PNAS
www.pnas.org

Supplementary Information for

Landscape analysis of adjacent gene rearrangements reveals BCL2L14-ETV6 gene fusions in more aggressive triple-negative breast cancer

Sanghoon Lee[a,b,1], Yiheng Hu[a,b,c,d,e,f,1], Suet Kee Loo[a,b,c], Ying Tan[d,e,f], Rohit Bhargava[c], Michael T. Lewis[d,e,f,g], Xiao-Song Wang[a,b,c,d,e,f,2]

* These authors contributed equally to this work.

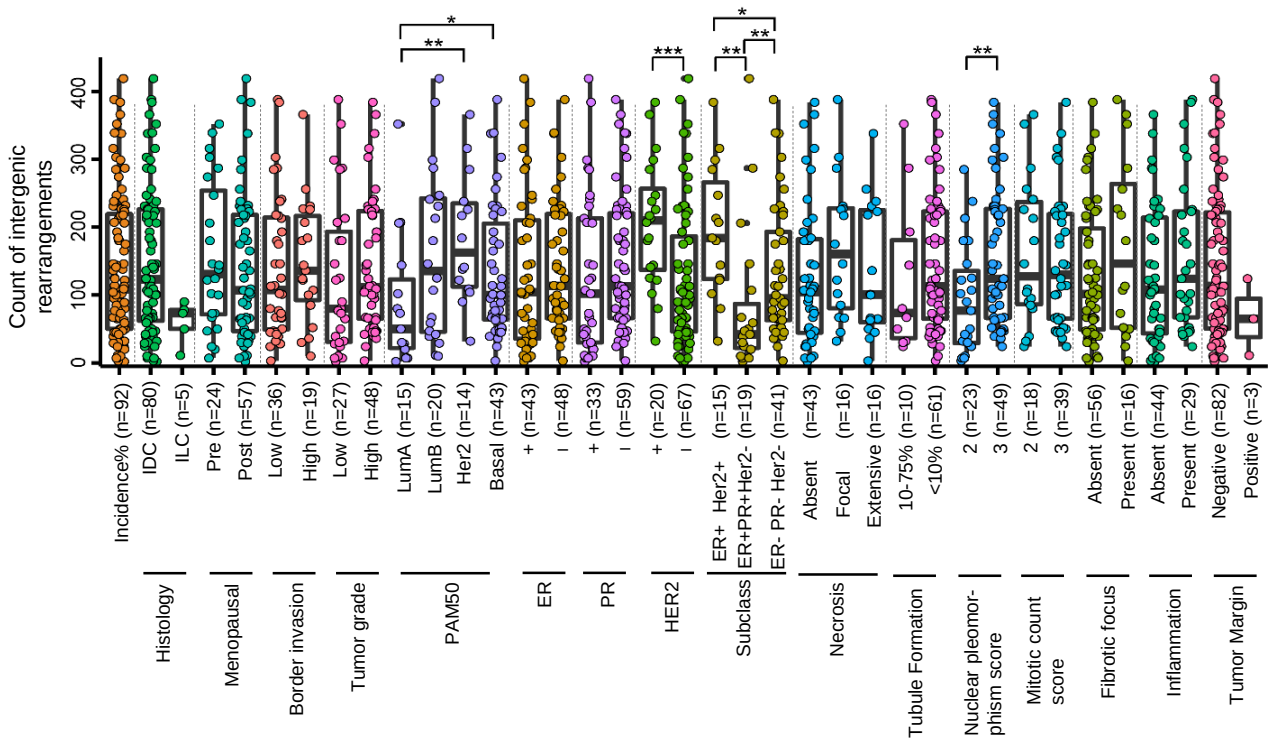
Corresponding Author:

Xiao-Song Wang, M.D., Ph.D.

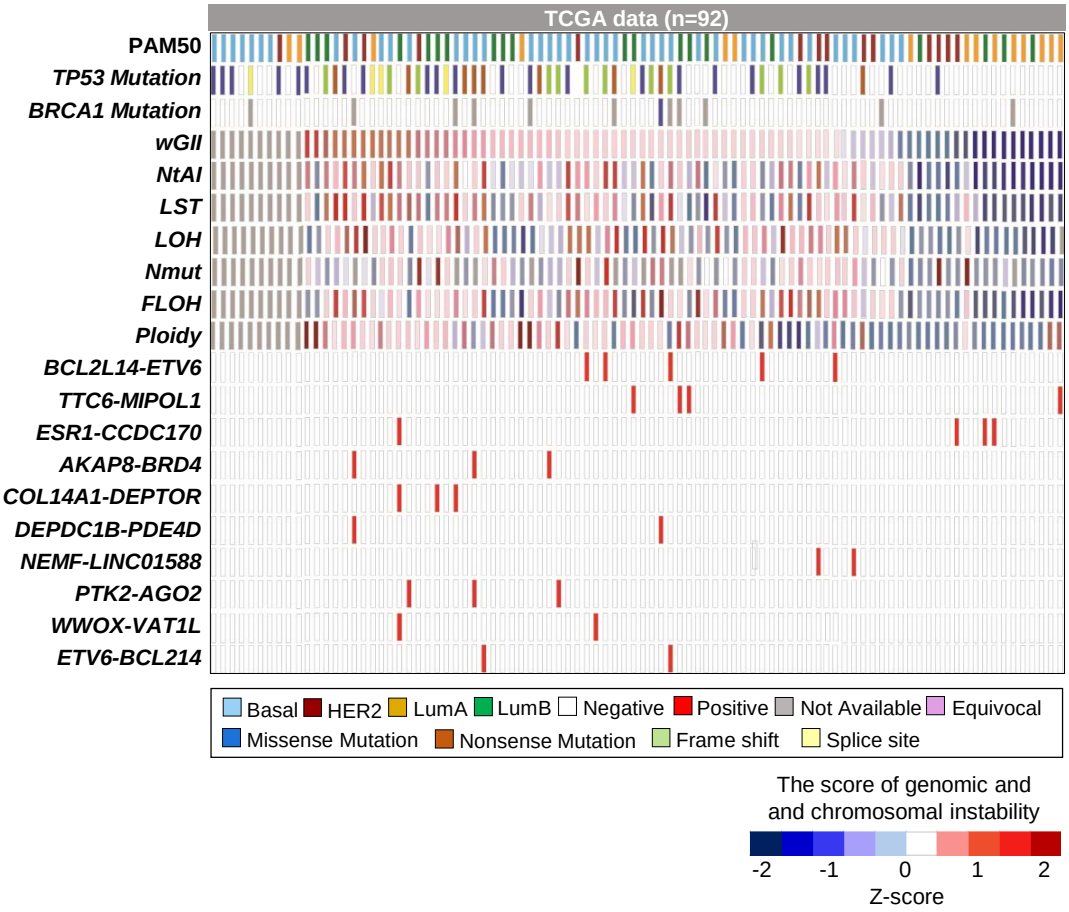
Email: xiaosongw@pitt.edu

This PDF file includes:

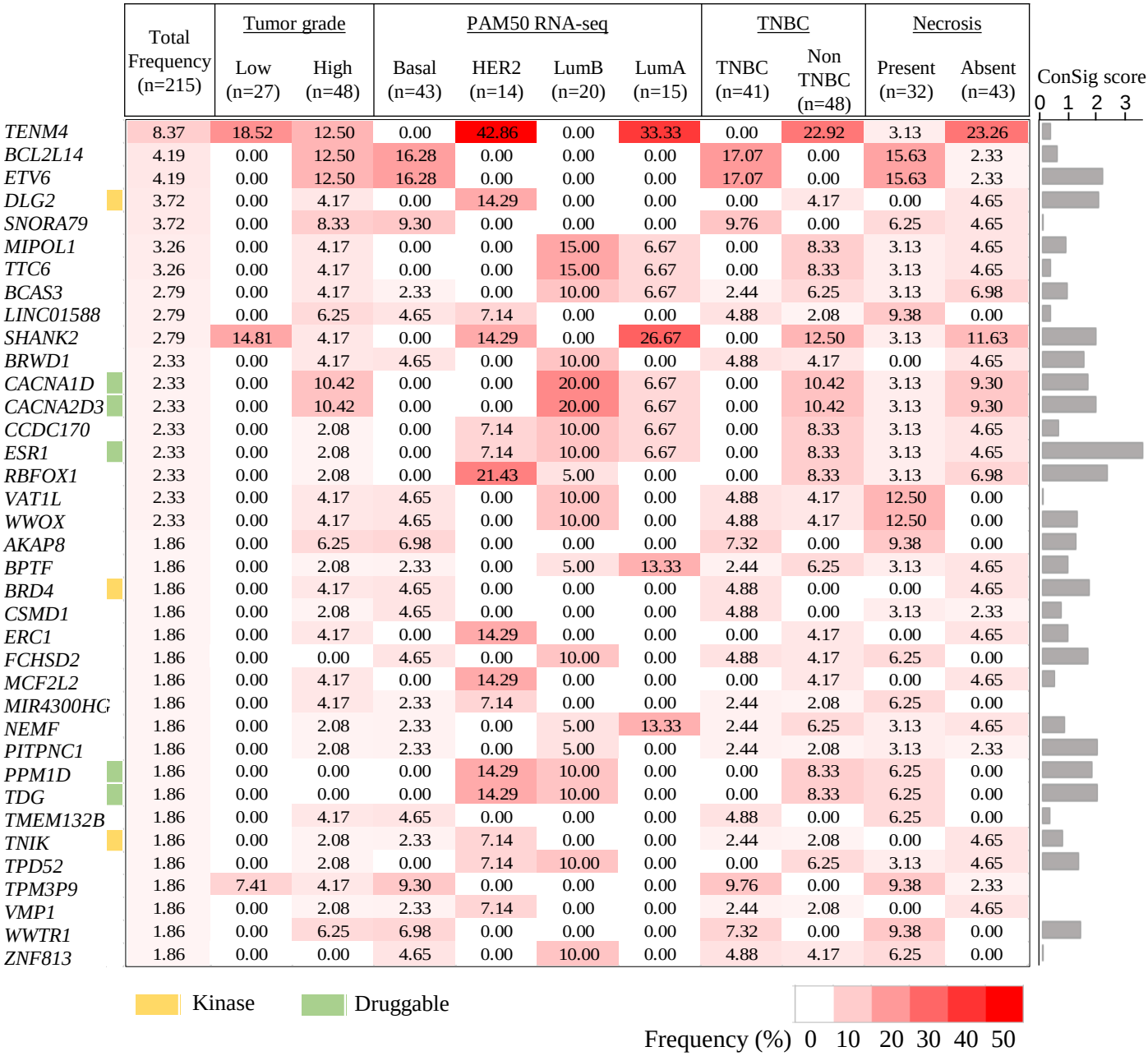
SI Appendix: Supplementary Figures



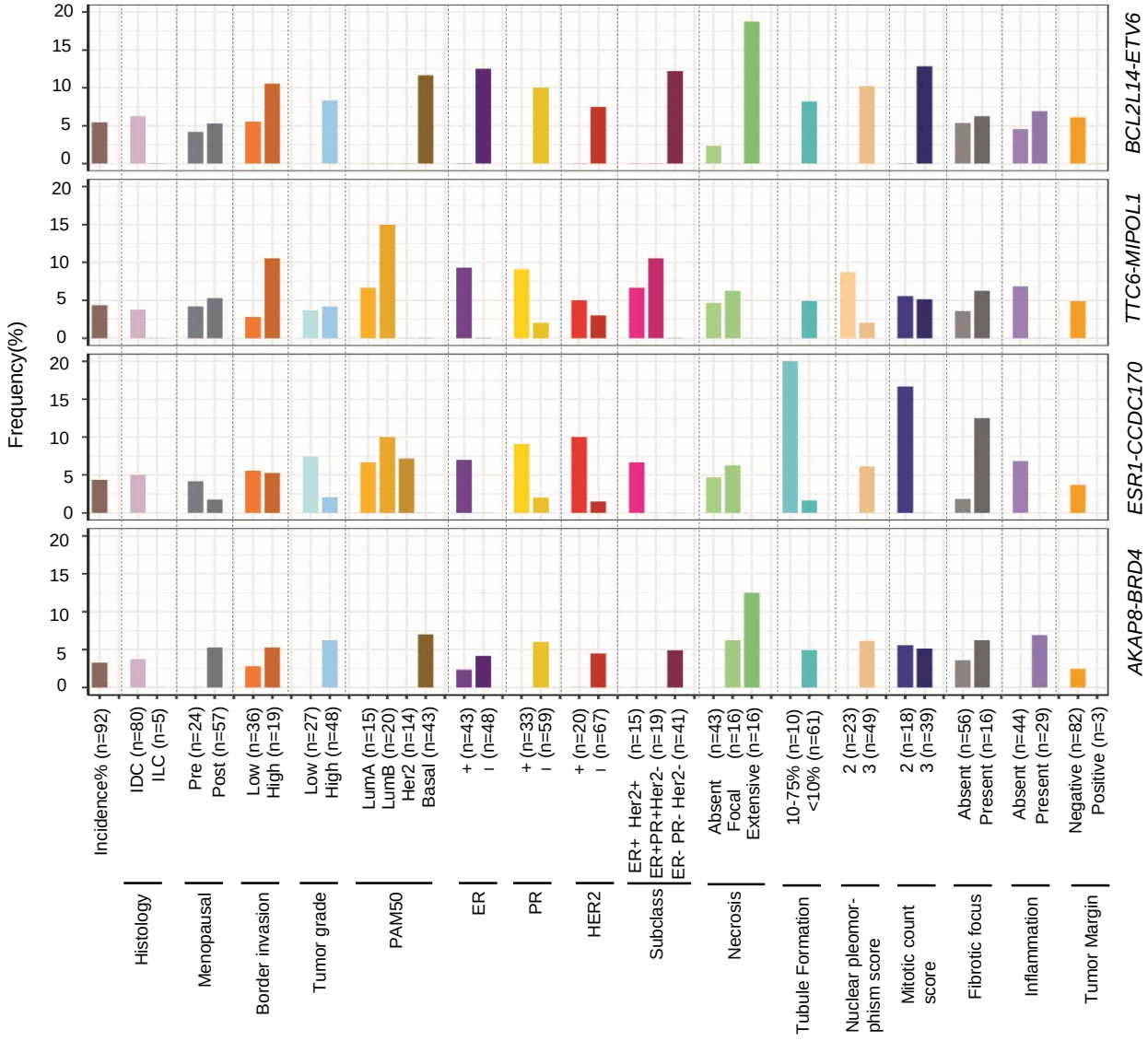
Supplementary Figure 1. Clinicopathological associations of the total number of intergenic rearrangements. Boxplot showing the total number of intergenic rearrangements in the different clinicopathological subtypes of breast tumors. A total of 92 TCGA breast tumors included in the ICGC dataset have available clinical and histopathological data obtained from Heng et al. (PMID: 27861902). *P<0.05, **P<0.01, ***P<0.001 (unpaired Wilcoxon Rank Sum Test).



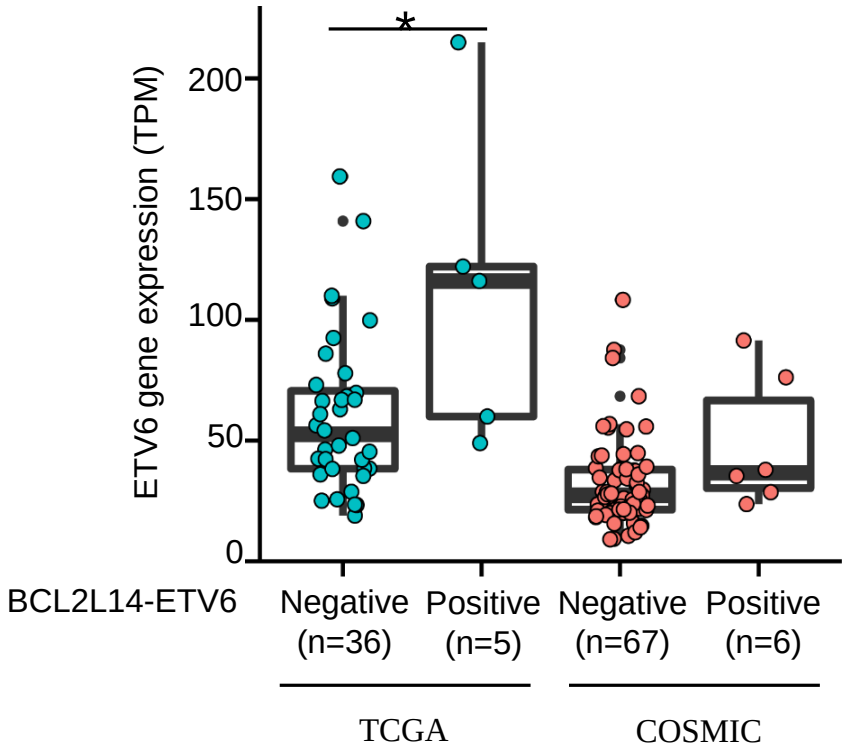
Supplementary Figure 2. Correlation of the top recurrent AGRs with genomic instability index and DNA Damage Repair (DDR) scores. The top AGRs detected in at least two TCGA tumors and >1% of all ICGC tumors are shown in the figure. The weighted genome integrity index (wGII) and DDR deficiency scores are from Marquard et al. (PMID: 26015868). BRCA1 mutation are based on Yost et al. (PMID: 31360904). NtAI, telomeric allelic imbalance; LST, large scale transition; LOH, loss of heterozygosity; Nmut, total number of mutations per sample; FLOH, frequency of LOH.



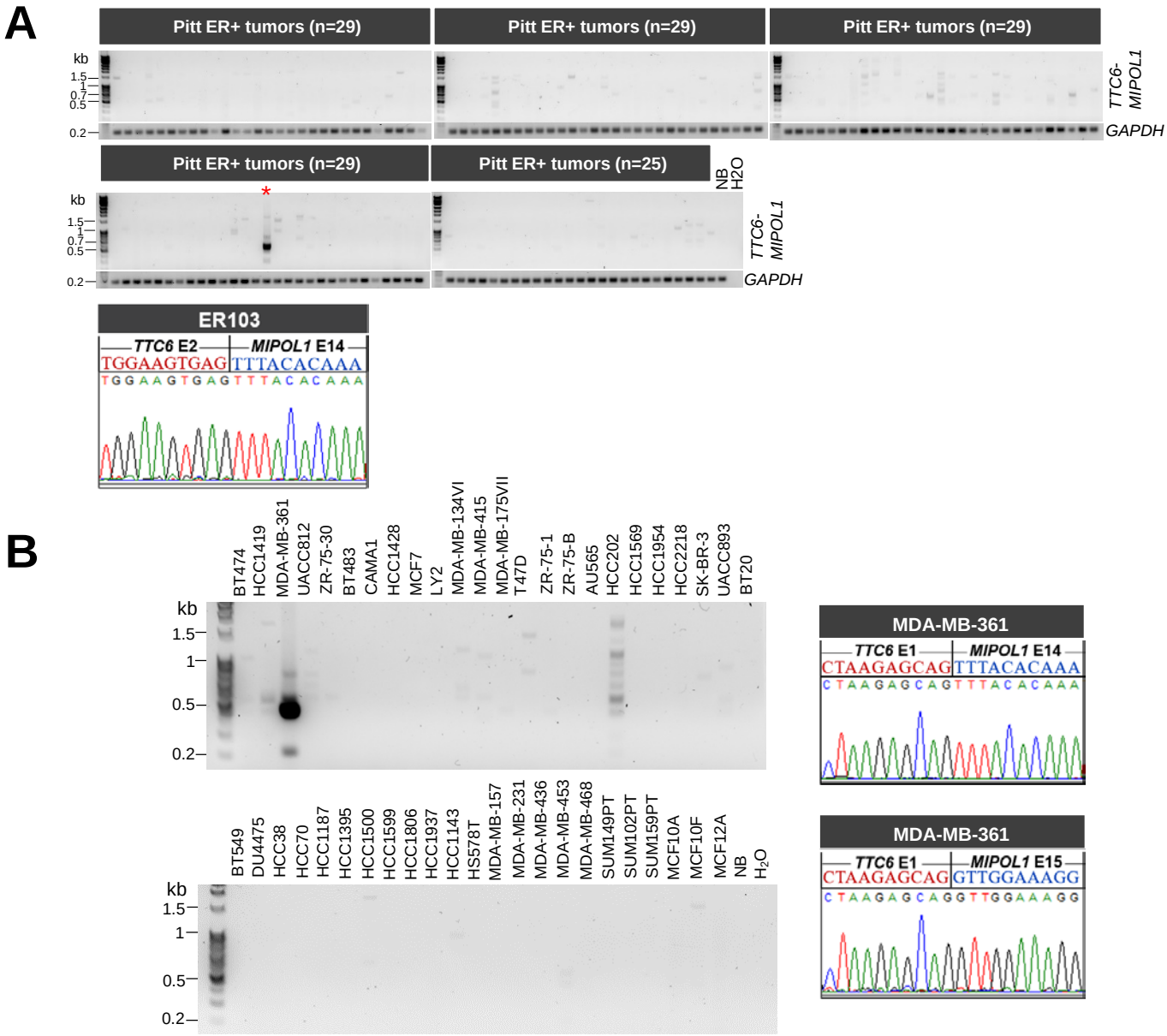
Supplementary Figure 3. The landscape of recurrent fusion partner genes in breast cancer. The incidence (%) of fusion partner genes in TCGA clinicopathological tumor entities are shown in the figure. Only the cases that harbor nonprivate fusions are counted. The partner genes with total frequency count > 4 (1.86 %) were displayed in the figure.



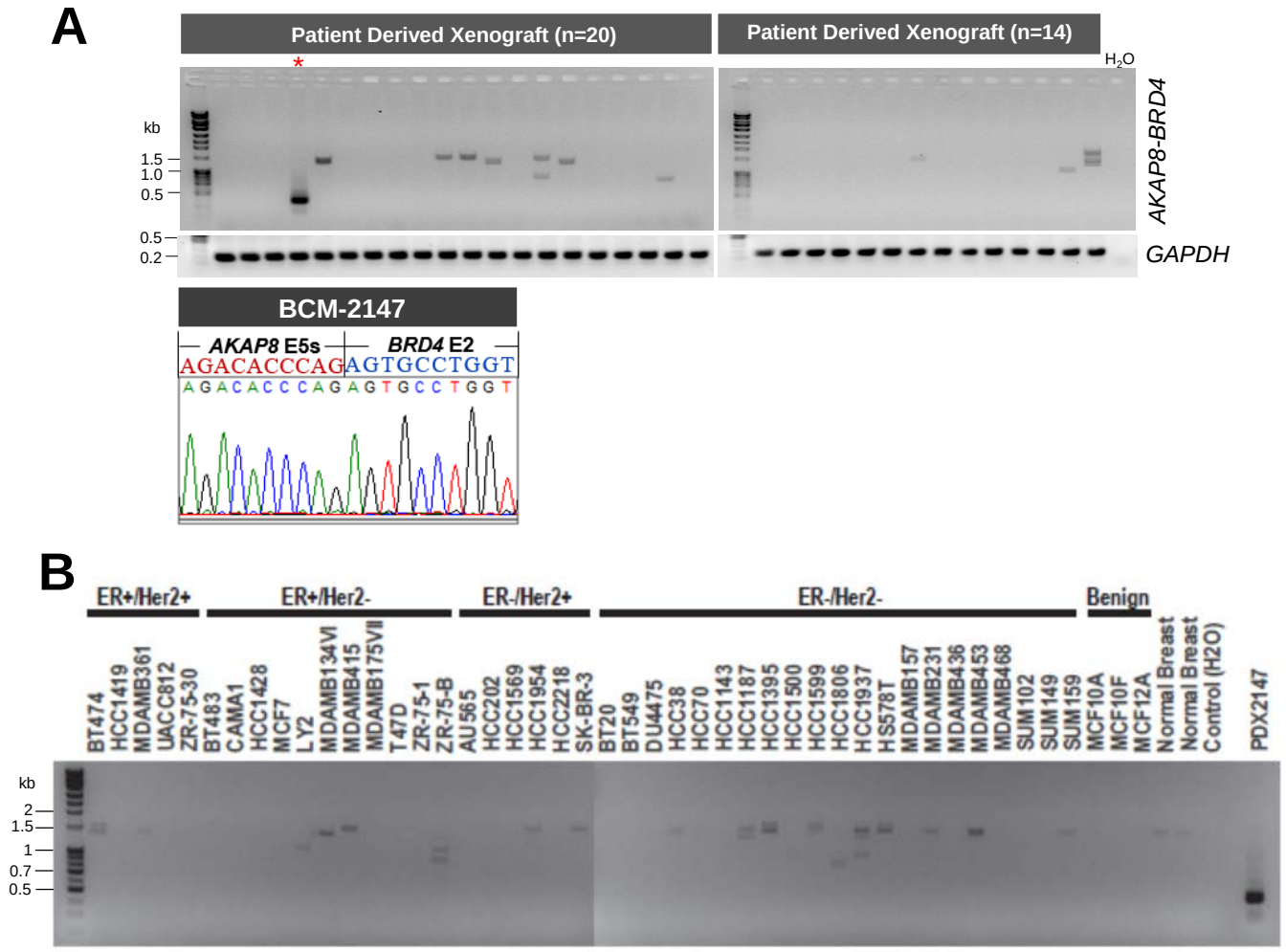
Supplementary Figure 4. Clinicopathological associations with fusion frequency in the four most frequent AGRs. The frequency of the top four AGRs were calculated in each clinical data type of the 92 TCGA breast tumors. The clinical and histopathological data were obtained from Heng et al. (PMID: 27861902).



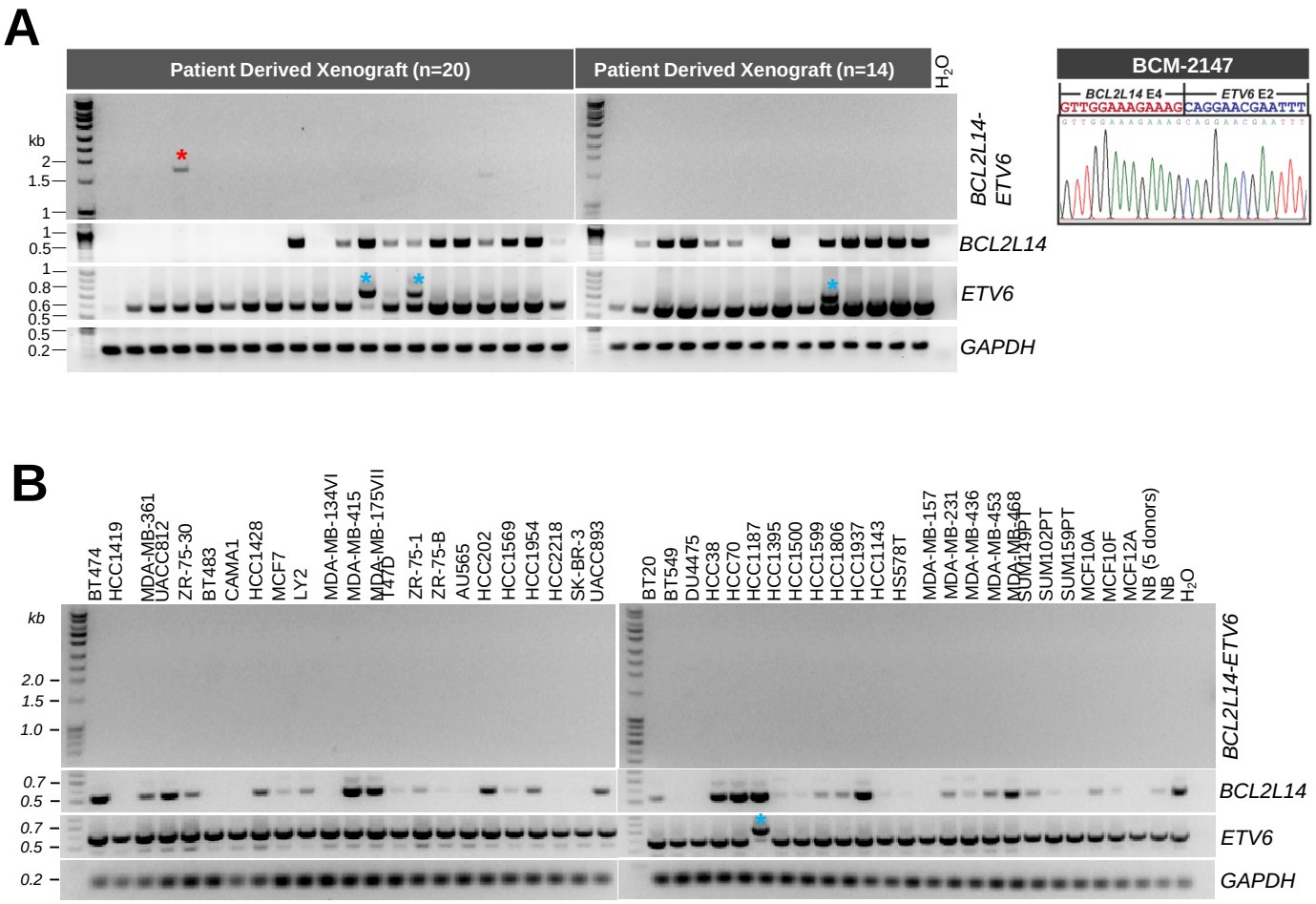
Supplementary Figure 5. ETV6 expression in BCL2L14-ETV6 negative or positive TNBC tumors in TCGA and COSMIC cohorts. *P<0.05 (unpaired Wilcoxon Rank Sum Test).



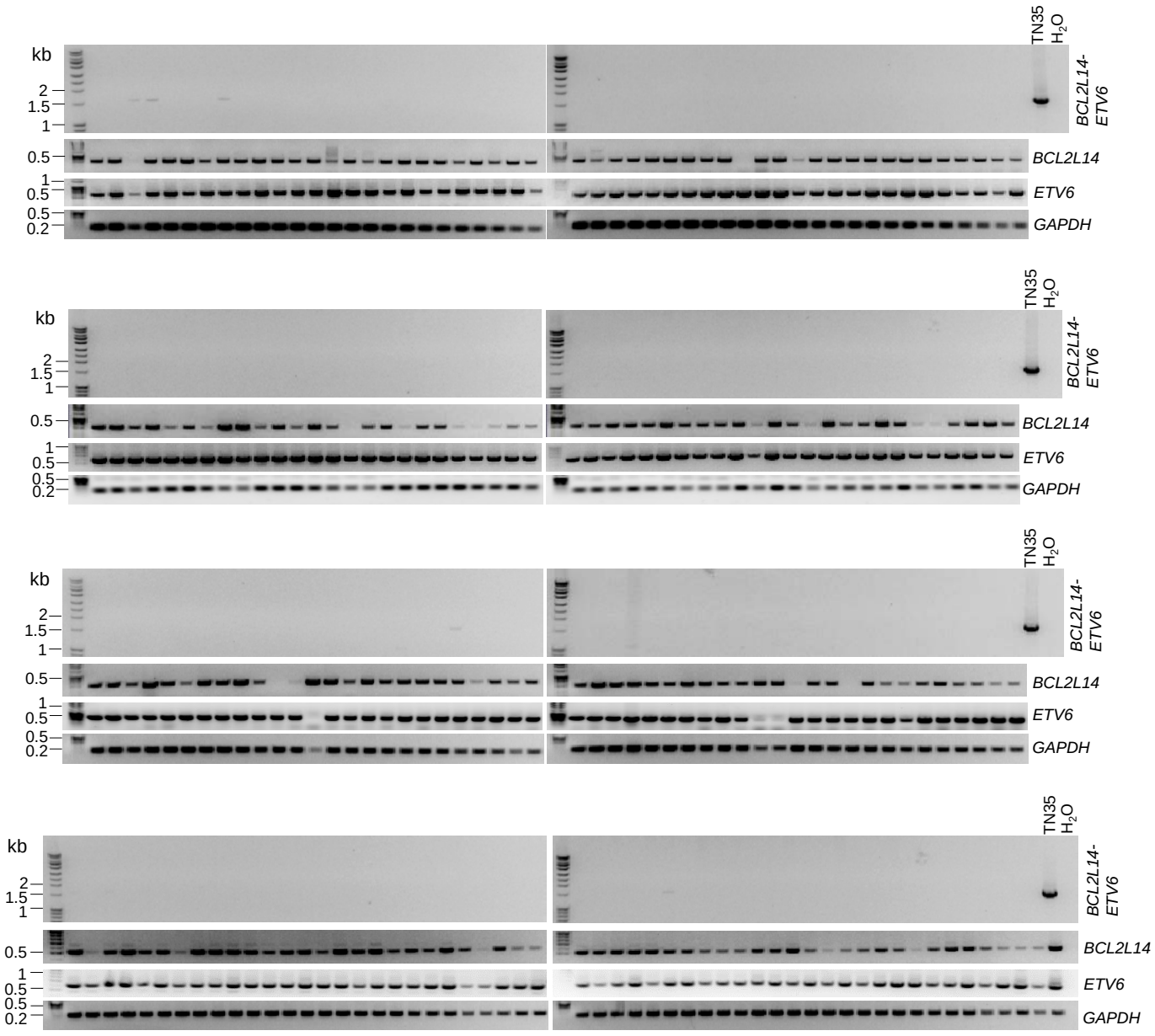
Supplementary Figure 6. Detecting *TTC6-MIPOLI* by RT-PCR in breast cancer cell lines and tumors. (A) RT-PCR analyses of *TTC6-MIPOL1* fusion in a panel of 141 ER+ breast tumors from the University of Pittsburgh cohort, with *GAPDH* as the loading control. Chromatogram in the lower panel shows the junction sequence of *TTC6-MIPOL1I* fusion variant detected in the ER103 tumor sample. Red asterisk denotes ER103. (B) RT-PCR analyses of *TTC6-MIPOL1* fusion in a panel of 44 breast cancer cell lines. Chromatograms in the lower panel show the junction sequences of two *TTC6-MIPOL1I* fusion variants detected in MDA-MB-361.



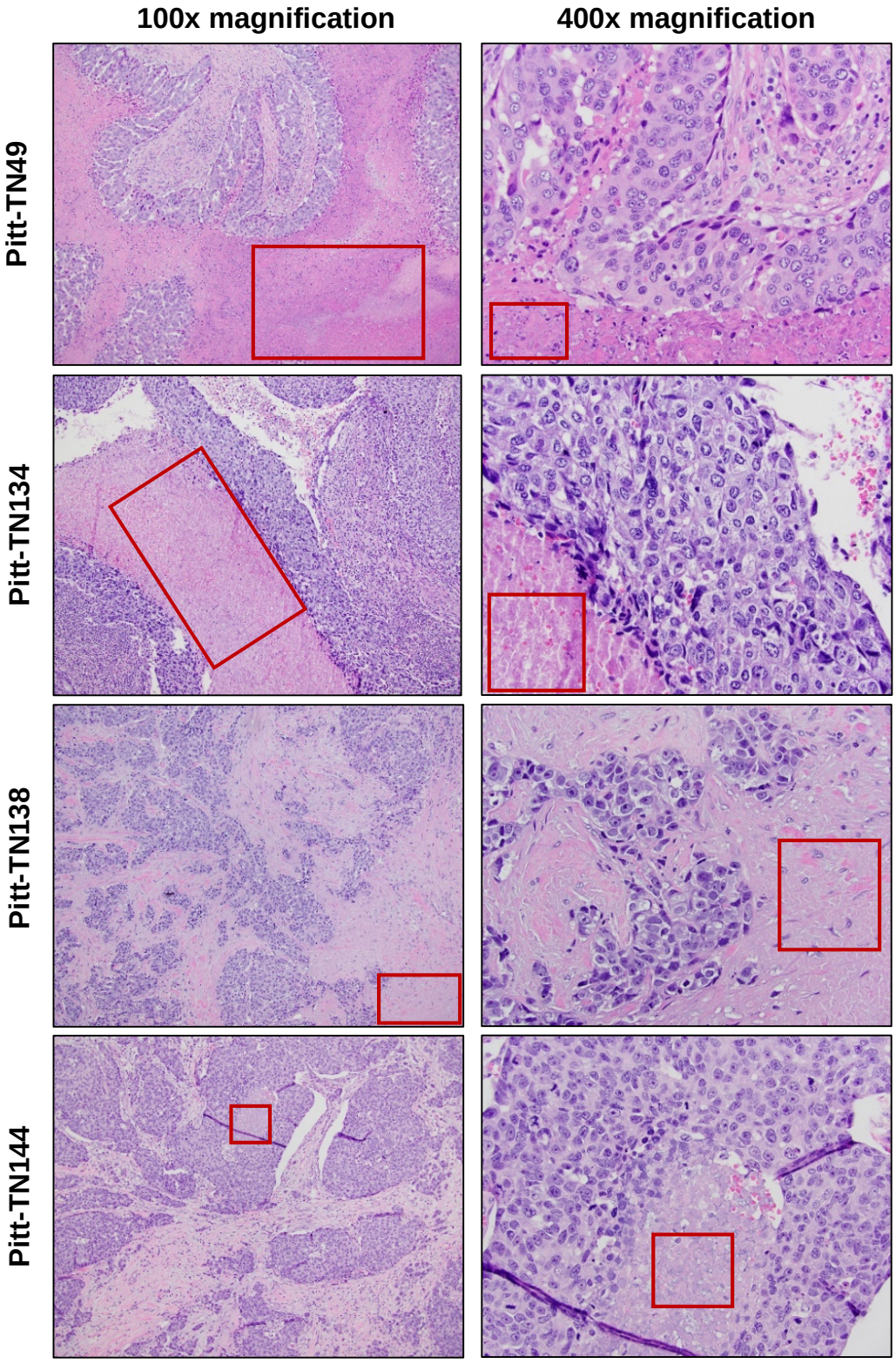
Supplementary Figure 7. *AKAP8-BRD4* expression in patient-derived xenografts and breast cancer cell lines. (A) RT-PCR analyses of *AKAP8-BRD4* fusion in a panel of patient-derived xenografts with *GAPDH* as the control. Chromatogram in the lower panel shows the junction sequence of *AKAP8-BRD4* fusion variant detected in the BCM-2147 PDX sample. Red asterisk denotes BCM-2147. (B) RT-PCR analyses of *AKAP8-BRD4* fusion in a panel of breast cancer cell lines.



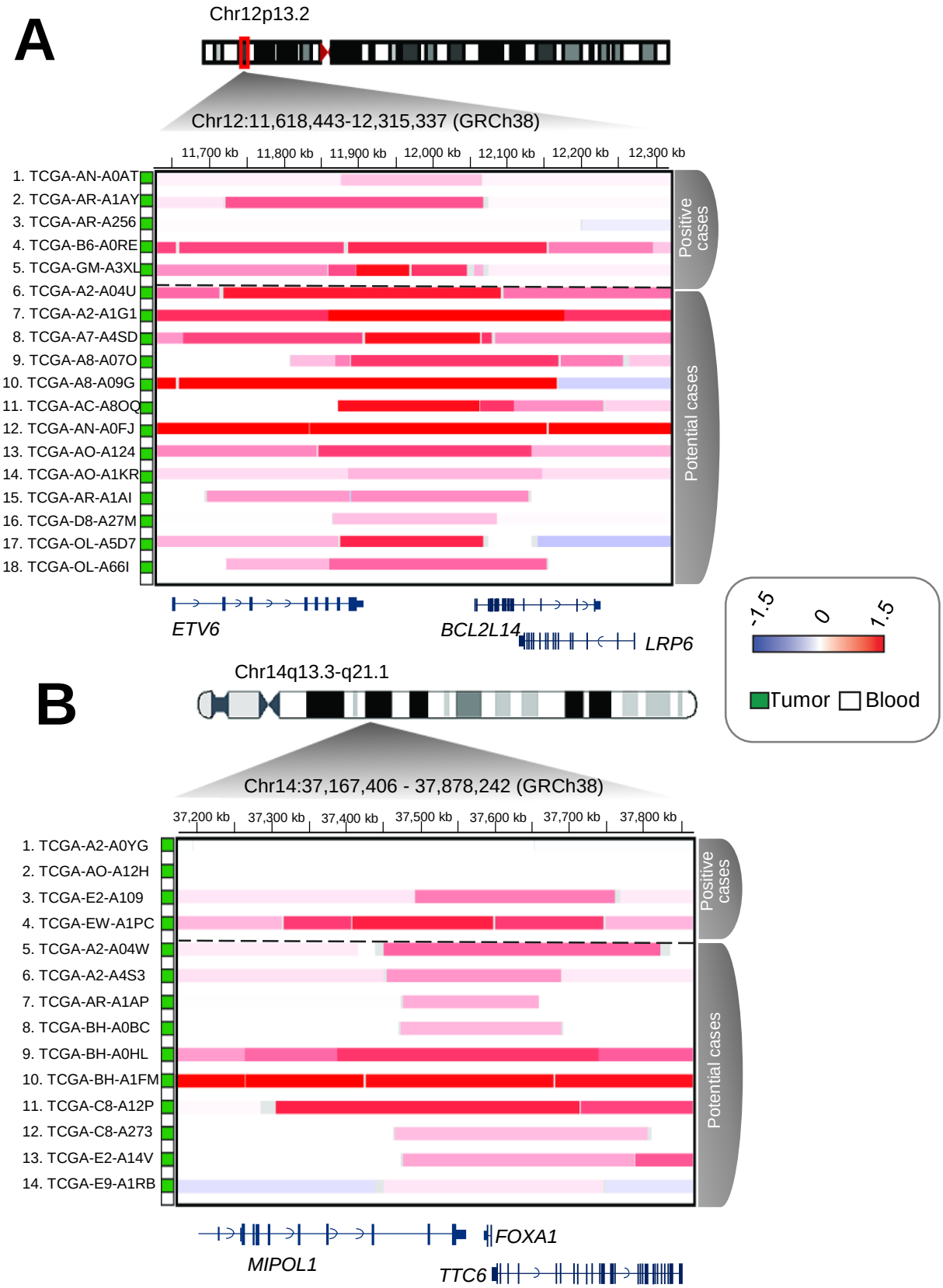
Supplementary Figure 8. (A) RT-PCR analyses of *BCL2L14-ETV6* fusion, wild-type (WT) *BCL2L14* and *ETV6*, and *GAPDH* in 34 triple-negative PDX breast tumors. The *BCL2L14-ETV6*-positive PDX is marked in red asterisks (BCM-2147). Chromatogram on the right shows the junction sequence of the fusion transcript detected in BCM-2147. For wt*ETV6*, blue asterisks denote cases with *ETV6* exon duplications, BCM-3611, BCM-3807 and BCM-5998, from left to right, respectively. (B) RT-PCR Screening of *BCL2L14-ETV6* fusion in a panel of 44 breast cancer cell lines. No cell line was identified with the fusion existence. Blue asterisk denotes the cell line with *ETV6* exon duplication.



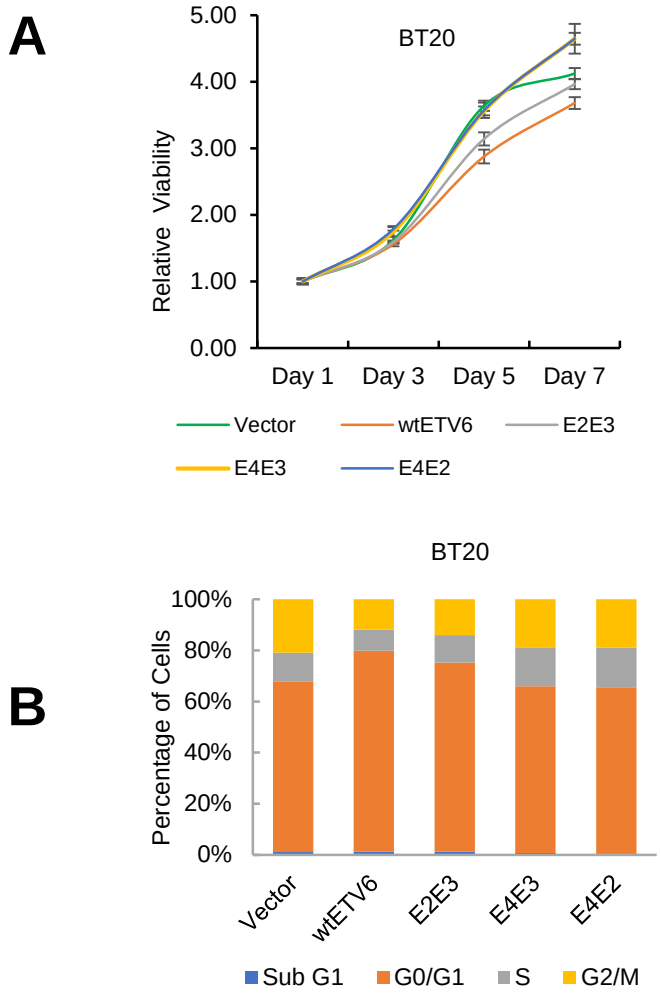
Supplementary Figure 9. RT-PCR analyses of *BCL2L14-ETV6* fusion, wide-type *BCL2L14*, *ETV6*, and *GAPDH* in 200 ER-positive breast tumors from the BCM patient cohort.



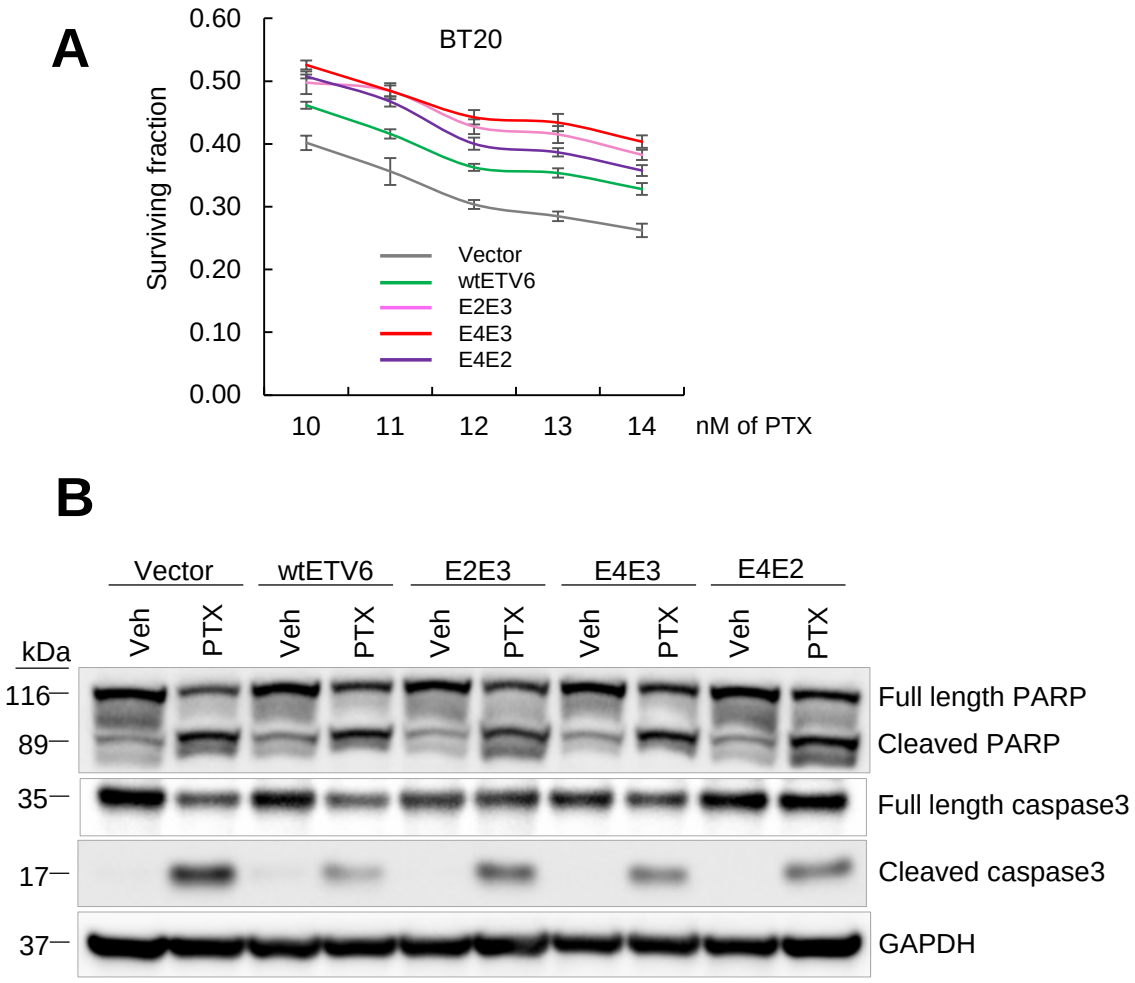
Supplementary Figure 10. Histopathology of *BCL2L14-ETV6* fusion-positive cases from Pitt cohort. Hematoxylin and eosin (H&E) images showing extensive necrosis in two fusion positive case, Pitt-TN49, Pitt-TN134, and focal necrosis in Pitt-TN138 and Pitt-TN144. Regions in the red boxes indicate necrosis areas. All tumors show high nuclear grade.



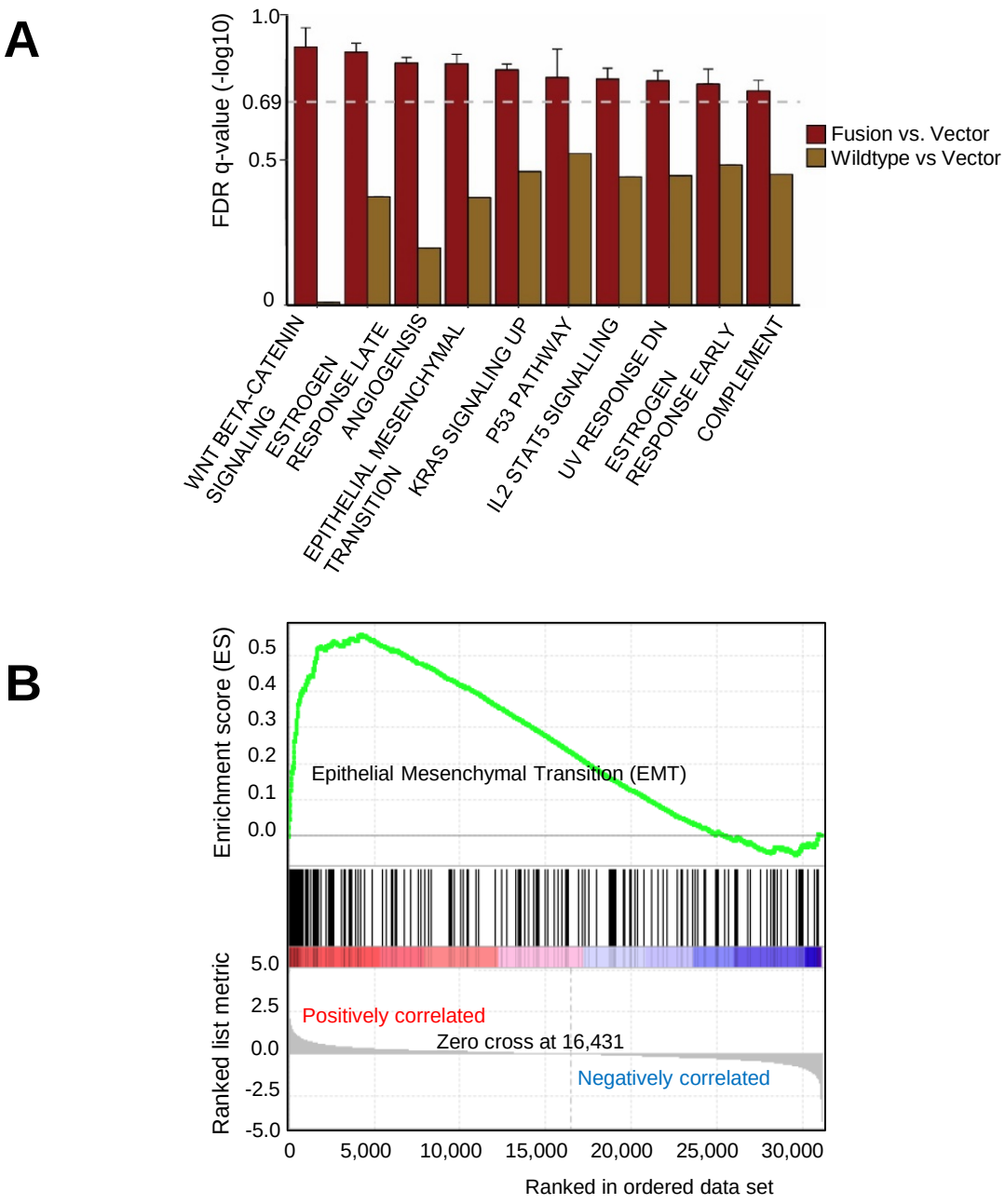
Supplementary Figure 11. Copy number data at the *ETV6*/*BCL2L14 and TTC6-MIPOL1* loci in the fusion positive TCGA cases, and in the TCGA cases that harbor duplications delineating the fusion partner genes. Log2 transformed copy number data for breast tumors and paired normal blood samples are from TCGA. The fusion positive cases detected by WGS data are positioned above the dash line.



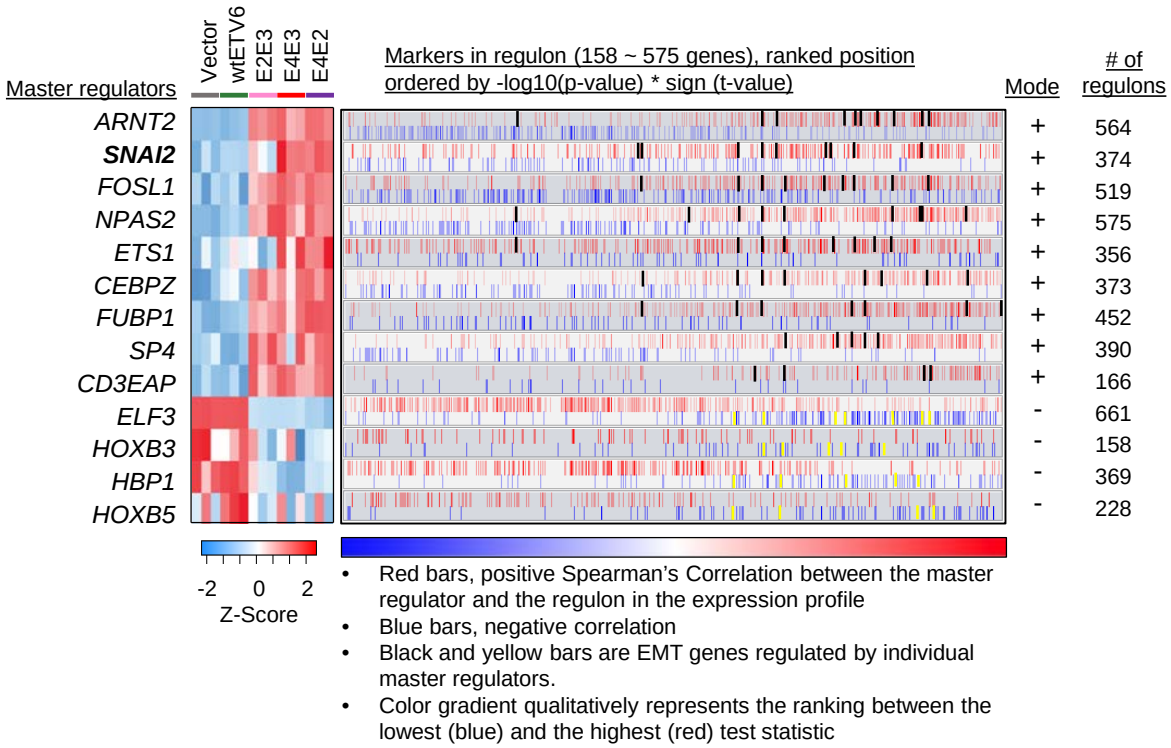
Supplementary Figure 12. The effect of ectopic expression of BCL2L14-ETV6 fusion variants in BT20 on cell viability and cell cycle progression. (A-B) Ectopic expression of BCL2L14-ETV6 fusion variants in BT20 did not result in significant changes in cell viability (A) or cell cycles (B).



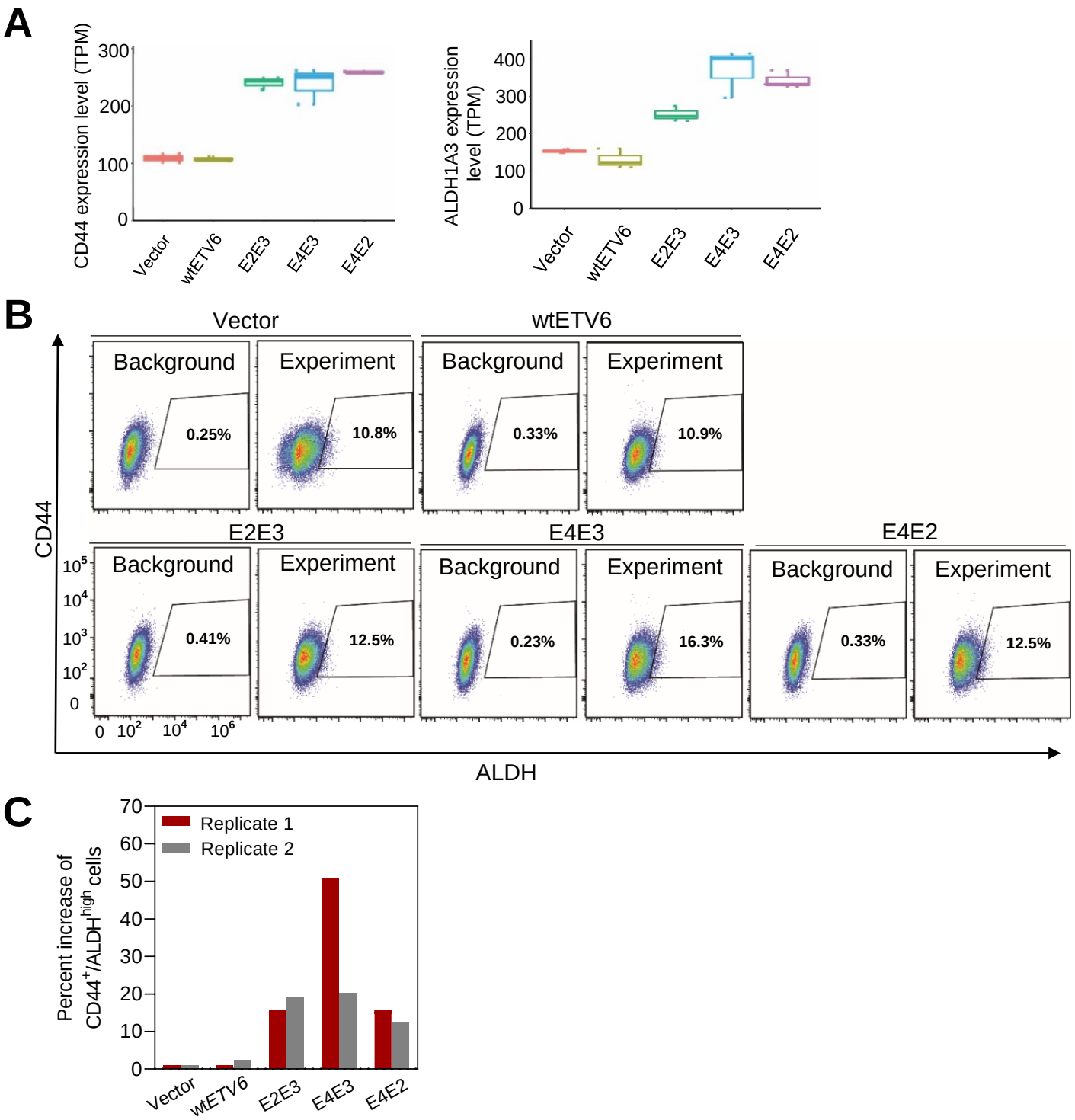
Supplementary Figure 13. The effect of paclitaxel treatment on the viability and apoptosis of the engineerred BT20 cells. (A) BT20 cells overexpressing BCL2L14-ETV6 but not vector- or wtETV6 showed increased resistance to paclitaxel in short-term (72 h) treatment. (B) Apoptotic biomarkers were detected by immunoblotting in the engineered BT20 cells following vehicle (DMSO) or paclitaxel treatment for 48 hours. Veh, vehicle. PTX, paclitaxel.



Supplementary Figure 14. The characteristic of pathway signatures in BCL2L14-ETV6 expressing BT20 cells. (A) Top enriched pathways characteristic of BCL2L14-ETV6 expressing BT20 cells revealed by GSEA. The FDR q-values (-log10) comparing the engineered BT20 cells expressing BCL2L14-ETV6 fusion variants or wtETV6 with the vector control are shown. The 10 pathways shown in the chart have significant FDR q-value < 0.2 (>0.69 in -log10 number) in the comparison between BCL2L14-ETV6 fusion variant vs. vector expressing BT20 cells, but not in the comparison between wtETV6 vs. vector expressing BT20 cells. (B) The enrichment plot of the Epithelial Mesenchymal Transition (EMT) pathway characteristic of the BT20 cells expressing BCL2L14-ETV6 fusion variants, compared with the vector control. The EMT gene signature is from Hallmark gene sets.



Supplementary Figure 15. Heatmap of the expression pattern of the top master regulators. 13 transcription factors were predicted by MRA as master regulators of which expression levels were altered by BCL2L14-ETV6 gene fusion in BT20 cells. SNAI2 was identified as one of the top master regulators that regulate EMT gene signatures in BCL2L14-ETV6 fusion variant expressing breast cancer cells. The heatmap shows the different gene expression levels between vector, wtETV6 and BCL2L14-ETV6 fusion variant expressing BT20 cells. The bar graph shows the distribution of positively (red) or negatively (blue) correlated target genes in the individual master regulators (MR) (Spearman's correlation between the expression levels of the MR and its targets). The black bars within the heatmap indicate EMT genes. The mode explains whether BCL2L14-ETV6 fusion variants positively (+) or negatively (−) affected the expression of the individual MRs.



Supplementary Figure 16. Expression of breast cancer stem cell markers CD44 and ALDH1A3 in BCL2L14-ETV6 fusion-expressing BT20 cells. (A) Box plots showing the expression level of CD44 and ALDH1A3 transcripts by RNA-seq analysis in vector-, wt*ETV6*-, or *BCL2L14-ETV6* fusion variant-expressing BT20 cells. CD44 and ALDH1A3 were over-expressed in *BCL2L14-ETV6* fusion-expressing BT20 cells compared to the vector or wt*ETV6* controls. (B) Representative density plot for detection of CD44 surface marker and ALDH activity by flow cytometry to reveal breast cancer stem cell populations in the engineered BT20 cells. CD44-high and ALDH-high cells are gated as trapeziums and indicated in percentages. (C) Percentages of cells expressing CD44 ($CD44^+$) and cells with high ALDH activity ($ALDH^{high}$) cells in wt*ETV6*- and fusion-expressing BT20 cells, relative to vector control.

Supplementary Materials and Methods

Analyses of whole genome sequencing data. To systematically catalog recurrent AGRs in breast cancer, we analyzed the somatic structural mutation (StSM) data cataloged from WGS data for 215 breast tumor patient cohort released by the ICGC. The StSM variant calling files (.vcf) are downloaded from ICGC portal (https://dcc.icgc.org/repositories, files labeled "dRanger_snowman" or "svfix2"). Using customized Perl scripts, the somatic structural mutations annotated as "PASS" in the "FILTER" column were first mapped with the human exome to reveal the genes and exons affected by the rearrangements (genome build GRCh37), then the fusion partners were determined based on the strands and genomic regions retained in the rearrangements. For mapping the exons, we created a merged exon database based on the exon annotations from GENCODE (https://www.gencodegenes.org/) and UCSC genome browser (https://genome.ucsc.edu/) (V27lift37). The exon numbers for each are assigned based on their starting and ending positions with the exon closest to 5' of the gene assigned as exon 1. The promoter region for each gene is defined as 3kb upstream of its transcription starting site. As authentic recurrent gene fusions usually present distinct genomic breakpoints in different patients, we assessed the median absolute deviations of the genomic breakpoint locations for each recurrent gene fusion. The gene fusions with breakpoint deviations of less than 10bp on each fusion partner gene are excluded from the following analyses, which are likely the result of misalignments. The gene fusions between known homolog genes are also excluded from the following analyses. The resulting recurrent gene fusions were then classified as AGRs, distant intra-chromosomal rearrangements, or inter-chromosomal rearrangements. AGRs are defined as intra-chromosomal rearrangements involving genes of less than 500Kb apart.

Next, we ranked the resulting gene rearrangements by their incidence in the ICGC breast cancer patient cohort, and their concept signature (ConSig) scores (http://www.cagenome.org/consig/, release 2) which indicate their functional relations underlying cancer computed based on the molecular concepts characteristic of known cancer genes, including ontologies, pathways, interactions, and domains (1). Here the max ConSig score of the two fusion partner genes is used to represent each gene fusion. Next, we selected the 92 TCGA cases from the 215 ICGC breast cancer cases and explore the clinicopathological associations of these recurrent gene fusions. For these cases we obtained PAM50 subtype and receptor status from Xena Browser data hub (https://xenabrowser.net/), histopathological classifications from Heng et al. (2), weighted genomic instability index (GII) and DDR deficiency scores from Marquard, et al. (3), TP53, PIK3CA mutation data from cBioPortal (http://www.cbioportal.org/), and BRCA1 mutation from Yost et al. (4). The tumor grade are deduced for TCGA tumors using the Nottingham metric (5). Using the same pipeline described above, we also analyzed the somatic structural rearrangements detected by WGS data for 516 breast tumors, which are provided by the Catalogue of Somatic Mutations in Cancer (COSMIC) (6, 7). We obtained TCGA TNBC subtyping data from Lehmann et al. (8) and Bareche et al. (9) studies. For COSMIC TNBC subtyping, we applied the online tool, TNBCtype (10), on the gene expression data of COSMIC tumors following the TNBC4 subtyping system (BL1, BL2, M, and LAR) (8).

Tissue procurement and RNA extraction. 45 triple-negative and 200 ER+ breast tumor tissues were obtained from the Tumor Bank of Lester and Sue Smith Breast Center at Baylor College of Medicine. 34 triple-negative patient-derived xenografts were kindly provided by Dr. Michael Lewis (11). 89 triple-negative and 141 ER+ breast tumors were gained from the Health Sciences Tissue Bank of University of Pittsburgh. Total RNA for normal breast tissues (5-Donor Pool) was purchased from BioChain. Cell lines' RNA were prepared from the breast cancer cell lines previously obtained from the NCI-ATTC ICBP 45 cell line kit. Total RNA was extracted from the tissues or cell lines using TRIzol reagent (Invitrogen) according to the manufacturer's instruction.

RT-PCR and genomic PCR. Complementary DNA was synthesized using SuperScript IV Reverse Transcriptase (Invitrogen). For amplification of *GAPDH*, RT-PCR was performed with GoTaq G2 DNA Polymerase (Promega), for amplification of *BCL2L14*, *ETV6*, *AKAP8-BRD4* and *TTC6-MIPOL1*, RT-PCR was performed using Platinum Taq DNA Polymerase High Fidelity (Invitrogen), for amplification of *BCL2L14-ETV6* fusions, RT-PCR or genomic PCR was performed with Expand Long Range dNTPack (Roche). PCR products from genomic PCR were purified for capillary sequencing (Macrogen). The primer sequences and PCR conditions are provided in **Table S10**.

Cell culture. MCF10A human breast epithelial cells and BT20 breast cancer cells were obtained from and authenticated by American Type Culture Collection (ATCC). 293FT cells used for lentivirus packaging were purchased from Invitrogen. MCF10A and 293FT cells were cultured as previously described (12). BT20 cells were cultured in EMEM (ATCC) with 10 % fetal bovine serum (FBS, HyClone).

Stable BCL2L14-ETV6 expression vector and stable cell lines. The full-length cDNAs of BCL2L14-ETV6 fusion variants (E2E3, E4E3 and E4E2) containing the full-length ORFs were amplified from fusion-positive tumors (BCM-TN13, BCM-TN35 and BCM-2147), using Expand Long Range dNTPack (Roche) and cloning primer sequences provided in **Table S10**. Wild-type ETV6 full-length cDNA was amplified from ETV6 (NM_001987) human cDNA clone (sc118922, OriGene) using Phusion Hot Start Flex DNA Polymerase (NEB) and cloning primers (**Table S10).** The BCL2L14-ETV6 fusion or wtETV6 cDNA was subcloned into a lentiviral pLenti7.3 vector (Invitrogen). A control *lacZ* gene-containing pLenti7.3 vector was provided by the manufacturer (Invitrogen). After validation by capillary sequencing (Eurofins), these constructs were infected by lentivirus into MCF10A or BT20 cells, and stable cell lines containing the constructs were selected using Flow cytometry sorting against GFP selection marker.

Western blot. For immunoblot analysis, total proteins were extracted by homogenizing the cells in NP40 Lysis Buffer supplemented with complete protease inhibitor cocktail tablet (Roche), 1mM DTT, and 1mM PMSF. 20~50 micrograms of protein extracts were denatured in sample buffer, separated by SDS-PAGE, and transferred onto a PVDF membrane (GE). The membranes were blocked and then incubated for 1 h at room temperature or overnight at 4ºC with primary antibodies, followed by incubation with respective horseradish peroxidase-conjugated secondary antibody. The signals were then visualized by the enhanced chemiluminescence system (Clarity Western ECL Substrate and ChemiDoc imaging system, Bio-Rad). The list of antibodies used for western blots is available in **Table S11**.

Cellular fractionation assay. Engineered stable MCF10A and BT20 cells transduced with lacZ gene, wt*ETV6* or *BCL2L14-ETV6* fusion-containing vectors were freshly harvested for cellular fractionation assay. Cytoplasmic and nuclear proteins of the cells were separated and extracted using NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Fisher Scientific) as per the manufacturer's instructions. The extracted proteins were then used for immunoblot analysis.

Transwell cell migration and Matrigel invasion assays. After serum starvation for 24 h in the starvation medium of DMEM/F12 containing 100 ng/ml cholera toxin, 500 ng/ml hydrocortisone and 2% of horse serum, stable MCF10A cells were then seeded at $3.5X10^4$ cells for migration or $4X10^5$ cells for invasion assay in the reduced growth medium of DMEM/F12 containing 100 ng/ml cholera toxin, 500 ng/ml hydrocortisone and 0.1% BSA in the Boyden chamber insert without or with Matrigel coating (Corning 354480), respectively. Serum-enriched medium (DMEM/F12 containing150 ng/ml cholera toxin, 750 ng/ml hydrocortisone, 30 ng/ml EGF, 0.015 mg/ml human insulin and 10% horse serum) was added to the bottom well of the 24-well plate as attractant. Stable BT20 cells were directly seeded at $2.5X10^4$ cells for migration or $5X10^4$ cells for invasion assay in the reduced growth medium of EMEM containing 0.1% BSA in the upper Boyden chamber without or with Matrigel coating (Corning 354480), respectively. Serum-enriched medium (EMEM containing 20% FBS) was added to the bottom well of the 24-well plate. After 18 h of incubation, migrated/invaded MCF10A or BT20 cells were stained with 0.1% crystal violet in 50% methanol for counting using CCD camera associated microscopy (Olympus) and ImageJ software.

Cell proliferation and clonogenic assays. Engineered stable BT20 cells were seeded at a density of 3,000 cells/well in a 96-well plate. Cell proliferation was measured by MTS assay at different time points using CellTiter 96 AQueous One Solution Cell Proliferation Assay (Promega). For paclitaxel dose curve, stable BT20 cells were seeded at a density of 5,000 cells/well in a 96-well plate and treated with vehicle or different doses of paclitaxel. Cell proliferation was measured by MTS assay after 72 hours of treatment. For clonogenic assay, stable BT20 or MCF10A cells were seeded at a density of 10,000 cells/well in a 24-well plate. After attachment to the plate, cells were treated with 0.1% DMSO (vehicle) or paclitaxel at 5 nM for BT20 cells for 6 days or 15 nM for MCF10A cells for 5 days before replacement of the chemical with fresh growth medium. The remaining colonies were growing in the plate for one month and then stained with 0.5% crystal violet in 50%

ethanol and counted using ChemiDoc photography (Bio-Rad) and ImageJ.

Flow cytometry. For cell cycle analysis, cells were stained with propidium iodide (Sigma) and analyzed using Accuri C6cell analyzer (BD Biosciences). Cell cycle phases were then calculated using FlowJo software. Assessment for the presence of breast cancer stem cells in MCF10A or BT20 cells stably expressing the vector, wtETV6 or BCL2L14-ETV6 fusion was performed via FACS analysis using the AldeRed ALDH detection assay (Millipore Sigma) for detection of ALDH activity and subsequent staining for CD44 cell surface marker using anti-CD44, clone IM7 (eFluor 450, ThermoFisher Scientific) according to the manufacturers' protocols. Following the staining process, cells were then analyzed with LSRFortessa cell analyzer (BD Biosciences) and FlowJo software.

RNA sequencing and data analysis. We used the standard procedure of Qiagen RNeasy kit to extract total RNA from the BT20 cells stably expressing BCL2L14-ETV6 variants, wt*ETV6* cDNA or pLenti7.3 vector containing the *lacZ* gene as control in triplicate experiments. The NovaSeq 6000 library for DNA sequencing was prepared using TruSeq Stranded mRNA Library Prep Kit (Illumina) following the protocol provided by the manufacturer. The final libraries were normalized by quantification with LightCycler 480 II (Roche Applied Science, Indianapolis, IN, USA) and quantification with Bioanalyzer (Agilent, Palo Alto, CA, USA). Final loading concentration was adjusted to 10 pM following the NovaSeq 6000 loading protocol and NovaSeq 6000 S2 Reagent Kit (Illumina) was used for paired-end reads (2×150 bp) sequencing reactions. Sequencing data was given as raw data with a Phred Q30 score of 80 or better.

For analysis we used Rsubread (Bioconductor release 3.8) (13) to align sequence reads to reference genome and used edgeR (14) and limma (15) R packages (Bioconductor release 3.8) to normalize gene expression level to log2 transcripts per million (TPM) (16). We aligned sequence reads to GRCh38 human genome reference sequence and mapped the aligned sequences to Entrez Genes. After normalization, we removed genes of which expression level is zero across all samples to get 31,084 genes for further pathway analysis.

Principle component, clustering, and pathway analyses. To explore the expression clusters of the engineered BT20 cells, we performed unsupervised hierarchical clustering analysis and Principal Component Analysis (PCA). We used Euclidean distance metric in hierarchical clustering, and the first three components in PCA. In addition, we performed gene set enrichment analysis (GSEA) (17) to identify the signaling pathways characteristic of the BT20 cells expressing BCL2L14-ETV6 variants. We performed GSEA analyses comparing BCL2L14-ETV6 variants vs. pLenti73 vector in pairwise, or wtETV6 vs pLenti73 vector using the Hallmark and canonical pathways (C2CP) downloaded from Molecular Signature DataBase (MSigDB) (18). We calculated the mean of normalized enrichment score (NES) and false discovery ate (FDR) from the pairwise GSEA and set the mean FDR q-value to 0.2 (20%) as the threshold to identify significantly enriched pathways.

Master regulator analysis (MRA). We constructed breast cancer cell line BT20-specific interactome by aggregating microarray or RNA-seq samples publicly available. We obtained a total of 13 data sets from GEO (including our own study, GSE120919), which are comprised of 50 microarray samples, 39 RNA-seq samples, and 12 beadchip samples. For the data normalization, we used SCAN.UPC (19) R package (release 3.8) on Affymetrix microarray platform datasets, and used Rsubread (13), edgeR (14), and Limma (15) R packages (release 3.8) on Illumina HiSeq platform datasets as described above. We combined the expression profile datasets with common genes across all samples and corrected batch effects (20). The combined BT20 expression profile data is available through GEO (GSE123917). We collected human TFs from Animal Transcription Factor Database 2.0 (21), and used ARACNe algorithm (22) to construct breast cancer cell line BT20-specific interactome. MRA-Fisher's exact test (FET) (23) inferred the candidate master regulators that regulate EMT gene signature.

Statistical analysis. The associations between *BCL2L14-ETV6* fusion and different clinicopathological features of the 516 breast tumors available in COSMIC were analyzed via Fisher's exact test and P-values were calculated with two-tails. Group wise mutual exclusivity test for the lead recurrent AGRs shown in Fig. 1E was performed with the "Discover" package (24), using the exclusivity statistics and all somatic gene rearrangements as background. The results of all *in vitro* experiments were analyzed by Student's t-tests, and all data are shown as mean ± standard deviation.

Availability of data and materials. The RNA-seq data on BT20 models and combined BT20 expression profile data are available through Gene Expression Omnibus (GSE120919 and GSE123917, respectively). The protocols, codes, and materials used in this study are available upon request to the corresponding author.

References:

1. Wang X-S, *et al.* (2009) An integrative approach to reveal driver gene fusions from paired-end sequencing data in cancer. *Nature biotechnology* 27(11):1005.
2. Heng YJ, *et al.* (2017) The molecular basis of breast cancer pathological phenotypes. *J Pathol* 241(3):375-391.
3. Marquard AM, *et al.* (2015) Pan-cancer analysis of genomic scar signatures associated with homologous recombination deficiency suggests novel indications for existing cancer drugs. *Biomark Res* 3:9.
4. Yost S, Ruark E, Alexandrov LB, & Rahman N (2019) Insights into BRCA Cancer Predisposition from Integrated Germline and Somatic Analyses in 7632 Cancers. *JNCI Cancer Spectr* 3(2):pkz028.
5. Galea MH, Blamey RW, Elston CE, & Ellis IO (1992) The Nottingham Prognostic Index in primary breast cancer. *Breast cancer research and treatment* 22(3):207-219.
6. Nik-Zainal S, *et al.* (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 534(7605):47-54.
7. Forbes SA, *et al.* (2016) COSMIC: High-Resolution Cancer Genetics Using the Catalogue of Somatic Mutations in Cancer. *Curr Protoc Hum Genet* 91:10 11 11-10 11 37.
8. Lehmann BD, *et al.* (2016) Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection. *PLoS One* 11(6):e0157368.
9. Bareche Y, *et al.* (2018) Unravelling triple-negative breast cancer molecular heterogeneity using an integrative multiomic analysis. *Annals of Oncology* 29(4):895-902.
10. Chen X, *et al.* (2012) TNBCtype: a subtyping tool for triple-negative breast cancer. *Cancer informatics* 11:CIN. S9983.
11. Neelakantan D, *et al.* (2017) EMT cells increase breast cancer metastasis via paracrine GLI activation in neighbouring tumour cells. *Nat Commun* 8:15773.
12. Veeraraghavan J, *et al.* (2014) Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers. *Nat Commun* 5:4577.
13. Liao Y, Smyth GK, & Shi W (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* 41(10):e108.
14. McCarthy DJ, Chen Y, & Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* 40(10):4288-4297.
15. Ritchie ME, *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43(7):e47.
16. Wagner GP, Kin K, & Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci* 131(4):281-285.
17. Subramanian A, *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102(43):15545-15550.
18. Liberzon A, *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27(12):1739-1740.
19. Piccolo SR, Withers MR, Francis OE, Bild AH, & Johnson WE (2013) Multiplatform single-sample estimates of transcriptional activation. *Proc Natl Acad Sci U S A* 110(44):17778-17783.
20. Johnson WE, Li C, & Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8(1):118-127.
21. Hu H, *et al.* (2019) AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Res* 47(D1):D33-D38.
22. Margolin AA, *et al.* (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7 Suppl 1:S7.

23. Lefebvre C, *et al.* (2010) A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Mol Syst Biol* 6:377.
24. Canisius S, Martens JW, & Wessels LF (2016) A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. *Genome Biol* 17(1):261.