A comprehensive screening strategy of breast cancer based on different penetrance susceptibility genes in healthy Han Chinese females

Ling Hu¹, Chengbi Wu¹, Ping Liu¹, Lanbin Jiang¹, Yuntao Guo³, Xin Jiang³, Nian Xiong¹,²*

¹ Wuhan Red Cross Hospital, Wuhan 430000, China
² Department of Neurology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430000, China
³ Medical Laboratory of Nantong Zhongke, Nantong 226000, China

Tel: +86-27-65606269, Fax: +86-27-65606269
E-mail: huling_hh@163.com

Abstract

Background: Morbidity and mortality associated with breast cancer (BC) had increased rapidly in China. Early screening and intervention could greatly reduce the risk of hereditary BC. Several risk models had been utilized over the last decades to predict individual BC risk, but most of them didn’t assess the polygenic risk of low-penetrant genes. A novel screening method integrating different penetrance susceptibility genes was eagerly needed.

Methods: Twenty-three variants of high and moderate penetrance susceptibility genes (HVs) and twenty loci of low penetrance susceptibility genes (LVs) were selected from previous studies. Genotyping of these mutations were conducted among 3777 healthy Chinese Han women and 401 BC subjects. Based on the mutation profiles, we raised a comprehensive screening strategy using HVs and LVs to evaluate the polygenic risk score (PRS) in healthy individuals.

Results: Three HVs in BRCA1, BRCA2 and PALB genes mutated in the study population, which
suggests the necessity of applying genetic determination to healthy Han Chinese women. LVs were widely carried in objects and their frequencies differed greatly between Han Chinese females and the west population. After quality control and lasso dimension reduction, nineteen of twenty-three LVs were involved to construct a logistic regression model to evaluate the cumulative genetic risk score. The area under curve (AUC), sensitivity and specificity of the model is 0.993, 0.9676 and 0.9617 respectively, indicating that it is robust. Finally, a screening strategy using HVs and LVs was put forward to evaluate the risk of BC in normal objects.

Conclusions: The distribution of HVs and LVs differed greatly between Han Chinese females and the west population. A screening strategy combining HVs and LVs showed strong efficacy in distinguishing high risk individuals from healthy women.

Key words: breast cancer, susceptibility gene, Han Chinese, risk

Trial registration: ChiCTR, ChiCTR2000038558. Registered 24 September 2020 - Retrospectively registered, http://www.chictr.org.cn/showproj.aspx?proj=60543

Background

Breast cancer (BC) is the most common invasive cancer and the leading cause of cancer-related deaths in women worldwide and in Asian countries [1]. Morbidity and mortality associated with BC have increased rapidly in China [2]. Although the precise mechanisms underlying this heterogeneous disease have not been fully elucidated, increasing evidence indicated that genetic variants may contribute to the heritable risk of breast cancer besides the environmental and lifestyle factors [3].

Over the past 30 years, at least 170 susceptible genes or loci with different population penetration
and incidence were identified in BC studies [4]. Variants of high penetrance susceptibility genes, like BRCA1, BRCA2, TP53 et al. [5-7], increased the risk of BC by more than 5 fold changes, which explained approximately 20−25% of familial risk [8]. Further 2−3% of familiar breast cancer are caused by rare mutations with moderate penetrance in genes such as CHEK2, ATM, PALB2 et al., which increased the risk by 2−5 fold changes [9-12]. Pathogenic variants in high and moderate penetrance genes (HVs) are associated with up to 85% lifetime risk and have high individual effect, but they are quite rare in the general population [13, 14]. With the application of Genome wide association studies (GWAS), numerous common polymorphisms in low penetrance susceptibility genes (LVs) have been proved to be associated with small increases in BC risk, including FGFR2, ESR1, C5orf67, et al., which jointly explain about 30% of the remaining familial risk [15-17]. Although many carriers do not express the malignant phenotype since these low penetrant genes are associated with small risk increase individually, the genetic risk is still a threat.

Early screening and intervention can greatly reduce the risk of hereditary BC, so accurate and individualized risk assessment is essential for decisions regarding the level of prevention. Several risk models have been utilized over the last decades to predict individual BC risk, such as the Breast Cancer Risk Assessment Tool (BCRAT, also referred to as the Gail model), the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm model (BOADICEA), the International Breast Cancer Intervention Study model (IBIS) et al [18-20]. All these models assign a weight to each factor and use algorithms to calculate breast cancer risk. They differ in accuracy, discriminatory power and calibration and they all have inherent weaknesses. Some underestimate the risk involved in nulliparous women and late menarche age, while others are only suitable for specific ethnicity and age [21, 22]. Moreover, most models do not assess the polygenic risk of low-
penetrant genes. Although some studies already combine genetic and clinical information and create multiplicative models which generated increased ROC values, these models’ discriminative accuracy is still insufficient for clinical exploitation [23]. In addition, it is still unclear which current gene panel available for clinical use provides the best BC-population risk stratification.

Distinct differences exist in certain BC predisposition variants between the East and West population. For example, overall, BRCA2 mutations appear to be more common than BRCA1 mutations in the East Asian population in contrast to white populations, in which BRCA1 mutations may be as common as or more frequent than BRCA2 mutations [24]. To refine BC risk stratification, it is essential to validate these single nucleotide polymorphisms (SNPs) in specific ethnic groups before assessing the polygenic risk scores [25]. To calculate the frequency of polymorphisms associated to BC risk and evaluate the PRS in healthy Han Chinese women (HCW) based on different penetrance susceptibility genes, we analyzed the frequencies of HVs and LVs in 3777 healthy Han Chinese females and developed a comprehensive screening strategy to assess the risk of healthy individuals.

Materials and Methods

Subjects

Study subjects were recruited from the Wuhan Region of China, from September 2018 to December 2019. All subjects were healthy HCW without family history of breast cancer and provided written informed consent before participating in the study. The name, age and clinical information were collected according to the privacy standard of WuHan Red Association Hospital.
**BC related SNPs selection**

Twenty HVs were obtained from Clinvar database (https://www.ncbi.nlm.nih.gov/clinvar/, version 20200210) [26]. All mutations were selected by the criterion: 1) the mutation was pathogenic or likely pathogenic; 2) the mutation was included in GnomAD or 1000 genome database [27, 28]. Twenty-three LVs were collected from GWAS Catalog database (http://www.ebi.ac.uk/gwas/) [29], CNKI database (http://cnki.net/) and NCBI Pubmed database (https://www.ncbi.nlm.nih.gov/pubmed/). All mutations were selected by the criterion: 1) The odds ratio (OR) is above 1; 2) The p value in the related study was smaller than 1E-6; 3) The SNP was carried by Chinese; 4) There was no significant linkage disequilibrium (LD) between any two loci; 5) There were researches about HCW covering these mutations.

**DNA extraction and genotyping**

Matrix-assisted laser analytical ionization time of flight mass spectrometry (MADI-TOF MS) was used for this test. The MassARRAY system was used for genotyping. Blood Genomic DNA Mini Kit was used to extract the whole Blood DNA of the tested person and storeed in the refrigerator at -20℃. The volume of PCR reaction system was 5.0 μl, containing H2O (1.3 μl), 10*buffer(0.5 μl), MgCl2 (0.4 μl), dNTP (0.1 μl), Extension Primer Mix (1 μl), enzyme (0.2 μl) and DNA (1.5 μl). The reaction programs followed the steps: 95℃ 2 min , 95℃ 30s , 56℃ 30s , 72℃ 60 s (45 cycles), 72℃ 5min , 4℃ hold. The SAP reaction system (7.0 μl) was constructed with PCR production (5 μl), H2O (1.53 μl), SAP*buffer (0.17 μl) and SAP enzyme (0.3 μl) , and processed with 37℃ 40min , 85℃ 5min , 4℃ hold step by step. The UEP reaction system (9.0 μl) mixed by PCR&SAP production (7 μl), H2O (0.619 μl), Gold*buffer (0.2 μl), termination mix (0.94 μl) and enzyme
(0.041 μl) was parsed with three steps: 94°C 30s, 94°C 5s; 52°C 5s, 80°C 5s; 5 cycles; Step1 & Step2: 40 cycles; Step 3: 72°C 3min, 4°C hold.

Reaction products of resin were purified after PCR extension, SAP digestion and single nucleotide base extension. Genotype analysis was carried out by MALDI-TOF-MS MS time-of-flight mass spectrometry analyzer. The molecular weight of the single nucleotide extension product was analyzed by the MassARRAY system to obtain the site typing results, and the quality value of the typing results was determined according to the signal to noise ratio, peak tip position and peak width of the mass spectrum peak.

**Data analysis**

All the tables and texts were parsed by in house Perl (version 5.18) scripts. Statistical analyses and image render were performed using R software (ver.3.6.1; http://www.r-project.org/). Plink (version 1.07-x86_64) was used to analyze the linkage between loci [30], any two sites with \( r^2 > 0.5 \) were judged to be linkage disequilibrium (LD). The Hardy-Weinberg equilibrium (HWE) test was also analyzed with Plink (--hardy). Tumor dataset GSE118527 was downloaded from the NCBI Gene Expression Omnibus [31]. Genotypes of LVs in tumor were calculated by the B-Alele frequency (BAF) with the criterion: BAF<0.1 wild, 0.1≤BAF≤0.95 heterozygosis, BAF>0.95 homozygosis. Then the tumor genotype information was transformed to vcf format and imputed by Beagle software (version 5.1) [32]. The reference dataset of haplotype was obtained from 1000 genome phased data (version 2). After imputation, loci with allele frequency bigger than 0.1 were kept for further analysis. The Least Absolute Shrinkage and Selectionator operator (LASSO) was employed to select the key LVs influencing the status of testers. A publicly available R package “glmnet” was applied to perform LASSO.
for 1000 iterations [33]. The logistic model was constructed by R function “glm”. To evaluate the
generalization capability of the model, K fold cross validation (K=5) was employed using R package
“caret” [34]. The AUC and ROC were calculated by R package “pROC” [35].

Results

Information of variants and population in the study

All 3777 subjects involved were healthy Han Chinese females without familial breast cancer history.
The median age was 40 (19-86), among which 1733 were below 40 years old, and 2044 were above
40 years old. The distribution of age was shown in Fig. 1a. Total 43 sites were selected for further
study, including 20 HVs and 23 LVs (Fig. 1b). HVs belong to BRCA1, BRCA2, TP53, PALB2,
ATM and CHEK2 genes and include three mutation types which are nonsense mutation, missense
mutation and frame shift mutation. All the HVs were located in the exon region (Fig. 1c). Detailed
information was in Supplemental Materials S1. All the 23 LVs were introns or intergenic region
variants, which were derived from the case-control study of breast cancer in GwasCatalog database
(Supplemental Materials S2). The OR values of LVs range from 1 to 2, and the maximum LD
coefficient between any two LVs was 0.36 (Fig. 1d). Genotype of all objects were measured by
MassARRAY platform. All genotyping data were list in Supplemental Materials S3.
Fig. 1 Statistics of study population and candidate variants. a The distribution of objects ages. b The genome location of 20 HVs and 23 LVs. c Exon function of 23 HVs. d The Linkage Disequilibrium Analysis result of 23 LVs, $r^2$ between any two loci was less than 0.5.

**Variants of high and moderate penetrance genes**

Three in the 20 HVs were detected in the study population, as shown in Table 1. Notably rs180177110 (PALB1) had a high carrying frequency, with 10 heterozygous and 1 homozygous. In addition, two heterozygous pathogenic mutation were detected in rs80356978(BRCA1) and rs80358600(BRCA2) respectively. HVs in TP53, CHEK2 and ATM were not detected in this study.

**Table 1 High and moderate pathogenic variants detected in Chinese Han population**

| SNP ID  | HGVS                                         | ExonicFunction | Mut_Het | Mut_Hom | MAF    |
|---------|----------------------------------------------|----------------|---------|---------|--------|
| rs180177110 | PALB2:NM_024675:c.2257C>T:p.R753X | stopgain       | 10      | 1       | 0.001589 |
| rs80356978  | BRCA1:NM_007297:                | stopgain       | 1       | 0       | 0.000132 |
ExonicFunction was the impact on protein of a SNP. Mut_Het and Mut_Hom were the total number of heterozygous and homozygous individuals separately. MAF was the frequency of a mutation allele.

**Variants of low penetrance genes**

Based on the genotype information of the study population, 23 LVs were tested for HWE test, among which rs10941679, rs7072776 and rs1436904 did not conform to the law (BH adjust p value < 0.05), as shown in Supplemental Materials S4.

**Fig. 2** Frequencies of LVs in different population. a The correlation of LVs’ frequencies between the east and west populations. b The differences of LVs’ frequency profiles between the east and west populations.

The allele frequencies of the 20 LVs meeting the HWE were highly correlated with the East Asian population frequency in 1000 genomes and gnomAD database ($r^2 > 0.98$), but differed greatly with the west populations ($r^2 < 0.6$). The frequencies of rs10474352, rs3960987 and rs9485372 were all over 0.2 higher than the mean of the west populations, while the frequency profiles of rs2380205,
rs889312 and rs7107217 was significantly lower than that of the west populations (shown in Table 2).

| SNP ID     | Han mean | difference | p.value  | p.adjust  |
|------------|----------|------------|----------|-----------|
| rs6788895  | 0.384432 | 0.319332   | 0.08846887 | 0.08846887 |
| rs10474352 | 0.48663  | 0.31461    | 0.0086707  | 0.00963411 |
| rs3760982  | 0.860604 | 0.284444   | 7.82E-05   | 0.00017366 |
| rs9485372  | 0.42547  | 0.22727    | 0.0019622  | 0.00245275 |
| rs12355688 | 0.251787 | 0.173067   | 0.01486858 | 0.01565114 |
| rs6001930  | 0.236431 | 0.115531   | 0.00095601 | 0.00147079 |
| rs10411161 | 0.292295 | 0.105855   | 0.00030353 | 0.00050588 |
| rs9485372  | 0.48663  | 0.17202    | 0.0086707  | 0.00963411 |
| rs3760982  | 0.860604 | 0.284444   | 7.82E-05   | 0.00017366 |
| rs12355688 | 0.251787 | 0.173067   | 0.01486858 | 0.01565114 |
| rs6001930  | 0.236431 | 0.115531   | 0.00095601 | 0.00147079 |
| rs10411161 | 0.292295 | 0.105855   | 0.00030353 | 0.00050588 |
| rs9485372  | 0.48663  | 0.17202    | 0.0086707  | 0.00963411 |
| rs3760982  | 0.860604 | 0.284444   | 7.82E-05   | 0.00017366 |

Construction of a logistic model based on LVs to predict genetic risk score of BC

Through analyzing the microarray data of GSE118527 breast cancer patients, high-quality genotype information of 217612 SNP loci of 401 BC patients was gained. Genotype imputation was performed using 1000 genome data as a reference to obtain genome-wide SNP information. After filtering out the site which do not fit HWE or of which MAF is less than 0.05 , 19 high quality LVs were remained to be used for subsequent analysis. With the genotype of 401 BC patients as the test dataset and objects older than 40 years old (1880) as the training set, 1000 iterations lasso regression
were conducted to address the major independent variables. The results showed that the numbers of LVs involved in the model range from 8 to 19 and the 19 LVs models have the highest proportion, detail in Supplemental Materials S5. Therefore, all 19 LVs were included to construct the logistic regression model, with genotype of LVs as independent variables and the status of disease as dependent variable. The threshold value \( y = 0.215 \) was calculated with the max AUC. With this threshold, the classifier works the best with balanced sensitivity and specificity (0.962 and 0.968) (Fig. 3a). Further more, in order to evaluate the generalization ability of the model, we employed the 5-fold cross-validation and randomly took eighty percent data once for testing. After 10 iterations, the average performance parameter of the classifier was calculated. It’s suggested that the average sensitivity of the model was 0.9676±0.0023, the specificity was 0.9617±0.0039, and the AUC was 0.993±0.0006, indicating a robust classification effect. The top three significant factors in the model were rs3760982, rs2046210 and rs10474352, detail in Fig. 3b. The polygenic risk scores of all objects were calculated by the logistic model and listed in Supplemental Materials S6.

A comprehensive screening strategy using HVs and LVs to predict PRS of BC

A comprehensive screening strategy was adopted to decide the PRS to objects (Fig. 3c). The PRS of testers was defined to be 1 whenever the HVs were detected, no matter it’s heterozygous or homozygous, further genetic consultation would be provided for these ones (13/3777). Afterwards, PRS of testers with the wild genotype of HVs would be assessed by the LVs logistic model. Ones with high PRS (148/3777) were divided to high risk group, continuous follow-up would be carried
out to determine the accuracy of the strategy.

Fig. 3 Correlation of LVs’ frequencies in different population and logistic model of LVs in BC

Discussion

To investigate the distribution of HVs, genotypes of twenty HVs were detected in 3777 healthy HCW, which locate in exons of BRCA1, BRCA2, TP53, CHEK2 and ATM genes. In this study population, rs80356978(BRCA1) and rs80358600(BRCA2) were detected. Pathogenic mutations in BRCA1 and BRCA2 genes are associated with a 10 to 20 fold increased risk of breast cancer, which corresponds to a cumulative risk of 55–65% of developing breast cancer by 70 years of age for BRCA1 mutation carriers and 45–47% for BRCA2 mutation carriers [36-38]. It should be noted
that rs180177110 (PALB2) mutated in 11 women, including ten heterozygote and one homozygote, which is a pathogenic variant that may contribute to over 4fold more risk of BC for carriers than normal. PALB2, a partner and localizer of BRCA2, was identified to be a definite susceptibility gene of BC [39]. Loss-of-function mutations in PALB2 are an important cause of hereditary breast cancer, with respect both to the frequency of cancer-predisposing mutations and to the risk associated with them. The BC risk for PALB2 mutation carriers may overlap with that for BRCA2 mutation carriers [40]. Frequency of the site in gnomAD database is 0.000024, while it’s relatively high (0.001589) in Han Chinese population. It suggested that we should pay close attention to this loci in the screening of BC in Han Chinese. Other HVs in TP53, CHEK2 and ATM were not detected in study population, indicating they might be rare carried in HCW. In summary, we detected several pathogenic mutations in BRCA1, BRCA2 and PALB2 genes in relation to breast cancer risk in HCW, suggesting that it’s necessary to conduct genetic screening in healthy women. Patients with pathogenic variants in any of these genes are eligible for increased surveillance for cancer or other preventive measures.

LVs were widely carried in HCW with frequency ranging form 0.005957 to 0.860604. As is expected, the mutation profiles of Han are similar to frequencies of the East Asia population in 1000 genome and gnomAD project, but vastly differ with the west populations. Several LVs change a lot between the east and the west. For example, the frequencies of rs10474352, rs3960987 and rs9485372 are all more than 0.2 higher than the average of the west populations, while rs2380205, rs889312 and rs7107217 have significantly lower mutation rate than that of the west population. The BAF of rs2380205 and rs889312 in the GWAS study of LiHai et al [41] about BC in Chinese women is consistent with us [41]. The BAF of rs10474352 (ARRDC3) is also the same with the
study of SBCGS-1 population which is a large scale GWAS study of Chinese BC [42]. It infers that our statistical results are credible.

Combined associations of SNPs can be summarized by a polygenic risk score. Women in the top 1% of the newly derived 313-SNP polygenic risk score have a four-fold increased risk of breast cancer than women at population-average risk [43]. In this study, a logistic model of 19 high quality LVs was constructed to predict the genetic risk score of objects with high sensitivity and specificity. Compared with other risk assessment models, our PRS model shows higher AUC and detection efficiency in Han Chinese population [44-46]. rs3760982-A has the highest coefficients in the model, which locate in the 1.1 kb upstream of KCNN4 genes on chromosome 19q13.31 and has the highest allele frequency (0.86) among LVs. Liu Y et al found that its A allele is predicted to enhance RUNX binding and effect the expression of ZNF404. The A/A genotype at rs3760982 is predicted to increase RUNX binding and is correlated with higher expression of ZNF404 in breast cancer tumor samples and breast tissue [47]. The rs3760982-A allele is associated with an increased risk (odds ratio of 1.06) of breast cancer [48]. Thus higher expression of ZNF404 correlates with increased breast cancer risk. Another study focusing on the eQTLs of BC showed rs3760982 was closely related with the expression of ZNF155 [49]. All evidence indicated that rs3760982 may affect the transcript regulation of ZNF gene family in BC. In a GWAS study on Chinese women, Zheng W et al found that rs2046210 on chromosome 6q25.1 showed a strong correlation with BC [16]. The OR (odds ratio) adjusted for AG and AA genotypes were 1.36 and 1.59, respectively, and a similar correlation was found in the European population. This locus is located upstream of estrogen receptor 1 (ESR1), which encodes estrogen receptor a (ERa) regulating signal transduction of estrogen, a sex hormone that has a central role in the etiology of BC. Elevated estrogen levels have
been shown to be associated with an increased risk of breast cancer in previous studies [50]. It is hypothesized that rs2046210 together with ESR1 may play an essential role in estrogen-dependent mammary epithelial proliferation, which is associated with increased mammographic density and subsequent risk to BC [51-53]. Other loci, such as rs10474352 [54], rs6788895 [55], rs7107217 et al [56], were also identified in previous studies in HCW and have the potential to be a predictive variable to assess the risk of BC.

In this study, we performed a comprehensive assessment of PRS based on HVs and LVs for breast cancer susceptibility in HCW. The genetic screening for breast cancer can be carried out by two steps: the first stage is high penetrance mutation screening, to determine whether there is a need for special detection of breast cancer and long-term disease prevention and tracking, so as to make breast cancer screening more targeted. In the second stage, low-penetrance mutation screening can effectively reduce the risk of disease by paying close attention to the improvement of living habits and living environment and conducting routine physical examination for people with more mutations [57]. With the development of molecular techniques, the next generation sequencing technology would be more and more commonly used. Relevant disease data generated in hospital can make the breast cancer health data more complete and the risk assessment of disease more accurate. However, many problems remain to be solved when it comes to use genetic testing for breast cancer screening work and it requires joint efforts of various aspects.

The main objective of our studies is to evaluate the mutation profile of HVs and LVs in HCW, and provide a novel synthetic screening strategy to assess the risk in healthy individuals. It should be noted that this study has examined only PRS in the study population. In fact, many other factors, like age, lifestyle, environment et al., also play a significant role in the developing of BC. Further
and large scale screening test and continuously follow-up should be done to check the effect and accuracy of the strategy.

Conclusions

The distribution of selected HVs and LVs differed vastly between Han Chinese population and the west population. A comprehensive screening strategy based on HVs and LVs might have powerful efficacy in distinguishing high risk individuals from healthy people.

Additional file

S1 : Mutations of high and moderate penetrance genes related HVs.
S2 : Mutations of low penetrance genes related LVs.
S3 : The genotyping information of study objects.
S4 : The HWE test results.
S5 : The lasso results.
S6 : The PRS score of study objects.

Abbreviations

BC : breast cancer; HVs : variants of high and moderate penetrance susceptibility genes; LVs: loci of low penetrance susceptibility genes; PRS : polygenic risk score; HCW : healthy Chinese Han women; GWAS : Genome wide association studies; SNP : single nucleotide polymorphisms; OR : odd ratio; BAF : B-Allele frequency; LASSO : least absolute shrinkage and selectionator operator; LD : linkage disequilibrium.

Authors’ contributions

Nian Xiong designed the study; Ling Hu, Chengbi Wu, Ping Liu and Lanbin Jiang performed the experiments and prepared the figures; Ling Hu, Yuntao Guo and XinJiang contributed to drafting the manuscript. All authors read and approved the final manuscript.
Authors details

1 Wuhan Red Cross Hospital, Wuhan 430000, China
2 Department of Neurology, Union Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430000, China
3 Medical Laboratory of Nantong Zhongke, Nantong 226000, China

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Written informed consent was obtained from all study participants. For minors/children, it was obtained from next of kin, caretakers, or guardians by written form. The protocol was approved by the Wuhan Red Cross Hospital’s ethics committee.

Funding

This work was supported by grants 2016YFC1306600 (to NX) and 2018YFC1314700 (to NX) from the National Key R&D Program of China, grants 81873782 (to NX) from the National Natural Science Foundation of China, Grant 2016CFB624 from Natural Science Foundation of Hubei Province (to NX), Grant 2017050304010278 from The Youth Science and technology morning light program of Wuhan City (to NX), 2018 Hubei medical research project WJ2019F030 (to NX), 2018 Wuhan medical research project S201802140011 (to NX), 2018 Wuhan Young and Middle-aged medical Talents Program (to NX) and 2017 Hubei provincial Party Committee Organization Department the second batch of Hubei youth elite development plan (to NX).
References

1. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2019. *CA Cancer J Clin* 2019, 69(1):7-34.

2. Li T, Mello-Thoms C, Brennan PC: Descriptive epidemiology of breast cancer in China: incidence, mortality, survival and prevalence. *Breast Cancer Res Treat* 2016, 159(3):395-406.

3. Fachal L, Dunning AM: From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. *Curr Opin Genet Dev* 2015, 30:32-41.

4. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J, Bolla MK, Shu XO, Lu Y, Cai Q et al: A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat Genet* 2018, 50(7):968-978.

5. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, Ding W et al: A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 1994, 266(5182):66-71.

6. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Manigton J, Collins N, Gregory S, Gumbs C, Micklem G: Identification of the breast cancer susceptibility gene BRCA2. *Nature* 1995, 378(6559):789-792.

7. Malkin D, Li FP, Strong LC, Fraumeni JF, Jr., Nelson CE, Kim DH, Kassel J, Gryka MA, Bischoff FZ, Tainsky MA et al: Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* 1990, 250(4985):1233-1238.

8. Melchor L, Benitez J: The complex genetic landscape of familial breast cancer. *Hum Genet* 2013, 132(8):845-863.

9. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K et al: ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet* 2006, 38(8):873-875.

10. Meijers-Heijboer H, van den Ouweland A, Klijn JG, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M et al: Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat Genet* 2002, 31(1):55-59.

11. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T et al: PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet* 2007, 39(2):165-167.

12. Sud A, Kinnersley B, Houlston RS: Genome-wide association studies of cancer: current insights and future perspectives. *Nat Rev Cancer* 2017, 17(11):692-704.

13. Thompson D, Easton DF, Breast Cancer Linkage C: Cancer Incidence in BRCA1 mutation carriers. *J Natl Cancer Inst* 2002, 94(18):1358-1365.

14. Breast Cancer Linkage C: Cancer risks in BRCA2 mutation carriers. *J Natl Cancer Inst* 1999, 91(15):1310-1316.

15. Han MR, Long J, Choi JY, Low SK, Kweon SS, Zheng Y, Cai Q, Shi J, Guo X, Matsuo K et al: Genome-wide association study in East Asians identifies two novel breast cancer susceptibility loci. *Hum Mol Genet* 2016, 25(15):3361-3371.

16. Zheng W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL et al: Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet* 2009, 41(3):324-328.
Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, Schmidt MK, Chang-Claude J, Bojesen SE, Bolla MK et al: Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet 2013, 45(4):353-361, 361e351-352.

Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ: Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989, 81(24):1879-1886.

Antoniou AC, Cunningham AP, Petö J, Evans DG, Lalloo F, Narod SA, Risch HA, Eyfjord JE, Hopper JL, Southey MC et al: The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. British Journal of Cancer 2008, 98(8):1457-1466.

Tyer J, Duffy SW, Cuzick J: A breast cancer prediction model incorporating familial and personal risk factors. Stat Med 2004, 23(7):1111-1130.

Terry MB, Liao Y, Whitemore AS, Leoece N, Buchsbaum R, Zeinomar N, Dite GS, Chung WK, Knight JA, Southey MC et al: 10-year performance of four models of breast cancer risk: a validation study. The Lancet Oncology 2019, 20(4):504-517.

Amir E, Freedman OC, Seruga B, Evans DG: Assessing women at high risk of breast cancer: a review of risk assessment models. J Natl Cancer Inst 2010, 102(10):680-691.

Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, Tyer JP, Chen TH, Wang Q, Bolla MK et al: Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet 2019, 104(1):21-34.

Ashour M, Ezzat Shafik H: Frequency of germline mutations in BRCA1 and BRCA2 in ovarian cancer patients and their effect on treatment outcome. Cancer Manag Res 2019, 11:6275-6284.

Yap YS, Lu YS, Tamura K, Lee JE, Ko EY, Park YH, Cao AY, Lin CH, Toi M, Wu J et al: Insights Into Breast Cancer in the East vs the West: A Review. JAMA Oncol 2019.

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W et al: ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 2018, 46(D1):D1062-D1067.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB et al: Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016, 536(7616):285-291.

Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA et al: A global reference for human genetic variation. Nature 2015, 526(7571):68-74.

Buniello A, MacArthur JAL, Cezaro M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E et al: The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res 2019, 47(D1):D1005-D1012.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007, 81(3):559-575.

Jiang YZ, Ma D, Suo C, Shi J, Xue M, Hu X, Xiao Y, Yu KD, Liu YR, Yu Y et al: Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies. Cancer Cell 2019, 35(3):428-440 e425.

Browning BL, Zhou Y, Browning SR: A One-Penny Imputed Genome from Next-
Generation Reference Panels. *Am J Hum Genet* 2018, **103**(3):338-348.

33. Friedman J, Hastie T, Tibshirani R: Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010, **33**(1):1-22.

34. Kuhn M: Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* 2008.

35. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Muller M: pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011, **12**:77.

36. Kurian AW: BRCA1 and BRCA2 mutations across race and ethnicity: distribution and clinical implications. *Curr Opin Obstet Gynecol* 2010, **22**(1):72-78.

37. Sun J, Meng H, Yao L, Lv M, Bai J, Zhang J, Wang L, Ouyang T, Li J, Wang T et al: Germline Mutations in Cancer Susceptibility Genes in a Large Series of Unselected Breast Cancer Patients. *Clin Cancer Res* 2017, **23**(20):6113-6119.

38. Kwong A, Shin VY, Ho JC, Kang E, Nakamura S, Teo SH, Lee AS, Sng JH, Ginsburg OM, Kurian AW et al: Comprehensive spectrum of BRCA1 and BRCA2 deleterious mutations in breast cancer in Asian countries. *J Med Genet* 2016, **53**(1):15-23.

39. Chen P, Liang J, Wang Z, Zhou X, Chen L, Li M, Xie D, Hu Z, Shen H, Wang H: Association of common PALB2 polymorphisms with breast cancer risk: a case-control study. *Clin Cancer Res* 2008, **14**(18):5931-5937.

40. Antoniou AC, Casadei S, Heikkinen T, Barrowdale D, Pylkas K, Roberts J, Lee A, Subramanian D, De Leeneer K, Fostira F et al: Breast-cancer risk in families with mutations in PALB2. *N Engl J Med* 2014, **371**(6):497-506.

41. Li H, Beeghly-Fadiel A, Wen W, Lu W, Gao YT, Xiang YB, Cai Q, Long J, Shi J, Chen K et al: Gene-environment interactions for breast cancer risk among Chinese women: a report from the Shanghai Breast Cancer Genetics Study. *Am J Epidemiol* 2013, **177**(2):161-170.

42. Cai Q, Zhang B, Sung H, Low SK, Kweon SS, Lu W, Shi J, Long J, Wen W, Choi JY 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**(8):886-890.

43. Kapoor PM, Mavaddat N, Choudhury PP, Wilcox AN, Lindstrom S, Behrens S, Michaillidou K, Dennis J, Bolla MK, Wang Q et al: Combined associations of a polygenic risk score and classical risk factors with breast cancer risk. *J Natl Cancer Inst* 2020.

44. Pal Choudhury P, Wilcox AN, Brook MN, Zhang Y, Ahearn T, Orr N, Coulson P, Schoemaker MJ, Jones ME, Gail MH et al: Comparative Validation of Breast Cancer Risk Prediction Models and Projections for Future Risk Stratification. *J Natl Cancer Inst* 2020, **112**(3):278-285.

45. Lee A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, Babb de Villiers C, Izquierdo A, Simard J, Schmidt MK et al: BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors. *Genet Med* 2019, **21**(8):1708-1718.

46. Dankova Z, Zubor P, Grendar M, Zelnova K, Jagelkova M, Stastny I, Kapinova A, Vargova D, Kasajova P, Dvorska D et al: Predictive accuracy of the breast cancer genetic risk model based on eight common genetic variants: The BACKSIDE study. *J Biotechnol* 2019, **299**:1-7.

47. Liu Y, Walavalkar NM, Dozmorov MG, Rich SS, Civelek M, Guertin MJ: Identification of
breast cancer associated variants that modulate transcription factor binding. *PLoS Genet* 2017, 13(9):e1006761.

48. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, Lemacon A, Soucy P, Glubb D, Rostamianfar A et al: Association analysis identifies 65 new breast cancer risk loci. *Nature* 2017, 551(7678):92-94.

49. Li Q, Stram A, Chen C, Kar S, Gayther S, Pharoah P, Haiman C, Stranger B, Kraft P, Freedman ML: Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. *Hum Mol Genet* 2014, 23(19):5294-5302.

50. Key T, Appleby P, Barnes I, Reeves G, Endogenous H, Breast Cancer Collaborative G: Endogenous sex hormones and breast cancer in postmenopausal women: reanalysis of nine prospective studies. *J Natl Cancer Inst* 2002, 94(8):606-616.

51. Mariapun S, Ho WK, Kang PC, Li J, Lindstrom S, Yip CH, Teo SH: Variants in 6q25.1 Are Associated with Mammographic Density in Malaysian Chinese Women. *Cancer Epidemiol Biomarkers Prev* 2016, 25(2):327-333.

52. Li X, Yao GY, Li FX, Qiu YR, Yang XX: Validation of associations between ESR1 variants and breast cancer risk in Chinese cohorts. *Genet Mol Res* 2016, 15(3).

53. He Y, Liu H, Chen Q, Sun X, Liu C, Shao Y: Relationship between five GWAS-identified single nucleotide polymorphisms and female breast cancer in the Chinese Han population. *Tumour Biol* 2016, 37(7):9739-9744.

54. Xu M, Xu Y, Chen M, Li Y, Li W, Zhu J, Zhang M, Chen Z, Zhang X, Liu J et al: Association study confirms two susceptibility loci for breast cancer in Chinese Han women. *Breast Cancer Res Treat* 2016, 159(3):433-442.

55. Zhang B, Li Y, Zheng X, Zuo X, Zhou F, Liang B, Zhu J, Li P, Ding Y, Huang Z et al: A common variant in the SIAH2 locus is associated with estrogen receptor-positive breast cancer in the Chinese Han population. *PLoS One* 2013, 8(11):e79365.

56. Long J, Cai Q, Sung H, Shi J, Zhang B, Choi JY, Wen W, Delahanty RJ, Lu W, Gao YT et al: Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer. *PLoS Genet* 2012, 8(2):e1002532.

57. Tung NM, Boughley JC, Pierce LJ, Robson ME, Bedrosian I, Dietz JR, Dragun A, Gelpi JB, Hofstatter EW, Isaacs CJ et al: Management of Hereditary Breast Cancer: American Society of Clinical Oncology, American Society for Radiation Oncology, and Society of Surgical Oncology Guideline. *J Clin Oncol* 2020, 38(18):2080-2106.