Discovery of structural deletions in breast cancer susceptibility genes using whole genome sequencing data from >2,000 African ancestry women

Zhishan Chen
Vanderbilt University Medical Center
Xingyi Guo
Vanderbilt University Medical Center
Jirong Long
Vanderbilt University Medical Center
Jie Ping
Vanderbilt University Medical Center https://orcid.org/0000-0001-9907-5819
Bingshan Li
Vanderbilt University https://orcid.org/0000-0003-2129-168X
Mary Kay Fadden
Meharry Medical College
Thomas Ahearn
National Cancer Institute - Division of Cancer Epidemiology and Genetics https://orcid.org/0000-0003-0771-7752
Daniel Stram
University of Southern California
Xiao-Ou Shu
Vanderbilt University Medical Center
Guochong Jia
Vanderbilt University Medical Center
Jonine Figueroa
University of Edinburgh https://orcid.org/0000-0002-5100-623X
Julie Palmer
Boston University
Maureen Sanderson
Meharry Medical College
Christopher Haiman
University of Southern California
William Blot

International Epidemiology Institute
Montserrat Garcia-Closas
National Cancer Institute https://orcid.org/0000-0003-1033-2650
Qiuyin Cai
Vanderbilt University Medical Center
Wei Zheng (wei.zheng@vanderbilt.edu)
Vanderbilt University Medical Center https://orcid.org/0000-0003-1226-070X

Article

Keywords: breast cancer, cancer, BRCA1, BRCA2

Posted Date: August 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-50862/v1

Abstract

Structural deletions in breast cancer susceptibility genes could confer to cancer risk, but remain poorly characterized. Here, we conducted in-depth whole genome sequencing (WGS) in germline DNA samples from 1,340 invasive breast cancer cases and 675 controls of African ancestry to discover such deletions. We identified 33 deletions, including five protein-truncating deletions in *BRCA1*, *BRCA2*, *RAD51C, GEN1*, and *BRIP1*, were observed only in cases but not in controls. Three deletions, including one protein-truncating deletion in *TP53*, were found to have a higher frequency in cases than in controls. In total, 4.6% of cases and 0.6% of controls carried any of these 36 deletions, resulting in an odds ratio (OR) of 8.0 (95%CI = 2.94 - 30.41). In addition, we identified a low-frequency deletion in *NF1* associated with breast cancer risk (OR = 1.93, 95%CI = 1.14 - 3.42). These findings have significant implications for genetic testing for this common cancer.

Introduction

Breast cancer is the most common malignancy and the second leading cause of cancer death among females in United States. Genetic factors contribute significantly to the etiology of breast cancer. To date, 11 breast cancer susceptibility genes, including *BRCA1, BRCA2, ATM, TP53, CHEK2, PALB2, PTEN, CDH1, STK11, NF1*, and *NBN* have been identified [1–8]. Approximately 30 additional genes, including *BARD1*, *CDKNA2A*, *MRE11A*, *MSH1*, *MSH2* and *MSH6*, have been linked to breast cancer risk, although their etiologic roles have not yet been well established [1,9–11]. Virtually all previous studies have focused on evaluating breast cancer risk associated with putative pathogenic single nucleotide variants (SNVs) and short insertion/deletions (Indels) [1,12,13].

Structure variants (SVs) from deletions can lead to the loss of genomic DNA fragments ranging from a few hundred to a million bases [14–16]. Consequently, SV deletions could have a larger impact on gene functions than SNVs and Indels [15,17]. However, it is challenging to systematically investigate SV deletions using the targeted sequencing assays that have been implemented in most previous studies. Deep whole-genome sequencing (WGS) techniques enable a systematic survey of the whole genome to identify all genetic variants in the genome. We recently reported the identification of six novel deletions in breast cancer susceptibility genes in a small WGS study conducted among 128 patients of Asian and European descent [10]. To expand the study and evaluate breast cancer genetic risk variants in African ancestry women, we analyzed WGS data generated in a large study conducted as part of the African American Breast Cancer Genetic (AABCG) consortium and the Ghana Breast Health Study (GBHS).

Results

The demographic and clinical characteristics of 1,340 invasive breast cancer cases and 675 controls selected from the AABCG consortium and the GBHS are presented in Table 1. The mean age at diagnosis of breast cancer cases was 54.8 years old (standard deviation (SD) = 7.7 years), and the mean age of controls was 58.0 years old (SD = 6.9 years). Approximately 14.8% of cases and 1.0% of controls reported

a family history of breast cancer. Approximately 38.1% of the cases had hormone receptor negative tumors.

The summary of the identified potential loss-of-fucntion deletions

We identified a total of 80 deletions in the intragenic regions of the 29 established or suspected susceptibility genes (Supplementary Data 4). The median length of these deletions was 405 bp (from 50 bp to 32.8 kb) and the majority of them were low frequency or rare (78.8% with a frequency of deletion carriers < 0.01). Of these deletions, 33 were only presented in the cases (n = 44), but not in controls. In contrast, 16 deletions were only presented in the controls (n = 17), but not in cases. The majority of these 49 deletions were located in the intronic regions (83.7%). Although overall there was no significant case-control difference in the frequency of carriers of the deletions detected in this study, cases were significantly more likely to carry the deletions either in the coding/exonic regions or in the intronic reigions with the evidence of epigenetic signals ($P < 0.05$; 2.6% in cases and 1.2% in controls). We also observed that four particular deletions showed a higher frequency in cases than in controls (odds ratio (OR) > 1.5 for each deletion; Supplementary Data 4).

Potential loss-of-function deletions only presented in cases

Of the 33 deletions that were only presented in cases (n = 44, 3.3% of 1,340 cases), we found five putative pathogenic deletions in the exonic or coding regions of *BRCA1*, *BRCA2*, *RAD51C*, *GEN1*, and *BRIP1*, and 28 in the intronic regions of 18 other genes (Table 2; Supplementary Data 4). Several deletions were found in all participating studies (Supplementary Data 5 and 6). No cases carried more than one deletion. Eleven of these deletion were identified in seven established breast cancer susceptibility genes, including *BRCA1*, *BRCA2*, *PTEN*, *CDH1*, *NF1*, *STK11* and *CHEK2*, and the remaining 22 were identified in the 15 putative breast cancer susceptibility genes (Table 2). One of these deletions, located in the *PTEN* gene, had been reported in our previous study conducted in Asian and European descendants [10]; in the present study, the deletion was found in nine cases and no controls ($P = 0.03$) (Supplementary Data 4).

Three cases carried potential loss-of-function deletions in the *BRCA1 or BRCA2* genes, accounting for 0.22% of cases under study (Table 2). Of them, two putative pathogenic deletions are located in the coding region, a 3.34 kb deletion in the *BRCA1* gene involving three exons (Fig. 1a) and a 32.8kb deletion in the *BRCA2* gene, with a loss of nine exons (Fig. 1c). The third deletion (3.41 kb) is located in the first intron of the *BRCA1* gene. This deletion may involve functional elements with the evidence of ChIP-seq enriched peaks (Fig. 1b).

Sixteen cases carried potential loss-of-function deletions in five other established breast cancer susceptibility genes, including *PTEN*, *CHEK2*, *NF1*, *CDH1*, and *STK11,* accounting for 1.2% of the total investigated cases (Table 2). Of these deletions, three (~ 0.9 kb, 7.1 kb, and 0.5 kb) were identified in *PTEN*, one (~ 3.0 kb) in *CHEK2*, one (~ 9.8 kb) in *NF1*, two (~ 3.8 kb and 1.1 kb) in *CDH1*, and one (~ 0.3 kb) in *STK11*. All of these deletions can lead to the loss of intronic sequences of these breast cancer susceptibility genes. In addition, these deletion regions are likely to involve potential regulatory elements

with the evidence of epigenetic signals, such as histone modifications, DNase I hypersensitive sites and ChIP-seq enriched peaks (Supplementary Data 4).

Twenty-five cases carried potential loss-of-function deletions in 15 putative breast cancer susceptibility genes, accounting for 1.9% of the total investigated cases. Of the 22 deletions observed in these genes, three putative pathogenic deletions involve exonic or coding sequences, including one (~ 140 bp) for *BRIP1*, one (~ 82 bp) for *GEN1* and one (~ 4.9 kb) for *RAD51C* (Table 2). All other 19 deletions can lead to the loss of intronic sequences of these breast cancer susceptibility genes (Table 2). Of them, 11 deletion regions are located in nine genes, including *BMPR1A, POLE*, *AKT1*, *RAD51D*, *MSH2*, *MSH6*, *XRCC2*, *FANCM* and *FANCC*, which may involve regulatory elements supported by epigenetic signals (Supplementary Data 4).

Rare deletions with higher frequency in cases than controls

We identified three potential loss-of-function deletions in breast cancer susceptibility genes, with each showing a higher frequency in cases than in controls (Table 2; Supplementary Data 4). The deletion in *TP53* (~ 1.6 kb in the coding region) results in a loss of the whole last exon and the 3' untranslated region (UTR). This deletion was observed in six cases and one control in the present study and has been reported in our previous study [10]. The other two deletions were observed in the intronic regions of *GEN1* and *MSH6* and were observed to have epigenetic signals (Table 2; Supplementary Data 4).

Taken together, the above three potential loss-of-function deletions, together with those only observed in cases, were presented in 4.6% cases (n = 61) and 0.6% of controls (n = 4). Carrying any one of these 36 deletions was associated with an 8.0-fold increased risk of breast cancer (95%CI = 2.94 - 30.41, $P = 1.5 \times 10^{-7}$) (Table 2). We observed significant different frequencies of the deletion carriers among studies (e.g. 3.6% and 1.4% for African Americans and Ghanaians, respectively; $P = 0.01$; Supplementary Data 6).

A low-frequency deletion associated with breast cancer risk

We identified a low-frequency deletion (~ 72 bp) in the intronic region of *NF1* associated with breast cancer risk (OR = 1.93, 95%CI = 1.14 - 3.42, $P = 0.01$) (Table 3). The deletion region may involve regulatory elements with the evidence of the epigenetic signals, including ChIP-seq enriched peak (Supplementary Data 4). We observed that deletion accounts for 5.3% of cases and 2.8% of controls.

Discussion

This is the first large study that uses deep WGS technology to systematically search for putative pathogenic SV deletions in the established and putative breast cancer susceptibility genes in African ancestry populations. We identified 37 potental loss-of-function deletions that likely confer to breast cancer susceptibility, including several putative pathogenic deletions with strong evidence of the loss of coding or exonic regions in the susceptibility genes. Of them, 33 deletions were identified in seven established and 15 putative susceptibility genes that were only presented in cases. Of them, 36 deletions

were seen in 4.6% of the cases and 0.6% controls under study, resulting in an 8-fold elevated risk of breast cancer among deletion carries. One of these deletions, located in *PTEN*, was associated with an increased risk of breast cancer in our previous study of European populations [10]. These identified potental loss-of-function deletions should be considered in genetic testing or in determining cancer therapies, such as platinum-based chemotherapy and inhibitors of poly(ADP-ribose) polymerase (PARP) commonly recommended for patients with BRCA1/2 mutations.

Two of the observed deletions that result in a deletion of the coding sequences of *BRCA1* (3.34 kb, covering three exons) and *BRCA2* (~ 32.8Kb, covering nine exons) are highly likely to be functionally significant as they disrupt gene function. In addition to these two well-known established breast cancer susceptibility genes, we showed that the loss of coding or exonic sequences for putative breast cancer genes, including *BRIP1*, *GEN1*, and *RAD51C*, could affect their gene functions, thus contributing to breast cancer susceptibility. Of the 31 potental loss-of-function deletions identified in the intronic regions of breast cancer susceptibility genes, 23 (74.2%) deletions may lead to the loss of potential regulatory elements, consequently disrupting the regulation of gene expression. For example, we observed H3K4me1, DNase clusters and TF ChIP-seq binding sites in the deletion in the intronic sequence of *PTEN* (Fig. 2). Future functional exploration and large-scale case-control association studies are warranted to confirm the deletions for breast cancer susceptibility.

We searched public SV datasets from the gnomAD database for the 37 SV deletions identified in our study and found that 28 deletions (75.7%) have not been reported in this database. Of the remaing nine SVs that overlapped with gnomeAD, no studies have reported their potential functional roles in disease susceptibility (Supplementary Data 7). In addition to our reported 37 SV deletions, we also identified three deletions, located in the coding or exonic regions of *POLE* (~ 505 bp), *PPM1D* (~ 107 bp) and *STK11* (~ 1 kb), that were only presented in controls (Supplementary Data 4). Interestingly, the deletion carriers did not show either relatively young age or family history. It's possible these deletions may not significantly confer breast cancer susceptibility, while follow-up studies are needed to confrm these findings.

It should be noted that our computationally predicted deletions lacked technical validation. However, we demonstrated the validity of our deletion calling in our previous study [10]. Specifically, to enhance the accuracy of SV deletion determination, we applied multiple commonly used SV detection algorithms to generate a consensus set from at least two overlapping SV callers. In addition, we developed comprehensive bioinformatics analysis strategies by analyzing informative reads to re-genotype these identified deletions. In particular, the number of supporting reads, and a ratio of mapped reads in the deletion region relative to its flanking region, were applied to remove some potential false positive findings. Of the 80 identified potental loss-of-function deletions, 76.3% (61/80) were identified consistently by both LUMPY and Manta, providing some assurance for the validity of SV calling in our study. Furthermore, we manually reviewed each of the identified deletions using the Integrative Genomics Viewer (IGV) and samplot. As demonstrated in Figs. 1 and 2, the sequencing reads were observed to be significantly reduced in the deletion regions compared to the flanking regions. These results provide strong evidence that the identified deletions were reliable.

In conclusion, this is the largest study using WGS data to search for potential loss-of-function deletions, especially putative pathogenic deletions, in breast cancer susceptibility genes. We provided strong evidence for the presence of putative pathogenic SV deletions in breast cancer genes. Our study reveals a large number of newly-identified potential loss-of-function and putative pathogenic SV deletions in African ancestry women, and these novel findings may improve clinical testing and selection of cancer treament.

Methods

Study populations

This study included 1,340 invasive breast cancer cases and 675 cancer-free controls from five studies, the Southern Community Cohort Study (SCCS, N = 685), the Nashville Breast Health Study (NBHS, N = 99), the Southern Tri-State Breast Health Study (STSBHS, N = 415), the Ghana Breast Health Study (GBHS, N = 443) and the Multiethnic Cohort Study (MEC, N = 373) (Supplementary Data 1). Detailed descriptions of these studies have been previously published and are briefly described below [18–22]. The SCCS is a prospective cohort study that recruited approximately 86,000 study participants aged 40–79 years between 2002 and 2009 from 12 states in the southeastern U.S. Approximately 32,500 of the SCCS participants were African American women. Cancer cases, including those diagnosed with breast cancer, were identified via linkage to state cancer registries. The NBHS is a population-based case-control study conducted in the Nashville metropolitan area. Participants were recruited between 2001 and 2008. Breast cancer cases were identified through the Tennessee State Cancer Registry and five major hospitals in Nashville that provide medical care for breast cancer patients. Controls were identified via random digit dialing of households in the same geographic area as the cases. 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. African American study participants were recruited from the Los Angeles area. 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. The STSBHS is a population-based case-only study conducted in Tennessee, South Carolina and Georgia. Participants have been recruited since 2012. Breast cancer cases were identified through the Tennessee Cancer Registry, the South Carolina Central Cancer Registry and the Georgia Comprehensive Cancer Registry, also a SEER registry. The GBHS is a population-based case-control study conducted in Accra and Kumasi, Ghana, between 2013 and 2015. Breast cancer cases were identified through three major hospitals and controls were frequency matched to cases by age and district of residence.

In the present study, we selected cases and controls according to the following criteria. For both cases and controls, these criteria were used: 1) blood or saliva samples available; 2) with an African ancestry estimate of > 70% if the estimate is available. For breast cancer cases, the following additional criteria were used: 1) with ER status or tissue availiable for ER measurement; 2) for ER + cases, age of diagnosis < 65 or with family history of breast cancer; 3) for ER- cases, age of diagnosis < 70 or with family history

of breast cancer; 4) for both ER + and ER- cases, excluding potential BRCA carrier cases with age of diagnosis < 45 and with family history of breast cancer. For controls, the following additional criteria were used: 1) without diagnosis of any cancer except non-melanoma skin cancer and and without family history of any cancer at the last follow-up; 2) with age at the last follow-up > 60, 50% of them in 60–64 yrs and 50% of them in 65–70 yrs.

Whole genome sequencing library construction

The whole-genome sequencing was performed using the Illunima HiSeq X Ten and BGIDEQ-500 platforms. For HiSeq X Ten sequencing, 1 μg genomic DNA was randomly fragmented by Covaris, followed by purification by an AxyPrep Mag PCR clean up kit. The fragments were end repaired by End Repair Mix and purified afterwards. The repaired DNAs were combined with A-Tailing Mix, then the Illumina adaptors were ligated to the Adenylate 3'Ends DNA and followed by the products' purification. The products were selected based on the insert size. Several rounds of PCR amplification with PCR Primer Cocktail and PCR Master Mix were performed to enrich the Adapter-ligated DNA fragments. After purification, the library was qualified by the Agilent Technologies 2100 bioanalyzer and ABI StepOnePlus Realtime PCR System and were sequenced pair-end using HiSeq X Ten. For BGISEQ sequencing, 1μg genomic DNA was randomly fragmented by Covaris. The fragmented DNA was selected by Agencourt AMPure XP-Medium kit to an average size of 200-400bp. The selected fragments were sequenced through end-repair, 3' adenylated, adapters-ligation, PCR Amplifying and the products were recovered by the AxyPrep Mag PCR clean up Kit. The double stranded PCR products were heat denatured and circularized by the splint oligo sequence. The single strand circle DNA (ssCir DNA) was formatted as the final library and qualified by QC. The qualified libraries were sequenced on a BGISEQ-500 platform.

Whole genome sequencing data processing

All of the sequencing samples reached the average sequencing depth with at least 30X. Bcl2fastq2 Conversion Software (Illumina) was used to generate de-multiplexed FASTQ files. FASTQ files were processed according to the GATK best practices pipeline [23]. The sequencing reads were aligned to the human reference genome (GRCh38) using the Burrows-Wheeler Aligner (BWA) program (version 0.79a) [24]. The mapped reads were further processed by removing the duplicated reads using MarkDuplicates from the Picard tool (http://picard.sourceforge.net/) and recalibrating the base quality scores using BaseRecalibrator. The mapped reads from each analysis-ready BAM file for each sample was generated for subsequent analysis of SV deletion calling.

SV deletion calling and quality control

We systematically searched SV deletions using six SV callers, including GenomeSTRiP [25], LUMPY [26], DELLY [27], Manta [28], Pindel [29] and Canvas [30]. For the SV caller GenomeSTRiP, we performed the SV deletion discovery using the 'SVDiscovery' script. The genotype for each identified SV deletion was then generated across all samples using the 'SVGenotyper' script. For the other five SV callers, we performed SV deletion discovery for each sample using the recommended parameters listed in Supplementary Data

2. We developed a pipeline to merge and compare SVs from different callers. To generate a consensus call from different SV callers, the initial merging step was implemented using the SURVIVOR tool using recommended parameters, including 1) the maximum distance of 1kb for breakpoints, 2) detected by at least two SV callers, 3) being on the same strand and 4) the minimum length of 50 bp for SV deletion [31]. For a consensus set of SV deletion for each sample, we further performed re-genotyping by analyzing informatics reads (i.e., split reads) from the analysis-ready BAM file using the SVTyper tool [32]. To remove a potential false-positive SV deletion, we filtered deletions if there were less than seven mapped reads around the deletion. We also calculated a ratio of mapped reads in the deletion region relative to its flanking 1kb region using the duphold tool and filtered deletions if the ratio > 0.7 [33]. We further merged the remaining SV deletions across all samples using SURVIVOR, with the recommended parameters. In the end, we generated a Variant Call Format (VCF) file of the high-confident SV deletions for the study.

Characterization of identified SV deletions

To systematically annotate SV deletions in an intragenic region, we used the tool 'GeneOverlapAnnotator' from the GenomeSTRiP (17) to detect the overlaps between deletions and annotated genes. Using the human transcriptome annotation Gencode version 33 (GRh38) from the GENCODE (https://www.gencodegenes.org/human/release_33.html), we assigned each SV deletion a gene body region, including the coding, exonic, and intronic regions. Deletions between 50 bp and 1 Mb in length were evaluated in 40 genes, including 11 established breast cancer susceptibility genes and 29 putative breast cancer susceptibility genes (Supplementary Data 3). The function of SV deletions in intronic regions was also accessed through the UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgGateway). The epigenetic landscape of histone markers H3K4Me1, H3K4Me3, and H3K27Ac, DNase I hypersensitive sites, and chromatin immunoprecipitation sequencing (ChIP-seq) binding sites of transcription factors on all available ENCODE cell lines [34] were examined through layered tracks from the UCSC Genome Browser. We estimated the association of each deletion with breast cancer risk using Fisher's exact test. The analysis was implemented using R (version 3.4.3).

Comparison of identified SV deletions with public SV databases

We compared the SV deletions identified in this study with SV deletions from large genomic databases, including the genome Aggregation Database (gnomAD v2.1.1) [35]. We downloaded a SV dataset generated using whole genome sequence data from 10,738 unrelated individuals in gnomAD v2.1.1 through http://gnomad.broadinstitute.org/. The intersection of SV deletions identified in the study and the above databases were analyzed using the 'bedtools intersect' function. A SV deletion between our study and gnomAD was defined as the same one if their reciprocal overlap (RO) was more than 80%.

Declarations

Data availability

Access to the whole genome sequencing data could be requested by submission of an inquiry to Dr. Wei Zheng.

Code availability

Access to the custom code could be requested by submission of an inquiry to Dr. Wei Zheng.

Acknowledgements

The authors thank the study participants and research staff for their contributions and support for this project. The study was supported primarily by NIH grant R01CA202981. The data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University. The GBHS was funded by Intramural Funds of the National Cancer Institute, USA. Samples from the GBHS were processed at the Cancer Genomics Research Laboratory, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Gaithersburg, MD.

Author Contributions

Study design: W.Z. and X.G.; Data analysis: Z.C., X.G., J.P.; Data interpretation: J.L., Q.C., X.-O.S., W.Z.; Drafting of the manuscript: Z.C., X.G., W.Z.; Review of the manuscript: Z.C., X.G., J.L., Q.C., X.-O.S., P.J.R., C.A.H., M.G.-C., W.Z.; Administrative, technical, or material support: B.L., F.M.K., T.U.A., S.D.O., G.J., J.F., the GBHS team, P.J.R., M.S., C.A.H., W.J.B., M.G.-C., W.Z..

Corresponding author

Correspondence to Wei Zheng

Competing Interests

The authors declare no competing interests.

References

1. Easton, D. F. *et al.* Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med* **372**, 2243–2257, doi:10.1056/NEJMsr1501341 (2015).
2. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66–71, doi:10.1126/science.7545954 (1994).
3. Renwick, A. *et al.* ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet* **38**, 873–875, doi:10.1038/ng1837 (2006).
4. Rahman, N. *et al.* PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet* **39**, 165–167, doi:10.1038/ng1959 (2007).
5. Gonzalez, K. D. *et al.* Beyond Li Fraumeni Syndrome: clinical characteristics of families with p53 germline mutations. *J Clin Oncol* **27**, 1250–1256, doi:10.1200/JCO.2008.16.6959 (2009).

6. Tan, M. H. *et al.* Lifetime cancer risks in individuals with germline PTEN mutations. *Clin Cancer Res* **18**, 400–407, doi:10.1158/1078-0432.CCR-11-2283 (2012).
7. Guo, X. *et al.* Discovery of a Pathogenic Variant rs139379666 (p. P2974L) in ATM for Breast Cancer Risk in Chinese Populations. *Cancer Epidemiol Biomarkers Prev* **28**, 1308–1315, doi:10.1158/1055-9965.EPI-18-1294 (2019).
8. Guo, X. *et al.* Discovery of rare coding variants in OGDHL and BRCA2 in relation to breast cancer risk in Chinese women. *Int J Cancer* **146**, 2175–2181, doi:10.1002/ijc.32825 (2020).
9. Lu, H. M. *et al.* Association of Breast and Ovarian Cancers With Predisposition Genes Identified by Large-Scale Sequencing. *JAMA Oncol* **5**, 51–57, doi:10.1001/jamaoncol.2018.2956 (2019).
10. Guo, X. Y. *et al.* Use of deep whole-genome sequencing data to identify structure risk variants in breast cancer susceptibility genes. *Hum Mol Genet* **27**, 853–859, doi:10.1093/hmg/ddy005 (2018).
11. Zeng, C. *et al.* Evaluation of pathogenetic mutations in breast cancer predisposition genes in population-based studies conducted among Chinese women. *Breast Cancer Res Treat* **181**, 465–473, doi:10.1007/s10549-020-05643-0 (2020).
12. Couch, F. J. *et al.* Associations Between Cancer Predisposition Testing Panel Genes and Breast Cancer. *Jama Oncol* **3**, 1190–1196, doi:10.1001/jamaoncol.2017.0424 (2017).
13. Lu, H. M. *et al.* Association of Breast and Ovarian Cancers With Predisposition Genes Identified by Large-Scale Sequencing. *Jama Oncol*, doi:10.1001/jamaoncol.2018.2956 (2018).
14. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu Rev Med* **61**, 437–455, doi:10.1146/annurev-med-100708-204735 (2010).
15. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**, 363–376, doi:10.1038/nrg2958 (2011).
16. Guo, X. *et al.* Variant discovery and breakpoint region prediction for studying the human 22q11.2 deletion using BAC clone and whole genome sequencing analysis. *Hum Mol Genet* **25**, 3754–3767, doi:10.1093/hmg/ddw221 (2016).
17. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81, doi:10.1038/nature15394 (2015).
18. Kolonel, L. N., Altshuler, D. & Henderson, B. E. The multiethnic cohort study: exploring genes, lifestyle and cancer risk. *Nat Rev Cancer* **4**, 519–527, doi:10.1038/nrc1389 (2004).
19. Zheng, W. *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet* **41**, 324–328, doi:10.1038/ng.318 (2009).
20. Signorello, L. B. *et al.* Southern community cohort study: establishing a cohort to investigate health disparities. *J Natl Med Assoc* **97**, 972–979 (2005).
21. Brinton, L. A. *et al.* Design considerations for identifying breast cancer risk factors in a population-based study in Africa. *Int J Cancer* **140**, 2667–2677, doi:10.1002/ijc.30688 (2017).
22. Figueroa, J. D. *et al.* Reproductive factors and risk of breast cancer by tumor subtypes among Ghanaian women: A population-based case-control study. *Int J Cancer*, doi:10.1002/ijc.32929

(2020).

23. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11 10 11–11 10 33, doi:10.1002/0471250953.bi1110s43 (2013).
24. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, doi:10.1093/bioinformatics/btp324 (2009).
25. Handsaker, R. E., Korn, J. M., Nemesh, J. & McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet* **43**, 269–276, doi:10.1038/ng.768 (2011).
26. Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol* **15**, R84, doi:10.1186/gb-2014-15-6-r84 (2014).
27. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333-i339, doi:10.1093/bioinformatics/bts378 (2012).
28. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222, doi:10.1093/bioinformatics/btv710 (2016).
29. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871, doi:10.1093/bioinformatics/btp394 (2009).
30. Roller, E., Ivakhno, S., Lee, S., Royce, T. & Tanner, S. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics* **32**, 2375–2377, doi:10.1093/bioinformatics/btw163 (2016).
31. Jeffares, D. C. *et al.* Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat Commun* **8**, 14061, doi:10.1038/ncomms14061 (2017).
32. Chiang, C. *et al.* SpeedSeq: ultra-fast personal genome analysis and interpretation. *Nat Methods* **12**, 966–968, doi:10.1038/nmeth.3505 (2015).
33. Pedersen, B. S. & Quinlan, A. R. Duphold: scalable, depth-based annotation and curation of high-confidence structural variant calls. *Gigascience* **8**, doi:10.1093/gigascience/giz040 (2019).
34. Consortium, E. P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, doi:10.1038/nature11247 (2012).
35. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. 531210, doi:10.1101/531210 %J bioRxiv (2019).

Tables

Table 1. Demographic and clinical characteristics of study participants

Characteristics	Cases	Controls
Number of participants	1340	675
Age, No. (%)		
Mean (SD)	54.8 (7.7)	58.0 (6.9)
< 45	107 (8.0)	0 (0.0)
45-54	562 (42.0)	212 (31.4)
55-64	570 (42.5)	365 (54.1)
>=65	101 (7.5)	98 (14.5)
Family history of breast cancer[a], No. (%)		
Yes	198 (14.8)	7 (1.0)
No	748 (55.8)	517 (76.6)
Molecular Subtypes[b], No. (%)		
ER+/PR+	390 (29.1)	NA
ER+/PR-	97 (7.2)	NA
ER-/PR+	53 (4.0)	NA
ER-/PR-	511 (38.1)	NA
ER-/PR-/Her 2+	98 (7.3)	NA
ER-/PR-/Her 2-	342 (25.5)	NA

SD: standard deviation; ER: estrogen receptor; PR: progesterone receptor; HER2: human epidermal growth factor receptor 2; [a,b]Among participants with available data only; The frequencies by study are presented in Supplementary Data 1.

Table 2. Potential loss-of-function SV deletions in breast cancer susceptibility genes.

Genes	No. of SV deletions	Deletions carriers		Chr	Start	End	Deletion length (kb)	Annotation
		No. (%) of patients	No.(%) of controls					
Group 1[a]	**33**	**44 (3.3)**	**0 (0.0)**					
BRCA1	2	1	0	chr17	43097359	43100701	3.34	CDS
		1	0	chr17	43141278	43144686	3.41	Intron
BRCA2	1	1	0	chr13	32328593	32361397	32.80	CDS
PTEN	3	9	0	chr10	87893066	87893965	0.90	Intron
		1	0	chr10	87898130	87905186	7.06	Intron
		1	0	chr10	87924200	87924684	0.49	Intron
CDH1	2	1	0	chr16	68767457	68771267	3.81	Intron
		1	0	chr16	68797747	68798883	1.13	Intron
NF1	1	1	0	chr17	31125126	31134974	9.84	Intron
STK11	1	1	0	chr19	1194104	1194416	0.31	Intron
CHEK2	1	1	0	chr22	28738189	28741177	2.99	Intron
BMPR1A	1	1	0	chr10	86860908	86860959	0.05	Intron
RECQL	1	1	0	chr12	21492924	21494178	1.25	Intron
POLE	2	1	0	chr12	132644231	132646226	1.99	Intron
		2	0	chr12	132684032	132685248	1.22	Intron
FANCM	1	3	0	chr14	45166046	45166452	0.41	Intron
AKT1	1	1	0	chr14	104783395	104786595	3.20	Intron
RAD51D	2	1	0	chr17	35110247	35116767	6.52	Intron
		1	0	chr17	35114413	35115113	0.70	Intron
RAD51C	1	1	0	chr17	58701274	58706194	4.92	CDS
BRIP1	1	1	0	chr17	61680184	61680322	0.14	Exon
SMAD4	1	1	0	chr18	51062688	51062766	0.08	Intron
GEN1	1	1	0	chr2	17787692	17787773	0.08	Exon
MSH2	3	1	0	chr2	47467407	47467714	0.31	Intron
		1	0	chr2	47469338	47469636	0.30	Intron
		1	0	chr2	47570627	47570678	0.05	Intron
MSH6	2	1	0	chr2	47731689	47731923	0.23	Intron
		1	0	chr2	47778330	47782670	4.34	Intron
RINT1	1	1	0	chr7	105545750	105546049	0.30	Intron
XRCC2	1	1	0	chr7	152653978	152655793	1.82	Intron
FANCC	3	1	0	chr9	95334681	95335119	0.44	Intron
		1	0	chr9	95397519	95402310	4.79	Intron
		1	0	chr9	95410433	95410514	0.08	Intron
Group 2[b]	**3**	**17 (1.3)**	**4 (0.6)**					
TP53	1	6	1	chr17	7661327	7662878	1.55	CDS
GEN1	1	5	1	chr2	17778323	17778385	0.06	Intron
MSH6	1	6	2	chr2	47744589	47744639	0.05	Intron

[a]Group 1 – Deletions only among patients. [b]Group 2 – Deletions at a higher frequency in patients than controls. A total of 61 patients (4.6%) and 4 controls (0.6%) carried any one of these 36 SV deletions (OR = 8.0, 95%CI = 2.94-30.41, $P = 1.49 \times 10^{-8}$).

Table 3. A low-frequency deletion in the *NF1* gene associated with breast cancer risk.

Chr	Start	End	Deletion length (kb)	Annotation	Deletion carriers		OR (95% CI)	*P*
					No. (%) of patients	No. (%) of controls		
chr17	31290891	31290962	0.07	Intron	71 (5.3)	19 (2.8)	1.93 (1.14; 3.42)	0.012

Figures

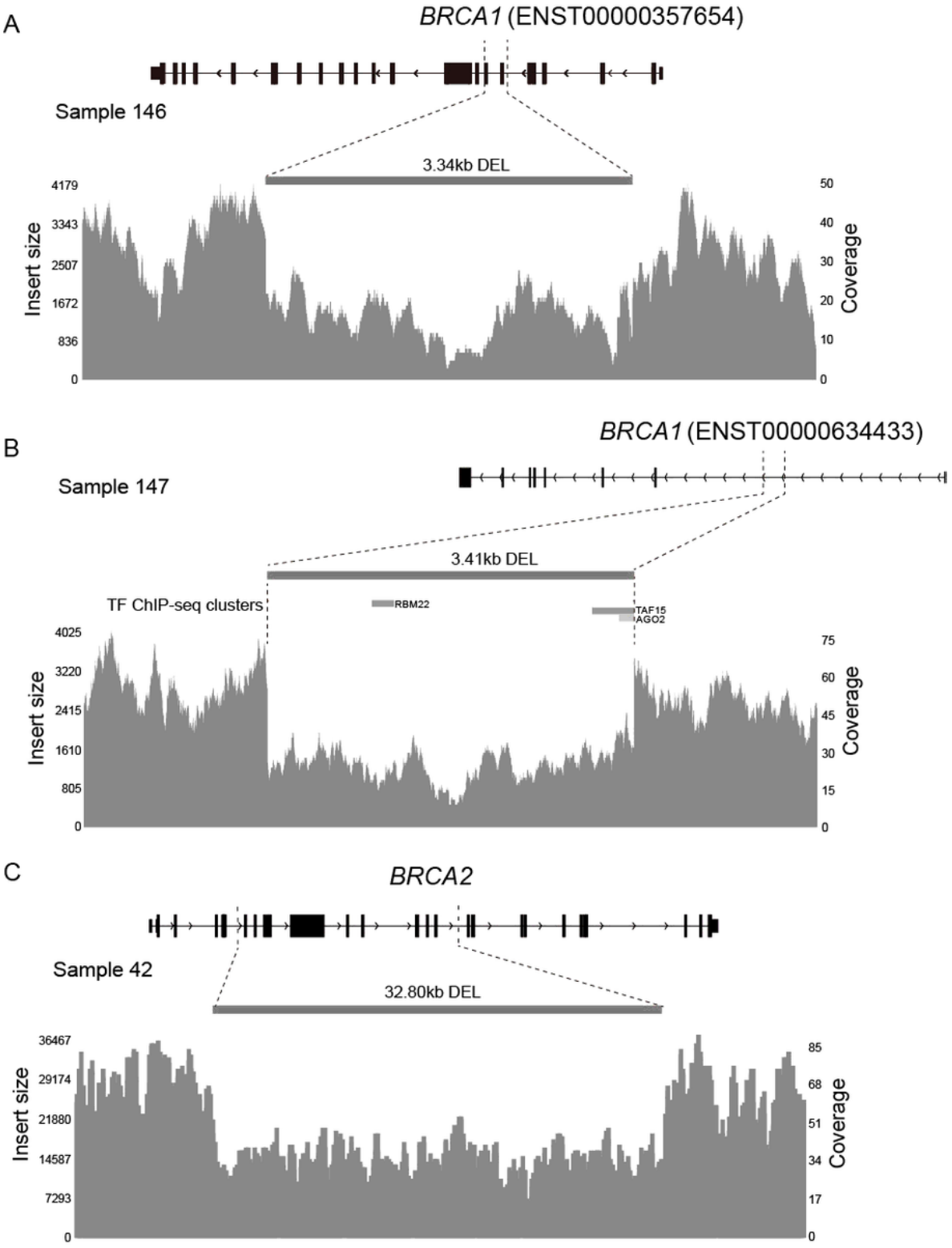

Figure 1

SV deletions in the BRCA1 and BRCA2 genes. a An SV deletion in the coding region of BRCA1 (transcript: ENST00000357654) in one patient. b An SV deletion in the intronic region of BRCA1 (transcript: ENST00000634433) in one patient. The epigenetic evidence of TF ChIP-seq clusters was observed in this region. c An SV deletion in the coding region of BRCA2 in one patient.

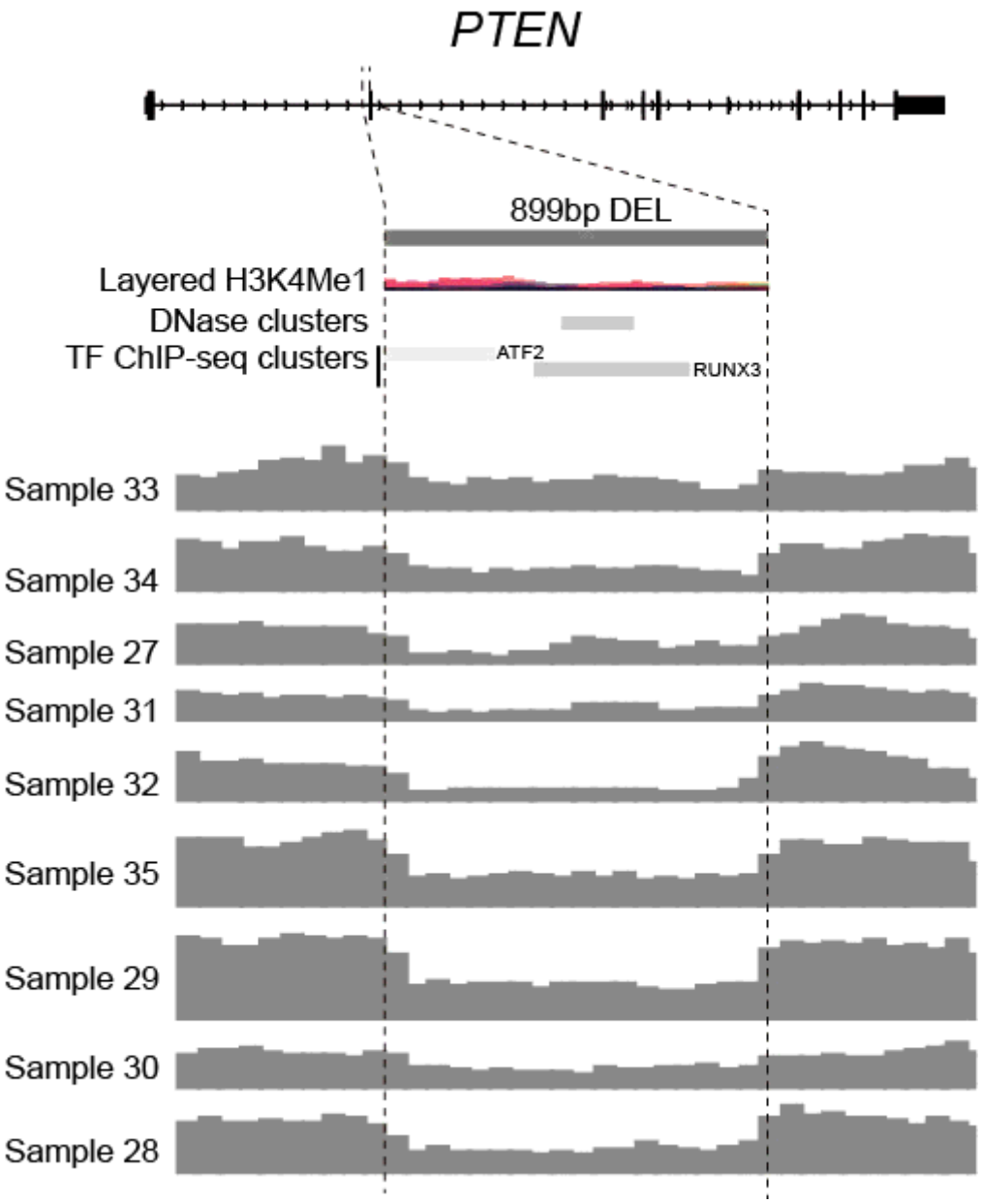

Figure 2

SV deletion in the intronic region of the PTEN gene in nine cases. From top to bottom, gene structure; layered H3K4Me1; DNase clusters; clustered ChIP-seq binding sites; The signals of layer H3K4 methylation from different ENCODE cell lines are shown in different colors.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarydata.xlsx