Am J Cancer Res 2021;11(10):5111-5125
www.ajcr.us /ISSN:2156-6976/ajcr0138211

Original Article

Different members of the APOBEC3 family of DNA mutators have opposing associations with the landscape of breast cancer

Mariko Asaoka[1,3], Santosh K Patnaik[2,4], Takashi Ishikawa[3], Kazuaki Takabe[1,3,4,5,6]

[1]Department of Breast Surgery, Roswell Park Comprehensive Cancer Center, Buffalo, New York, USA; [2]Department of Thoracic Surgery, Roswell Park Comprehensive Cancer Center, Buffalo, New York, USA; [3]Department of Breast Surgery and Oncology, Tokyo Medical University, Tokyo, Japan; [4]Department of Surgery, Jacobs School of Medicine and Biomedical Sciences, State University of New York, Buffalo, New York, USA; [5]Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan; [6]Department of Surgery, Yokohama City University, Yokohama, Japan

Received August 11, 2021; Accepted September 14, 2021; Epub October 15, 2021; Published October 30, 2021

Abstract: APOBEC enzymes are strong mutagenic factors. In breast cancer, expression of APOBEC3B is increased and associated with mutation load and poor outcome. Other APOBEC3s can also mutate DNA but their clinical significance in breast cancer and its underpinnings have not been comprehensively studied. In our examination of 1,091 breast carcinoma cases, high expression of *APOBEC3A* or *APOBEC3B* genes was associated with greater tumor burden of mutations and other genomic aberrations. Expression of none of the five *APOBEC3C-H* genes had any correlation with these features, including T[C-T/G]W mutations, but their high expression levels indicated a robust anti-cancer immune response within tumors, with elevated $CD8^+$ T cell abundance, T cell receptor diversity, and immune cytolytic activity. Concordantly, survival analyses of this and two other cohorts with > 3,000 patients each showed favorable prognostic benefit of high *APOBEC3C-H* expression for both cancer progression and mortality. A detrimental prognostic value was observed for *APOBEC3A* and *APOBEC3B*. Single-cell data revealed cancer epithelial and stromal immune cells as major sources of *APOBEC3B* and *APOBEC3C-H* expression in tumors, respectively. These observations on opposing associations with breast cancer of different APOBEC3s highlight the contrasting roles of these enzymes, promoting cancer through mutagenesis while antagonizing it through immune response.

Keywords: APOBEC, breast cancer, mutation, prognosis, RNA editing, tumor immune microenvironment

Introduction

Analyses of the sequence context of mutations that arise in cancer show that single nucleotide C-to-T or -G mutations in the TCW trinucleotide context (W = A or T) comprise a significant fraction of mutations in a number of solid tumors, especially those of bladder, breast, cervix, head & neck, and lung [1, 2]. These mutations are believed to arise from the activity of APOlipoprotein B mRNA Editing Catalytic polypeptide-like (APOBEC) proteins. APOBECs are zinc-coordinating enzymes with one or more cytidine deaminase (CD) domains that can convert cytosine (C) bases in single-stranded DNA or RNA to uracil (U). Within cells, DNA deamination by APOBEC3s generates C-to-T as well as C-to-G mutations, the latter arising consequent to cellular repair processes. Among the 11 human APOBEC proteins, expression of APOBEC1, APOBEC2, APOBEC4, and Activation-Induced Deaminase (AID) is tissue-specific, respectively restricted to liver, muscle, gonads, and lymphocytes [3], and APOBEC2 and APOBEC4 lack any known ability to deaminase C bases [4]. Thus, mutations with APOBEC signature occurring in solid tumors likely originate from the activity of the remaining seven APOBECs, APOBEC3A-D and APOBEC3F-H (there is no APOBEC3E), which are encoded by a cluster of seven *APOBEC3A-H* genes on chromosome 22. APOBEC3A, 3C, and 3H have a single CD domain, whereas APOBEC3B, 3D, 3F, and 3G have double CD domains. All seven APOBEC3s have the ability to convert C bases of DNA to U, especially in the TCW context [5-7]. However,

C-to-U conversion in RNA (RNA editing) has been reported for only APOBEC3A, 3B, and 3G [8]. DNA mutagenesis by APOBEC3s is widespread in human cancers [1], and is believed to contribute to cancer progression and therapy resistance. Localized hyper-mutation in cancer cell genomes, termed kataegis, which is characterized by clusters of C-to-U or -G mutations in preferential TCW context is also supposed to be a result of APOBEC activity [1, 9, 10].

Among the seven APOBEC3s, APOBEC3B has been studied the most in human cancer, especially that of breast. APOBEC3B has been described as a strong driver of breast cancer [2]. Expression of the *APOBEC3B* gene is upregulated in most primary breast cancer tumors and is associated with kataegis, aggressive clinical and pathological features like high nuclear grade and Ki67 index [11, 12], and poor prognosis [12-17]. Oddly, higher *APOBEC3B* gene expression in breast cancer was significantly associated with achievement of pathological complete response to neoadjuvant chemotherapy in a cohort of 274 patients who underwent tumor resection, though the gene expression level was not prognostic of relapse-free interval or disease-specific survival [12]. Unlike APOBEC3B, the significance of the five APOBEC3C-H proteins in breast cancer is unclear. APOBEC3C-H also possess the C-to-U DNA mutating ability, which is believed to underlie their known role in inhibiting infections by viruses such as human immune deficiency virus type 1 (HIV-1), hepatitis B, and Epstein-Barr virus through deamination of viral cDNAs [3, 18-20]. Here, to investigate the relevance of APOBEC3C-H in breast cancer, we performed systematic cancer genomics and transcriptomics association studies.

Materials and methods

Patient cohorts

Clinical and *APOBEC3* gene expression data of three breast cancer patient cohorts are examined in this study (cohort characteristics in Table S1). The study focuses on the cohort of 1,097 breast carcinoma cases (BRCA) of The Cancer Genome Atlas (TCGA) project that undertook molecular analyses of treatment-naive tumors diagnosed during 1978-2013 (median 2009) [21]. The Sweden Cancerome Analysis Network - Breast (SCAN-B) cohort of 3,273 patients who were diagnosed during 2010-2015 and had transcriptome profiling of resected tumors is from an on-going study. The latest publicly available clinical data of these patients was obtained from resources noted in a recent SCAN-B study [22]. The "Kaplan-Meier (KM) Plotter" set of 5,134 patients is a meta-cohort of subjects and normalized tumor gene expression data that was developed for breast cancer research [23] and currently encompasses 35 studies (2016.10.13 update).

Gene expression data

For 1,091 fresh-frozen tumor and 112 normal tissues of the 1,092 subjects of the TCGA-BRCA project for whom RNA sequencing data was available, gene-level read counts were obtained in April 2018 from National Cancer Institute (USA) Genome Data Commons portal with TCGAbiolinks Bioconductor package for R (version 2.5.9) [24], and used to generate transcripts per million (TPM) values, with transcript lengths as per release 92 of Ensembl human gene annotations. Count data was also processed with edgeR Bioconductor package (version 3.20.9) [25] for normalization with the trimmed mean of M-values (TMM) method to generate gene expression data in counts per million (CPM), excluding genes without count of ≥ 2 in > 56 tissues. Values for rowsum.filter and prior.df parameters in edgeR estimateCommonDisp and estimateTagwiseDisp functions were set at 224 and 0.25, respectively. Data without a Human Genome Organization (HUGO) gene symbol was removed from the final set of gene expression values. For $\log_2$-transformation, the expression values were padded with 0.25. For the 523 fresh-frozen tumors for which mRNA transcriptome had also been profiled with Agilent microarrays, a $\log_2$-transformed, normalized, microarray-based gene expression dataset was obtained from Broad Institute's Firebrowse (version 1.1.40) [26]. Log_2-transformed RNA sequencing-based gene expression measurements of SCAN-B tumors, in terms of fragments per kb per million reads (FPKM), were obtained in July 2020 from Gene Expression Omnibus (dataset GSE96058). RNA sequencing-based gene expression measurements in TPM units for 55 human breast cancer cell-lines of the Cancer Cell Line Encyclopedia [27], 16 human tissues of the Illumina Body Map [28], and 27 types of human periph-

eral blood cells of the BLUEPRINT Epigenome [29] projects were obtained from the European Bioinformatics Institute Expression Atlas [30] in June 2018. Normalized single-cell gene expression and cell identity data of dissociated cells of six human breast cancer tumors examined by Karaayvaz et al. [31] were obtained from Gene Expression Omnibus in June 2020 (GSE118389).

Single nucleotide mutation data

Counts of somatic single nucleotide substitutions in exomes of TCGA-BRCA tumors (n = 1,014) were estimated from publicly available mutation annotation format (MAF) files of the TCGA project [32]. Helmsman software (version 1.1.0) [33] was used to derive the trinucleotide contexts of the mutations, and deconstructSigs Bioconductor package (version 1.8.0) [34] was used for context-based mutational signature analysis. C-to-G or -T mutations in TCW trinucleotide context were considered to be APOBEC-mediated. For breast cancer cell-lines, APOBEC-mediated mutation counts were obtained from the Jarvis et al [35].

Other TCGA-BRCA data

Nottingham histological scores were manually collated from pathology reports by us for a recently published study [36]. Scores could be retrieved for 588 subjects. Relative abundances of 22 types of tumor-infiltrating immune cells were conjectured from tumor gene expression data using the CIBERSORT deconvolution method (online version 1.06) [37] with TPM gene expression values, the LM22 signature matrix, and 100 permutations. Samples (n = 176) with p value > 0.05 in the CIBERSORT examination were excluded from downstream analyses. Abundances for supersets of cell types were generated by adding abundance values of their constituents [38]. Immune cytolytic activity in tumors was quantified from TPM data of *GZMB* and *PRF1* genes as the CYT score [39]. Quantifications of other tumor immune features such as T cell receptor (TCR) diversity and leukocyte infiltration were from Thorsson et al [38]. Clinical outcome data was obtained from the Pan-Cancer Clinical Data Resource [21], a standardized, curated, and filtered dataset for survival endpoints for TCGA cases. Prediction Analysis of Microarray (PAM50) subtype data was from Berger et al [40]. Information for other clinical and pathologic variables such as age, tumor stage, and status for estrogen receptor (ER) expression or human epidermal growth factor receptor 2 (HER2) overexpression was obtained through the cBio Cancer Genomics Portal [41]. Measurements of tumor genome characteristics like aneuploidy, copy number alterations (CNA), clonal heterogeneity quantified by ABSOLUTE method, and homologous recombination defects were from the study of Ellrott et al [38].

Comparisons of groups identified by APOBEC3 gene expression in tumors

Using within-cohort tertiles, patients were grouped by their tumor *APOBEC3* gene expression into high (top tertile) and low (bottom tertile) expressors. Comparisons of these groups of patients for survival outcomes was performed with the survival package (version 3.1-12) in R; Kaplan-Meier method with log-rank test was used. Survival analyses of the KM Plotter cohort for similar comparison of within-cohort tertile groups using log-rank test was performed with the KM Plotter online tool at http://kmplot.com in August 2020. Enrichment of genesets in tumors of high relative to low expressors was analyzed with GSEA desktop software (version 4.0.2) using the 50-geneset Hallmark [42] collection of the Molecular Signatures Database (MSigDB; version 7.1). Genesets with normalized enrichment score > 1.5 and false discovery rate (FDR) q value < 0.05 were considered significant in the enrichment analyses.

Other

For the TCGA cohort, TPM values of *APOBEC3* gene expression were used for inter-APOBEC3 comparisons, and in comparison of tumor and adjacent normal tissues; CPM values were used in other analyses. Statistical analyses and data plotting were performed using R (version 3.6.2) and Prism (version 8.4.3; GraphPad Software®, San Diego, USA). Unless noted otherwise, software options were default ones, and a threshold of 0.05 was used to deem significance from p values of statistical tests.

Results

All seven APOBEC3 genes are expressed in breast cancer tumors

Data of the Illumina Body Map project [28], in which transcriptomes of 16 healthy human tis-

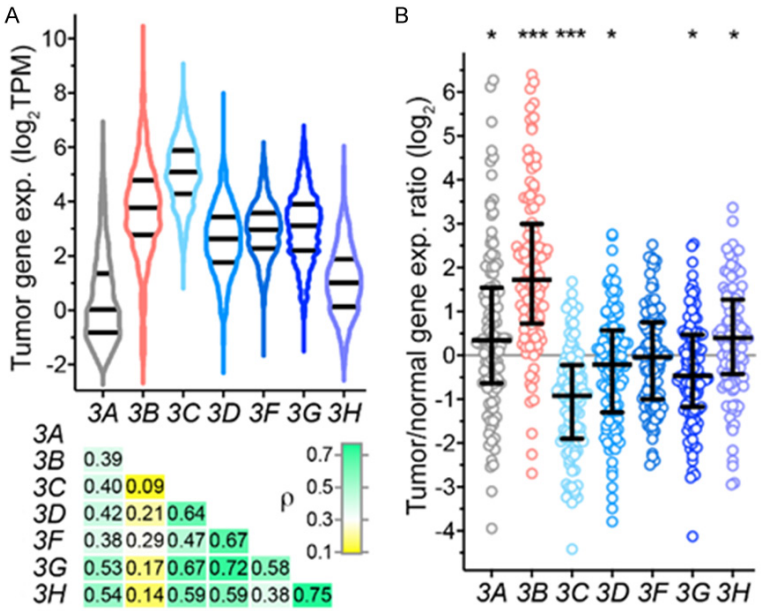

Figure 1. *APOBEC3* gene expression in breast cancer tissues of TCGA. A. Violin plots of gene expression in 1,091 tumors with median and inter-quartile range (IQR) are shown for the seven *APOBEC3* genes (3A-3H). Underneath is a heatmap of Spearman coefficients (ρ) of correlations between the various *APOBEC3*s. TPM, transcripts per million. B. Ratio of *APOBEC3* gene expression of tumors and their matched adjacent normal breast tissue is plotted along with median and IQR values for 112 patients with data available for both tissues. Significant *p* values in paired tumor vs. normal comparison by Welch t test are indicated (*, < 0.05; ***, < 0.001).

sues were profiled by RNA sequencing, shows that all seven *APOBEC3* genes are expressed in breast, with *APOBEC3C* expressed the most compared to other *APOBEC3s* (Figure S1A). Expression of all *APOBEC3* genes, and the highest expression for APO*BEC3C*, are also observed among the 1,091 treatment-naive breast carcinoma tumors of The Cancer Genome Atlas project (TCGA-BRCA) for whom RNA sequencing data was available (**Figure 1A**). There is good correlation among the six *APOBEC3A* and *APOBEC3C-H* genes for their expression levels in the tumors (Spearman ρ = 0.38-0.75), but their correlations with *APOBEC3B* levels are poor, with ρ of 0.09-0.29. Comparison of tumors with their matched normal adjacent tissues, for the 112 subjects with available data, shows that expression of *APOBEC3A*, *3B*, and *3H* is significantly higher in tumors compared to normal, by 1.5-, 3.9-, and 1.3-fold on average, with paired t test *p* values of 0.002, < 0.0001, and 0.005, respectively (**Figure 1B**). On the other hand, significantly reduced expression in tumor compared to normal tissue is observed for *APOBEC3C*, *3D*, and *3G*, with fold-changes of 0.5, 0.8, and 0.8 (P < 0.0001, 0.02, and 0.001), respectively. It should be noted, though, that the 112 patients varied widely for *APOBEC3* gene expression differences between their tumor and normal tissues, and similar tumor-normal gene expression changes in > 75% of subjects were seen for only *APOBEC3B* and *3C* (**Figure 1B**).

Unlike APOBEC3A and 3B, APOBEC3C-H gene expression is not strongly associated with adverse cancer pathological features

The positive association of high *APOBEC3B* gene expression in breast cancer with adverse pathological features has been demonstrated previously, with data of the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort of ~2,000 patients [11]. In the TCGA cohort too, *APOBEC3B* expression was higher among tumors that lack expression of ER or progesterone receptor (PR), are triple negative (TNBC) for ER, PR, and HER2 status, or are of elevated Nottingham histological grade (Welch t test p < 0.001, after adjustment for multiple testing with Benjamini-Hochberg method, for all comparisons; **Figure 2**). However, unlike for METABRIC, *APOBEC3B* gene expression did not differ by status for HER2 or lymph node metastasis in TCGA (**Figure 2**). Associations of *APOBEC3A* with tumor pathological features were like those of *APOBEC3B*. While each of the five *APOBEC3C-H* genes had higher expression in ER-negative (except for *APOBEC3F*) or TNBC tumors, similar to *APOBEC3B*, they had no association with histological grade (except for *APOBEC3H*) or node involvement. For *APOBEC3C* only, expression was significantly higher among HER2-negative tumors. These assessments of gene expression and tumor pathological features show a separation of *APOBEC3C-H* from *APOBEC3A* and *3B*, with the

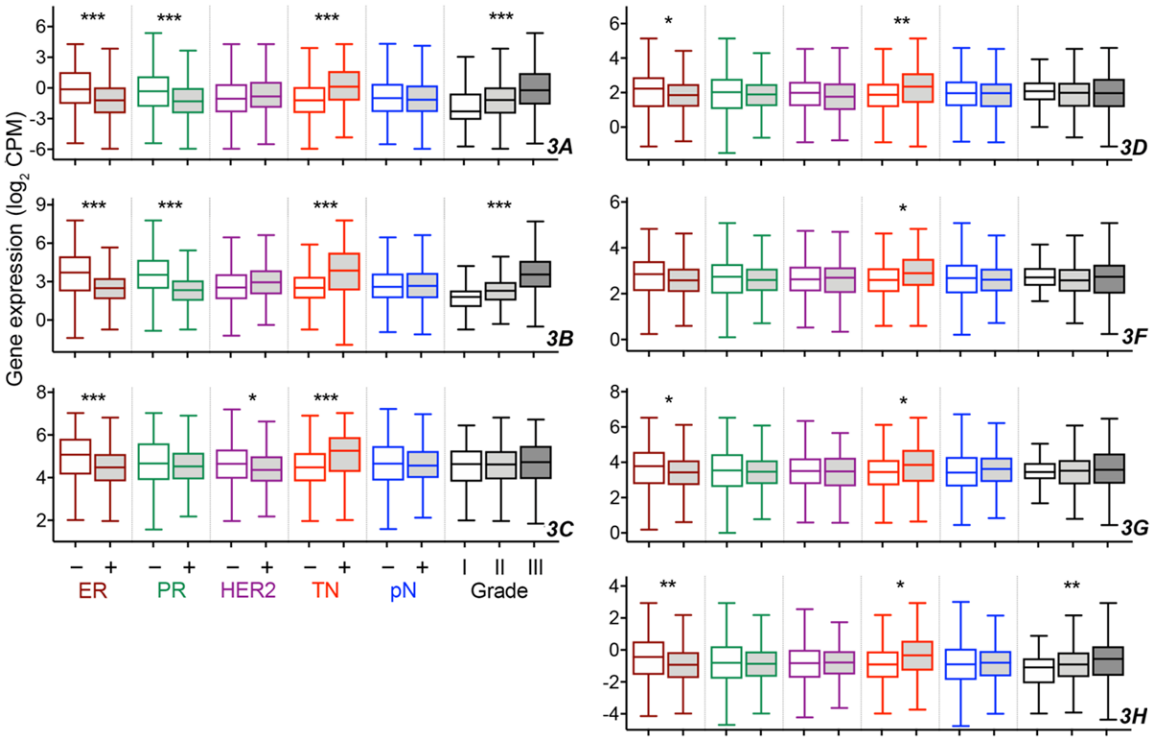

Figure 2. Association of *APOBEC3* gene expression with pathological features of TCGA breast cancer. Tukey boxplots of *APOBEC3* gene expression are shown for tumors of different histological grade (data available for 77, 268, and 235 tumors for grades I, II, and III, respectively), and status for estrogen receptor (ER; 237 negative, 802 positive), progesterone receptor (PR; 342 negative, 694 positive), human epidermal growth factor receptor 2 (HER2; 764 negative, 185 positive), triple-negativity for the three receptors (TN; 858 false, 160 true), and pathologic lymph node metastasis (pN; 513 negative, 556 positive). Significant *p* values, adjusted for multiple testing with Benjamini-Hochberg method, are noted and were calculated with one-way Welch ANOVA test (*, < 0.05; **, < 0.005; ***, < 0.0005).

former lacking a strong association with features that are prognostically adverse.

Expression of APOBEC3A and 3B but not APOBEC3C-H in breast cancer associates with genomic aberrations

To investigate the extent to which different APOBEC3s may reflect cancer-associated genomic changes in breast cancer, we examined the association of their gene expression with various genome features. APOBEC3A and 3B are believed to be the main drivers of APOBEC signature-bearing single nucleotide mutations (C-to-G or -T in the TCW trinucleotide context) [9, 13, 15, 43, 44]. Consistent with this, *APOBEC3A* and *3B* gene expression of tumors correlated positively with their load of all exonic single nucleotide mutations, with Spearman ρ of 0.33 and 0.35 respectively (P < 0.001 for both; **Figure 3A**). Similarly, positive correlations were observed for the subset of APOBEC signature mutations (i.e., likely to be APOBEC-mediated), with *ρ* values of 0.33 and 0.31 respectively (both P < 0.01). On the other hand, Spearman coefficients were poor (0.10-0.21) for *APOBEC3C-H* in analyses of all and APOBEC-mediated mutations. In examination of the 55 human breast cancer cell-lines that have been characterized in the Cancer Cell Line Encyclopedia (CCLE) project, *APOBEC3C-H* expression levels of the cell-lines had similarly poor correlation with their load of APOBEC-mediated mutations (ρ = -0.28-0.03), whereas correlations for *APOBEC3A* and *3B* were 0.52 and -0.11 respectively. As might be expected from their correlations with mutation burden, both *APOBEC3A* and *3B* gene expression had good correlation with neoantigen load. Copy number alteration, aneuploidy, homologous recombination defects, and clonal heterogeneity of tumors too correlated positively with *APOBEC3B* level (ρ = 0.18-0.41), whereas such

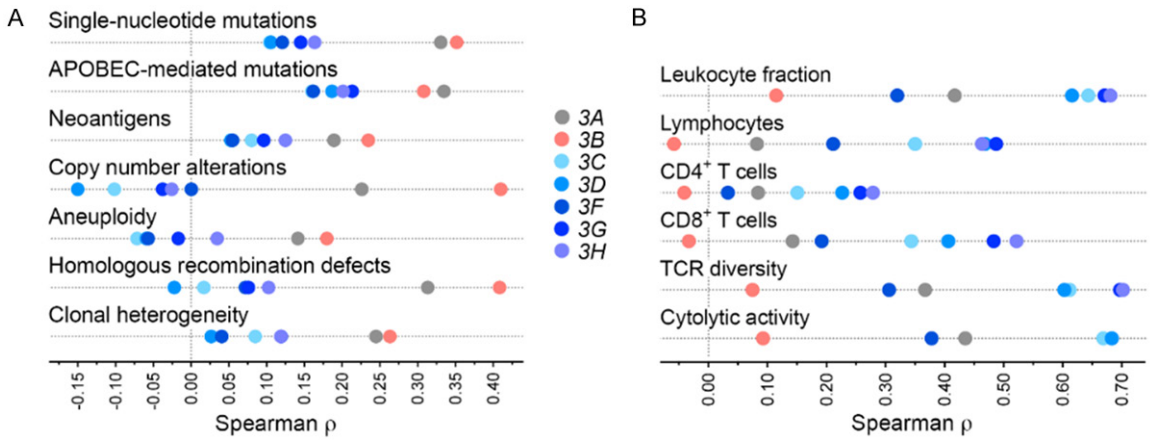

Figure 3. Association of *APOBEC3* gene expression of breast tumors of The Cancer Genome Atlas with their cancer genome and immune characteristics. Shown are Spearman coefficient (ρ) values in analyses of correlation of gene expression of tumors with various features. A. For genome features, total and APOBEC-mediated single-nucleotide exonic mutation counts (n = 1,014 tumors with available data), counts of neoantigens predicted to arise from single-nucleotide mutations and indels (856), segments affected by copy number alteration (1,062), scorings for aneuploidy (1,034) and homologous recombination defects (1,035), and clonal heterogeneity (1,016) are analyzed. B. For immune features, coefficients are shown for infiltrating leukocyte fraction (n = 1,064 tumors with available data) and relative abundance of lymphocytes, $CD4^+$, and $CD8^+$ T cells among infiltrating immune cells (915), T cell receptor (TCR) Shannon diversity (1,038), and immune cytolytic activity measured as the CYT score (1,091).

correlations were weak or absent for *APOBEC3C-H* (ρ = -0.15-0.12; **Figure 3A**).

Immune activity is increased in tumors with high APOBEC3C-H expression

For each of *APOBEC3C-H*, tumor gene expression correlated positively with infiltration of the tumors with leukocytes (ρ = 0.32-0.68) and their lymphocyte subset (ρ = 0.21-0.49), whereas the correlations were poor for *APOBEC3B* (ρ = 0.11 and -0.06, respectively) (**Figure 3B**). Correlations of *APOBEC3C-H* levels with relative abundances of many major tumor-infiltrating immune cell subsets, such as B cells, macrophages, natural killer cells, and neutrophils were poor (Figure S2), positive associations of *APOBEC3C-H* expression, though modest, were observed for relative abundance of $CD8^+$ T cells (ρ = 0.19-0.52). Concordantly, expression of each of these genes correlated positively with estimates of TCR diversity (ρ = 0.31-0.70) as well as immune cytolytic activity [39] within tumors (ρ = 0.38-0.78). For *APOBEC3B*, the correlations of gene expression with these features were much weaker (**Figure 3B**). Evidence of the positive association of *APOBEC3C-H* gene expression with increased anti-cancer immune activity within tumors was also noted in examination of the global tumor transcriptome with geneset enrichment analysis. For this, we scrutinized the 50 genesets of the mSigDB Hallmark collection [42] for enriched expression among tumors with high relative to tumors with low *APOBEC3* gene expression, respectively identified as those in the top and bottom tertiles. With significance thresholds of 0.05 for FDR (q value) and 1.5 for normalized enrichment score, between 7 and 15 genesets were associated with each of the seven *APOBEC3*s. Six genesets were common to the five *APOBEC3C-H* genes as well as *APOBEC3A*. All of these genesets are related to immune activity (allograft rejection, complement, IL2-STAT5 and IL6-JAK-STAT3 signaling pathways, inflammatory response, and interferon gamma response). On the other hand, six of the 14 enriched genesets for *APOBEC3B* are associated with cell proliferation (E2F targets, G2M checkpoint, mitotic spindle, mTORC1 signaling, and Myc targets v1 and v2). None of those 14 genesets is for immune-related processes. **Figure 4** depicts the functional enrichment of immune and proliferative pathways respectively in tumors with high *APOBEC3C-H* and *APOBEC3B* levels.

Breast cancer patients with high tumor APOBEC3C-3H expression have better survival

Consistent with its activity as a DNA mutator, high expression of *APOBEC3B* gene in breast

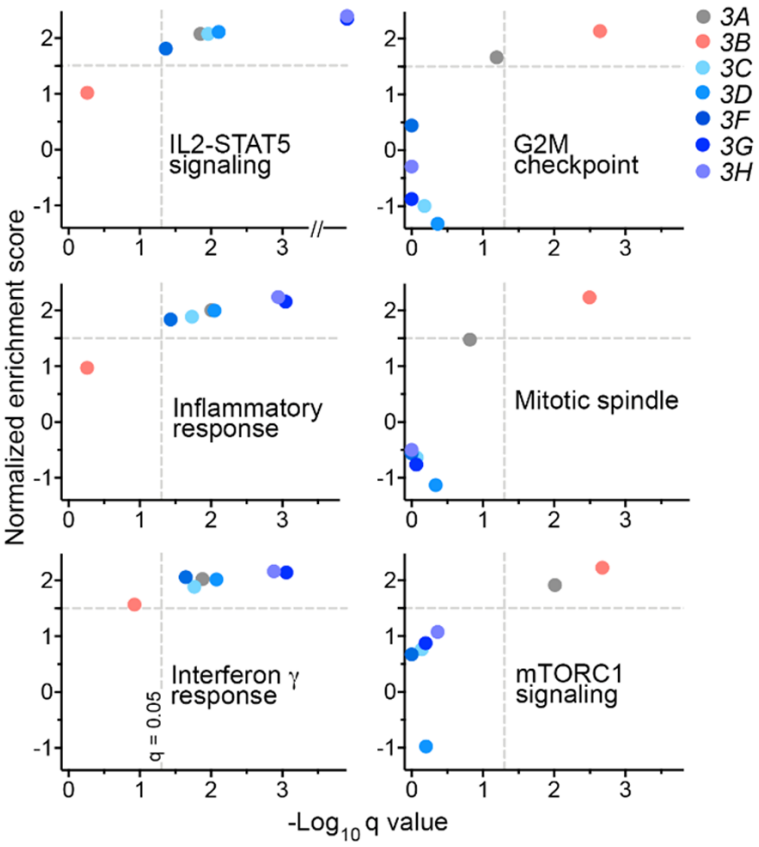

Figure 4. Breast cancer tumors with high *APOBEC3C-H* transcript levels have enriched expression of genesets for immune but not cell proliferation processes. Subjects (n = 1,091) were binned into high and low groups (top and bottom tertiles) by their tumor *APOBEC3* gene expression to compare the two groups in geneset enrichment analysis. Normalized enrichment score and q values, generated from nominal *p* values after correction for false discovery, are plotted for each *APOBEC3* gene for three each of immune- and cell proliferation-related mSigDB Hallmark genesets.

cancer tumors has been associated with worse disease-free survival in the METABRIC cohort [11]. Because the other six APOBEC3s are capable of mutating DNA [3], we had hypothesized that their tumor gene expression too will have a negative impact on disease outcome. On the other hand, the positive associations of *APOBEC3C-H* with favorable tumor genome and immune features that we observed suggest a good prognosis for patients with high tumor *APOBEC3C-H* expression. To clarify this, we examined overall (OS) and disease-specific (DSS) survivals of the TCGA breast cancer patients, comparing the top and bottom one-thirds of patients defined by their tumor gene expression for each *APOBEC3*. Uniformly curated and filtered TCGA data for survival endpoints was used for this analysis [21]. In univariate analyses, though both OS and DSS were worse for high *APOBEC3B* expressors compared to low expressors, with hazard ratios (HR) of 1.32 and 1.57 respectively, the differences were not statistically significant (log-rank test p > 0.05) (**Figure 5**). Similarly, no significant association between *APOBEC3A* expression and survival was observed. However, for each of *APOBEC3C-H*, both OS and DSS were significantly better among high expressors, with hazard ratios of 0.43 to 0.66 (**Figure 5**).

We analyzed two independent and large breast cancer cohorts to validate the prognostic benefit of high *APOBEC3C-H* expression that was observed for TCGA. The SCAN-B cohort was comprised of 3,273 patients of the on-going Swedish initiative [45] for whom clinical and RNA sequencing-based gene expression data is currently publicly available. The other cohort, KM Plotter, is an integrated clinical and hybridization microarray-based tumor gene expression dataset of 5,134 patients of 35 studies [23]. These two cohorts significantly differ from TCGA for certain characteristics such as age at diagnosis and tumor ER status (Table S1). Nevertheless, comparison of OS between within-cohort *APOBEC3* gene expression-based top and bottom tertiles of patients showed a survival benefit of high *APOBEC3C-G* expression in one or both of the KM Plotter and SCAN-B cohorts, with HR values for comparisons with log-rank test *p* values < 0.05 associated with high gene expression ranging from 0.64 to 0.96 (**Figure 6**). For *APOBEC3H*, there was no association with OS in SCAN-B (P = 0.97); measurements of this gene are unavailable in the KM Plotter dataset. Both datasets lack DSS information. Unlike for TCGA, a significantly (P < 0.05) unfavorable association of high expression was noted for

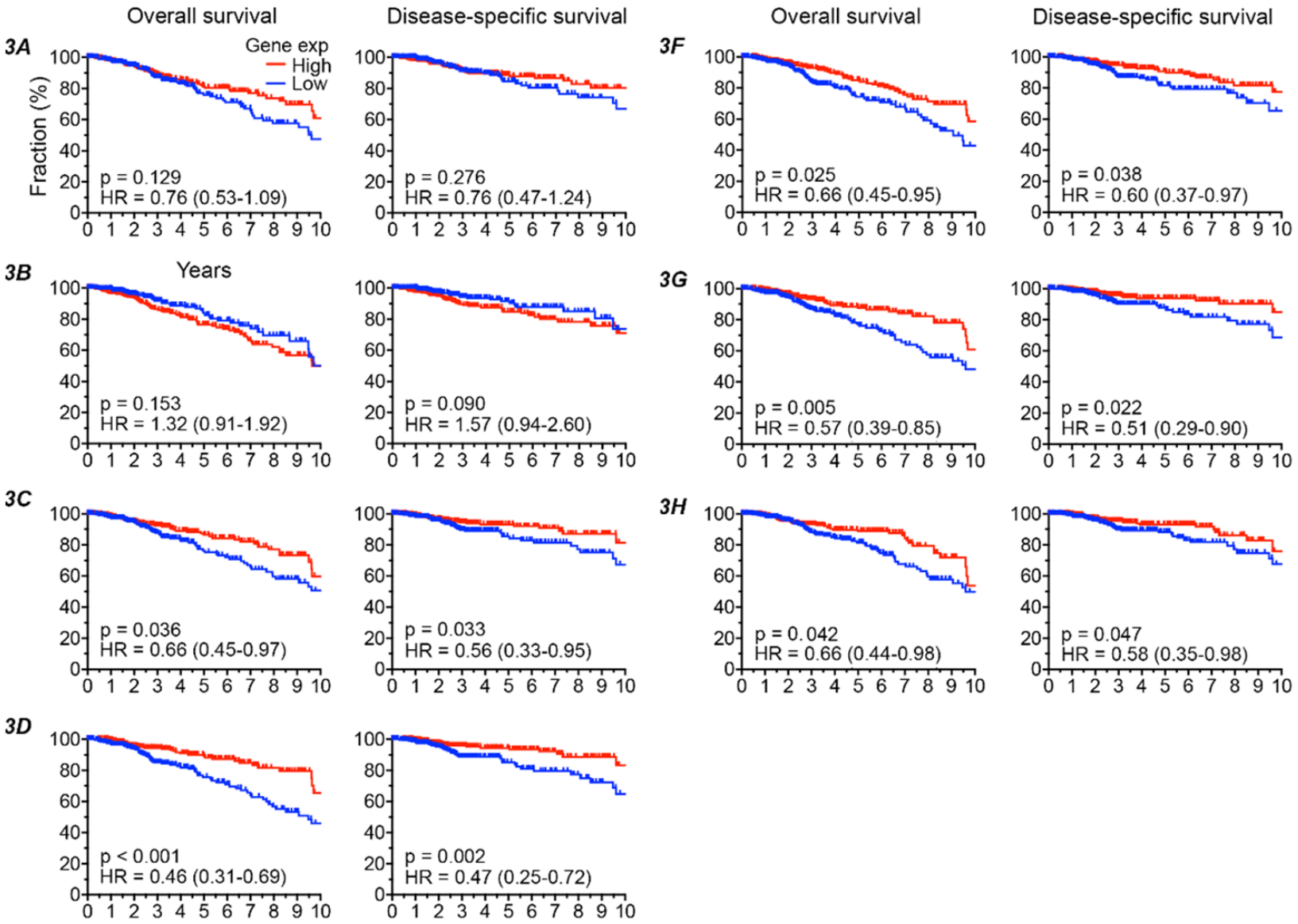

Figure 5. High *APOBEC3C-H* gene expression in breast tumors of The Cancer Genome Atlas is associated with improved survival. The 1,091 subjects with available tumor gene expression data were binned into high and low groups (top and bottom tertiles) by their *APOBEC3* gene expression. Survival plots comparing the high and low groups along with hazard ratio (HR) with its 95% confidence interval and *p* values in log-rank tests are shown for association of *APOBEC3* gene expression with overall and disease-specific survival of patients. The axis for time since cancer diagnosis is truncated at 10 years. Group-sizes in these analyses vary from 346 to 362 because of unavailability of accurate survival data for some subjects.

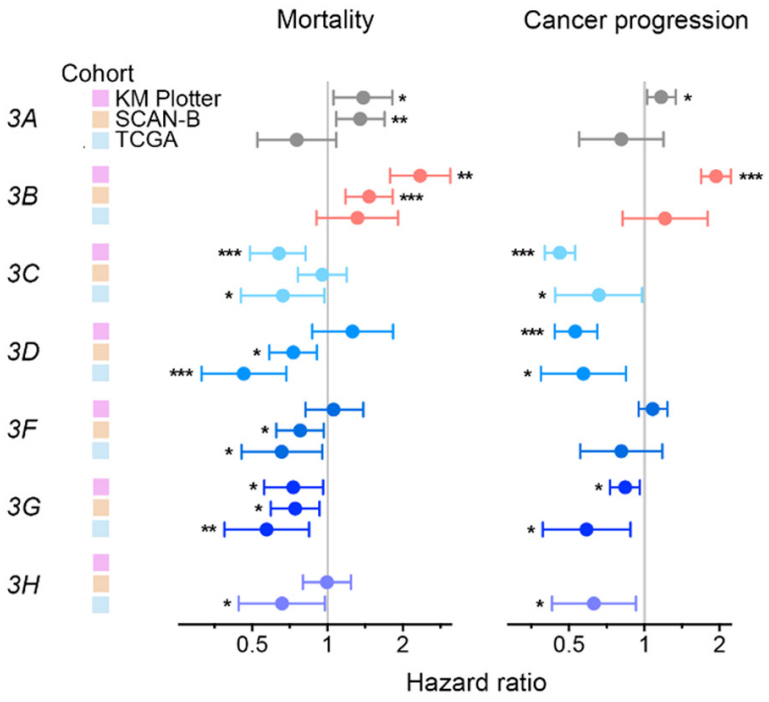

Figure 6. Favorable prognostic value of high tumor *APOBEC3C-H* gene expression in multiple breast cancer cohorts. For three independent cohorts, Kaplan-Meier (KM) Plotter (n = 5,134), Sweden Cancerome Analysis Network-Breast (SCAN-B; 3,273), and The Cancer Genome Atlas (TCGA; 1,091), subjects were binned into high and low groups by their *APOBEC*3 gene expression (within-cohort top and bottom tertiles). The groups were compared for mortality (overall survival) and cancer progression (disease recurrence or progression for KM Plotter and TCGA, respectively) in Kaplan-Meier analysis. Shown are hazard ratios and their 95% confidence intervals, with indicators of *p* values in log-rank tests (* < 0.05; ** < 0.01; *** < 0.001). The KM Plotter dataset lacks *APOBEC3H* values.

APOBEC3A and *3B* in OS analysis of both KM Plotter and SCAN-B data, with HR values of 1.35 to 2.35. In these survival analyses, we ignored patient treatment because detailed data for this aspect was unavailable for all three cohorts. Besides surgery, a majority of patients receive neoadjuvant and/or adjuvant treatments of various types and variable efficacies (chemotherapy, hormonal therapy, radiation, etc.; Table S1).

We also compared high and low *APOBEC3C-H* gene expressors for disease outcomes other than survival. For two of the three patient cohorts, time data was available for cancer progression - time to disease recurrence for KM Plotter, and disease progression-free interval for TCGA. For four of the five *APOBEC3C-H* genes, a protective benefit of high expression against cancer progression was observed in these cohorts, with HR values from 0.46 to 0.84 (**Figure 6**). In the KM Plotter but not the TCGA cohort, high *APOBEC3B* expression was had significant association with likelihood of disease progression. Taken together, these analyses demonstrate the favorable prognostic value of tumor *APOBEC3C-H* and the opposing detrimental value of *APOBEC3A* and *3B* levels. In joint analyses of tumor *APOBEC3* expression levels with various patient demographic and tumor pathological features using multivariable Cox proportional-hazards regression models, none of the seven *APOBEC3*s was associated with cancer mortality or progression independently of histological grade, and node, ER, and PR status, which are routinely evaluated in patients (all Wald test P > 0.05; Table S2). Thus, while the associations of tumor *APOBEC*3 gene expression with disease outcome are of biological interest, tumor *APOBEC3* gene expression does not appear to be of practical translational value as prognostic biomarkers in breast cancer.

APOBEC3C-H expression primarily occurs in immune cells of tumors

To obtain an insight on the source of *APOBEC3C-H* gene expression associated with survival benefit in breast cancer patients, we evaluated data from a number of previous studies. *APOBEC3* gene expression is well-known to be high in peripheral blood leukocytes [14]. Illumina Body Map data shows that expression of all seven *APOBEC3*s is higher in these cells compared to normal breast (Figure S1A). Data of the BLUEPRINT Epigenome project [29] shows that *APOBEC3C-G* are expressed in both myeloid and lymphoid leukocyte subsets in peripheral blood (Figure S1B). On the other hand, expression of *APOBEC3B* is stronger than *APOBEC3C-H* in the CCLE data for 55

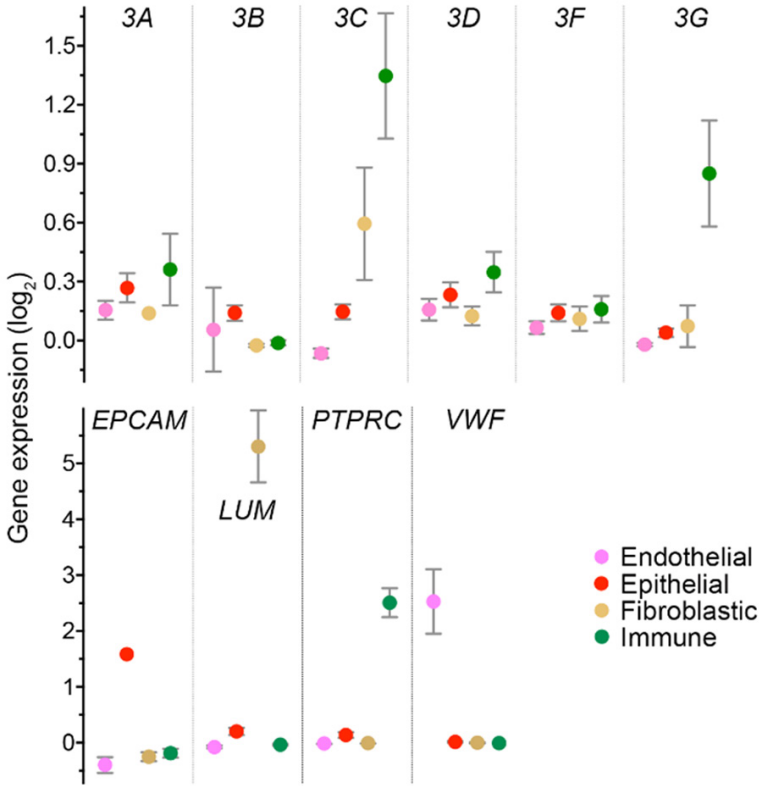

Figure 7. *APOBEC3* gene expression in various components of breast cancer tumors. Gene expression in endothelial (n = 14), epithelial (868), fibroblastic (94), and immune (136) cells of six triple-negative breast cancer tumors are shown for the *APOBEC3* genes, and for well-known gene markers of these types of cells. Mean and its 95% confidence interval are plotted for normalized gene expression data from the single-cell RNA sequencing study of Karaayvaz et al. *APOBEC3H* data was not captured for any of the cells. EPCAM, epithelial cell adhesion molecule; LUM, lumican; PTPRC, protein tyrosine phosphatase, receptor type, C (CD45); VWF, von Willebrand factor.

breast cancer cell-lines (Figure S1C). These observations suggest that cancer epithelial and immune cells largely account for tumor *APOBEC3B* and *APOBEC3C-H* gene expression, respectively. To further explore the cellular origin of *APOBEC3C-H* expression within breast cancer tumors, we examined publicly available single-cell RNA sequencing data of breast cancer tumors. There is only one such study in which all cells were profiled without selection of specific cell-types [31]. Six triple-negative tumors were profiled in the study and cellular identities were established through their expression of canonical markers of cell-types, including *VWF*, *EPCAM*, *LUM*, and *PTPRC* (CD45) genes, respectively, for endothelial, epithelial, fibroblastic, or immune cells. As has been reported with immunohistochemical analysis of tumors [46], the study's data also shows cancer epithelial cells to be the main source of *APOBEC3B* in breast cancer tumors (**Figure 7**). On the other hand, expression levels of *APOBEC3C* and *3G* were much higher in immune compared to cancer cells. *APOBEC3H* was not detected in any type of cell and levels of *APOBEC3D* and *3F*, though low, were higher in immune compared to cancer cells.

Discussion

In contrast to the other APOBECs, the seven human APOBEC3 proteins have demonstrable cytidine deamination activity and less restrictive tissue specificity of expression. These APOBEC3s are therefore likely to be the source of the APOBEC pattern of mutations that are prevalent in cancers of multiple organs [1, 2]. A number of observations indicate that among the APOBEC3s, APOBEC3B is the cause of these mutations in breast cancer. Expression of *APOBEC3B* is higher in tumors relative to adjacent normal breast tissue, and in cancer relative to normal breast epithelial cell-lines, artificial overexpression of the gene in cell-lines generates mutations with APOBEC pattern, and there is a positive association between *APOBEC3B* gene expression of tumors and mutation and kataegis burdens [15, 44]. In line with this, *APOBEC3B* gene expression in breast cancer is elevated in tumors with adverse pathological features and worse clinical outcome [11, 16]. Unlike for APOBEC3B, the significance in breast cancer of the other APOBEC3s, which too are capable DNA mutators, has remained unclear, and we sought to address this with our study.

We find that in most breast cancer tumors of the TCGA cohort, and in many breast cancer cell-lines, all *APOBEC3* genes are expressed, and among these genes, the RNA levels of

APOBEC3C and not *APOBEC3B* are the highest (**Figures 1A**, S1C). However, unlike *APOBEC3A* and *APOBEC3B*, for which expression is on average > 3x higher in tumor compared to normal breast, *APOBEC3C-H* expression is either lower in tumors (*APOBEC3C*) or not significantly altered (**Figure 1B**). Similarity of *APOBEC3A* and *APOBEC3B*, and their dissimilarity from *APOBEC3C-H* are also seen in the associations of gene expression with tumor pathological features. We observed strong associations with prognostically adverse features like ER, PR, or triple negativity, and high tumor grade only for *APOBEC3A* and *APOBEC3B* and not *APOBEC3C-H*. However, contradicting what has been described for the METABRIC cohort [11], *APOBEC3B* expression was not associated with node involvement or positive HER2 status in the TCGA cohort. Consistent with their associations with pathological features, high *APOBEC3C-H* expression in tumors generally correlated with improved disease outcome for both progression and mortality while high expression of either *APOBEC3A* or *APOBEC3B* prognosticated worse outcomes (**Figures 5**, **6**). We conclude so from survival analyses of three large and independent breast cancer patient cohorts. There was variation among the cohorts for some results of these analyses, such as the OS hazard reduction of high *APOBEC3D* or *APOBEC3F* expression reaching significance threshold ($P < 0.05$) for the SCAN-B and TCGA but not KM Plotter cohorts (**Figure 6**). This could be reflective of not only systemic differences between the cohorts, but also the platforms that were used for measuring gene expression (Table S1). Sequence similarity is high among *APOBEC3* genes [47], which can reduce precision of measurement made using microarrays that rely on oligonucleotide probe hybridization, as is the case of KM Plotter and METABRIC cohorts. Indeed, an analysis of 523 TCGA-BRCA tumors whose transcriptomes were profiled using both RNA sequencing and microarray shows that correlation between the methods is either poor or modest for four of the seven *APOBEC3*s, with Spearman ρ of 0.80, 0.21, 0.46, and 0.81 for *APOBEC3A*, *3D*, *3F*, and *3H*, respectively (Figure S3).

The clinical benefit that we observe for high tumor *APOBEC3C-H* expression is supported from its association with markers of heightened anti-cancer immune response within tumors, especially elevated abundance of $CD8^+$ T cells, TCR diversity, immune cytolytic activity, and expression of genesets of immune processes (**Figures 3B**, **4**). For *APOBEC3B*, these associations were absent or weak, and enriched expression of genesets for cell proliferation processes was observed, as has been reported for METABRIC [11] and in concordance with association of high *APOBEC3B* expression with tumor grade (**Figure 2**) and Ki67 level [17].

As expected, in both breast cancer cell-lines and TCGA tumors, correlations of gene expression of tumors with their mutation burden and other facets of cancer genome aberrations, including neoantigenicity, were strongest for *APOBEC3B*, and *3A*, and weak or absent for *APOBEC3C-H* (**Figure 3A**). The same was seen for the tumor load of mutations of APOBEC pattern. Although APOBEC3C-H have the ability to act as DNA mutators [5, 6], association of *APOBEC3C-H* gene expression with mutations of APOBEC pattern was not observed in either the tumors (**Figure 3A**) or cell-lines. This does not necessarily indicate that APOBEC3B is the major cause of APOBEC signature mutations in breast cancer because the mutations arose in the past, when gene expression levels may have been different. Mutations can drive cancer forward, for instance, by dysregulating checkpoints for cell proliferation, but they can also generate antigens that stimulate anti-cancer immune responses. Cancer mutagenesis occurs over long periods of time, and while it is accumulative, clonally generated mutations are also lost to selection pressure, including that from the anti-cancer immune response. This and the fact that immune cells also express APOBEC3s can complicate the examination of the biological and clinical relevance of tumor APOBEC3 activity. Indeed, our examination of gene expression of individual cells of breast cancer tumors suggests that epithelial cancer and immune cells are respectively the major sources of *APOBEC3B* and *APOBEC3C-H* expression in these tumors (**Figure 7**).

A comprehensive examination of all APOBEC3s covering multiple aspects of cancer biology such as immune features and patient outcome has not been previously described. Though we did not generate new data with experiment work, we utilized multiple large patient cohorts in our examination to strengthen the conclu-

sions that may be drawn from its observations. Our study did not examine the APOBEC3 members at a finer level. For instance, haplotype I of APOBEC3H protein has increased nuclear localization and may therefore be more mutagenic compared to haplotype II [48], and a germline variation with deletion of the *APOBEC3B* coding region generates an APOBEC3A-3B hybrid that may be hypermutagenic [49]. It also remains possible that levels of APOBEC3s at the protein or enzyme activity level do not correlate with that at the RNA level, which is used for all analyses of this study. Clearly, additional experimental work is necessary to confirm some of the observations that we have presented here (such as immunohistochemistry or RNA in situ hybridization to identify cellular sources of APOBEC3s). We hope that our analyses of existing data guides such future work. Observations similar to those that we have made about APOBEC3s for breast cancer have been noted for other cancers. For instance, a survival benefit of high tumor *APOBEC3G* expression, associating with high levels of tumor-infiltrating T cells, has been made for ovarian cancer [50].

In conclusion, we show that while the associations of each of the seven APOBEC3s in breast cancer tumors with their genomic, immune, clinical features are concordant, the associations for APOBEC3C-H are favorably prognostic and opposite to the detrimental values of APOBEC3A and APOBEC3B. The underlying basis for the prognostic benefit of high APOBEC3C-H levels in tumors remains unknown. It is possible that this merely reflects the infiltration of tumors by immune cells that appear to be the major tumor source of expression for these genes. On the other hand, functional activities of the APOBEC3C-H proteins within immune cells may be important for anticancer immune response. Besides mutating DNA, many APOBEC3 enzymes can also cause C-to-U RNA editing to affect activity and level of proteins, and we have shown that APOBEC3-mediated C-to-U RNA editing occurs in breast cancer tumors and is associated with improved disease outcome [8]. APOBEC3s may also have molecular activities other than cytidine deamination [51]. Identification of these activities will help us understand the sharply contrasting roles of these enzymes, promoting cancer through mutagenesis while antagonizing it through immune response.

Acknowledgements

This research was supported by grants R01CA160688 to K.T. and P30-CA016056 to Roswell Park Comprehensive Cancer Center from National Institutes of Health, USA.

Disclosure of conflict of interest

None.

Address correspondence to: Kazuaki Takabe, Department of Breast Surgery, Roswell Park Comprehensive Cancer Center, Buffalo, New York, USA. E-mail: kazuaki.takabe@roswellpark.org; Santosh K Patnaik, Department of Thoracic Surgery, Roswell Park Comprehensive Cancer Center, Buffalo, New York, USA. E-mail: santosh.patnaik@roswellpark.org

References

[1] Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, Kiezun A, Kryukov GV, Carter SL, Saksena G, Harris S, Shah RR, Resnick MA, Getz G and Gordenin DA. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet 2013; 45: 970-976.

[2] Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinski M, Jager N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otin C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdes-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ and Stratton MR. Signatures of mutational processes in human cancer. Nature 2013; 500: 415-421.

[3] Harris RS and Dudley JP. APOBECs and virus restriction. Virology 2015; 479-480: 131-145.

[4] Rogozin IB, Basu MK, Jordan IK, Pavlov YI and Koonin EV. APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy) cytidine deaminases predicted by computational analysis. Cell Cycle 2005; 4: 1281-1285.

[5] Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, Shafi S, Johnson DH, Mitter R, Rosenthal R, Salm M, Horswell S, Escudero M, Matthews N, Rowan A, Chambers T, Moore DA, Turajlic S, Xu H, Lee SM, Forster MD, Ahmad T, Hiley CT, Abbosh C, Falzon M, Borg E, Marafioti T, Lawrence D, Hayward M, Kolvekar S, Panagiotopoulos N, Janes SM, Thakrar R, Ahmed A, Blackhall F, Summers Y, Shah R, Joseph L, Quinn AM, Crosbie PA, Naidu B, Middleton G, Langman G, Trotter S, Nicolson M, Remmen H, Kerr K, Chetty M, Gomersall L, Fennell DA, Nakas A, Rathinam S, Anand G, Khan S, Russell P, Ezhil V, Ismail B, Irvin-Sellers M, Prakash V, Lester JF, Kornaszewska M, Attanoos R, Adams H, Davies H, Dentro S, Taniere P, O'Sullivan B, Lowe HL, Hartley JA, Iles N, Bell H, Ngai Y, Shaw JA, Herrero J, Szallasi Z, Schwarz RF, Stewart A, Quezada SA, Le Quesne J, Van Loo P, Dive C, Hackshaw A and Swanton C; TRACERx Consortium. Tracking the evolution of non-small-cell lung cancer. N Engl J Med 2017; 376: 2109-2121.

[6] Jarmuz A, Chester A, Bayliss J, Gisbourne J, Dunham I, Scott J and Navaratnam N. An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. Genomics 2002; 79: 285-296.

[7] Dang Y, Wang X, Esselman WJ and Zheng YH. Identification of APOBEC3DE as another antiretroviral factor from the human APOBEC family. J Virol 2006; 80: 10522-10533.

[8] Asaoka M, Ishikawa T, Takabe K and Patnaik SK. APOBEC3-mediated RNA editing in breast cancer is associated with heightened immune activity and improved survival. Int J Mol Sci 2019; 20: 5621.

[9] Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jonsson G, Garber JE, Silver D, Miron P, Fatima A, Boyault S, Langerod A, Tutt A, Martens JW, Aparicio SA, Borg A, Salomon AV, Thomas G, Borresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ and Stratton MR. Mutational processes molding the genomes of 21 breast cancers. Cell 2012; 149: 979-993.

[10] Pavri R and Nussenzweig MC. AID targeting in antibody diversity. Adv Immunol 2011; 110: 1-26.

[11] Cescon DW, Haibe-Kains B and Mak TW. APOBEC3B expression in breast cancer reflects cellular proliferation, while a deletion polymorphism is associated with immune activation. Proc Natl Acad Sci U S A 2015; 112: 2841-2846.

[12] Fujiki Y, Yamamoto Y, Sueta A, Yamamoto-Ibusuki M, Goto-Yamaguchi L, Tomiguchi M, Takeshita T and Iwase H. APOBEC3B gene expression as a novel predictive factor for pathological complete response to neoadjuvant chemotherapy in breast cancer. Oncotarget 2018; 9: 30513-30526.

[13] Nikkila J, Kumar R, Campbell J, Brandsma I, Pemberton HN, Wallberg F, Nagy K, Scheer I, Vertessy BG, Serebrenik AA, Monni V, Harris RS, Pettitt SJ, Ashworth A and Lord CJ. Elevated APOBEC3B expression drives a kataegic-like mutation signature and replication stress-related therapeutic vulnerabilities in p53-defective cells. Br J Cancer 2017; 117: 113-123.

[14] Refsland EW and Harris RS. The APOBEC3 family of retroelement restriction factors. Curr Top Microbiol Immunol 2013; 371: 1-27.

[15] Burns MB, Lackey L, Carpenter MA, Rathore A, Land AM, Leonard B, Refsland EW, Kotandeniya D, Tretyakova N, Nikas JB, Yee D, Temiz NA, Donohue DE, McDougle RM, Brown WL, Law EK and Harris RS. APOBEC3B is an enzymatic source of mutation in breast cancer. Nature 2013; 494: 366-370.

[16] Sieuwerts AM, Willis S, Burns MB, Look MP, Meijer-Van Gelder ME, Schlicker A, Heideman MR, Jacobs H, Wessels L, Leyland-Jones B, Gray KP, Foekens JA, Harris RS and Martens JW. Elevated APOBEC3B correlates with poor outcomes for estrogen-receptor-positive breast cancers. Horm Cancer 2014; 5: 405-413.

[17] Tokunaga E, Yamashita N, Tanaka K, Inoue Y, Akiyoshi S, Saeki H, Oki E, Kitao H and Maehara Y. Expression of APOBEC3B mRNA in primary breast cancer of Japanese women. PLoS One 2016; 11: e0168090.

[18] Cheng AZ, Yockteng-Melgar J, Jarvis MC, Malik-Soni N, Borozan I, Carpenter MA, McCann JL, Ebrahimi D, Shaban NM, Marcon E, Greenblatt J, Brown WL, Frappier L and Harris RS. Epstein-Barr virus BORF2 inhibits cellular APOBEC3B to preserve viral genome integrity. Nat Microbiol 2019; 4: 78-88.

[19] Suspene R, Guetard D, Henry M, Sommer P, Wain-Hobson S and Vartanian JP. Extensive editing of both hepatitis B virus DNA strands by APOBEC3 cytidine deaminases in vitro and in vivo. Proc Natl Acad Sci U S A 2005; 102: 8321-8326.

[20] Harris RS, Hultquist JF and Evans DT. The restriction factors of human immunodeficiency virus. J Biol Chem 2012; 287: 40875-40883.

[21] Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC,

Levine DA, Lee AV, Omberg L, Wolf DM, Shriver CD, Thorsson V and Hu H. An integrated TCGA Pan-cancer clinical data resource to drive high-quality survival outcome Analytics. Cell 2018; 173: 400-416 e411.

[22] Brueffer C, Gladchuk S, Winter C, Vallon-Christersson J, Hegardt C, Häkkinen J, George AM, Chen Y, Ehinger A, Larsson C, Loman N, Malmberg M, Rydén L, Borg Å and Saal LH. The mutational landscape of the SCAN-B real-world primary breast cancer transcriptome. EMBO Mol Med 2020; 12: e12118.

[23] Győrffy B, Lanczky A, Eklund AC, Denkert C, Budczies J, Li Q and Szallasi Z. An online survival analysis tool to rapidly assess the effect of 22,277 genes on breast cancer prognosis using microarray data of 1,809 patients. Breast Cancer Res Treat 2010; 123: 725-731.

[24] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G and Noushmehr H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 2016; 44: e71.

[25] Robinson MD, McCarthy DJ and Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010; 26: 139-140.

[26] Lee JS. Exploring cancer genomic data from the cancer genome atlas project. BMB Rep 2016; 49: 607-611.

[27] Ghandi M, Huang FW, Jane-Valbuena J, Kryukov GV, Lo CC, McDonald ER 3rd, Barretina J, Gelfand ET, Bielski CM, Li H, Hu K, Andreev-Drakhlin AY, Kim J, Hess JM, Haas BJ, Aguet F, Weir BA, Rothberg MV, Paolella BR, Lawrence MS, Akbani R, Lu Y, Tiv HL, Gokhale PC, de Weck A, Mansour AA, Oh C, Shih J, Hadi K, Rosen Y, Bistline J, Venkatesan K, Reddy A, Sonkin D, Liu M, Lehar J, Korn JM, Porter DA, Jones MD, Golji J, Caponigro G, Taylor JE, Dunning CM, Creech AL, Warren AC, McFarland JM, Zamanighomi M, Kauffmann A, Stransky N, Imielinski M, Maruvka YE, Cherniack AD, Tsherniak A, Vazquez F, Jaffe JD, Lane AA, Weinstock DM, Johannessen CM, Morrissey MP, Stegmeier F, Schlegel R, Hahn WC, Getz G, Mills GB, Boehm JS, Golub TR, Garraway LA and Sellers WR. Next-generation characterization of the cancer cell line encyclopedia. Nature 2019; 569: 503-508.

[28] Nellore A, Jaffe AE, Fortin JP, Alquicira-Hernandez J, Collado-Torres L, Wang S, Phillips RA, III, Karbhari N, Hansen KD, Langmead B and Leek JT. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. Genome Biol 2016; 17: 266.

[29] Martens JH and Stunnenberg HG. BLUEPRINT: mapping human blood cell epigenomes. Haematologica 2013; 98: 1487-1489.

[30] Petryszak R, Burdett T, Fiorelli B, Fonseca NA, Gonzalez-Porta M, Hastings E, Huber W, Jupp S, Keays M, Kryvych N, McMurry J, Marioni JC, Malone J, Megy K, Rustici G, Tang AY, Taubert J, Williams E, Mannion O, Parkinson HE and Brazma A. Expression Atlas update--a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. Nucleic Acids Res 2014; 42: D926-932.

[31] Karaayvaz M, Cristea S, Gillespie SM, Patel AP, Mylvaganam R, Luo CC, Specht MC, Bernstein BE, Michor F and Ellisen LW. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat Commun 2018; 9: 3588.

[32] Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, Hess J, Ma S, Chiotti KE, McLellan M, Sofia HJ, Hutter C, Getz G, Wheeler D and Ding L; MC3 Working Group; Cancer Genome Atlas Research Network. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst 2018; 6: 271-281, e277.

[33] Carlson J, Li JZ and Zollner S. Helmsman: fast and efficient mutation signature analysis for massive sequencing datasets. BMC Genomics 2018; 19: 845.

[34] Rosenthal R, McGranahan N, Herrero J, Taylor BS and Swanton C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol 2016; 17: 31.

[35] Jarvis MC, Ebrahimi D, Temiz NA and Harris RS. Mutation signatures including APOBEC in cancer cell lines. JNCI Cancer Spectr 2018; 2: pky002.

[36] Asaoka M, Patnaik SK, Zhang F, Ishikawa T and Takabe K. Lymphovascular invasion in breast cancer is associated with gene expression signatures of cell proliferation but not lymphangiogenesis or immune response. Breast Cancer Res Treat 2020; 181: 309-322.

[37] Chen B, Khodadoust MS, Liu CL, Newman AM and Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. Methods Mol Biol 2018; 1711: 243-259.

[38] Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, Porta-Pardo E, Gao GF, Plaisier CL, Eddy JA, Ziv E, Culhane AC, Paull EO, Sivakumar IKA, Gentles AJ, Malhotra R, Farshidfar F, Colaprico A, Parker JS, Mose LE, Vo NS, Liu J, Liu Y, Rader J, Dhankani V, Reynolds SM, Bowlby R, Califano A, Cherniack AD, Anastassiou D, Bedognetti D, Rao A, Chen K, Krasnitz A, Hu H, Malta TM, Noushmehr H, Pedamallu CS, Bullman S, Ojesina AI, Lamb A, Zhou W, Shen H, Choueiri TK, Weinstein JN, Guinney J, Saltz J, Holt RA, Rabkin CE, Lazar AJ, Serody JS, Demicco EG, Disis ML, Vincent BG

and Shmulevich L. The immune landscape of cancer. Immunity 2018; 48: 812-830, e814.
[39] Rooney MS, Shukla SA, Wu CJ, Getz G and Hacohen N. Molecular and genetic properties of tumors associated with local immune cytolytic activity. Cell 2015; 160: 48-61.
[40] Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, Liu Y, Fan H, Shen H, Ravikumar V, Rao A, Schultz A, Li X, Sumazin P, Williams C, Mestdagh P, Gunaratne PH, Yau C, Bowlby R, Robertson AG, Tiezzi DG, Wang C, Cherniack AD, Godwin AK, Kuderer NM, Rader JS, Zuna RE, Sood AK, Lazar AJ, Ojesina AI, Adebamowo C, Adebamowo SN, Baggerly KA, Chen TW, Chiu HS, Lefever S, Liu L, MacKenzie K, Orsulic S, Roszik J, Shelley CS, Song Q, Vellano CP, Wentzensen N, Weinstein JN, Mills GB, Levine DA and Akbani R. A comprehensive Pan-cancer molecular study of gynecologic and breast cancers. Cancer Cell 2018; 33: 690-705, e699.
[41] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C and Schultz N. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012; 2: 401-404.
[42] Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP and Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 2015; 1: 417-425.
[43] Harris RS. Molecular mechanism and clinical impact of APOBEC3B-catalyzed mutagenesis in breast cancer. Breast Cancer Res 2015; 17: 8.
[44] Burns MB, Temiz NA and Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat Genet 2013; 45: 977-983.
[45] Saal LH, Vallon-Christersson J, Hakkinen J, Hegardt C, Grabau D, Winter C, Brueffer C, Tang MH, Reutersward C, Schulz R, Karlsson A, Ehinger A, Malina J, Manjer J, Malmberg M, Larsson C, Ryden L, Loman N and Borg A. The Sweden Cancerome Analysis Network - Breast (SCAN-B) initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Med 2015; 7: 20.
[46] Brown WL, Law EK, Argyris PP, Carpenter MA, Levin-Klein R, Ranum AN, Molan AM, Forster CL, Anderson BD, Lackey L and Harris RS. A rabbit monoclonal antibody against the antiviral and cancer genomic DNA mutating enzyme APOBEC3B. Antibodies (Basel) 2019; 8: 47.
[47] Conticello SG, Thomas CJ, Petersen-Mahrt SK and Neuberger MS. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. Mol Biol Evol 2005; 22: 367-377.
[48] Starrett GJ, Luengas EM, McCann JL, Ebrahimi D, Temiz NA, Love RP, Feng Y, Adolph MB, Chelico L, Law EK, Carpenter MA and Harris RS. The DNA cytosine deaminase APOBEC3H haplotype I likely contributes to breast and lung cancer mutagenesis. Nat Commun 2016; 7: 12918.
[49] Pan JW, Zabidi MMA, Chong BK, Meng MY, Ng PS, Hasan SN, Sandey B, Bahnu S, Rajadurai P, Yip CH, Rueda OM, Caldas C, Chin SF and Teo SH. Germline APOBEC3B deletion increases somatic hypermutation in Asian breast cancer that is associated with Her2 subtype, PIK-3CA mutations, and immune activation. Int J Cancer 2021; [Epub ahead of print].
[50] Leonard B, Starrett GJ, Maurer MJ, Oberg AL, Van Bockstal M, Van Dorpe J, De Wever O, Helleman J, Sieuwerts AM, Berns EM, Martens JW, Anderson BD, Brown WL, Kalli KR, Kaufmann SH and Harris RS. APOBEC3G expression correlates with T-cell infiltration and improved clinical outcomes in high-grade serous ovarian carcinoma. Clin Cancer Res 2016; 22: 4746-4755.
[51] Willems L and Gillet NA. APOBEC3 interference during replication of viral genomes. Viruses 2015; 7: 2999-3018.

Table S1. Characteristics of examined breast cancer patient cohorts[a]

	KM Plotter	*SCAN-B*	*TCGA-BRCA*
Patients	5,134	3,273	1,097
Country of patient care		Sweden	USA (> 90%)
Period of diagnoses		2010-2015	1978-2013[b]
Age (years) at diagnosis (mean ± SD)		62.7 ± 13.1	48.4 ± 13.2
Race[c] (%)			
African			18.1
Asian			6.2
Caucasian			75.3
Non-surgical treatment (%)			
Chemotherapy	42.2	39.8	77.4[d]
Hormone therapy	44.6	78.2	
Radiation			44.2
Follow-up (months) from diagnosis (mean ± SD)	86.6 ± 47.3	73.0 ± 20.1	41.6 ± 39.9
Deaths (%)	24.2	14.5	13.9
Estrogen receptor-positive (%)	67.9	92.2	77.2
Progesterone receptor-positive (%)	48.1	86.9	67.0
HER2-positive (%)	22.2	13.2	19.5
Lymph node-positive[e] (%)	39.2	36.6	52.9
Nottingham histological grade (%)			
1	14.9	15.3	13.3
2	42.3	47.9	46.3
3	42.8	36.8	40.4
PAM50 subtype (%)			
Basal	17.1	9.9	17.7
HER2-enriched	6.5	8.7	7.6
Luminal A	48.7	48.0	51.8
Luminal B	27.7	27.9	19.3
Normal-like		3.5	3.7
Platform for tumor gene expression profiling	Affymetrix microarray	Illumina RNA sequencing	Illumina RNA sequencing

[a]Data for some characteristics or molecular assays is unavailable or incomplete for some cohorts. *HER2*, human epidermal growth factor receptor 2; *KM Plotter*, Kaplan-Meier Plotter meta-cohort of 35 patient and tumor transcriptome datasets (2016.10.13 update); *PAM50*, Prediction Analysis of Microarray 50; *SCAN-B*, Sweden Cancerome Analysis Network-Breast; *TCGA-BRCA*, The Cancer Genome Atlas, Breast Carcinoma. [b]For all TCGA patients including non-BRCA ones. [c]African and Asian designations include mixed races with these components. [d]Chemotherapy and/or hormone therapy. eN stage > 0.

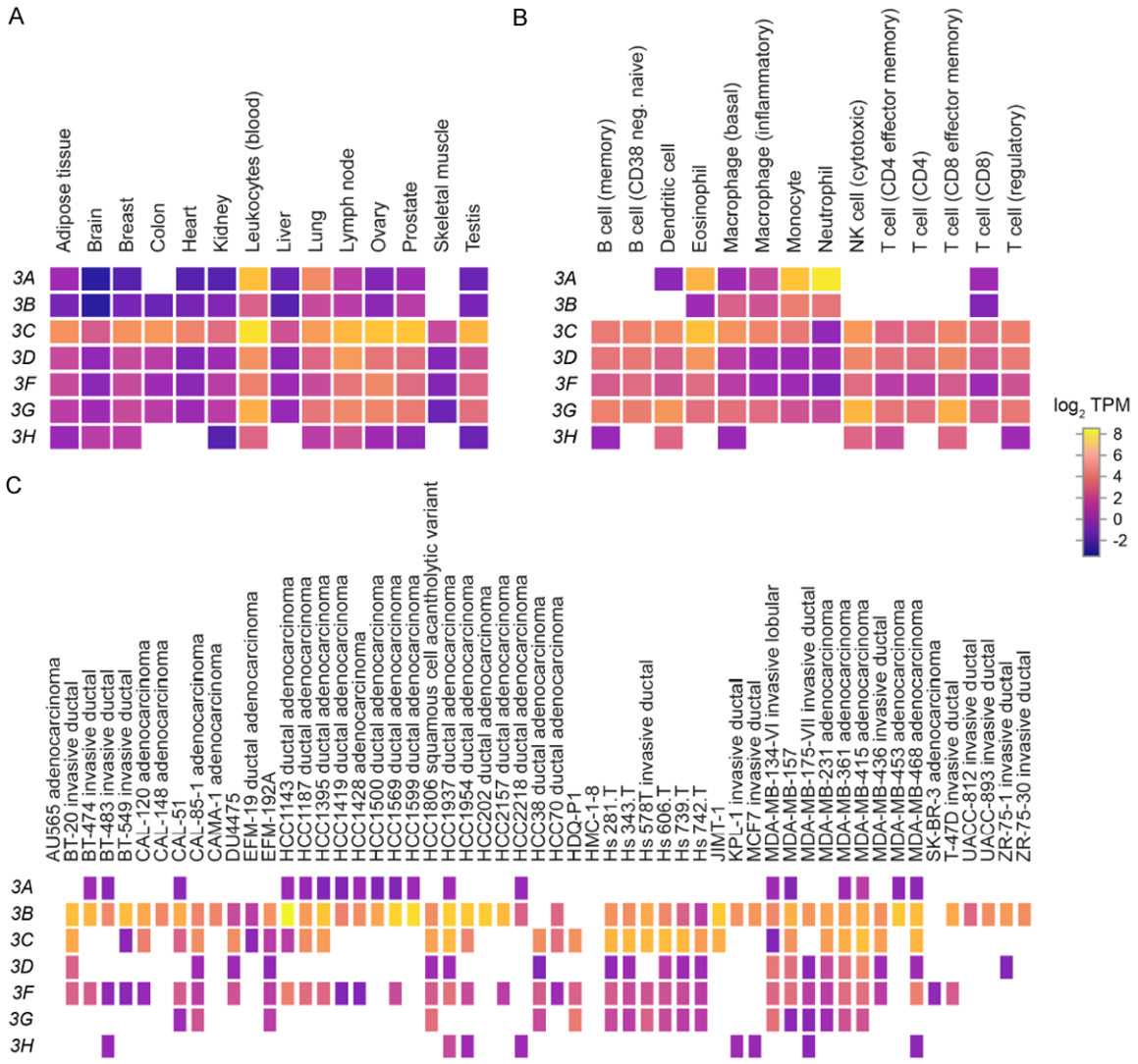

Figure S1. *APOBEC3 gene expression in normal human tissues and breast cancer cell-lines*. Heatmaps, with the same color scale, are shown for average gene expression in (A) 16 human tissues of the Illumina Human Body Map, (B) 14 peripheral leukocyte subsets of the Blueprint Epigenome, and (C) 55 human breast cancer cell-lines of the Cancer Cell Line Encyclopedia projects. *TPM*, transcripts per million.

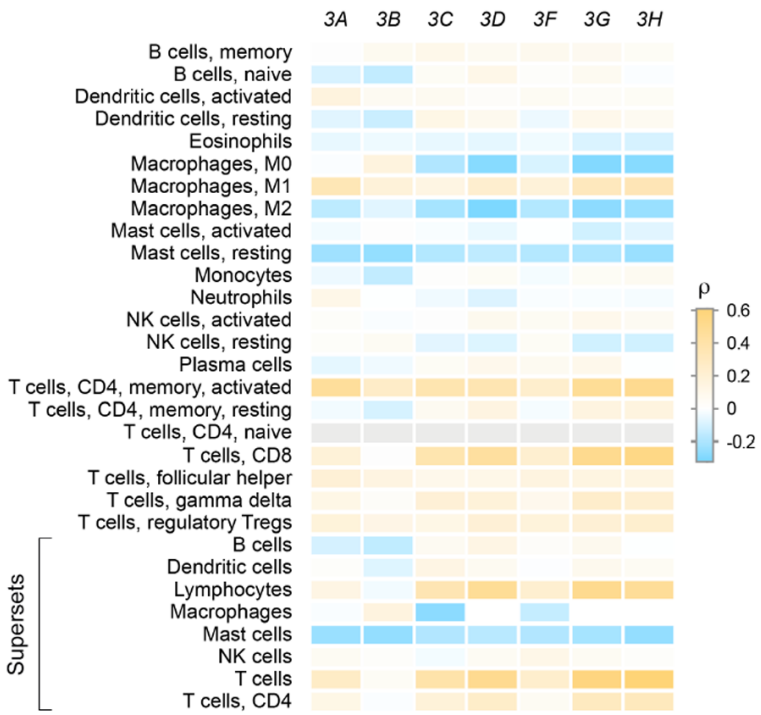

Figure S2. *Correlations of TCGA breast cancer tumor APOBEC3 gene expression with relative abundances of tumor-infiltrating immune cell types*. Heatmap of Spearman correlation coefficients (r) is shown for the 915 tumors for which abundance values were determined for 22 types of cells by CIBERSORT method with p < 0.05. Abundance of $CD4^+$ naive T cells was 0 for all tumors. Correlations for supersets of cell types are also plotted.

Table S2. Multivariable Cox proportional-hazards regression analyses of TCGA breast cancer tumor *APOBEC3* gene expression and cancer progression and mortality[a]

Variable	*Value*	*Disease-specific survival*			*Progression-free survival*		
		LRT[b] *p*	*Wald p*	*HR (95% CI)*	*LRT p*	*Wald p*	*HR (95% CI)*
Demographic and pathologic variables-Univariable analyses							
Age at diagnosis[c]		0.165	0.163	1.01 (1.00-1.03)	0.635	0.635	1.00 (0.99-1.02)
Race (vs. Caucasian)	African	0.709	0.450	1.23 (0.72-2.11)	0.730	0.463	1.17 (0.77-1.78)
	Asian		0.647	1.31 (0.41-4.21)		0.690	1.20 (0.49-2.96)
Tumor grade (vs. I)	II	*0.045*	0.385	1.95 (0.43-8.81)	*0.026*	0.291	1.78 (0.61-5.19)
	III		0.068	3.88 (0.9-16.71)		*0.035*	3.06 (1.08-8.69)
pN (vs. negative)	Positive	*2.52E-07*	*2.79E-06*	3.63 (2.12-6.21)	*5.73E-06*	*1.27E-05*	2.26 (1.57-3.26)
ER (vs. negative)	Positive	*0.011*	*0.008*	0.53 (0.33-0.85)	*0.007*	*0.005*	0.60 (0.42-0.86)
PR (vs. negative)	Positive	*0.004*	*0.004*	0.51 (0.33-0.80)	*0.001*	*0.001*	0.55 (0.40-0.78)
HER2 (vs. negative)	Positive	0.887	0.888	0.95 (0.46-1.94)	0.741	0.743	0.92 (0.55-1.54)
Tumor APOBEC3 gene expression-Multivariable analyses[d]							
APOBEC3A (vs. low)	High	*0.001*	0.894	0.93 (0.35-2.52)	*0.020*	0.970	0.99 (0.46-2.12)
APOBEC3B (vs. low)	High	*0.007*	0.831	0.90 (0.35-2.34)	*0.015*	0.866	0.94 (0.44-2.01)
APOBEC3C (vs. low)	High	*0.011*	0.336	0.65 (0.26-1.58)	*0.016*	0.257	0.67 (0.34-1.33)
APOBEC3D (vs. low)	High	*0.010*	0.355	0.63 (0.23-1.69)	0.090	0.383	0.73 (0.36-1.49)
APOBEC3F (vs. low)	High	0.143	0.424	0.71 (0.31-1.64)	0.261	0.432	0.78 (0.41-1.46)
APOBEC3G (vs. low)	High	0.086	0.347	0.61 (0.22-1.69)	0.118	0.189	0.62 (0.30-1.27)
APOBEC3H (vs. low)	High	0.182	0.565	0.76 (0.30-1.94)	0.205	0.743	0.89 (0.45-1.78)

[a]*P* values below 0.05 are italicized. *CI*, confidence interval; ER, estrogen receptor; *HER2*, human epidermal growth factor receptor 2; *HR*, hazard ratio; *LRT*, likelihood ratio test; *pN*, pathologic lymph node status; *PR*, progesterone receptor. [b]Full model's *p* value. [c]Continuous value. [d]Single *APOBEC3* gene expression categorized as high or low based on within-cohort tertile grouping (top or bottom) in a model with tumor Nottingham histological grade, and pN, ER, and PR status (with LRT p < 0.05 in univariable analyses) as other variables. Proportional hazard assumptions hold true (p > 0.05) for all variables except for PR in analyses of disease-specific survival in cases of *APOBEC3C*, *3F*, and *3G* (p = 0.01-0.04).

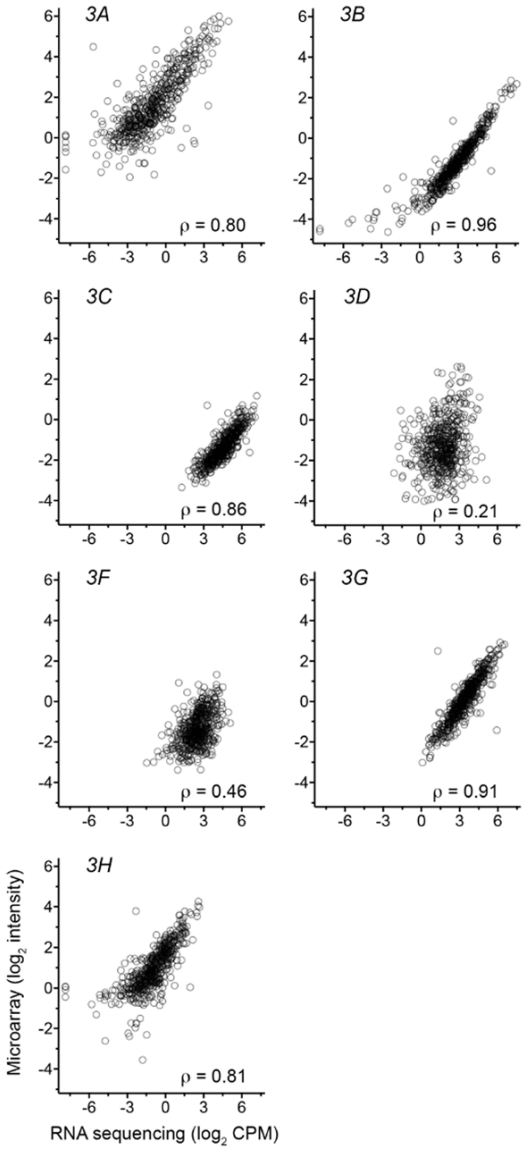

Figure S3. *Correlation between RNA sequencing- and Agilent microarray-based APOBEC3 gene expression measurements of TCGA breast cancer tumors.* Log2-transformed normalized measurements are plotted for the 523 tumors whose RNAs (from fresh-frozen samples) were assayed by both sequencing and microarray. Spearman correlation coefficient values (r) are shown. Measurements for microarray are from Broad Institute Firebrowse (version 1.1.40), while generation of the RNA sequencing-based measurements are described elsewhere in this publication. *CPM*, count per million.