina GAII sequencing platform with paired-end reads in length with 2 × 100 bp (mean read depth with 50×). The DNA sequencing reads for each sample were mapped to the human reference genome (hg19) using the Burrows–Wheeler Aligner BWA program (version 0.75).[26] Aligned reads marked as duplicates were removed using PICARD MarkDuplicates (http://picard.sourceforge.net/). The remaining aligned reads were then processed using the Genome Analysis ToolKit (GATKv3.2).[27] We performed additional data processes, including local realignment (GATK RealignerTargetCreator and IndelRealigner) and base qualities recalibration (GATK TableRecalibration), after the GATK procedure. We evaluated the sequencing mapping quality, including the mapping rate and coverage for each sample, using the QPLOT tool.[28] Germline variants calling, including both SNPs and Indels, was performed individually for each sample using the GATK HaplotypeCaller tool. We next performed GenotypeGVCFs on variants for all samples together to create a complete list of SNPs and indel VCFs. The Variant Quality Score Recalibration (VQSR) was then applied to filter variants of low quality, using the levels of truth sensitivity: –ts_filter_level of 99.5 for SNP calling and –ts_filter_level of 99.0 for Indel calling. Additionally, we removed variants with low depth of coverage (average < 8 per sample) and high missingness (>2%). Principal components analyses (PCA) were conducted based on approximately 10,000 ancestry informative markers (AIMs) using EIGENSTRAT[29] to identify population outliers, with the 1,000 Genomes Project data as a reference. We also estimated the pair-wise proportion of identity-by-descent (IBD) to identify potentially genetically identical samples, unexpected duplicate samples, or close relatives. After filtering eight samples with low-quality control (QC), we retained data from 831 cases and 839 controls, for downstream association analyses.

Genotyping using MEGA

We performed genotyping using the Illumina MEGA-Expanded Array (Illumina Inc., San Diego, CA), which included 1,554 potential breast cancer-associated coding variants, which we selected from our own WES and by searching other sequencing sources involving more than 12,000 samples. Raw genotype data were imported into GenomeStudio and genotypes were called using cluster definitions provided by Illumina. MEGA genotype calling was carried out using Illumina's GenTrain version 2.0 clustering algorithm in GenomeStudio version 2011.1. Cluster boundaries were determined using study samples. Clustering of the candidate variants for breast cancer risk was manually reviewed. We further conducted QC using PLINK,[30] and repeated the QC procedures conducted in WES. Samples were excluded if (*i*) the call rate <99%, (*ii*) the consistency rates with 1,000 Genomes data <99%, (*iii*) they were a heterozygosity outlier, (*iv*) they were an ethnic outlier (non-Han), (*v*) the samples were in close relationship, (*vi*) the consistency rates among duplicated samples <99% or (*vii*) the samples were of the wrong sex. After filtering samples with low QCs, we retained

genetic data from 1,628 breast cancer cases and 1,943 controls for downstream association analyses.

Examining allele frequency of variants from the genome aggregation database

The genome Aggregation Database (gnomAD v2.1.1) has provided summary data (i.e., allele counts) for germline variants from 125,748 WES and 15,708 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies, through the website browser http://gnomad.broadinstitute.org/. For the rare coding variants discovered in our study, we examined their AFs using population data only from general East Asian women (n = 4,664).

Variant annotation, bioinformatics and statistical analyses

The ANNOVAR tool was applied to annotate missense and disruptive variants.[31] Disruptive variants were defined by nonsense, splice-site and frameshift. To further evaluate the functional impact of missense variants, we annotated each variant with the possible impact of an amino acid substitution on the structure/function from five protein prediction algorithms, including Polyphen-2 HumDiv, Poplyphen HumVar, Sorting Intolerant From Tolerant (SIFT), logistic regression test scores and MutationTaster. All of these analyses were implemented using the WGS Annotator (WGSA) *via* Amazon Web Service (AWS).[32,33]

For gene-based analyses, we evaluated associations of breast cancer risk with all protein-coding genes for WES analysis and selected 29 genes for MEGA analysis. We considered three sets of variants for each tested gene in accordance with their predicted protein function impacts from benign to deleterious alleles, including the “missense,” deleterious “polyphen” and deleterious “strict” sets. The “missense” set included all missense variants as well as disruptive variants. The deleterious “polyphen” set included both the disruptive variants and the missense variants that were predicted to be deleterious by the PolyPhen-2 HumDiv tool. The deleterious “strict” set included both the disruptive variants and the missense variants that were predicted to be deleterious by all five tools, Polyphen-2 HumDiv, Poplyphen HumVar, Sorting Intolerant From Tolerant (SIFT), logistic regression test scores and MutationTaster. We evaluated the association of each set of variants for each gene with breast cancer risk.

For single variant association analyses, we evaluated the association of each coding variant with breast cancer risk using the Fisher exact test under the additive genetic model. The analysis was implemented using the Plink tool and R package.[30] For the gene-based analysis, we only included rare coding variants with MAF < 0.5% for downstream analysis. The gene-based burden analysis was performed to evaluate the association of different sets of variants, including the “missense,” “polyphen” and “strict” sets. All of the analyses were implemented in the RVTESTS package, with the adjustment of batch effect and the first five PCs.[34] To account for the gene-based analysis in the replication stage, we set the significance for our study at $p < 5.7 \times 10^{-4}$, a Bonferroni correction for the testing of a total of 29 genes in three tests.

Data availability

The WES data from 831 breast cancer cases and 839 controls and their clinical characterization in our study have been uploaded to the database of Genotypes and Phenotypes (dbGaP) under Sequence Read Archive (SRA) accession numbers "PRJNA560925" and "PRJNA557488" for sharing the data with the research community.

Results

Discoveries from whole-exome sequencing data

After performing QC procedures for WES data of 1,670 Chinese women, we identified a total of 269,055 rare missense and disruptive coding variants (MAF < 0.5%), including 257,169 missense, 3,017 splicings, 2,394 frameshift and 6,475 nonsense variants (see Materials and Methods).

Single rare coding variant association analyses revealed 71 missense and disruptive coding variants significantly associated with breast cancer risk at a nominal $p < 0.01$ (Supporting Information Table S2). Of them, a deleterious variant, rs201774196 (p. Arg650Trp, *AFAP1*), that was predicted by Polyphen-2 HumDiv and SIFT, showed the most significant association, with breast cancer risk with $p = 1.33 \times 10^{-5}$ (Supporting Information Table S2). Additionally, four nonsense and nine predicted "strict" deleterious variants (predicted by all five algorithms) were characterized by multiple genes (Supporting Information Table S2). Although none of these variants remained significant with an adjustment for the exome-wide multiple comparisons, they suggest valuable candidates of rare coding variants for breast cancer risk.

A gene-based burden association analysis on rare missense and disruptive variants (defined as a "missense" set) revealed that 130 genes, including the known breast cancer susceptibility *BRCA2* gene, were associated with breast cancer risk at $p < 0.01$ (Supporting Information Table S3). Further analyses of "strict" deleterious sets of variants, the "polyphen" set, showed that a total of 81 genes (62.3% of the 130 genes), were associated with breast cancer risk at $p < 0.05$ (Supporting Information Table S3).

A gene-based burden association analysis of the "strict" set showed that 29 were associated with breast cancer risk at $p < 0.05$. Of them, we observed that both *BRCA2* and *OGDHL* genes showed an association at $p < 0.05$ in both "polyphen" and "strict" set analysis (Supporting Information Table S3). In particular, *OGDHL*, functioning as a putative tumor suppressor, involved in regulating the AKT signaling pathway and carbon metabolism,[35–38] was significantly associated with increased breast cancer risk. We considered using the criteria of burden tests for each gene with rare missense at $p < 0.01$ and a relax threshold for predicted deleterious variants at $p < 0.05$, due to relatively fewer variants included in the latter set. Using the above criteria, we selected 29 genes showing promising associations with breast cancer risk for further replication.

Replication of the promising genes in an independent study

Of the 71 rare variants that showed promising associations with breast cancer risk in the discovery stage, we further evaluated their associations with breast cancer risk using data generated by the MEGA. Of the 56 investigated rare variants that were designed by MEGA, three (p.Val97Ile, *TMED3*; p.Pro318Leu, *MRPL33*; p.Ile2077Ser, *ITPR3*) were associated with breast cancer risk at $p < 0.05$ in our validation study (Supporting Information Table S4). In addition, we examined the AFs of these variants using the data of East Asian Women from the gnomAD. Two of these variants (p.Val97Ile, *TMED3*; p.Pro318Leu, *MRPL33*;) were rare, with a frequency of <0.1% in >4,500 East Asian women from gnomAD, providing additional support that they are possible candidates for breast cancer risk (Supporting Information Table S4).

Using a gene-based burden association analysis, we replicated the associations for both the *OGDHL and BRCA2* genes at a Bonferroni-correction of $p < 0.05$. The analyses of the deleterious "polyphen" and "strict" sets showed that the *OGDHL* gene was associated with an increased breast cancer risk ($p = 3.0 \times 10^{-4}$, OR = 2.3 and $p = 8.3 \times 10^{-3}$, OR = 1.9; Table 1). Our single-variant analysis using data from MEGA, revealed three deleterious rare missense variants in association with an increased breast cancer risk at a nominal $p < 0.01$ (Table 2; Supporting Information Table S5). Of them, two missense variants (p.Pro839Leu, p.Phe836Ser) were predicted to be deleterious by all five algorithms. The missense variant, p.Pro839Leu, showed the strongest association, with an increased breast risk ($p = 3.6 \times 10^{-3}$, OR = 12.1; Table 2). The remaining variant, p.Val827Met, predicted to be deleterious by both Polyphen-2 HumDiv and SIFT, showed an association with an increased breast risk, at $p = 0.01$ (OR = 9.8) (Table 2; Supporting Information Table S5). Notably, based on the analysis WES data, our results showed that the variant p.Val827Met was only detected in cases, but none of the controls, while the other two former variants were not present in either cases or controls (Table 2; Supporting Information Table S5). We performed additional analyses by including the data of East Asian Women from the gnomAD. Our results showed that the two deleterious variants, p.Pro839Leu and p.Val827Met, were not present in East Asian women population, while no data is available for the remaining variant, p.Phe836Ser (Supporting Information Table S5).

We analyzed a total of 160 rare coding variants in *BRCA2* from either sequencing and/or MEGA data (Supporting Information Table S6). Gene-based burden association analyses on deleterious "polyphen" and "strict" sets from these variants showed that *BRCA2* had a $p = 1.8 \times 10^{-3}$ and $p = 3.4 \times 10^{-4}$, respectively. We further characterized a total of 27 disruptive variants, including 18 nonsense, six frameshift and three splicing variants, whereas they were only detected in cases but not in controls (Supporting Information Table S6). In particular, of these nonsense variants, four, including p.Tyr1894Ter, p.Ser2984Ter, p.Gln1037Ter and p.Ser2120Ter, were only detected in cases but not in controls from both the sequencing data and MEGA data (Table 3). The results for these variants were further supported by additional analyses, including data from the gnomAD. Specifically, our results showed that two nonsense variants, p.Tyr1894Ter and p.Ser2984Ter, were not present in East Asian women from the gnomAD, and the variant p.Gln1037Ter was observed with a frequency of <0.1% (Supporting Information Table S6). Notably, no data is available for the

remaining variant, p.Ser2120Ter. Furthermore, the evidence of the pathogenicity of these four nonsense variants was also supported by the ClinVar database (https://www.ncbi.nlm.nih.gov/clinvar/).

Discussion

In our study, we provide evidence that a newly identified gene *OGDHL* was associated with breast cancer risk in Chinese populations. The *OGDHL* gene, encoding a component of the multienzyme OGDH complex (OGDHC), has been suggested to be a putative tumor suppressor by previous studies.[35–38] Hypermethylation at CpG sites in the promoter region of the *OGDHL* gene were observed in multiple cancer types, including breast cancer.[35,36] These *in vitro* findings provided additionally biological evidence of *OGDHL* being a putative breast cancer susceptibility gene.

Based on variant functional annotation and association analyses, we characterized four nonsense variants in *BRCA2*, associated with an increased breast cancer risk. Of note, all of the identified nonsense variants have been documented as pathogenic variants in the ClinVar database. In addition, we also characterized additional 23 disruptive variants, especially some of these variants that were only detected in cases, but none of the controls. A further in-depth functional investigation is needed to confirm their potential pathogenic roles in breast cancer. Characterization of these variants is important for genetic testing to identify women at high-risk for breast cancer.

Rare coding variants may account for a substantial proportion of missing heritability for breast cancer. Our study has conducted WES to discover all possible coding variants in 831 breast cancer cases and 839 controls. However, it remains challenging to identify novel rare coding variants for breast cancer risk due to insufficient statistical power. The sample size for our discovery stage is relatively small, and thus many very rare coding variants could have been missed in our study. To overcome this limitation, we included additional variants reported from other publicly available sequencing data sources in our replication stage. However, these additional variants were primarily identified in nonbreast cancer populations and prevented us from capturing pathogenic variants that are only presented in breast cancer patients. it should be noted that three deleterious rare missense variants, in our reported gene *OGDHL*, are significantly associated with breast cancer risk, which was found using data from MEGA in our single-variant analysis (Table 2). However, based on our WES data, we observed that only the variant rs767116963 was present in cases but not in controls. Two other variants were not detected in both cases and controls, which may be due to a limited sample size. In addition, the collapsed variants in our tested genes for gene-based association analysis varied between the WES and MEGA data, which may introduce additional noise for statistical significance. Nevertheless, our results using data in both WES and MEGA provide valuable candidates of deleterious and disruptive rare coding variants and susceptibility genes for breast cancer risk. Future studies with larger sample sizes with deep target sequencing of these promising variants and genes and searching for additional disease susceptible genes for breast cancer in Chinese women are warranted.

In conclusion, we discovered a putative tumor suppressor, *OGDHL*, to be associated with an increased breast cancer risk. We also characterized four nonsense variants in the known cancer susceptibility *BRCA2* gene that could be potentially included in genetic testing of this gene. The identification of variants associated with an elevated risk of breast cancer has important implications in genetic testing to identify high-risk women to reduce breast cancer risk.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The authors thank the study participants and research staff for their contributions and support for this project. We thank Regina Courtney and Jie Wu for laboratory assistance, and Marshal Younger for assistance with editing and manuscript preparation. We also thank Jing He, Jiajun Shi and Yingchang Lu for data QC and preparation. The data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University. This research is supported by the grant from the US National Institutes of Health grant R01CA158473 to W.Z., and by the research development fund from Vanderbilt University Medical Center to X.G.

Grant sponsor: Vanderbilt University Medical Center; **Grant sponsor:** US National Institutes of Health; **Grant number:** R01CA158473

Abbreviations:

AF	allele frequency
BRCA2	BRCA2 DNA Repair Associated
dbGaP	database of Genotypes and Phenotypes
GATK	Genome Analysis ToolKit
gnomAD	Genome Aggregation Database
GWAS	genome-wide association studies
MEGA	Multi-Ethnic Global Array
OGDHL	Oxoglutarate Dehydrogenase Like
SBCGS	Shanghai Breast Cancer Genetic Study
SWHS	Shanghai Women Health Study
WES	whole-exome sequencing

References

1. Nathanson KL, Wooster R, Weber BL. Breast cancer genetics: what we know and what we need. Nat Med 2001;7:552–6. [PubMed: 11329055]
2. Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007;447: 1087–93. [PubMed: 17529967]
3. Long J, Cai Q, Sung H, et al. Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer. PLoS Genet 2012;8: e1002532.

4. Michailidou K, Beesley J, Lindstrom S, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet 2015;47:373–80. [PubMed: 25751625]
5. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet 2013;45: 353–61.e1–2. [PubMed: 23535729]
6. Turnbull C, Ahmed S, Morrison J, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 2010;42: 504–7. [PubMed: 20453838]
7. Michailidou K, Lindstrom S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. Nature 2017;551:92–4. [PubMed: 29059683]
8. Milne RL, Kuchenbaecker KB, Michailidou K, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet 2017;49:1767–78. [PubMed: 29058716]
9. Cai Q, Zhang B, Sung H, et al. Genome-wide association analysis in east Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. Nat Genet 2014;46:886–90. [PubMed: 25038754]
10. Zheng W, Long J, Gao YT, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 2009;41: 324–8. [PubMed: 19219042]
11. Zheng W, Zhang B, Cai Q, et al. Common genetic determinants of breast-cancer risk in east Asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls. Hum Mol Genet 2013;22:2539–50. [PubMed: 23535825]
12. Apostolou P, Fostira F. Hereditary breast cancer: the era of new susceptibility genes. Biomed Res Int 2013;2013:747318.
13. Tan MH, Mester JL, Ngeow J, et al. Lifetime cancer risks in individuals with germline PTEN mutations. Clin Cancer Res 2012;18:400–7. [PubMed: 22252256]
14. 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;42:410–4. [PubMed: 20400964]
15. Gonzalez KD, Noltner KA, Buzin CH, et al. Beyond Li Fraumeni syndrome: clinical characteristics of families with p53 germline mutations. J Clin Oncol 2009;27:1250–6. [PubMed: 19204208]
16. Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. Nat Genet 2008;40: 17–22. [PubMed: 18163131]
17. Pujana MA, Han JD, Starita LM, et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nat Genet 2007;39: 1338–49. [PubMed: 17922014]
18. Rahman N, Seal S, Thompson D, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet 2007; 39:165–7. [PubMed: 17200668]
19. Seal S, Thompson D, Renwick A, et al. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nat Genet 2006;38:1239–41. [PubMed: 17033622]
20. Renwick A, Thompson D, Seal S, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet 2006;38:873–5. [PubMed: 16832357]
21. Pharoah PD, Guilford P, Caldas C. International gastric cancer linkage C. incidence of gastric cancer and breast cancer in CDH1 (E-cadherin) mutation carriers from hereditary diffuse gastric cancer families. Gastroenterology 2001;121: 1348–53. [PubMed: 11729114]
22. Couch FJ, Shimelis H, Hu CL, et al. Associations between cancer predisposition testing panel genes and breast cancer. JAMA Oncol 2017;3:1190–6. [PubMed: 28418444]
23. Lu HM, Li S, Black MH, et al. Association of Breast and Ovarian Cancers with Predisposition Genes Identified by large-scale sequencing. JAMA Oncol 2018;5:51–4.
24. Easton DF, Pharoah PD, Antoniou AC, et al. Gene-panel sequencing and the prediction of breast-cancer risk. N Engl J Med 2015;372: 2243–57. [PubMed: 26014596]
25. Guo X, Lin W, Bao J, et al. A comprehensive cis-eQTL analysis revealed target genes in breast cancer susceptibility loci identified in genome-wide association studies. Am J Hum Genet 2018;102: 890–903. [PubMed: 29727689]

26. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 2009;25:1754–60. [PubMed: 19451168]
27. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–8. [PubMed: 21478889]
28. Li B, Zhan X, Wing MK, et al. QPLOT: a quality assessment tool for next generation sequencing data. Biomed Res Int 2013;2013:865181.
29. Price AL, Patterson NJ, Plenge RM, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–9. [PubMed: 16862161]
30. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75. [PubMed: 17701901]
31. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010;38:e164.
32. Liu X, White S, Peng B, et al. WGSA: an annotation pipeline for human genome sequencing studies. J Med Genet 2016;53:111–2. [PubMed: 26395054]
33. Thibodeau SN, French AJ, McDonnell SK, et al. Identification of candidate genes for prostate cancer-risk SNPs utilizing a normal prostate tissue eQTL data set. Nat Commun 2015;6:8653. [PubMed: 26611117]
34. Zhan X, Hu Y, Li B, et al. RVTESTS: an efficient and comprehensive tool for rare variant association analysis using sequence data. Bioinformatics 2016;32:1423–6. [PubMed: 27153000]
35. Ostrow KL, Park HL, Hoque MO, et al. Pharmacologic unmasking of epigenetically silenced genes in breast cancer. Clin Cancer Res 2009;15: 1184–91. [PubMed: 19228724]
36. Sen T, Sen N, Noordhuis MG, et al. OGDHL is a modifier of AKT-dependent signaling and NF-kappaB function. PLoS One 2012;7:e48770.
37. Guerrero-Preston R, Hadar T, Ostrow KL, et al. Differential promoter methylation of kinesin family member 1a in plasma is associated with breast cancer and DNA repair capacity. Oncol Rep 2014; 32:505–12. [PubMed: 24927296]
38. Fedorova MS, Kudryavtseva AV, Lakunina VA, et al. Downregulation of OGDHL expression is associated with promoter hypermethylation in colorectal cancer. Mol Biol (Mosk) 2015;49: 678–88. [PubMed: 26299868]

What's new?

Many rare pathogenic coding variants in breast cancer have recently been discovered by sequencing known cancer predisposition genes in European-ancestry populations. However, rare coding variants in known breast cancer susceptibility genes and other less well-characterized genes have not been adequately investigated in Asian populations. Using both whole-exome sequencing and array-based genotyping approaches, here the authors identified OGDHL as a novel breast cancer susceptibility gene and multiple disruptive variants of BRCA2 in Chinese women. The identification of variants associated with an elevated risk of breast cancer has important implications in genetic testing to identify high-risk women to reduce breast cancer risk.

Table 1.

Gene-based burden association results of rare coding variants in the *OGDHL* gene for breast cancer risk

Variant set	Number of cases/controls	Number of variants	T1 cases[a]	T1 controls[a]	Freq cases (%)[b]	Freq controls (%)[b]	OR	*P*[c]
Gene-based analysis from all cases and controls in sequencing (discovery stage)								
Missense	831/839	25	25	8	0.030	0.010	3.2	2.6×10^{-3}
Deleterious ('polyphen')	831/839	20	19	8	0.022	0.010	2.4	0.03
Deleterious ('strict')	831/839	16	16	6	0.019	0.007	2.7	0.03
Gene-based analysis from all cases and controls in MEGA (replication stage)								
Missense	1,628/1,943	22	51	27	0.028	0.013	2.20	4.2×10^{-4}
Deleterious ('polyphen')	1,628/1,943	17	48	24	0.027	0.012	2.3	3.0×10^{-4}
Deleterious (strict)	1,628/1,943	14	38	23	0.021	0.011	1.9	8.3×10^{-3}

Summary allele counts and carrier frequencies are presented. Variant sets included the selected deleterious variants predicted by five protein function algorithms (see Materials and Methods). Only variants with allele frequencies less than 0.5% were considered in burden analysis.

Abbreviations: OR, odds ratio

[a] Number of cases or controls carrying the risk alleles in the test

[b] Percentage of cases or controls carrying the risk alleles in the test

[c] *P* value derived from burden test after adjusted for batch effect and the first five PCs

Table 2.

Association results of rare coding variants in the *OGDHL* gene for breast cancer risk ($p < 0.01$)

rsID	Chr	Position (hg19)	Allele[a]	Sequencing analysis				MEGA analysis				Amino acid change	Functional annotation[d]
				Freq cases (%)[b]	Freq controls (%)[b]	OR	P[c]	Freq cases (%)[b]	Freq controls (%)[b]	OR	P[c]		
rs768928553	10	50946003	**G**/A	-	-	-	-	0.48	0.13	3.7	0.01	p.Phe836Ser	strict
rs200530979	10	50945994	**A**/G	-	-	-	-	0.31	0.03	12.1	3.6×10^{-3}	p.Pro839Leu	strict
rs767116963	10	50946031	**T**/C	0.06	0	-	0.50	0.25	0.03	9.8	0.01	p.Val827Met	polyphen

Abbreviations: “D”, deleterious allele; “-“, data not available

[a] Risk/reference allele; risk alleles are shown in bold

[b] Risk alleles frequency (%)

[c] *P* value derived from Fisher exact test under additive model

[d] Functional annotation by ANNOVA and multiple protein function algorithms; “strict” refers to a missense variant predicted to be deleterious by all algorithms

Table 3.

Nonsense variants in the *BRCA2* gene only presented in cases but not in controls

rsID	Chr	Position (hg19)	Allele[a]	Sequencing analysis			MEGA analysis			Amino acid change
				Freq cases (%)[b]	Freq controls (%)[b]	P[c]	Freq cases (%)[b]	Freq controls (%)[b]	P[c]	
rs41293497	13	32914174	**G**/C	0.06	0	0.50	0.15	0	0.02	p.Tyr1894Ter
rs80359146	13	32953650	**G**/C	0.06	0	0.50	0.03	0	0.46	p.Ser2984Ter
rs80358557	13	32911601	**T**/C	0.12	0	0.25	0.09	0	0.09	p.Gln1037Ter
rs397507845	13	32914851	**G**/C	0.12	0	0.25	0.06	0	0.21	p.Ser2120Ter

Abbreviations: "-", data not available

[a] Risk/reference allele; risk alleles are shown in bold

[b] Risk alleles frequency (%)

[c] *P* value derived from Fisher exact test under additive model

[d] Functional annotation by ANNOVA and five protein function algorithms (see Materials and Methods)