# Detection of structural variations and fusion genes in breast cancer samples using third-generation sequencing

**Taobo Hu** ( thuac@connect.ust.hk )
Peking University People's Hospital https://orcid.org/0000-0001-5124-7167
**Jingjing Li**
Nextomics Biosciences, Wuhan https://orcid.org/0000-0002-0142-5495
**Mengping Long**
Peking University Cancer Hospital
**Jinbo Wu**
Peking University People's Hospital
**Zhen Zhang**
The Chinese University of Hong Kong
**Fei Xie**
Peking University People's Hospital
**Jin Zhao**
Peking University People's Hospital
**Houpu Yang**
Peking University People's Hospital
**Qianqian Song**
Peking University
**Shen Lian**
Hong Kong University of Science and Technology
**Jiandong Shi**
Hong Kong University of Science and Technology
**Xueyu Guo**
GrandOmics Inc
**Daoli Yuan**
GrandOmics Inc
**Dandan Lang**
https://orcid.org/0000-0001-7912-0459
**Guoliang Yu**
GrandOmics Inc
**Baosheng Liang**

Peking University
**Xiao-Hua Zhou**
Peking University https://orcid.org/0000-0001-7935-1222
**Toyotaka Ishibashi**
Hong Kong University of Science and Technology
**Xiaodan Fan**
The Chinese University of Hong Kong
**Weichuan Yu**
Hong Kong University of Science and Technology
**Depeng Wang**
GrandOmics Biosciences
**Yang Wang**
GrandOmics Inc
**I-Feng Peng**
GrandOmics Inc
**Shu Wang**
Peking University Cancer Hospital

---

**Article**

**Keywords:** Long-read sequencing, Breast cancer, Structural variation, Fusion gene, Sequencing panel

**Posted Date:** October 4th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-953712/v1



---

# Abstract

Structural variations (SVs) are common genetic alterations in the human genome that could cause different phenotypes and various diseases including cancer. However, the detection of structural variations using the second-generation sequencing was limited by its short read-length which in turn restrained our understanding of structural variations. In this study, we analyzed structural variations in 28 breast cancer-related genes through long-read genomic and transcriptomic sequencing of tumor, para-tumor and blood samples in 19 breast cancer patients. Our results showed that some somatic SVs were recurring among the selected genes, though the majority of them occurred in the non-exonic region. We found evidence supporting the existence of hotspot regions for SVs, which extended our previous understanding that they exist only for single nucleotide variations. In conclusion, we employed long-read genomic and transcriptomic sequencing in identifying SVs from breast cancer patients and proved that this approach holds great potential in clinical application.


# Introduction

Breast cancer is the most common malignancy in women. Genome instability is one of the important molecular characteristics of breast cancer, whereas structural variation is a direct manifestation of genome instability[1]. Structural variations (SVs) including insertion, deletion, duplication, inversion, and translocation affects nucleotides in a much larger scale over single nucleotide variations (SNVs)[2]. SVs are common variations in the general population as shown by the 1000 genome project[3, 4], where specific variations are known to be responsible for the development of a number of genetic diseases and cancers[5, 6, 7, 8]. Previous studies of structural variation influence on gene structure and expression have greatly deepened our understanding of tumorigenesis[9]. Many oncogenes have been proven to be the products of chromosomal translocations and can be served as therapeutic targets. However, it remains challenging to identify SVs in the cancer genome, due to the limitation of the next-generation sequencing (NGS), i.e. short-read length and sequence preference in PCR, which both hinder NGS from detecting complex SVs. Moreover, algorithms trying to identify SVs from NGS data of short-read length showed a high false-negative rate[10]. Third-generation sequencing (TGS) techniques, including Single-Molecule Real-Time (SMRT) sequencing of Pacific Biosciences (PacBio) and the Nanopore long-read single-molecule sequencing of Oxford Nanopore Technologies (ONT) have shown higher sensitivity and specificity in structural variation detection, and have been applied in tumor research including breast cancer research[11, 12, 13].

Despite the fact that SVs of breast cancer in SKBR-3 cell line and patient-derived organoids have been widely studied[14], more proof is needed to illustrate the relationship of SVs and cancer. Nevertheless, the emerging TGS technologies with long-read capability have demonstrated their strengths in cancer study, which allows us to analyze the haploid genome at unprecedented precision, and could provide valuable insights into precision medicine, as in the case of double in-cis PIK3CA mutations showing high sensitivity for alpesilib[13].

In this study, we aim to accurately detect DNA structural variations of a 28-gene panel in breast cancer tissue, matched by para-tumor tissues and blood samples via both ONT and PacBio TGS platforms. To the best of our knowledge, this study was the first to comprehensively analyze structural variation in breast cancer tissue directly via multiple TGS technologies.

## Results

# Target regions capturing and coverage

Several approaches have been used to examine the genomic and transcriptional signatures in breast cancer patients. We recruited 19 breast cancer patients as well as 7 control cases in this long-read study during 2019-2020 (Figure 1). All experimental designs and procedures abide by the regulations from the Institutional Review Board of Peking University People's Hospital. Multiple subtypes of breast cancer were selected as research subjects in this study, including four invasive subtypes (Luminal A, Luminal B, HER-2 enriched, and Triple Negative Breast Cancer (TNBC) cases previously classified by immunohistochemical staining) and Ductal Carcinoma *in situ* (DCIS) cases (Figure 1). Three sets of samples (blood, para-tumor, and tumor) were obtained from all patients. Long-read DNA and RNA information was obtained for a 28-gene panel using the PacBio platform and the ONT full-length whole transcriptome platform, respectively (refer to methods). In addition, blood samples from 7 healthy control donors were processed with the same procedures (Figure 1).

By the combination of a full-length panel approach and long-read sequencing tools, it was possible to explore not only SNPs but also most SVs within these genes, regardless of their locations at either exons or introns. The panel in our study focuses on two gene types: twenty genes associated with a high risk of breast cancer and also participated in homologous recombination repair (HRR), and eight genes involved in the precision medicine during breast cancer treatment (Supplementary Table 1). Probes were designed to cover the whole genome regions of these genes, which are in total about 5M bases. Our results shown in Figure 2 and Supplementary Table 2 summarized some basic characteristics of this panel plus a long-read approach: sufficient depth of sequencing, long reads (N50 is around 3500 bases), and high target coverage (> 99.5%). There were no significant differences in these basic characteristics among the three types of samples and no obvious disparity between samples from patients and healthy controls (Figure 2 & Supplementary Table 2).

# Analysis of germline SVs in breast cancer patients

In our panel study, germline SVs were detected in the blood sample of 12 patients (12/19, 63%) against the healthy controls (Figure 3). The number of SV carried by a single patient vary from one to six (left subset, Figure 3A). Based on their locations, these SVs could be classified into exons, introns, upstream or downstream regions, untranslated regions (UTRs) at 3' or 5' side, flanking regions of genes within 2 kilo-bases, or multiple-hit sites which means more than one of the previous categories. Only a few SVs were found at exonic regions (6/33, bright blue blocks, upper inset in Figure 3A), which is in agreement with previous studies[15]. From another perspective, SVs could be found in HRD genes like *RAD51B* and *BRIP1*

or treatment-related genes like *ERBB4* and *EGFR4*. The distribution of germline SVs in these genes was sporadic and no obvious high-frequent genes were counted, presumably due to the relatively small number of samples.

It is noteworthy that our long-read plus full-length-gene approach allowed us to detect SVs at locations that were hardly detected by the conventional short-read techniques (Figure 3B). For instance, an about 250-base insertion at 3' UTR of EGFR was detected in patient RM65B, but not in healthy control RMH3. This UTR region is close to the centromere of Chromosome 7 and contains many TA repeats. Meanwhile, among individual reads, the locations of this insertion and its size are slightly different as shown in Figure 3B, which further demonstrates the complexity of this mutation site.

# Potential hotspot of somatic SVs revealed in tumor tissue

The somatic SVs could be identified by annotating the unique SVs in tumor tissues against the ones in either blood samples or para-tumor tissues. When comparing SVs detected from para-tumor samples with that from matching blood samples, it was found that most of them were shared by both control samples (43 to 55 per patient, upper plot in Figure 4), implying that they were likely to have occurred before the tissue differentiation during the development. Meanwhile, the existence of unique SVs in para-tumor samples (0 to 10, middle plot) and in blood samples (2 to 8, bottom plot) were possibly caused by loss of heterozygosity (LOH). It indicated that the para-tumor tissue which was histologically normal had already been genetically altered in terms of SVs. This is consistent with findings from SNV studies[13]. In our study, we used blood samples as a reference for tumor tissues to find cancer-driven SVs. The somatic SVs in tumor tissue affecting the 28 breast cancer-related genes were identified and displayed in Figure 5. Our results showed that each patient carried none or only a few somatic SVs (0 to 3, 13 out of 19 patients had SVs $\geq 1$, upper inset in Figure 5A) in this 28-gene panel study. Meanwhile, somatic SVs were detected in 12 out of 28 genes (12/28, 43%, right inset of Figure 5A). SVs had been classified into exonic and intronic SVs according to their locations. Consistent with previous studies, most of the SVs were identified within the intronic region[16]. Among the 12 genes, *ERBB2* had the highest SV frequency, which was detected in 4 patients and being all intronic SVs, followed by *NF1* and *RAD51B*. Figure 5B summarized the four cases of SVs in *ERBB2*: two insertions and two duplications. Noteworthy, the starting sites for three cases were very close, resulting in a certain degree of overlap among the following sequences (RM73T, RM75T and RM80T). These patients were clinically divided into three different groups (Luminal B, TNBC, and DCIS). As far as we know, this region was AT-rich and had not been reported to cause disease. However, such high-frequency somatic SVs at the same site in *ERBB2* (3 out of 19 independent patients) could imply that abnormalities in this area are associated with breast cancer.

# Full-length transcriptome analysis of tumor and para-tumor

Changes at the transcriptional level could provide supports for genomic mutation and also independent evidence in the changes of carcinogenesis. The cDNA of para-tumor and tumor were sequenced using a Nanopore PromethION platform to get the full-length transcriptome data. Data points below 7 in Read Quality (accuracy lower than 85%) were excluded, and the valid points were scattered based on their

length in Figure 6A. Our results showed that the average read quality was about 10, and the mean and median for read length was 1.3k and 1.9k bp, respectively. Principal component analysis revealed that the para-tumor and tumor tissues could be efficiently distinguished based on their transcriptomic data (red and green dots, respectively, Figure 6B). The density plot of reads per gene per 10,000 reads (RPG10K) showed that tumor tissues had shorter reads per gene than para-tumor tissues (Figure 6C). With all mentioned characteristics taken together, it suggests that long-read sequencing on transcriptome could potentially be a good candidate technique for diagnostic application in the future.

# Gene fusions with both genomic and transcriptomic evidence

The accumulation of fusion genes is one of the patterns commonly found in tumor tissues[17]. However, how the fusion genes contribute to or are formed during cancer progression is barely documented. Due to their long-read sequencing characteristics, PacBio and ONT platforms might bring their advantages into studying fusion genes. In a total of 19 cases, we reported that there were seven fusion genes observed in six patients (Table 2). One case found in RM64 showed that a fusion gene at *RECQL5* in Chromosome 17 contained two other segments from Chromosomes 8 and 7 (Figure 7A), showing a particular case of a three-segment fusion gene. The confidence of this fusion gene is supported by its high coverage (depth > 30X, Figure 7B) of reads which were obtained by the high confidential PacBio HiFi platform. In addition, transcripts of this fusion gene were also obtained from ONT Platform (Figure 7C). It appeared that a certain degree of alternative splicing was processed (indicated by dash lines between Figures 7B and C), resulting in a missing of Chromosome-7 segment as well as shorter length in transcripts.

Table 1
Clinicopathological features of breast cancer patients recruited

| Classification | Patient # | Tumor Size (cm) | Lymph Node Metastasis | Ki-67 |
|---|---|---|---|---|
| DCIS | RM65 | 3*2 | 0 | 10% |
| | RM80 | 0.5 | 0 | 10% |
| HER2 | RM62 | 3*1 | 0 | 30% |
| | RM63 | 3*1.5 | 0 | 15% |
| | RM71 | 2.5*2 | 2 | 30% |
| Luminal A | RM66 | 1.7*1.4 | 0 | 10% |
| | RM70 | 2.4*2 | 3 | 10% |
| | RM74 | 4.9*4*2.4 | 1 | 10% |
| | RM77 | 1.7*1.2 | 0 | 5% |
| Luminal B | RM72 | 2.5*2.5*1.8 | 2 | 40% |
| | RM73 | 3*2.8 | 6 | 50% |
| | RM76 | 9.5*7.5*1.8 | 1 | 20% |
| | RM78 | 0.6 | 0 | 40% |
| | RM79 | 4.3*3.7*2.7 | 2 | 20% |
| TNBC | RM64 | 2.7*2.2 | 0 | 70% |
| | RM67 | 1.6*1.5 | 3 | 70% |
| | RM68 | 2.8*2*1.9 | 1 | 90% |
| | RM69 | 1.5*1*1 | 0 | 70% |
| | RM75 | 2.5*2 | 0 | 20% |

Table 2
Fusion genes detected in tumor samples

| Sample | Chr | Start | SV type | SV ID | Gene | Location |
|---|---|---|---|---|---|---|
| RM64 | chr17 | 75646925 | BND | pbsv.BND.chr17:75646925-chr7:9488789_1 | RECQL5;SMIM6 | intronic |
| RM64 | chr9 | 95119452 | BND | pbsv.BND.chr9:95119452-chr19:36870553_1 | FANCC | intronic |
| RM71 | chr17 | 61860680 | BND | pbsv.BND.chr17:61860680-chr17:72564876_1 | BRIP1 | intronic |
| RM76 | chr17 | 58714578 | BND | pbsv.BND.chr17:58714578-chr17:57731280_1 | RAD51C | intronic |
| RM78 | chr16 | 68793666 | BND | pbsv.BND.chr16:68793666-chr16:72915551_1 | CDH1 | intronic |
| RM79 | chr17 | 31350606 | BND | pbsv.BND.chr17:31350606-chr1:248935729_1 | NF1 | intronic |
| RM80 | chr6 | 151943280 | BND | pbsv.BND.chr6:151943280-chr17:38647479_1 | ESR1 | intronic |

# Discussion

Previous study demonstrates that PacBio long reads could detect over 20,000 SVs in a typical whole human genome[18]. However, whole genome third-generation sequencing is rather expensive which limited its application to the clinic. To address this issue, we applied to the best of our knowledge the first clinical TGS panel using PacBio HiFi platform to breast cancer samples. We conducted a comprehensive analysis on structural variations across 28 breast cancer-related genes through long-read genomic and transcriptomic sequencing of paired breast cancer tissue and blood. Our results suggested that germline and somatic SVs were common in the selected genes among breast cancer patients, though the majority of them occurred in the non-exonic region. We also identified a potential hotspot region for somatic SVs. Taking together, our results demonstrated that SVs are potentially important in the tumorigenesis of breast cancer. Indeed, the International Cancer Genome Consortium (ICGC) previously showed that driver SVs are more prevalent than point mutations in breast adenocarcinomas (6.4 SVs compared with 2.2 point mutations on average)[19].

The traditional NGS platforms have poor mapping to repetitive elements including tandem repeats and interspersed repeats, which has made a substantial fraction of most genomes inaccessible and limited its ability to detect SVs[20]. One of representative type of interspersed repeats is Alu element which accounted for 11% of the human genome sequences on average, it belongs to a class of retroelements termed short interspersed elements (SINEs) and often causes SVs through homologous recombination[21]. An important reason that we developed the 28-gene TGS panel for illuminating the full landscape of SVs in breast

cancer is to overcome the limitations of NGS in detecting SVs around repetitive elements. The repetitive elements are abundant in the 28-gene panel which contains most of the breast cancer-related genes, for instance, the *BRCA1* gene has around 40% of Alu family repetitive elements in its DNA sequences[22, 23].

In this paper, by acquiring paired blood, paratumor and tumor tissue from patients, we delineated germline and somatic mutations which were both reported to be responsible for carcinogenesis. Interestingly, we found a potential somatic SV hotspot in the AT-rich region of *ERBB2* gene. Although this region is not belonging to interspersed repeats which often causes SVs through homologues recombination, there are proofs in previous studies that SV hotspots could exist in regions other than SINE elements and DNA transposons[24]. Hence, our method of fine-scale characterization of genomic structural variations using TGS holds great potential to elucidate the full landscape of SV in breast cancer.

We have also systematically examined the paratumor tissues which was used as control samples to identify somatic mutations in tumor. During the process of carcinogenesis, somatic mutations continuously accumulated within the tumor tissue, turning the genomic structure different from surrounding paratumor tissues[25]. It is important to figure out how different is paratumor compared to the blood and to the tumor. We have shown that most SVs were the same in both blood and paratumor tissues, but different from those in the breast cancer tissues. This is in accordance with previous study that demonstrated copy number variations mostly occurred between paratumor and tumor[26].

Our 28-gene TGS panel also showed great promise in identify casual SVs of breast cancer. *NF1* is one of the 12 breast cancer predisposition genes identified to date, however, virtually all previous studies have focused on evaluating breast cancer risk associated with putative pathogenic SNVs and small InDels[27, 28]. We have successfully identified two exonic SVs in two breast tumor tissues, which proves that our TGS panel is useful for detecting cancer-related SVs. Moreover, our TGS panel is robust in identifying SVs, as indicated by the concordant results between long-read genomic and transcriptomic sequencing in identifying fusion genes.

Our findings that somatic SVs are abundant in the cancer genome suggest that they may play an important role in the process of tumorigenesis and development. This is especially important for breast cancer, since the pan-cancer studies by ICGC found that the driver SVs is most evidently prevalent in breast cancer compared to driver point mutations[19]. Taking together, our clinical TGS panel shown here is an accurate and robust method to detect SVs in breast cancer, which is both important for breast cancer research and holds great potential for further clinical application.

# Methods

# DNA extraction

Genomic DNA was extracted from the frozen tissue/blood specimens using the standard phenol/chloroform extraction protocol. Briefly, the tissue specimens were fully ground with liquid nitrogen. For blood, 1 ml whole-blood samples were added with equal amount of ice-cold cell lysis buffer (1.28 M sucrose, 40 mM Tris hydrochloride, 20 mM $MgCl_2$, 4% Triton X-100 [pH 7.5]) and 3 volumes of ice-cold distilled water. This mixture was incubated for 10 minutes on ice, and nuclear pellet was collected by centrifugation (6,000 rpm, 5 min, 4°C). The nuclei both from tissue and blood samples were suspended in extraction buffer (1 M sodium chloride, 100 mM Tris, and 50 mM EDTA, buffered at pH 8.0) containing 2% sodium dodecyl sulphate (SDS) and proteinase K (2 mg/ml final concentration). The suspended nuclei were incubated at 56°C for 2 hours, extracted once with phenol-chloroform-isoamyl alcohol (28:24:1 by volume), one more time with chloroform-isoamyl alcohol (24:1 by volume), and precipitated with 0.7 volume of isopropyl alcohol at -20°C for 40 minutes. The DNA precipitates were washed in ice-cold 80% ethanol twice, collected by centrifugation (12,000 rpm, 15 min, 4°C), dried under vacuum, and finally resuspended in 100 ul of EB (10 mM Tris hydrochloride [pH 8.0]) (#19086, Qiagen). The quantity and quality of DNA samples were measured by NanoDrop One (ND-ONE-W, Thermo Fisher Scientific Inc.) and on 1% agarose gel electrophoresis.

# Target regions capturing and sequencing

DNA probes of 120 bases were designed to cover full-length genes of interest as a custom-made DNA-Cap Panel, and were synthesized by Boke Biotechnologies (Wuxi, Jiangsu, China). During the design of the probes, the Repeat Masker dataset was used to remove probes corresponding to repetitive sequences in the human genome. Capture and enrichment of regions of interest was performed following the manufacturer's protocol. Briefly, 3 ug genomic DNA was sheared to around 5-6 kb fragments by a g-TUBE (#520079, Covaris, Woburn, MA, USA) centrifugation (15,000 g, 2 min, twice). End-repair and dA-tailing of DNA fragments according to protocol recommendations were performed using the Ultra II End Prep module (#E7546, NEB) through pre-capture amplification. Targeted sequence capture was conducted by pooling indexed PCR products and hybridized with custom-made probes. Captured DNA fragments were amplified by PCR again using universal primer. After purification, the prepared target DNA was sequenced using the Pacific Biosciences (PacBio, Menlo Park, CA, USA) SMRT sequencing technology according to protocol recommendations. The PacBio single-molecule real-time (SMRT) Bell™ sequencing library was constructed using a SMRTbell Express Template Prep Kit 2.0 (#100-938-900, PacBio), and finally, sequencing was performed on the PacBio Sequel II platform according to the manufacturer's instructions.

# Data quality control and detection/annotation of SVs

Raw sequencing data (also called raw polymerase reads) were first tested in a standard quality control protocol by using the SMRTlink 8.0 (PacBio) in order to remove low quality reads and adapters resulting in subreads. The minimum polymerase reads accuracy was 0.75. The read quality (RQ) was marked as 0.8 if passed the quality control or as 0 if failed in the filtering. Subreads were obtained by the above filtering. Circular Consensus Sequence (CCS) was used to get CCS reads, and Lima was used for barcode splitting. PBMarkDUP (PacBio) was used to removed potential copies in CCS reads, and PBMM2 (PacBio) was used to compare CCS reads to the reference genome hg38. PBSV (V9.0,

https://www.pacb.com/support/software-downloads/) was used to detect SVs, and DeepVariant[29] (V1.0.0, https://github.com/google/deepvariant) was used to detect SNP and InDel. Detected mutations were be annotated by Annovar[30] (http://nar.oxfordjournals.org/content/38/16/e164) if the following criteria have been met. For SVs, (1) number of supported reads with mutations ≥ 2, (2) mutation frequency among tumor samples ≥ 0.1, (3) mutation frequency = 0 in reference, and (4) screening mutations at interested regions. For SNP and InDel, (1) number of reads covering mutation sites ≥ 5, (2) number of reads with mutations ≥ 2, (3) mutation frequency among samples ≥ 0.05; (4) number of reads covering mutation sites ≥ in reference control ≥ 0; (5) the frequency ration between reference and tumor samples < 0.143, and (6) screening mutations at interested regions.

# RNA sample preparation, cDNA library construction and sequencing

Total RNA from each tissue sample was extracted using the RNeasy Plus Mini Kit (Qiagen, Germany). The RNA purity was checked using the NanoDrop™ One (Thermo Fisher Scientific, USA). RNA degradation and contamination were monitored using 1% agarose gels. The RNA concentration was measured using the Qubit® RNA Assay Kit in the Qubit® 3.0 Fluorometer (Life Technologies, CA, USA). The RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). The RNA quality criteria for the RNA samples was RIN >8.0 (RNA Integrity Number) and 2.0 < OD 260/280 < 2.2. Qualified RNAs were used for Nanopore library preparation. First, reverse transcription of qualified RNA, PCR amplification and adapter ligation were performed using the library preparation kit SQK- PCS109 (Oxford Nanopore Technologies) following the recommended protocol. Then prepared libraries were sequenced on a Nanopore PromethION platform using flowcell R9.4.1.

# Preprocessing of sequencing reads and genome mapping

For the raw sequencing reads, reads of which quality score is lower than 7 or length is shorter than 200 bp were discarded using quality control tool Nanofilt[31] (https://github.com/wdecoster/nanofilt). Then full-length reads were identified and oriented from sequencing reads by the pychopper tool(https://github.com/nanoporetech/pychopper) with default parameters. Then full-length reads were aligned to hg38 reference genome using minimap2[32] (-ax splice -uf –junc-bed). Genome mapping results of full length reads were visualized using the Integrative Genome Viewer[33].

# Prediction of coding sequences and fusion transcript identification

Prediction of coding sequences and protein sequence was performed in all novel isoforms using the ANGEL software[34] (https://github.com/PacificBiosciences/ANGEL). Fusion transcripts were identified using fusion_finder.py from software cDNA_Cupcake (https://github.com/Magdoll/cDNA_Cupcake). Specifically, an identified fusion transcript must meet the following criteria: (1) fusion transcripts map to two or more loci in the genome; (2) each mapped locus must align with at least 95% identity and at least

5% coverage; (3) total aligned coverage of the fusion transcript must be above 99%; (4) each mapped locus must be at least 10kb apart.

# Declarations


### Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant No. 2021YFE0203200), the National Natural Science Foundation of China (Grant No. 82002979), the Beijing Municipal Natural Science Foundation (Grant No. 7202212), the Research and Development Funds of Peking University People's Hospital (Grant No. RDY2020-16) and the Young Investigator Program of Peking University Health Science Center (Grant No. BMU2020PYB022, BMU2021PYB013).


### Author Contributions

TH, YW, IFP and SW conceived and initiated the study; TH, ML, JW, FX, JZ, HY and SW organized and collected the clinical samples and data; TH, JL, ZZ, CG, DY, QS, SL, JS, BL, XZ, DL, GY, TI and SW analyzed the data; and TH, XF, WY, DW, YW, IFP and SW wrote the paper.

### Competing Financial Interests

The authors declare no competing financial interests.

# References


1. Duijf PHG, Nanayakkara D, Nones K, Srihari S, Kalimutho M, Khanna KK. Mechanisms of Genomic Instability in Breast Cancer. *Trends Mol Med* **25**, 595–611 (2019).
2. Sudmant PH, *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
3. Iafrate AJ, *et al.* Detection of large-scale variation in the human genome. *Nat Genet* **36**, 949–951 (2004).
4. Sebat J, *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
5. Li Y, *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).
6. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet* **7**, 85–97 (2006).
7. Sharp AJ, Cheng Z, Eichler EE. Structural variation of the human genome. *Annu Rev Genomics Hum Genet* **7**, 407–442 (2006).
8. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**, 363–376 (2011).

9. Hollox EJ, Zuccherato LW, Tucci S. Genome structural variation in human evolution. *Trends Genet*, (2021).
10. Mills RE, *et al.* Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
11. Aganezov S, *et al.* Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing. *Genome Res* **30**, 1258–1273 (2020).
12. Nattestad M, *et al.* Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res* **28**, 1126–1135 (2018).
13. Vasan N, *et al.* Double PIK3CA mutations in cis increase oncogenicity and sensitivity to PI3Kalpha inhibitors. *Science* **366**, 714–723 (2019).
14. Zhuang Y, *et al.* Establishment and characterization of immortalized human breast cancer cell lines from breast cancer patient-derived xenografts (PDX). *NPJ Breast Cancer* **7**, 79 (2021).
15. Sakamoto Y, Sereewattanawoot S, Suzuki A. A new era of long-read sequencing for cancer genomics. *J Hum Genet* **65**, 3–10 (2020).
16. Tuzun E, *et al.* Fine-scale structural variation of the human genome. *Nat Genet* **37**, 727–732 (2005).
17. Matsushige T, *et al.* Detection of Disease-specific Fusion Genes of Soft Tissue Tumors Using Formalin-fixed Paraffin-embedded Tissues; Its Diagnostic Usefulness and Factors Affecting the Detection Rates. *Yonago Acta Med* **62**, 115–123 (2019).
18. Chaisson MJ, *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
19. Consortium ITP-CAoWG. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
20. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. *Nat Rev Genet* **19**, 329–346 (2018).
21. Doronina L, Reising O, Schmitz J. Gene Conversion amongst Alu SINE Elements. *Genes (Basel)* **12**, (2021).
22. Sobczak K, Krzyzosiak WJ. Structural determinants of BRCA1 translational regulation. *J Biol Chem* **277**, 17349–17358 (2002).
23. Ewald IP, Ribeiro PL, Palmero EI, Cossio SL, Giugliani R, Ashton-Prolla P. Genomic rearrangements in BRCA1 and BRCA2: A literature review. *Genet Mol Biol* **32**, 437–446 (2009).
24. Lin YL, Gokcumen O. Fine-Scale Characterization of Genomic Structural Variation in the Human Genome Reveals Adaptive and Biomedically Relevant Hotspots. *Genome Biol Evol* **11**, 1136–1151 (2019).
25. Shoshani O, *et al.* Chromothripsis drives the evolution of gene amplification in cancer. *Nature* **591**, 137–141 (2021).
26. Hu T, *et al.* Forward and reverse mutations in stages of cancer development. *Hum Genomics* **12**, 40 (2018).

27. Hu C, *et al.* A Population-Based Study of Genes Previously Implicated in Breast Cancer. *N Engl J Med* **384**, 440–451 (2021).
28. Chen Z, *et al.* Discovery of structural deletions in breast cancer predisposition genes using whole genome sequencing data from > 2000 women of African-ancestry. *Hum Genet*, (2021).
29. Poplin R, *et al.* A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* **36**, 983–987 (2018).
30. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
31. De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).
32. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
33. Robinson JT, *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24–26 (2011).
34. Shimizu K, Adachi J, Muraoka Y. ANGLE: a sequencing errors resistant program for predicting protein coding regions in unfinished cDNA. *J Bioinform Comput Biol* **4**, 649–664 (2006).

# Figures

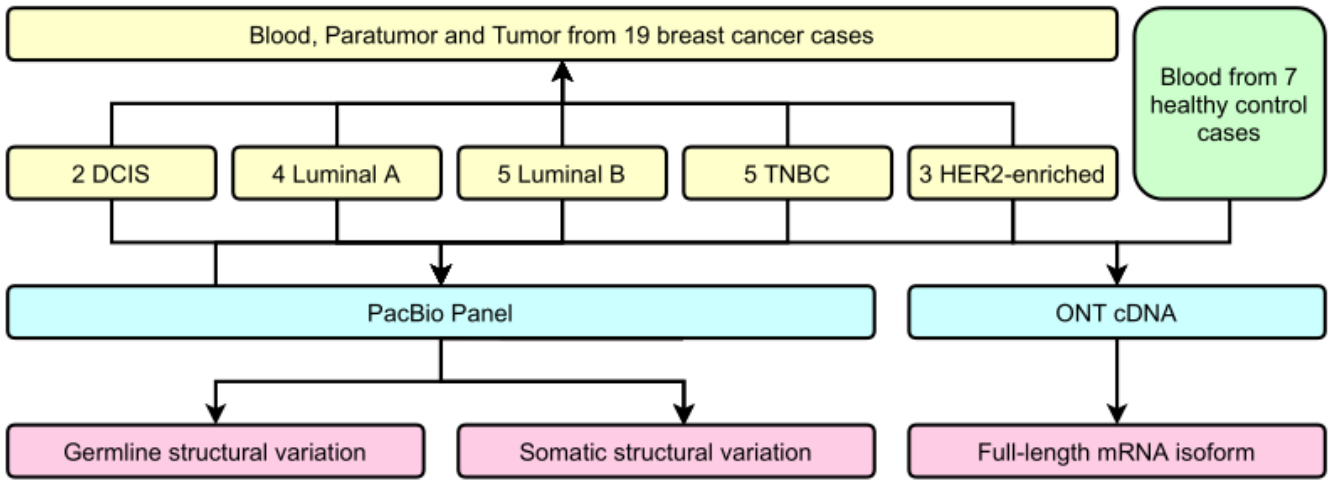
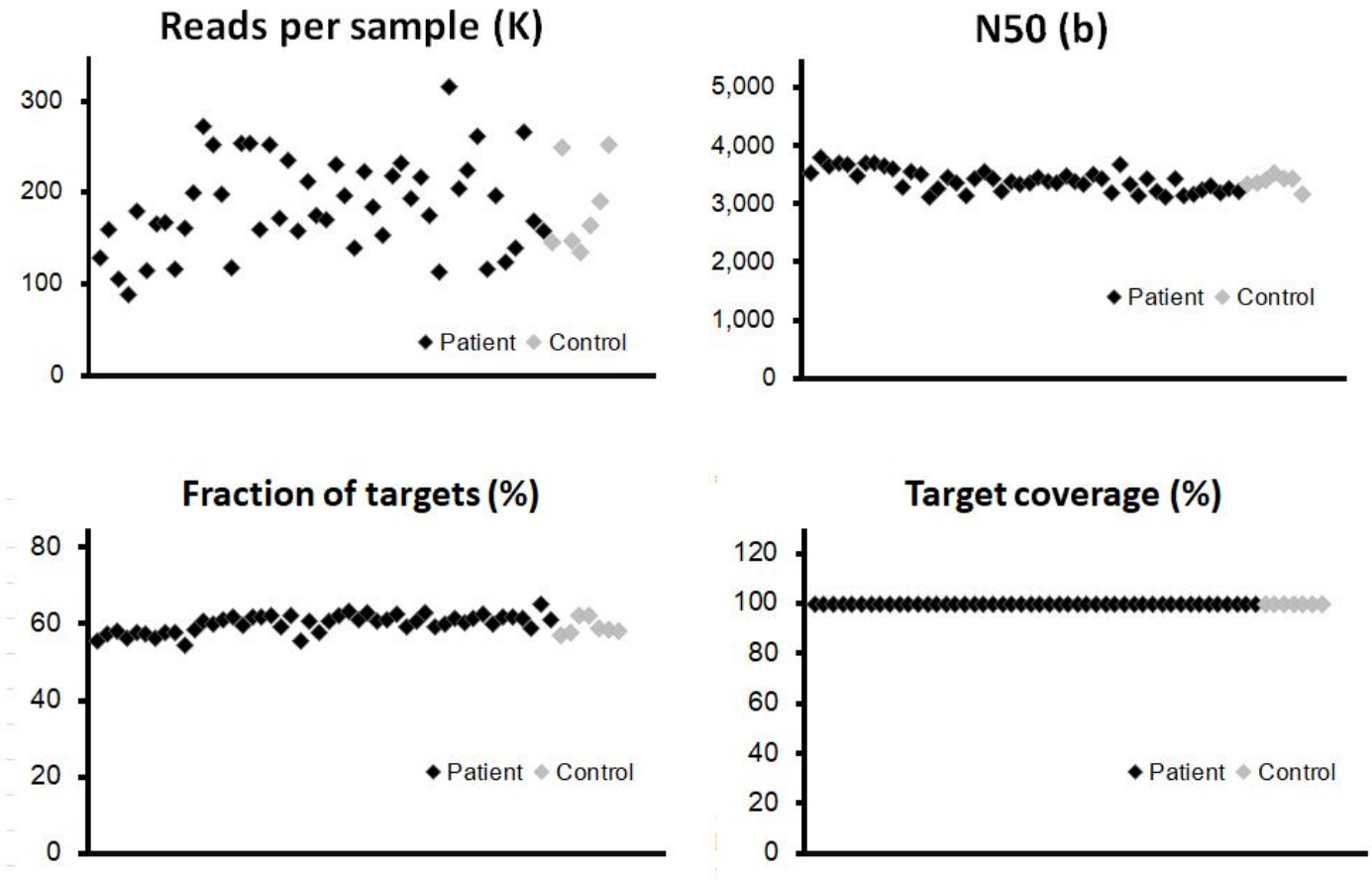


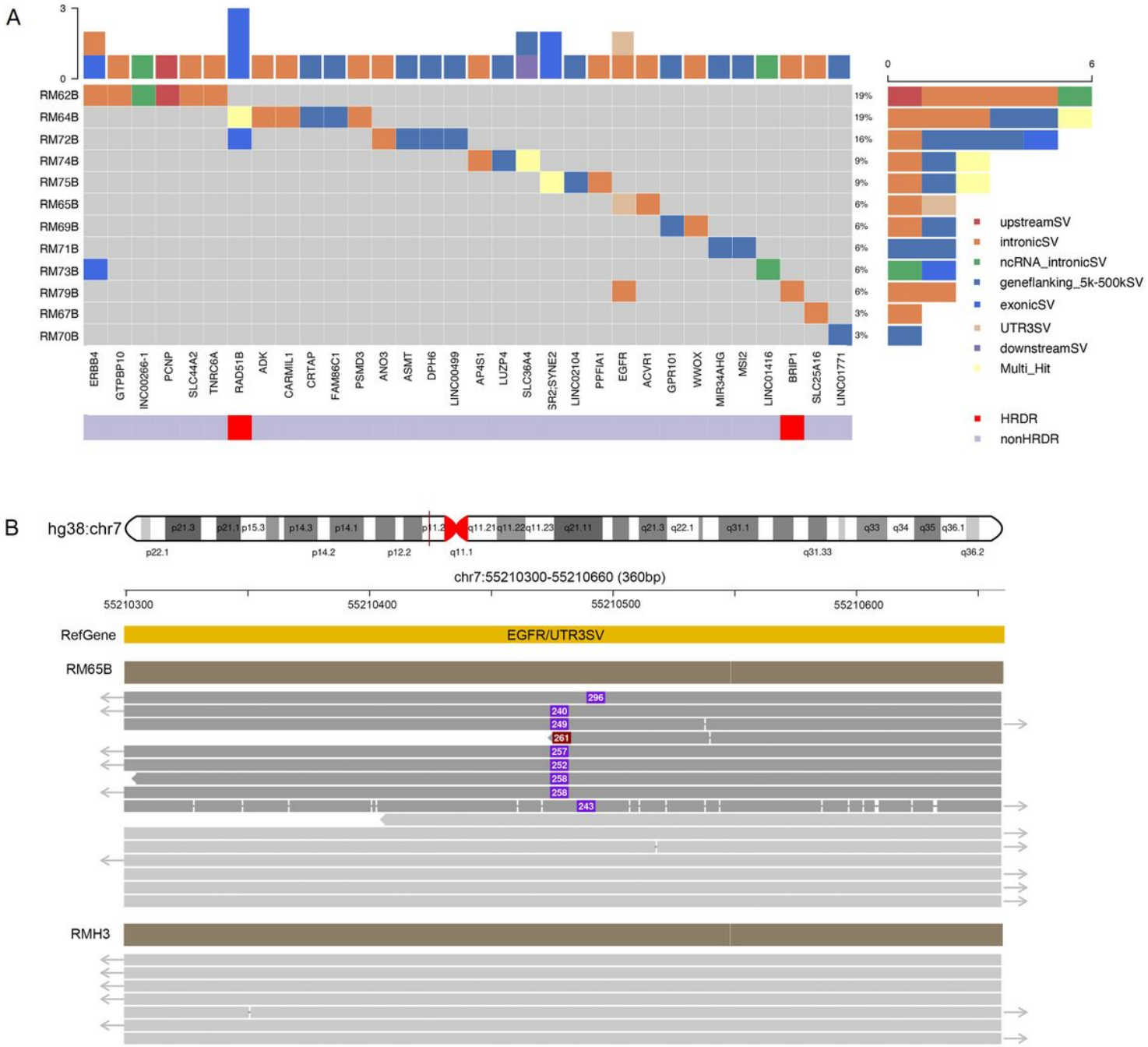

**Figure 1**

Flow chart of study design.




## Figure 2

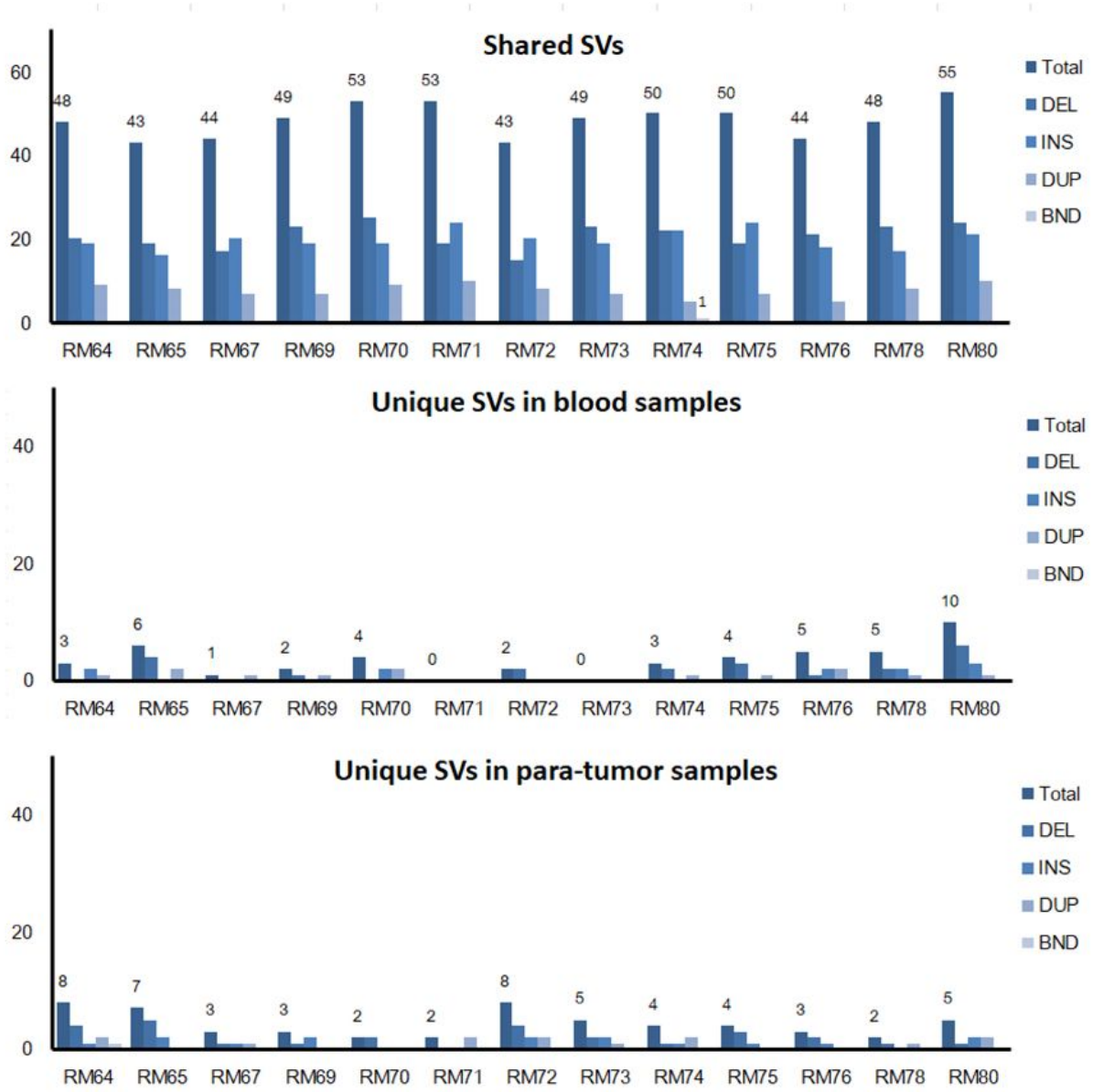
Quality control of long-read sequencing for the panel of 28 genes. The probes were designed to cover the whole genome regions of all panel genes, which are around 5M-base coverage of detection. The vertical axis of each point illustrated the quantitative information from individual blood sample, para-tumor tissue, or tumor tissue. The effective read numbers were around 100 kb to 300 kb per sample, and the N50s were around 3,000 to 4,000. No obvious differences could be detected between patient and control groups (black and gray points, respectively). After the alignment process, the fraction of targets among different samples was around 55% with a slightly fluctuation. The coverage of the target region was above 99% in all tested samples.




## Figure 3

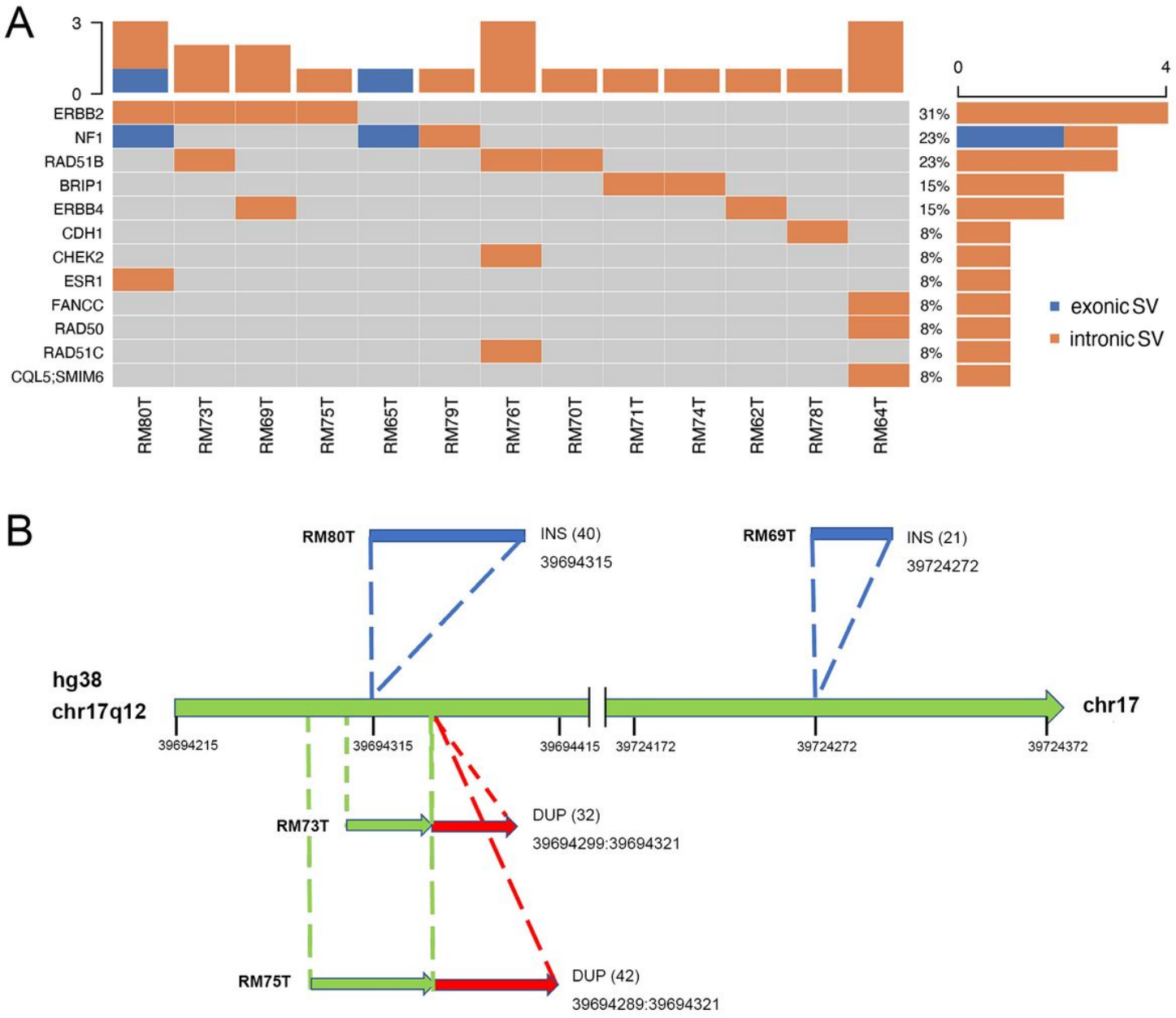
Excessive germline structural variants occurred in non-exonic regions in breast cancer patients. (A) Summary of germline SVs in specific genes and patients. Individual patients' blood samples (starting with "RM" labeling, plus patient number, and end with B for blood samples) were examined against to seven healthy samples as control. The calling of SVs would be classified to one of the following categories: exons, introns, upstream or downstream regions, un-translated regions (UTRs) at 3' or 5' side, flanking regions of genes within 2 kilo-bases, or multiple-hit sites. The scales in the top and right insets illustrated the cumulative numbers of SVs in particular genes and patients, respectively. Most SVs were located at non-exonic regions. (B) SVs identified in EGFR. In RM65B, an insertion (~280 bp) was identified

at the UTR 3’ region of EGFR genes. The dark red solid line in hg38:chr7 pointed to a 360-bp region as expanded below. Representative reads from RM65B and RMH3 (control) were aligned accordingly. The purple boxes and inside numbers showed the locations and sizes of this insertion in individual reads. Such insertion was identified only in a part of reads in RM65B (24/49) but not in RMH3 (0/83).



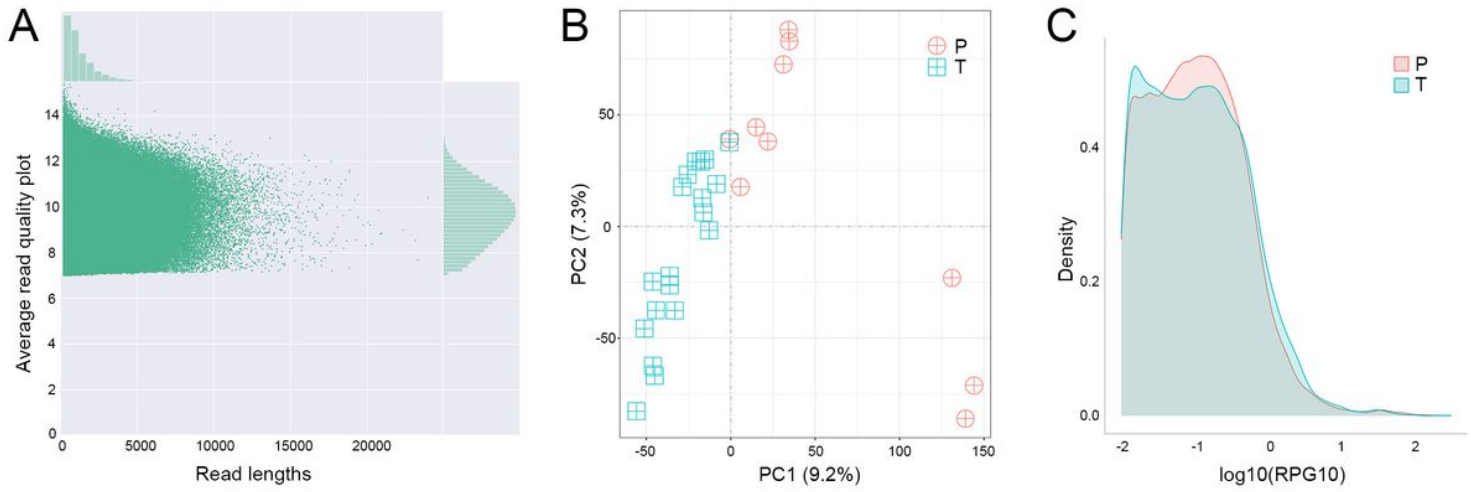

## Figure 4

Shared and unique SVs in para-tumor tissue and blood samples. A comparison of shared and unique SVs between two kinds of samples. Numbers above individual bars showed the number of total SVs. DEL, deletion; INS, insertion; DUP, duplication; and BND, Breakpoint notation. Most SVs were found in both tissues, while a few unique SVs were only observed in one of the tissues.




## Figure 5

Excessive somatic SVs in non-exonic region (A) A summary of somatic SVs in particular genes and patients. Patients' tumor samples were examined against to their blood samples as control. SVs are sorted into exonic and intronic types based on their locations. Similarly, most SVs were found at non-exonic regions. Among thirteen patients, four of them had been identified to carry SVs at ERBB2. (B) Hotspot of SVs in ERBB2. There were two INS SVs (shown in blue) and two DUP SVs (red) found in ERBB2. Numbers inside parentheses indicated the sizes of SVs. A conserved region with SV occurrence was exposed among three independent patients (RM80T, RM73T and RM75T).




## Figure 6

Distinguishable transcriptomes of tumor samples from others in long-read sequence. (A) Quality control of transcriptome analysis. Transcriptome was built-up based on their cDNA library construction followed by ONT nanopore sequencing. Plot showed the quality of individual data points (main figure) as well as their distributions (top and left insets). Data points below 7 in Average Read Quality (i.e., lower than 85% accuracy from ONT manual instruction) were excluded from analysis. (B) Principal component (PC) analyses. Each symbol represented one clinical sample from para-tumor tissues (P) and tumor tissues (T). Clearly the majority of samples were quite distinguishable from ones from the other group. (C) Density of reads per gene10 (RPG10) plot. Differential patterns from different types of samples revealed that tumor tissues had smaller reads per genes than para-tumor tissues.

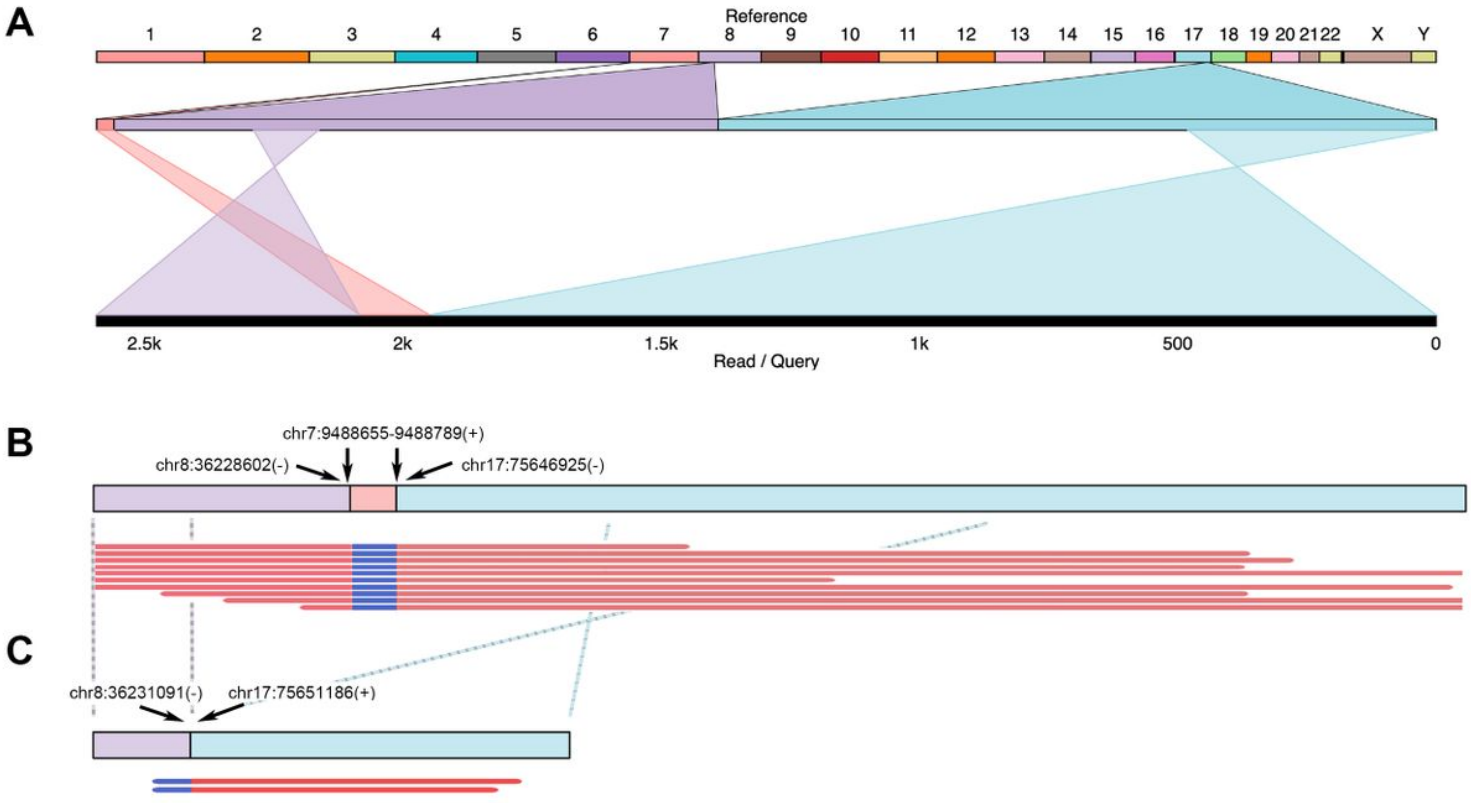


**Figure 7**

Example of fusion genes and transcripts observed in patient tumor tissues. In Patient RM64, several structural variants (including DEL and BND) were identified at FANCC, RAD50 and RECQL5 regions. A three-segment fusion gene observed at RECQL5 was illustrated. (A) Top line: reference genome from hg38. Numbers and letters indicated individual chromosomes. Middle line: expansion of sequenced regions from chromosomes 7, 8 and 17. Bottom line: Illustration of one genomic structure in RM64T samples, containing regions from chromosomes 8, 7 and 17. Crossed projection lines from middle to bottom lines represented reversions occurred during the fusion process. (B) Top line: a plot of this fusion gene at genomic DNA. Breakpoint notation locations were labeled by arrows. Plus and minus symbols showed forward and reverse directions, respectively. Bottom lines: representative data selected from individual reads. (C) Transcripts observed in this area. Top line: a cartoon demonstration of corresponding mRNA. Crossed dash lines showed a reversion observed after a comparison to reference. Bottom lines: two reads continuously from chr8 to chr17 were shown. The chr7 segment in this fusion gene did not have detectable mRNA reads.

# Supplementary Files

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

- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx