ORIGINAL ARTICLE

WILEY Cancer Science

Integrative molecular profiling identifies a novel cluster of estrogen receptor-positive breast cancer in very young women

Charny Park[1] | Kyong-Ah Yoon[2] | Jihyun Kim[1] | In Hae Park[3] | Soo Jin Park[3] | Min Kyeong Kim[4] | Wooyeong Jang[1] | Soo Young Cho[1] | Boyoung Park[5] | Sun-Young Kong[4,5] [iD] | Eun Sook Lee[3,5]

[1]Clinical Genomics Analysis Branch, Research Institute, National Cancer Center, Goyang, Korea

[2]Laboratory of Biochemistry, College of Veterinary Medicine, Konkuk University, Seoul, Korea

[3]Center for Breast Cancer, Hospital, National Cancer Center, Goyang, Korea

[4]Translational Cancer Research Branch, Division of Translational Science, National Cancer Center, Goyang, Korea

[5]Graduate School for Cancer Science and Policy, National Cancer Center, Goyang, Korea

Correspondence
Sun-Young Kong, Department of Cancer Biomedical Science, Graduate School of Cancer Science and Policy, Translational Cancer Research Branch, Research Institute, National Cancer Center, Goyang-si, Korea.
Email: ksy@ncc.re.kr
and
Eun Sook Lee, Center for Breast Cancer, Hospital, Department of Cancer Biomedical Science, Graduate School of Cancer Science and Policy, National Cancer Center, Goyang-si, Korea.
Email: eslee@ncc.re.kr

Very young breast cancer patients are more common in Asian countries than Western countries and are thought to have worse prognosis than older patients. The aim of the current study was to identify molecular characteristics of young patients with estrogen receptor (ER)-positive breast cancer by analyzing mutations and copy number variants (CNV), and by applying expression profiling. The whole exome and transcriptome of 47 Korean young breast cancer (KYBR) patients (age <35) were analyzed. Genomic profiles were constructed using mutations, CNV and differential gene expression from sequencing data. Pathway analyses were also performed using gene sets to identify biological processes. Our data were compared with young ER+ breast cancer patients in The Cancer Genome Atlas (TCGA) dataset. *TP53*, *PIK3CA* and *GATA3* were highly recurrent somatic mutation genes. APOBEC-associated mutation signature was more frequent in KYBR compared with young TCGA patients. Integrative profiling was used to classify our patients into 3 subgroups based on molecular characteristics. Group A showed luminal A-like subtype and IGF1R signal dysregulation. Luminal B patients were classified into groups B and C, which showed chromosomal instability and enrichment for APOBEC3A/B deletions, respectively. Group B was characterized by 11q13 (CCND1) amplification and activation of the ubiquitin-mediated proteolysis pathway. Group C showed 17q12 (ERBB2) amplification and lower ER and progesterone receptor expression. Group C was also distinguished by immune activation and lower epithelial-mesenchyme transition (EMT) degree compared with group B. This study showed that integrative genomic profiling could classify very young patients with breast cancer into molecular subgroups that are potentially linked to different clinical characteristics.

1 | INTRODUCTION

Breast cancer is the most commonly diagnosed cancer in women and shows high mortality rates worldwide. Although the highest incidence rates are found in Western countries, the frequency of breast cancer has been steadily increasing in Asian countries, including Korea, China and Japan.[1-3] A notably different pattern among Asian females compared with their Western counterparts is the age of onset. In contrast to the gradual increase in incidence according to age in Western women, older Asian women do not always demonstrate a higher rate of breast cancer.[4-6] In Korea, the age-specific rate of breast cancer peaks before the age of 50 and levels off thereafter.[4] Although breast cancer in very young women is not common, more women under the age of 35 are diagnosed with breast cancer in Asian countries than in Western countries. Nationwide survival data in Korea showed that the prognosis was worse for younger patients (≤35 years of age) than older patients (35-50 years of age), especially among those in hormone receptor-positive groups.[7] Similarly, poor outcomes, characterized by more advanced clinical stage and shorter survival, have also been reported for Chinese patients under the age of 35.[8] These worse outcomes may be associated with unique biological and genetic characteristics that lead to differences in clinical responses to treatment.

Breast cancer is classified into 4 intrinsic subtypes (luminal A, luminal B, triple-negative and HER type) according to the expression status of the hormone receptors estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2). Notably, these subtypes are closely related to clinical features. It has been reported that the triple-negative subtype is enriched in the younger aged group, whereas the luminal A type is less frequent.[9,10] However, because of the heterogeneity, assignment to subgroups is not sufficient to establish clinical management strategies. To gain a better understanding of the molecular characteristics underlying this heterogeneity, researchers have extensively studied breast cancer using genomic and proteomic profiling approaches. In this context, The Cancer Genome Atlas (TCGA) projects have identified major genetic and epigenetic abnormalities in breast cancer, including somatic mutations, altered gene expression and copy number aberrations.[11,12] Recent advances in proteogenomics have also identified significant signaling pathways as well as somatic mutations.[13]

In this study, we sought to identify unique molecular features by investigating Korean young breast cancer (KYBR) patients, aged 35 and younger, using whole exome sequencing (WES) and RNA-sequencing (RNA-seq) analyses. To limit the heterogeneity of the patient population, and, thus, minimize the complexity of our analysis, we focused on estrogen receptor (ER)-positive breast cancer patients. We profiled somatic mutations, germline variants, copy-number variants (CNV)

and differentially expressed genes (DEG), and compared our results to those in TCGA ER-positive young and old age patients. Finally, we classified ER-positive patients into 3 subgroups (Group A, B and C) according to molecular characteristics, and defined separate subgroups among the luminal B subtype. Our results suggest a more elaborate classification of breast cancer in very young women.

2 | METHODS

2.1 | Study objectives and specimens

This study included 47 patients with histologically confirmed breast cancer, aged 35 years or younger, treated at the National Cancer Center in Korea. All patients underwent surgical resection; patients who received neoadjuvant chemotherapy were excluded. Demographic characteristics, including age and family history of cancer, were also collected. Tumor and adjacent normal samples were obtained from surgically resected specimens, and blood samples were collected from the patients. Genomic DNA and RNA were extracted from tissue specimens and blood samples using an AllPrep DNA/RNA Mini Kit and a QIAamp DNA Blood Mini Kit, according to the manufacturer's protocol (Qiagen, Valencia, CA, USA). The concentration and integrity of RNA were assessed using a NanoDrop spectrophotometer (Thermo Fisher Scientific, Waltham, MA) and an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The participants voluntarily signed an informed consent form that was approved by the Institutional Review Board (IRB No. NCCNCS 13717).

2.2 | Clinical data

We retrospectively reviewed the medical and pathology records of all patients to collect histological diagnoses of surgical specimens, tumor staging, and follow-up data. The expression of hormone receptors, including ER, PR and HER2, was assessed by immunohistochemical staining and evaluation by pathologists according to American Society of Clinical Oncology (ASCO)/College of American Pathologists (CAP) guidelines.[14,15] All patients were followed up with an average interval of 3 months after surgery and median follow up of 127 months. The overall survival after surgery was calculated from the date of surgery until the date of death or last follow up.

2.3 | Whole exome sequencing analysis for somatic mutations, germline variants and copy number variants

Whole exome sequencing data were generated from genomic DNA obtained from the tumor tissues and blood of 47 patients using the

Agilent SureSelect Human All Exon V5 Target Enrichment kit, according to the manufacturer's standard protocol. Genomic DNA was amplified and processed for sequencing using the HiSeq 2500 platform (Illumina, San Diego, CA, USA).

Low-quality reads were trimmed by processing sequencing reads using Trimmomatic v0.36.[16] Sequence reads were aligned to the hg19 reference using BWA v0.7.3 software; sorting, marking of duplicated reads and realigning around indel regions were performed using Picard v1.128 and GATK v3.3.[17,18] To aid in calling somatic mutations, we collected somatic mutations from 2 mutation callers: Mutect v2 and Strelka v1.0.14.[19,20] Off-target mutations were eliminated by reference to SureSelect target regions. Somatic mutations were converted to MAF format and annotated using Oncotator v1.8.[21]

Germline SNV were called using the GATK Haplotype caller, and low-quality SNV were eliminated using GATK VariantFiltration. The pathogenicity of total germline SNV was confirmed using ClinVar v20161102, and truncation variants of known cancer genes were additionally chosen.[22]

In order to refer the origin of the somatic mutational process, we clustered mutation signatures using non-negative matrix factorization (NMF) from somatic SNV using the R package, SomaticSignatures.[23] We ran NMF for 9 signature clusters from the mutation count matrix and selected 2 primary signatures based on maximum cophenetic-correlation coefficients. Identical signature numbers and associated mutational process were referred from COSMIC signature.

We also analyzed CNV using EXCAVATOR2 v1.1.2 and called CNV peak regions by using GISTIC v2.0.22 with a $P = 0.95$.[24,25] Germline CNV analyses were performed using 2 R packages CODEX and cn.mops, which support normal-pooled sample analysis.[26] Finally, a reliable overlap of over 80% between the results of 2 methods was confirmed for germline CNV deletion regions.

To identify regions of chromosomal instability (CIN), we counted copy number alterations in arm-level CNV, calculated from GISTIC analyses, and confirmed differences among the 3 groups (Figure S1). This measure was validated by performing a signature-based CIN analysis (CIN70 score) as previously reported.[27]

2.4 | APOBEC3A/B germline deletion

Read counts of deleted regions were calculated from BAM files and identified as belonging to 3 APOBEC3A/B classes (homozygous deletion, heterozygous deletion or wild-type) using k-means clustering. Germline CNV in the TCGA breast cancer database were also investigated by filtering using that same strategy as used for our dataset. Therefore, we filtered out deletion regions that were too long (>30 kb). Next, we confirmed the remaining regions using the TCGA SNP 6.0 level 3 dataset. Germline deletions in APOBEC3A/B were previously reported to be in the chr22:39,363,619-39,375,307 (hg19) region, based on a 24-probe Affymetrix SNP6.0 array.[28] We also focused on 711 cancer-related genes that were curated in COSMIC, including APOBEC3A/B. Germline exonic deletion was confirmed in TCGA level 3 exon-level expression data and tested using Student's t test (P-value < 0.05).

APOBEC3A/B deletion was validated by copy number analysis and genotyping. APOBEC3B and RNase P, an endogenous control gene, were amplified by quantitative RT-PCR using fluorescent probes. The copy-number status of APOBEC3B was calculated from differences in threshold cycles (Ct values) between APOBEC3B and RNaseP.[29] APOBEC3A/B deletion was also confirmed by genotyping SNP rs12628403, which is known to be in strong linkage disequilibrium (LD) with the APOBEC3A/B deletion allele.[30] The strong LD was confirmed by comparing the results of copy number status and genotypes of rs12628403.

2.5 | Gene expression profile analysis using RNA-seq data

RNA-seq data were generated from tumor RNA from 47 patients, prepared using the TruSeq stranded Total RNA LT Kit (Illumina). Double-stranded cDNA libraries were prepared, obtaining strand specificity, and after indexing adapters ligation, were sequenced using an Illumina sequencing platform.

The RNA-Seq analysis workflow for quantification of gene expression follows the TCGA GDC pipeline. Low-quality RNA reads were trimmed with Trimmomatic v0.36, and sequences were aligned to the reference hg19 using Mapsplice v2.0.1.9.[31] The aligned reads were filtered to remove indels, large inserts and zero mapping quality reads. Finally, gene expression was quantified using RSEM 1.1.13, referring to known UCSC gene models.[32]

2.6 | Molecular subtype classification

Molecular subtypes were classified using the NMF clustering method generally used in previous TCGA studies. More refined gene sets associated with each subtype were obtained by applying a gene-interaction network-based submodule analysis approach using BioNet.[33] Network submodules were identified based on a false-discovery rate (FDR) < .025, ultimately yielding a DEG gene set comprising 1463 genes. The biological functions of submodules were analyzed and visualized using the Cytoscape Reactome FI plugin.[34]

Specific characteristics of immune system and epidermal-mesenchyme transition (EMT) status were also examined. A stromal and immune cell admixture was inferred using the ESTIMATE method.[35] EMT status was inferred from principal component analysis (PCA) by reference to a previous method based on the expression of 315 previously identified EMT-related genes; principal component 2 (PC2) clearly divided molecular subtypes into group C and others.[36]

2.7 | Detection and validation of fusion genes

Fusion genes were identified using RNA-Seq fasta files. Three fusion callers (deFuse v0.6.2, PRADA v1.2 and STAR-fusion v1.0.0) were used, with reference to hg19 and ENSEMBL release 69.[37-39] False-positive candidates were eliminated based on the following criteria, described in the TCGA Pan-Cancer Fusion Database: (i) gene homology (e-value > .01); (ii) multiple different breakpoints (n > 2); and (iii)

sample recurrence (n > 2).[40] Ultimately accepted fusions were those identified by at least 2 caller programs and FusionInspector; final results were annotated using Pegasus.[41] In general, fusion genes have low recurrence, but different fusion genes comprise common molecular pathways, like BRAF fusions spanning different partner genes. Recurrent molecular pathways of fusion genes were investigated by gene set enrichment analysis (GSEA) using the R package "GSVA." To validate candidate fusion genes before GSEA, we additionally explored the consistency of fusion gene and segmentation regions' break point and low correlated genes were eliminated.[42]

Estrogen receptor 1 (ESR1) fusion variants detected by RNA-seq data analysis were confirmed by RT-PCR analysis using primers designed to amplify the coding sequences of the ESR1 fusion junction. ESR1-ARMT1 (acidic residue methyltransferase 1) fusion was examined using specific primers for ESR1 and ARMT1 (forward, 5′-CAG ATG GTC AGT GCC TTG TT-3′; reverse, 5′-AGA AAG GAG AGA GAT AGC TT-3′).

2.8 | Comparative analysis using The Cancer Genome Atlas breast cancer project database

The Cancer Genome Atlas data consisting of 796 patients with ER-positive breast cancer were downloaded from the GDC database (https://portal.gdc.cancer.gov) for comparison with KYBR data. Because of differences in age-prevalence and sample size between Korean and TCGA patients, young and old patients in the TCGA database were defined as those aged 40 years or younger and 75 years or older, respectively, based on a previous study.[43] TCGA level-3 results were compared with KYBR data with respect to mutation burden, somatic mutations and somatic CNV. The germline CNV analysis method precisely followed our WES-based normal-pooling method; segmentation analysis results were used for additional validation. NMF clustering for molecular subtype identification was performed by expression profiling using exactly the same method and gene sets with young breast cancer (YBR). Immune, stromal and CIN70 scores were also calculated using the exact same methods and gene signatures with the YBR dataset.

3 | RESULTS

3.1 | Young patients with estrogen receptor-positive breast cancer and molecular subtypes

A total of 47 ER-positive KYBR patients (≤35 years of age) were analyzed. Intrinsic molecular subtypes of patients were identified based on hormone receptor status and differential expression of 50 genes (PAM50 classifier). Demographic features of KYBR patients were compared with those of TCGA subjects, as described in Table 1. According to the clinical profile of TCGA ER-positive patients (n = 794), Asians are predisposed to ER-positive breast cancer at a younger age. The average age of Asian patients was 50.5, while patients of other races were significantly older (African American 57.5 and White 59.2; t test, P = 3.84e-05). There were no ER-positive Asians over 75, in contrast to African Americans (8.4%) and White people (13.1%).

Molecular subtypes were defined by performing an integrative investigation. Somatic and germline variants, including point mutations and CNV, were identified, and KYBR patients were classified into 3 subgroups, A (23%), B (41%) and C (36%), derived from an NMF clustering using gene expression profiling (Figure 1A). An estimation of chromosomal instability based on arm-level segmentation count and CIN70 score[27] revealed that group A clearly belonged to the chromosomal-stable type, whereas group C showed high chromosomal instability (Figure S1). An investigation of PAM50 status showed that group A was enriched in luminal A type (91%), group B was a mixture of luminal A and B (89%), and group C included HER2-enriched and luminal B types (64%) (Figure 1B). Based on clinical profile, histological grades and lymphatic invasion gradually increased from group A to C. Specific variants and associated pathways for each subtype are described below. Individual clinical information and important genomic results for patients are presented in Table S1.

When classifying TCGA ER-positive patients into our 3 molecular subtypes, scoring status for immune-infiltration, stromal cells and chromosomal instability were significantly similar for YBR (Figure S2). However, prevalence ratios of subtypes were different for age groups. Group A (13.7% in young age) increased in patients over 40 years old (26.0%) and, inversely, group B and C decreased in patients over 40 years. As the results of the 5-year survival analysis, young ER-positive patients in TCGA dataset seem to have better prognosis than older patients (P = 0.001; Figure S2D). However the survival rate of group B rapidly decreased after 5 years and the prognosis of groups A and B converged similarly to the worst prognosis group C in 10 years (P = 0.04; Figure S2D).

Notably, expression of genes for the important breast cancer markers ESR1 and Ki-67 exhibited a clear change with age (Figure S3). In contrast to findings of a previous investigation of young breast cancer patients,[44] we found that PR, HER2 and epidermal growth factor receptor (EGFR) mRNA expression were not significantly different among TCGA age groups. Despite the diagnosis of ER-positivity in young breast cancer patients, expression of ESR1 was 2.7-fold higher in older individuals than younger individuals, whereas expression of Ki-67 was lower, with a fold change of .75. Further details about somatic variants or pathways are discussed in the following subsections.

3.2 | Mutations and copy number alterations

A total of 5765 somatic mutations were identified: 1971 missense, 125 nonsense, 96 splice site, 87 frame-shift indels and 17 in-frame indels. A complete mutation list is provided in Table S2. Our result implied relatively higher mutation rates (KYBR, 2.4 mutations/Mb; TCGA young, 1.12 mutations/Mb; TCGA old, 2.20 mutations/Mb) and fewer cases with a hypermutation rate greater than 10 (KYBR, 4%; TCGA young, 8.9%; TCGA old, 9.3%). Mutational processes of tumor samples were revealed by highly occurring mutation signature analysis. We identified 2 dominant known signatures, APOBEC enzyme activity (signature 13) and age-associated C > T transitions (signature 1) (Figure 2A). A comparison with age suggested that signature 1 steadily increased with senescence. The proportion of signature 1 increased with aging in both

KYBR (r = .37) and young TCGA (r = .37) patients, but a higher proportion of mutations in old-age TCGA patients (r = −.13) consistently consisted of aging-associated signatures (>95% of patients) (Figure 2C).

Recurrent mutations in 3 genes (*TP53* [23%], *PIK3CA* [21%] and *GATA3* [21%]) were detected in 53% of KYBR patients. *TP53* mutations prevailed in group C (P = 0.001) and *GATA3* mutations predominated in group B (P = 0.28; 60% of mutated cases). All *GATA3* mutations consisted of frameshift indels and splicing site mutations that resulted in loss of function. *PIK3CA* mutations occurred in 2 hotspots (p.H1407R/L [13%] and p.E542K [6%]) and the only *AKT1* mutation identified was p.E17K (9%). Rare mutations in various breast cancer markers and genes encoding proteins involved in DNA repair were sporadically distributed among subtypes. These included *BRCA1* (2%) and *BRIP* (6%), involved in the homologous recombination pathway. An *ERBB2* mutation (2%) was discovered in 1 HER2-negative case, and 1 *ESR1* mutation (2%) was detected at a site crucial for estrogen activity. Although *SMARCA4* (4%), *AKT1* (8%) and *ESR1* (2%) genes were rarely mutated, it was confirmed that these genes were enriched in KYBR and young TCGA patients (Figure S4). In contrast, mutations in the frequently mutated gene, *TP53*, showed no age association within TCGA ER-positive breast cancer.

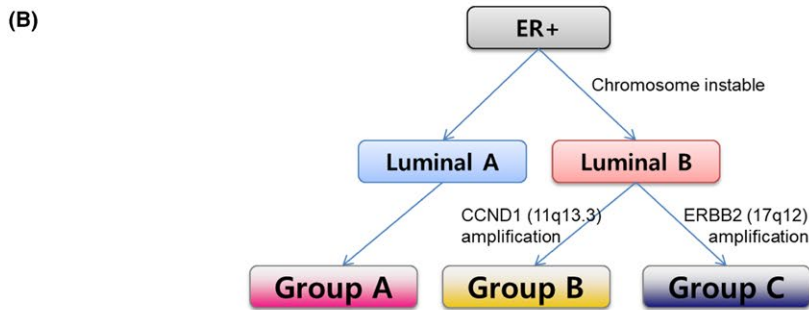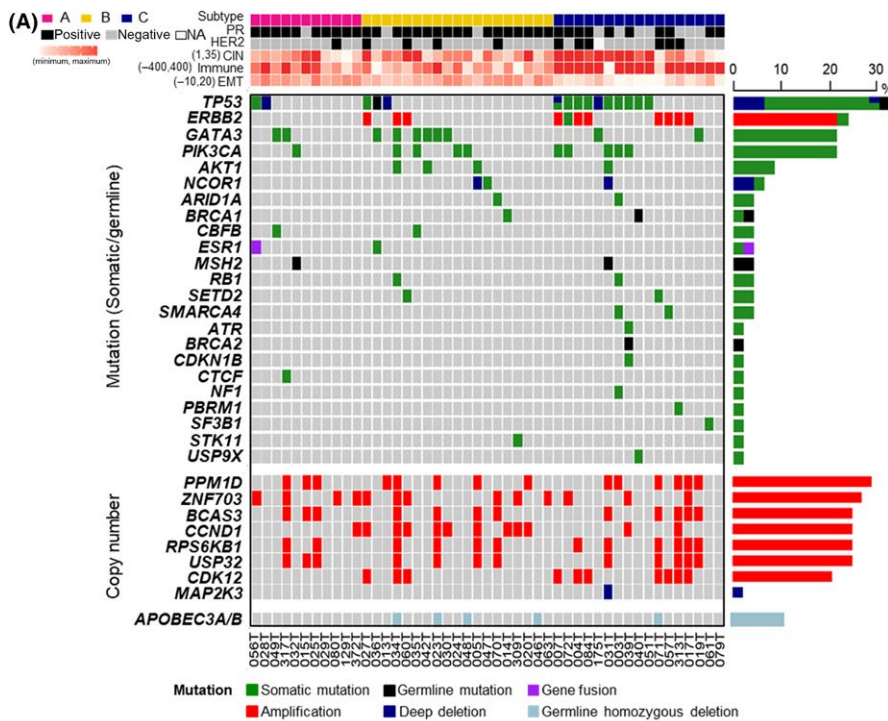In total, 11% of patients harbored pathogenic germline mutations and the frequency is similar with the previously reported 10.7% in an investigation of 25 cancer susceptibility genes.[45] Known pathogenic or truncation germline mutations were discovered in genes encoding *MSH2* (4%), *BRCA1* (2%), *BRCA2* (2%) and *TP53* (2%), which are known to play a role in DNA repair (Table S3). Somatic or germline deficiencies in 4 DNA repair pathway genes (*BRCA1*, *BRCA2*, *TP53* and *MSH2*) accumulated in 34% of KYBR patients.

We identified somatic CNV peaks, compared with previous studies, and revealed that CNV genes were a strongly associated subtype.[46] First, known amplified peak regions were identified in 11q13.3 (*CCND1*; q-value = 8.46×10^{-9}), 17q12 (*ERBB2*; q-value = 2.03×10^{-11}), 17q23 (*RPS6KB1*; q-value = 3.81×10^{-8}) and 8p11.23 (*ZNF703*; q-value = 4.56×10^{-7}) (Figure S5). Moreover, part of a deep-deletion gene was discovered in *TP53* (9%), *NCOR1* (4%) and *MAP2K3* (2%). HER2 (ERBB2) amplification was consistent with immunohistochemistry (IHC) results (P = 1.4×10^{-5}; Fisher's exact test). Our CNV peaks strongly accorded with 3 molecular subgroups based on gene expression. A peak 11q13.3 harboring *CCND1* was mainly amplified in group B (P = 7.084×10^{-3}; Fisher's exact test) and 17q12 of *ERBB2* (P = 2.324×10^{-2}; Fisher's exact test) in subgroup C. Regions of *RPS6KB1* (17q23) and *ERBB2* (17q12) showed more than 2-fold amplification in young patients of both KYBR and TCGA (Figure S5). In contrast, *CCND1* (11q13.3)

TABLE 1 Demographic characteristics of breast cancer patients

	KYBR	TCGA_young	TCGA_old	*P*-value
Patients (n)	47	51	104	
Age (average, y)	31.8 (31.1-32.5)	35.1 (26-39)	81.2 (76-90)	
Hormonal receptor status (n, %)				
ER-positive	47 (100%)	51 (100%)	104 (100%)	-
PR-positive	37 (78.7%)	44 (86.3%)	82 (78.8%)	0.06
HER2-positive	11 (23.4%)	9 (17.6%)	14 (13.5%)	0.31
PAM50 (n, %)				
Basal-like	2 (4.3%)	2 (3.9%)	1 (1.0%)	0.28
HER2-enriched	9 (19.1%)	6 (11.8%)	6 (5.8%)	0.04
Luminal A	24 (51.1%)	22 (43.1%)	64 (61.5%)	0.08
Luminal B	12 (25.5%)	21 (41.2%)	33 (31.7%)	0.25
TNM stage (n, %)				
Stage IA	12 (25.5%)	7 (13.7%)	23 (22.1%)	0.31
Stage IIA	16 (34.0%)	11 (21.6%)	26 (25.0%)	0.36
Stage IIB	13 (27.7%)	15 (29.4%)	21 (20.2%)	0.35
Stage IIIA	3 (6.4%)	14 (27.5%)	13 (12.5%)	0.01
Stage IIIB	1 (2.1%)	0 (0%)	6 (5.8%)	0.17
Stage IIIC	2 (4.3%)	4 (7.8%)	9 (8.7%)	0.73
Race				
Asian	47 (100%)	7 (13.7%)	0 (0%)	
Black	0 (0%)	11 (21.6%)	9 (8.7%)	
White	0 (0%)	32 (62.7%)	74 (71.2%)	
N/A	0 (0%)	1 (2.0%)	21 (20.2%)	

TCGA_young and TCGA_old groups were defined as the patients aged 40 or younger and 75 or older, respectively. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; PR, progesterone receptor.

FIGURE 1 Genomic profiling and integrative summary of molecular characteristics. A, Genomic features heatmap of Korean young breast cancer tumors (n = 47). Three molecular subtypes, and progesterone receptor and HER2 immunohistochemical status; immune scores and epithelial-mesenchyme transition scores inferred from gene expression analysis; chromosomal instability status determined by counts of arm-level alterations. Somatic and germline-level mutations, copy number variants, and fusion variants in 32 genes. Below: APOBEC3A/B homozygous germline deletion status. Right panel: Frequency of variants of each gene, depicted as a bar plot; variant types are discriminated by color. B, Hierarchical classification based on 3 molecular subtypes. Group A belongs to luminal A and is genomically stable; groups B and C are subdivisions of luminal B, classified according to amplification region. ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; NA, not available; PR, progesterone receptor

and *ZNF703* (8p11.23) regions showed no age-dependent changes.

We identified germline deletions on *APOBEC3A/B* (11%), *DMBT1* (deleted in malignant brain tumors; 14%), *GSTM1* (glutathione S-transferase mu; 55%) and *GSTT1* (glutathione S-transferase theta 1; 57%) after stringent filtering to consider homozygous deletion and concordant exonic mRNA expression difference (Table S4). *APOBEC3A/B* deletion status was strongly correlated (cosine-similarity, .86) with the C > T mutation-dominant COSMIC mutation signature 13[47] (Figure 2). The *APOBEC3A/B* frequency of 40% in our KYBR patients is concordant with previous

reports for East Asian populations (approximately 37%), and is much higher than that among Europeans (approximately 8%).[28,30] In particular, APOBEC homozygous or heterozygous deletions were absent in group A and homozygous deletion patients frequently existed in group B.

3.3 | Comparison of genomic characteristics among subgroups

mRNA expression levels, including those of the main diagnostic markers, highlight the subgroup-specific heterogeneity of ER-positive

breast cancer patients (Figure S6). Despite the ER-positive diagnosis of all patients, ER ($P = 1.08 \times 10^{-4}$) and PR ($P = 2.25 \times 10^{-5}$) mRNA were clearly downregulated in group C. The additional prognostic marker Ki-67 ($P = 1.93 \times 10^{-4}$) clearly showed low expression in the good-prognosis group A. Although 10 of the 11 HER2-positive samples (90.9%) were enriched in group B and C; mRNA expression in these samples was more diverse than expected ($P = 0.15$).

To reveal subtype-specific biological functions, we investigated specific pathways based on GSEA (Figure 3). Group A was characterized by IGF1R, and ER-alpha pathways' activation and PLK1 downregulation ($P < 5.0 \times 10^{-4}$). PLK1 (polo like kinase 1), a key regulator of mitosis, cooperates with ER-dependent gene transcription, and its overexpression in cancer cells is associated with poor prognosis.[48] PLK1 downregulation enriched in group A seems to be associated with good-prognosis, high proportion of stromal cells and low chromosome instability ($P = 7.513 \times 10^{-5}$, t test; FC = 2.88; Figure S1).[49]

Truncation mutations of GATA3 frequently existed in group B and the DNA double strand break pathway was dysregulated in group B ($P < 1.59 \times 10^{-6}$).[50] Group B was characterized by activation of EMT and chromosome instability, and inactivation of immune pathways. A survival analysis demonstrated differences among the 3 groups. Group B patients (5-year survival rate, .78; average DFS, 30.6) showed a trend toward poorer prognosis than group A (5-year survival rate, .91; average DFS, 41.5 months) and a shorter disease-free survival compared with patients in group C (5-year survival rate, .79; average DFS, 35.0 months; Figure 3D).

Various immune-related pathways, including tumor necrosis factor (TNF), interferon (IFN)-γ, T-cell receptor and co-stimulation by CD28 family proteins, were consistently activated in group C ($P < 4.54 \times 10^{-6}$). Group C showed the highest immune system activation scores ($P = 3.54 \times 10^{-5}$; t test). In addition, EMT and immune scores were mutually exclusive, and discriminated group C from other groups (Figure S7). We next sought to identify specific molecular functions associated with lower EMT scores ($P = 1.12 \times 10^{-8}$) in group C. The metastasis and cancer stem cell markers aldo-keto reductase family 1 member B10 (AKR1B10), C-C motif chemokine ligand 8 (CCL8), CD24 and prostate stem cell antigen (PSCA) were consistently upregulated in group C (Figure S8).
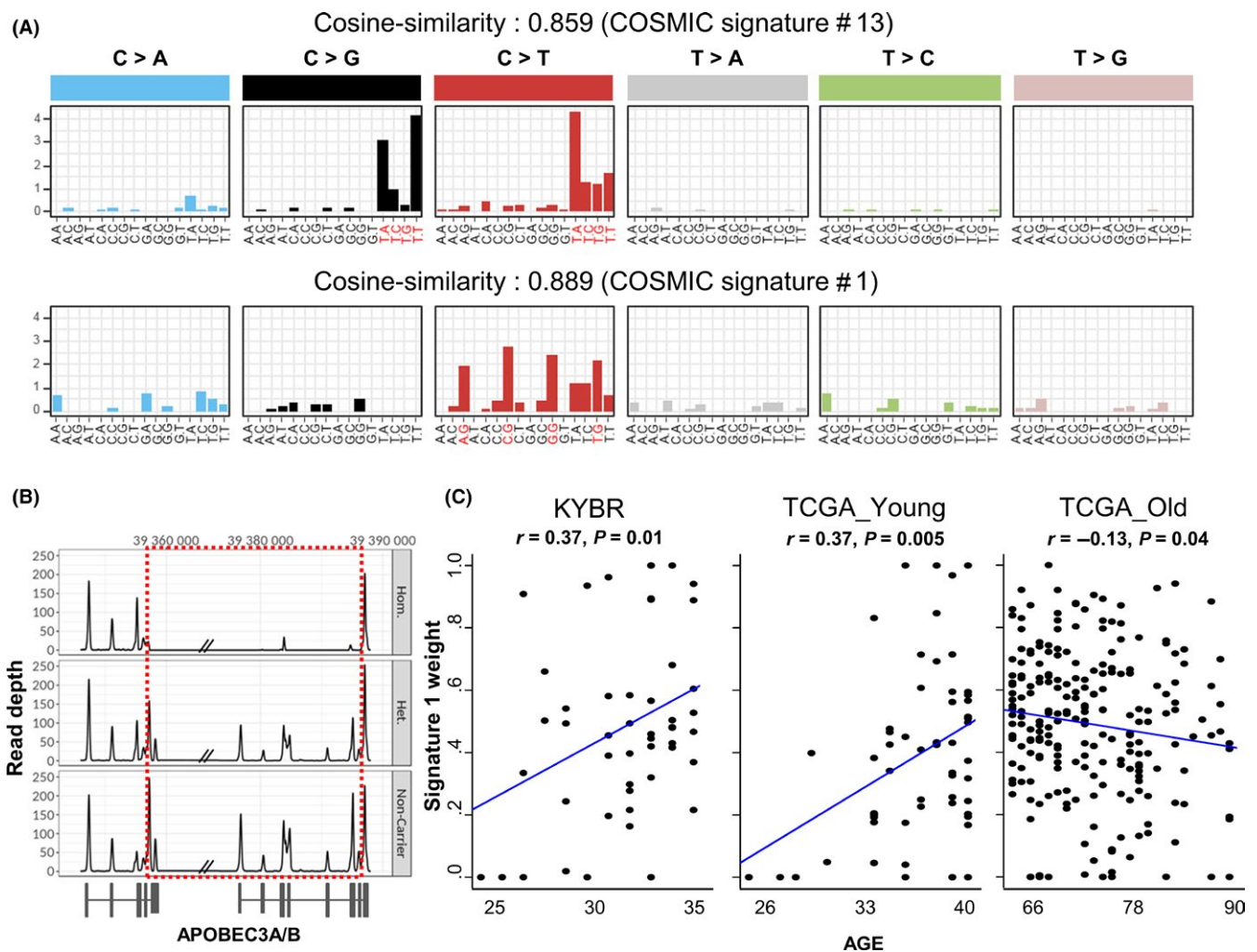

FIGURE 2 Two dominant mutation signatures were identified in our Korean young breast cancer (KYBR) patients. A, Signature 13 of APOBEC mutagenesis and age-associated signature 1. B, Read depth of ABPOBEC3A/B regions. C, Signature 1 increased in KYBR and The Cancer Genome Atlas young age patients. CNV, copy number variant

3.4 | Detection of fusion genes in estrogen receptor-positive breast cancer

We identified fusion proteins from RNA-Seq read alignments using strict calling steps and the validation process described in the Methods (Figure 4 and Table S5). A total of 170 fusions encompassing 272 genes were detected in 35 patients and included 40 in-frame fusions. Fusion transcripts of ESR1 (2%) and ERBB2 (2%) were also detected (Figure 4A). We investigated the possibility of fusions around CNV segmentation breakpoints.[51,52] Loss-of-function fusions of the autophagy regulator vacuole membrane protein 1 (VMP1; 10%) and ERα coactivator breast carcinoma amplified sequence 3 (BCAS3; 8%) were repeatedly observed around CNV peak 17q23, a finding similar to that reported in a previous study.[52,53] Other fusion genes identified, including chemokine signal, PI3K-Akt signal, IGF1R signal and FOXM1 transcription factor, among others (Table S6), have consistently been linked to breast cancer-associated pathways. Of particular note is the novel fusion ESR1-ARMT, an intra-chromosomal short fusion located in 6q25.1 (Figure 4B), a region known to be a strong breast cancer susceptibility candidate.[54] This fusion was detected in an HER2-negative patient in group A.

4 | DISCUSSION

Young patients with breast cancer face clinical issues of poor prognosis, treatment resistance and diminished quality of life. Here, we sought to identify molecular characteristics of breast cancer in very young women (≤35 years of age) using exome and transcriptome profiling. We supposed that aging could influence the occurrence of somatic variant based on mutation rate and mutation signature analysis. We found rare germline mutations in more than 10% of cancer susceptibility genes in our KYBR patients, a finding consistent with a previous study.[45] Diversity of ESR1 variants implicated the heterogeneity of ER-positive breast cancer. We identified its somatic mutation and fusion genes (4%). In addition, BCAS3 fusions could interrupt regular ERα coactivation. Similarly, patients (4%) also harbored various variants of mutation and fusions in ERBB2 (HER2). This highlights the importance of considering resistance to endocrine therapy in this patient population, and suggests that identifying complex genetic variants in ER-positive breast cancer patients would aid in the development of precise, personalized treatment strategies.

We identified hierarchical molecular subtypes within ER-positive breast cancer and confirmed interconnections among gene expression, mutations and CNV. As expected in luminal A type, group A showed better prognosis compared with other groups. We could categorize luminal B cases as group B and C, defined based on immune cell infiltration status. Notably, CCND1 amplifications and GATA3 mutations were prominently detected in group B, and mRNA expression of ubiquitin-mediated proteolysis pathway-related genes was confirmed in this group. Immune-activation group C was characterized by activated immune cells, including CD8+ T cells and M1-type macrophage. We further found that PLK1 conferring chromosome stability was a
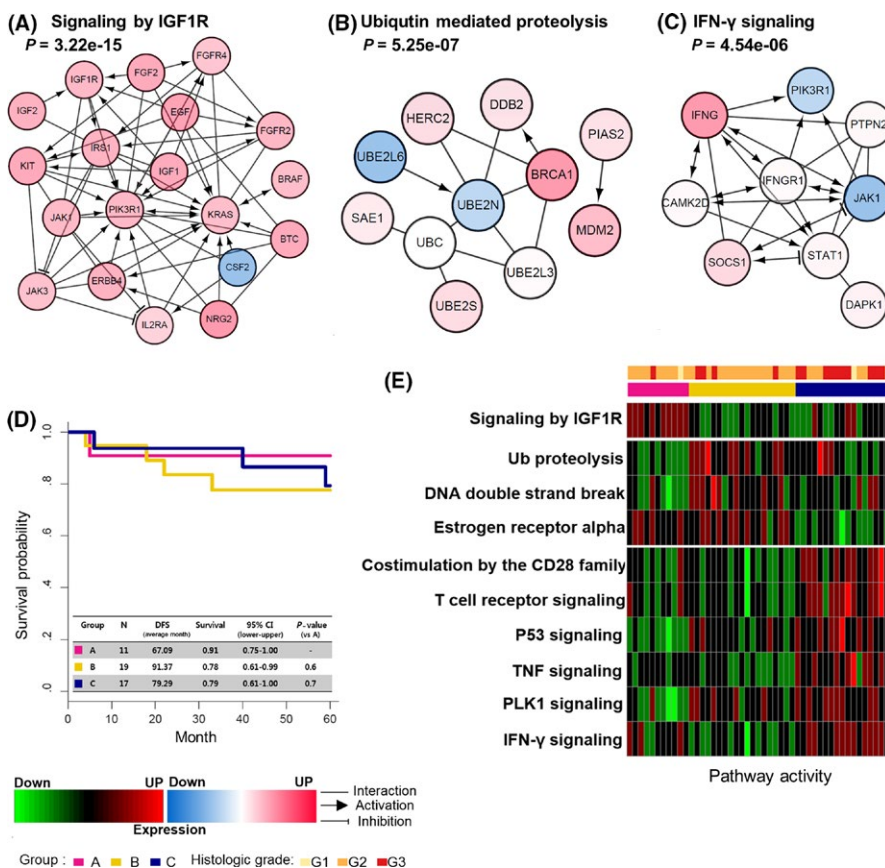

FIGURE 3 Representative pathways for each molecular subtype, inferred from gene expression profiles. A-C, Dysregulated pathways and gene set enrichment analysis (GSEA) P-values for differentially expressed genes for subtype groups A, B and C. Network nodes are rendered in colors based on gene expression profiles. Bar plots summarize total gene set expression for each group. D, Survival plotted according to molecular subtype. A log-rank test was performed for group A. Survival rate, 95% confidential intervals and P-values are summarized in the table; E, Signature gene expression heatmap of the corresponding group A-B pathway. F, Significant pathways identified by GSEA
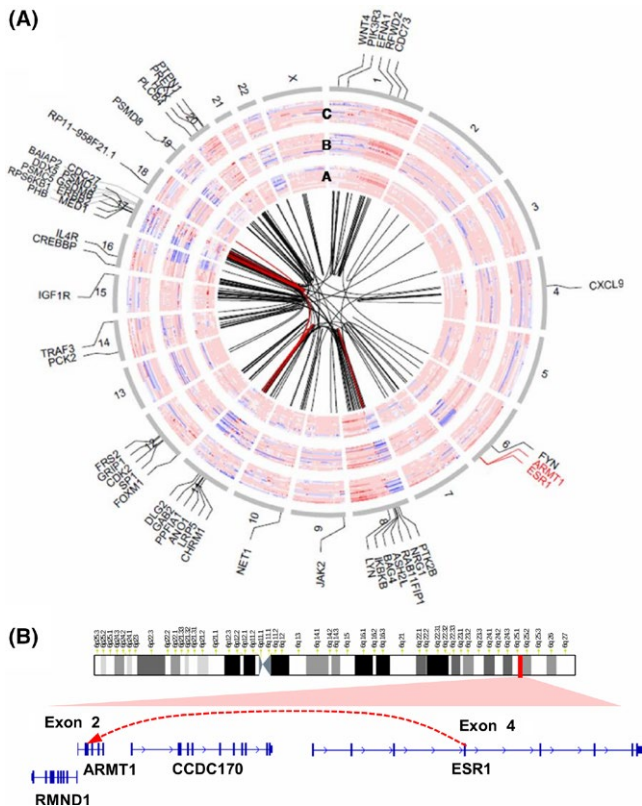
(A)

(B)

FIGURE 4 Fusion genes identified in our Korean young breast cancer patients. A, Circos plot that includes fusion gene breakpoints, a chromosomal copy number variant (CNV) segmentation heatmap (red, amplification; blue, deletion) and names of genes that satisfy gene set enrichment analysis and gene expression evidence. Three CNV tracks were divided according to molecular subtype A, B and C. Fusions, including break points in amplification peak regions 8p11.23, 11q13.3 and 17q12, are highlighted (red lines). B, ESR1-ARMT1 fusion structure. This fusion is a frameshift located in the chr6.q25.1 gene cluster region near the known prognosis-associated fusion, ESR1-CCDC170

potential strong therapeutic potential marker with the ability to discriminate luminal A and B. GATA3 loss-of-function mutations uniquely discriminated group B and offered a potential therapeutic target within luminal B type. Finally, CD8+ T cell-associated immunotherapy could be appropriate for group C patients.

Fusion genes were also consistent features of the mutation, CNV, and gene expression profile landscape. Interestingly, the breakpoints or partner genes identified here are different from those of previous reported fusions. Thus, it is important to validate consistent fusion genes associated with the PIK3CA-Akt pathway, including those involving Janus kinase 2 (JAK2), PIK3RC, RPS6KB1 and IGF1R. In the current study, ESR1 fusion was a rare finding (2%), detected in a single HER2-negative patient in group A. By comparison, a previous study investigating recurrent ESR1-CCDC170 fusions suggested a degree of enrichment of such fusions in HER-positive patients (luminal A .9%, luminal B 2.9% and HER2 3.1%).[55] We also detected an intra-chromosomal ERBB2-ORMDL3 frameshift fusion within 230 kbp of 17q12, a distance longer than the reported 106-kbp ERBB2 amplicon region.[56]

The different patterns of mutation types among subgroups also suggested potential activation of pathways that could affect treatment efficiency. Prognosis was predicted to be worse for patients in groups B and C, characterized by chromosome instability, than in luminal A patients, a finding that could be consistent with the poor prognosis of young patients.[57] Group C demonstrated highly activated immune scores that could be applicable for immune therapy. However, owing to the limited number of patients, we were unable to detect a statistically significant difference in survival between groups. Therefore, additional studies using a much larger number of patients will be necessary to elucidate the clinical implications of the observed molecular differences among subgroups in young patients.

This study demonstrated mutation signatures and the somatic mutations that were enriched in young patients. Integrative genomic profiling could classify very young patients with breast cancer into 3 subgroups based on distinct molecular features that revealed the biological aspects. Each subgroup was characterized by the different signaling of IGF1R, PLK1 and ubiquitin-mediated proteolysis. Chromosomal instability, activated EMT and inactivation of immune pathways were important features of clustering, suggesting different clinical manifestations of each subgroup.

ACKNOWLEDGMENTS

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interests.

ORCID

Sun-Young Kong 🔟 https://orcid.org/0000-0003-0620-4058

REFERENCES

1. Shin HR, Joubert C, Boniol M, et al. Recent trends and patterns in breast cancer incidence among Eastern and Southeastern Asian women. *Cancer Causes Control.* 2010;21:1777-1785.
2. Jemal A, Center MM, DeSantis C, Ward EM. Global patterns of cancer incidence and mortality rates and trends. *Cancer Epidemiol Biomarkers Prev.* 2010;19:1893-1907.
3. Sung H, Rosenberg PS, Chen WQ, et al. Female breast cancer incidence among Asian and Western populations: more similar than expected. *J Natl Cancer Inst.* 2015;107:djv109.
4. Jung KW, Won YJ, Kong HJ, et al. Cancer statistics in Korea: incidence, mortality, survival, and prevalence in 2012. *Cancer Res Treat.* 2015;47:127-141.
5. Liu L, Zhang J, Wu AH, Pike MC, Deapen D. Invasive breast cancer incidence trends by detailed race/ethnicity and age. *Int J Cancer.* 2012;130:395-404.
6. Ly D, Forman D, Ferlay J, Brinton LA, Cook MB. An international comparison of male and female breast cancer incidence rates. *Int J Cancer.* 2013;132:1918-1926.

7. Ahn SH, Son BH, Kim SW, et al. Poor outcome of hormone receptor-positive breast cancer at very young age is due to tamoxifen resistance: nationwide survival data in Korea–a report from the Korean Breast Cancer Society. *J Clin Oncol*. 2007;25:2360-2368.

8. Wei X-Q, Li X, Xin X-J, Tong Z-S, Zhang S. Clinical features and survival analysis of very young (age < 35) breast cancer patients. *Asian Pac J Cancer Prev*. 2013;14:5949-5952.

9. Cancello G, Maisonneuve P, Rotmensz N, et al. Prognosis and adjuvant treatment effects in selected breast cancer subtypes of very young women (<35 years) with operable breast cancer. *Ann Oncol*. 2010;21:1974-1981.

10. Peng R, Wang S, Shi Y, et al. Patients 35 years old or younger with operable breast cancer are more at risk for relapse and survival: a retrospective matched case-control study. *Breast*. 2011;20:568-573.

11. Cancer Genome Atlas N. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490:61-70.

12. Ciriello G, Gatza ML, Beck AH, et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*. 2015;163:506-519.

13. Mertins P, Mani DR, Ruggles KV, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature*. 2016;534:55-62.

14. Hammond ME, Hayes DF, Dowsett M, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. *J Clin Oncol*. 2010;28:2784-2795.

15. Hammond ME, Hayes DF, Wolff AC. Clinical Notice for American Society of Clinical Oncology-College of American Pathologists guideline recommendations on ER/PgR and HER2 testing in breast cancer. *J Clin Oncol*. 2011;29:e458.

16. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30:2114-2120.

17. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2010;26:589-595.

18. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297-1303.

19. Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*. 2013;31:213-219.

20. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics*. 2012;28:1811-1817.

21. Ramos AH, Lichtenstein L, Gupta M, et al. Oncotator: cancer variant annotation tool. *Hum Mutat*. 2015;36:E2423-E2429.

22. Kanchi KL, Johnson KJ, Lu C, et al. Integrated analysis of germline and somatic variants in ovarian cancer. *Nat Commun*. 2014;5:3156.

23. Gehring JS, Fischer B, Lawrence M, Huber W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics*. 2015;31:3673-3675.

24. D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A. Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. *Nucleic Acids Res*. 2016;44:e154.

25. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*. 2011;12:R41.

26. Klambauer G, Schwarzbauer K, Mayr A, et al. cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res*. 2012;40:e69.

27. Buccitelli C, Salgueiro L, Rowald K, Sotillo R, Mardin BR, Korbel JO. Pan-cancer analysis distinguishes transcriptional changes of aneuploidy from proliferation. *Genome Res*. 2017;27:501-511.

28. Kidd JM, Newman TL, Tuzun E, Kaul R, Eichler EE. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genet*. 2007;3:e63.

29. Middlebrooks CD, Banday AR, Matsuda K, et al. Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat Genet*. 2016;48:1330-1338.

30. Nik-Zainal S, Wedge DC, Alexandrov LB, et al. Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. *Nat Genet*. 2014;46:487-491.

31. Wang K, Singh D, Zeng Z, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res*. 2010;38:e178.

32. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323.

33. Beisser D, Klau GW, Dandekar T, Muller T, Dittrich MT. BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics*. 2010;26:1129-1130.

34. Wu G, Dawson E, Duong A, Haw R, Stein L. ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis. *F1000Res*. 2014;3:146.

35. Yoshihara K, Shahmoradgoli M, Martinez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun*. 2013;4:2612.

36. Tan TZ, Miow QH, Miki Y, et al. Epithelial-mesenchymal transition spectrum quantification and its efficacy in deciphering survival and drug responses of cancer patients. *EMBO Mol Med*. 2014;6:1279-1293.

37. McPherson A, Hormozdiari F, Zayed A, et al. deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. *PLoS Comput Biol*. 2011;7:e1001138.

38. Torres-Garcia W, Zheng S, Sivachenko A, et al. PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics*. 2014;30:2224-2226.

39. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*. 2013;29:15-21.

40. Yoshihara K, Wang Q, Torres-Garcia W, et al. The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*. 2015;34:4845-4854.

41. Abate F, Zairis S, Ficarra E, et al. Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC Syst Biol*. 2014;8:97.

42. Hanzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;14:7.

43. Azim HA Jr, Nguyen B, Brohee S, Zoppoli G, Sotiriou C. Genomic aberrations in young and elderly breast cancer patients. *BMC Med*. 2015;13:266.

44. Anders CK, Hsu DS, Broadwater G, et al. Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression. *J Clin Oncol*. 2008;26:3324-3330.

45. Tung N, Lin NU, Kidd J, et al. Frequency of germline mutations in 25 cancer susceptibility genes in a sequential series of patients with breast cancer. *J Clin Oncol*. 2016;34:1460-1468.

46. Pereira B, Chin SF, Rueda OM, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat Commun*. 2016;7:11479.

47. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature*. 2013;500:415-421.

48. Liu Z, Sun Q, Wang X. PLK1, a potential target for cancer therapy. *Transl Oncol*. 2017;10:22-32.

49. Wierer M, Verde G, Pisano P, et al. PLK1 signaling in breast cancer cells cooperates with estrogen receptor-dependent gene transcription. *Cell Rep*. 2013;3:2021-2032.

50. Takaku M, Grimm SA, Wade PA. GATA3 in breast cancer: tumor suppressor or oncogene? *Gene Expr*. 2015;16:163-168.

51. Edgren H, Murumagi A, Kangaspeska S, et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biol*. 2011;12:R6.

52. Inaki K, Hillmer AM, Ukil L, et al. Transcriptional consequences of genomic structural aberrations in breast cancer. *Genome Res*. 2011;21:676-687.

53. Barlund M, Monni O, Weaver JD, et al. Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. *Genes Chromosom Cancer*. 2002;35:311-317.

54. Wang Y, He Y, Qin Z, et al. Evaluation of functional genetic variants at 6q25.1 and risk of breast cancer in a Chinese population. *Breast Cancer Res*. 2014;16:422.

55. Veeraraghavan J, Tan Y, Cao XX, et al. Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers. *Nat Commun*. 2014;5:4577.

56. Ferrari A, Vincent-Salomon A, Pivot X, et al. A whole-genome sequence and transcriptome perspective on HER2-positive breast cancers. *Nat Commun*. 2016;7:12222.

57. Tang LC, Jin X, Yang HY, et al. Luminal B subtype: a key factor for the worse prognosis of young breast cancer patients in China. *BMC Cancer*. 2015;15:201.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.