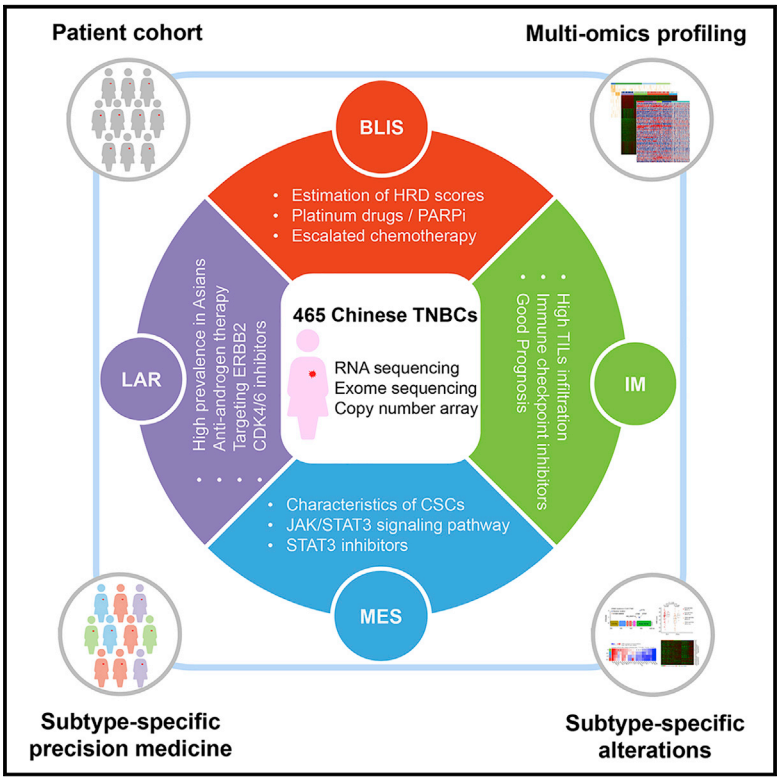
Article

Cancer Cell

Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies

Graphical Abstract



Highlights

- We build the genomic and transcriptomic landscape of 465 primary TNBCs
- Chinese TNBC cases demonstrate more *PIK3CA* mutations and LAR subtype
- Transcriptomic data classify TNBCs into four subtypes
- Multi-omics profiling identifies potential targets within specific TNBC subtypes

Authors

Yi-Zhou Jiang, Ding Ma, Chen Suo, ..., Leming Shi, Wei Huang, Zhi-Ming Shao

Correspondence

wangpeng@picb.ac.cn (P.W.), lemingshi@fudan.edu.cn (L.S.), huangwei@chgc.sh.cn (W.H.), zhimingshao@yahoo.com (Z.-M.S.)

In Brief

Jiang et al. characterize primary Chinese triple-negative breast cancer (TNBC) and classify it into four subtypes. They find that these TNBCs have more frequent *PIK3CA* mutations and chromosome 22q11 copy-number gains than non-Asian TNBCs and that the LAR subtype has more *ERBB2* somatic mutations and *CDKN2A* loss.

Jiang et al., 2019, Cancer Cell *35*, 428–440
March 18, 2019
https://doi.org/10.1016/j.ccell.2019.02.001

CellPress

CellPress

Cancer Cell
Article

Genomic and Transcriptomic Landscape of Triple-Negative Breast Cancers: Subtypes and Treatment Strategies

Yi-Zhou Jiang,[1,14] Ding Ma,[1,14] Chen Suo,[2,3,14] Jinxiu Shi,[4,14] Mengzhu Xue,[5,14] Xin Hu,[1,14] Yi Xiao,[1] Ke-Da Yu,[1] Yi-Rong Liu,[1] Ying Yu,[2] Yuanting Zheng,[2] Xiangnan Li,[2] Chenhui Zhang,[4] Pengchen Hu,[4] Jing Zhang,[4] Qi Hua,[4] Jiyang Zhang,[2] Wanwan Hou,[2] Luyao Ren,[2] Ding Bao,[2] Bingying Li,[2] Jingcheng Yang,[2] Ling Yao,[1] Wen-Jia Zuo,[1] Shen Zhao,[1] Yue Gong,[1] Yi-Xing Ren,[1] Ya-Xin Zhao,[1] Yun-Song Yang,[1] Zhenmin Niu,[4] Zhi-Gang Cao,[1,13] Daniel G. Stover,[6] Claire Verschraegen,[6] Virginia Kaklamani,[7] Anneleen Daemen,[8] John R. Benson,[9] Kazuaki Takabe,[10] Fan Bai,[11] Da-Qiang Li,[1] Peng Wang,[12,*] Leming Shi,[2,*] Wei Huang,[4,*] and Zhi-Ming Shao[1,15,*]

[1]Department of Breast Surgery, Precision Cancer Medicine Center, Fudan University Shanghai Cancer Center, 270 Dong'an Road, Shanghai 200032, P.R. China

[2]State Key Laboratory of Genetic Engineering, School of Life Sciences and Human Phenome Institute, Fudan University, 2005 Songhu Road, Shanghai 200438, P.R. China

[3]Department of Epidemiology, School of Public Health, Fudan University, Shanghai 200032, P.R. China

[4]Shanghai-MOST Key Laboratory of Health and Disease Genomics, Chinese National Human Genome Center at Shanghai (CHGC) and Shanghai Industrial Technology Institute (SITI), 250 Bibo Road, Shanghai 201203, P.R. China

[5]SARI Center for Stem Cell and Nanomedicine, Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, P.R. China

[6]The Ohio State University Comprehensive Cancer Center, Columbus, OH 43210, USA

[7]Division Hematology/Oncology, University of Texas Health Science Center San Antonio, San Antonio, TX 78284, USA

[8]Department of Bioinformatics & Computational Biology, Genentech Inc., 1 DNA Way, South San Francisco, CA 94080, USA

[9]Cambridge Breast Unit, Addenbrooke's Hospital, Hills Road, Cambridge CB2 0QQ, UK

[10]Division of Breast Surgery, Department of Surgical Oncology, Roswell Park Cancer Institute, Buffalo, NY 14263, USA

[11]Biodynamic Optical Imaging Center (BIOPIC), School of Life Sciences, Peking University, Beijing 100871, P.R. China

[12]Bio-med Big Data Center, CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, P.R. China

[13]Present address: AME Publishing Company, Kings Wing Plaza 1, No. 3 On Kwan Street, Hong Kong, P.R. China

[14]These authors contributed equally

[15]Lead Contact

*Correspondence: wangpeng@picb.ac.cn (P.W.), lemingshi@fudan.edu.cn (L.S.), huangwei@chgc.sh.cn (W.H.), zhimingshao@yahoo.com (Z.-M.S.)

https://doi.org/10.1016/j.ccell.2019.02.001

SUMMARY

We comprehensively analyzed clinical, genomic, and transcriptomic data of a cohort of 465 primary triple-negative breast cancer (TNBC). *PIK3CA* mutations and copy-number gains of chromosome 22q11 were more frequent in our Chinese cohort than in The Cancer Genome Atlas. We classified TNBCs into four transcriptome-based subtypes: (1) luminal androgen receptor (LAR), (2) immunomodulatory, (3) basal-like immune-suppressed, and (4) mesenchymal-like. Putative therapeutic targets or biomarkers were identified among each subtype. Importantly, the LAR subtype showed more *ERBB2* somatic mutations, infrequent mutational signature 3 and frequent *CDKN2A* loss. The comprehensive profile of TNBCs provided here will serve as a reference to further advance the understanding and precision treatment of TNBC.

Significance

Triple-negative breast cancer (TNBC) is highly heterogeneous. Due to the limited number of TNBCs that have been analyzed, molecular events driving subtypes and prognosis are not firmly established, and little is known regarding TNBC in non-Caucasian patients. Our findings advance the understanding of TNBC subtypes, subdivide the established transcriptome-based subtypes in search of more targeted therapeutic strategies, and offer potential insights to guide subtype-specific therapy. East Asian TNBCs demonstrated higher frequencies of *PIK3CA* mutations and the luminal androgen receptor subtype, implying a need for the adjustment of clinical management. This large collection of comprehensively profiled TNBC tumors with well-documented clinical information will serve as a reference to further advance the understanding of TNBC.

INTRODUCTION

Breast cancers that do not express the estrogen receptor (ER) or the progesterone receptor (PR) and have no *ERBB2* (commonly referred to as human epidermal growth factor receptor 2 [HER2]) amplification are categorized as triple-negative breast cancers (TNBCs) and comprise 10%–20% of all breast cancers (Dent et al., 2007; Venkitaraman, 2010). TNBCs occur more frequently in younger patients and the tumors are usually larger in size, of higher grade, more likely to show lymph node involvement at diagnosis, and biologically more aggressive (Carey et al., 2010; Dent et al., 2007; Yin et al., 2009). Women with TNBC have a higher rate of early distant recurrence and a worse 5-year prognosis than do women with other subtypes of breast cancer (Carey et al., 2010; Dignam et al., 2009; Rouzier et al., 2005). Although PARP inhibitors are now approved for *BRCA*-mutant TNBC (Robson et al., 2017) and immune modulators are showing great promise (Schmid et al., 2018), targeted therapeutic treatments for TNBC are still at their early stage, while chemotherapy remains the standard treatment (Bianchini et al., 2016).

In view of the limited clinical benefit observed with targeted therapies in unselected TNBC patients, there has been a research focus on discovering actionable targets through molecular subtyping of TNBC (Baselga et al., 2013). Because actionable somatic mutations are generally low-frequency events in TNBC (Bareche et al., 2018; Lehmann and Pietenpol, 2015; Shah et al., 2012), additional genomic analyses and studies covering a variety of ethnic backgrounds are needed to discover subtypes with clinically meaningful targets.

To date, genomic studies of TNBC have mainly focused on mRNA expression and DNA copy number (Burstein et al., 2015; Curtis et al., 2012; Lehmann et al., 2011), which have provided limited insight into the biological underpinnings, especially the genomic signatures, of this disease. Furthermore, the initial breast cancer study reported in The Cancer Genome Atlas (TCGA) in 2012 assessed just over 100 TNBC tumors using 6 different technological platforms without the formal evaluation of any clinical association (Cancer Genome Atlas Network, 2012). To provide a broader molecular profile of TNBC, we performed multi-omic profiling of a large Chinese TNBC cohort.

RESULTS

Patient Samples and Clinical Data

Primary tumor tissue and blood samples were obtained from 504 consecutive female Chinese patients with TNBC treated at Fudan University Shanghai Cancer Center (FUSCC). Detailed information on patient selection and sample preparation are described in the STAR Methods. After discarding samples that failed the quality check, a final cohort of 465 patients were available for analysis (Figure S1; Table S1). Among these 465 patients, 279 had whole exome sequencing (WES) data on primary tumor tissue and paired blood samples, 401 had copy-number alteration (CNA) data and 360 had RNA sequencing data on primary tumor tissue. These patients underwent surgery at FUSCC between 2007 and 2014 with a median duration of follow-up of 45.8 months. A total of 65 patients experienced relapse and/or metastatic progression during this period (Tables S1 and S2).

Somatic Genomic Alterations in Chinese TNBC

Among the TNBCs with WES data, 25,713 somatic mutations were identified, comprising 24,025 single-nucleotide variants (SNVs) and 1,688 insertions or deletions (INDELs). These tumors harbored a median of 46 nonsynonymous SNVs and 4 INDELs, which are similar to the results for the TCGA TNBC cohort. As demonstrated in Figures 1A and 1B, the most prominent cancer-related variations (An et al., 2016; Futreal et al., 2004) observed in this cohort were *TP53* mutations (found in 74% of tumors), followed by *PIK3CA* (18%), *KMT2C* (7%), and *PTEN* (6%) mutations. Interestingly, *PIK3CA*, *PTEN*, and *PIK3R1* mutations were strongly associated with the luminal androgen receptor (LAR) subtype, and a similar pattern was observed for *HRAS* and *ERBB2* mutations despite their low prevalence (2%) in the entire cohort (see Table S3, details on subtypes will be discussed below). In particular, all five *ERBB2* mutations occurred in a group of patients designated as the LAR subtype (Figure 1B).

To assess somatic copy number alterations (CNAs), reported oncogenes and tumor suppressor genes observed in the top rank (residual $q < 1 \times 10^{-4}$) GISTIC peaks were examined (Figure 1C; Table S4) (Bareche et al., 2018; Nik-Zainal et al., 2016). *MYC* was the gene most frequently affected by somatic CNAs (gained in 81% of patient samples), with frequent gains in *E2F3* (55%), *IRS2* (49%), *CCNE1* (47%), *EGFR* (47%), *NFIB* (44%), *CCND1* (44%), and *MYB* (41%) and frequent losses in *CHD1* (lost in 71% of samples), *PTEN* (58%), *RB1* (54%), and *CDKN2A* (43%).

We further investigated the differences in genomic features between Chinese and TCGA TNBC cohorts. Chinese TNBCs had a higher *PIK3CA* mutation rate than do TCGA TNBCs overall (18% versus 10%, respectively), which appeared to be driven primarily by the differences between African American TNBCs and Chinese TNBCs (5% versus 18%, respectively, $p = 0.03$). Compared with Caucasian patients, the Chinese patients did not show a significantly higher *PIK3CA* mutation rate (Figure 2A). In terms of somatic CNA differences, the Chinese TNBCs had more frequent somatic copy-number gains on chromosome 22q11 than TCGA TNBCs (median alteration frequency, Chinese versus TCGA African American versus TCGA Caucasian: 44% versus 19% versus 15%, respectively; false discovery rate [FDR] < 0.1; Figure 2B; Table S5), spanning 115 genes that include 5 known tumor-associated genes: *BCR* (GTPase-activating protein), *MAPK1* (MAP kinase pathway), *CLTCL1* (clathrin heavy chain family), *LZTR1* (leucine-zipper-like transcription regulator 1, located in the Golgi apparatus), and *SEPT5* (septin gene family of nucleotide-binding proteins). Expression levels of *BCR*, *MAPK1*, *LZTR1*, and *SEPT5* were also significantly elevated in those cases with copy-number gain or amplification (Table S5).

In addition to the genomic differences, we also extrapolated our mRNA-based TNBC subtypes (discussed below; Figure S2) to TCGA using a nearest shrunken centroids method (Tibshirani et al., 2002) (Table S6). Compared with TCGA TNBC patients, Chinese patients had a higher prevalence of the LAR subtype (23% in Chinese versus 9% in TCGA African Americans, $p = 0.008$; versus 12% in TCGA Caucasians, $p = 0.026$; Figure 2C; Table S6). Interestingly, while only nine of the TCGA TNBC patients were Asian, three of them (33%) were identified as the LAR subtype.

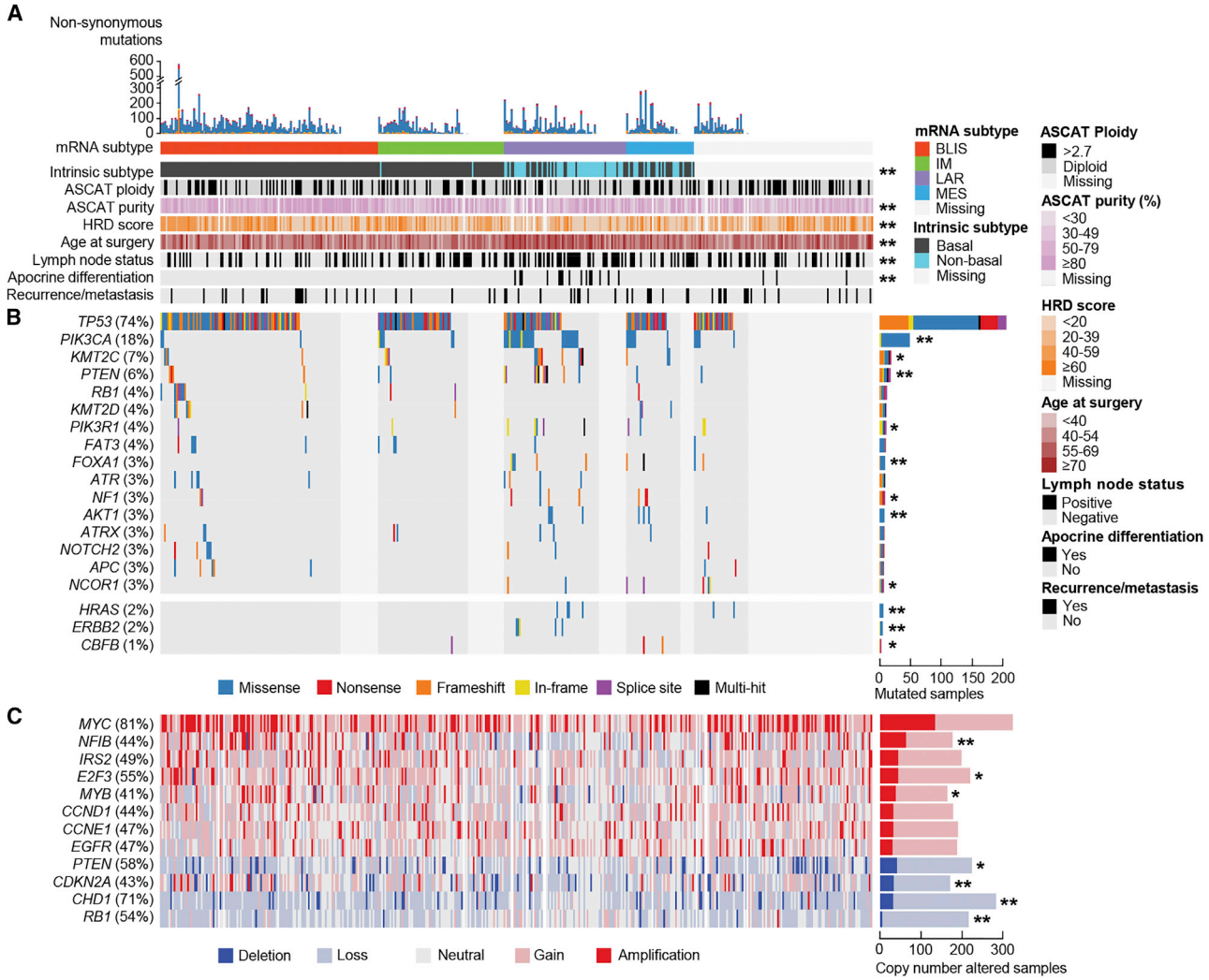


Figure 1. The Genomic Landscape of Chinese TNBC

(A) Four hundred and nineteen TNBC samples with mutation and/or copy-number alteration data are ordered by mRNA subtypes and mutation profile, with clinical and molecular features annotated below.

(B) Known cancer-related genes (An et al., 2016; Futreal et al., 2004) that were mutated in at least 2.5% of the cases (upper) or differentially mutated per mRNA subtypes (lower).

(C) Known cancer-related genes located in significant GISTIC peaks with residual q < 1×10^{-4} (amplification, gain, neutral, loss, and deletion were defined as GISTIC +2, +1, 0, −1, and −2, respectively).

(A–C) Asterisks indicate associations with mRNA subtypes (intrinsic subtype, ASCAT ploidy, lymph node status, and somatic copy-number alterations were tested using Pearson's chi-square test; ASCAT purity and homologous recombination deficiency [HRD] score were tested using the Kruskal-Wallis test; age at surgery was tested using analysis of variance; apocrine differentiation and somatic mutations were tested using Fisher's exact test. **p < 0.01, *p < 0.05).

See also Figure S1 and Tables S1–S4.

TNBC Subtypes Based on Multi-omics Data

We performed consensus clustering by resampling (1,000 iterations) randomly selected tumor profiles (Figure S2A). According to the "elbow" point in the relative change in area (Δ) under the consensus distribution function plot (Figure S2B), we chose four as the number of subtypes. Based on these findings, we classified 360 tumors into 4 separate subtypes using unsupervised k-means clustering based on the top 2,000 most differentially expressed coding genes (Figures 3A, S2C, and S2D). These four subtypes consisted of the following: (1) a luminal androgen receptor (LAR) subtype (23%) characterized by androgen receptor signaling; (2) an immunomodulatory (IM) subtype (comprising 24% of tumors) with high immune cell signaling and cytokine signaling gene expression; (3) a basal-like and immune-suppressed (BLIS) (39%) subtype characterized by upregulation of cell cycle, activation of DNA repair, and downregulation of immune response genes; and (4) a mesenchymal-like (MES) subtype (15%) enriched in mammary stem cell pathways (Figure 3A; Table S7).The results were correlated with the TNBC subtypes defined by Lehmann and Pietenpol (2015) (Figure 3A). Several

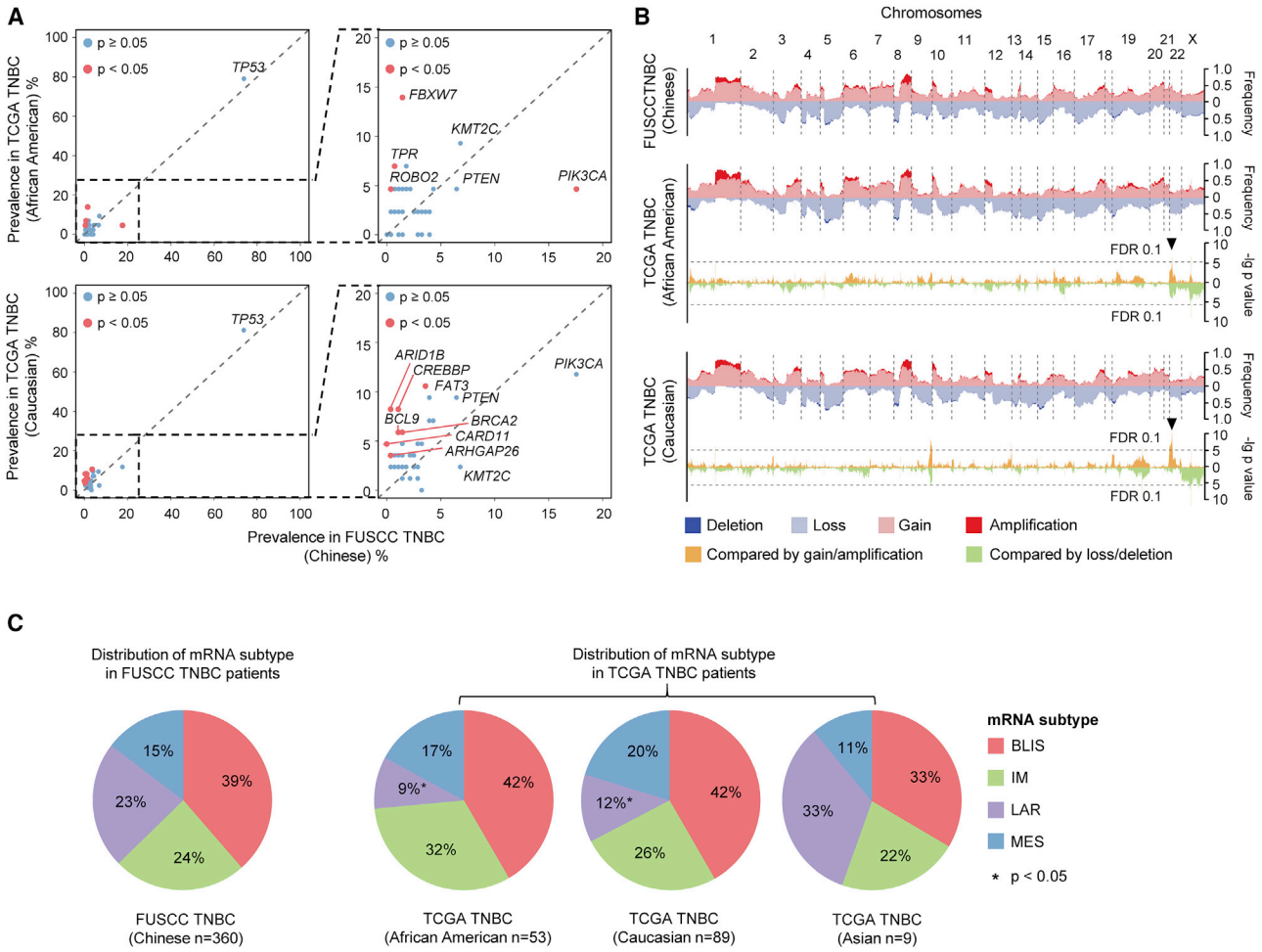


Figure 2. Population-Specific Genomic Events in Chinese TNBC

(A) Mutation prevalence of cancer-related genes in Chinese TNBC and TCGA African American and Caucasian cohorts.

(B) Somatic copy-number alterations (CNAs) between Chinese and TCGA datasets. TCGA panels consisted of somatic CNA frequency (upper) and –log10 unadjusted p values for comparison with the FUSCC TNBC somatic CNA profile (lower, calculated using Fisher's exact test). Regions enriched in Chinese TNBC patients (at least five continuous genes with false discovery rate [FDR] < 0.1 were considered significant) are marked with a triangle.

(C) Distribution of TNBC mRNA subtypes in Chinese and TCGA TNBCs. *p < 0.05 (luminal androgen receptor subtype or not, compared with the Chinese TNBC patients using Fisher's exact test).

See also Tables S5 and S6.

candidate gene targets were identified within each subtype (Table 1), indicating the potential application of future subtype-specific precision treatments (details on each subtype will be discussed below).

DNA changes, including somatic mutations and CNAs, not only serve as "drivers" for boosting tumor cell growth but also contain footprints attributable to specific cellular processes, reflecting the innate biological characteristics of tumors. Using k-means clustering and consensus clustering, methods similar to mRNA clustering, six clusters based on CNA peaks were identified (Figure 3B; Table S8): CNA subtype 1, frequent 9p23 amplification (Chr9p23 amp); CNA subtype 2, frequent 12p13 amplification (Chr12p13 amp); CNA subtype 3, frequent Chr13q34 amplifications (Chr13q34 amp); CNA subtype 4, frequent Chr20q13 amplification (Chr20q13 amp); CNA subtype 5, frequent Chr8p21 loss (Chr8p21 del); and CNA subtype 6, somatic CNA lacking a CN cluster but with low chromosomal instability (CIN) (low CIN).

In addition, four subtypes based on Catalog of Somatic Mutation In Cancer mutational signatures evaluated by deconstructSigs were identified (Figure 3C) (Alexandrov et al., 2015; Rosenthal et al., 2016): mutation subtype 1, which was dominated by APOBEC-related signatures 2 and 13 (APOBEC); mutation subtype 2, which was highlighted by homologous recombination deficiency (HRD)-related signature 3 (HRD); mutation subtype 3, with clock-like signatures 1 and 5 (clock-like); and mutation subtype 4, with no dominant signature (mixed).

To more comprehensively understand the interaction of these subtypes and mutational signature approaches, we investigated the specific associations among expression, copy number, and mutational signature-based subtypes (Figure 3D). For example, the majority of the LAR subtype samples (85%) were classified

as either low CIN or Chr8p21 del, while simultaneously the LAR subtype was the minority in the mutation HRD subtype (7%). Similar results were observed when using iClusterPlus (Figures S3A and S3B; Table S1). The iCluster cluster 8 (iC8) consisted almost exclusively of the mRNA LAR subtype (32/33, 97%), and the majority of iC8 also corresponded to the CNA low-CIN subtype (25/33, 76%), while only one case was classified as the mutational HRD subtype (1/33, 3%). Despite the lack of a specific one-to-one correspondence between the other subtypes, specific relationships were still observed. For example, most Chr9p23 amp and Chr12p13 amp tumors (61% and 74%, respectively) were of the BLIS subtype, and the BLIS subtype also constituted 65% of the mutational HRD subtype based on the mutational signatures.

The potential prognostic value of these subtype strategies was investigated. Neither the copy number nor the mutational clustering reliably predicted relapse-free survival (RFS). No overall difference in RFS was found between the four mRNA subtypes ($p = 0.122$). However, the IM tumors showed a significantly better prognosis after adjusting for lymph node status and tumor size (versus BLIS, hazard ratio: 0.34; 95% confidence intervals: 0.13–0.85; $p = 0.021$; Figure 3E). The patients who were classified into these four TNBC subtypes also demonstrated distinct clinical characteristics (Table S2). Specifically, the LAR patients were older (84% were 51 years of age or older), and 28% of the BLIS patients and 55% of the LAR patients had histologically positive lymph nodes.

Activated ERBB2 and Cell-Cycle Signaling in the LAR Subtype

To explore potential therapeutic targets in the four TNBC subtypes, distinct genomic alterations were identified. Five LAR patients (9% in LAR versus 0% in other subtypes, $p < 0.001$; Figures 1B and 4A; Table S9) were found to harbor *ERBB2* nonsynonymous mutations. Similarly, in the TCGA TNBC cohort, four of the six TNBCs harboring nonsynonymous *ERBB2* mutations were classified as LAR (20% in LAR versus 2% in other subtypes, $p = 0.004$; Figure 4A; Table S9). This was also significantly higher than the *ERBB2* mutation frequency in non-TNBC cases in the TCGA breast cancer dataset (2% in non-TNBC, $p = 0.001$). The *ERBB2* mutations found in our cohort comprised of activating mutations V777L (in two cases), D769Y (in one case), and L755S (in one case), which inferred not only ERBB2 activation but also resistance to trastuzumab and lapatinib and a previously unreported E698_P699delinsA (Figure 4B). None of these five breast cancers were clinically ERBB2-positive according to standard immunohistochemistry (IHC) and fluorescence *in situ* hybridization evaluation (Table S9). Nonetheless ERBB2 signature scores, defined by gene set variation analysis (Hanzelmann et al., 2013), revealed that patients with *ERBB2* mutations also exhibited relative activation of the ERBB2 pathway (Desmedt et al., 2008, 2016) (Figure 4C). Upon further searching the patient database at our center (Zuo et al., 2016), 8 of 58 TNBCs with positive IHC for the androgen receptor also exhibited *ERBB2* somatic mutations.

Despite having fewer CNAs, the LAR subtype was enriched with Chr9p21 loss, which influences *CDKN2A*, a crucial gene in regulating the cell cycle (Figures 4D, S4A, and S4B). *CDKN2A* losses/deletions were noted in 65% of LAR cases versus in 36% of other subtypes ($p < 0.001$). Other cell-cycle-related genes residing at GISTIC peaks in our cohort were also checked. *RB1* losses/deletions (28% in LAR versus 58% in the other subtypes), and *CCND1* and *E2F3* gains/amplifications, were infrequent in LAR tumors (29% versus 46% and 37% versus 57%, respectively, in the other subtypes). Expression analyses also showed decreased mRNA expression levels for *CDKN2A* and *E2F3* in LAR tumors (Figure 4D). These data suggested potential sensitivity of LAR TNBCs to CDK 4/6 inhibitors or other cell-cycle inhibitors.

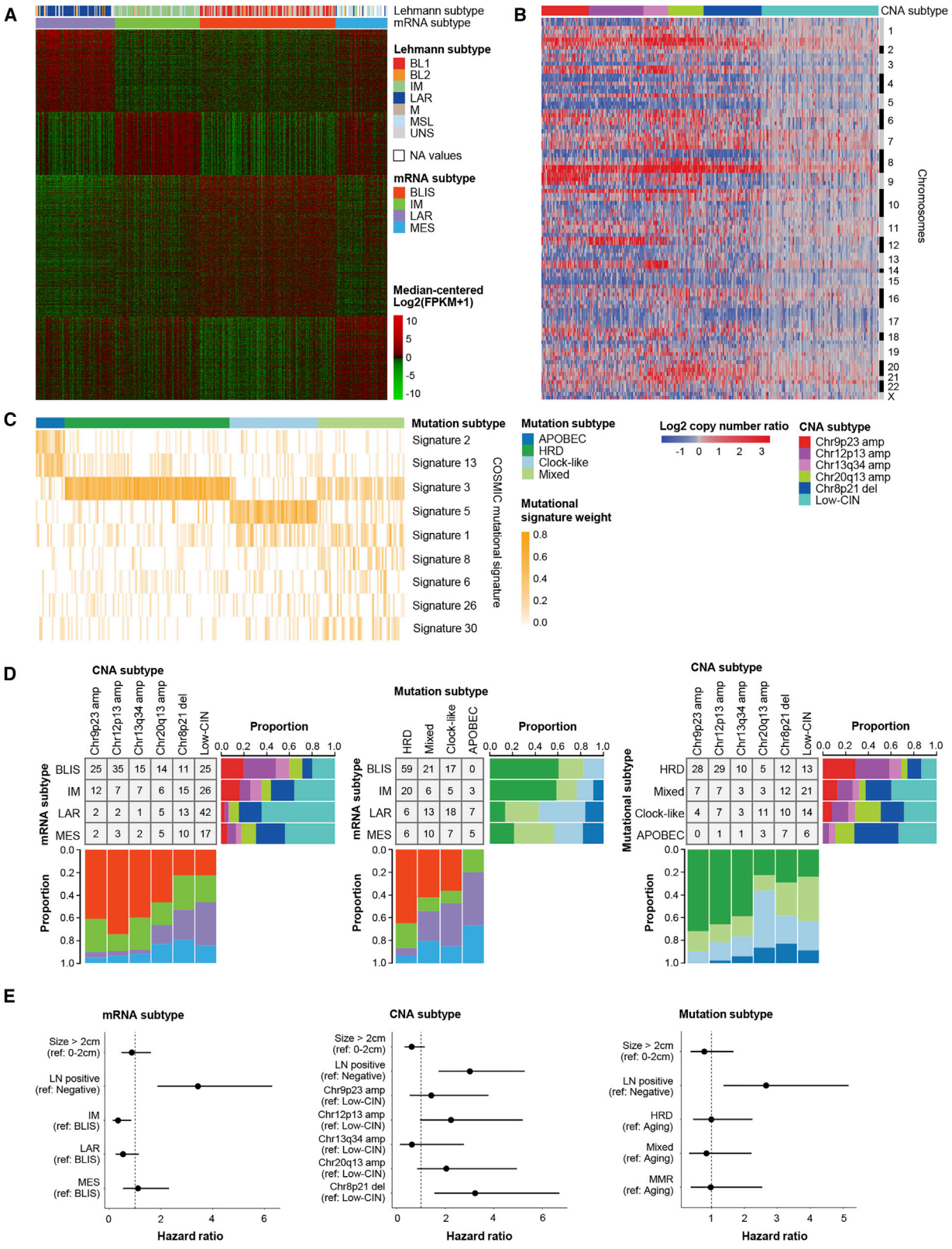
Potential Application of an Immune Checkpoint Blockade in the IM Subtype

Despite the relatively favorable outcome of the IM subtype, 6% of patients in this group experienced recurrence and/or metastases within 5 years after surgery. This subtype was characterized by elevated immune cell signaling observed in the gene expression data. H&E staining confirmed the high prevalence of both stromal and intratumoral tumor-infiltrating lymphocytes (TILs) (Figure 5A). Although the mutational load was not significantly higher in the IM subtype than in other subtypes (Figures S4C–S4G), gene set enrichment analysis (GSEA) demonstrated the activation of the adaptive immune system and interferon γ-related pathways between IM versus other subtypes (Figure 5B; Table S7). In addition, a combined analysis with CIBERSORT (Newman et al., 2015) and differential expression profiling revealed that the IM subtype was enriched for both immune-activated cells and immuno-stimulators (Angelova et al., 2015) (Figure 5C).

Both clinical and omic features have confirmed that immune recognition is activated in the IM subtype, and, thus, the mechanisms by which these tumors achieved immune escape likely involves the recruitment of immune suppressive cells or the activation of immune checkpoint molecules. While the number of immune suppressive cells was not elevated in the IM subtype, expression profiling demonstrated that immuno-inhibitors (Angelova et al., 2015), particularly IDO1, were significantly overexpressed in this subtype, providing additional rationale for the use of an immune checkpoint blockade as a therapeutic approach (Figure 5D).

Further Classification of the Basal-like Immune Suppressive Subtype Based on the HRD Score

BLIS indicates a group of TNBCs with a worse prognosis and lacks immune activation, hence is not likely to benefit from immune checkpoint inhibitors. Thus, other treatment strategies should be adopted in this group of patients. As shown in Figure 3D, BLIS constituted 65% of the mutation HRD subtype, inferring that biomarkers representing HRD might provide further insights into this group of TNBCs. The HRD score proposed by Timms et al. is a copy-number-based biomarker facilitating the identification of patients (including but not restricted to germline *BRCA1* or *BRCA2* mutation carriers) who might benefit from DNA-damaging agents (Telli et al., 2016; Timms et al., 2014). As demonstrated in Figure 6A, while in other subtypes germline *BRCA1* or *BRCA2* mutation carriers (Lang et al., 2017) had significantly higher HRD scores ($p = 0.039$), BLIS patients generally had higher HRD scores (compared with patients of other subtypes, $p = 0.005$) irrespective of germline *BRCA1* or *BRCA2*



(legend on next page)

Table 1. Highlights of Genomic, Clinical and Potential Treatment Strategies for TNBC Subtypes

Subtype	BLIS	IM	LAR	MES
Clinical	poor prognosis (5-year RFS, 84%)	good prognosis (5-year RFS, 94%)	high prevalence in Asians; poor prognosis (5-year RFS, 88%); elderly patients; related to apocrine differentiation; androgen receptor-positive	poor prognosis (5-year RFS, 79%)
Mutation	*TP53* (77%); no other frequent mutation; enrichment of the HRD mutation signature	*TP53* (81%) no other frequent mutation; enrichment of the HRD mutation signature	*TP53* (61%) PI3K-AKT pathway (~70%) *ERBB2* (9%)	mutation profile between LAR and the other two groups
Copy number	high chromosomal instability; frequent 9p23 and 12p13 amplification	relatively high chromosomal instability	low chromosomal instability; *CDKN2A/B* loss (*RB1* neutral)	copy-number profile between LAR and the other two groups
Treatment	low HRD score: escalated chemotherapy, intensive monitoring; high HRD score: platinum drugs	immune checkpoint inhibitors	endocrine therapy; targeting ERBB2 and CDK4/6 inhibitors	targeting CSCs; STAT3 inhibitor

BLIS, basal-like immune suppressed; CSCs, cancer stem cells; HRD, homologous recombination deficiency; IM, immunomodulatory; LAR, luminal androgen receptor; MES, mesenchymal; RFS, relapse-free survival.

status. A median HRD score of 41.64 was adopted to further classify the TNBCs in the BLIS subtype into high-HRD BLIS and low-HRD BLIS subgroups. Patients having TNBCs in the low-HRD BLIS subgroup had a worse prognosis than those having TNBCs in the high-HRD subgroup (5-year RFS of 73% and 95%, respectively, $p = 0.002$; Figure 6B). Interestingly, while only one of three patients having high-HRD BLIS TNBC who experienced recurrence received subsequent treatment in our center, her liver and brain metastases were highly sensitive to both platinum-based chemotherapy treatment and radiotherapy, and the patient obtained a complete remission (Figure 6C).

Furthermore, high-HRD BLIS TNBCs and low-HRD BLIS TNBCs had different genomic characteristics (Figure 6D). The high-HRD BLIS TNBCs had a higher proportion of Chr9p23 amp and Chr13q34 amp subtypes (distribution of CNA subtypes, $p < 0.001$). In contrast, the low-HRD BLIS TNBCs were more likely to exhibit whole-genome doubling (defined as allele-specific copy-number analysis of tumors [ASCAT] [Van Loo et al., 2010] ploidy >2.7; 51% in low HRD versus 27% in high HRD, $p = 0.011$). Thus, incorporating HRD into the expression-based BLIS subtype suggests that those with high HRD may derive substantial benefit from DNA-damaging therapies, such as platinum-based chemotherapy, whereas low-HRD BLIS patients have no specific treatment available and should be considered for clinical trials upon recurrence.

JAK/STAT3 Signaling Pathway Upregulation in the MES Subtype

The MES subtype harbored driver genomic events, such as *E2F3* gain (59% in BLIS, 59% in IM, 37% in LAR, and 44% in MES; Figure 1C) and *PIK3CA* mutation (3% in BLIS, 11% in IM, 50% in LAR and 22% in MES; Figure 1B) at a frequency between LAR and the other two subtypes. Investigation of the CNA/mutation subtypes also demonstrated that the MES subtype had an intermediate genomic profile and cannot be discriminated from the other subtypes using a specific genomic feature (Figure 3D).

Gene expression profiling revealed that the MES subtype displayed characteristics of breast cancer stem cells (CSCs) (Figure S5A). To further understand the alterations in this subtype, the JAK/STAT3 signaling pathway, which plays a crucial role in maintenance of breast CSCs, was investigated (Balko et al., 2016; Yu et al., 2014). The results showed that this subtype exhibited a higher expression of *JAK1* and *IL6*, which are important drivers of JAK/STAT3 activation (Figure S5B). The activated or

Figure 3. Subtyping Chinese TNBCs with Multi-omics Data

(A) mRNA-based clustering results. Heatmap with expression characteristics of the four mRNA subtypes. The top 2,000 most differentially expressed genes used for clustering are plotted. Samples are also annotated on top by the subtypes defined by Lehmann and colleagues (Lehmann et al., 2011; Lehmann and Pietenpol, 2015).

(B) Copy-number-based clustering based on GISTIC peaks. Heatmap with log2 copy-number ratio values across the genome.

(C) Mutational signature-based clustering results. Heatmap with the contribution of mutational signatures to the four mutation subtypes. A total of 230 patients with at least 30 single-nucleotide variants were enrolled in this analysis.

(D) Relationships between mRNA subtypes and CNA subtypes (left), mRNA subtypes and mutation subtypes (middle), and CNA subtypes and mutation subtypes (right).

(E) Association of the mRNA subtypes (left), CNA subtypes (middle), and mutation subtypes (right) with relapse-free survival. Cox regression analyses adjusting for lymph node status and tumor size were used. The hazard ratios are shown with 95% confidence intervals.

See also Figures S2 and S3 and Tables S7 and S8.

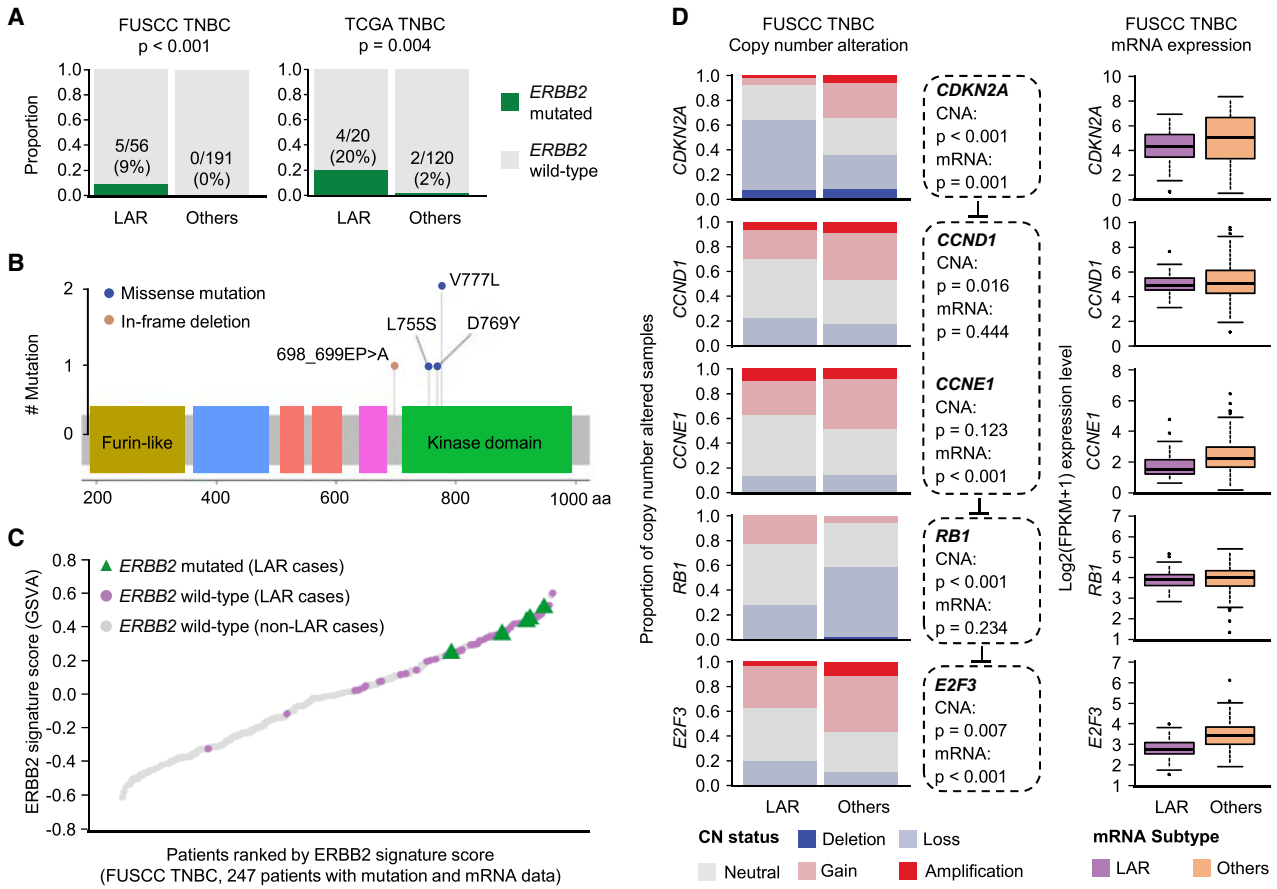


Figure 4. Activated ERBB2 and Cell-Cycle Signaling in the LAR Subtype

(A) Distribution of *ERBB2* nonsynonymous mutations among TNBC mRNA subtypes in FUSCC (left) and TCGA (right) cohorts.

(B) *ERBB2* mutations discovered in Chinese TNBC samples. Mutations were labeled in a diagram of the *ERBB2* coding region, the heights of the "lollipop" sticks indicate the number of the indicated mutation.

(C) *ERBB2* signature score (gene set variation analysis) in 247 TNBCs for whom both mutation and mRNA data were available. The samples are annotated by *ERBB2* mutation status and mRNA subtype (LAR or not).

(D) Cell-cycle-related genes at GISTIC peaks compared between LAR TNBCs and the other TNBCs. Copy-number alterations are presented on the left and log2(FPKM +1) expression values are shown on the right. For the boxplots: line in the box indicates the median; boxes correspond with the first and third quartiles; whiskers extend 1.5 times the interquartile range; outlier data are shown as dots.

See also Figure S4 and Table S9.

tyrosine-phosphorylated STAT3 (pSTAT3) gene signature score (Sonnenblick et al., 2015) was higher in the MES subtype than in other subtypes (Figures S5C and S5D). These results collectively highlight the upregulation of the JAK/STAT3 signaling pathway in the MES subtype. Despite recent evidence of the limited activity of the JAK1/2 inhibitor ruxolitinib in metastatic TNBC (Stover et al., 2018), these data suggest that the identification of MES subtype may be beneficial in a population with JAK-STAT activation to target with STAT3 inhibitors as a potential treatment strategy (Table 1) (Li et al., 2015).

DISCUSSION

While primary TNBC continues to be typically treated as a single disease, genomic findings suggest marked heterogeneity, and our data reinforce the distinct mutational, copy number, and transcriptional features of TNBC subtypes.

This cohort of Chinese patients with TNBC displayed a mutation spectrum similar to that reported in other studies. The *PIK3CA* mutation rate was higher in our cohort than that in the TCGA dataset, particularly when compared with African American patients. This result is consistent with one of our recent studies (Chen et al., 2018) and the observed difference in mRNA subtype given that LAR TNBCs are more likely to be non-basal (69%) and have *PIK3CA* mutations (Cancer Genome Atlas Network, 2012). Furthermore, Chinese TNBC displayed a higher frequency of copy-number gains in chromosome 22q11.

Genomic-based analysis revealed unique genomic alterations in receptor tyrosine kinase and cell-cycle pathways in the LAR subgroup. The LAR subtype was enriched with *ERBB2* mutations, inferring ERBB2 activation and potential resistance to trastuzumab and lapatinib. Tumor cells harboring these mutations can be inhibited by neratinib, an irreversible tyrosine kinase inhibitor (Bose et al., 2013; Zuo et al., 2016). Because of the

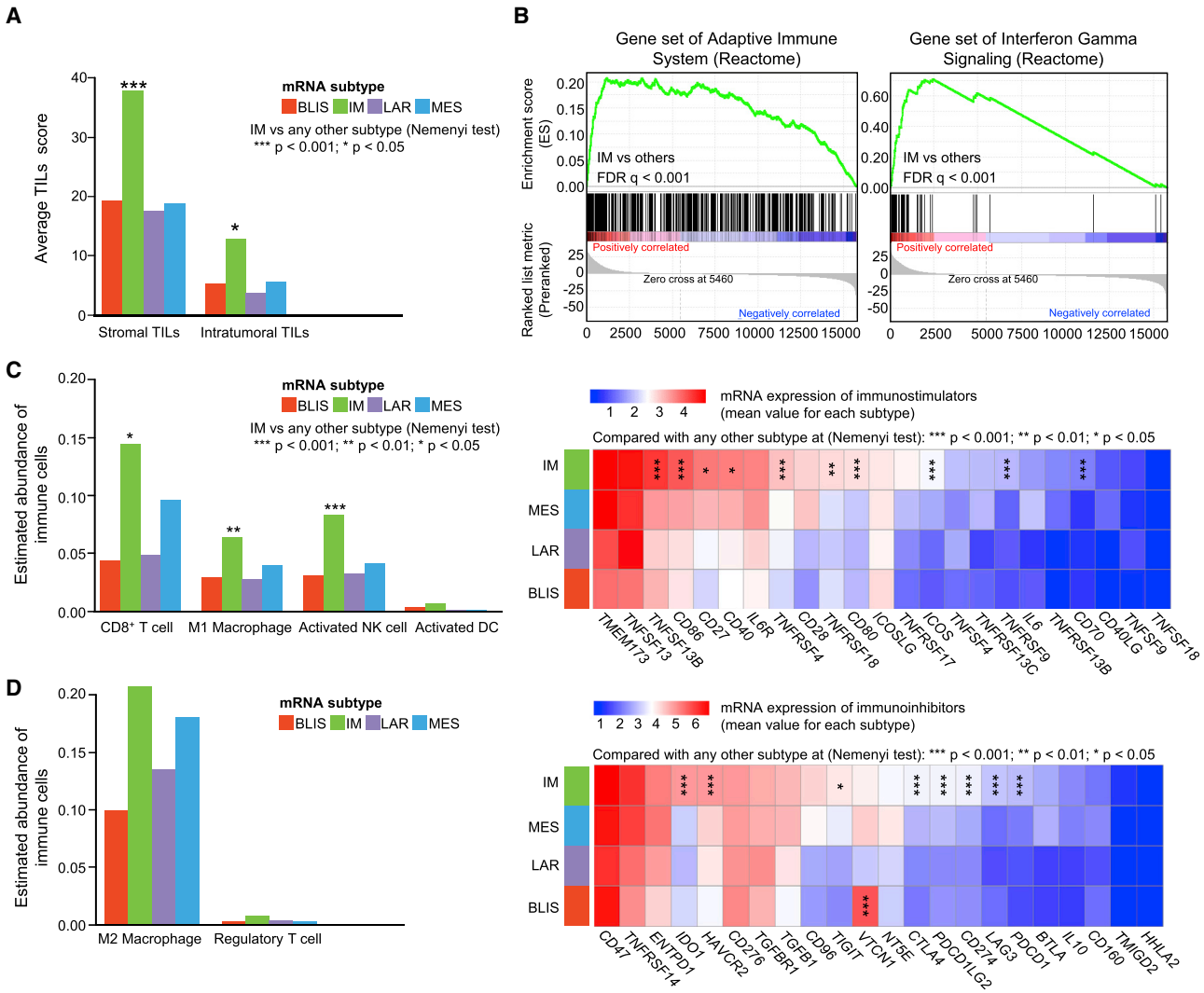


Figure 5. Potential Application of Immune Checkpoint Blockade in the IM Subtype

(A) Average stromal and intratumoral tumor-infiltrating lymphocyte (TIL) score in the four mRNA subtypes.

(B) Representative gene set enrichment analysis plot showing upregulated adaptive immune system (left) and interferon γ-related pathway (right) in the IM subtype versus the other subtypes.

(C) Relative number of immunostimulatory cells in the four mRNA subtypes (calculated using the CIBERSORT algorithm; left) and the expression of immunostimulatory molecules in four mRNA subtypes (right).

(D) Relative number of immunosuppressive cells (calculated using the CIBERSORT algorithm; left) and the expression of immune checkpoint genes (right) in four mRNA subtypes. When the Kruskal-Wallis test reached significance ($p < 0.05$), the Nemenyi test was conducted and annotated as follows (compared with any other subtype): ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

See also Figure S4.

relatively low mutation rate in unselected primary breast cancer (Bose et al., 2013; Zuo et al., 2016), the discovery of high *ERBB2* mutation rates can guide effective screening in the LAR TNBC subtype and subsequent prospective clinical trials using *ERBB2* gene-sequencing results. In addition, approximately 70% of LAR tumors exhibited somatic mutations in the phosphatidylinositol 3-kinase (PI3K) signaling pathway, suggesting a potential benefit of PI3K and AKT inhibitors, as previously reported in preclinical models (Asghar et al., 2017). LAR tumors also retained *RB1* and showed frequent *CDKN2A* alterations. Because *RB1* and *CDKN2A* have been associated with responses to CDK4/6 inhibitors (O'Leary et al., 2016), patients diagnosed with LAR tumors may be potential candidates for treatment with CDK4/6 inhibitors or other cell-cycle inhibitors.

Elevated immune cell signaling and TILs were hallmarks of the IM subtype. High mRNA expression levels of immune checkpoint inhibitor genes such as PD1, PDL1, CTLA4, and IDO1, were observed. The high expression of immune-related signatures suggests that patients with these types of tumors might potentially benefit from immune checkpoint inhibitors.

The BLIS subtype was characterized by high genomic instability, which suggests that these tumors may be sensitive to

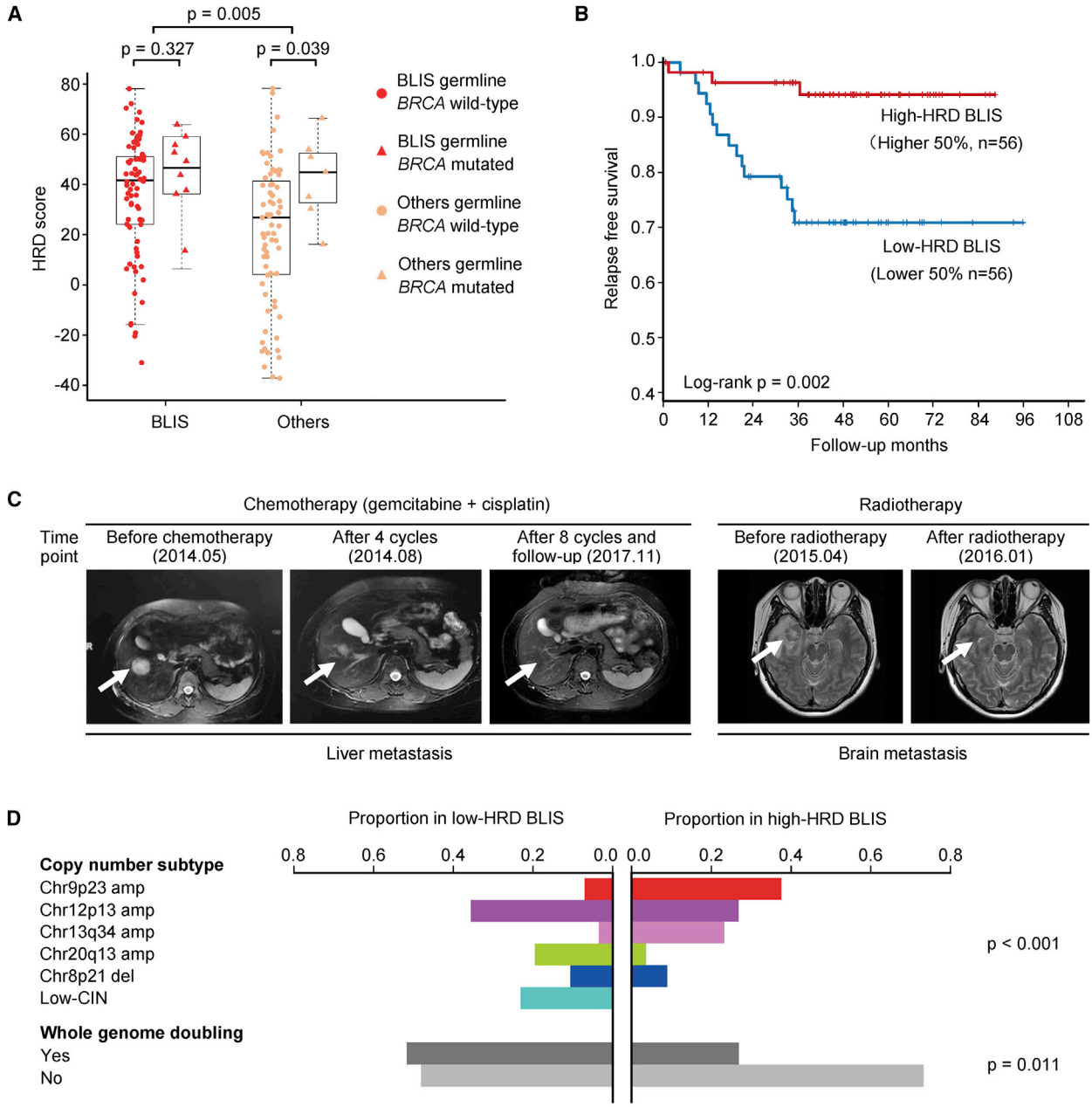


Figure 6. Further Classification of the BLIS Subtype According to HRD Score

(A) Distribution of allele-specific copy-number profile-based HRD scores in BLIS and the other subtypes according to germline *BRCA1/2* mutation status (Lang et al., 2017). The p values were calculated using the Mann-Whitney Wilcoxon test. For the boxplots: line in the box indicates the median; boxes correspond with the first and third quartiles; whiskers extend 1.5 times the interquartile range; dots/triangles show all data values.

(B) RFS of patients with high-HRD BLIS versus low-HRD BLIS tumors. The p value was based on the log rank test.

(C) Example of one patient with high-HRD BLIS, *BRCA* wild-type TNBC (FUSCCTNBC012, HRD score = 68). This patient received platinum-based treatment (liver metastasis, left) and radiotherapy (brain metastases, right) after the diagnosis of metastatic disease.

(D) Overview of the CNA subtypes and characteristics that are significantly different between low-HRD and high-HRD BLIS tumors. WGD, whole-genome doubling, defined as ASCAT ploidy higher than 2.7. The p values were calculated using Fisher's exact test (for CNA subtypes) and Pearson's chi-square test (for WGD).

PARP inhibitors and other DNA-damaging agents. The BLIS subtype could possibly be sorted into two subgroups based on HRD scores, which could affect clinical practice. TNBC patients with high HRD scores tended to have a favorable prognosis and may benefit from DNA-damaging chemotherapy or DNA repair inhibitors. Conversely, low-HRD patients had a poorer prognosis, and the optimal treatment for these patients warrants treatment in clinical trials. Better defining the classification of

BLIS TNBC in terms of low- or high-HRD status might help improve the current understanding, detection, and follow-up of the disease and lead to informed and targeted treatment selection.

Despite the lack of distinctive genomic alterations, tumors in the MES subtype displayed an overexpression of stem cell-related genes. The STAT3 signaling pathway was enriched in this subtype, showing a high level of the pSTAT3 pathway signature score. These data imply that targeting the STAT3 pathway may be an option specifically for the MES subtype.

We also classified TNBC according to mutational signatures and studied the relationships between these genomic ''footprints'' and mRNA-based TNBC subtypes. While the clock-like and mixed mutation subtypes consisted of rather similar proportions of each mRNA subtypes, the majority of the HRD subtype (65%) were classified as BLIS. This suggests that BLIS patients should be preferentially considered for clinical trials targeting DNA repair. By classifying TNBC using CNA data, we proposed that although high chromosomal instability has been accepted as a hallmark of TNBC, distinct copy-number signatures can be extracted. Notably, we observed certain CNA subtypes dominating the low-HRD BLIS, a group of diseases with dismal outcome for which no specific treatment strategy is available. Subsequent experimental studies focusing on these CNA peaks (Chr12p13, Chr20q13 and Chr8p21) might help with revealing essential molecules and potential targets.

This study is a large single-center study concerning multi-omics profiling of TNBC, demonstrating the genomic and transcriptomic landscape of Chinese TNBC patients. We studied the unique molecular features of Chinese TNBCs, subtyped TNBC by both CNA and mutational signature and interpreted how the established TNBC mRNA subtypes interact with these genomic signatures. While representing an important validation of previous molecular profiling studies on TNBC, our study highlights some potential targets and biomarkers within specific subgroups that might have been missed by smaller studies. For example, several interesting points were noted after exploring mRNA subtype-based treatment strategies, such as the enriched *ERBB2* mutation in the LAR subtype and the clinical value of the HRD score in the BLIS subtype.

Further studies could be planned. First, as we attempt to build the molecular landscape of Chinese TNBC, patients from different medical centers and different geographic regions in China should be included. Second, our conclusions and hypothesis were mainly based on analyses of genomic and transcriptomic data; both experimental and prospectively collected clinical evidence should be added before we translate our results into clinical practice. For instance, NCT02445391 (clinicaltrials.gov Identifier: NCT02445391) is a study of the efficacy of platinum-based chemotherapy in residual basal-like TNBC after neoadjuvant treatment. Incorporating the BLIS subtyping and HRD score evaluation in such well-designed clinical trials would prospectively validate our hypothesis, and could provide valuable information for patient stratification and selection, eventually translating our results into clinical practice.

In conclusion, increasing genomic precision by categorizing the TNBC population into corresponding subtypes would lead to improved targeted therapies that would benefit more TNBC patients (Li et al., 2015; Robson et al., 2017; Schmid et al., 2018; Sonnenblick et al., 2015). The genomic and transcriptomic profiles (somatic mutations, CNAs, and gene expression data) obtained through this large Chinese TNBC dataset helped identify unique molecular subtypes and demonstrated a substantial biological heterogeneity, which is characteristic of TNBCs. This dataset enabled the identification of several candidate targets and suggested hitherto unrecognized therapeutic possibilities. Moreover, this study presents a large collection of comprehensively profiled TNBC tumors to date and could serve as an important reference for further exploration of the biology of TNBC.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Biospecimen and Pathological Data Collection
  - RNA Sequencing
  - Copy Number Alterations (CNVs)
  - DNA Sequencing
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes five figures and nine tables and can be found with this article online at https://doi.org/10.1016/j.ccell.2019.02.001.

ACKNOWLEDGMENTS

This manuscript was written on behalf of AME Breast Cancer Collaborative Group. We are grateful to the patients and their families who contributed to this study. This study was funded by grants from the National Natural Science Foundation of China (81874112, 81572583, 81372848, 81370075, 31720103909, 31671368, and 31671380), the Training Plan of Excellent Talents in Shanghai Municipality Health System (2017YQ038), the ''Chen Guang'' project supported by Shanghai Municipal Education Commission and Shanghai Education Development Foundation (17CG01), Shanghai Pujiang Program (18PJD007), Shanghai Leading Talent Training Program, the Training Plan of Excellent Talents of FUSCC (YJYQ201602), the Municipal Project for Developing Emerging and Frontier Technology in Shanghai Hospitals (SHDC12010116), the Cooperation Project of Conquering Major Diseases in Shanghai Municipality Health System (2013ZYJB0302), the Innovation Team of Ministry of Education (IRT1223), the Shanghai Key Laboratory of Breast Cancer (12DZ2260100), the National Key R&D Project of China (2016YFC0901704, 2017YFC0907502, 2017YFC0907503, 2017YFC0907000, and 2017YFF0204600), Shandong Province Key R&D Project (2017CXGC1209), and the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01, 18JC1420100, and 15411953300). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. We also thank Dr. Stephanie Fortier, Dr. Charles Perou, and Dr. Mien-Chie Hung for editing the manuscript.

AUTHOR CONTRIBUTIONS

Z.-M.S., W.H., L.S., and P.W. outlined the manuscript. Y.-Z.J., D.M., C.S., J.S., M.X., X.H., Y.X., K.-D.Y., Y.-R.L., Y.Y., Y.Z., X.L., J.Z., W.H., B.L., X.L., Y.-X.Z., L.R., D.B., B.L., J.Y., W.H., S.Z., Y.G., Y.-X.R., C.Z., Z.N., Z.-G.C., F.B.,

D.-Q.L., P.W., L.S., W.H., and Z.-M.S. contributed to the literature search, data collection, and data analysis. Y.-Z.J. and D.M. provided the figures and drafted the manuscript, with additional input from all authors. D.G.S., C.V., V.K., A.D., J.R.B., and K.T. helped with data interpretation and manuscript editing. All authors approved the final manuscript. Y.-Z.J., D.M., C.S., J.S., M.X., and X.H. contributed equally to this work.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 19, 2018
Revised: January 16, 2019
Accepted: February 4, 2019
Published: March 7, 2019

REFERENCES

Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., and Stratton, M.R. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. *47*, 1402–1407.

An, O., Dall'Olio, G.M., Mourikis, T.P., and Ciccarelli, F.D. (2016). NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. Nucleic Acids Res. *44*, D992–D999.

Angelova, M., Charoentong, P., Hackl, H., Fischer, M.L., Snajder, R., Krogsdam, A.M., Waldner, M.J., Bindea, G., Mlecnik, B., Galon, J., and Trajanoski, Z. (2015). Characterization of the immunophenotypes and antigenomes of colorectal cancers reveals distinct tumor escape mechanisms and novel targets for immunotherapy. Genome Biol. *16*, 64.

Asghar, U.S., Barr, A.R., Cutts, R., Beaney, M., Babina, I., Sampath, D., Giltnane, J., Lacap, J.A., Crocker, L., Young, A., et al. (2017). Single-cell dynamics determines response to CDK4/6 inhibition in triple-negative breast cancer. Clin. Cancer Res. *23*, 5561–5572.

Balko, J.M., Schwarz, L.J., Luo, N., Estrada, M.V., Giltnane, J.M., Davila-Gonzalez, D., Wang, K., Sanchez, V., Dean, P.T., Combs, S.E., et al. (2016). Triple-negative breast cancers with amplification of JAK2 at the 9p24 locus demonstrate JAK2-specific dependence. Sci. Transl. Med. *8*, 334ra353.

Bareche, Y., Venet, D., Ignatiadis, M., Aftimos, P., Piccart, M., Rothe, F., and Sotiriou, C. (2018). Unravelling triple-negative breast cancer molecular heterogeneity using an integrative multiomic analysis. Ann. Oncol. *29*, 895–902.

Baselga, J., Gomez, P., Greil, R., Braga, S., Climent, M.A., Wardley, A.M., Kaufman, B., Stemmer, S.M., Pego, A., Chan, A., et al. (2013). Randomized phase II study of the anti-epidermal growth factor receptor monoclonal antibody cetuximab with cisplatin versus cisplatin alone in patients with metastatic triple-negative breast cancer. J. Clin. Oncol. *31*, 2586–2592.

Bianchini, G., Balko, J.M., Mayer, I.A., Sanders, M.E., and Gianni, L. (2016). Triple-negative breast cancer: challenges and opportunities of a heterogeneous disease. Nat. Rev. Clin. Oncol. *13*, 674–690.

Bose, R., Kavuri, S.M., Searleman, A.C., Shen, W., Shen, D., Koboldt, D.C., Monsey, J., Goel, N., Aronson, A.B., Li, S., et al. (2013). Activating HER2 mutations in HER2 gene amplification negative breast cancer. Cancer Discov. *3*, 224–237.

Burstein, M.D., Tsimelzon, A., Poage, G.M., Covington, K.R., Contreras, A., Fuqua, S.A., Savage, M.I., Osborne, C.K., Hilsenbeck, S.G., Chang, J.C., et al. (2015). Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer. Clin. Cancer Res. *21*, 1688–1698.

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70.

Carey, L., Winer, E., Viale, G., Cameron, D., and Gianni, L. (2010). Triple-negative breast cancer: disease entity or title of convenience? Nat. Rev. Clin. Oncol. *7*, 683–692.

Chen, L., Yang, L., Yao, L., Kuang, X.Y., Zuo, W.J., Li, S., Qiao, F., Liu, Y.R., Cao, Z.G., Zhou, S.L., et al. (2018). Characterization of PIK3CA and PIK3R1 somatic mutations in Chinese breast cancer patients. Nat. Commun. *9*, 1357.

Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature *486*, 346–352.

Dent, R., Trudeau, M., Pritchard, K.I., Hanna, W.M., Kahn, H.K., Sawka, C.A., Lickley, L.A., Rawlinson, E., Sun, P., and Narod, S.A. (2007). Triple-negative breast cancer: clinical features and patterns of recurrence. Clin. Cancer Res. *13*, 4429–4434.

Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., Delorenzi, M., Piccart, M., and Sotiriou, C. (2008). Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin. Cancer Res. *14*, 5158–5165.

Desmedt, C., Zoppoli, G., Gundem, G., Pruneri, G., Larsimont, D., Fornili, M., Fumagalli, D., Brown, D., Rothe, F., Vincent, D., et al. (2016). Genomic characterization of primary invasive lobular breast cancer. J. Clin. Oncol. *34*, 1872–1881.

Dignam, J.J., Dukic, V., Anderson, S.J., Mamounas, E.P., Wickerham, D.L., and Wolmark, N. (2009). Hazard of recurrence and adjuvant treatment effects over time in lymph node-negative breast cancer. Breast Cancer Res. Treat. *116*, 595–602.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat. Rev. Cancer *4*, 177–183.

Gendoo, D.M., Ratanasirigulchai, N., Schroder, M.S., Pare, L., Parker, J.S., Prat, A., and Haibe-Kains, B. (2016). Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics *32*, 1097–1099.

Hammond, M.E., Hayes, D.F., Dowsett, M., Allred, D.C., Hagerty, K.L., Badve, S., Fitzgibbons, P.L., Francis, G., Goldstein, N.S., Hayes, M., et al. (2010). 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. *28*, 2784–2795.

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

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576.

Lang, G.T., Shi, J.X., Hu, X., Zhang, C.H., Shan, L., Song, C.G., Zhuang, Z.G., Cao, A.Y., Ling, H., Yu, K.D., et al. (2017). The spectrum of BRCA mutations and characteristics of BRCA-associated breast cancers in China: screening of 2,991 patients and 1,043 controls by next-generation sequencing. Int. J. Cancer *141*, 129–142.

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics *28*, 882–883.

Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., and Pietenpol, J.A. (2011). Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J. Clin. Invest. *121*, 2750–2767.

Lehmann, B.D., Jovanovic, B., Chen, X., Estrada, M.V., Johnson, K.N., Shyr, Y., Moses, H.L., Sanders, M.E., and Pietenpol, J.A. (2016). Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection. PLoS One *11*, e0157368.

Lehmann, B.D., and Pietenpol, J.A. (2015). Clinical implications of molecular heterogeneity in triple negative breast cancer. Breast *24* (*Suppl 2*), S36–S40.

Li, Y., Rogoff, H.A., Keates, S., Gao, Y., Murikipudi, S., Mikule, K., Leggett, D., Li, W., Pardee, A.B., and Li, C.J. (2015). Suppression of cancer relapse and metastasis by inhibiting cancer stemness. Proc. Natl. Acad. Sci. U S A *112*, 1839–1844.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41.

Mo, Q., Wang, S., Seshan, V.E., Olshen, A.B., Schultz, N., Sander, C., Powers, R.S., Ladanyi, M., and Shen, R. (2013). Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc. Natl. Acad. Sci. U S A *110*, 4245–4250.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat. Methods *12*, 453–457.

Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47–54.

Nik-Zainal, S., and Morganella, S. (2017). Mutational signatures in breast cancer: the problem at the DNA level. Clin. Cancer Res. *23*, 2617–2629.

O'Leary, B., Finn, R.S., and Turner, N.C. (2016). Treating cancer with selective CDK4/6 inhibitors. Nat. Rev. Clin. Oncol. *13*, 417–430.

Robson, M., Im, S.A., Senkus, E., Xu, B., Domchek, S.M., Masuda, N., Delaloge, S., Li, W., Tung, N., Armstrong, A., et al. (2017). Olaparib for metastatic breast cancer in patients with a germline BRCA mutation. N. Engl. J. Med. *377*, 523–533.

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., and Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. *17*, 31.

Rouzier, R., Perou, C.M., Symmans, W.F., Ibrahim, N., Cristofanilli, M., Anderson, K., Hess, K.R., Stec, J., Ayers, M., Wagner, P., et al. (2005). Breast cancer molecular subtypes respond differently to preoperative chemotherapy. Clin. Cancer Res. *11*, 5678–5685.

Salgado, R., Denkert, C., Demaria, S., Sirtaine, N., Klauschen, F., Pruneri, G., Wienert, S., Van den Eynden, G., Baehner, F.L., Penault-Llorca, F., et al. (2015). The evaluation of tumor-infiltrating lymphocytes (TILs) in breast cancer: recommendations by an International TILs Working Group 2014. Ann. Oncol. *26*, 259–271.

Schmid, P., Adams, S., Rugo, H.S., Schneeweiss, A., Barrios, C.H., Iwata, H., Dieras, V., Hegg, R., Im, S.A., Shaw Wright, G., et al. (2018). Atezolizumab and Nab-paclitaxel in advanced triple-negative breast cancer. N. Engl. J. Med. *379*, 2108–2121.

Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., et al. (2012). The clonal and mutational evolution spectrum of primary triple-negative breast cancers. Nature *486*, 395–399.

Sonnenblick, A., Brohee, S., Fumagalli, D., Vincent, D., Venet, D., Ignatiadis, M., Salgado, R., Van den Eynden, G., Rothe, F., Desmedt, C., et al. (2015). Constitutive phosphorylated STAT3-associated gene signature is predictive for trastuzumab resistance in primary HER2-positive breast cancer. BMC Med. *13*, 177.

Stover, D.G., Gil Del Alcazar, C.R., Brock, J., Guo, H., Overmoyer, B., Balko, J., Xu, Q., Bardia, A., Tolaney, S.M., Gelman, R., et al. (2018). Phase II study of ruxolitinib, a selective JAK1/2 inhibitor, in patients with metastatic triple-negative breast cancer. NPJ Breast Cancer *4*, 10.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U S A *102*, 15545–15550.

Telli, M.L., Timms, K.M., Reid, J., Hennessy, B., Mills, G.B., Jensen, K.C., Szallasi, Z., Barry, W.T., Winer, E.P., Tung, N.M., et al. (2016). Homologous recombination deficiency (HRD) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. Clin. Cancer Res. *22*, 3764–3773.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. U S A *99*, 6567–6572.

Timms, K.M., Abkevich, V., Hughes, E., Neff, C., Reid, J., Morris, B., Kalva, S., Potter, J., Tran, T.V., Chen, J., et al. (2014). Association of BRCA1/2 defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. Breast Cancer Res. *16*, 475.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Van Loo, P., Nordgard, S.H., Lingjaerde, O.C., Russnes, H.G., Rye, I.H., Sun, W., Weigman, V.J., Marynen, P., Zetterberg, A., Naume, B., et al. (2010). Allele-specific copy number analysis of tumors. Proc. Natl. Acad. Sci. U S A *107*, 16910–16915.

Venkitaraman, R. (2010). Triple-negative/basal-like breast cancer: clinical, pathologic and molecular features. Expert Rev. Anticancer Ther. *10*, 199–207.

Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods *11*, 333–337.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573.

Yin, W.J., Lu, J.S., Di, G.H., Lin, Y.P., Zhou, L.H., Liu, G.Y., Wu, J., Shen, K.W., Han, Q.X., Shen, Z.Z., and Shao, Z.M. (2009). Clinicopathological features of the triple-negative tumors in Chinese breast cancer patients. Breast Cancer Res. Treat. *115*, 325–333.

Yu, H., Lee, H., Herrmann, A., Buettner, R., and Jove, R. (2014). Revisiting STAT3 signalling in cancer: new and unexpected biological functions. Nat. Rev. Cancer *14*, 736–746.

Zuo, W.J., Jiang, Y.Z., Wang, Y.J., Xu, X.E., Hu, X., Liu, G.Y., Wu, J., Di, G.H., Yu, K.D., and Shao, Z.M. (2016). Dual characteristics of novel HER2 kinase domain mutations in response to HER2-targeted therapies in human breast cancer. Clin. Cancer Res. *22*, 4859–4869.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Tumor and normal tissue samples (breast cancer patients)	This study	FUSCCTNBC
Critical Commercial Assays		
MiRNeasy mini kit	Qiagen	217004
Ribo-Zero Gold rRNA Removal Kit	Illumina	MRZG12324
SuperScript II Reverse Transcriptase	Invitrogen	18064071
TruSeq Stranded Total RNA LT Sample Prep Kit - Set A/B	Illumina	RS-122-2301/2
OncoScan CNV Assay	Affymetrix/Thermo Fisher	902695
Agilent SureSelectXT Library Prep Kit	Agilent	G9621B
Agilent SureSelect Human All Exon V6	Agilent	5190-8873
Deposited Data		
RNA sequencing data (FUSCC TNBC)	This study	NODE: OEP000155; SRA: SRP157974
Whole exome sequencing data (FUSCC TNBC)	This study	NODE: OEP000155; SRA: SRP157974
OncoScan CNV array data (FUSCC TNBC)	This study	NODE: OEP000155; GEO: GSE118527
TCGA breast cancer data	TCGA (https://portal.gdc.cancer.gov)	TCGA-BRCA
Software and Algorithms		
ConsensusClusterPlus	(Wilkerson and Hayes, 2010)	http://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html
ASCAT v2.4.3	(Van Loo et al., 2010)	
pheatmap	NA	https://www.rdocumentation.org/packages/pheatmap/versions/1.0.2
Genefu v2.4.2	(Gendoo et al., 2016)	https://www.rdocumentation.org/packages/genefu/versions/2.4.2
VarScan v2.4.2	(Koboldt et al., 2012)	http://dkoboldt.github.io/varscan/
pamr v1.55	(Tibshirani et al., 2002)	https://www.rdocumentation.org/packages/pamr/versions/1.55
Sentieon TNscope v2017.11	Sentieon	http://www.sentieon.com
Sentieon TNseq v2017.11	Sentieon	http://www.sentieon.com
ANNOVAR v2016Feb01	(Wang et al., 2010)	http://annovar.openbioinformatics.org/
GISTIC 2 v2.0.22	(Mermel et al., 2011)	http://software.broadinstitute.org/software/cprg/?q=node/31
CIBERSORT	(Newman et al., 2015)	cibersort.stanford.edu
Gene Set Enrichment Analysis (GSEA) v3.0	(Liberzon et al., 2011)	http://software.broadinstitute.org/cancer/software/genepattern/
R statistical packages version 3.4.2	R development core team	https://www.r-project.org
Tophat v2.0.8		PMID: 19289445
deconstructSigs	(Rosenthal et al., 2016)	PMID: 26899170
Cufflinks	(Trapnell et al., 2012)	https://cole-trapnell-lab.github.io/cufflinks/
iClusterPlus V1.6.0	(Mo et al., 2013)	https://www.mskcc.org/departments/epidemiology-biostatistics/biostatistics/icluster
DESeq2	(Love et al., 2014)	https://bioconductor.org/packages/release/bioc/html/DESeq2.html/
Other		
Firehose, FireBrowse	The Broad Institute	https://gdac.broadinstitute.org/http://firebrowse.org/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be facilitated by the Lead Contact, Zhi-Ming Shao (zhimingshao@yahoo.com).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Patients diagnosed with malignant breast cancer and willing to participate in the present study were retrospectively selected. A total of 504 consecutive Chinese patients who were treated at the Department of Breast Surgery at Fudan University Shanghai Cancer Center (FUSCC) from January 1, 2007 to December 31, 2014 were enrolled according to the following defined criteria: 1) female patients diagnosed with unilateral invasive ductal carcinoma with an ER-, PR- and HER2- phenotype. 2) central pathologic examination of tumor specimens performed by the Department of Pathology at FUSCC (ER, PR and HER2 status were independently confirmed by two experienced pathologists based on immunochemical analysis and *in situ* hybridization). We used <1% positively stained cells as the cutoff for ER/PR negativity in immunohistochemistry testing according to the American Society of Clinical Oncology/College of American Pathologists guideline (Hammond et al., 2010); 3) no evidence of distant metastasis at diagnosis; and 4) sufficient frozen tissue available for further investigation. Patients with breast carcinoma *in situ* or with inflammatory breast cancer were excluded. Clinicopathological characteristics included age; tumor histologic type; tumor size; lymph node status; histologic grade; adjuvant therapies; and ER, PR, HER2 and Ki67 status. Extent of disease (assessed by chest computed tomography, bone scan, abdominal ultrasound, bilateral mammography, breast ultrasound and/or magnetic resonance imaging) was recorded.

Follow-up within this cohort of patients was completed on June 30, 2017 and the median length of follow-up was 45.8 months (interquartile range, 34.4-59.8 months). Relapse-free survival (RFS) was defined as the time from diagnosis to first recurrence or a diagnosis of contralateral breast cancer. Patients without events were censored from the time point of the last follow-up. All tissue samples included in the present study were obtained after approval of the research by the FUSCC Ethics Committee, and each patient provided written informed consent.

METHOD DETAILS

Biospecimen and Pathological Data Collection

Biospecimen Collection, Quality Control and Processing

Fresh frozen tumor tissues were macrodissected to avoid the influence of stromal tissues (<30% stromal tissue). The percentage of tumor cells was confirmed to be 50% or more in all the breast cancer specimens. Total DNA was isolated from fresh frozen TNBC samples using TGuide M24 (Tiangen, Beijing, China). The purity and quantity of total DNA were estimated by measuring the absorbance at 260 nm (A260) and 280 nm (A280) using a NanoDrop 2000 spectrophotometer (Thermo Scientific, Wilmington, DE, USA). The extracted DNA was considered pure and suitable for future experiments when the A260/A280 ratio was within the range 1.6-1.9.

A MiRNeasy mini kit (Qiagen, Hilden, Germany) was used for the purification of total RNA from tissues that had been previously stored in RNA-later solution. All RNA-seq libraries were prepared according to the Ribo-Zero protocol and then sequenced on an Illumina HiSeq platform. Paired-end reads with lengths of 150 nucleotides were generated.

Evaluation of Tumor Infiltrating Lymphocytes (TILs)

TILs were evaluated on hematoxylin and eosin–stained sections in a routine diagnostic setting. Stromal TILs (sTILs) were defined as lymphocytes located in stromal components of the tumor tissue, while intratumoral TILs (iTILs) were defined as lymphocytes located in epithelial components of the tumor tissue. Each patient case was evaluated by two pathologists, based on the related guideline (Salgado et al., 2015). We evaluated the sTILs and iTILs scores separately.

RNA Sequencing

Sample Preparation and Data Generation

RNA library preparation was performed as described in the Illumina TruSeq Stranded Total RNA LT sample preparation kit with Ribo-Zero Gold (Illumina Inc., San Diego, CA, USA). Briefly, ribosomal RNA was depleted from 0.3-1 μg of total RNA. Following depletion, the mRNA was fragmented to an average insert size of 200–400 bp at 94°C for 4 min. The cleaved RNA fragments were copied into first-strand cDNA using reverse transcriptase (Invitrogen, Carlsbad, CA, USA) and random primers. The first-strand cDNA was converted into double-stranded DNA in the presence of dUTP. The incorporation of dUTP in second-strand cDNA synthesis quenches the second strand during amplification, thus improving the strand specificity of the library. These cDNA fragments were subjected to the addition of a single 'A' base and subsequent ligation of the adapter. The products were purified and enriched via PCR to generate the final library. After testing quality using an Agilent 2100 Bioanalyzer (Agilent Technologies) with the DNA chip and further quantification using a Qubit® 3.0 fluorometer (Invitrogen, Carlsbad, CA, USA), the libraries were sequenced on the Illumina HiSeq platform (Illumina Inc., San Diego, CA, USA).

Fragments Per Kilobase of exon model per Million mapped fragments (FPKM) values were obtained using the tophat-cufflinks pipeline (Trapnell et al., 2012), where the fastq files were mapped to a human reference genome (Hg19, GRCh37_snp_tran). To

choose genes with accurate expression value, we removed genes whose FPKM was 0 in more than 30% samples prior to subsequent analyses.

Expression-Based Unsupervised Clustering

We performed k-means clustering (the "kmeans" function in R) and consensus clustering (the "ConsensusClusterPlus" package in R (Wilkerson and Hayes, 2010)) to determine the optimal number of TNBC subtypes in the mRNA data. Consensus clustering was used to assess the robustness of k-means clustering (1,000 iterations, 0.8 resampling). The optimal number of clusters was determined from the cumulative density function (CDF), which plots the corresponding empirical cumulative distribution, defined over a range between 0 and 1, and from a calculation of the proportion increase in the area under the CDF curve. The number of clusters was determined when any further increase in cluster number (k) did not lead to a corresponding remarkable increase in the CDF area (Figures S2A–S2C). In addition, to determine the optimal number of genes in k-means clustering, we tested the clustering results by choosing genes based on the standard error (SD) from the top 5% to the top 30% and finally chose the SD top 2000 genes to perform k-means clustering (Figure S2D).

Breast Cancer Intrinsic Subtype

Intrinsic subtype of our samples was determined using the "genefu" package (v2.12.0) (Gendoo et al., 2016), before which we merged our data and the TCGA expression data (RSEM data were downloaded from http://gdac.broadinstitute.org/ and transformed by log2[RSEM+1]; We divided our RNA-seq data into two parts [n = 180 each] randomly to merge with the TCGA data, ensuring that 70-80% of the input samples were non-TNBC tumors in each run). Batch effects were removed using the "ComBat" function in package "sva" with TCGA TNBCs as the reference for batch adjustment (Leek et al., 2012).

Extrapolating the Clustering Results to TCGA TNBCs

We extrapolated our mRNA-based subtypes to the TCGA TNBCs using the R package "pamr" (Tibshirani et al., 2002) which used a nearest shrunken centroids method to perform sample classification. Before using the "pamr.train" and "pamr.predict" functions, the expression data from our study and the log2(RSEM+1) values of TCGA TNBCs were normalized within each data set (using the "scale" function in R with default parameters). As demonstrated in Table S6, our results were apparently correlated with Lehmann subtypes and the five subtypes defined by Bareche and colleagues (Bareche et al., 2018). Clear correspondence can be found between our subtyping strategy and the others, especially with the subtypes by Bareche and colleagues, summarized as follows: 1) LAR (FUSCC) vs. LAR (Bareche and colleagues), our LAR subtype contained all of the LAR tumors (100%) identified by Bareche and colleagues; 2) IM (FUSCC) vs. IM (Bareche and colleagues), 86.2% of Bareche's IM tumors were also identified as the IM subtype in our results; 3) BLIS (FUSCC) vs. BL1/M (Bareche and colleagues), we put the BL1 and M subtypes of Bareche's subtyping system together because both of them have been proven to be basal-like tumors lacking lymphocytic infiltration (Lehmann et al., 2016), and 79.7% of these two subtypes were identified as the BLIS subtype in our results; 4) MES (FUSCC) vs. MSL (Bareche and colleagues): 80% of Bareche's MSL tumors were identified as the MES subtype in our results.

Estimation of Immune Cell Numbers within Tumor Tissues

The "Cell-type Identification By Estimating Relative Subsets Of RNA Transcripts (CIBERSORT)" (Newman et al., 2015) tool (absolute mode) (https://cibersort.stanford.edu/) was used to calculate the abundance of 22 types of immune cell subsets in each sample. Cell subsets that were reported to have the function of killing tumor cells or promoting tumor development were extracted and compared among the FUSCC TNBC mRNA subtypes.

Gene Set Enrichment Analysis (GSEA) and Gene Set Variation Analysis (GSVA)

GSEA was performed using the GSEA software (v3.0) (Subramanian et al., 2005) and the Molecular Signature Database (v6.1) (Liberzon et al., 2011) (http://www.broad.mit.edu/gsea/) using the GSEA preranked function. Differential expression analysis outputs of DESeq2 (one subtype vs. the rest) (Love et al., 2014) were used to generate the ranked list file (.rnk file; ranked by [sign of log2FoldChange] x –log10[p value]). One thousand total permutations were used. The "gsva" function in the R package "GSVA" (Hanzelmann et al., 2013) was used to calculate the ERRB2 pathway score using a published ERRB2-related gene set (Desmedt et al., 2008, 2016).

Copy Number Alterations (CNVs)

Sample Preparation and Data Generation

Genome-wide copy number analysis was performed using an OncoScan CNV Assay Kit (Affymetrix, Santa Clara, CA, USA) according to the manufacturer's recommendations. Briefly, a total of 80 ng of DNA from each tumor sample was processed. Molecular inversion probes (MIPs) were added to the sample DNA and annealed at 58°C overnight. The annealed DNA was divided into two equal parts and incubated with AT or GC gap-fill master mixes for ligation. Then, the unincorporated, non-circularized MIPs and the remaining genomic template were removed through exonuclease treatment. The circularized MIPs were linearized with a cleavage enzyme, and the first PCR amplification was performed, followed by a second amplification. The amplified products were digested with HaeIII, and the small fragments containing the specific SNP genotype were hybridized onto arrays. The arrays were washed and stained using a GeneChip Fluidics Station 450 (Affymetrix, Santa Clara, CA, USA) and were scanned using a GeneChip Scanner 3000 7G (Affymetrix, Santa Clara, CA, USA). The fluorescence of clusters was measured to generate a DAT file. Cluster intensity values were automatically calculated using a built-in algorithm from DAT files using GeneChip Command Console software (Affymetrix, Santa Clara, CA, USA), and a CEL file was generated.

Analysis of SNP Array Data

An analysis of Affymetrix OncoScan CNV SNP probe assays was performed with OncoScan Console (v1.3) software (Affymetrix, Inc.). BioDiscovery Nexus Express™ for OncoScan 3 software was used to assess recurrent germline/potential false-positive calls by using a reference cohort of DNA from 23 randomly selected white blood cell samples from the mentioned patients (regions altered in 12 or more patients were defined as recurrent germline/potential false-positive calls for subsequent removal).

Probe-level output from the OncoScan Console was analyzed using ASCAT (v2.4.3) (Van Loo et al., 2010) to obtain segmented copy number calls, estimated tumor ploidy and estimated tumor purity results. Segments overlapping with previously described recurrent germline/potential false-positive calls were removed. The ASCAT segments were subsequently used to produce log2 ratios by dividing the total copy number (nAraw + nBraw, with zero values set to 0.05). These segments were used as the input of GISTIC2.0 (v2.0.22) (Mermel et al., 2011) and were used to study the recurrence of gene level CNVs in our sample set. GISTIC2.0 was run with the following parameters changed from the default settings (-ta 0.2 -td 0.2 -genegistic 1 -smallmem 1 -broad 1 -conf 0.95 -rx 0 –brlen 0.7 -cap 3.5 –armpeel 1).

Estimation of Homologous Recombination Deficiency (HRD) Score

As summarized in a previous study (Timms et al., 2014), HRD score was calculated as the sum of three scores using segments with integer copy number produced by ASCAT: allelic imbalance extending to the telomere (NtAI) score, loss of heterozygosity score (LOH) and large-scale state transition score (LST). Calculation of these three scores was described in the article of Timms et al. (Timms et al., 2014). Briefly, NtAI score was defined as the number of regions with allelic imbalance that are longer than 11 Mb and extend to one of the subtelomeres but do not cross the centromere. LOH score was defined as the number of LOH regions longer than 15 Mb but shorter than the whole chromosome. LST score was defined as the number of break points between regions longer than 10 Mb after filtering out regions shorter than 3 Mb. To diminish effect of ploidy, LST score was modified according to the formula: LSTm = LST – kP, where P is ploidy and k is a constant of 15.5.

CNA-Based Unsupervised Clustering

CNA-based clustering was performed using k-means clustering (‘‘kmeans’’ function in R) with consensus clustering (‘‘ConsensusClusterPlus’’ package in R) to determine the optimal number of subtypes, similar to the mRNA based clustering. Input data for each sample (n = 401) was the ‘‘actual copy change given’’ of each ‘‘peak region’’ obtained from the file ‘‘all_lesions.conf_95.txt’’, where values over 3.5 were set to 3.5 (consistent with the cap value in GISTIC 2.0). Enriched CNA in each CNA-based subtypes were checked using the Kruskal-Wallis test and Fisher’s exact test (where the threshold value for high-level amplification was manually changed to log2[5/2]) (see Table S8).

DNA Sequencing

Sample Preparation and Data Generation

Qualified genomic DNA from tissues and matched white blood cell samples was prepared for whole-exome sequencing (WES). A total of 300 ng of each DNA sample based on Qubit quantification was fragmented on a Bioruptor Plus sonication system (Diagenode, Liège, Belgium). Sheared DNA was used to perform end repair, A-tailing and adapter ligation with an Agilent SureSelectXT Library Prep Kit (Agilent Technologies, Santa Clara, CA, USA) according to the manufacturer’s protocol. Then, 750 ng of prepared DNA in a volume of 3.4 μl was captured using Agilent SureSelect Human All Exon V6 (Agilent Technologies) probes, followed by the amplification of the captured library with indexing primers. Quality control was performed using the Agilent 2100 Bioanalyzer (Agilent Technologies) with a DNA chip. After quantified with a Qubit® 3.0 fluorometer (Invitrogen, Carlsbad, CA, USA), the libraries were sequenced on an Illumina HiSeq platform (Illumina Inc., San Diego, CA, USA). For each library preparation from tissue, 12 samples were loaded in a single lane. For each library preparation from blood, 20 samples were loaded in a single lane.

Mutation Calling

Exome sequencing alignment and quality assessment: 279 matched tumor/normal pairs were present in our data set. The exome-sequenced reads were aligned using BWA-mem, and the resulting BAM files were preprocessed using version 201711 of Sentieon tools (https://www.sentieon.com/), which are drop-in improvements for the GATK Toolkit (McKenna et al., 2010). Sequencing quality statistics were obtained using FastQC and Samtools. The sequencing coverage was 100X-150X for tumor samples and 30X-50X for white blood cells.

Variant Calling. To identify all the variants in the samples, we used VarScan2 v2.4.2 (–min-coverage 3 –min-coverage-normal 3 –min-coverage-tumor 3 –min-var-freq 0.08 –p-value 0.10 –somatic-p-value 0.05 –strand-filter1 (Koboldt et al., 2012)), TNseq (sentieon driver -r –algo TNhaplotyper –dbsnp –pon –cosmic; https://www.sentieon.com/) and TNscope (sentieon driver -t –algo TNscope –dbsnp –pon –cosmic; https://www.sentieon.com/) for single nucleotide variants (SNVs) and indels. To improve specificity, a panel of normal (PON) samples filter was used to screen out expected germline variation and artifacts. This PON panel was based on 330 normal blood samples, from which a VCF file was created for the sites identified as variants by TNseq in more than one normal sample. Somatic mutation calls were annotated using ANNOVAR (Wang et al., 2010).

Specifically, processSomatic and somaticFilter (–min-coverage 10, –min-reads2 2, –min-strands2 1, –min-avg-qual 20, –p-value 0.1) was used to extract high-confidence somatic variants from the raw VarScan2 results and to remove clusters of false positives and SNV calls near indels. SNVs with VarScan2 minimum average base quality scores < 20 for variant-supporting reads were removed. Variants that failed VarScan2 internal filters due to ‘SS=0’ were removed. To identify variants exhibiting read direction bias, variants with > 90% support on one strand were filtered out from the variant calls.

To obtain the final set of mutation calls, we used a two-step approach, first removing any spurious variant calls arising as a consequence of sequencing artifacts and then making use of consensus mutations in at least two out of three callers (TNseq, TNscope and VarScan2) to identify somatic mutations. Second, additional filtering based on bam-readcount (https://github.com/genome/bam-readcount) was performed to reduce false positive calls: 1) variant allele frequency (VAF) $\geq$ 8%; 2) sequencing depth in the region $\geq$ 8; 3) sequence reads in support of the variant call $\geq$ 2. Only variants with the following functional classification were considered in this study: missense mutation, nonsense mutation, nonstop mutation, RNA mutation, silent mutation, variants at splice site or translation start site, insertion and deletion. Genes that were recorded as ''cancer genes'' in the Network of Cancer Genes (An et al., 2016) and were recorded as the ''Tier 1'' mutations in The Cancer Gene Census (Futreal et al., 2004) were considered known cancer-related genes.

Mutational Signature Analysis

As TNBC progresses, signatures characterized by specific patterns of nucleotide substitution are generated (Alexandrov et al., 2015). The package ''deconstructSigs'' (Rosenthal et al., 2016) was used to identify mutational signatures that presented in the TNBC samples (n = 230, only samples with at least 30 SNVs were included). The normalization method was set to 'exome2genome'. This approach organized sample information in the form of the fraction of mutations in each of 96 trinucleotides and determined the weighted combination of published signatures (Alexandrov et al., 2015) (https://cancer.sanger.ac.uk/cosmic/signatures) that most closely reconstructed the mutational profile. Only signatures that have been observed in human breast cancers were considered (Cosmic signature 1, 2, 3, 5, 6, 8, 13, 17, 18, 20, 26, 30) (Nik-Zainal et al., 2016; Nik-Zainal and Morganella, 2017).

Mutational Signature-based Unsupervised Clustering

Mutational signature clustering was performed using k-means clustering similar to the mRNA-based clustering and the CNA based clustering. Input data for each sample (n = 230, only samples with at least 30 SNVs were included) was the deconstructSigs determined signature ''weight'' of each signature, where only signatures appeared in at least 5% of the samples at a minimum ''weight'' of 0.2 were used (Signature 1, 2, 3, 5, 6, 8, 13, 26 and 30).

Integrative Clustering

We performed two integrated clustering approaches: iCluster (using the R package ''iClusterPlus'' (Mo et al., 2013)) and similarity network fusion (SNF; using the R package ''SNFtool'' (Wang et al., 2014)). The iCluster was performed according to the manual of iClusterPlus, where 37 frequently mutated genes (cancer-related genes that were mutated in at least 5 samples), 4,256 non-redundant copy number segments, and the top 2,000 variable gene expression values were used in this analysis. No apparent ''elbow'' was observed when k was within the range of 1-9 (Figure S3A), so we chose k = 9 (10 clusters) (Figure S3B). SNF allows for more flexible data input, and we used the exact same data (after normalization according to the manual of SNF) that we used in the single-platform clustering analyses. Optimal cluster numbers were 2 and 9, and we chose the latter to perform further analyses (Figure S3C). The results of both approaches were demonstrated in Table S1 and Figure S3.

QUANTIFICATION AND STATISTICAL ANALYSIS

Frequency tabulation and summary statistics were used to characterize the data distribution. Student's t-test, analysis of variance, the Mann-Whitney Wilcoxon test and the Kruskal-Wallis test were utilized to compare continuous variables and ordered categorical variables whilst Pearson's chi-square test and Fisher's exact test were employed for comparison of unordered categorical variables. Survival curves were constructed using the Kaplan–Meier product limit method and compared with the log-rank test. A Cox proportional hazard regression model adjusting or not adjusting for available prognostic clinical covariates was performed to calculate hazard ratios (HRs) and 95% confidence intervals. The p values were adjusted to false discovery rate (FDR) using the Benjamini-Hochberg procedure in multiple comparisons (except for the Kruskal-Wallis test for which the Nemenyi test was chosen as the post-hoc test). All analyses were performed using R packages version 3.4.2 (https://cran.r-project.org/).

DATA AND SOFTWARE AVAILABILITY

The accession number for all the data reported in this paper is NODE: OEP000155. All data can be viewed in The National Omics Data Encyclopedia (NODE) (http://www.biosino.org/node) by pasting the accession (OEP000155) into the text search box or through the URL: http://www.biosino.org/node/project/detail/OEP000155. Microarray data and sequence data have also been deposited in the NCBI Gene Expression Omnibus (OncoScan array; GEO: GSE118527) and Sequence Read Archive (WES and RNA-seq; SRA: SRP157974).