

Original Research

Non-Invasive Detection of Breast Cancer by Low-Coverage Whole-Genome Sequencing from Plasma

Li Peng^{1,†}, Ru Yao^{1,†}, Sihang Gao^{2,†}, Yang Qu¹, Li Qu², Jingbo Zhang², Yidong Zhou^{1,*}¹Department of Breast Surgery, Peking Union Medical College Hospital, Chinese Academy of Medical Science, 100026 Beijing, China²Department of Clinical Investigation, Beijing USCI Medical Laboratory, 100024 Beijing, China*Correspondence: zhouydpumch@126.com (Yidong Zhou)

†These authors contributed equally.

Academic Editor: Luca Roncati

Submitted: 7 January 2023 Revised: 30 March 2023 Accepted: 3 April 2023 Published: 26 July 2023

Abstract

Background: Breast cancer is the most common cancer in women worldwide. Here we aimed to develop an effective non-invasive method to screen for breast cancer and reduce mortality while still being curable. **Methods:** Here we propose a method that leverages the available data by incorporating information on copy number variations, mutation signature, and fragment size. Our approach adopted principal component analysis and a generalized linear model algorithm to distinguish between breast cancer and normal samples. **Results:** A total of 100 samples (85 tumor, 15 controls) were used for training, and 44 samples (37 tumor, 7 controls) were used to validate the proposed method based on whether the sample originated from breast cancer. Our model reached an area under the receiver operating characteristic curve reached 1.0 and 0.690 in the training set and in the validation set, respectively. **Conclusions:** Our method can differentiate between breast cancer patients and controls using non-invasive, cost-effective, low-coverage whole-genome sequencing technology that may provide new ideas for future breast cancer screenings.

Keywords: breast neoplasms; generalized linear model; non-invasive diagnosis; whole-genome sequencing

1. Introduction

Breast cancer (BC) has become the leading cause of cancer-related death among women worldwide, with an estimated more than 2 million new cases annually [1]. The incidence of BC has increased in China and the US [2,3]. Considering population expansion, experts estimate that by 2050, there will be roughly 3.2 million additional BC cases annually worldwide [4]. More importantly, the average age at the diagnosis of BC is decreasing worldwide [5]. Many factors may contribute to the increasing number and increasingly younger trend of patients with BC, including family history, individual differences, and living environment [6]. In order to reduce the mortality rate of breast cancer, early diagnosis and treatment are urgently needed, as breast cancer is prone to dissemination in the middle and late stages [7].

Most BC patients are not diagnosed until the disease has spread throughout their bodies [8]. However, breast tumours can concentrate in certain parts of the human body during their natural cycle, allowing them to be diagnosed and treated earlier [9]. BC mortality can be reduced by 40% in medium-risk women who undergo early screening, resulting in improved survival [10]. Based on these reports, the death rate from BC decreased by 40% from 1989 to 2017 [3], which can be attributed to advancements in treatment and early detection [11]. The purpose of early screening for breast tumours is to discover them at a stage when ther-

apy is more likely to be effective, giving patients a greater possibility of long-term survival [12]. In addition, both traditional and novel therapeutic approaches, including resection, irradiation, and pharmacological treatments, are more effective when breast tumour density is minimal [7–9].

Early screening methods have been examined as possible strategies to address this clinical gap. Among the most common screening methods, modern screening digital mammography has a sensitivity of 86.9%; however, not all types of BC are detectable on mammography [13], while certain breast tumours grow along with the natural breast architecture, making them difficult to detect using mammography [14]. The high false positive rate of mammography, particularly in women who are young or have thick breasts, results in unnecessary additional testing such as biopsies [15,16]. According to recent reports, indicators based on methylation analysis have been used for the early diagnosis of BC [17]. However, numerous methylation-based methods use distinct locus or Cytosine-phosphate-Guanine (CpG) site mechanization with a group of non-restrictive controls in which the methylation features of benign and cancerous tumours are mainly revealed [18]. Circulating tumour DNA (ctDNA) analysis based on whole-genome mutations with next-generation sequencing (NGS) has been used to predict early recurrence and guide drug treatment of a variety of malignancies. Unfortunately, the reliability of mutation-based ctDNA identification in BC is hindered by the lack of common variants [19]. Thus, an ef-



fective, fast, and economical method for the early detection of BC is urgently needed.

To address these problems, we attempted to develop a cost-effective early screening method for BC. In general, whole-genome sequencing (WGS) was performed on the patient's peripheral blood and the principal component analysis (PCA)-based method was applied to discriminate between tumour and normal cells. Specifically, for each sample, we built a vector that contained copy number variation (CNV) indications, error rates, and fragment sizes and used it for PCA training in which we extracted the first five principal components for each sample and used them to classify normal and tumour cells. The experimental results showed that our early screening algorithm performed well for BC.

2. Methods

2.1 Patient Selection and Ethics Statement

In this study, 144 samples (122 BC, 22 normal controls) were analysed. All participants were recruited from the Peking Union Medical College Hospital from April 2018 to June 2018. BC was diagnosed according to the National Comprehensive Cancer Network Clinical Practice Guidelines in Oncology for BC. Age, pathological type, and clinical stage information of BC cases were obtained from medical records (Table 1). Written informed consent was obtained from all patients. This study was approved by the Ethics Committee of Peking Union Medical College Hospital (HS-1949) and conducted according to the principles of the Declaration of Helsinki.

Patients were selected based on the following criteria: (1) age of 18 or above; (2) diagnosed as breast cancer; (3) a peripheral blood sample of 14–20 mL can be collected; (4) voluntary informed consent. Patients were excluded according to the following criteria: (1) with unqualified nucleic acid quality after DNA extraction; (2) diagnosed with a second primary malignant tumor; (3) who had previously received transplant surgery; (4) who had received an allogeneic blood transfusion or immunotherapy that may introduce foreign DNA.

2.2 Library Preparation and NGS

Peripheral blood was collected in Streck tubes (Streck, Inc., Omaha, NE, USA) and processed within 72 h of sample collection. The plasma was then isolated by centrifugation at 4 °C for 10 min at 1600 ×g, followed by centrifugation at 16,000 ×g for 10 min. A cell-free DNA (cfDNA) extraction from 4 mL of plasma was then performed using a QIAamp® Circulating Nucleic Acid Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. DNA concentrations were determined using a Qubit dsDNA High-Sensitivity Assay kit (Q32854, Invitrogen, Carlsbad, CA, USA).

The cfDNA from each sample was used for the library construction using a KAPA Hyper Prep Kit (KK8504,

Table 1. Clinical characteristics of breast cancer cases.

Characteristic	N (%)
Age (years)	
≥60	25 (20.49)
40–60	63 (51.64)
<40	34 (27.87)
Pathological type	
Benign	65 (53.28)
Luminal A	19 (15.57)
Luminal B	26 (21.31)
Her2+	7 (5.74)
TNBC	5 (4.10)
Clinical stage	
Benign	65 (53.28)
I	27 (22.13)
II	19 (15.57)
III	5 (4.10)
NA	6 (4.92)

TNBC, triple negative breast cancer; NA, not available.

KAPA Biosystems, Wilmington, MA, USA) according to the manufacturer's instructions. The final polymerase chain reaction (PCR) products were analysed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), Library Quant Kit (KK4824, Illumina GA revised primers-SYBR Fast Universal, KAPA Biosystems, Wilmington, MA, USA), and an ABI 7500 real-time PCR system (Life Technologies, Paisley, UK) to estimate insert size and concentration.

After quality control, the pooled library was sequenced using a NextSeq500 analyser (Illumina, San Diego, CA, USA) to generate 75-bp paired-end reads. The read depth of the entire genome was nearly 1 ×.

2.3 Patients' Characteristics and Vector Construction

2.3.1 Bioinformatics Analysis

The FASTQ files were generated using bcl2fastq2 (v.2.17.1.14, Illumina, San Diego, CA, USA). Raw reads containing the P5/P7 adapters were trimmed. Reads that contained more than one N or five consecutive nucleotides with an average Phred score of <20 were removed. The raw reads were mapped to the human reference genome GRCh37 (<https://hgdownload.soe.ucsc.edu/goldenPath/hg19/chromosomes/>) using the BWA MEM algorithm (v.0.7.17; <http://bio-bwa.sourceforge.net/>). PCR duplicates were removed using Picard v.1.119 (<https://broadinstitute.github.io/picard/>), and only uniquely mapped reads were retained for further analysis. Each Binary Alignment/Map (BAM) file was converted to Sequence Alignment/Map (SAM) using SAMtools (version 1.7, <https://github.com/samtools/samtools>) and then filtered using AWK to retain mutant reads containing only a single point mutation. SAMtools mpileup was used to identify the point muta-

tions. ANNOVAR software (version 2018-04-16, <https://annovar.openbioinformatics.org/en/latest/>) was used to annotate the variants using RefSeq (<http://www.ncbi.nlm.nih.gov/refseq/>) and dbSNP 151 (<https://ncbi.nlm.nih.gov/snp/>). Mutations were annotated as concordant or discordant. Indels were removed from the variant call format (VCF) mutation using grep. Annotated and filtered VCF files were read into R (version 4.1.2; <https://www.rstudio.com/products/rstudio/download/>), and mutations were annotated with single-base substitution contexts using the MutationalPatterns package (version 1.10.0; <http://bioconductor.org/packages/MutationalPatterns>).

2.3.2 CNV Vector

We used 100 Kb as the optimal bin size and overlapping bins with two adjoining bins overlapped by 50 Kb. Bins that carried putative CNV at the population level were removed. We controlled for GC and chromosomal biases. We applied a smoothing spline method to remove the GC bias. We then used 15 samples from healthy people to compute the mean and variance of the GC-corrected read number in each bin and regressed out the mean using weighted linear regression, with the inverse of the variance as the weight to obtain the bin dosages of the reference and test samples. After normalization, the bin dosages were used as the CNV vector.

2.3.3 Mutation Signature Vector

Several studies have shown that rearrangement mutational signatures are correlated with clinical characteristics, such as C > T transitions at the NCG, and C > T and T > C have been associated with age. Based on these studies, we inferred that the ratio of each mutational signature may be informative, indicating whether the samples came from patients with BC. We selected loci with a low population frequency (<1% or >99%), obtained counts of each mutational signature, and then computed the ratios of each mutational signature as the mutation signature vector.

2.3.4 Fragment Size Vector

In addition, several studies have shown that plasma cfDNA concentrations and fragment sizes can be used to differentiate between cancer and normal samples and thus be reflected in clinical characteristics. As our data were obtained from low-coverage WGS, it was difficult to deduce the cfDNA concentrations. Thus, we considered only cfDNA fragment size. As previous studies demonstrated, the main difference in the fragment size distribution between cancer and control samples was the second peak. Therefore, we computed the fragment size density among the second peaks (approximately 251–450 bp) as the fragment size vector.

2.3.5 Vector Construction

We combined the CNV, mutation signature, and fragment size vectors as the entire sample vector. The range of each feature varies significantly. We normalized the data and considered only the degree of feature deviation to make the features of the vector comparable to one another (Fig. 1).

$$x = \left(\frac{x - \text{mean}(x)}{s.d.(x)} \right)^2$$

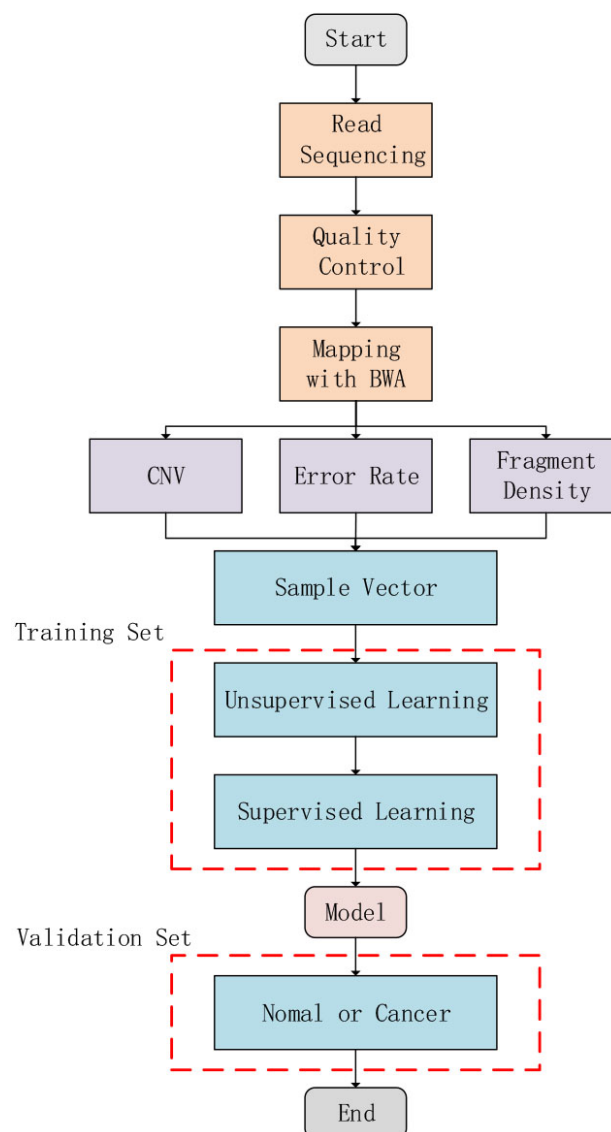


Fig. 1. Schematic illustration of study design. Three different feature types, including fragment size ratio, error rate, and copy number variation (CNV), were calculated using mapped sequencing reads. The most important five principal components were chosen, which was subsequently used by a general linearized model (GLM) algorithm to train the final model.

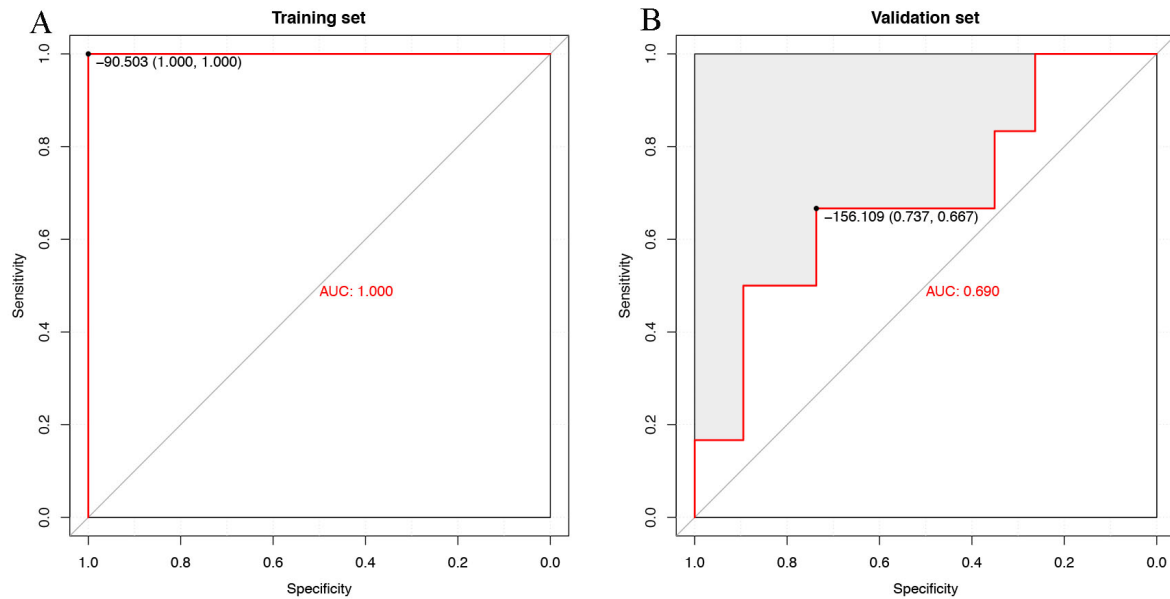


Fig. 2. Evaluation of the breast cancer (BC) risk score. (A) Receiver operating characteristic (ROC) curves evaluating the overall performance of the predictive model for distinguishing BC patients from healthy controls in the training set. (B) ROC curves evaluating the overall performance of the predictive model for distinguishing BC patients from healthy controls in the validation set. AUC, area under the receiver operating characteristic curve.

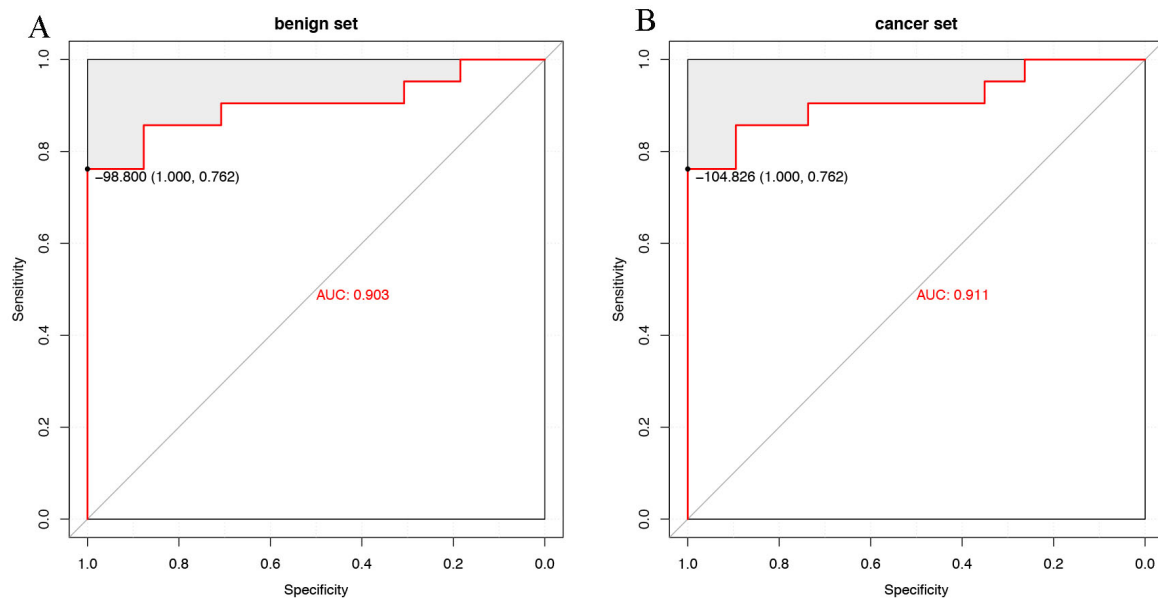


Fig. 3. Evaluation of the breast cancer risk score. (A) ROC curves evaluating the overall performance of the predictive model in distinguishing benign patients from healthy controls. (B) ROC curves evaluating the overall performance of the predictive model in distinguishing BC patients from healthy controls. AUC, area under the receiver operating characteristic curve; ROC, receiver operating characteristic.

2.4 Sampling Training

To maximize training input, we used a combination of unsupervised and supervised learning.

2.4.1 Unsupervised Learning

Between the observations (samples), some informative factors of the vectors were expected to indicate whether a sample originated from a healthy person or a BC patient. PCA is a technique that orthogonally transforms observed variances to convert the original informative variables into new linearly uncorrected parameters called principal com-

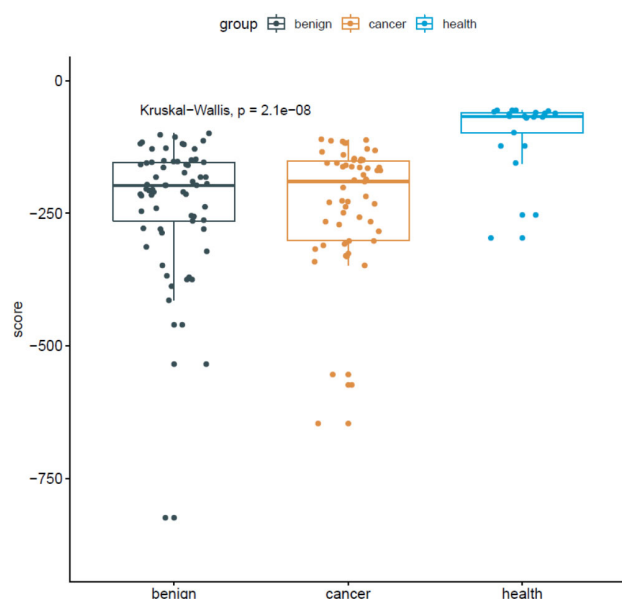


Fig. 4. Distribution of breast cancer risk scores for benign, cancer, and healthy samples.

ponents (PCs). We trained 22 healthy individuals and 40 patients with BC using PCA and obtained 61 PCs. These PCs were ranked by the variance they could explain, and the five most important PCs were selected as they accounted for a large portion of the non-random variance, which we considered correlated with information on whether samples were incorporated with cancer. The five most important PCA rotations were used to calculate the five PCs of the 61 samples in the training set for supervised learning.

2.4.2 Supervised Learning

The general linearized model (GLM) algorithm was used to fit GLM specified by providing a symbolic description of the linear predictor and a description of the error distribution. We used the GLM algorithm to model the correlation between the five most important PCs and the training set and obtained a BC risk score to indicate whether the sample came from a BC patient. We used the following formula: $y = 0.6063 \times \text{PC1} - 0.4668 \times \text{PC2} - 0.7961 \times \text{PC3} + 0.5623 \times \text{PC4} + 0.3139 \times \text{PC5}$.

3. Results

3.1 Sample Characteristics

The clinical characteristics of the BC and control participants are summarized in Table 1. All patients with BC were women with a mean age of 47.9 (20–80) years. Of the 122 patients, 34 (27.9%) were aged <40 years, 63 (51.6%) were aged 40–60 years, and 25 (20.5%) were aged ≥60 years. Patients with benign disease and that staged I, II, III, and NA (not available) accounted for 53.2% (65/122), 22.1% (27/122), 15.6% (19/122), 4.1% (5/122), and 4.9% (6/122), respectively. We also collected 22 healthy samples as controls.

3.2 Optimizing Scores for Predictions of Test and Validation Datasets

We obtained 22 control and 122 BC samples. We randomly selected 70% of the samples (15 controls, 85 BC samples) as the training set and the remaining samples (7 controls, 37 BC) as the validation set. Our model worked well in the training set, reaching an area under the receiver operating characteristic curve (AUC) of 1.0 in the training set and 0.690 in the validation set (Fig. 2A,B).

3.3 Optimizing Scores for Predictions among Healthy, Benign, and Cancer Samples

As our model aimed to identify cancer patients at an early stage, we attempted to differentiate between benign patients and controls to determine the performance. The model obtained an AUC of 0.903, similar to that differentiating between cancer patients and controls (AUC, 0.911; Fig. 3A,B). The median risk scores were −197.98 (Q25–75, −264.35 to −153.81), −190.239 (Q25–75, −301.95 to −151.36), −67.69 (Q25–75, −98.51 to −60.22) for the benign, cancer, and healthy sets, respectively (Fig. 4).

3.4 BC Risk Score Effects by Clinical Characteristics

Associations among BC risk score, stage, and pathological type were studied. The median risk scores were −227.90 (Q25–75, −324.38 to −158.28), −185.73 (Q25–75, −269.17 to −160.02), and −164.71 (Q25–75 −238.38 to −147.42) for stages I, II and III, respectively. Moreover, there was no statistically significant difference between BC risk score and clinical stage. The median risk scores were −163.23 (Q25–75, −314.17 to −140.96), −195.95 (Q25–75, −266.19 to −158.40), −232.57 (Q25–75 −284.78 to −176.14), and −185.73 (Q25–75 −310.77 to −150.02) for Luminal A, Luminal B, Her2+, and triple negative breast cancer (TNBC), respectively. There was no correlation between BC risk score and pathological type (Fig. 5).

4. Discussion

BC is the most common cancer affecting women worldwide. In 2017, 255,180 new cases of BC were diagnosed in the USA; in 2018, the number of deaths due to BC reached 627,000 [3,4]. The goal of early screening is to detect BC early while it is still curable to decrease mortality rates. Previous studies demonstrated that early detection with screening mammography significantly reduces BC deaths by 20–40% [16–20], which might be very useful in the diagnosis of BC.

Several studies have reported that certain factors may contribute to the early detection of cancer. Urothelial carcinoma can be detected in a sample as it has more CNV than a control sample [17–19]. Based on this information, Urine Exfoliated Cells Copy Number Aberration Detector (UroCAD) was developed [17–19]. In the current study, a clear correlation was observed in the fragment size distribution determined by the paired ends generated by NGS. In

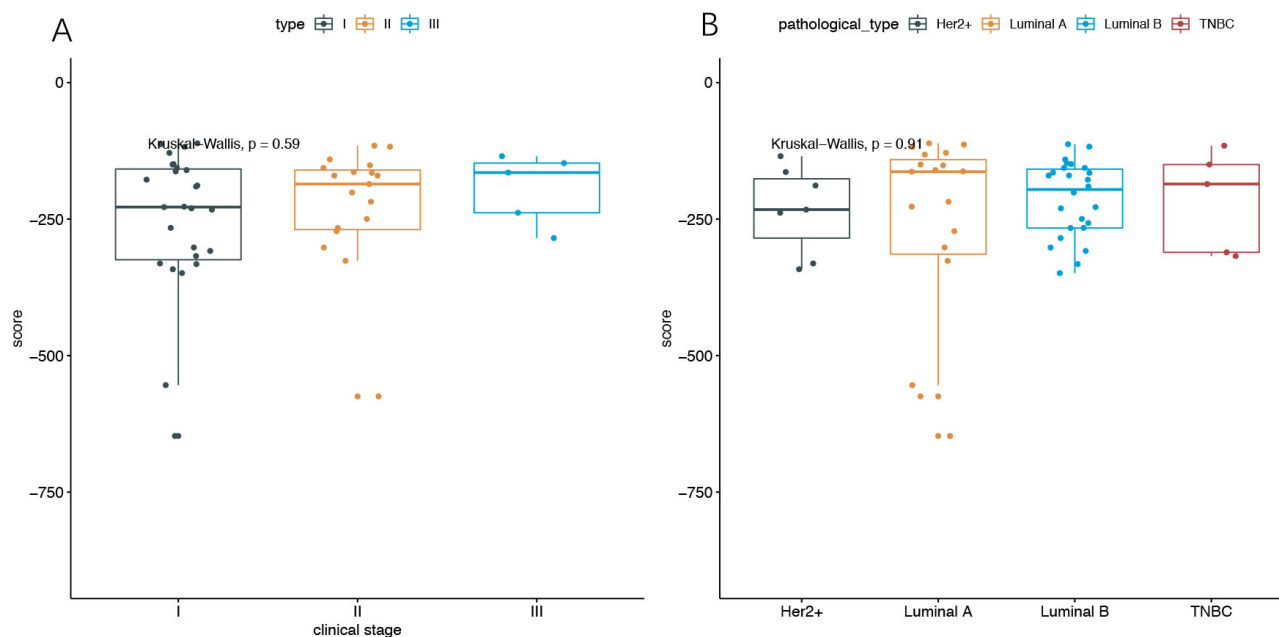


Fig. 5. Distribution of BC risk scores. (A) Distribution of BC scores of stage I, II, III samples. (B) Distribution of BC scores of Luminal A, Luminal B, Her2+, and TNBC samples. BC, breast cancer; TNBC, triple negative breast cancer.

addition, the second peak exhibited the most distinct distribution. Thus, we also considered a second peak (251–450 bp). In addition, patients with BC are likely to have loss of heterozygosity, so it may be helpful to add the mutation signature to our model. To enhance the performance of our model, we included all three factors.

We considered three factors to construct the vector for the samples and then used PCA and the GLM algorithm for the following analysis. Sample quality and selection were assessed. We randomly selected approximately 70% (15 healthy, 85 cancer) of the samples as the training set and the remaining samples as the validation set. Our model achieved an AUC of 1.0 in the training set and an AUC of 0.690 in the validation set. We also tested whether our model could differentiate benign samples from controls and obtained an AUC of 0.903. Consistent with previous studies of non-small cell lung, gastric, ovarian, and pancreatic cancer, this proof-of-concept study of low-coverage plasma WGS provides insight into circulating signatures in plasma DNA and their potential utility for diagnosing cancer [20,21].

Moreover, we also attempted to observe the distribution of BC risk scores among patients with BC of clinical stages I, II, and III. We found no statistically significant differences between BC risk scores and clinical stages. We also found no statistical significance between BC risk score and pathological characteristics, consistent with the results of previous studies [21]. This study has the disadvantages of including a limited sample size and a large number of raw features before the PCA analysis, which may have resulted in model overfitting. Future studies using deep sequenc-

ing in plasma and the features from matched samples are needed to fully characterize the circulating signatures.

5. Conclusions

In conclusion, here we introduced the BC risk score, which incorporates the effects of CNV, mutation signature, and fragment size, and analysed its performance at screening for BC in a Chinese population. Our study showed that the BC risk score could differentiate among BC patients and controls using a non-invasive, cost-effective, and low-coverage WGS technology, which may provide new ideas for future BC screening.

Availability of Data and Materials

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

Author Contributions

YDZ designed, performed, and interpreted the experiments and wrote the manuscript, LP, RY and SHG are responsible for sample collection and experiments, YQ, LQ, JBJ headed all computational and analytical aspects, and all authors participated in proofreading. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript. All authors have participated sufficiently in the work to take public responsibility for appropriate portions of the content and agreed to be accountable for all aspects of the work in ensuring that questions related to its accuracy or integrity.

Ethics Approval and Consent to Participate

Approved by the Ethics Review Committee of Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, ethics review number: HS-1949. Written informed consent was obtained from all patients.

Acknowledgment

Thanks to all the peer reviewers for their opinions and suggestions. We thank the study subjects for their participation for their assistance in conducting the studies. We would like to thank Editage (<https://www.editage.cn>) for English language editing.

Funding

This project was Supported by the Non-profit Central Research Institute Fund of Chinese Academy of Medical Sciences (Grant No. 2019XK320019).

Conflict of Interest

The authors declare no conflict of interest.

References

- [1] Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. *CA: A Cancer Journal for Clinicians*. 2015; 65: 87–108.
- [2] Chen W, Zheng R, Baade PD, Zhang S, Zeng H, Bray F, *et al.* Cancer statistics in China, 2015. *CA: A Cancer Journal for Clinicians*. 2016; 66: 115–132.
- [3] DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, *et al.* Breast cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*. 2019; 69: 438–451.
- [4] Tao Z, Shi A, Lu C, Song T, Zhang Z, Zhao J. Breast Cancer: Epidemiology and Etiology. *Cell Biochemistry and Biophysics*. 2015; 72: 333–338.
- [5] DeSantis CE, Ma J, Jemal A. Trends in stage at diagnosis for young breast cancer patients in the United States. *Breast Cancer Research and Treatment*. 2019; 173: 743–747.
- [6] Zografos GC, Panou M, Panou N. Common risk factors of breast and ovarian cancer: recent view. *International Journal of Gynecological Cancer: Official Journal of the International Gynecological Cancer Society*. 2004; 14: 721–740.
- [7] Bonneau C, Gurard-Levin ZA, Andre F, Pusztai L, Rouzier R. Predictive and Prognostic Value of the TauProtein in Breast Cancer. *Anticancer Research*. 2015; 35: 5179–5184.
- [8] Ahlquist DA. Universal cancer screening: revolutionary, rational, and realizable. *NPJ Precision Oncology*. 2018; 2: 23.
- [9] Siegel RJ. Surgical pathology of lymph nodes in cancer staging: routine and specialized techniques. *Surgical Oncology Clinics of North America*. 1996; 5: 25–31.
- [10] Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. *British Journal of Cancer*. 2013; 108: 2205–2240.
- [11] Miller KD, Nogueira L, Mariotto AB, Rowland JH, Yabroff KR, Alfano CM, *et al.* Cancer treatment and survivorship statistics, 2019. *CA: A Cancer Journal for Clinicians*. 2019; 69: 363–385.
- [12] Cai Y, Li J, Gao Y, Yang K, He J, Li N, *et al.* A systematic review of recommendations on screening strategies for breast cancer due to hereditary predisposition: Who, When, and How? *Cancer Medicine*. 2021; 10: 3437–3448.
- [13] Lehman CD, Arao RF, Sprague BL, Lee JM, Buist DSM, Kerlikowske K, *et al.* National Performance Benchmarks for Modern Screening Digital Mammography: Update from the Breast Cancer Surveillance Consortium. *Radiology*. 2017; 283: 49–58.
- [14] Seely JM, Alhassan T. Screening for breast cancer in 2018-what should we be doing today? *Current Oncology*. 2018; 25: S115–S124.
- [15] Nelson HD, Pappas M, Cantor A, Griffin J, Daeges M, Humphrey L. Harms of Breast Cancer Screening: Systematic Review to Update the 2009 U.S. Preventive Services Task Force Recommendation. *Annals of Internal Medicine*. 2016; 164: 256–267.
- [16] Oeffinger KC, Fontham ETH, Etzioni R, Herzig A, Michaelson JS, Shih YCT, *et al.* Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. *The Journal of the American Medical Association*. 2015; 314: 1599–1614.
- [17] Liu MC, Oxnard GR, Klein EA, Swanton C, Seiden MV, CCGA Consortium. Sensitive and specific multi-cancer detection and localization using methylation signatures in cell-free DNA. *Annals of Oncology*. 2020; 31: 745–759.
- [18] Luo H, Zhao Q, Wei W, Zheng L, Yi S, Li G, *et al.* Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. *Science Translational Medicine*. 2020; 12: eaax7533.
- [19] Cohen JD, Li L, Wang Y, Thoburn C, Afsari B, Danilova L, *et al.* Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*. 2018; 359: 926–930.
- [20] Wan JCM, Stephens D, Luo L, White JR, Stewart CM, Rousseau B, *et al.* Genome-wide mutational signatures in low-coverage whole genome sequencing of cell-free DNA. *Nature Communications*. 2022; 13: 4953.
- [21] Ye LP, Mao XL, Zhou XB, Wang Y, Xu SW, He SQ, *et al.* Cost-effective low-coverage whole-genome sequencing assay for the risk stratification of gastric cancer. *World Journal of Gastrointestinal Oncology*. 2022; 14: 690–702.