Award Number: W81XWH-11-1-0014

TITLE: Mutation of Breast Cancer Cell Genomic DNA by APOBEC3B

PRINCIPAL INVESTIGATOR: Michael Bradley Burns

CONTRACTING ORGANIZATION: University of Minnesota
Minneapolis, MN 55455

REPORT DATE: September 2012

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE
1 September 2012

2. REPORT TYPE
ANNUAL SUMMARY

3. DATES COVERED
01 Sep 2012 – 31 Aug 2012

4. TITLE AND SUBTITLE
Mutation of Breast Cancer Cell Genomic DNA by APOBEC3B

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-11-1-0014

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Michael Burns

E-Mail: burn0230@umn.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Regents of the University of Minnesota

MINNEAPOLIS, MN 55455

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Many breast cancers have somatic mutation spectra dominated by C-to-T transitions[1-9]. In the course of my training, funded by this grant, I discovered that the DNA cytosine deaminase APOBEC3B (A3B) likely generates a substantial portion of these mutations. *A3B* mRNA is up-regulated in the majority of primary breast tumors and breast cancer cell lines. Endogenous A3B protein is predominantly nuclear and is the only detectable source of DNA C-to-U editing activity in breast cancer cell line extracts. Knockdown experiments show that endogenous A3B is responsible for elevated levels of genomic uracil, increased mutation frequencies, and C-to-T transitions. Furthermore, induced A3B over-expression causes cell cycle deviations, cell death, DNA fragmentation, gamma-H2AX accumulation, and C-to-T mutations. The preferred deamination signature of recombinant A3B explains a major proportion (~20%) of the entire breast cancer base substitution mutation load. These data suggest a model in which A3B-catalyzed deamination provides a chronic source of DNA damage in breast cancer that explains how some cancers evolve rapidly and manifest gross molecular and clinical heterogeneity.

15. SUBJECT TERMS
Breast cancer, APOBEC3B, Mutation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	61	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Introduction:

Cancer as a disease is defined by mutation. Without mutation, cancer cannot occur. In the case of breast cancer, while there are a few known causative agents in sub-types of the malignancy, the source of the molecular and clinical heterogeneity remains a mystery. The purpose of this research is to determine the impact of the endogenous DNA mutating enzyme, APOBEC3B (A3B), in human breast cancer. Thus far, I have generated sufficient data to clearly demonstrate that A3B can cause DNA mutations in cell culture models and, by mining the publicly available datasets, that A3B up-regulation is directly correlated with increased mutation frequency in breast cancer patients. These findings have been collected in a manuscript and are currently in review prior to publication.

Body:

Over the course of this training grant period, several of the specific aims were addressed. Aim 1 was completed by culturing 46 breast cell lines (cancerous lines as well as normal-like control lines) and profiled by qRT-PCR for APOBEC gene expression. The results of these tests can be found in the appended manuscript (**Fig. 1a**, **Fig. S2**, and **Table S1**). The finding is described in detail in the manuscript, but the general point is that A3B is found significantly over-expressed in breast cancer cell lines, but not in normal-like controls. This finding is in line with the data presented in the original grant application, with the added benefit of the new cell lines (45/46) having been procured directly from ATCC to ensure their identities and origins. This new finding makes clear the distinction that A3B, among all 11 APOBEC family members, is the only one that is consistently up-regulated in breast cancer cell lines.

In addition to the cell line work, I acquired 52 matched breast cancer and normal samples as well as 28 reduction mammoplasty samples in order to profile A3B levels in primary patient tissues. The findings are, again, described in detail in the manuscript (see **Fig. 4**, **Fig. S8**, and **Table S2**) with the major finding being that, as with cell lines, the up-regulation of A3B is significantly associated with breast cancer samples when compared to patient-matched normal tissue and is not seen in otherwise normal reduction mammoplasty samples.

Difficulties were encountered when attempting to determine the protein levels of A3B in breast cancer cell lines and tissues. This is due to the high sequence homology found among the different APOBEC family members at the nucleotide and amino acid levels. There are currently no antibodies available commercially (despite the manufacturers' claims) or academically that are capable of specifically detecting endogenously expressed A3B. In order to address this critical pitfall, I utilized a enzyme activity assay in conjunction with A3B mRNA knock-down in order to assess the levels of active A3B present in breast cancer cell lines. This allowed me to discover (along with fluorescent microscopy of transiently transfected A3B-GFP) that A3B is localized to the nucleus of breast cancer cells and is the only source of C-to-U deamination activity in these cells (see **Fig. 1c-1f**, **Fig. S5**, and **Fig. S6**). In other words, this combination of techniques, used in lieu of Western blotting, allowed me to determine not just that the enzyme was actually translated into protein, but that it is localized to the nucleus and catalytically active.

The work described above have been compiled as part of a much larger manuscript and are following addressing of reviewers' comments is again in review at *Nature*. Publication of this work will signal completion of MILESTONE 1. A final decision on the manuscript is anticipated to occur within the next 2 weeks (01 Oct 2012 – 14 Oct 2012).

As follow-up to reviewers' comments, the next set of experiments focused entirely on demonstrating that A3B up-regulation in breast cancer 1) *can* damage genomic DNA and 2) *does* correlate with mutation load and mutational signature in patient samples. These experiments were essential for publication of this work in a high-impact journal and clearly demonstrate the significance of this research (*i.e.* this enzyme *can* and *does* generate mutation in breast cancer). To this end, I, along with several Harris lab members and other collaborators (see manuscript author list), worked to ask and answer the following questions (each described in detail in the appended manuscript):

1) Does endogenous up-regulation of A3B in breast cancer cells result in a higher steady-state level of genomic uracil (A3B's enzymatic product)?

2) Does endogenous up-regulation of A3B in breast cancer cells allow them to mutate their genome and escape toxic drug treatment?
3) Can we see higher levels of random mutation generated in the breast cancer genome of cells that express high levels of A3B?
4) Does A3B over-expression result in detectable DNA damage signals?
5) Do publicly available sequencing datasets contain an A3B mutation signature?
6) Do recently released TCGA RNAseq and exome sequencing confirm a mutation load and A3B expression correlation?

Rather than cut and paste the text that addresses each of these questions directly from the manuscript I present here a summary table that includes the question, the technique used, the result, and a reference to the appropriate figure in the appended manuscript.

Table 1. Experimental questions, approaches, and outcomes.

	Question	Technique	Result	Reference
1)	Does A3B generate genomic uracil?	HPLC-MS/MS analysis of breast cancer cell line DNA +/- A3B knock-down	Yes, A3B expression increases the steady-state level of genomic uracil	**Fig. 2a-2c**
2)	Can A3B mutate a target gene to escape drug treatment?	Genetic selection assay using the drug-sensitivity gene *thymidine kinase* (*TK*) and the associated drug, ganciclovir	Yes, A3B expression resulted in a higher frequency of cancer cell escape from ganciclovir treatement.	**Fig. 2d-2f**
3)	Is there an increase in random genomic mutations in cells with high A3B?	3D-PCR and sequencing was performed on several genes from cancer cell lines that endogenously over-express A3B and from the same cells lines that have had A3B knock-down.	Yes, A3B expression resulted in higher numbers of random transition mutations.	**Fig. 2g-2h**
4)	Does A3B over-expression damage genomic DNA?	A3B and catalytically dead A3B were induced under a tetracycycline controlled promoter followed by cell cycle profiling, colony formation assays, gamma-H2AX measurements, comet assays, and 3D-PCR/sequencing.	Each of the assays demonstrated that genomic DNA is damaged (lethally) when catalytically active A3B is over-expressed.	**Fig. 3**
5)	Is A3B's mutation signature present in public datasets?	The mutational kinetics of A3B were characterized and used to determine an A3B signature. Whole-genome and exome sequencing data were mined to see if this identifier was present.	A3B's muation signature was detected in breast cancer samples, but not in liver cancer or melanoma.	**Fig. 4c-4e**
6)	Do A3B expression and mutation load correlate in TCGA datasets?	Mining of recently released (July 2012) breast cancer RNAseq and exome sequencing data at The Cancer Genome Atlas website.	A3B expression and mutation load and signature are correlated in the reported primary breast cancers.	**Fig. S14**

Specific Aim 2 has been greatly advanced by screening several shRNA constructs (pursuant to Aim 1) to generate a more robust knock-down than was demonstrated in **Fig. 2** of the original grant proposal. As can be seen in **Fig. 1d**, **Fig. 2b**, and several others, the new shRNA routinely decreases A3B mRNA by >85%. I have generated subclones (>3/per line per condition) of cell lines HCC1569, MDA-MB-453, and MDA-MB-468 with either control shRNA or

A3B-knock-down shRNA. These lines will next be stably transduced with firefly luciferase and prepared for xenograft experiments.

As is evident from the results involving artificial over-expression of A3B in Aim 1, generating MCF-10A cells that stably express A3B is problematic. We have not been able to generate stable clones that constitutively express active A3B in breast cancer cell lines. We have, however, generated cell lines that express A3B in a tetracycline-inducible system. This experiment subsequently allowed us to see that the reason we were unable to generate constitutively expressing clones is because massive over-expression of A3B is genotoxic to cells. This also causes issues when attempting to grow cells in soft agar since even in an inducible system, the expression will kill cells expressing A3B within 10 days (Burns and Lackey, data not shown). To address this concern, I have generated MCF-10A clones that express the tet repressor and am engineering a weak CMV promoter to replace the original CMV promoter in the tet-responsive plasmid. This is anticipated to alleviate some of the toxic effects of A3B expression by decreasing the "dose," as it were, of the deaminase.

Specific Aim 4 is likewise well underway. The transgene was successfully constructed and used to generate transgenic mice (IACUC protocol# 1002A77540). Our current mouse colony for this project contains 55 mice descended from the A3B founder. Once the mouse colony has reached a large enough size to ensure that the line is stable and matches the power analysis described in the IACUC protocol, we will cross A3B mice to commercially available MMTV-cre mice.

As this project moves forward, there are still unanswered questions. Aims 2 and 3 are well-underway and are anticipated to produce meaningful results in line with the successful outcome of Aim 1. A major gap in our understanding of A3B's role in breast cancer is what is driving its up-regulation in tumors. To that end, I will be recommending a change in the SOW formally (following submission of this Annual Report) to add an ongoing research focus on addressing what the mechanism is that drives A3B expression. This is a critical question to answer since, once the mechanism is understood, we may be able to target it to decrease the mutational capacity of A3B-high breast cancers and potentially decrease the chance that they will escape therapy.

Key Research Accomplishments:

- Definitive demonstration that A3B is up-regulated in breast cancer cell lines and primary tumors
- Discovery that A3B is not only expressed in the nucleus of breast cancer cells, but is also catalytically active
- Discovery that A3B up-regulation drives mutation in breast cancer cell lines
- Finding that publicly available datasets show:
    - A3B mutation signature in breast cancer
    - A3B expression correlated with mutation load in breast cancer

Reportable Outcomes:

Over the course of this grant period, this research has resulted in a manuscript (appended to this report - currently in review at *Nature*), several clonal breast cancer cell lines (A3B-knock down and control), MCF-10A tet-repressor, a cell system that can express A3B or catalytically dead A3B under control of the tet repressor, and *TK*-containing MDA-MB-453 and HCC1569 cells containing either control shRNA or A3B knockdown shRNA. Additionally, a mouse model for the study of A3B was generated.

References:

(see appended manuscript)

Appendices:

The appendices include the full version of the most recent manuscript (2nd revision 07 Sep 2012) and addendum (14 Sep 2012) as sent to *Nature* for re-review.

APOBEC3B is an enzymatic source of mutation in breast cancer

Michael B. Burns[1-4,*], Lela Lackey[1-4,*], Michael A. Carpenter[1-4], Anurag Rathore[1-4], Allison M. Land[1-4], Brandon Leonard[2-5], Eric W. Refsland[1-4], Delshanee Kotandeniya[2,6], Natalia Tretyakova[2,6], Jason B. Nikas[2], Douglas Yee[2], Nuri A. Temiz[7], Duncan E. Donohue[7], Rebecca M. McDougle[1-4], William L. Brown[1-4], Emily K. Law[1-4] & Reuben S. Harris[1-5,#]

[1]Biochemistry, Molecular Biology and Biophysics Department, [2]Masonic Cancer Center, [3]Institute for Molecular Virology, [4]Center for Genome Engineering, [5]Microbiology, Cancer Biology and Immunology Graduate Program, [6]Department of Medicinal Chemistry,
University of Minnesota, Minneapolis, MN 55455, USA
[7]*In Silico* Research Centers of Excellence, Advanced Biomedical Computing Center, Information Systems Program, SAIC-Frederick Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 21702 USA

[*] Equal primary contributions.

[#] Correspondence to R.S.H. (rsh@umn.edu).

250 word abstract, 1916 word main text, 4 figures, and 30 references. Supplementary online information consists of discussion, methods, 11 tables, and 13 figures.

Multiple mutations are required for cancer development, and genome sequencing has revealed that several cancers, including breast, have somatic mutation spectra dominated by C-to-T transitions[1-9]. The majority of these mutations are unlikely to be spontaneous because they occur at hydrolytically disfavored[10] non-methylated cytosines throughout the genome, and they are sometimes clustered (kataegis)[8]. Here, we show that the DNA cytosine deaminase APOBEC3B (A3B) is the most likely source of these mutational events. ***A3B*** **mRNA is up-regulated in the majority of primary breast tumors and breast cancer cell lines, but only detected at low levels in normal breast tissues and mammary epithelial cell lines. Endogenous A3B protein is predominantly nuclear and the only detectable source of DNA C-to-U editing activity in breast cancer cell line extracts. Knockdown experiments show that endogenous A3B is responsible for elevated levels of genomic uracil, increased mutation frequencies, and C-to-T transitions. Furthermore, induced A3B over-expression causes cell cycle deviations, cell death, DNA fragmentation, γ-H2AX accumulation, and C-to-T mutations. The preferred deamination signature of recombinant A3B explains a major proportion (~20%) of the entire breast cancer base substitution mutation load. Our data suggest a model in which A3B-catalyzed deamination provides a chronic source of DNA damage in breast cancer that helps explain how some cancers evolve rapidly and manifest gross molecular and clinical heterogeneity. As a non-essential protein[11] with possible overlapping function in innate immunity, A3B may be an attractive marker and therapeutic target for breast cancer and, possibly, non-mammary neoplasms where C-to-T mutations and kataegis also manifest[1,2,4-7,12].**

Spontaneous hydrolytic deamination of DNA cytosine to uracil (C-to-U) or methyl-cytosine to thymine is a well-established pro-mutagenic process, which, by DNA replication or misrepair, can result in C-to-T transition mutations[13]. Interestingly, methyl-cytosine is 4.5-fold more labile than normal cytosine[10], yet in many cancers, such as breast, only a minority of mutations occur within methyl-CpG contexts[1,3,9]. This incongruence is compounded by reports of localized TC-to-TT biased mutation clusters in breast[8] and other neoplasms[12]. These observations strongly suggest the existence of an underlying non-spontaneous mechanism.

Most humans encode a total of eleven polynucleotide cytosine deaminase family members that might contribute to mutation in cancer – APOBEC1, activation-induced deaminase (AID), APOBEC2, APOBEC3s (A, B, C, D, F, G, and H), and APOBEC4. APOBEC2 and APOBEC4 have not shown activity. APOBEC1 and AID are confined developmentally and implicated in cancers of those tissues, liver and B lymphocytes, respectively[14,15]. We therefore hypothesized that one or more of the seven APOBEC3s may be responsible for C-to-T mutation in other human cancers. This possibility is supported by hybridization experiments suggesting *APOBEC3* up-regulation in some cancers, such as breast[16], and qPCR data demonstrating broad *APOBEC3* expression profiles[17] (**Fig. S1**).

To identify the contributing APOBEC3, we quantified mRNA levels for each of the 11 family members in a panel of breast cancer cell lines (**Fig. S2**). Surprisingly, only *APOBEC3B* (*A3B*) mRNA trended toward up-regulation. This analysis was expanded to include a total of 38 independent breast cancer cell lines. *A3B* was up-regulated by $\geq$3 s.d. relative to controls in 28/38 lines, with levels exceeding 10-fold in 12/38 lines (**Fig. 1a & Table S1**). MDA-MB-453, MDA-MB-468, and HCC1569, representative lines used below, showed 20-, 21-, and 61-fold up-regulation, respectively. These results correlate with cell line microarray data (p=0.027; **Fig. 1b & Supplementary Discussion**). No positive correlation was evident for any other deaminase family member (**Fig. S2**). *A3B* up-regulation is most likely due to an upstream signal transduction event because it is not a frequent site of rearrangement or copy number variation (http://dbCRID.biolead.org) and sequencing failed to reveal promoter activating mutations or CpG islands indicative of epigenetic regulation (**Fig. S3**).

Epitope-tagged A3B localizes to the nucleus of several transfected cell types[18]. To ask whether this is also a property of breast cancer cell lines, an A3B-eGFP construct was transfected into MDA-MB-453, MDA-MB-468, and HCC1569. Live cell images of A3B-eGFP showed nuclear localization, in contrast to the cytoplasmic A3F-eGFP or A3G-eGFP (**Fig. 1c & S4**). Corroborating data were obtained for HA-tagged proteins in fixed breast cancer cell lines (**Fig. S4**). To study *endogenous* A3B activity and subcellular compartmentalization, we used a fluorescence-based DNA C-to-U assay. We first found that nuclear fractions of several breast cancer cell lines contain DNA editing activity, which could be ablated by shRNA knockdown of

A3B (**Fig. 1d-e & S5**). Identical results were obtained with an independent shA3B construct (not shown). Protein extracts were then used to assess endogenous A3B's local dinucleotide deamination preference. Similar to retroviral hypermutation signatures caused by A3B over-expression[19], endogenous A3B showed a strong preference for editing cytosines in the TC dinucleotide context (**Fig. 1f**). No deaminase activity was observed for extracts from MCF10A (A3B-low epithelial line) or SK-BR-3 (A3B-null cancer line), although it could be conferred by transfecting an A3B expression construct (**Fig. 1f & S6**). Only A3B-eGFP and A3A-eGFP elicited measurable TC-to-TU activity in lysates from transfected HEK293T cells (**Fig. S7**). Since *A3A* is myeloid lineage-specific[20] and non-detectable in breast cancer cell lines (**Figs. S1 & S2**), our studies demonstrate that, of the entire APOBEC/AID family, A3B is the only enzyme appropriately positioned to deaminate breast cancer genomic DNA.

To address whether endogenous A3B damages genomic DNA, we employed a combination of biophysical and genetic assays. We first used a mass spectrometry-based approach to quantify levels of genomic uracil in MDA-MB-453 and HCC1569 with high levels of endogenous A3B (shControl) versus knock-down levels of A3B (shA3B) (**Fig. 2a**). Genomic uracil loads decreased by 40% in HCC1569 expressing shA3B, corresponding with weaker knockdown, and by 70% in MDA-MB-453 expressing shA3B, which had stronger knockdown (**Fig. 2b-c**). Although these relative differences may seem modest, 20 and 10 uracils per Mbp, respectively, this equates to approximately 60,000 and 30,000 A3B-dependent uracils per haploid genome. These values may underestimate the actual number of A3B-catalyzed pro-mutagenic lesions because mismatch repair and several base excision repair pathways undoubtedly work to counteract this damage.

Second, we used a thymidine kinase-positive (TK^{plus}) to TK^{minus} fluctuation analysis[20] to determine whether up-regulated *A3B* and elevated uracil loads lead to higher levels of mutation (**Fig. 2d**). MDA-MB-453 and HCC1569 cells were engineered to express herpes simplex virus *TK*, which confers sensitivity to the drug ganciclovir. TK^{plus} lines were transduced with shA3B or shControl constructs and limiting dilution was used to generate single cell shA3B and shControl sub-clones, respectively (**Fig. 2e**). Expanded sub-clones were subjected to ganciclovir selection, and resistant cells were grown to visible colonies. Colony counts enabled the median mutation frequencies to be determined, which revealed that cells with up-regulated A3B accumulate 3-to-5-fold more mutations (**Fig. 2f**).

Third, 3D-PCR[20,21] was used to ask whether C/G-to-T/A transition mutations accumulate at three genomic loci from cells transduced with shA3B and shControl viruses resulting in A3B^{low} and A3B^{high} pools of HCC1569 cells. This technique enables qualitative estimates of genomic mutation within a population of cells because DNA sequences with higher T/A content amplify at lower denaturation temperatures than parental sequences. Lower temperature amplicons were detected for *TP53* and *c-MYC*, but not *CDKN2B* (**Fig. 2g**). Individual low temperature amplicons were cloned and sequenced, and more C/G-to-T/A transition mutations were observed in A3B^{high} in comparison to A3B^{low} samples (**Fig. 2h**). Other types of base substitution mutations were rare. Some C/G-to-T/A transitions were still evident in the A3B^{low} samples, possibly due to residual deaminase activity and/or amplification of spontaneous events.

Next, we asked whether A3B triggers additional hallmarks of cancer[22]. We first tried and failed to stably express A3B in MCF10A and HEK293 cells. To circumvent toxicity, we constructed a panel of HEK293 clones with doxycycline (Dox)-inducible A3B, A3B-E68A-E255Q, A3A, or A3A-E72A eGFP fusions. As measured by flow cytometry, A3-eGFP levels were barely detectable without Dox and induced in nearly 100% of cells with Dox (**Fig. 3a**). A3A over-expression caused rapid S-phase arrest, cytotoxicity, and γ-H2AX focus formation, as reported[23] (**Fig. 3b-f**). In comparison, A3B induction caused a delayed cell cycle arrest, a more pronounced formation of abnormal anucleate and multinucleate cells, and eventual cell death (**Fig. 3b-e**). A3B induction also caused γ-H2AX focus formation, DNA fragmentation, as evidenced by visible comets, and C/G-to-T/A mutations (**Fig. 3e-h**).

We next asked whether our cell-based results could be extended to primary tumors. First, we quantified mRNA levels for each of the 11 family members in 21 randomly chosen breast tumor specimens, in parallel with matched normal tissue procured simultaneously from an adjacent area or the contralateral breast. Only *A3B* was expressed preferentially in tumors (p=0.0003) (**Fig. S8**). We confirmed this analysis by measuring *A3B* levels in 31 additional tumor/normal matched tissue sets. In total, *A3B* was up-regulated by ≥3 s.d. in 20/52 tumors in comparison to the patient-matched normal tissue mean, and in 44/52 tumors in comparison to the reduction mammoplasty tissue mean (**Fig. 4a**, p=7.1x10^{-7} and p=2x10^{-5}; **Table S2** for patient information). These values, though highly significant and broadly reflected by microarray data (**Supplementary Discussion, Fig. S9 & Tables S3-S10**), are underestimates because tumor specimens have varying fractions of non-*A3B* expressing normal cells. Some of the matched 'normal' samples may also be contaminated by tumor cells, as judged by comparisons to mean levels in mammoplasty samples (**Fig. 4a**; p=0.002). The related deaminase, *A3G*, was not expressed differentially in the same tumor panel, indicating that these observations are not due to immune cells known to express multiple A3s[17] (**Fig. 4b**; p=0.591).

Finally, we determined the impact of A3B on the breast tumor genome by correlating recombinant A3B's deamination signature *in vitro* and the somatic mutation spectra accumulated during tumor development *in vivo*. Using a series of single-stranded DNA substrates varying only at the immediate 5' or 3' position relative to the target cytosine (underlined), we found that recombinant A3B prefers T$\underline{C}$>C$\underline{C}$>G$\underline{C}$=A$\underline{C}$ (**Fig. S10**; similar to endogenous A3B in **Fig. 1f**) and $\underline{C}$A=$\underline{C}$G=$\underline{C}$T>$\underline{C}$C (**Fig. 4c**). These local sequence contexts were then compared to those for C-to-T transitions reported for breast[8,9], melanoma[24], liver[25], and lung[26] tumors. Consistent with non-spontaneous origins, C-to-T transition loads are much greater in melanoma (~80%) and breast (~40%) than liver (~20%) and lung (~20%) tumors (**Fig. 4d**). The local sequence contexts for C-to-T transitions are even more striking, with flanking T and C in melanoma (T$\underline{C}$C) and T and A in breast cancer (T$\underline{C}$A) (**Fig. 4e & S11**). The melanoma pattern is expected due to error-prone DNA synthesis (A insertion) opposite UV-induced pyrimidine dimers. In contrast, the preferred context of C-to-T transitions in two independent breast cancer somatic mutation data sets, one including kataegis, mirrors the *in vitro* preference of recombinant A3B (**Fig. 4e & S11**).

Taken together, we conclude that A3B is the only DNA deaminase family member with expression and activity profiles consistent with a role in deaminating the breast cancer genome and causing the reported C-to-T mutation biases and localized kataegis events[1,3,8,9]. Other mutational patterns are evident in breast cancers suggesting additional processes at work[8]. However, some of these other patterns may also be due to A3B activity via further processing by 'repair' enzymes into transitions, transversions, and even DNA breaks that may precipitate larger-scale rearrangements (**Fig. S12**). Future work is needed to understand the regulation of *A3B* and its interplay with other oncogenes and tumor suppressors. For example, a possible mechanistic linkage between *A3B* and *TP53* is suggested by a near-significant correlation (p=0.071) between *A3B* up-regulation and *TP53* inactivation in the ATCC breast cancer cell line panel, whereas other common markers do not correlate (**Tables S1, S2 & Fig. S13**). *TP53* inactivation helps reconcile our observations that many tumor cells and cell lines are able to tolerate up-regulated A3B, whereas other cell lines succumb to toxic genomic DNA damage upon enforced A3B over-expression (*e.g.*, **Fig. 3**).

Although the A3B-dependent mechanism described here may impact the majority of breast cancers, this is not always the case as an *A3B* deletion[11] is homozygous in 2/38 breast cancer cell lines and 1/52 tumors (**Figs. 1 & 4**). This natural variation creates opportunities for clinical studies. For instance, one study reported a *non-significant* negative correlation with *A3B* status and breast cancer in Japan[27], where *A3B* deletions are prevalent[11]. Large cohorts will be needed to address whether *A3B* status and/or expression levels correlate with molecular and clinical features of breast cancer. It will be interesting to ask whether the 2-to-3-fold higher incidence of breast cancer in the United States versus Japan[28] is partly attributable to the differential prevalence of A3B.

We provide the first direct evidence for active involvement of the DNA deaminase A3B in breast cancer. Conceptually supportive of the original mutator hypothesis[29], A3B-catalyzed genomic DNA deamination could provide a major source of genetic fuel for cancer development, metastasis, and even resistance to therapy. We propose that A3B is a dominant underlying factor that contributes to tumor heterogeneity by broadly affecting multiple pathways and phenotypes. A3B represents a new marker for breast cancer and a strong candidate for targeted intervention, especially given its non-essential nature[11]. A3B inhibition in evolving tumors may decrease mutation rates and thereby stabilize the cellular targets of existing therapeutics and anti-cancer immune responses (*e.g.*, A3B inhibitors used in combination with therapies that are frequently undermined by resistance mutations).

METHODS SUMMARY

Flash frozen breast tumor and matched normal tissue pairs were obtained from the University of Minnesota Tissue Procurement Facility. Samples were chosen randomly with breast cancer and available matched normal tissue being the only selection criteria. Mammary reduction samples were used as non-cancer controls. These studies were performed in accordance with IRB guidelines (IRB study number 1003E78700). The breast cancer cell line panel 30-4500K was obtained from the ATCC and cultured as recommended. RNA isolation, cDNA synthesis, and qPCR procedures were performed as reported[17] (**Table S11**). Knockdown and control shRNA constructs were obtained from Open Biosystems. Microscopy, cellular fractionation and deaminase activity assays were done as described[18,20]. Genomic uracil was quantified by treating DNA samples with uracil DNA glycosylase, purifying the nucleobase from the remaining DNA and analyzing the samples by mass spectrometry. The TK assay and 3D PCR have been described and were modified for use with breast cancer cell lines[20]. Dox-inducible cells were obtained from Invitrogen and stables were created with the indicated constructs. These lines were analyzed for cell cycle arrest using propidium iodide staining and cell viability with crystal violet staining and the MTS assay. DNA damage was measured by the comet assay and by flow cytometry and microscopy of cells immunostained for γ-H2AX. Recombinant A3B195-382-mycHis was purified and used for deamination kinetics as described[30] using 5'-ATTATTATTATNCNAATGGATTTATTTATTTATTTATTTATTT-6-FAM (NCA and TCN for 5' and 3' preference experiments, respectively). The somatic single nucleotide mutation frequencies with local sequence contexts were determined from published primary tumor genomes[8,9,24-26]. Potential mechanistic overlap with hydrolytic deamination of 5-methyl-cytosines was avoided by excluding CpG dinucleotides from mutational preference calculations.

AUTHOR CONTRIBUTIONS

M.B.B. worked with R.S.H. on all aspects of experimental design, project management, and manuscript preparation. M.B.B., E.W.R., and B.L. generated mRNA expression profiles, L.L. and E.L. performed microscopy, L.L. and A.R. performed biochemical fractionations and DNA deaminase assays, M.B.B. performed uracil quantifications, A.M.L. performed TK fluctuations, A.R. generated 3D-PCR sequences, and L.L., A.L., A.R., and M.A.C. determined the impact of induced A3B over-expression. M.A.C. performed deaminase assays with recombinant protein, and M.A.C., D.K., and N.T. assisted with UPLC-MS set-up. N.A.T. and D.E.D. contributed bioinformatic analyses. J.B.N. conducted the bioinformatic analysis of microarray data and developed a normalization algorithm for this analysis. R.S.H. drafted the manuscript and all authors contributed to manuscript revisions.

ACKNOWLEDGEMENTS

We thank J. Hultquist for assistance with statistics, M. Li for help with DNA deaminase assays, R. Nzara for promoter sequencing, V. Polunovsky for hTERT-HMEC, V. Simon for an shRNA construct, S. Kaufmann, C. Lange, and D. Largaespada for expert consultation, and the University of Minnesota Masonic Cancer Center for purchasing the ATCC breast cancer panel. Tissues were obtained from the University of Minnesota Tissue Procurement Facility, which is part of BioNet, supported by the Academic Health Center and NIH grants P30 CA77598 (D. Yee), P50 CA101955 (D. Buchsbaum) and KL2 RR033182 (B. Blazar). M.B.B. was supported in part by a Cancer Biology Training Grant (NIH NCI T32 CA009138) and a Department of Defense Breast Cancer Research Program Predoctoral Fellowship (BC101124). L. Lackey was supported in part by an NSF Predoctoral Fellowship and by a position on the Institute for Molecular Virology Training Grant NIH T32 AI083196. A.M. Land was supported by a CIHR Postdoctoral Fellowship. Computational analyses (N.A.T, D.E.D) were supported by federal funds from the National Cancer Institute, National Institutes of Health, CBIIT/caBIG ISRCE yellow task #09-260. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. Harris laboratory operational support was provided in part by the Children's Cancer Research Fund (Minneapolis, MN), the University of Minnesota College of Biological Sciences, and a seed grant from the University of Minnesota Clinical and Translational Science Institute (supported by NIH 1UL1RR033183).

REFERENCES

1 Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158 (2007).

2 Jones, S. *et al.* Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* **330**, 228-231 (2010).

3 Sjöblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-274 (2006).

4 Kumar, A. *et al.* Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proc Natl Acad Sci U S A* **108**, 17087-17092 (2011).

5 Parsons, D. W. *et al.* The genetic landscape of the childhood cancer medulloblastoma. *Science* **331**, 435-439 (2011).

6 Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-220 (2011).

7 Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-1160 (2011).

8 Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993 (2012).

9 Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400-404 (2012).

10 Ehrlich, M., Norris, K. F., Wang, R. Y., Kuo, K. C. & Gehrke, C. W. DNA cytosine methylation and heat-induced deamination. *Biosci Rep* **6**, 387-393 (1986).

11 Kidd, J. M., Newman, T. L., Tuzun, E., Kaul, R. & Eichler, E. E. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS Genet* **3**, e63 (2007).

12 Roberts, S. A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol Cell* **46**, 424-435 (2012).

13 Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709-715 (1993).

14 Pavri, R. & Nussenzweig, M. C. AID targeting in antibody diversity. *Adv Immunol* **110**, 1-26 (2011).

15 Yamanaka, S. *et al.* Apolipoprotein B mRNA-editing protein induces hepatocellular carcinoma and dysplasia in transgenic animals. *Proc Natl Acad Sci U S A* **92**, 8483-8487 (1995).

16 Harris, R. S., Petersen-Mahrt, S. K. & Neuberger, M. S. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol Cell* **10**, 1247-1253 (2002).

17 Refsland, E. W. *et al.* Quantitative profiling of the full *APOBEC3* mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction. *Nucleic Acids Res* **38**, 4274-4284 (2010).

18 Lackey, L. *et al.* APOBEC3B and AID have similar nuclear import mechanisms. *J Mol Biol* **419**, 301-314 (2012).

19 Albin, J. S. & Harris, R. S. Interactions of host APOBEC3 restriction factors with HIV-1 in vivo: implications for therapeutics. *Expert Rev Mol Med* **12**, e4 (2010).

20 Stenglein, M. D., Burns, M. B., Li, M., Lengyel, J. & Harris, R. S. APOBEC3 proteins mediate the clearance of foreign DNA from human cells. *Nat Struct Mol Biol* **17**, 222-229 (2010).

21 Suspène, R. *et al.* Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. *Proc Natl Acad Sci U S A* **108**, 4858-4863 (2011).

22 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674 (2011).

23 Landry, S., Narvaiza, I., Linfesty, D. C. & Weitzman, M. D. APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. *EMBO Rep* **12**, 444-450 (2011).

24 Wei, X. *et al.* Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat Genet* **43**, 442-446 (2011).

25 Zhang, J. *et al.* International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database* **2011** (2011).

26 Lee, W. *et al.* The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**, 473-477 (2010).

27 Komatsu, A., Nagasaki, K., Fujimori, M., Amano, J. & Miki, Y. Identification of novel deletion polymorphisms in breast cancer. *International journal of oncology* **33**, 261-270 (2008).

28 Forouzanfar, M. H. *et al.* Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. *Lancet* **378**, 1461-1484 (2011).

29 Loeb, L. A., Springgate, C. F. & Battula, N. Errors in DNA replication as a basis of malignant changes. *Cancer Res* **34**, 2311-2321 (1974).

30 Carpenter, M. A. *et al.* Methyl- and normal-cytosine deamination by the foreign DNA restriction enzyme APOBEC3A. *J Biol Chem* **in press** (online August 2012).

FIGURE LEGENDS

Figure 1. *A3B* up-regulation and activity in breast cancer cell lines.

a, *A3B* levels in the indicated breast cancer cell lines (red circles, n=40 including two sister pairs SK-BR-3/AU565 and MDA-MB-453/MDA-kb2) and non-cancerous cell lines (blue squares, n=6 including one sister pair MCF10A/F). Each data point is the mean *A3B* level of three qPCR reactions presented relative to mRNA levels of the constitutive housekeeping gene *TBP* (s.d. shown unless smaller than the data point). Data are arranged from lowest to highest *A3B* expression level. Cell lines used in mechanistic studies are labeled.

b, Positive correlation between *A3B* qPCR data and microarray data (n=39; see Supplementary Discussion). Cell lines used in mechanistic studies are labeled.

c, A3B-eGFP (green) co-localizes with nuclear DNA (Hoescht-stained blue), whereas A3F-eGFP is cytoplasmic, in the indicated breast cancer cell lines. MDA-MB-468 shows some cytoplasmic A3B-eGFP localization, but is still predominantly nuclear.

d, *A3B* mRNA levels in the indicated breast cancer cell lines stably transduced with shControl or shA3B lentiviruses.

e, Nuclear DNA C-to-U activity in extracts from the indicated breast cancer cell lines transduced as in (d) (n=3). s.d. shown unless smaller than data point.

f, Intrinsic dinucleotide DNA deamination preference of enzyme(s) in soluble extracts from the indicated cell lines (n=3; s.d. is smaller than each data point).

Figure 2. A3B-dependent uracil lesions and mutations in breast cancer genomic DNA.

a, Workflow for genomic uracil quantification by UPLC-MS.

b, *A3B* mRNA levels in the indicated breast cancer cell lines stably transduced with shControl or shA3B lentiviruses.

c, Steady-state genomic uracil loads per mega-basepair (Mbp) in the indicated breast cancer cell lines expressing shControl or shA3B constructs.

d, Workflow for TK fluctuation analysis.

e, *A3B* mRNA levels in TK^{plus} MDA-MB-453 and HCC1569 lines expressing shControl or shA3B constructs.

f, Dot plots depicting the TK^{minus} mutation frequencies of MDA-MB-453 and HCC1569 subclones expressing shControl or shA3B constructs. Each dot corresponds to one subclone, and *median* values are indicated for each condition.

g, Agarose gel analysis of 3D-PCR amplicons obtained using primers specific for the indicated target genes and genomic DNA prepared from HCC1569 cells expressing shControl or shA3B constructs. The denaturation temperature range is indicated above each gel.

h, Pie charts depicting the C/G-to-T/A mutation load in 3D-PCR products after cloning and sequencing ($n \geq 35$ per condition). Charts align with target genes labeled in (g).

Figure 3. Cancer phenotypes triggered by inducing A3B over-expression.

a, The percent fluorescence for the indicated HEK293-derived A3-eGFP cell lines in absence or presence of Dox (corresponding immunoblot below).

b, Cell cycle status of the indicated cell lines at the indicated time points (comparisons made relative to each line uninduced).

c, Cell viability of the indicated cell populations using the MTS assay at the indicated time points post-Dox addition (comparisons made relative to each line uninduced).

d, Toxicity of induced A3A-eGFP and A3B-eGFP. Colonies were stained with crystal violet 8 days post-Dox treatment.

e & f, Representative fields of cells imaged for γ-H2AX and A3A-eGFP (1 day) or A3B-eGFP (3 days) post-Dox treatment and quantitative γ-H2AX flow cytometry over several days. Abnormal, multinuclear cell clusters are typical of induced A3B-eGFP expression (highlighted by white arrows; classified as abnormal cells in ‘b’).

g, Representative images of comets due to DNA fragmentation induced by A3A-eGFP or A3B-eGFP.

h, Pie diagram of C/G-to-T/A mutations in *TP53* detected by sequencing 3D-PCR products from A3A- or A3B-eGFP expressing cells, 2 or 4 days post-Dox treatment, respectively (n≥12 clones per condition).

Figure 4. *A3B* up-regulation and mutational signatures in breast tumors.

a, *A3B* mRNA levels in reduction mammoplasty samples (green triangles, n=28) and matched sets of breast tumor in comparison to adjacent or contralateral normal tissue (red circles and blue squares, respectively, n=52). Each data point is the mean *A3B* level of three qPCR reactions presented relative to mRNA levels of the constitutive housekeeping gene *TBP* (s.d. shown unless smaller than data point). Mean *A3B* levels are 0.025 (reduction), 0.26 (tumor), and 0.072 (normal). Data are arranged from lowest to highest *A3B* expression level.

b, *A3G* levels in the same samples as in (a). Average *A3G* levels are 1.00 (reduction), 1.45 (tumor), and 1.31 (normal) relative to *TBP*.

c, A3B catalytic domain deamination kinetics using single-stranded DNA substrates that vary only at the 3' position relative to the target cytosine.

d, Histogram depicting the percentage C-to-T of the total number of somatic mutations in the indicated tumors (total somatic mutations: lung 50489, liver 654879, melanoma 2798, breast 183916, breast triple negative 6964).

e, Nucleotide frequencies proportional to font size immediately 5' and 3' of genomic cytosines (expected) or the C-to-T mutated cytosine in cancers of the liver, skin, breast, or breast (triple negative). The next nucleotide preferences for recombinant A3B are derived experimentally from catalytic efficiencies (panel c & **Fig. S10**; additional comparisons in **Fig. S11**).

METHODS (to accompany the manuscript online)

RNA isolation, cDNA synthesis, and qPCR. Matched tumor/normal breast tumors and mammary reduction samples from the University of Minnesota TPF and breast cancer cell lines 30-4500K from the ATCC were used for RNA isolation, cDNA synthesis and qPCR as described[17]. Tissue RNA was from 100 mg flash-frozen tissue disrupted by a 2 h water bath sonication in 1 mL of Qiazol Lysis Reagent (RNeasy, Qiagen). Cell RNA was made using Qiashredder (RNeasy, Qiagen). qPCR was performed on a Roche Lightcycler 480 instrument. The housekeeping gene *TBP* was used for normalization. Statistical analyses for matched tissues were done using the Wilcoxson signed-rank test, and unmatched sets with the Mann-Whitney U-test (Graphpad Prism). Primer and probe sequences are listed in **Table S11**.

Knockdown constructs. *A3B* shRNA and shControl lentiviral constructs were from Open Biosystems through the BMGC RNAi Core (TRCN0000157469, TRCN0000140546, and scramble). Helper plasmids pdelta-NRF, containing HIV-1 *gag*, *pol*, *rev*, and *tat* genes, and pMDG, containing the VSV-G *env* gene, were co-transfected in HEK293T cells. Cell-free supernatants were harvested and concentrated by centrifugation (14,000 g x 2 h). Stable transductants were selected with puromycin (1 µg/ml).

Cell fractionation and DNA deaminase activity assays. Cellular fractionation was performed as described by syringe treatment of 10^7 cells in 0.5 mL of hypotonic buffer[31]. Nuclei were lysed by sonication in lysis buffer (25 mM Hepes, pH7.4, 250 mM NaCl, 10% glycerol, 0.5% Triton X-100, 1 mM EDTA, 1 mM $MgCl_2$, 1 mM $ZnCl_2$). Anti-histone H3 (1:2000; Abcam) and anti-tubulin (1:10,000; Covance) followed by anti-mouse 800 or anti-rabbit 680 (1:5000; Licor) immunoblots were used to assess fractionation. Lysates were tested in a fluorescence-based deaminase activity assay[20]. Dilutions were incubated 2 h at 37°C with a DNA oligonucleotide 5'-(6-FAM)-AAA-TT$\underline{\text{C}}$-TAA-TAG-ATA-ATG-TGA-(TAMRA). Fluorescence was measured on SynergyMx plate reader (BioTek). Local dinucleotide preferences in extracts were analyzed similarly using 5'-AC, CC, GC, or TC at the NN position of 5'-(6-FAM)-ATA-A$\underline{\text{NN}}$-AAA-TAG-ATA-AT-(TAMRA).

Genomic uracil quantifications. Genomic DNA was prepared from shA3B of shControl cells transduced and cultured for 21 days. Samples were spiked with heavy (+6)-labeled uracil (C^{13} and N^{15}; Cambridge Isotopes) and treated with UDG (NEB). Uracil was purified using 3,000 MWCO columns (Pall Scientific) and SPE (Carbograph, Grace). Samples were resuspended in water containing 0.1% formic acid. Analyses were performed on a capillary HPLC-ESI+-MS/MS (Thermo-Finnigan Ultra TSQ mass spectrometer, Waters nanoACQUITY HPLC). The MS was operated in positive ion mode, with 3.0 kV typical spray voltage, 250ºC capillary temperature, 67 V tube lens offset, and nitrogen sheath gas (25 counts). Argon collision gas was used at 1.1 mTorr. MS/MS analyses were performed with a scan width of 0.4 m/z and scan time of 0.1 s. The Hypercarb HPLC column (0.5 mm x100 mm, 5 µm, Thermo Scientific) was maintained at 40ºC and a flowrate of 15 µL/min. Solvents were 0.1% formic acid and acetonitrile. A linear gradient of 0% to 8% acetonitrile in 8 min was used, followed by an increase to 80% acetonitrile over 7 min. Uracils eluted at 11.5 min. Selected reaction monitoring was conducted with collision energy of 20V using the transitions: m/z 113.08 [M+H+]→70.08 [M-CONH]+ and m/z 96.08 [M-NH2]+ for uracil, while the internal standard ([15N-2, 13C-4]-uracil) was monitored by the transitions m/z 119.08 [M+H+]→m/z 74.08 [M-CONH]+ and m/z 101.08 [M-NH2]+ respectively. Internal standards were used for quantification.

TK fluctuations. *TK-neo* was introduced into MDA-MB-453 and HCC1569 cells as described[20]. TK^{plus} cells were transduced with shA3B or shControl lentiviruses and subcloned by limiting dilution. 10^6 cells from each expanded subclone population were subjected to ganciclovir and incubated until colonies outgrew. Frequencies were determined by applying the method of the median[32].

3D-PCR and sequencing. DNA was harvested from Ugi-expressing[33] T-REx-293 clones or HCC1569 cells transduced with shA3B or shControl lentiviruses. 3D-PCR was done using Taq (Denville Scientific) as described[20]. Primers sequences available upon request. PCR products were analyzed by gel electrophoresis with ethidium bromide, PCR purified (Epoch), blunt-end cloned into pJET (Fermentas), sequenced with T7 primer (BMGC), and aligned and analyzed with Sequencher software (Gene Codes Corporation).

Cell cycle experiments. T-REx-293 cells (Invitrogen) were transfected with pcDNA5/TO A3-GFP using TransIT-LT1 (Mirus) followed by clone selection using hygromycin. Cells were induced with 1 µg/mL Dox (MP Biomedicals 198955) for the indicated times then trypsinized and fixed with 4% paraformaldehyde in PBS. Cell pellets were resuspended in 0.1% Triton X 100, 20 µg/mL propidium iodide and 40 µg/mL RNase A (Qiagen) in PBS for 30 min and the DNA content and GFP induction measured by flow cytometry (BD Biosciences FACS Canto II) and analyzed with FlowJo and GraphPad Prism.

Cell viability assays. Cells were plated into multiple 96 well plates (2500 cells/well) and measured at the days indicated. The MTS reagent and PMS reagents were used as directed (Promega, Celltiter Aq 96). Absorbance was measured at 490 nm (PerkinElmer 1420 Victor 3V). The results were normalized to untreated cells. For crystal violet staining wells of a six well plate were plated with 2 x 10^5 cells. Half of the wells were induced with 1 ug/mL Dox. A crystal violet (0.5%), methanol (49.5%), water (50%) solution was used to stain cells after seven days.

DNA damage experiments. Flow cytometric analysis of γ-H2AX foci was adapted[34]. Fixed cells were incubated overnight in 0.2% Triton-X 100, 1 % BSA in PBS (blocking buffer) with 1:100 rabbit anti-γ-H2AX (Bethyl A300-081A). Secondary incubation was with goat anti-rabbit TRITC (Jackson 111025144) for 3 hrs before flow cytometry (BD Biosciences FACS Canto II) and analysis (FloJo and GraphPad). For microscopy, HEK293 cells were induced with 1 ug/mL of Dox before fixation with 4% paraformaldehyde and incubation with 1:50 anti-γ-H2AX conjugated to Alexa 647 (Cell Signaling 20E3) in blocking buffer for 3 hours. The cells were stained with 0.1% Hoechst dye and imaged at 20x or 60x (Deltavision) and deconvolved (SoftWoRx, Applied Precision).

Comet Assays. As described[35], microscope slides were coated with 1.5% agarose and dried. Low melting agarose (0.5% in PBS) was combined 1:1 with HEK293T cells transfected with A3A-eGFP (1 d) or A3B-eGFP (6 d). 10,000 cells were added to coated slides and the cells were lysed overnight in 10 mM Tris, 100 mM EDTA, 2.5 M NaCl, 1% Triton X-100. Slides were incubated for 10 min in running buffer (300 mM NaOH, 1 mM EDTA pH 13.1) then run at 0.75 V/cm 30 min. Gels were neutralized with 0.4 M Tris-Cl pH 7.5 and treated with RNase A (Qiagen). The microgels were allowed to dry and comets were visualized using propidium iodide.

References:

31 Shlyakhtenko, L. S. *et al.* Atomic force microscopy studies provide direct evidence for dimerization of the HIV restriction factor APOBEC3G. *J Biol Chem* **286**, 3387-3395 (2011).

32 Lea, D. E. & Coulson, C. A. The distribution of the numbers of mutants in bacterial populations. *Journal of Genetics* **49**, 264-285 (1949).

33 Di Noia, J. & Neuberger, M. S. Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. *Nature* **419**, 43-48 (2002).

34 Huang, X. & Darzynkiewicz, Z. Cytometric assessment of histone H2AX phosphorylation: a reporter of DNA damage. *Methods Mol Biol* **314**, 73-80 (2006).

35 Fairbairn, D. W., Olive, P. L. & O'Neill, K. L. The comet assay: a comprehensive review. *Mutat Res* **339**, 37-59 (1995).

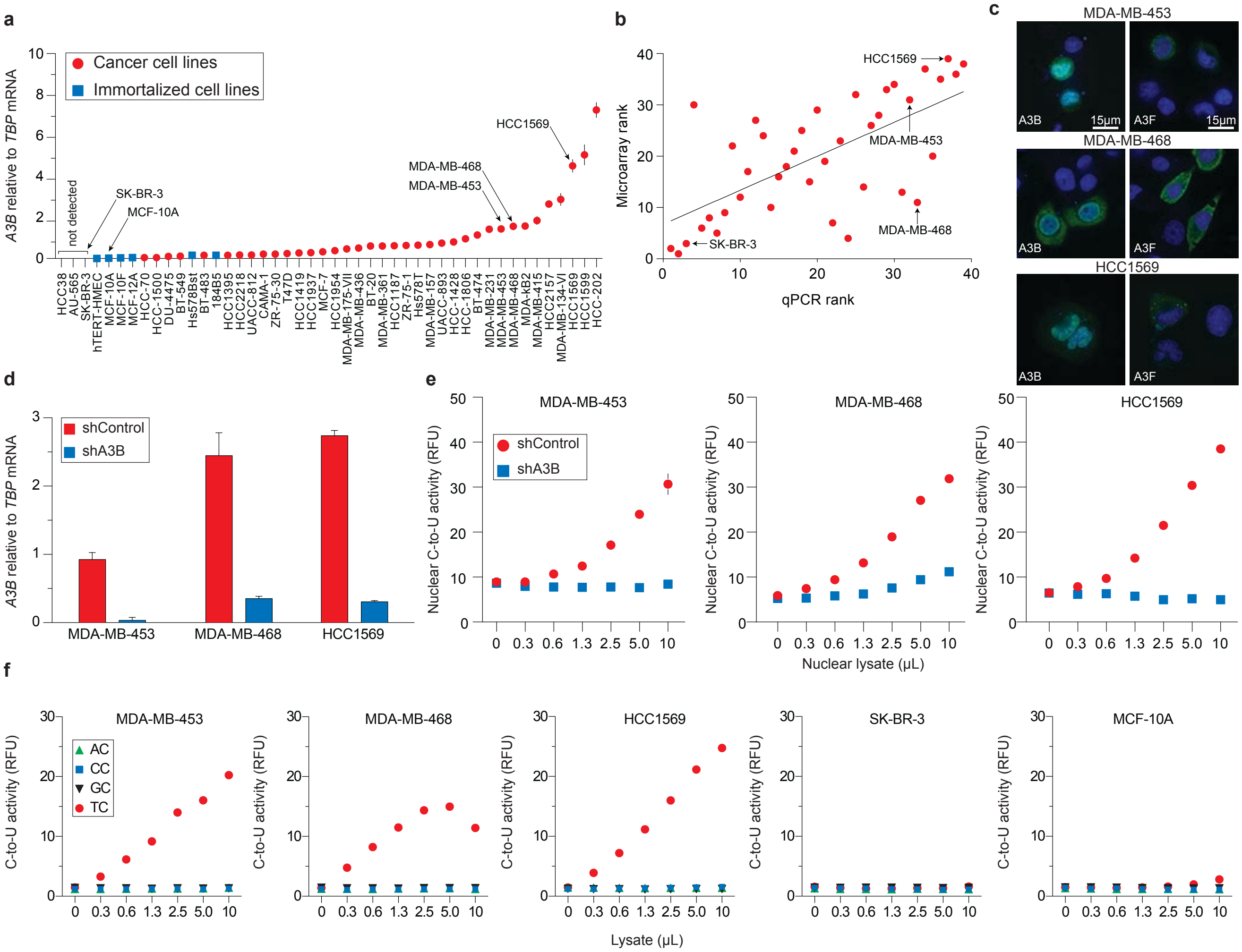

a
10
8
6
4
2
0
A3B relative to TBP mRNA
Cancer cell lines
Immortalized cell lines
not detected
SK-BR-3
MCF-10A
MDA-MB-468
MDA-MB-453
HCC1569
HCC38
AU-565
SK-BR-3
hTERT-HMEC
MCF-10A
MCF-10F
MCF-12A
HCC-70
HCC-1500
DU-4475
BT-549
Hs578Bst
BT-483
184B5
HCC1395
HCC2218
UACC-812
CAMA-1
ZR-75-30
T47D
HCC1419
HCC1937
MCF-7
HCC1954
MDA-MB-175-VII
MDA-MB-436
BT-20
MDA-MB-361
HCC1187
ZR-75-1
Hs578T
MDA-MB-157
UACC-893
HCC-1428
HCC-1806
BT-474
MDA-MB-231
MDA-MB-453
MDA-MB-468
MDA-kB2
MDA-MB-415
HCC2157
MDA-MB-134-VI
HCC1569
HCC1599
HCC-202
b
40
30
20
10
0
Microarray rank
HCC1569
MDA-MB-453
SK-BR-3
MDA-MB-468
0
10
20
30
40
qPCR rank
c
MDA-MB-453
A3B
15µm
A3F
15µm
MDA-MB-468
A3B
A3F
HCC1569
A3B
A3F
d
3
2
1
0
A3B relative to TBP mRNA
shControl
shA3B
MDA-MB-453
MDA-MB-468
HCC1569
e
50
40
30
20
10
0
Nuclear C-to-U activity (RFU)
MDA-MB-453
shControl
shA3B
0
0.3
0.6
1.3
2.5
5.0
10
MDA-MB-468
HCC1569
Nuclear lysate (µL)
f
30
20
10
0
C-to-U activity (RFU)
MDA-MB-453
AC
CC
GC
TC
MDA-MB-468
HCC1569
SK-BR-3
MCF-10A
0
0.3
0.6
1.3
2.5
5.0
10
Lysate (µL)

Fig. 1

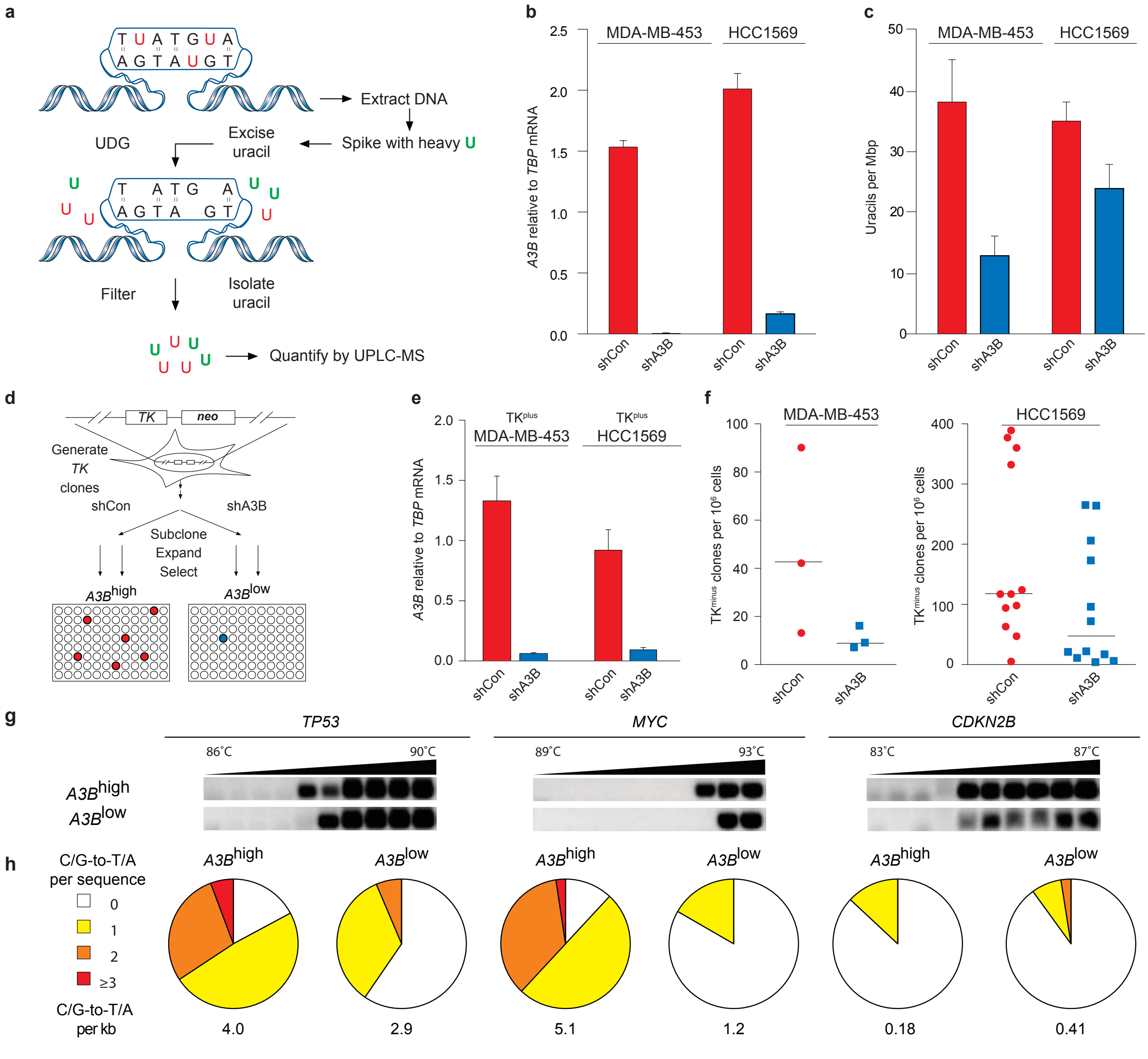

a
T U A T G U A
A G T A U G T
Extract DNA
Spike with heavy U
Excise
uracil
UDG
U
U U
T A T G A
A G T A G T
U U
U
Filter
Isolate
uracil
U U U
U U U
Quantify by UPLC-MS
b
MDA-MB-453
HCC1569
A3B relative to TBP mRNA
2.5
2.0
1.5
1.0
0.5
0.0
shCon
shA3B
shCon
shA3B
c
MDA-MB-453
HCC1569
Uracils per Mbp
50
40
30
20
10
0
shCon
shA3B
shCon
shA3B
d
TK
neo
Generate
TK
clones
shCon
shA3B
Subclone
Expand
Select
A3Bhigh
A3Blow
e
TKplus
MDA-MB-453
TKplus
HCC1569
A3B relative to TBP mRNA
2.0
1.5
1.0
0.5
0.0
shCon
shA3B
shCon
shA3B
f
MDA-MB-453
TKminus clones per 10^6 cells
100
80
60
40
20
0
shCon
shA3B
HCC1569
TKminus clones per 10^6 cells
400
300
200
100
0
shCon
shA3B
g
TP53
86˚C
90˚C
MYC
89˚C
93˚C
CDKN2B
83˚C
87˚C
A3Bhigh
A3Blow
h
C/G-to-T/A
per sequence
0
1
2
≥3
C/G-to-T/A
per kb
A3Bhigh
A3Blow
A3Bhigh
A3Blow
A3Bhigh
A3Blow
4.0
2.9
5.1
1.2
0.18
0.41

Fig. 2

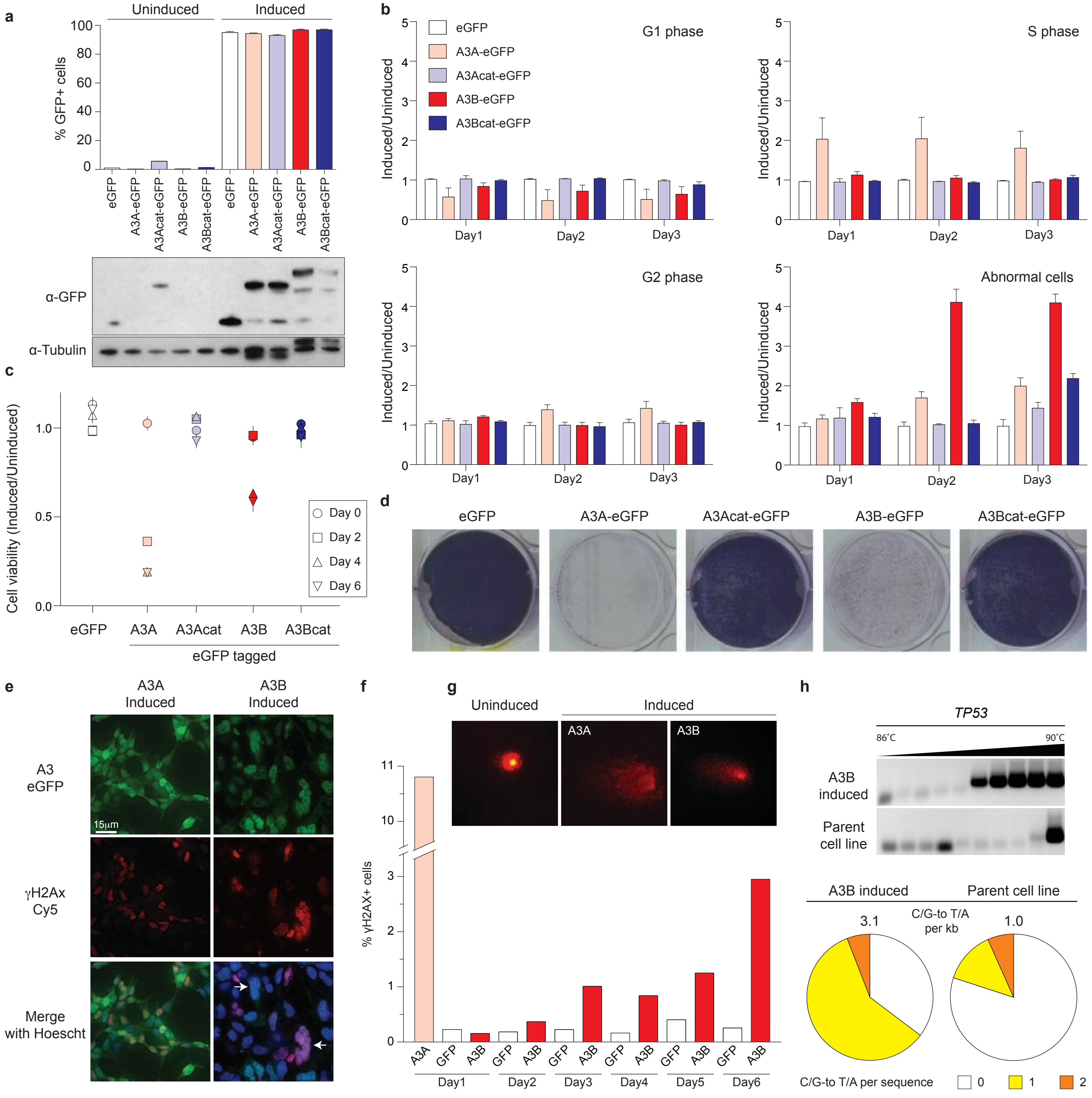

a
Uninduced
Induced
% GFP+ cells
100
80
60
40
20
0
eGFP
A3A-eGFP
A3Acat-eGFP
A3B-eGFP
A3Bcat-eGFP
eGFP
A3A-eGFP
A3Acat-eGFP
A3B-eGFP
A3Bcat-eGFP
α-GFP
α-Tubulin
b
eGFP
A3A-eGFP
A3Acat-eGFP
A3B-eGFP
A3Bcat-eGFP
G1 phase
S phase
G2 phase
Abnormal cells
Induced/Uninduced
0
1
2
3
4
5
Day1
Day2
Day3
c
Cell viability (Induced/Uninduced)
0.0
0.5
1.0
○ Day 0
□ Day 2
△ Day 4
▽ Day 6
eGFP
A3A
A3Acat
A3B
A3Bcat
eGFP tagged
d
eGFP
A3A-eGFP
A3Acat-eGFP
A3B-eGFP
A3Bcat-eGFP
e
A3A Induced
A3B Induced
A3 eGFP
15μm
γH2Ax Cy5
Merge with Hoescht
f
% γH2AX+ cells
0
1
2
3
10
11
A3A
GFP
A3B
Day1
Day2
Day3
Day4
Day5
Day6
g
Uninduced
Induced
A3A
A3B
h
TP53
86°C
90°C
A3B induced
Parent cell line
A3B induced
Parent cell line
3.1
C/G-to T/A per kb
1.0
C/G-to T/A per sequence
□ 0
🟨 1
🟧 2

Fig. 3

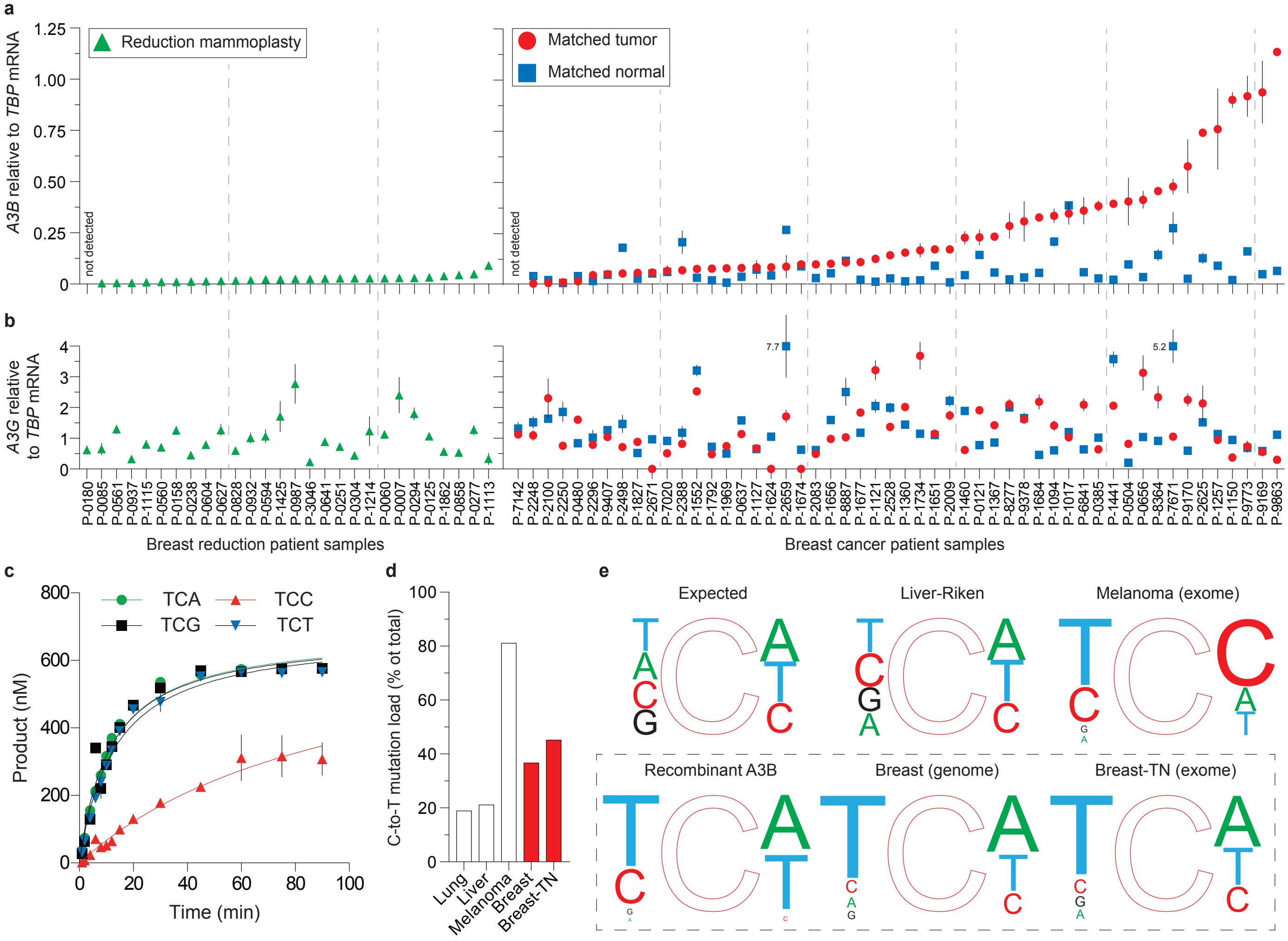

a
A3B relative to TBP mRNA
1.25
1.00
0.75
0.50
0.25
0
not detected
Reduction mammoplasty
not detected
Matched tumor
Matched normal
b
A3G relative to TBP mRNA
4
3
2
1
0
7.7
5.2
P-0180
P-0085
P-0561
P-0937
P-1115
P-0560
P-0158
P-0238
P-0604
P-0627
P-0828
P-0932
P-0594
P-1425
P-0987
P-3046
P-0641
P-0251
P-0304
P-1214
P-0060
P-0007
P-0294
P-0125
P-1862
P-0858
P-0277
P-1113
P-7142
P-2248
P-2100
P-2250
P-0480
P-2296
P-9407
P-2498
P-1827
P-2671
P-7020
P-2388
P-1552
P-1792
P-1969
P-0637
P-1127
P-1624
P-2659
P-1674
P-2083
P-1656
P-8887
P-1677
P-1121
P-2528
P-1360
P-1734
P-1651
P-2009
P-1460
P-0121
P-1367
P-8277
P-9378
P-1684
P-1094
P-1017
P-6841
P-0385
P-1441
P-0504
P-0656
P-8364
P-7671
P-9170
P-2625
P-1257
P-1150
P-9773
P-9169
P-9863
Breast reduction patient samples
Breast cancer patient samples
c
Product (nM)
800
600
400
200
0
0
20
40
60
80
100
Time (min)
TCA
TCC
TCG
TCT
d
C-to-T mutation load (% of total)
100
80
60
40
20
0
Lung
Liver
Melanoma
Breast
Breast-TN
e
Expected
Liver-Riken
Melanoma (exome)
Recombinant A3B
Breast (genome)
Breast-TN (exome)

Fig. 4

Supplementary online materials for

APOBEC3B is an enzymatic source of mutation in breast cancer

Michael B. Burns[1-4,*], Lela Lackey[1-4,*], Michael A. Carpenter[1-4], Anurag Rathore[1-4], Allison M. Land[1-4], Brandon Leonard[2-5], Eric W. Refsland[1-4], Delshanee Kotandeniya[2,6], Natalia Tretyakova[2,6], Jason B. Nikas[2], Douglas Yee[2], Nuri A. Temiz[7], Duncan E. Donohue[7], Rebecca M. McDougle[1-4], William L. Brown[1-4], Emily K. Law[1-4] & Reuben S. Harris[1-5,#]

[1]Biochemistry, Molecular Biology and Biophysics Department, [2]Masonic Cancer Center, [3]Institute for Molecular Virology, [4]Center for Genome Engineering, [5]Microbiology, Cancer Biology and Immunology Graduate Program, [6]Department of Medicinal Chemistry,
University of Minnesota, Minneapolis, MN 55455, USA
[7]*In Silico* Research Centers of Excellence, Advanced Biomedical Computing Center, Information Systems Program, SAIC-Frederick Inc., Frederick National Laboratory for Cancer Research, Frederick, MD 21702 USA

[*] Equal primary contributions.

[#] Correspondence to R.S.H. (rsh@umn.edu).

This section contains Supplementary Discussion, Methods, Tables S1-S11 and Figures S1-S13.

SUPPLEMENTARY DISCUSSION

Why has *A3B* eluded identification as an oncogene prior to this study? The most likely explanation is that the *A3B* gene shares a high level of sequence identity (in some regions nearly 100%) with the 10 other APOBEC family members. Therefore, the short oligonucleotides used as probes on microarrays are not capable of identifying any single *APOBEC*, simply an overall total for different cross-hybridizing mRNA species. This issue is illustrated in tabular format in **Tables S3-S10**. For instance, the commonly used Affymetrix Genechip Human Genome Array U133A has 11 probes intended for *A3B* detection (**Table S3 & S5**). Of these probes, *nine are not specific*, with 22/25 or 23/25 nucleotides identity to *A3A* and/or *A3G*. Similar non-specificities (and even complete off-target designs) were evident for the other *APOBEC3* probe clusters (**Tables S3-S10**).

Nevertheless, with knowledge of these limitations, useful information can still be derived from published microarray data sets. In particular, robust comparisons with microarray data become possible for breast cancer cell lines, which are clonal and do not express *A3A* (this gene is only expressed in myeloid lineage cells[1-4]) (**Figs. S1 & S2**). A strong, positive correlation is evident between our *A3B* qPCR measurements and reported microarray values for *A3B* in the ATCC breast cancer cell line panel (**Fig. 1b**; Cancer Cell Line Encyclopedia, http://www.broadinstitute.org/ccle/home).

However, the situation is more complex for microarray studies of human neoplasms, which are invariably a montage of tumor and multiple surrounding/infiltrating normal cell types. Moreover, depending on the stringency of hybridization and the particular sample being analyzed, *A3A* and *A3G* sequences may easily outcompete potential *A3B* target sequences (*e.g.*, *A3G* is higher than *A3B* in most samples that we analyzed; **Figs. 4 & S8**). Regardless, in comparisons of large published microarray data sets, we were still able to detect significant *A3B* up-regulation in tumor versus normal tissues (n=285 and n=22; p-value $<10^{-6}$; **Table S3**). As expected by the non-specificity of several probe sets, highly significant differences were also seen for the "*A3A*" and "*A3F,G*" probe sets, which are both predicted to cross-hybridize with *A3B* mRNA (**Tables S4 & S7**). In comparison, probe sets with low identity to *A3B* showed no significant correlation (*e.g.*, *A3C*; **Table S6**). As shown in **Fig. S9**, near-identical expression values for 62 housekeeping genes between different microarray data sets provides strong confidence that this approach is detecting over-expression of an *APOBEC3* gene in tumor versus normal samples. This situation mirrors our original hybridization results[5]. However, combined with the data sets shown here, we are confident in our conclusion that this up-regulated *APOBEC3* gene is indeed only *A3B*.

A secondary explanation for why *A3B* has proven elusive up to now is that the short read lengths generated during deep-sequencing (RNA-Seq) are difficult to assign unambiguously to members of repetitive gene families such as the *APOBEC3*s resulting in sequence mis-assignment or exclusion at the grooming stage of bioinformatic analysis. A final explanation is that the *A3B* gene is not a hotspot for gross chromosome abnormalities (database of Chromosomal Rearrangements In Diseases[6], http://dbCRID.biolead.org), which might have been found by classical cytogenetic techniques[7] or, more recently, by deep sequencing[8,9].

SUPPLEMENTARY METHODS

Microarray comparisons. Affymetrix GeneChip microarray data were reported previously by others. Tripathi *et al.*[10] (GEO ID GSE9574) and Graham *et al.*[11] (GEO ID GSE20437) reported data for 15 and 7 reduction mammoplasty samples, respectively. Tabchy *et al.*[12] (GEO ID GSE20271) reported data for 178 stage I-III breast cancers (procured at 6 sites worldwide), and Lasham *et al.*[13] (GEO ID GSE36771) reported data for 107 primary breast tumors. NCBI GEO resources were used to obtain raw data sets for additional analyses (CEL files). Next, we used the RMA algorithm (510K FDA approved) of the Expression Console Software (Affymetrix) with the standard settings to re-analyze the data for all 307 subjects. Since data sets from multiple independent studies were used, we normalized all tumor data with respect to the normal data in order to be able to perform comparisons. More specifically, we projected all tumor data into the space of the normal data by performing a non-linear normalization employing the following mathematical function:

$$\mathbf{X_n} = \frac{\mathbf{R_n}}{\mathbf{1+e^{\left(\frac{X_o-m}{R_o}\right)}}} + \mathbf{N}_{\min} \qquad (1)$$

In Eq. (1), X_n is the new, normalized variable; X_o is the old variable; R_n is the magnitude of the range of the new space; R_o is the magnitude of the range of the old space; m is the median of the old variable; and N_{min} is the minimum of the range of the new space. 62 housekeeping genes were used to assess these normalization methods, and a strong positive correlation was found between each independent data set (*e.g.*, **Fig. S9**). Having performed the same normalization method to all *APOBEC3* genes, we were able to obtain expression data for the tumor versus normal samples (**Table S3**). As previously[14-18], we assessed statistical significance using three different methods: i) t-Test (Mann-Whitney for non-parametric variables) with the significance level adjusted to $\alpha = 0.007143$ to account for seven comparisons, ii) fold-change defined as the ratio of the mean expression of the cancer group over the mean expression of the normal group (FC=C/N), and iii) ROC AUC. We performed ROC curve analysis on all seven *APOBEC3* probe clusters to assess their discriminating power with respect to the two groups (cancer versus normal). As can be seen in **Table S3**, the probe sets corresponding to *A3A*, *A3B*, and *A3(F,G)* are deemed to have significant differential expression according to all three methods.

SUPPLEMENTARY REFERENCES

1 Peng, G. *et al.* Myeloid differentiation and susceptibility to HIV-1 are linked to APOBEC3 expression. *Blood* **110**, 393-400 (2007).

2 Chen, H. *et al.* APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. *Curr Biol* **16**, 480-485 (2006).

3 Refsland, E. W. *et al.* Quantitative profiling of the full *APOBEC3* mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction. *Nucleic Acids Res* **38**, 4274-4284 (2010).

4 Stenglein, M. D., Burns, M. B., Li, M., Lengyel, J. & Harris, R. S. APOBEC3 proteins mediate the clearance of foreign DNA from human cells. *Nat Struct Mol Biol* **17**, 222-229 (2010).

5 Harris, R. S., Petersen-Mahrt, S. K. & Neuberger, M. S. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol Cell* **10**, 1247-1253 (2002).

6 Kong, F. *et al.* dbCRID: a database of chromosomal rearrangements in human diseases. *Nucleic Acids Res* **39**, D895-900 (2011).

7 Edwards, P. A. Fusion genes and chromosome translocations in the common epithelial cancers. *J Pathol* **220**, 244-254 (2010).

8 Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993 (2012).

9 Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400-404 (2012).

10 Tripathi, A. *et al.* Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. *Int J Cancer* **122**, 1557-1566 (2008).

11 Graham, K. *et al.* Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile. *British journal of cancer* **102**, 1284-1293 (2010).

12 Tabchy, A. *et al.* Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer. *Clin Cancer Res* **16**, 5351-5361 (2010).

13 Lasham, A. *et al.* YB-1, the E2F pathway, and regulation of tumor cell growth. *J Natl Cancer Inst* **104**, 133-146 (2012).

14 Nikas, J. B., Boylan, K. L., Skubitz, A. P. & Low, W. C. Mathematical prognostic biomarker models for treatment response and survival in epithelial ovarian cancer. *Cancer Inform* **10**, 233-247 (2011).

15 Nikas, J. B. & Low, W. C. ROC-supervised principal component analysis in connection with the diagnosis of diseases. *Am J Transl Res* **3**, 180-196 (2011).

16 Nikas, J. B. & Low, W. C. Application of clustering analyses to the diagnosis of Huntington disease in mice and other diseases with well-defined group boundaries. *Comput Methods Programs Biomed* **104**, e133-147 (2011).

17 Nikas, J. B. & Low, W. C. Linear discriminant functions in connection with the micro-RNA diagnosis of colon cancer. *Cancer Inform* **11**, 1-14 (2012).

18 Nikas, J. B., Low, W. C. & Burgio, P. A. Prognosis of treatment response (pathological complete response) in breast cancer. *Biomark Insights* **7**, 59-70 (2012).

Supplementary Table S1. Breast cell line information.

Cell Line	Derivation	Site of Origin	ER	PR	Her2/neu	TP53
hTERT-HMEC	Immortalized	Mammary gland	n.a.	n.a.	n.a.	normal
MCF-10A (MCF-10F)*	Immortalized	Mammary Gland	n.a.	n.a.	n.a.	normal
MCF-10F (MCF-10A)*	Immortalized	Mammary Gland	n.a.	n.a.	n.a.	normal
MCF-12A	Immortalized	Mammary Gland	n.a.	n.a.	n.a.	normal
Hs578Bst	Immortalized	Mammary Gland	-	n.a.	n.a.	normal
184B5	Immortalized	Mammary Gland	n.a.	n.a.	n.a.	normal
HCC38	Cancer	Primary Ductal Carcinoma	-	-	-	mutant
AU-565 (SK-BR-3)*	Cancer	Metastatic Adenocarcinoma; Pleural Effusion	n.a.	n.a.	+	mutant
SK-BR-3 (AU-565)*	Cancer	Adenocarcinoma; Pleural effusion	n.a.	n.a.	n.a.	mutant
HCC70	Cancer	Primary Ductal Carcinoma	+	-	-	mutant
HCC1500	Cancer	Primary Ductal Carcinoma	+	+	-	normal
DU4475	Cancer	Mammary Gland	n.a.	n.a.	n.a.	normal
BT-549	Cancer	Papillary, Invasive Ductal Tumor	n.a.	n.a.	n.a.	mutant
BT-483	Cancer	Ductal Carcinoma	n.a.	n.a.	n.a.	mutant
HCC1395	Cancer	Primary Ductal Carcinoma	-	-	-	mutant
HCC2218	Cancer	Primary Ductal Carcinoma	-	n.a.	+	mutant
UACC-812	Cancer	Primary Ductal Carcinoma	-	-	+	normal
CAMA-1	Cancer	Pleural Effusion	n.a.	n.a.	n.a.	mutant
ZR-75-30	Cancer	Ascites	n.a.	n.a.	n.a.	normal
T47D	Cancer	Ductal Carcinoma	+	+	-	mutant
HCC1419	Cancer	Primary Ductal Carcinoma	-	-	+	mutant
HCC1937	Cancer	Primary Ductal Carcinoma	-	-	-	mutant
MCF-7	Cancer	Adenocarcinoma; Pleural Effusion	+	+	-	normal
HCC1954	Cancer	Primary Ductal Carcinoma	-	-	+	mutant
MDA-MB-175-VII	Cancer	Metastic Ductal Carcinoma; Pleural Effusion	n.a.	n.a.	n.a.	normal
MDA-MB-436	Cancer	Metastatic Adenocarcinoma; Pleural Effusion	n.a.	n.a.	n.a.	mutant
BT-20	Cancer	Mammary Gland Carcinoma	-	n.a.	n.a.	mutant
MDA-MB-361	Cancer	Metastatic Adenocarinoma	n.a.	n.a.	n.a.	mutant
HCC1187	Cancer	Primary Ductal Carcinoma	n.a.	-	-	mutant
ZR-75-1	Cancer	Ascites	+	n.a.	n.a.	normal
Hs578T	Cancer	Mammary Gland Carcinoma	-	n.a.	n.a.	mutant
MDA-MB-157	Cancer	Medulallary Carcinoma	n.a.	n.a.	n.a.	mutant
UACC-893	Cancer	Primary Ductal Carcinoma	-	-	+	mutant
HCC1428	Cancer	Adenocarcinoma; Pleural Effusion cells	n.a.	n.a.	-	mutant
HCC1806	Cancer	Primary Squamous Cell Carcinoma	-	-	-	mutant
BT-474	Cancer	Invasive Ductal Carcinoma	n.a.	n.a.	n.a.	mutant
MDA-MB-231	Cancer	Metastatic adenocarcinoma; Pleural Effusion	-	-	-	mutant
MDA-MB-453 (MDA-kb2)*	Cancer	Metastatic Pericardial Effusion	-	-	+	mutant
MDA-MB-468	Cancer	Metastatic Adenocarcinoma; Pleural Effusion	-	-	-	mutant
MDA-kb2 (MDA-MB-453)*	Cancer	Metastatic Pericardial Effusion	n.a.	n.a.	n.a.	mutant
MDA-MB-415	Cancer	Adenocarcinoma; Pleural Effusion	n.a.	n.a.	n.a.	mutant
HCC2157	Cancer	Primary Ductal Carcinoma	-	+	+	mutant
MDA-MB-134-VI	Cancer	Pleural Effusion	n.a.	n.a.	n.a.	mutant
HCC1569	Cancer	Primary Metaplastic Carcinoma	-	-	+	mutant
HCC1599	Cancer	Primary Ductal Carcinoma	-	-	-	mutant
HCC202	Cancer	Primary Ductal Carcinoma	-	-	+	mutant

*Related cell lines; n.a. = not available.

Supplementary Table S2. Breast cancer patient information.

Table S2

Patient ID	Age	Ethnicity	Age	ER	PR	Her2/neu	Type	Grade
P-7142	40	Caucasian	40	+	-	+	IDC	3
P-2248	51	African American	51	+	-	-	IDC	2
P-2100	75	Caucasian	75	+	+	-	IDC	2
P-2250	76	Caucasian	76	+	+	-	IDC	2
P-0480	51	Caucasian	51	-	-	-	IDC	3
P-2296	49	Caucasian	49	+	+	-	IDC	2
P-9407	38	Caucasian	38	+	+	-	IDC	2
P-2498	40	Caucasian	40	+	+	-	IDC	2
P-1827	37	Caucasian	37	+	-	-	IDC/ILC	2
P-2671	61	Caucasian	61	+	+	-	ILC	2
P-7020	40	Caucasian	40	+	+	-	IDC/ILC	1
P-2388	47	Caucasian	47	+	+	-	IDC	1
P-1552	58	Caucasian	58	+	+	-	IDC	3
P-1792	44	Caucasian	44	+	+	n.a.	DCIS	1
P-1969	77	Caucasian	77	+	-	+	IDC	2
P-0637	70	Caucasian	70	+	-	+	IDC	2
P-1127	68	Caucasian	68	+	+	n.a.	DCIS	1
P-1624	49	Caucasian	49	+	+	-	IDC	2
P-2659	58	Caucasian	58	+	-	+	IDC	3
P-1674	64	Caucasian	64	+	-	-	ILC	2
P-2083	39	Caucasian	39	+	+	-	IDC	2
P-1656	74	Caucasian	74	+	+	-	ILC	2
P-8887	45	Native American	45	+	+	-	LC	2
P-1677	49	Caucasian	49	+	+	-	IDC	2
P-1121	75	Caucasian	75	+	+	-	ILC	1
P-2528	51	Caucasian	51	+	+	-	ILC	2
P-1360	66	Caucasian	66	+	+	-	IDC	2
P-1734	47	Caucasian	47	+	+	-	IDC	2
P-1651	51	Caucasian	51	+	+	-	IMC	2
P-2009	62	Caucasian	62	+	+	-	IDC	1
P-1460	62	Caucasian	62	+	+	+	IDC	3
P-0121	77	Caucasian	77	+	+	-	IDC	2
P-1367	43	Caucasian	43	+	+	-	IDC	2
P-8277	54	Caucasian	54	+	-	-	IDC	1
P-9378	68	Caucasian	68	+	-	-	ILC	2
P-1684	45	Caucasian	45	+	+	-	IDC/ILC	2
P-1094	51	Caucasian	51	+	+	-	IDC	2
P-1017	40	Caucasian	40	+	+	+	IDC	3
P-6841	68	Caucasian	68	+	+	+	IDC	3
P-0385	56	Caucasian	56	+	+	-	ILC	2
P-1441	70	Caucasian	70	-	-	-	IDC	3
P-0504	56	Caucasian	56	+	-	-	IDC	2
P-0656	39	Caucasian	39	-	-	+	IDC	3
P-8364	42	Caucasian	42	+	-	n.a.	DCIS	1
P-7671	48	Caucasian	48	+	-	-	DCIS	1
P-9170	55	Caucasian	55	+	-	-	IDC	2
P-2625	72	Caucasian	72	+	+	+	IDC	3
P-1257	77	Caucasian	77	+	-	+	IDC	2
P-1150*	30	Caucasian	30	+	+	+	IDC	3
P-9773	37	Caucasian	37	-	-	-	IDC	3
P-9169	62	Caucasian	62	+	+	-	IDC/ILC	1
P-9863	46	Caucasian	46	+	+	-	IDC	2

*Listed in order from A3B^{null} to A3B^{high} as in **Fig. 4**. #Male patient; DCIS - Ductal carcinoma *in situ*; IDC - Invasive ductal carcinoma; ILC - Invasive lobular carcinoma; IDC/ILC - Invasive ductal carcinoma with lobular features; IMC - Invasive mucinous carcinoma; n.a. - Not available.

Supplementary Table S3. Microarray data summary.

Gene	**Normal** (n=22; mean ± SD)	**Cancer** (n=285; mean ± SD)	**t-Test P value**	**Fold Change (C/N)**	**ROC AUC**
210873_x_at (APOBEC3A)	3.554 ± 0.237	3.698 ±0.042	< 1x10^{-6}	1.041	0.836
206632_s_at (APOBEC3B)	4.049 ± 0.386	4.404 ± 0.082	< 1x10^{-6}	1.088	0.900
209584_x_at (APOBEC3C)	4.977 ± 0.226	4.901 ± 0.038	0.144	0.985	0.594
214995_s_at (APOBEC3F,G)	3.858 ± 0.190	4.012 ±0.037	1x10^{-6}	1.040	0.816
214994_at (APOBEC3F)	3.968 ± 0.228	3.894 ± 0.041	0.008	0.981	0.670
204205_at (APOBEC3G)	5.535 ± 0.491	5.422 ± 0.071	0.011	0.980	0.663
215579_at (APOBEC3G)	5.845 ± 0.187	5.897 ±0.037	0.001	1.010	0.705
House Gene 1	6.107 ± 0.312	6.039 ± 0.050	0.183	0.990	0.585
House Gene 2	3.053 ± 0.643	3.128 ± 0.080	0.438	1.025	0.550

Table S4. Affymetrix microarray HG-U133A A3A probe (cross)hybridization within the APOBEC3 family.

		Probe identity to APOBEC3A-H						
		# identities/probe length (%)						
Intended target gene* (RefSeq)	Probe set 210873_x_at	A	B	C	D*	F	G	H*
APOBEC3A NM_145699	GCTCACAGACGCCAGCAAAGCAGTA	25/25	22/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GACGCCAGCAAAGCAGTATGCTCCC	25/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GCAGTATGCTCCCGATCAAGTAGAT	25/25	22/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	AAAAAATCAGAGTGGGCCGGGCGCG	25/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GAGGCAGGAGAGTACGTGAACCCGG	24/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	AACTGAAAATTTCTCTTATGTTCCA	25/25	24/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	CTCTTATGTTCCAAGGTACACAATA	25/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GATTATGCTCAATATTCTCAGAATA	25/25	24/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	TTTGGCTTCATATCTAGACTAACAC	24/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GAATCTTCCATAATTGCTTTTGCTC	25/25	21/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	TAATTGCTTTTGCTCAGTAACTGTG	25/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25

* A3D and A3H are not represented intentionally in the U133 probe set.

Table S5. Affymetrix microarray HG-U133A A3B probe (cross)hybridization within the APOBEC3 family.

		Probe identity to APOBEC3A-H						
		# identities/probe length (%)						
Intended target gene* (RefSeq)	Probe set 206632_s_at	A	B	C	D*	F	G	H*
APOBEC3B NM_004900	CTACGATGAGTTTGAGTACTGCTGG	22/25	25/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	CACCTTTGTGTACCGCCAGGGATGT	23/25	25/25	≤20/25	≤20/25	≤20/25	23/25	≤20/25
	GAAATGCAAACGAGCCGTTCACCAC	22/25	22/25	≤20/25	≤20/25	≤20/25	22/25	≤20/25
	ACCAGCAAAGCAATGTGCTCCTGAT	≤20/25	25/25	≤20/25	≤20/25	≤20/25	22/25	≤20/25
	AGCAATGTGCTCCTGATCAAGTAGA	22/25	25/25	≤20/25	≤20/25	≤20/25	22/25	≤20/25
	ATGTGCTCCTGATCAAGTAGATTTT	22/25	25/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	TGTTCCAAGTGTACAAGAGTAAGAT	22/25	25/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	TTATGCTCAATATTCCCAGAATAGT	23/25	25/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	ATTCCCAGAATAGTTTTCAATGTAT	23/25	25/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GAAGTGATTAATTGGCTCCATATTT	≤20/25	25/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	TAATTGGCTCCATATTTAGACTAAT	≤20/25	25/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25

* A3D and A3H are not represented intentionally in the U133 probe set.

Table S6. Affymetrix microarray HG-U133A A3C probe (cross)hybridization within the APOBEC3 family.

		Probe identity to APOBEC3A-H						
		# identities/probe length (%)						
Intended target gene* (RefSeq)	Probe set 209584_x_at	A	B	C	D*	F	G	H*
APOBEC3C NM_14508	AAGGGGTCGCTGTGGAGATCATGGA	≤20/25	≤20/25	25/25	≤20/25	≤20/25	≤20/25	≤20/25
	TAATGAGCCATTCAAGCCTTGGGAA	≤20/25	≤20/25	24/25	23/25	23/25	≤20/25	≤20/25
	CCAACTTTCGACTTCTGAAAAGAAG	≤20/25	≤20/25	25/25	25/25	≤20/25	≤20/25	≤20/25
	AAGAAGGCTACGGGAGAGTCTCCAG	≤20/25	≤20/25	25/25	24/25	≤20/25	≤20/25	≤20/25
	GGGAGAGTCTCCAGTGAGGGGTCTC	≤20/25	≤20/25	25/25	24/25	22/25	≤20/25	≤20/25
	CTCCCCAGCATAACCAAATCTTACT	≤20/25	≤20/25	25/25	23/25	≤20/25	≤20/25	≤20/25
	TTACTAAACTCATGCTAGGCTGGGC	≤20/25	≤20/25	24/25	≤20/25	≤20/25	≤20/25	≤20/25
	TAGGCTGGGCATGGTGACTCACGCC	≤20/25	≤20/25	25/25	22/25	≤20/25	≤20/25	≤20/25
	GGTGGGAGAATCGCGTGAGCCCAGG	≤20/25	≤20/25	25/25	23/25	23/25	≤20/25	≤20/25
	AGCCCAGGAGTTCCAGACCAGGCTG	≤20/25	≤20/25	25/25	≤20/25	22/25	≤20/25	≤20/25
	TCCAGACCAGGCTGGGTCACATGAC	≤20/25	≤20/25	25/25	≤20/25	≤20/25	≤20/25	≤20/25

* A3D and A3H are not represented intentionally in the U133 probe set.

Table S7. Affymetrix microarray HG-U133A A3F/A3G (1) probe (cross)hybridization within the APOBEC3 family.

		Probe identity to APOBEC3A-H						
		# identities/probe length (%)						
Intended target gene* (RefSeq)	Probe set 214995_s_at	A	B	C	D*	F	G	H*
APOBEC3F, APOBEC3G NM_145298, NM_021822	GAAAGTGAAACCCTGGTGCTCCAGA	≤20/25	≤20/25	≤20/25	≤20/25	25/25	25/25	≤20/25
	GGTGCTCCAGACAAAGATCTTAGTC	≤20/25	≤20/25	≤20/25	≤20/25	25/25	25/25	≤20/25
	AGATCTTAGTCGGGACTAGCCGGCC	≤20/25	≤20/25	≤20/25	≤20/25	25/25	25/25	≤20/25
	GGGACTAGCCGGCCAAGGATGAAGC	≤20/25	≤20/25	≤20/25	≤20/25	25/25	25/25	≤20/25
	GAAGCCTCACTTCAGAAACACAGTG	≤20/25	≤20/25	≤20/25	≤20/25	25/25	25/25	≤20/25
	AGTGGAGCGAATGTATCGAGACACA	≤20/25	23/25	≤20/25	23/25	25/25	25/25	≤20/25
	ACACATTCTCCTACAACTTTTATAA	≤20/25	≤20/25	≤20/25	≤20/25	25/25	25/25	≤20/25
	TATAATAGACCCATCCTTTCTCGTC	≤20/25	≤20/25	≤20/25	≤20/25	25/25	25/25	≤20/25
	CTTTCTCGTCGGAATACCGTCTGGC	≤20/25	≤20/25	≤20/25	≤20/25	25/25	25/25	≤20/25
	TACCGTCTGGCTGTGCTACGAAGTG	≤20/25	≤20/25	≤20/25	≤20/25	25/25	25/25	≤20/25
	GGACGCAAAGATCTTTCGAGGCCAG	≤20/25	≤20/25	≤20/25	≤20/25	25/25	25/25	≤20/25

* A3D and A3H are not represented intentionally in the U133 probe set.

Table S8. Affymetrix microarray HG-U133A A3F/A3G (2) probe (cross)hybridization within the APOBEC3 family.

		Probe identity to APOBEC3A-H						
		# identities/probe length (%)						
Intended target gene* (RefSeq)	Probe set 214994_at	A	B	C	D*	F	G	H*
APOBEC3F NM_145298	CACCACATGGGACAGCGCAGGTCCA	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	CACATGGGACAGCGCAGGTCCAGTG	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	CCAGCTGACCGCAGGCAGGGAACAA	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GGCAGGGAACAAGGCAGACCCTAGA	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	AAGGCAGACCCTAGAGGGCCAGGCC	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	TGCCAGAATTCACGCATGAGGCTCT	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GCATGAGGCTCTGAACAGGGCTGGG	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	TGAACAGGGCTGGGAAAACTTCCAA	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	AAGCTCATGTCTTGGTGCACTTTGT	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	CACTTTGTGATGATGCTTCAACAGC	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GCTTCAACAGCAGGACTGAGATGGG	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25

* A3D and A3H are not represented intentionally in the U133 probe set.

Table S9. Affymetrix microarray HG-U133A A3G (1) probe (cross)hybridization within the APOBEC3 family.

		Probe identity to APOBEC3A-H						
		# identities/probe length (%)						
Intended target gene* (RefSeq)	Probe set 204205_at	A	B	C	D*	F	G	H*
APOBEC3G NM_021822	GCCCGCATCTATGATGATCAAGGAA	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	25/25	≤20/25
	AAGATGTCAGGAGGGGCTGCGCACC	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	25/25	≤20/25
	ACCAGCAAAGCAATGCACTCCTGAC	≤20/25	22/25	≤20/25	≤20/25	≤20/25	25/25	≤20/25
	GCAATGCACTCCTGACCAAGTAGAT	≤20/25	22/25	≤20/25	≤20/25	≤20/25	25/25	≤20/25
	GCACTCCTGACCAAGTAGATTCTTT	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	25/25	≤20/25
	ATTAGAGTGCATTACTTTGAATCAA	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	25/25	≤20/25
	TAAAGTACTAAGATTGTGCTCAATA	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	25/25	≤20/25
	GTTTCAAACCTACTAATCCAGCGAC	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	25/25	≤20/25
	AAACCTACTAATCCAGCGACAATTT	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	25/25	≤20/25
	ATCCAGCGACAATTTGAATCGGTTT	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	25/25	≤20/25
	GAATCGGTTTTGTAGGTAGAGGAAT	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	25/25	≤20/25

* A3D and A3H are not represented intentionally in the U133 probe set.

Table S10. Affymetrix microarray HG-U133A A3G (2) probe (cross)hybridization within the APOBEC3 family.

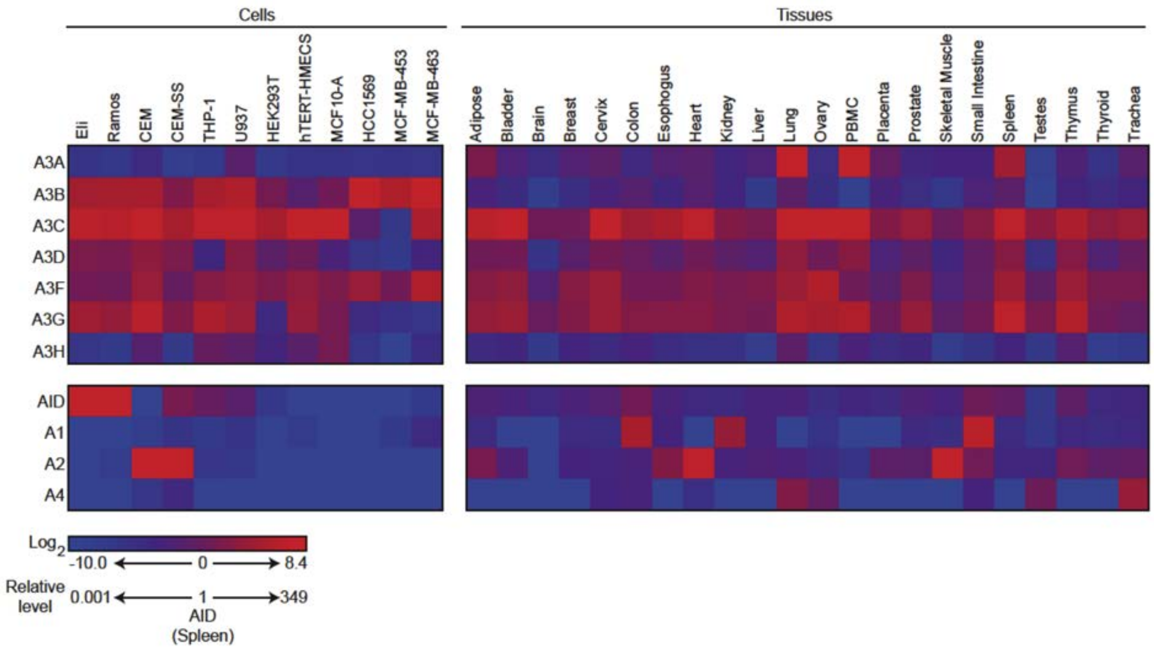
		Probe identity to APOBEC3A-H						
		# identities/probe length (%)						
Intended target gene* (RefSeq)	Probe set 215579_at	A	B	C	D*	F	G	H*
APOBEC3G NM_021822	TTTCCAAATACAGCCACCCTTTGAG	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	ACAGCCACCCTTTGAGGGAGCGGGG	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	TGAGGGAGCGGGGGTTAAGGCTTCA	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GGGGGTTAAGGCTTCAATACATTGA	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	AGAAACAGTGAAGGCCACGGCAAGA	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	AGAAGCTGCAGTCATTGTGGGCGGG	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	TTCCCAGGGGAGTCCTGACCTGACT	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	TCTGGGGTCCGGACATGACCCCTCA	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GTCCTATCAAAGGTGGCATCCTCCC	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GCCTCTGCACTGGGTGCTAATAATT	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25
	GGGTGCTAATAATTCACTTTTACCT	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25	≤20/25

* A3D and A3H are not represented intentionally in the U133 probe set.

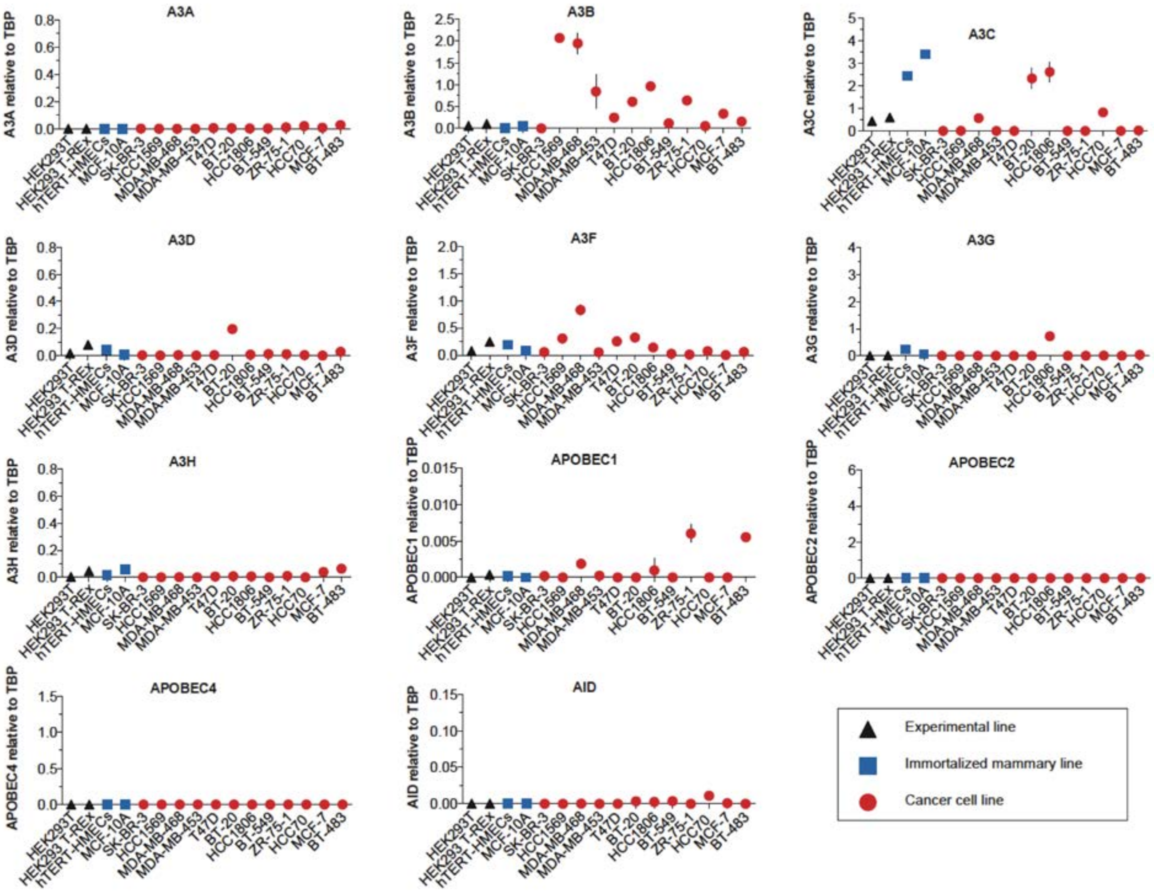
Supplementary Table S11. Quantitative PCR primer and probe information.

Gene Symbol	mRNA NCBI	5' Primer		3' Primer		Probe	
	Accession	Name	Seq (5'-3')	Name	Seq (5'-3')	Name	Seq[a]
APOBEC3s							
APOBEC3A	NM_145699	RSH2742	gagaagggacaagcacatgg	RSH2743	tggatccatcaagtgtctgg	UPL26	ctgggctg
APOBEC3B	NM_004900	RSH3220	gaccctttggtccttcgac	RSH3221	gcacagccccaggagaag	UPL1	cctggagc
APOBEC3C	NM_014508	RSH3085	agcgcttcagaaaagagtgg	RSH3086	aagtttcgttccgatcgttg	UPL155	ttgccttc
APOBEC3D	NM_152426	RSH2749	acccaaacgtcagtcgaatc	RSH2750	cacatttctgcgtggttctc	UPL51	ggcaggag
APOBEC3F	NM_145298	RSH2751	ccgtttggacgcaaagat	RSH2752	ccaggtgatctggaaacactt	UPL27	gctgcctg
APOBEC3G	NM_021822	RSH2753	ccgaggacccgaaggttac	RSH2754	tccaacagtgctgaaattcg	UPL79	ccaggagg
APOBEC3H	NM_181773	RSH2757	agctgtggccagaagcac	RSH2758	cggaatgtttcggctgtt	UPL21	tggctctg
AID	NM_020661	RSH3066	gactttggttatcttcgcaataaga	RSH3067	aggtcccagtccgagatgta	UPL69	ggaggaag
APOBEC1	NM_001644	RSH3068	gggaccttgttaacagtggagt	RSH3069	ccaggtgggtagttgacaaaa	UPL67	tgctggag
APOBEC2	NM_006789	RSH3070	aagtagggcaactgggcttt	RSH3071	ggctgtacatgtcattgctgtc	UPL74	ctgctgcc
APOBEC4	NM_203454	RSH3072	ttctaacacctggaatgtgatcc	RSH3073	tttactgtcttctagctgcaaacc	UPL80	cctggaga
Reference Gene							
TBP	NM_003194	RSH3231	cccatgactcccatgacc	RSH3232	tttacaaccaagattcactgtgg	UPL51	ggcaggag

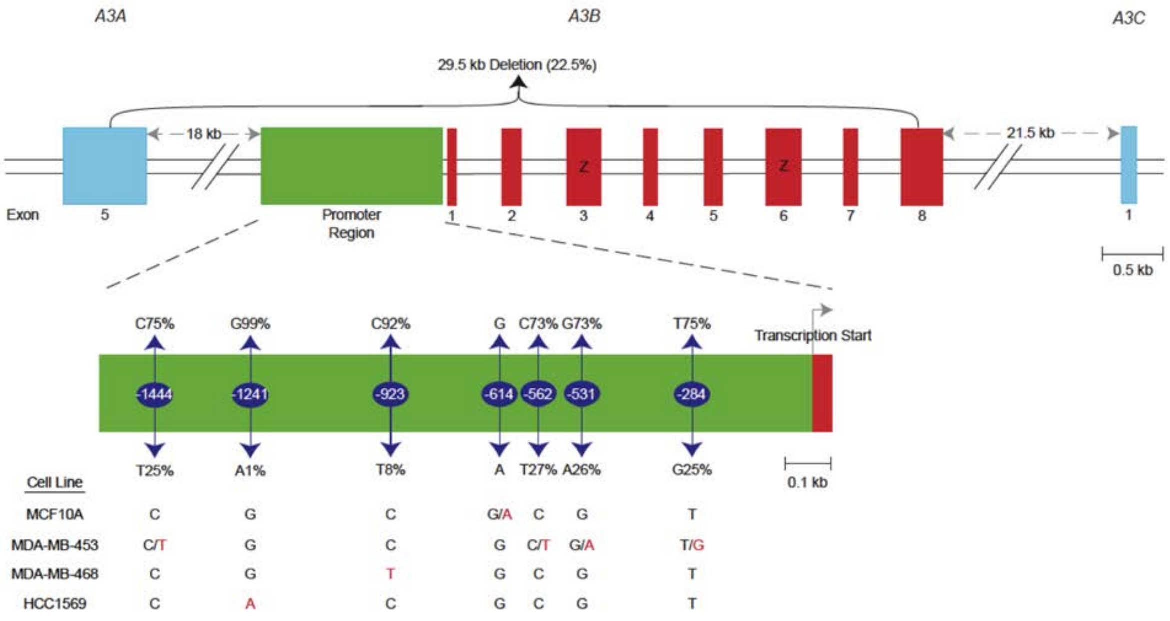
(a) It is not known whether probes from the Universal Probe Library (UPL) correspond to the coding or template DNA strands of their target sequences (Roche proprietary information).



Supplementary Figure S1. Expression profiles for *APOBEC* family members in human cell lines and tissues. A heat-map summary of qPCR data showing relative *APOBEC3* (*A3)*, *AID*, *APOBEC1* (*A1*), *APOBEC2* (*A2*), and *APOBEC4* (*A4*) mRNA expression levels in the indicated cell lines and tissues. The data are relative to the median *AID* mRNA level in spleen and presented in $\log_2$ format. The average of three independent qPCR reactions was used for each condition. Data for the normal tissues, excluding PBMCs and breast tissue, were reported previously [Refsland *et al*. (Ref. 17)]. They were recalculated and presented here in $\log_2$ format for comparative purposes and to emphasize the general observation that *A3B* is low or almost undetectable in every normal tissue that we have examined to date.

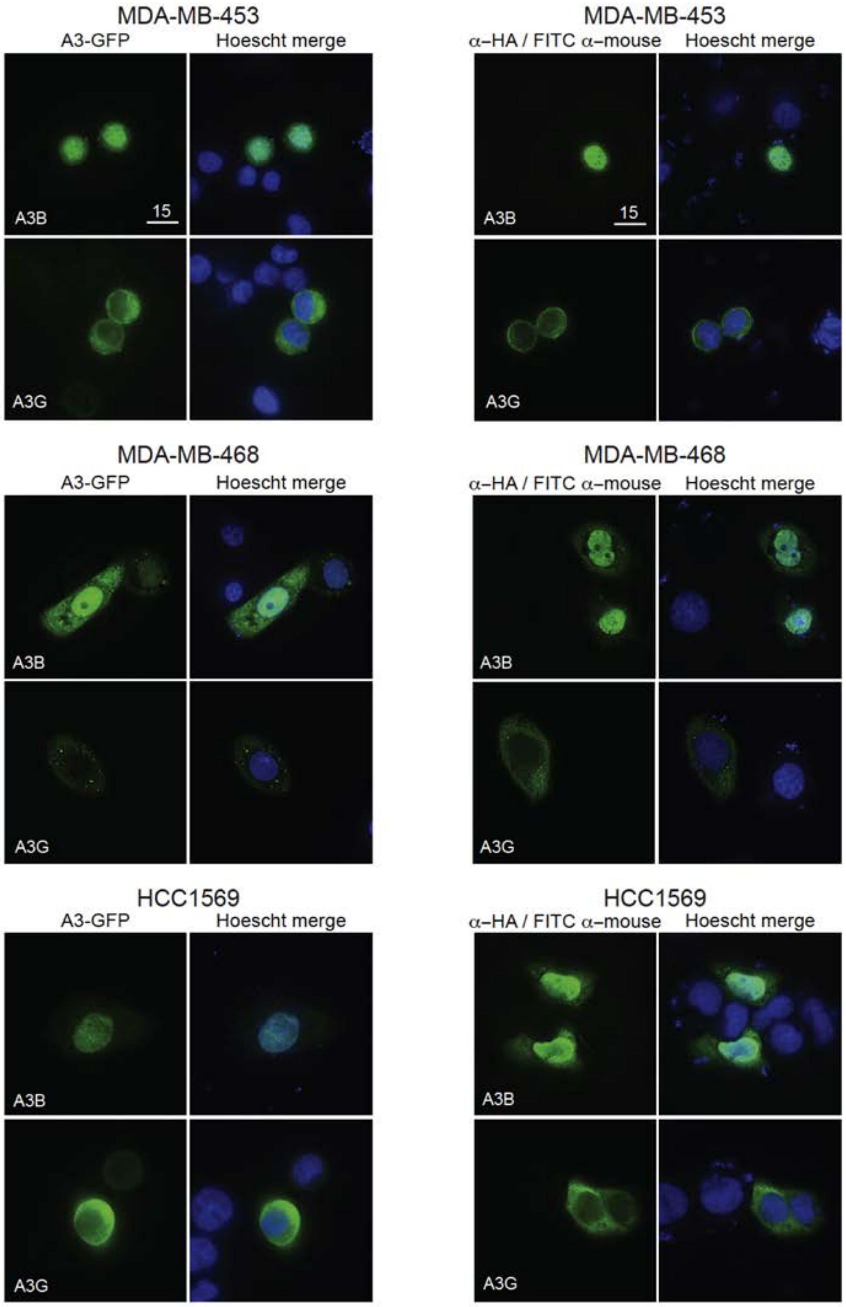


Supplementary Figure S2. Full expression profiles for *APOBEC* family members in a panel of representative cell lines.
The indicated cell lines were used to generate cDNA for qPCR analyses of the full human *APOBEC* repertoire. Each data point is mean mRNA level of three qPCR reactions presented relative to mRNA levels of the constitutive housekeeping gene *TBP* (s.d. shown as a bar unless smaller than the data point). Relevant *A3B* data are also presented in Fig. 1a in the context of the full panel of normal and breast cancer cell lines.

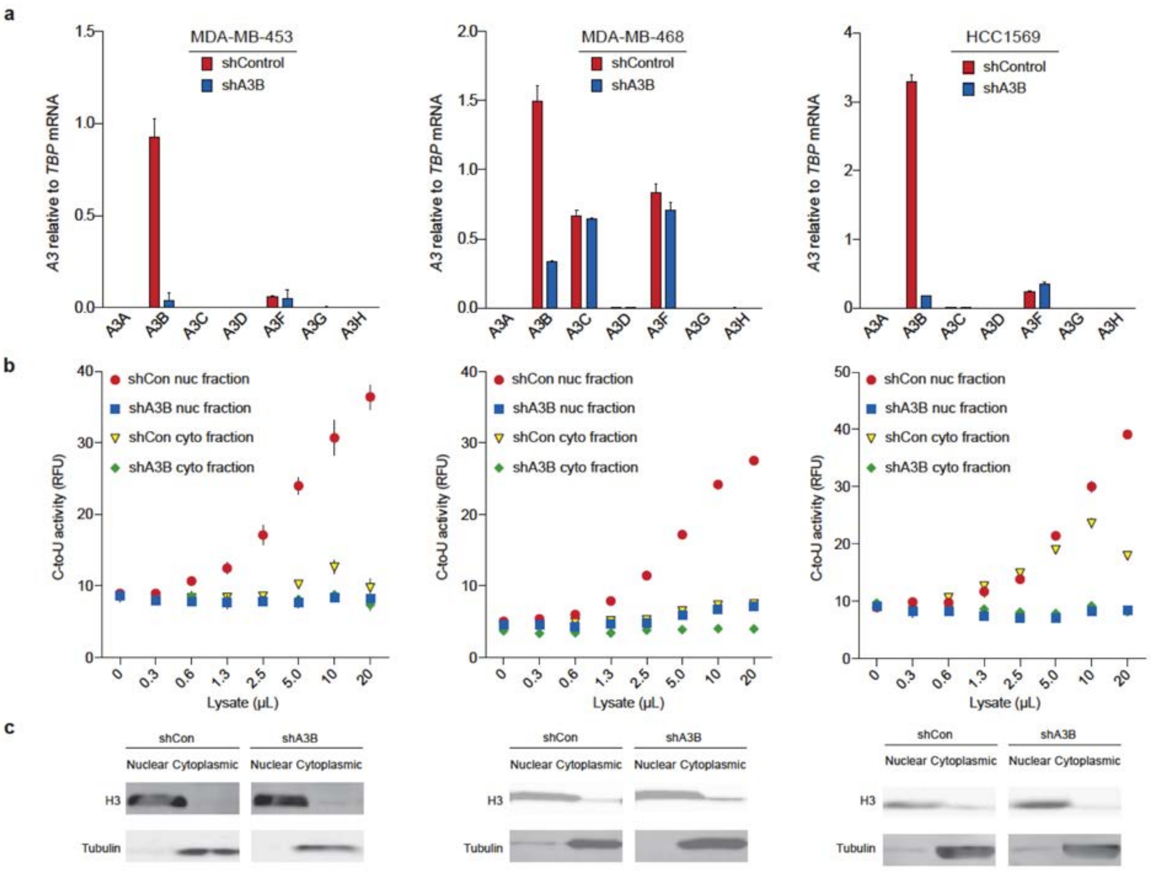


Cell Line	-1444	-1241	-923	-614	-562	-531	-284
MCF10A	C	G	C	G/A	C	G	T
MDA-MB-453	C/T	G	C	G	C/T	G/A	T/G
MDA-MB-468	C	G	T	G	C	G	T
HCC1569	C	A	C	G	C	G	T

Supplementary Figure S3. ***A3B*** **promoter region sequence analysis.**
A schematic of the *A3B* genomic locus depicting flanking genes (blue), exons (red), deaminase domain exons (red with Z label), promoter region (green), and position of the 29.5kb deletion allele. Below, an enlarged schematic of the *A3B* promoter region showing the most common SNPs (above) and minor alleles (below). Allele frequencies are indicated as percentages (www.ncbi.nlm.nih.gov/projects/SNP/). Nucleotide positions are labeled relative to the transcription start site (+1). The promoter regions of the indicated cell lines are identical except at the nucleotides shown.



Supplementary Figure S4. Additional live and fixed breast cancer cell localization data. A3B-eGFP (green) co-localizes with nuclear DNA (Hoescht-stained blue), whereas A3G-eGFP is cytoplasmic, in the indicated breast cancer cell lines. MDA-MB-468 shows some cytoplasmic A3B-eGFP localization, but is still predominantly nuclear. A3B-HA, A3G-HA, and A3F-HA (not shown) in fixed cells have localization patterns similar to those of live cell eGFP-tagged proteins. In many cases, A3B-HA is more nuclear, perhaps owing to background caused by internal translation initiation and cell-wide expression of the eGFP protein alone.

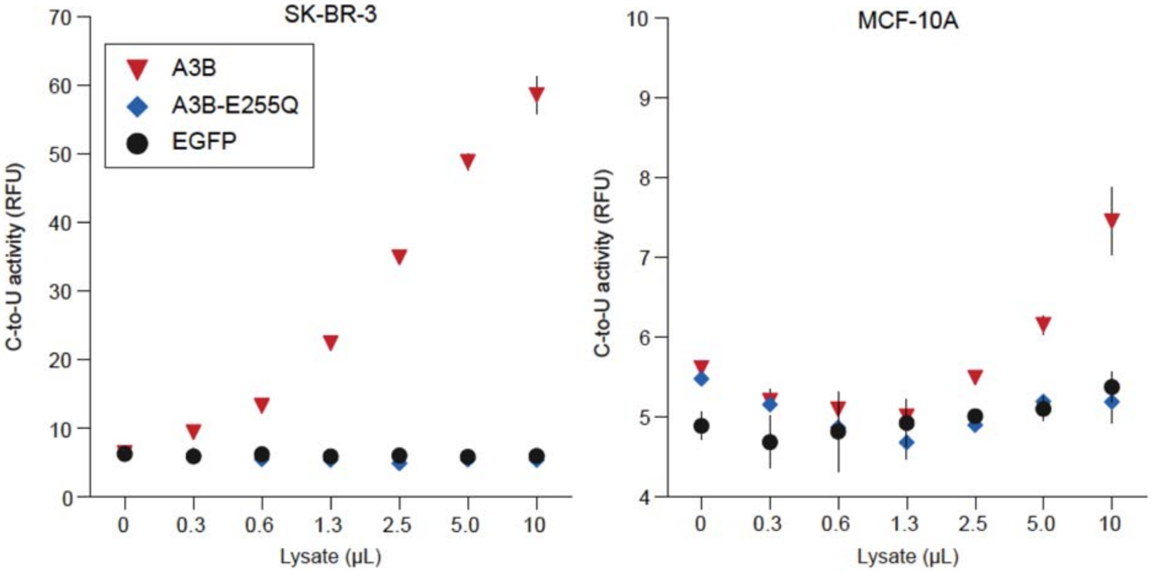


Supplementary Figure S5. A3B is active in the nuclear protein fraction of multiple breast cancer cell lines.

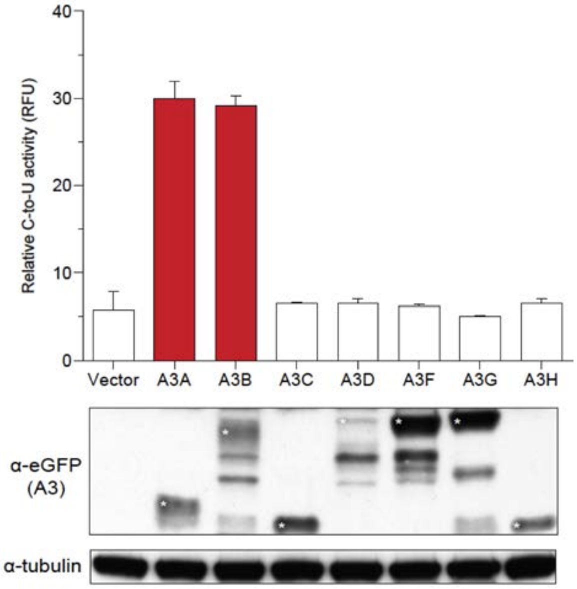
a, *A3* mRNA levels in the indicated breast cancer cell lines. Each column is mean +/- s.d. of three qPCR reactions presented relative to mRNA levels of the constitutive housekeeping gene *TBP*. Red and blue bars represent expression data from cells stably transduced with shControl or shA3B lentivirus, respectively.

b, A3B-dependent DNA deaminase activity in the nuclear (Nuc) and cytoplasmic (Cyt) fractions obtained from the cell lines in (a). The fractionation was cleaner in MDA-MB-453 and MDA-MB-468 lines than HCC1569, but all detectable deaminase activity was still dependent on A3B.

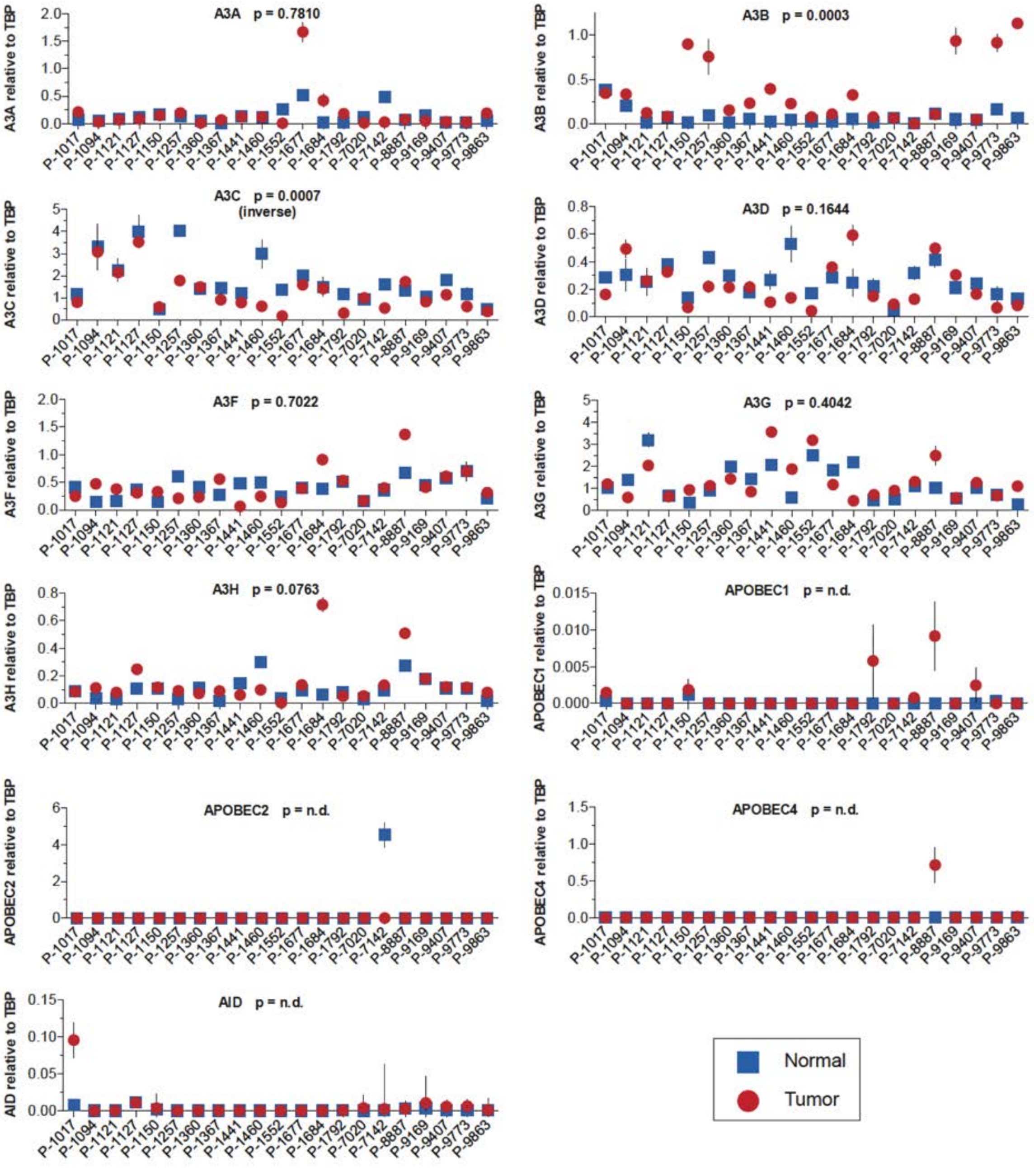
c, Immunoblots showing the distribution of histone H3, a nuclear protein, and tubulin, a cytoplasmic protein, in the protein preparations used in (b) to confirm efficient sub-cellular fractionation.



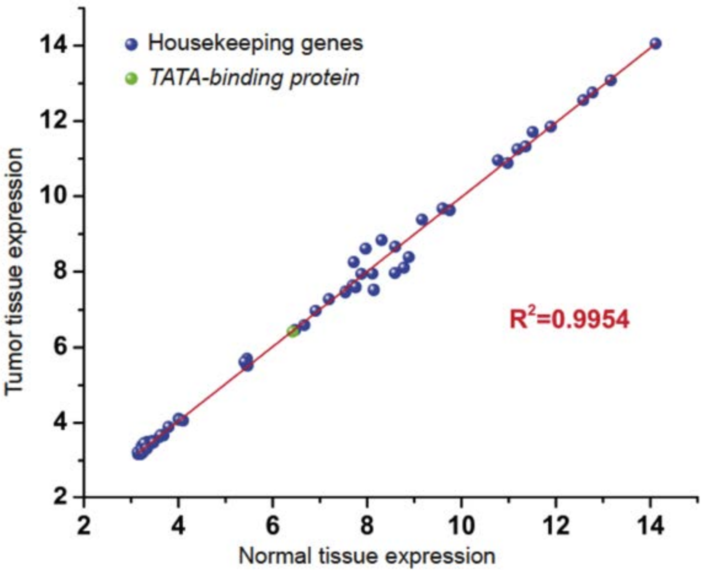
Supplementary Figure S6. DNA deaminase activity in A3B-low cell types.
Nuclear DNA C-to-U activity in extracts from SK-BR-3 and MCF10A transfected transiently with A3B-eGFP, A3B-E255Q-eGFP, or eGFP expression constructs. The higher activity levels in SK-BR-3 nuclear lysates are due to higher transfection efficiencies (30-40%), in comparison to MCF10A (1-5%). Mean values are shown with s.d. indicated unless smaller than the symbol (n=3).



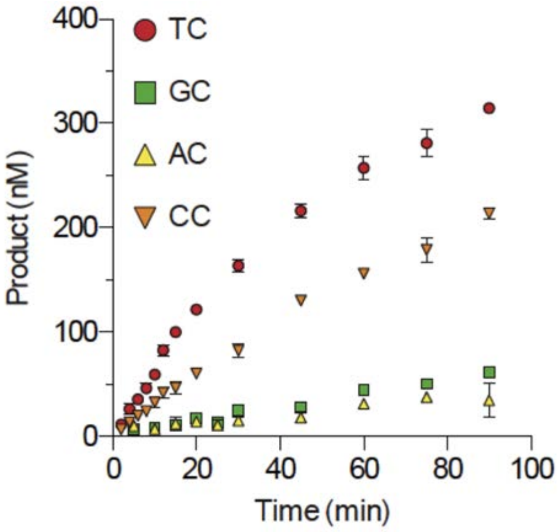
Supplementary Figure S7. Deaminase activity of HEK293T cell extracts with individual over-expressed A3 proteins.
Mean DNA C-to-U activity in whole cell extracts of HEK293T cells transfected with the indicated A3-eGFP expression constructs (n=3 per condition; s.d. shown). Activity was only detected in lysates from cells transfected with A3A- or A3B-eGFP. The corresponding anti-GFP immunoblot shows levels of each A3 (white asterisks), and the anti-tubulin blot indicates similar protein levels in each lysate.



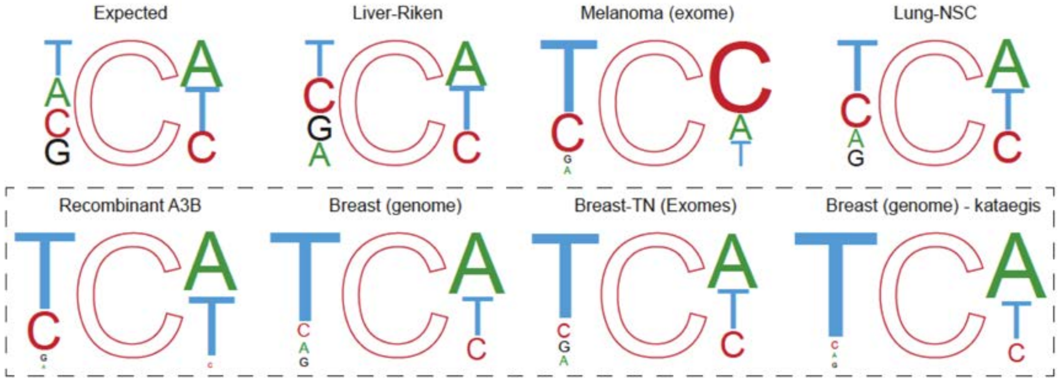
Supplementary Figure S8. Discovery data set - APOBEC family member expression profiles for 21 randomly selected sets of matched breast tumor and normal tissue. 21 representative breast tumor samples and the matched normal control tissues were used to synthesize cDNA for qPCR analyses of the full human *APOBEC* repertoire. Each data point is the mean mRNA level of three qPCR reactions presented relative to mRNA levels of the constitutive housekeeping gene *TBP* (s.d. shown as a bar unless smaller than the data point). P-values are indicated except those *AID, A1*, *A2*, and *A4* where the majority of samples had no detectable mRNA for these targets. *A3B* emerges as the only differentially up-regulated family member in tumor versus matched normal tissues. *A3C* shows an inverse correlation. Samples are presented in order of an arbitrarily assigned patient number. The *A3B* and *A3G* data were merged with 31 validation set samples for presentation in Fig. 4.



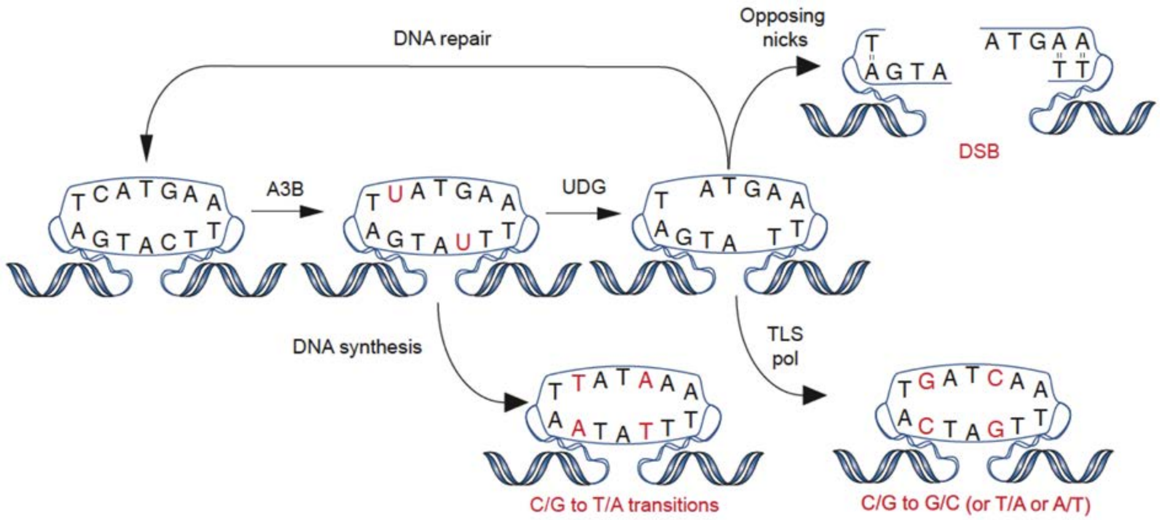
Supplementary Figure S9. Microarray housekeeping gene comparisons. See Supplementary Discussion and Methods for details.



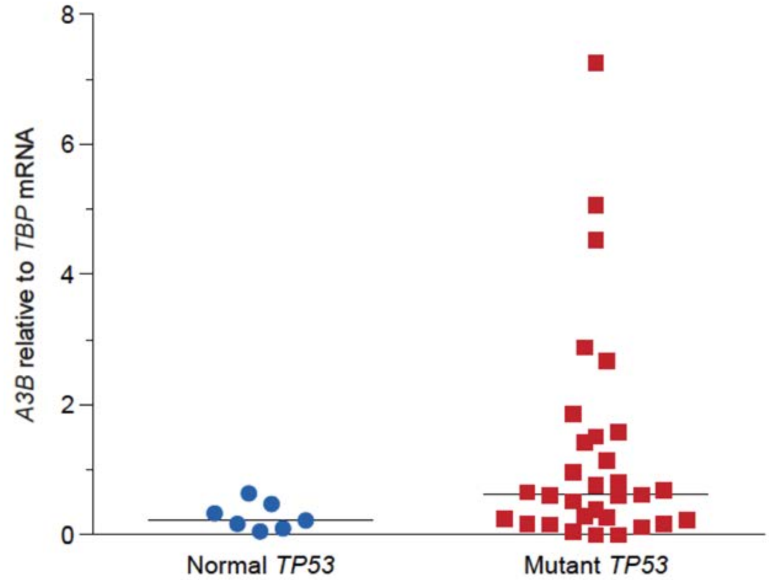
Supplementary Figure S10. A3B catalytic domain local deamination preferences.



Supplementary Figure S11. C-to-T transition mutation contexts *in vivo* versus *in vitro*. Data for non-small cell (NSC) lung cancer and breast cancer genomic kataegis events are included here, in addition to data from Fig. 4e, for comparison. The distribution of C-to-T transition mutation contexts in melanoma (Ref. #24) is consistent with the established mechanism of error-prone DNA polymerase misinsertion of A's opposite UV-induced pyrimidine dimers, which through replication or repair become transition mutations. In contrast, the distribution of C-to-T transition mutation contexts in liver and lung cancers (Refs. #25 & 26) strongly resembles the actual distribution of cytosines in the human genome (*i.e.*, the C-to-T transition mutations in these tumors appear context independent consistent with other underlying mutational mechanisms). In further contrast, the distribution of C-to-T transition mutation contexts in 21 full breast tumor genome sequences (Ref. #8) and 100 triple-negative breast tumor exome sequences (Ref. #9) strongly resembles the local deamination preference of recombinant A3B with prominent 5'-T and 3'-A or T (not C) flanking nucleotide biases. This bias is even further exaggerated in regions of kataegis, which may represent exposed single-stranded DNA regions *in vivo* that may be subjected to high-frequency A3B-dependent DNA deamination and different repair mechanisms.



Supplementary Figure S12. DNA deamination model for A3B in cancer.
Deamination of genomic DNA cytosines by up-regulated A3B leads to uracil lesions, which can be repaired faithfully or can lead to at least three possible outcomes: i) C-to-T transitions by direct DNA synthesis (lower outcome), ii) DNA double-stranded breaks by uracil excision and opposing abasic site cleavage (upper right outcome), and iii) transversions or transition mutations by error-prone DNA synthesis or aberrant repair (TLS pol = translesion synthesis DNA polymerase).



Supplementary Figure S13. ***A3B*** **up-regulation and** ***TP53*** **inactivation in the ATCC breast cancer cell line panel.**

A simple plot of *A3B* mRNA levels in *TP53* positive versus *TP53* mutant breast cancer cell lines from the ATCC (n=38; full list of cell lines in Supplementary Table S1).

14th September 2012

Angela Eggleston, Ph.D.
Senior Editor, *Nature*
a.eggleston@us.nature.com

Subject: Nature manuscript 2012-02-01827C/D

Dear Dr. Eggleston,

In addition to the revisions submitted one week ago, we have now been able to successfully mine some recently released TCGA data sets. We provide this new information here as a single pdf addendum to our revised manuscript (1 line of new main text and one new supplemental figure --- **Fig. S14**).

These data **FULLY SUPPORT** our current findings, as follows (letters match those in panels of the new **Fig. S14**):

a) A3B is up-regulated in ~50% of tumor vs matched normal tissues using RNAseq data as quantitative measures of gene expression levels;
b) A3G is not significantly different using the same data sets;
c) Tumor exome mutation loads correlate with A3B expression levels;
d) A3B mutation signature is enhanced in tumors with up-regulated A3B; and
e) A3B up-regulation correlates significantly with *TP53* inactivation.

We apologize for not discovering a way to do this sooner, but this delay is partly due to the fact that most of these data did not become publicly accessible until very recently (*i.e.*, to the best of our knowledge, the data were quietly released to the public sometime within the last 2 months). In any event, it is better late than never, and we strongly feel that these new supporting data combine to address the remaining comments of Reviewers 2 and 3.

Please forward this addendum to the reviewers, as these additional data add further clarity and confidence to our work.

We look forward to moving forward and communicating this story to the audiences of *Nature*.

Yours sincerely,
Reuben Harris

We have added one sentence (blue) to the following main text paragraph (page #6):

"Finally, we determined the impact of A3B on the breast tumor genome by correlating recombinant A3B's deamination signature *in vitro* and the somatic mutation spectra accumulated during tumor development *in vivo*. Using a series of single-stranded DNA substrates varying only at the immediate 5' or 3' position relative to the target cytosine (underlined), we found that recombinant A3B prefers TC>CC>GC=AC (**Fig. S10**; similar to endogenous A3B in **Fig. 1f**) and CA=CG=CT>CC (**Fig. 4c**). These local sequence contexts were then compared to those for C-to-T transitions reported for breast[8,9], melanoma[24], liver[25], and lung[26] tumors. Consistent with non-spontaneous origins, C-to-T transition loads are much greater in melanoma (~80%) and breast (~40%) than liver (~20%) and lung (~20%) tumors (**Fig. 4d**). The local sequence contexts for C-to-T transitions are even more striking, with flanking T and C in melanoma (TCC) and T and A in breast cancer (TCA) (**Fig. 4e & S11**). The melanoma pattern is expected due to error-prone DNA synthesis (A insertion) opposite UV-induced pyrimidine dimers. In contrast, the preferred context of C-to-T transitions in two independent breast cancer somatic mutation data sets, one including kataegis, mirrors the *in vitro* preference of recombinant A3B (**Fig. 4e & S11**). These results are fully supported by TCGA breast cancer data, which reveal positive correlations between *A3B* expression levels and exome mutation loads, overall C-to-T mutational signature, and *TP53* inactivation (**Fig. S14**)."

We have also added Fig. S14 (final page of this pdf) together with the following legend:

Supplementary Figure S14. Evidence for A3B up-regulation and genomic mutation in TCGA breast cancer data sets.

(**a**) *A3B* expression relative to *TBP* in 98 breast tumors (red circles) and matched normal tissues (blue squares) as determined by RNAseq data in the TCGA database (https://tcga-data.nci.nih.gov/tcga/). The red dashed line indicates 3 s.d. above the mean *A3B* relative to *TBP* level in the normal samples. Based on this stringent cut-off, *A3B* is up-regulated in 50% of tumors (p<0.0001 by Wilcoxon-signed-rank). These data sets were selected based on availability of matched tumor and normal RNAseq data and tumor exome sequencing data . *A3B* levels relative to *TBP* were calculated using the normalized counts from RNAseq for each gene.

(**b**) *A3G* expression in the same tumor and normal samples, determined as above. *A3G* expression is not significantly different between tumor and matched normal samples (p=0.0680 by Wilcoxon signed-rank).

(**c**) Dot plots of the total somatic mutation load per exome for the 98 samples in (a) and (b). The data were divided into three groups, as shown, from A3B-low to A3B-high. The median mutation loads for the lowest and the middle groups are not significantly different, whereas the median between the lowest and the highest A3B-expressing group is significantly different (median values of 32 and 68 somatic mutations per exome, respectively; p values determined by Mann Whitney U test).

(**d**) Nucleotide frequencies proportional to font size immediately 5' and 3' of genomic cytosines (expected) or the C-to-T mutated cytosine in the bottom and top one-third of A3B-expressing tumors from above. There are significant increases in both 5'T and 3'T in the A3B-high tumors, which is a signature that strongly resembles the A3B preferences *in vitro* and is apparent in other breast cancer genomic mutation data sets (Figs. 4 & S11).

(**e**) *A3B* levels are significantly higher in tumors with *TP53* mutations (p values determined by Mann Whitney U test). Exome sequences (MAF files) were used to identify potentially inactivating mutations in the *TP53* gene, which enabled the A3B data points to be separated into two groups. Silent mutations were ignored, and other mechanisms for TP53 inactivation were not considered (*e.g.*, MDM2 over-expression). These data are analogous to similar trends observed with breast cancer cell lines (Fig. S13).

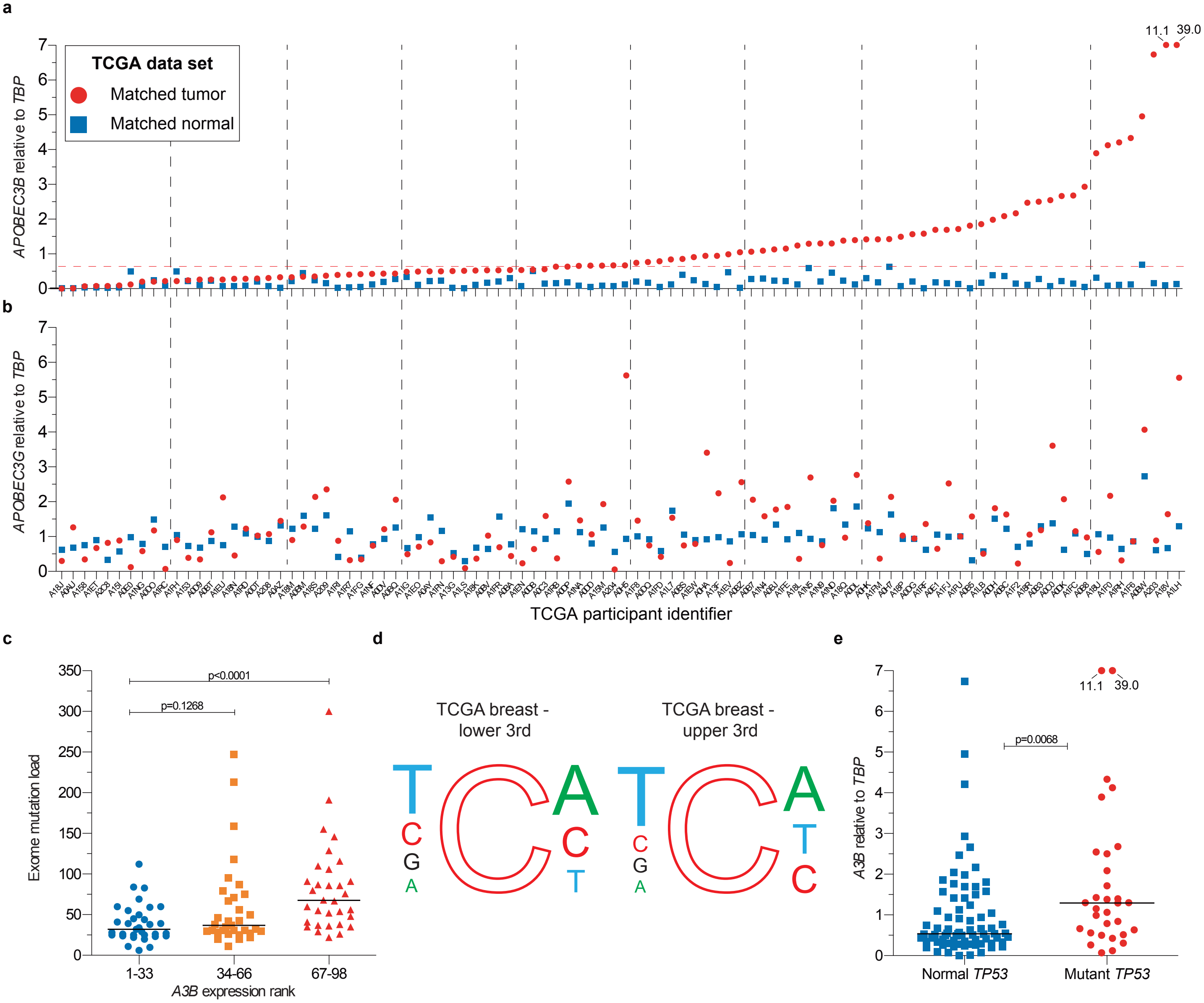

a
TCGA data set
Matched tumor
Matched normal
APOBEC3B relative to TBP
11.1
39.0
b
APOBEC3G relative to TBP
TCGA participant identifier
c
Exome mutation load
p<0.0001
p=0.1268
1-33
34-66
67-98
A3B expression rank
d
TCGA breast - lower 3rd
TCGA breast - upper 3rd
e
A3B relative to TBP
p=0.0068
11.1
39.0
Normal TP53
Mutant TP53

Fig. S14