AD_____

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 2013

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**
September 2013	Annual Summary	1 September 2012 – 31 August 2013

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 Bradley Burns

E-Mail: burn0230@umn.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

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

Over the course of my tenure on this training grant, I have discovered that an endogenous enzyme, APOBEC3B, is up-regulated in the majority of breast cancers. This holds true for both primary cancer tissues as well as cell lines. I have mechanistically characterized the enzyme's activity, sub-cellular localization, preferred sequence substrate and kinetics, and impact on the cancer cell genome. In addition, I have expanded my research to include all of the available cancer types represented in the publicly accessible portion of The Cancer Genome Atlas (TCGA). Using the massive data repository, I have found that the same mutational phenomenon that I have identified in breast cancer is also at work in 5 other types of cancer: head & neck cancer, cervical cancer, bladder cancer, squamous cell lung cancer and lung adenocarcinomas.

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	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U	UU	64	

Table of Contents

Introduction:

Cancer as a disease is defined by mutation. Without mutation, cancer cannot occur. In the specific 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 was to determine the impact of the endogenous DNA mutating enzyme, APOBEC3B (A3B), in human breast cancer and it has done so. The definitive manuscript characterizing the enzyme and connecting its mis-regulation to genetic heterogeneity in breast cancer was published in *Nature* on 21 February 2013. As part of this research was to uncover how, in breast cancer, this phenomenon is operating, I compared the mutation signatures and expression patterns of APOBEC3B-driven breast cancers to those of 18 other cancer types with data available from The Cancer Genome Atlas (TCGA). In doing so, I discovered that 5 other cancers are subject to the same mutagenic mechanism. This work was published in *Nature Genetics* 28 August 2013.

Body:

Over the course of this training grant, 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 *Nature* article. 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 *Nature* manuscript 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. 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.

Summary of Specific Aim 1: *Quantify the levels of APOBEC3B in 45 different breast cancer cell lines and 38 matched normal and cancer primary tissue samples* – All the stated goals for aim 1 were completed, in addition to substantial additional supporting work that was required to allow publication in a top-tier journal.

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. This portion of the research project, upon the defense of

my thesis, was passed along to another graduate student in the Harris lab, Mr. Brandon Leonard. This ensures that although the funding does not support Mr. Leonard or his research, the project is not discontinued with the training grant. Together we have developed a collaboration with Dr. Douglas Yee, the head of the Masonic Cancer Center, to perform the Xenograft assays outlined in this aim.

Summary of Specific Aim 2: *Determine whether APOBEC3B knockdown/knockout alters/abrogates/diminishes the tumor generating capacity of APOBEC3B overexpressing tumor cell lines* – This aim is currently in progress and will likely yield results within the next 6 months (despite the discontinuation of the grant) under the guidance of Mr. Brandon Leonard in the Harris lab.

Specific Aim 3 involved the over-expression of APOBEC3B in an epithelial cell line system. As was stated in the last report, there are reasons why this is an experiment that has logistical difficulties. Stable, forced over-expression of APOBEC3B in epithelial cell lines is lethal. Generation of clonal lines that constitutively over-express APOBEC3B is not possible, though we have generated cell lines that express it under the control of a tetracycline-inducible promoter. In these cases, over-expression results in cell death within eight days. We have reported these data in the appended *Nature* manuscript.

Summary of Specific Aim 3: *Ask whether APOBEC3B overexpression accelerates cancer progression* – As written, this SA is not tenable. We have reported the results (still useful in that they demonstrate the damage that unregulated APOBEC3B can do to the cellular genome) as part of the *Nature* manuscript described for Aim 1.

Specific Aim 4 has had some setbacks. The single female founder mouse used for our initial APOBEC3B transgenic colony was found to have leaky expression from the transgene. Our initial quality control testing was done using HEK293 cells with and without expression of cre recombinase. In that testing, the expression control was precise. It appears that, though the contruct works perfectly in human cells, there is likely a cryptic promoter when used in murine tissue. This unfortunate set-back has led us to re-design the original transgene using a different backbone. We are performing these experiments in collaboration with Dr. Hilde Nilsen at the University of Oslo. As this grant is terminating, this project has passed on the a scientist in the Harris lab, Ms. Emily Law. Again, though this funding source is no longer going to support me, the project will still continue.

Summary of Specific Aim 4: *Determine the ability of APOBEC3B to generate breast cancer in a transgenic mouse model* – Unforeseen technical details have slowed progress on this avenue of the project, but it is still in progress and will continue after this grant has finished.

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
- Discovery of the APOBEC3B mutagenic signature in 5 additional cancer types

Reportable Outcomes:

Over the course of this grant, this research has resulted two high-profile published manuscripts (one in *Nature* and the other in *Nature Genetics*), numeous 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. I presented my work at the Mechanisms and Models of Cancer conference at Cold Spring Harbor in 2012 (as a poster). It has also resulted in the awarding of my Ph.D. (tentative official award date of 30

September 2013). I was also awarded a position as a Howard Hughes Medical Institute (HHMI) post-doctoral fellow under Dr. Robin Wright at the University of Minnesota, in part, due to the experience awarded by this funding.

Conclusion:

Over the past two years of funding, I have been able to discover and characterize a previously unappreciated source of mutation in human breast cancer, as well as in several other cancers. Aside from the obvious clinical implications, these findings were important enough to warrant publication in two of the premier scientific journals and to result in my being awarded my Ph.D. and an HHMI post-doctoral fellowship for the coming year. While not all of the specific aims were addressed in their entirety, the main goal of the research project, characterizing APOBEC3B and determining its role as a mutator in breast cancer, was a phenomenal success.

References:

Manuscripts

1) Burns, M. B. *et al*., APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* 494, 366-370, doi:10.1038/nature11881 (2013).
2) Burns, M. B., Temiz, N. A. & Harris, R. S. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nature genetics*, doi:10.1038/ng.2701 (2013).

Conference Abstracts

3) Burns M.B., Enzyme catalyzed DNA deamination in multiple human cancers. (Poster) Mechanisms and Models of Cancer, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, August 14-18th, 2012.

Abstract

Multiple mutations are required for cancer development, and deep sequencing has revealed that many cancers harbor staggering numbers of mutations. The underlying mutation spectra are often dominated by both dispersed and clustered C/G-to-T/A transition mutations, suggesting a common and non-spontaneous origin. We present evidence for APOBEC3-dependent DNA deamination in multiple human cancers. Gene expression profiling shows APOBEC3 expression preferentially in tumors in comparison to normal tissues. Knockdown experiments demonstrate that APOBEC3 causes elevated levels of steady state genomic uracil, increased mutation frequencies, and hallmark C/G-to-T/A genomic mutations. Our data are consistent with a model in which enzyme-catalyzed DNA deamination provides mutational fuel for multiple human cancers.

Appendices:

The appendices include the full versions of two manuscripts – one published in *Nature* and the other in *Nature Genetics*. Full supplemental materials are also included.

LETTER

doi:10.1038/nature11881

APOBEC3B is an enzymatic source of mutation in breast cancer

Michael B. Burns[1,2,3,4]*, Lela Lackey[1,2,3,4]*, Michael A. Carpenter[1,2,3,4], Anurag Rathore[1,2,3,4], Allison M. Land[1,2,3,4], Brandon Leonard[2,3,4,5], Eric W. Refsland[1,2,3,4], 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[1,2,3,4], William L. Brown[1,2,3,4], Emily K. Law[1,2,3,4] & Reuben S. Harris[1,2,3,4,5]

Several mutations are required for cancer development, and genome sequencing has revealed that many cancers, including breast cancer, have somatic mutation spectra dominated by C-to-T transitions[1–9]. Most of these mutations occur at hydrolytically disfavoured[10] non-methylated cytosines throughout the genome, and are sometimes clustered[8]. Here we show that the DNA cytosine deaminase APOBEC3B is a probable source of these mutations. *APOBEC3B* messenger RNA is upregulated in most primary breast tumours and breast cancer cell lines. Tumours that express high levels of *APOBEC3B* have twice as many mutations as those that express low levels and are more likely to have mutations in *TP53*. Endogenous APOBEC3B 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 APOBEC3B correlates with increased levels of genomic uracil, increased mutation frequencies, and C-to-T transitions. Furthermore, induced APOBEC3B overexpression causes cell cycle deviations, cell death, DNA fragmentation, γ-H2AX accumulation and C-to-T mutations. Our data suggest a model in which APOBEC3B-catalysed deamination provides a chronic source of DNA damage in breast cancers that could select *TP53* inactivation and explain how some tumours evolve rapidly and manifest heterogeneity.

Most humans encode a total of 11 polynucleotide cytosine deaminase family members that could contribute to mutation in cancer—APOBEC1, activation-induced deaminase (AID), APOBEC2, APOBEC3 proteins (known as A3A, A3B, A3C, A3D, A3F, A3G and A3H), and APOBEC4. APOBEC2 and APOBEC4 have not shown activity. APOBEC1 and AID are expressed tissue specifically and implicated in cancers of those tissues, hepatocytes and B cells, respectively[11,12]. We therefore proposed that one or more of the seven APOBEC3 proteins may be responsible for the C-to-T mutations in other human cancers. This possibility is consistent with hybridization[13] and expression studies[14] (Supplementary Fig. 1).

To identify the contributing APOBEC3 protein, we quantified mRNA levels for each of the 11 family members in breast cancer cell lines (Supplementary Fig. 2). Surprisingly, only *APOBEC3B* mRNA trended towards upregulation. This analysis was expanded to include a total of 38 independent breast cancer cell lines. *APOBEC3B* was upregulated by ≥3 s.d. relative to controls in 28 out of 38 lines, with levels exceeding tenfold in 12 out of 38 lines (Fig. 1a and Supplementary Table 1). Of the representative cell lines used, MDA-MB-453, MDA-MB-468 and HCC1569 showed 20-, 21- and 61-fold upregulation, respectively. These results correlate with cell-line microarray data (Supplementary Fig. 3, Supplementary Tables 2–9 and Supplementary Discussion). *APOBEC3B* upregulation is probably 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 (Supplementary Fig. 4).

Epitope-tagged APOBEC3B (A3B) localizes to the nucleus of several transfected cell types[15]. To determine whether this is also a property of breast cancer lines, a construct encoding A3B fused to enhanced green fluorescent protein (A3B–eGFP) was transfected into MDA-MB-453,

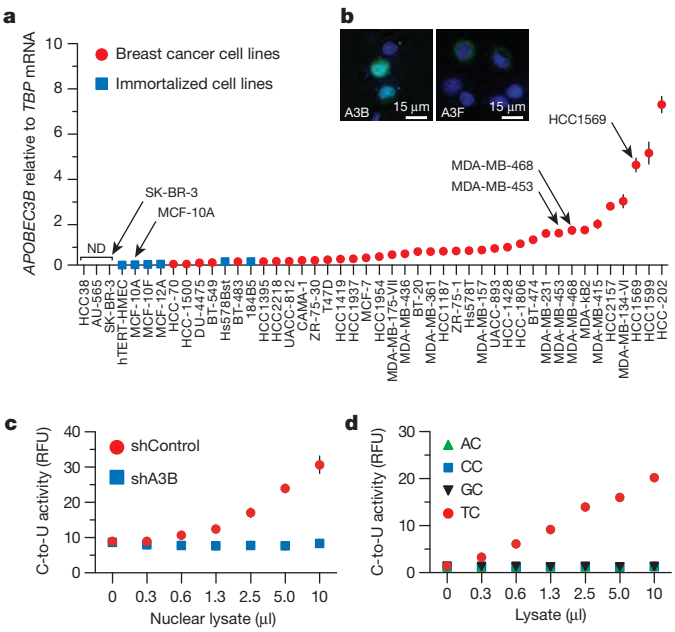


Figure 1 | *APOBEC3B* upregulation and activity in breast cancer cell lines. **a**, *APOBEC3B* levels in indicated cell lines. Each point represents the mean of three reactions presented relative to *TBP* (s.d. shown unless smaller than symbol). ND, not detected. **b**, A3B–eGFP or A3F–eGFP localization in MDA-MB-453 cells (nuclei are blue). **c**, Nuclear DNA C-to-U activity in extracts from MDA-MB-453 transduced with shControl or shA3B lentiviruses ($n = 3$; s.d. shown unless smaller than symbol). RFU, relative fluorescence units. **d**, Intrinsic dinucleotide DNA deamination preference of endogenous A3B in extracts from MDA-MB-453 cells ($n = 3$; s.d. smaller than symbols).

[1]Biochemistry, Molecular Biology and Biophysics Department, University of Minnesota, Minneapolis, Minnesota 55455, USA. [2]Masonic Cancer Center, University of Minnesota, Minneapolis, Minnesota 55455, USA. [3]Institute for Molecular Virology, University of Minnesota, Minneapolis, Minnesota 55455, USA. [4]Center for Genome Engineering, University of Minnesota, Minneapolis, Minnesota 55455, USA. [5]Microbiology, Cancer Biology and Immunology Graduate Program, University of Minnesota, Minneapolis, Minnesota 55455, USA. [6]Department of Medicinal Chemistry, University of Minnesota, Minneapolis, Minnesota 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, Maryland 21702, USA.
*These authors contributed equally to this work.

MDA-MB-468 and HCC1569 cells. Live cell images showed nuclear localization of A3B–eGFP, in contrast to the cytoplasmic localization of an A3F–eGFP construct (Fig. 1b and Supplementary Fig. 5). Corroborating data were obtained for haemagglutinin (HA)-tagged proteins (Supplementary Fig. 5). To study endogenous A3B subcellular compartmentalization and activity, we used a fluorescence-based DNA C-to-U assay. We first found that nuclear, but not cytoplasmic, fractions of several breast cancer cell lines contain a robust DNA editing activity, which could be ablated by *APOBEC3B* knockdown (Fig. 1c and Supplementary Figs 6 and 7). Similar results were obtained with an independent knockdown construct (not shown). Protein extracts were then used to assess the local dinucleotide deamination preference of endogenous A3B. Similar to retroviral hypermutation signatures caused by A3B overexpression[16], endogenous A3B showed a strong preference for editing cytosines in the TC dinucleotide context (Fig. 1d and Supplementary Fig. 6). No deaminase activity was observed for extracts from MCF-10A ($A3B^{low}$) or SK-BR-3 ($A3B^{null}$) cells, although it could be conferred by transient A3B transfection (Supplementary Fig. 8). Both A3B–HA and A3A–HA could elicit measurable TC-to-TU activity in lysates from transfected HEK293T cells (Supplementary Fig. 9). However, because *APOBEC3A* mRNA is myeloid lineage-specific[17] and non-detectable in breast cancer cell lines (Supplementary Figs 1 and 2), our expression and activity studies indicated that A3B may be the only enzyme poised to deaminate breast cancer genomic DNA.

To address whether endogenous A3B damages genomic DNA, we used 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 cells with high levels of endogenous A3B versus knockdown levels of A3B (short hairpin RNA (shRNA) control versus shRNA against *APOBEC3B* (shA3B)) (Fig. 2a and Supplementary Fig. 10). Genomic uracil loads decreased by 30% in HCC1569 cells expressing shA3B and by 70% in MDA-MB-453 cells, in which knockdown was stronger (Fig. 2b and Supplementary Fig. 10). Although these relative differences may seem modest—10 and 20 uracils per megabases (Mb), respectively—this equates to 30,000 and 60,000 A3B-dependent uracils per haploid genome. The actual number of pro-mutagenic uracils may be even higher because several repair pathways may concurrently function to limit this damage.

Second, we used a thymidine kinase-positive (TK^{plus}) to -negative (TK^{minus}) fluctuation analysis[17] to determine whether upregulated A3B and increased uracil loads lead to higher levels of mutation. MDA-MB-453 and HCC1569 cells were engineered to express the herpes simplex virus type 1 *TK* gene, 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 subclones. Expanded subclones were subjected to ganciclovir selection and resistant cells were grown to visible colonies, which showed that cells with upregulated A3B accumulate 3–5-fold more mutations (Fig. 2c and Supplementary Fig. 10).

Third, differential DNA denaturation PCR (3D-PCR)[17,18] was used to determine whether C-to-T transition mutations accumulate differentially at three genomic loci 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 A/T content amplify at lower denaturation temperatures than parental sequences. Lower temperature amplicons were observed for *TP53* and *c-MYC*, but not *CDKN2B* (Fig. 2d and Supplementary Fig. 10). These amplicons were cloned and sequenced, and more C-to-T transition mutations were observed in $A3B^{high}$ compared with $A3B^{low}$ samples (Fig. 2d and Supplementary Fig. 10). *TP53* and *c-MYC* appeared more mutable than *CDKN2B*, suggesting that all genomic regions are not equally susceptible to enzymatic deamination. Other base substitution mutations were rare, and some C-to-T transitions were still evident in the $A3B^{low}$ samples, possibly owing to residual deaminase activity and/or amplification of spontaneous events.

To address whether A3B triggers other cancer hallmarks[19], we tried and failed to stably express A3B in several epithelial cell lines. We therefore 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 (Supplementary Fig. 11). A3A overexpression caused rapid S-phase arrest, cytotoxicity and γ-H2AX focus formation, as reported previously[20] (Fig. 3a–c and Supplementary Fig. 11). 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. 3a, b and Supplementary Fig. 11). A3B induction also caused γ-H2AX focus formation, DNA fragmentation, as evidenced by visible comets, and C-to-T mutations (Fig. 3c–e). A3B catalytic activity, as evidenced by the glutamate mutants, was required for the induction of these cancer phenotypes.

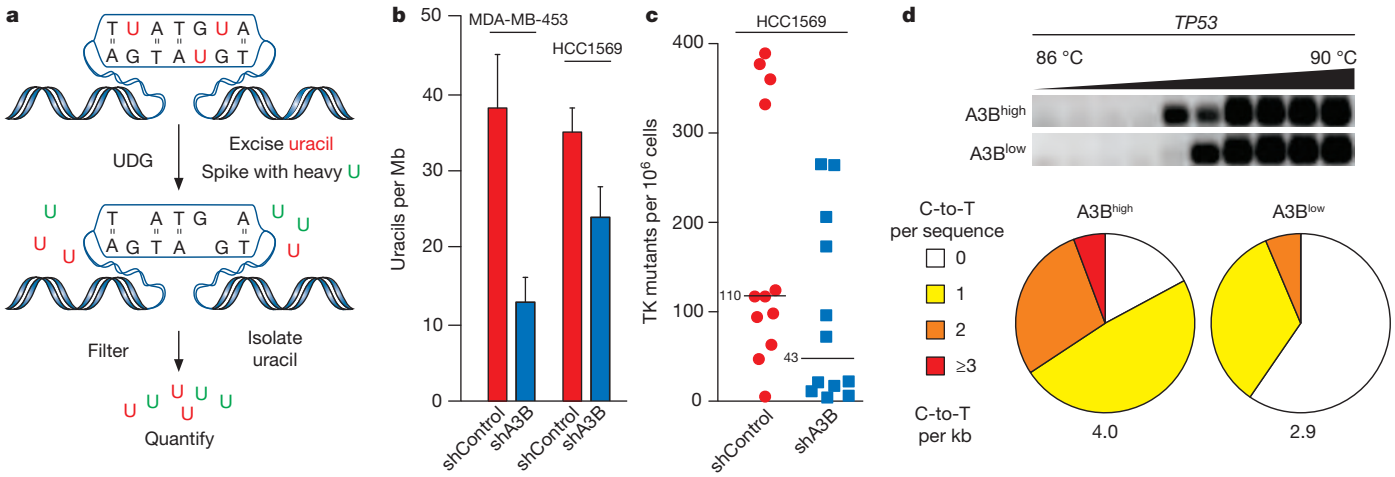


Figure 2 | A3B-dependent uracil lesions and mutations in breast cancer genomic DNA. **a**, Workflow for genomic uracil quantification by high-performance liquid chromatography–tandem mass spectrometry (HPLC-ESI-MS/MS). **b**, Average uracil loads in the indicated cell lines ($n = 3$; errors, s.d.). **c**, Dot plot representing thymidine kinase mutant frequencies of HCC1569 subclones expressing shControl or shA3B. Each dot corresponds to one subclone. Medians are labelled. **d**, Agarose gel and mutation analysis of *TP53* 3D-PCR amplicons from HCC1569 cells expressing shControl ($A3B^{high}$) or shA3B ($A3B^{low}$) ($n \geq 35$ sequences per condition). See Supplementary Fig. 10 for further data.

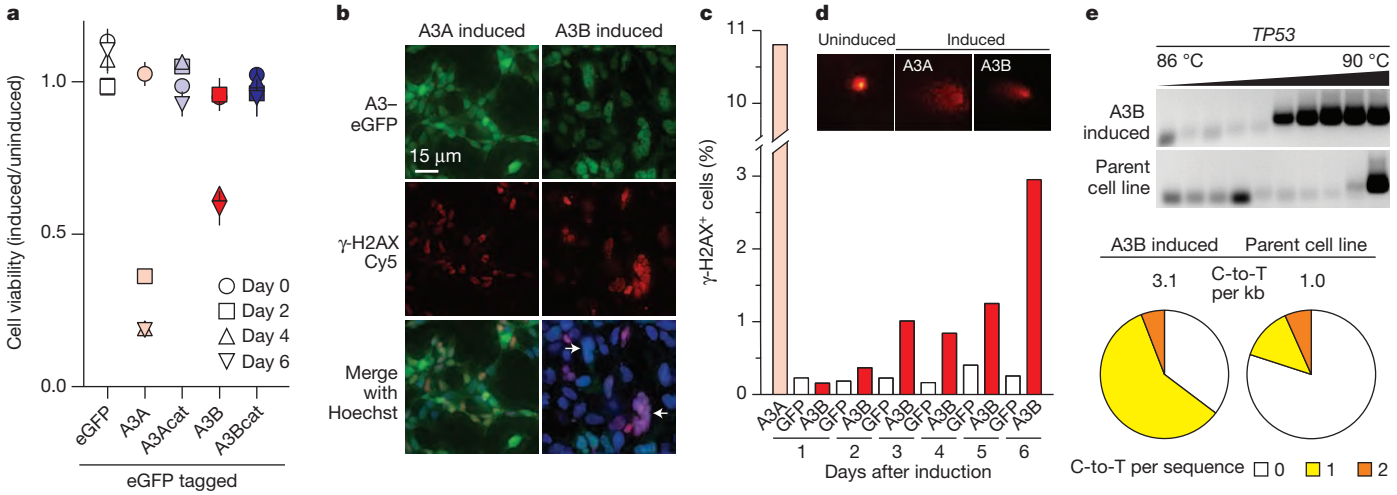


Figure 3 | Cancer phenotypes triggered by inducing A3B overexpression. **a**, Cell viability at indicated times after induction (mean and s.d. for $n = 3$ per condition). A3Acat and A3Bcat denote catalytically defective glutamate mutants. **b**, **c**, Representative fields of cells imaged for γ-H2AX and A3A–eGFP (1 day) or A3B–eGFP (3 days) after induction, and γ-H2AX quantification. Abnormal, multinuclear clusters are typical of induced A3B–eGFP (white arrows). **d**, Representative images of A3-induced DNA comets (original magnification, ×400). **e**, C-to-T mutations in *TP53* detected by sequencing 3D-PCR products 4 days after induction ($n > 12$ sequences per condition).

We next asked whether our cell-based results could be extended to primary tumours. First, we quantified mRNA levels for each of the 11 family members in 21 randomly chosen breast tumour specimens, in parallel with matched normal tissue procured simultaneously from an adjacent area or the contralateral breast. Only *APOBEC3B* was expressed preferentially in tumours ($P = 0.0003$) (Supplementary Fig. 12). We confirmed this analysis by measuring *APOBEC3B* levels in 31 additional tumour/normal matched tissue sets. In total, *APOBEC3B* was upregulated by ≥3 s.d. in 20 out of 52 tumours in comparison to the patient-matched normal tissue mean, and in 44 out of 52 tumours in comparison to the reduction mammoplasty tissue mean (Fig. 4a; $P = 7.1 \times 10^{-7}$ and $P = 2 \times 10^{-5}$, respectively; patient information in Supplementary Table 10). These are underestimates because tumour specimens have varying fractions of non-*APOBEC3B*-expressing normal cells. Some of the matched 'normal' samples may also be contaminated by tumour cells, as judged by the mean levels in mammoplasty samples (Fig. 4a; $P = 0.002$). The related deaminase, *APOBEC3G*, was not expressed differentially in these samples, indicating that these observations are not due to immune cells known to express several APOBEC3 proteins[14] ($P = 0.591$). Similar results were obtained by quantifying RNA-sequencing data for independent matched tumour and normal pairs[21], with ~50% showing upregulated A3B (defined as tumours with A3B levels >3 s.d. above the mean of the normal matched samples; $P < 0.0001$) (Fig. 4b).

Finally, we assessed the effect of A3B on the breast tumour genome by correlating the deamination signature of A3B *in vitro* and the somatic mutation spectra accumulated during tumour 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 and CA>CT = CC (Supplementary Fig. 13). These local sequence preferences were then compared to the expected distribution of cytosine in the human genome and the reported C-to-T mutation profiles for melanoma[22], liver[23] and breast[8,9,21] tumours. Consistent with a spontaneous origin, the C-to-T frequency is low in liver tumours (~20%) and mutational events appear random (Fig. 4c, d). As expected, C-to-T frequencies are high in melanomas (~80%) and focused at dipyrimidines consistent with ultraviolet-induced lesions and subsequent error-prone lesion bypass synthesis (Fig. 4c, d). Interestingly, the C-to-T frequency was intermediate in three independent breast tumour data sets (~40%) and largely focused at trinucleotides that mimic the preferred sites for A3B-dependent DNA deamination *in vitro* (Fig. 4c, d and Supplementary Fig. 13). The availability of both high-throughput RNA sequencing (RNA-seq) and somatic mutation data[21] also enabled the establishment of strong positive correlations between *APOBEC3B* expression levels and the C-to-T mutation load, overall base substitution mutation load, and *TP53* inactivation (Fig. 4e–g). Importantly, tumours expressing high A3B levels have twice as many mutations (Fig. 4e, f and Supplementary Fig. 14). This equates to 10 C-to-T and 30 total mutations per exome, or approximately 1,000 and 3,000 mutations per genome, being attributable to A3B.

Taken together, we conclude that A3B is an important mutational source in breast cancer accounting for C-to-T mutation biases and increased mutational loads. Moreover, the disproportional increase in overall base substitutions indicates that some of these other patterns may be due to further processing of U/G mispairs by 'repair' enzymes into transitions, transversions and DNA breaks that could precipitate chromosomal rearrangements (model in Supplementary Fig. 15 with similarities to AID-dependent antibody diversification mechanisms[24]). Future work is needed to understand A3B regulation and the potential interaction with other oncogenes and tumour suppressors. For example, although several common breast cancer markers do not correlate with *APOBEC3B* upregulation, a mechanistic linkage between increased *APOBEC3B* and inactivated *TP53* is evident in primary tumour data and cell lines (Fig. 4g and Supplementary Fig. 16). *TP53* inactivation may be required to allow cells to bypass DNA damage checkpoints triggered by A3B.

This is the first study, to our knowledge, to demonstrate upregulation of the DNA deaminase A3B in breast cancer and reveal it as a considerable source of enzymatic mutation. Conceptually supportive of the original mutator hypothesis[25], A3B-catalysed genomic DNA deamination could provide genetic fuel for cancer development, metastasis, and even therapy resistance. We propose that A3B is a dominant underlying factor that contributes to tumour heterogeneity by broadly affecting several pathways and phenotypes. A3B may represent a new marker for breast cancer and a strong candidate for targeted intervention, especially given its non-essential nature[26]. A3B inhibition may decrease the rate of tumour evolution and stabilize the targets of existing therapeutics.

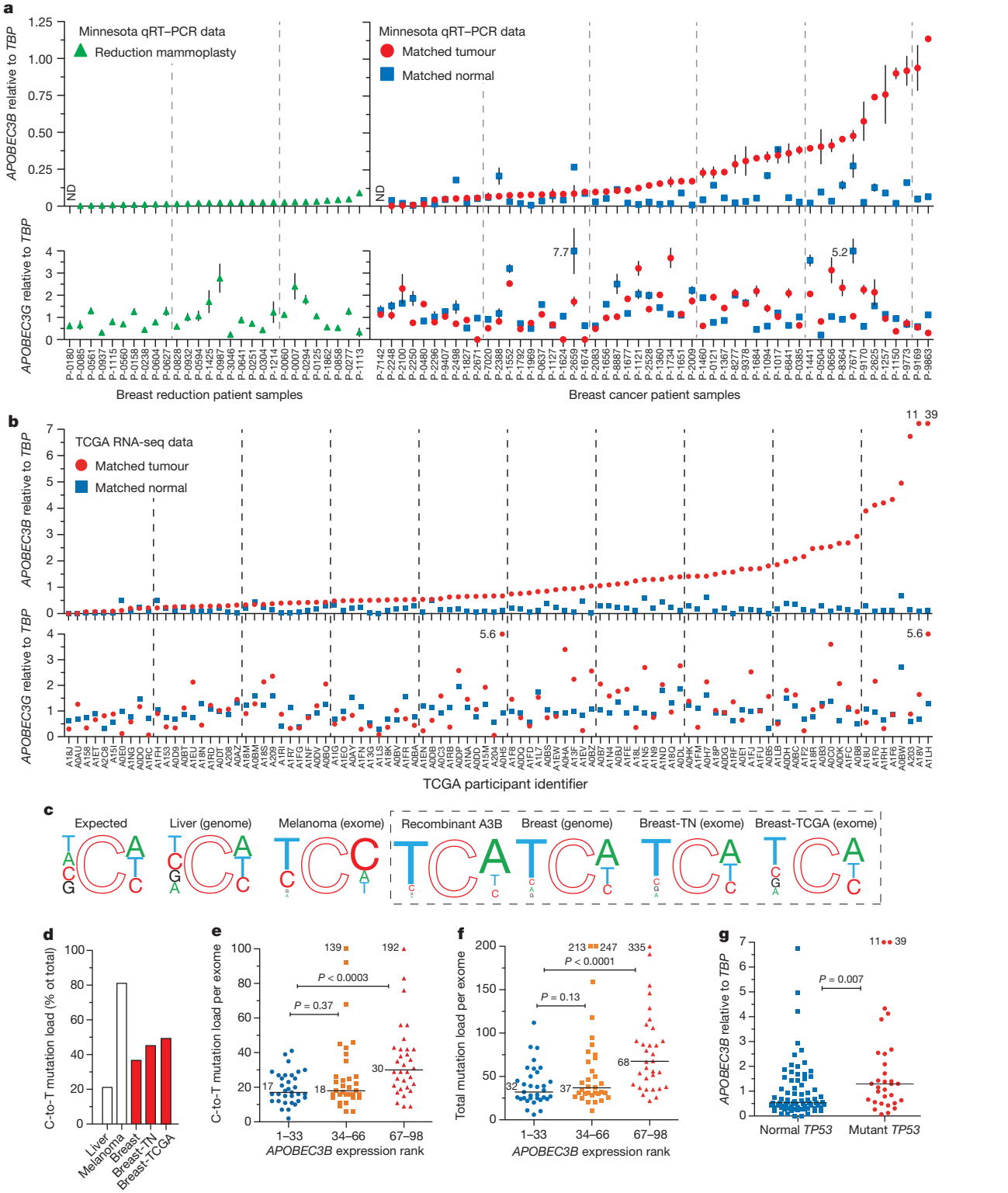


Figure 4 | *APOBEC3B* upregulation and mutation in breast tumours. **a**, *APOBEC3B* and *APOBEC3G* mRNA levels in the indicated tissues. Each symbol represents the mean mRNA level of three quantitative PCR with reverse transcription (qRT–PCR) reactions, presented relative to *TBP* (s.d. shown unless smaller than symbol). **b**, RNA-seq data for *APOBEC3B* and *APOBEC3G* in the indicated samples. TCGA, The Cancer Genome Atlas. **c**, Local sequence contexts for all genomic cytosines (expected), cytosines deaminated by recombinant A3B (Supplementary Fig. 13), and observed C-to-T transitions in the indicated cancers. Font size is proportional to nucleotide frequency. TN, triple-negative. **d**, Percentage of C-to-T mutations in the indicated tumours. **e**, **f**, C-to-T (**e**) and total (**f**) mutation counts for tumours in **b** grouped into lower, middle and upper thirds based on *APOBEC3B* levels (medians are labelled). **g**, Relationship between *APOBEC3B* level (RNA-seq data) and *TP53* status for tumours in **b**. Off-scale values in **a**, **b**, **e**–**g** are indicated numerically; *P* values in **e**–**g** are from Mann–Whitney U test.

METHODS SUMMARY

Flash-frozen tissues were obtained from the University of Minnesota Tissue Procurement Facility. Availability of both tumour and matched normal tissue was the only selection criteria. Mammary reduction samples were used as non-cancer controls. These studies were performed in accordance with Institutional Review Board (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, complementary DNA synthesis, and quantitative PCR procedures were performed as reported[14] (Supplementary Table 11). Knockdown and control shRNA constructs were obtained from Open Biosystems. Microscopy, cellular fractionation and deaminase activity assays were done as described[15,17]. Genomic uracil was quantified by treating DNA samples with uracil DNA glycosylase, purifying the nucleobase away from the remaining DNA, and analysing the samples by mass spectrometry. The *TK* and 3D-PCR mutation assays have been described and were modified for use with breast cancer cell lines[17]. Dox-inducible cells were obtained from Invitrogen and stable derivatives were created with the indicated constructs. These lines were analysed 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[27] 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 by compiling published primary tumour genomic, exomic or RNA sequencing data[8,9,21–23]. Potential mechanistic overlap with hydrolytic deamination of 5-methyl-cytosines was avoided by excluding CpG dinucleotides from mutational preference calculations.

Full Methods and any associated references are available in the online version of the paper.

Received 9 February; accepted 24 December 2012.
Published online 6 February 2013.

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. USA* **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. Pavri, R. & Nussenzweig, M. C. AID targeting in antibody diversity. *Adv. Immunol.* **110,** 1–26 (2011).
12. Yamanaka, S. *et al.* Apolipoprotein B mRNA-editing protein induces hepatocellular carcinoma and dysplasia in transgenic animals. *Proc. Natl Acad. Sci. USA* **92,** 8483–8487 (1995).
13. 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).
14. 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).
15. Lackey, L. *et al.* APOBEC3B and AID have similar nuclear import mechanisms. *J. Mol. Biol.* **419,** 301–314 (2012).
16. 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).
17. Stenglein, M. D., Burns, M. B., Li, M., Lengyel, J. & Harris, R. S. APOBEC3 proteins mediate the clearance of foreign DNA from human cells. *Nature Struct. Mol. Biol.* **17,** 222–229 (2010).
18. 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. USA* **108,** 4858–4863 (2011).
19. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
20. 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).
21. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
22. Wei, X. *et al.* Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nature Genet.* **43,** 442–446 (2011).
23. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database* **2011,** bar026 (2011).
24. Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* **76,** 1–22 (2007).
25. Loeb, L. A., Springgate, C. F. & Battula, N. Errors in DNA replication as a basis of malignant changes. *Cancer Res.* **34,** 2311–2321 (1974).
26. 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).
27. Carpenter, M. A. *et al.* Methylcytosine and normal cytosine deamination by the foreign DNA restriction enzyme APOBEC3A. *J. Biol. Chem.* **287,** 34801–34808 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank J. Hultquist and R. Vogel for statistics, T. Hwang for bioinformatic assistance, V. Polunovsky for hTERT-HMECs, V. Simon for shRNA, S. Kaufmann, C. Lange and D. Largaespada for consultation, and the Masonic Cancer Center Breast Cancer Research Fund for purchasing the ATCC breast cancer panel. Tissues were obtained from the Masonic Cancer Center Tissue Procurement Facility, which is part of BioNet, supported by the Academic Health Center and National Institutes of Health (NIH) grants P30 CA77598 (D.Y.), 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.L. was supported in part by a National Science Foundation Predoctoral Fellowship and by a position on the Institute for Molecular Virology Training Grant NIH T32 AI083196. M.A.C. was supported by an NIH postdoctoral fellowship (F32 GM095219). A.M.L. was supported by a CIHR postdoctoral fellowship. E.W.R. was supported by a position on the Institute for Molecular Virology Training Grant NIH T32 AI083196 and subsequently by an NIH predoctoral fellowship (F31 DA033186). Computational analyses (N.A.T. and D.E.D.) were supported by federal funds from the National Cancer Institute, NIH, CBIIT/caBIG ISRCE yellow task 09-260. The Harris laboratory was supported in part by NIH R01 AI064046, NIH P01 GM091743, the Children's Cancer Research Fund, and a seed grant from the University of Minnesota Clinical and Translational Science Institute (supported by NIH 1UL1RR033183).

Author Contributions R.S.H. conceived and managed the overall project. M.B.B. assisted R.S.H. with experimental design, project management and manuscript preparation. M.B.B., E.W.R. and B.L. generated mRNA expression profiles; L.L. and E.K.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 thymidine kinase fluctuations; A.R. generated 3D-PCR sequences; and L.L., A.M.L., A.R. and M.A.C. determined the effect of induced A3B overexpression. M.A.C. performed deaminase assays with recombinant protein; and M.A.C. and D.K. assisted with the HPLC-ESI-MS/MS set up. N.T. was involved in HPLC-ESI-MS/MS method development. J.B.N. conducted the search and performed the bioinformatic analysis of the microarray data and developed the normalization algorithm for this analysis. N.A.T., D.E.D. and M.B.B. contributed bioinformatic analyses. All authors contributed to manuscript revisions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to R.S.H. (rsh@umn.edu).

METHODS

RNA isolation, cDNA synthesis and qRT–PCR. Matched tumour/normal breast tumours and mammary reduction samples from the University of Minnesota Tissue Procurement Facility and breast cancer cell lines 30-4500K from the ATCC were used for RNA isolation, cDNA synthesis and qPCR as described[14]. 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 Light-cycler 480 instrument. The housekeeping gene *TBP* was used for normalization. Statistical analyses for matched tissues were done using the Wilcoxon signed-rank test, and unmatched sets with the Mann–Whitney U-test (Graphpad Prism). Primer and probe sequences are listed in Supplementary Table 11.

Knockdown constructs. *APOBEC3B* shRNA and shControl lentiviral constructs were from Open Biosystems (TRCN0000157469, TRCN0000140546 and scramble). Knockdown levels ranged from 80% to 95% by qRT–PCR. Helper plasmids pΔ-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 collected and concentrated by centrifugation (14,000*g* for 2 h). Stable transductants were selected with puromycin (1 μg ml^{-1}).

Cell fractionation and DNA deaminase activity assays. Subcellular activity analysis and dinucleotide preferences were measured as described[27,28]. In brief, cellular fractionation was performed by syringe treatment of 10^7 cells in 0.5 ml of hypotonic buffer[28]. Nuclei were lysed by sonication in lysis buffer (25 mM HEPES, pH 7.4, 250 mM NaCl, 10% glycerol, 0.5% Triton X-100, 1 mM EDTA, 1 mM $MgCl_2$ and 1 mM $ZnCl_2$). Anti-histone H3 (1:2,000; Abcam) and anti-tubulin (1:10,000; Covance) followed by anti-mouse 800 or anti-rabbit 680 (1:5,000; Licor) immunoblots were used to assess fractionation. Lysates were tested in a fluorescence-based deaminase activity assay[17]. Dilutions were incubated 2 h at 37 °C with a DNA oligonucleotide 5′-(6-FAM)-AAA-TTC-TAA-TAG-ATA-ATG-TGA-(TAMRA). Fluorescence was measured on SynergyMx plate reader (BioTek). Local dinucleotide preferences in extracts were analysed similarly using 5′-AC, CC, GC or TC at the NN position of 5′-(6-FAM)-ATA-ANN-AAA-TAG-ATA-AT-(TAMRA).

Genomic uracil quantifications. Genomic DNA was prepared from shA3B- or shControl-transduced cells cultured for 21 days. Samples were spiked with heavy (+6)-labelled uracil (^{13}C and ^{15}N; Cambridge Isotopes) and treated with uracil-DNA glycosylase (NEB). Uracil was purified using 3,000 molecular mass cut-off columns (Pall Scientific) and solid-phase extraction (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 mass spectrometer was operated in positive ion mode, with 3.0 kV typical spray voltage, 250 °C capillary temperature, 67 V tube lens offset, and nitrogen sheath gas (25 counts). Argon collision gas was used at 146.7 mPa. Tandem mass spectrometry 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 × 100 mm, 5 μm, Thermo Scientific) was maintained at 40 °C and a flow rate of 15 μl min^{-1}. Solvents were 0.1% formic acid and acetonitrile. A linear gradient of 0–8% 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 20 V using the transitions: *m*/*z* 113.08 $[M^+H^+] \rightarrow$70.08 $[M\text{-}CONH]^+$ and *m*/*z* 96.08 $[M\text{-}NH_2]^+$ for uracil, whereas the internal standard ($[^{15}N\text{-}2, ^{13}C\text{-}4]$-uracil) was monitored by the transitions *m*/*z* 119.08 $[M^+H^+] \rightarrow m/z$ 74.08 $[M\text{-}CONH]^+$ and *m*/*z* 101.08 $[M\text{-}NH_2]^+$, respectively. Internal standards were used for quantification.

TK fluctuations. TK-neo was introduced into MDA-MB-453 and HCC1569 cells as described[17]. TK^{plus} cells were transduced with shA3B or shControl lentiviruses and subcloned by limiting dilution. One-million 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[29].

3D-PCR and sequencing. DNA was collected from Ugi-expressing[30] T-REx-293 clones or HCC1569 cells transduced with shA3B or shControl lentiviruses. 3D-PCR was done using Taq (Denville Scientific) as described[17]. Primers sequences are available on request. PCR products were analysed by gel electrophoresis with ethidium bromide, PCR purified (Epoch), blunt-end cloned into pJET (Fermentas), sequenced with T7 primer (BMGC), and aligned and analysed 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^{-1} 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^{-1} propidium iodide and 40 μg ml^{-1} RNase A (Qiagen) in PBS for 30 min and the DNA content and GFP induction were measured by flow cytometry (BD Biosciences FACS Canto II) and analysed with FlowJo and GraphPad Prism.

Cell viability assays. Cells were plated into multiple 96-well plates (2,500 cells per well) and measured at the days indicated. The MTS 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 6-well plate were plated with 2×10^5 cells. Half of the wells were induced with 1 μg ml^{-1} Dox. A crystal violet (0.5%), methanol (49.5%), water (50%) solution was used to stain cells after 7 days.

DNA damage experiments. Flow cytometric analysis of γ-H2AX foci was adapted[31]. 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 h before flow cytometry (BD Biosciences FACS Canto II) and analysis (FloJo and GraphPad). For microscopy, HEK293 cells were induced with 1 μg ml^{-1} 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 h. The cells were stained with 0.1% Hoechst dye and imaged at ×20 or ×60 (Deltavision) and deconvolved (SoftWoRx, Applied Precision).

Comet assays. As described[32], 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 day) or A3B–eGFP (6 day). Ten-thousand cells were added to coated slides and the cells were lysed overnight in 10 mM Tris, 100 mM EDTA, 2.5 M NaCl and 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^{-1} for 30 min. Gels were neutralized with 0.4 M Tris-HCl, pH 7.5, and treated with RNase A (Qiagen). The microgels were allowed to dry and comets were visualized using propidium iodide.

Bioinformatic analyses. Primary tumour genomic, exomic or RNA sequencing data were obtained from public sources[8,9,21–23]. Liver tumour genomes had 654,879, melanoma exomes had 2,798, breast tumour genomes had 183,916, breast triple-negative study exomes had 6,964, and TCGA breast tumour exomes had 5,559 total single base substitution mutations. Local contexts were tabulated and presented as weblogo schematics. Complex mutational events and CpG motifs were excluded.

28. 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).
29. Lea, D. E. & Coulson, C. A. The distribution of the numbers of mutants in bacterial populations. *J. Genet.* **49,** 264–285 (1949).
30. Di Noia, J. & Neuberger, M. S. Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. *Nature* **419,** 43–48 (2002).
31. Huang, X. & Darzynkiewicz, Z. Cytometric assessment of histone H2AX phosphorylation: a reporter of DNA damage. *Methods Mol. Biol.* **314,** 73–80 (2006).
32. Fairbairn, D. W., Olive, P. L. & O'Neill, K. L. The comet assay: a comprehensive review. *Mutat. Res.* **339,** 37–59 (1995).

doi:10.1038/nature11881

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 S2-S9**. For instance, the commonly used Affymetrix Genechip Human Genome Array U133A has 11 probes intended for *A3B* detection (**Table S2 & S4**). 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 S2-S9**).

Nevertheless, with knowledge of these limitations, useful information can still be derived from published microarray data sets. In particular, 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* qRT-PCR measurements and reported microarray values for *A3B* in the ATCC breast cancer cell line panel (Spearman Rank Test, p=0.0001; **Fig. S3a**; 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; **Fig. 4a**). 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 S2**). As expected by the non-specificity of several probe sets, significant differences were also seen for the "*A3A*" and "*A3F,G*" probe sets, which are both predicted to cross-hybridize with *A3B* mRNA (**Table S2**). In comparison, probe sets with low identity to *A3B* showed no significant correlation (*e.g.*, *A3C*; **Table S2**). As shown in **Fig. S3b**, 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].

A secondary explanation for why *A3B* has proven elusive up to now 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]. Interestingly, however, *A3B* up-regulation is clear and highly significant in RNAseq data sets recently made available to the broader research community by TCGA (**Fig. 4b**). The quantification of RNAseq data is not as robust and specific as qRT-PCR but it is superior to microarrays, most likely because sequence reads are paired (at least for Illumina platforms) and each read is longer than most microarray probes.

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:

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

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. S3b**). Having performed the same normalization method to all *APOBEC3* genes, we were able to obtain expression data for the tumor versus normal samples (**Table S2**). 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 $\alpha = 0.0007143$ to account for seventy 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 S2**, 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. 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 S3. 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 S4. 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 S5. 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 S6. 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 S7. 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 S8. 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 S9. 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 S10. Breast cancer patient information.

Table S2

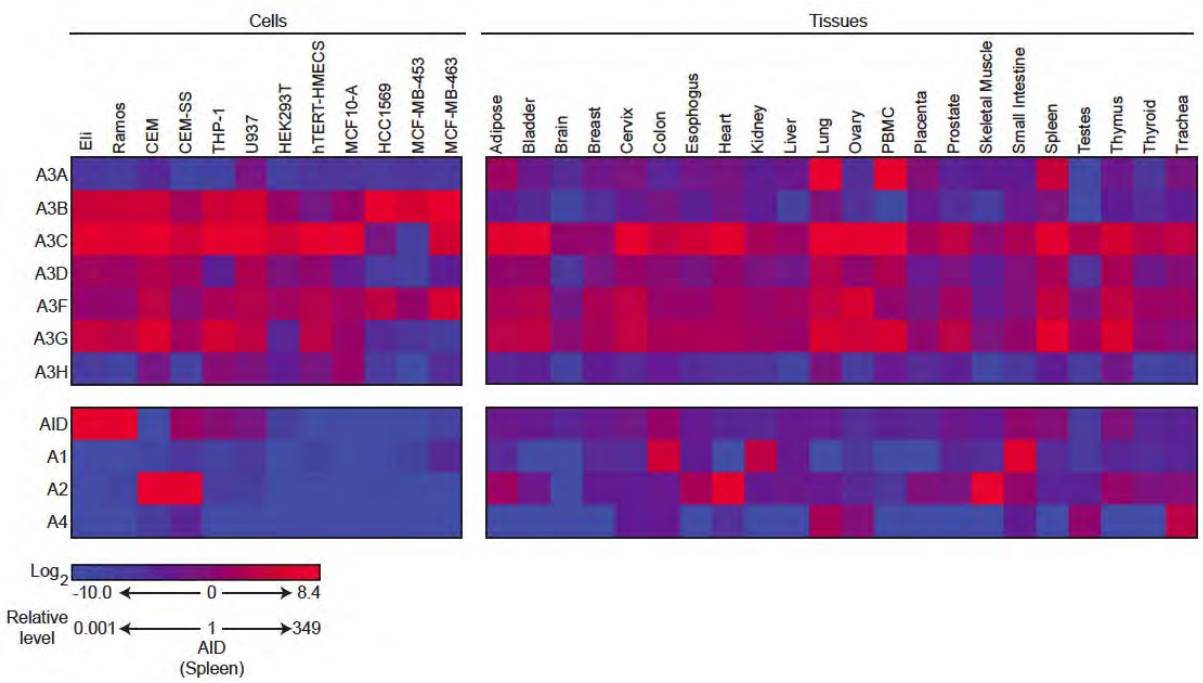
Patient ID	Age	Ethnicity	Age	ER	PR	Her2/neu	Type	Grade
P-7142	40	Caucasian	40	+	-	+	IDC	3
P-2248	51	African American	51	+	-	-	IDC	2
P-2100	75	Caucasian	75	+	+	-	IDC	2
P-2250	76	Caucasian	76	+	+	-	IDC	2
P-0480	51	Caucasian	51	-	-	-	IDC	3
P-2296	49	Caucasian	49	+	+	-	IDC	2
P-9407	38	Caucasian	38	+	+	-	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.

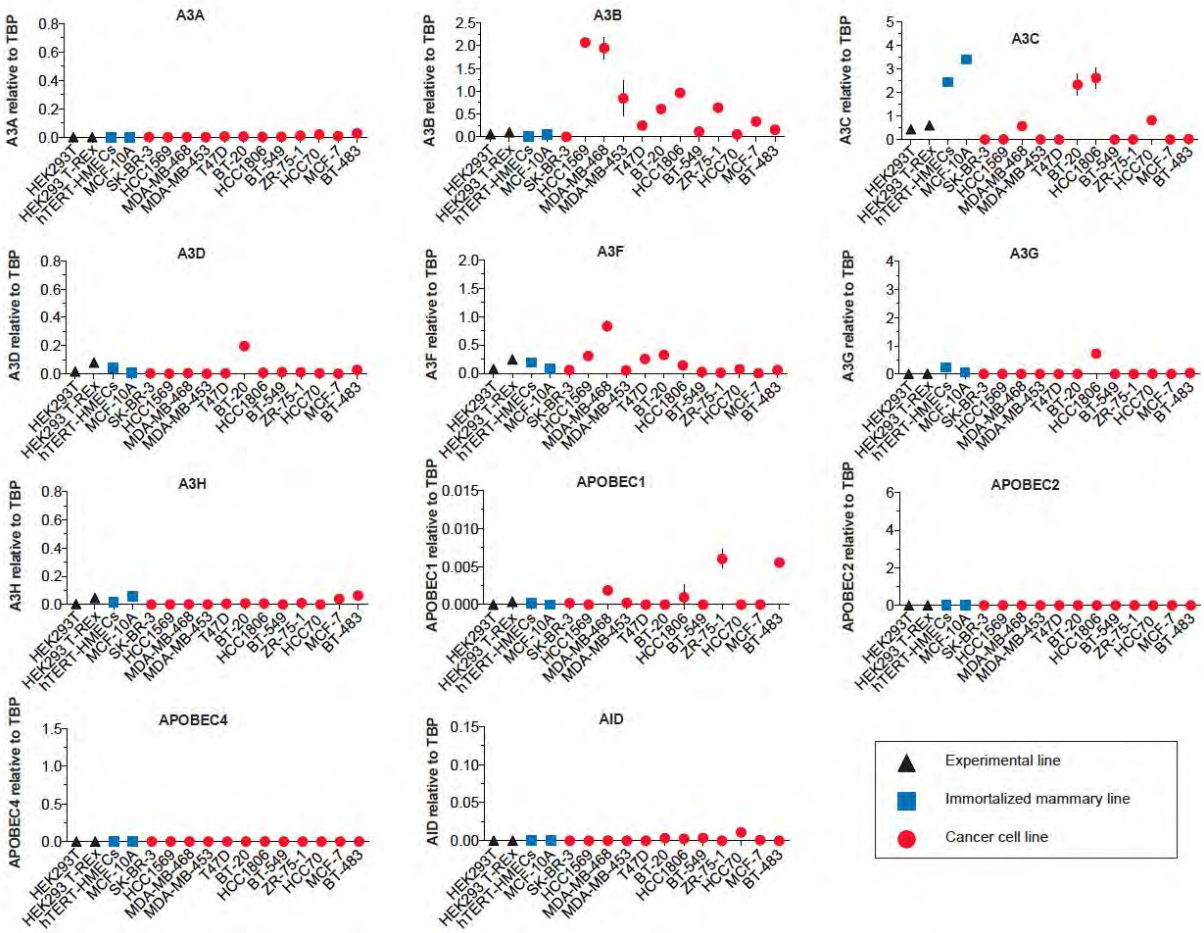
Supplementary Table S11. Quantitative PCR primer and probe information.

Gene Symbol	mRNA NCBI Accession	5' Primer Name	Seq (5'-3')	3' Primer Name	Seq (5'-3')	Probe Name	Seq[a]
APOBEC3s							
APOBEC3A	NM_145699	RSH2742	gagaagggacaagcacatgg	RSH2743	tggatccatcaagtgtctgg	UPL26	ctgggctg
APOBEC3B	NM_004900	RSH3220	gacccttggtccttcgac	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).

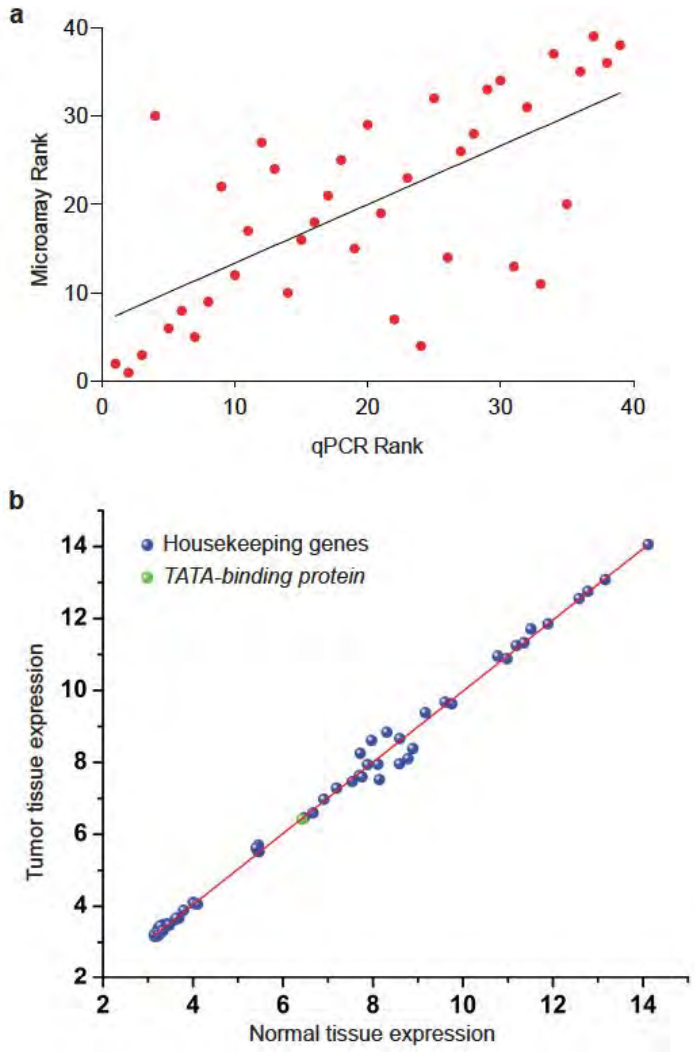


Supplementary Figure S1. Expression profiles for *APOBEC* family members in human cell lines and tissues. A heat-map summary of qRT-PCR data showing relative *APOBEC3* (*A3*), *AID*, *APOBEC1* (*A1*), *APOBEC2* (*A2*), and *APOBEC4* (*A4*) mRNA expression levels in the indicated cell lines and tissues. 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 *APOBEC3* expression in normal tissues, excluding PBMCs and breast tissue, were reported previously [Refsland *et al*. (Ref. 14)]. 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.

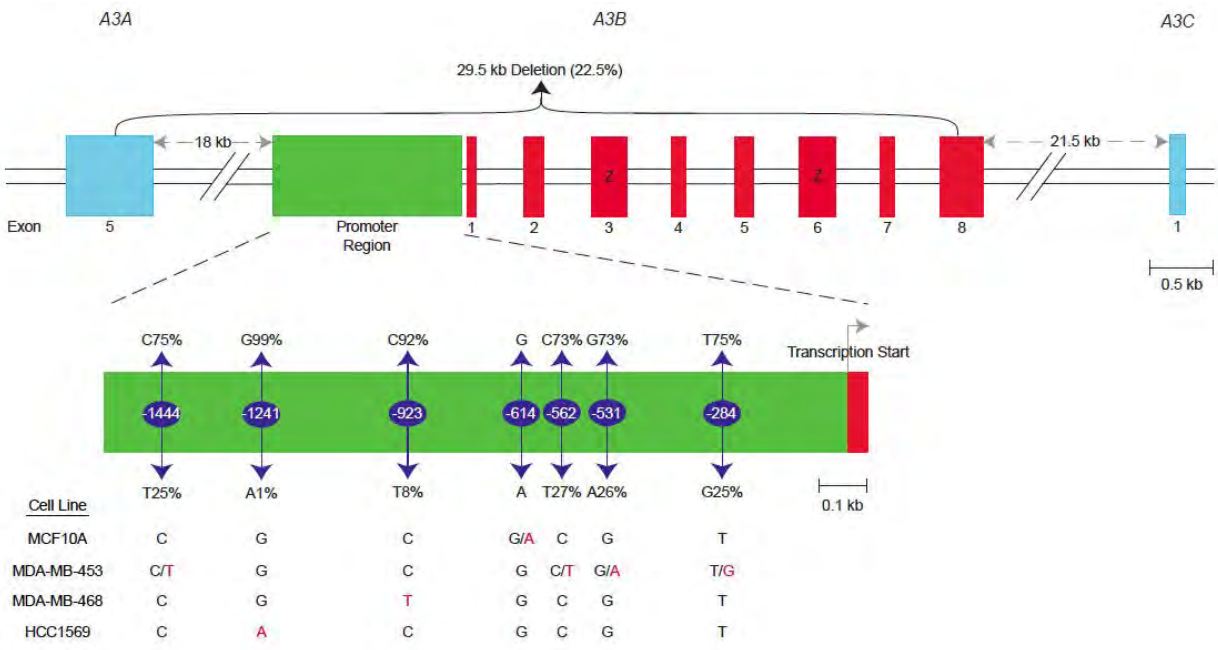


Supplementary Figure S2. Full expression profiles for *APOBEC* family members in a panel of representative cell lines.

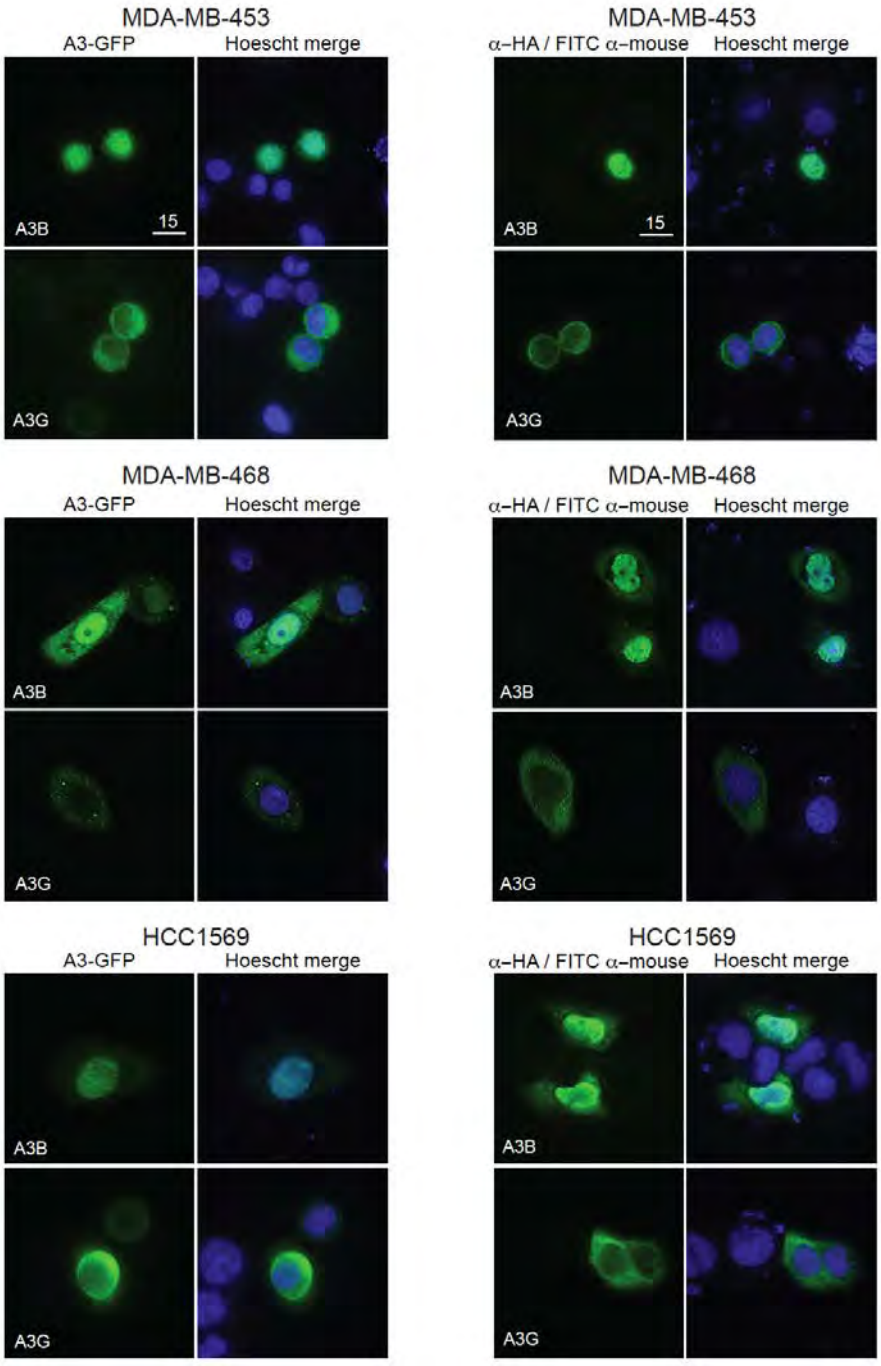
The indicated cell lines were used to generate cDNA for qPCR analyses of the full human *APOBEC* repertoire. Each data point is mean mRNA level of three qPCR reactions presented relative to mRNA levels of the constitutive housekeeping gene *TBP* (s.d. shown as a bar unless smaller than the data point).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



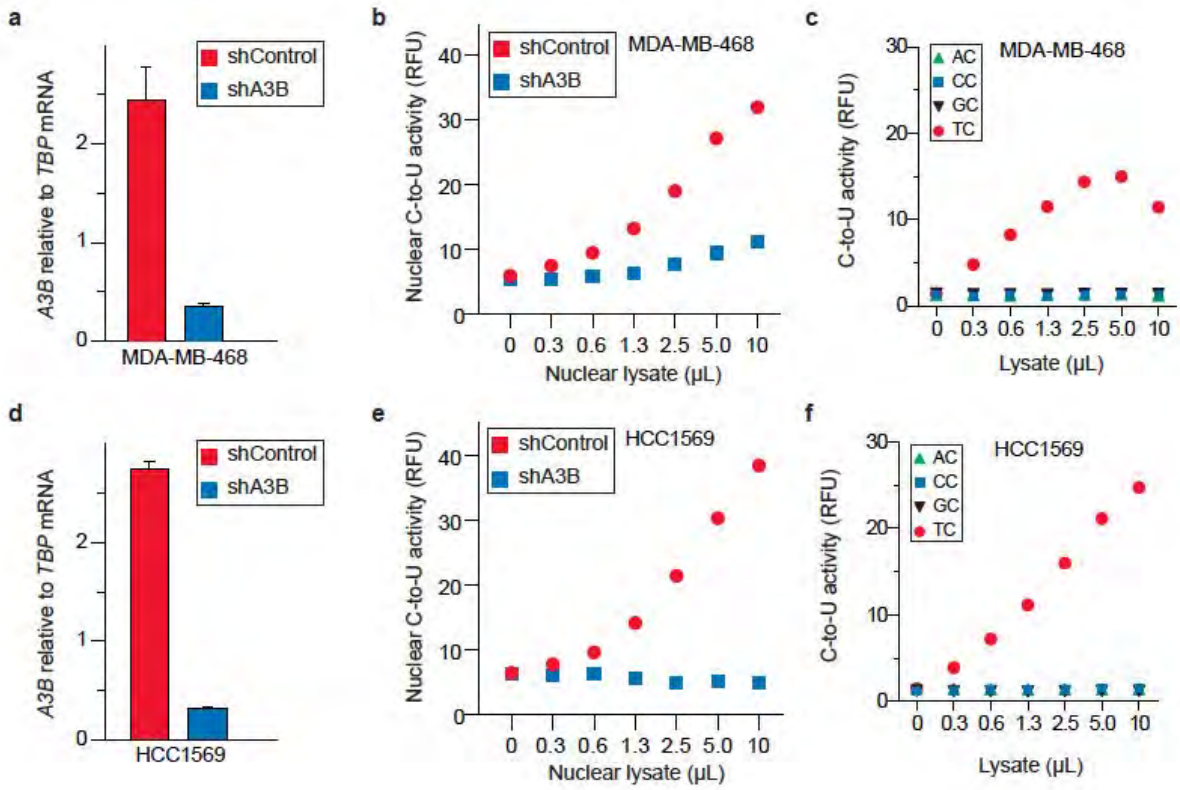
Supplementary Figure S3. Microarray information.
a, Positive correlation between *A3B* qRT-PCR data and microarray data (R^2=0.439). A total of 39 ATCC cell lines are common to both data sets. See Supplementary Discussion for additional details.
b, Housekeeping genes including *TBP* have near-identical expression levels in breast tumor and unrelated normal tissues (R^2=0.995).



Supplementary Figure S4. *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 S5. 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.

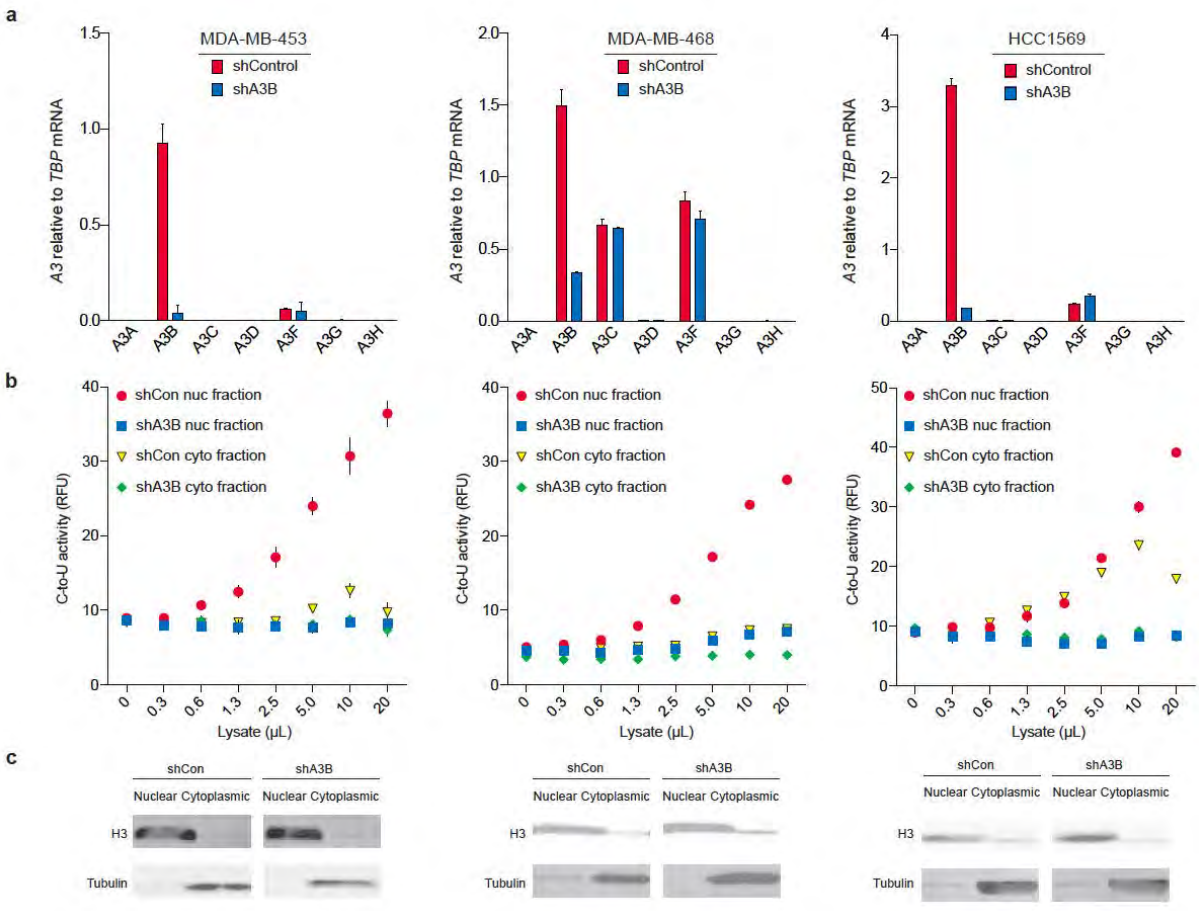


Supplementary Figure S6. DNA cytosine deaminase activity of endogenous A3B in breast cancer cell line nuclear extracts.

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

b & e, Nuclear DNA C-to-U activity in extracts from the indicated breast cancer cell lines transduced as in (a) (n=3; s.d. are smaller than data points).

c & f, Intrinsic dinucleotide DNA deamination preference of endogenous A3B in soluble nuclear extracts from the indicated cell lines (n=3; s.d. are smaller than each data points).

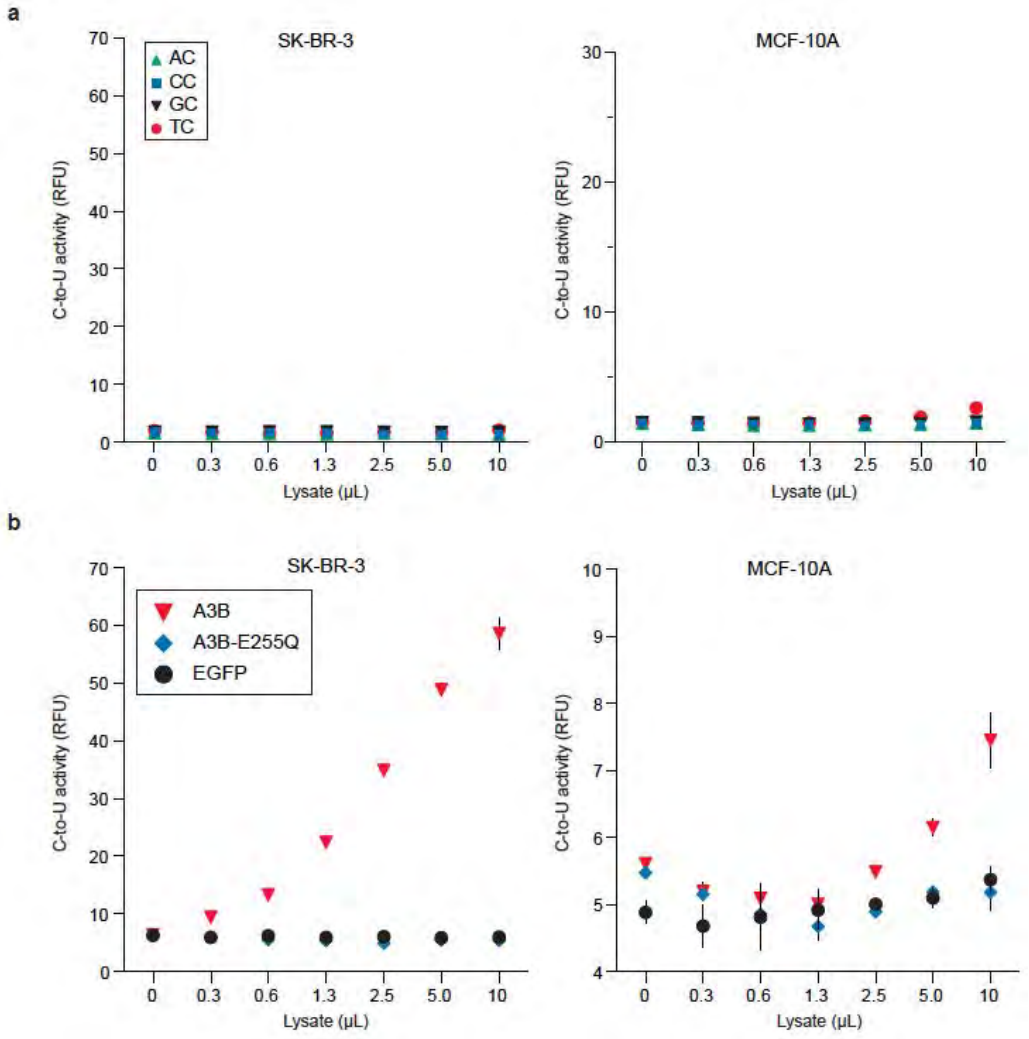


Supplementary Figure S7. A3B is active in the nuclear protein fraction of multiple breast cancer cell lines.

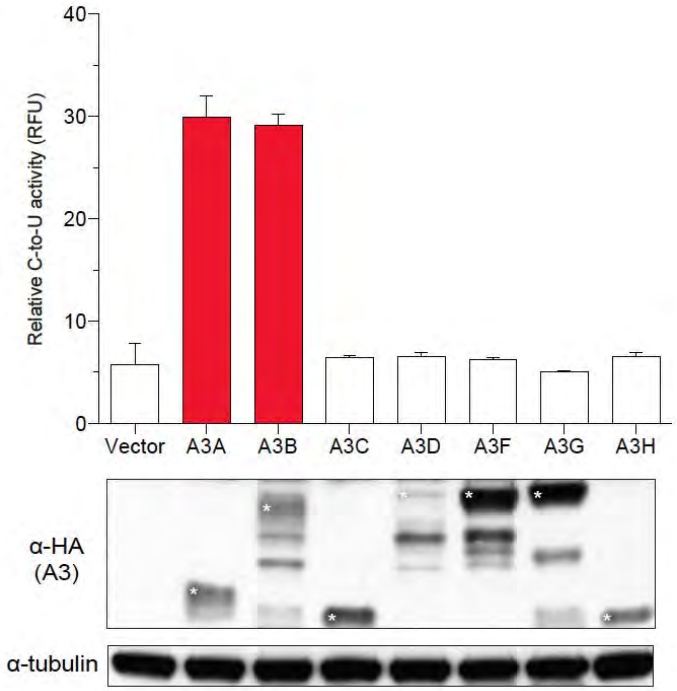
a, *A3* mRNA levels in the indicated breast cancer cell lines. Each column is mean +/- s.d. of three qPCR reactions presented relative to mRNA levels of the constitutive housekeeping gene *TBP*. Red and blue bars represent expression data from cells stably transduced with shControl or shA3B lentivirus,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

b, A3B-dependent DNA deaminase activity in the nuclear (Nuc) and cytoplasmic (Cyto) 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.

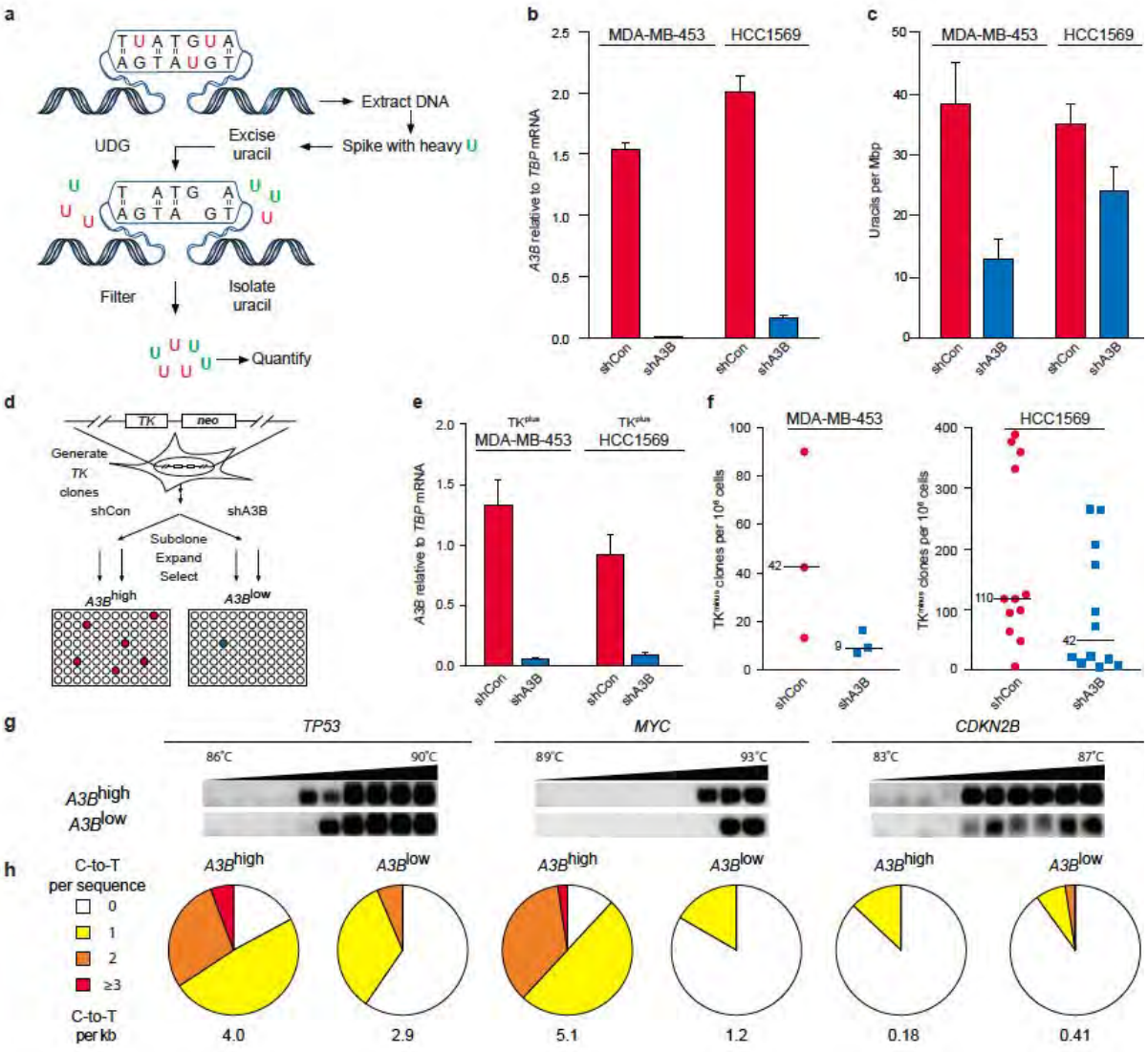
c, Immunoblots showing the distribution of histone H3, a nuclear protein, and tubulin, a cytoplasmic protein, in the protein preparations used in (b) to confirm efficient sub-cellular fractionation.



Supplementary Figure S8. DNA deaminase activity in A3B-low cell types.
a, Virtually no DNA C-to-U activity is observed in extracts from SK-BR-3 and MCF-10A cell lines using single-stranded DNA substrates with indicated dinucleotide targets.
b, DNA C-to-U activity in extracts from SK-BR-3 and MCF-10A transfected transiently with A3B-eGFP, A3B-E255Q-eGFP, or eGFP expression constructs. The higher activity levels in SK-BR-3 lysates are due to higher transfection efficiencies (30-40%), in comparison to MCF-10A (1-5%). Mean values are shown with s.d. indicated unless smaller than the symbol (n=3).



Supplementary Figure S9. 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-HA expression constructs (n=3 per condition; s.d. shown). Activity was only detected in lysates from cells transfected with A3A- or A3B-HA. The corresponding anti-HA immunoblot shows levels of each A3 (white asterisks), and the anti-tubulin blot indicates similar protein levels in each lysate.



Supplementary Figure S10. A3B-dependent uracil lesions and mutations in breast cancer genomic DNA (data from Fig. 2 reproduced here for comparison).

a, Workflow for genomic uracil quantification by HPLC-ESI+MS/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.

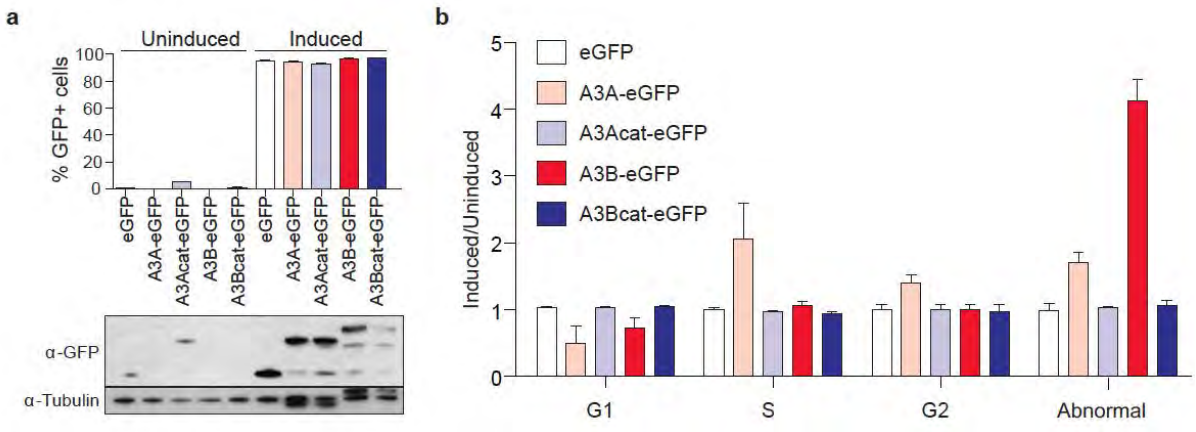
d, Workflow for TK fluctuation analysis.

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.

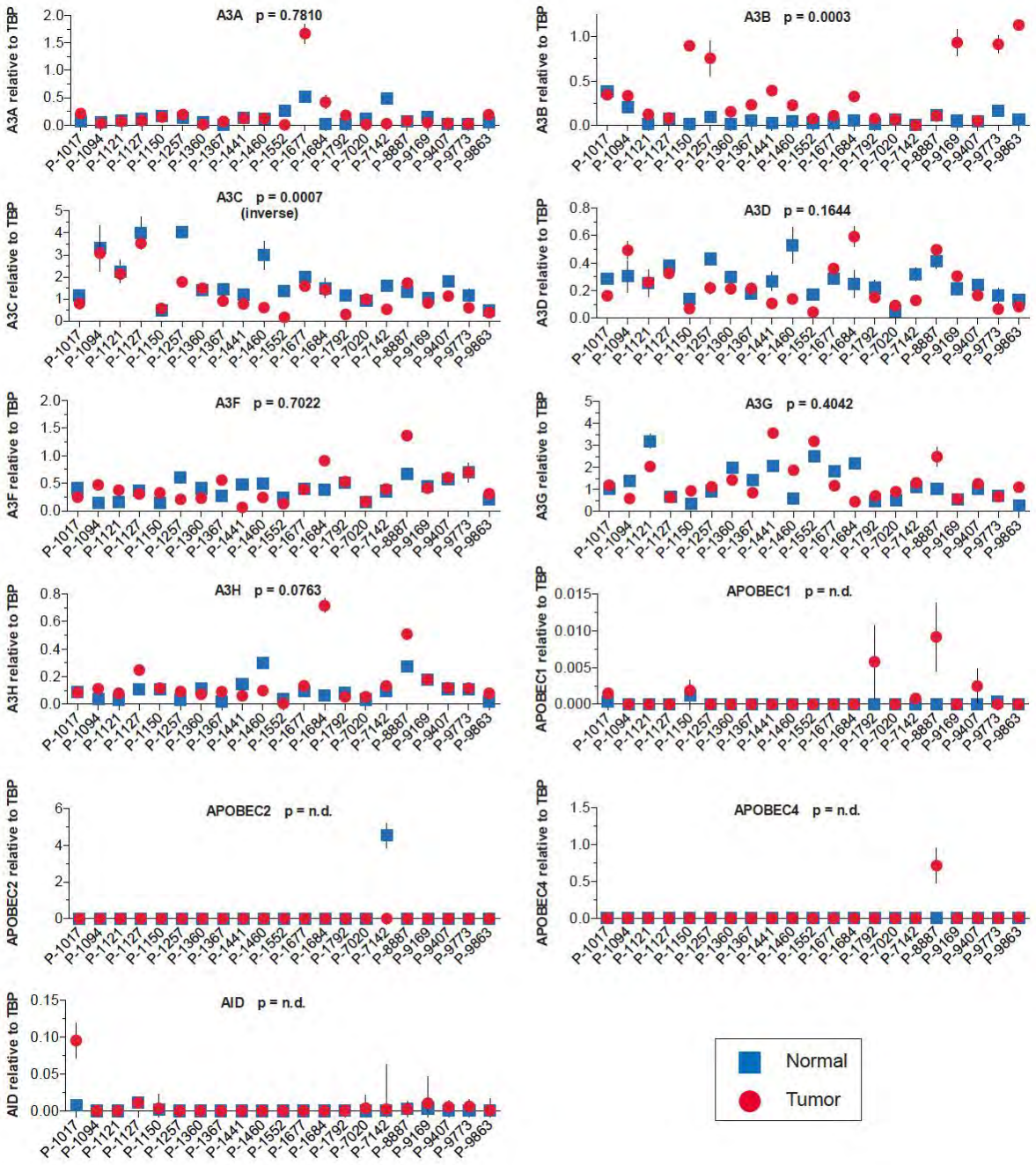
h, Pie charts depicting the C/G-to-T/A mutation load in 3D-PCR products after cloning and sequencing ($n \geq 35$ per condition). Charts align with target genes labeled in (g).



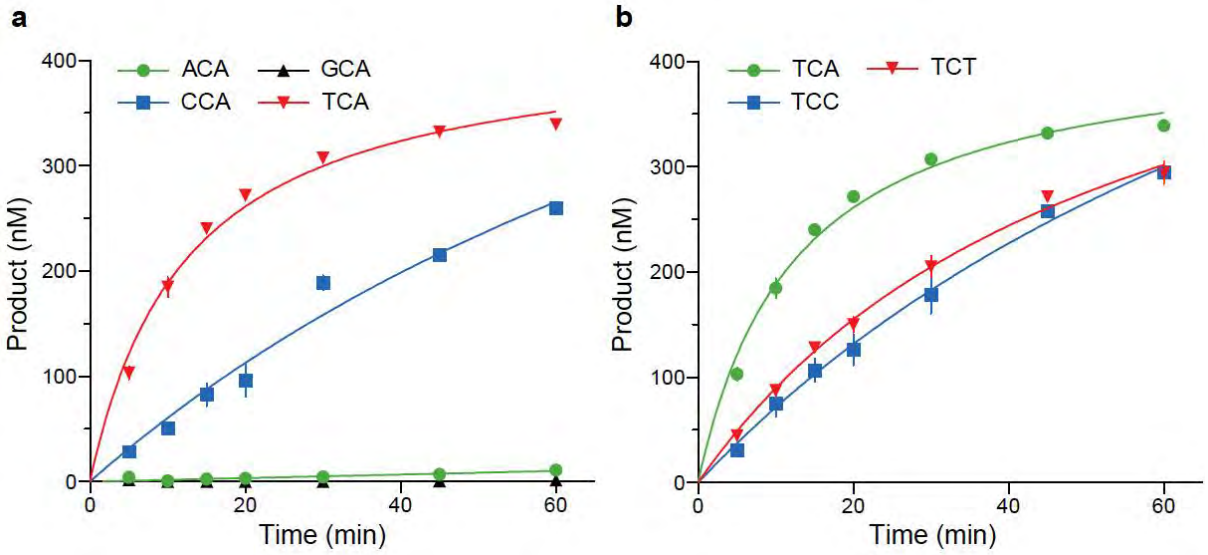
Supplementary Figure S11. Experimental system and cell cycle data for A3A and A3B induction.

a, The percent fluorescence for the indicated HEK293-derived A3-eGFP cell lines in absence or presence of Dox (corresponding anti-GFP immunoblot below along with an anti-tubulin blot to control for protein loading).

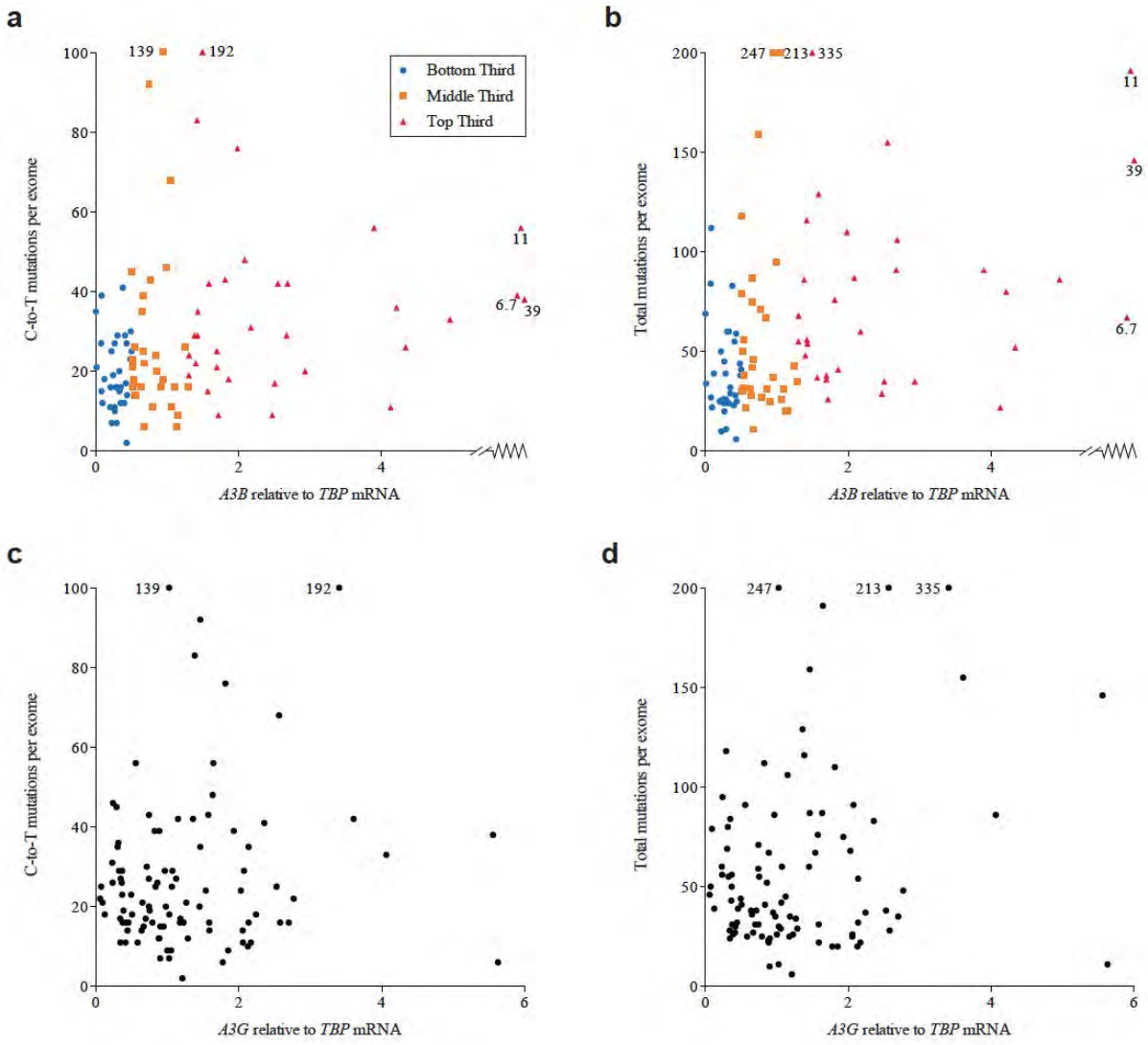
b, Cell cycle status 2 days post-induction (relative to indicated lines uninduced).



Supplementary Figure S12. 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 (n.d., not determined). *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. 4a.



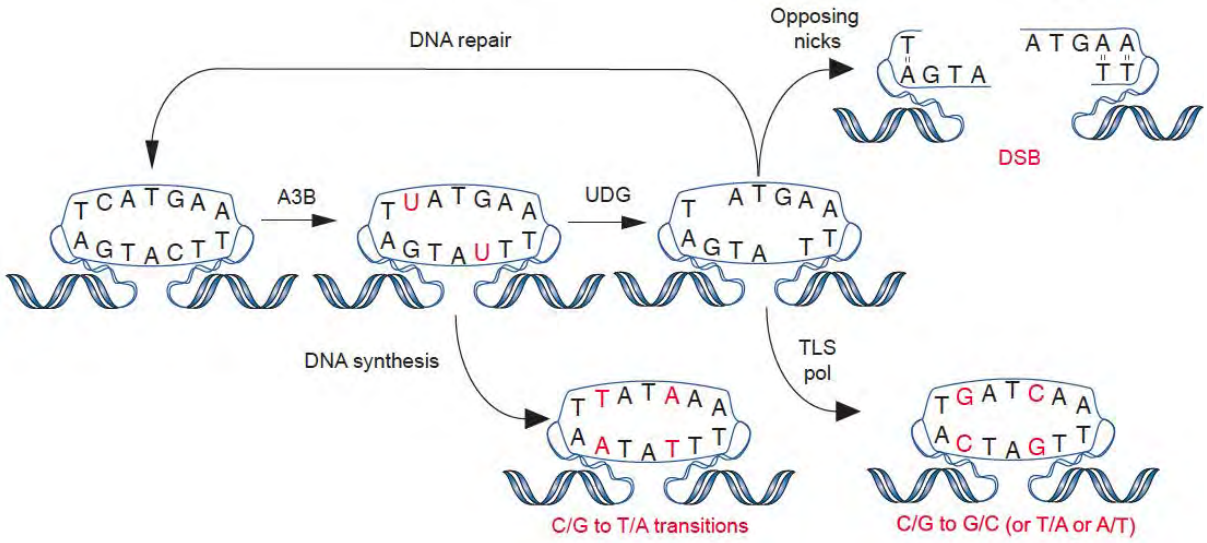
Supplementary Figure S13. A3B catalytic domain local deamination preferences.
a, A3B catalytic domain deamination kinetics using single-stranded DNA substrates that vary as shown at the 5' position relative to the target cytosine.
b, A3B catalytic domain deamination kinetics using single-stranded DNA substrates that vary as shown at the 3' position relative to the target cytosine. The 5'-TCA data in this panel are the same as those in (a) to facilitate direct comparisons. The potentially methylatable CpG dinucleotide substrate was not included in this analysis to avoid possible confusion with a hydrolytic, spontaneous deamination mechanism, as methyl-cytosines are more labile than normal cytosines.



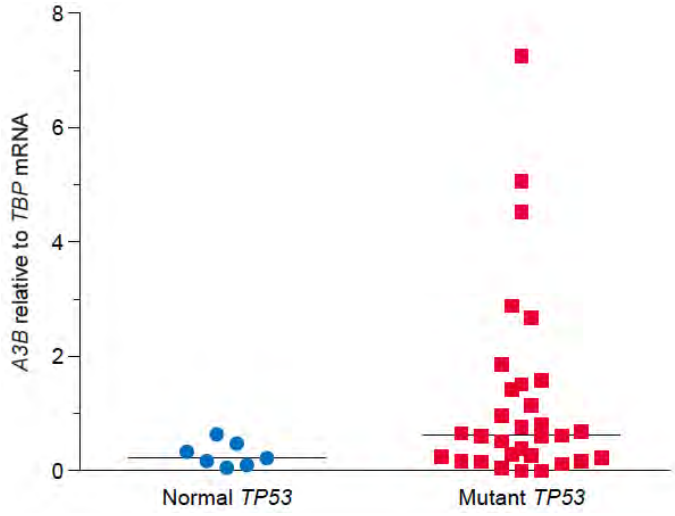
Supplementary Figure S14. Breast cancer mutation load and gene expression correlations.

a & b, Two-dimensional plots of C-to-T mutation loads and total mutation loads for each breast tumor vs. *A3B* expression level by RNA sequencing (Spearman r=0.34 and p=0.0006 for C-to-T and r=0.38 and p=0.0001 for total mutations). These are alternative presentations of the data shown in Fig. 4e & f.

c & d, Two-dimensional plots of C-to-T mutation loads and total mutation loads for each breast tumor vs. *A3G* expression level by RNA sequencing (Spearman r=0.018 and p=0.86 for C-to-T r=0.028 and p=0.78 for total mutations). *A3G* expression data are the same as those presented in Fig. 4b.



Supplementary Figure S15. DNA deamination model for A3B in cancer.
Deamination of genomic DNA cytosines by up-regulated A3B leads to uracil lesions, which may be repaired faithfully or lead to at least three possible outcomes: i) C-to-T transitions by direct DNA synthesis, ii) DNA double-stranded breaks by uracil excision and opposing abasic site cleavage (or, not shown, replication fork collapse at a single-stranded nick), and iii) transversions or transition mutations by error-prone DNA synthesis or aberrant repair (TLS pol = translesion synthesis DNA polymerase). The mechanism of AID-dependent antibody gene diversification provides several precedents for this model including the fact that DNA C-to-U deamination events at expressed antibody loci are essential precursors to a diverse array of final outcomes such as all types of base substitutions and isotype changes (Ref. 24).



Supplementary Figure S16. *A3B* up-regulation and *TP53* inactivation in the ATCC breast cancer cell line panel.
A dot 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).

nature genetics

Evidence for APOBEC3B mutagenesis in multiple human cancers

Michael B Burns[1–4], Nuri A Temiz[1,2] & Reuben S Harris[1–4]

Thousands of somatic mutations accrue in most human cancers, and their causes are largely unknown. We recently showed that the DNA cytidine deaminase APOBEC3B accounts for up to half of the mutational load in breast carcinomas expressing this enzyme. Here we address whether APOBEC3B is broadly responsible for mutagenesis in multiple tumor types. We analyzed gene expression data and mutation patterns, distributions and loads for 19 different cancer types, with over 4,800 exomes and 1,000,000 somatic mutations. Notably, *APOBEC3B* is upregulated, and its preferred target sequence is frequently mutated and clustered in at least six distinct cancers: bladder, cervix, lung (adenocarcinoma and squamous cell carcinoma), head and neck, and breast. Interpreting these findings in the light of previous genetic, cellular and biochemical studies, the most parsimonious conclusion from these global analyses is that APOBEC3B-catalyzed genomic uracil lesions are responsible for a large proportion of both dispersed and clustered mutations in multiple distinct cancers.

Somatic mutations are essential for normal cells to develop into cancers. Partial and full tumor genome sequences have shown the existence of hundreds to thousands of mutations in most cancers[1–10]. The observed mutation spectrum is the result of DNA lesions that either escaped repair or were misrepaired. This spectrum can be used to help determine the cause or source of the initial damage. For instance, the cytosine-to-thymine transition bias in skin cancers can be explained by a mechanism in which ultraviolet (UV)-induced lesions—cyclobutane pyrimidine dimers (C*C, C*T, T*C or T*T, where asterisks denote UV-induced lesions between adjacent pyrimidines)—are bypassed by DNA polymerase–catalyzed insertion of two adenine bases opposite each unrepaired lesion[11]. A second round of DNA replication or excision and repair of the pyrimidine dimer results in cytosine-to-thymine transitions. Notably, the nature of this type of DNA damage dictates that each resulting cytosine-to-thymine transition occurs in a dipyrimidine context, with each mutated cytosine invariably flanked on the 5′ or the 3′ side by a cytosine or thymine. A similar rationale combining observed mutation spectra and knowledge of biochemical mechanisms may be used to delineate other sources of DNA damage and mutation in human cancers.

Nonrandom mutation patterns, such as CG base pairs being more frequently mutated than AT base pairs[1–10] and the occurrence of strand-coordinated clusters of cytosine mutations[9,12,13] (strand-coordinated mutations are those that occur together on a single strand of the DNA double helix), are also observed in other types of cancer. Spontaneous hydrolytic deamination of cytosine to uracil may explain a subset of these events but not the majority, as most occur outside of CpG dinucleotide motifs that can be methylated (the sites most prone to spontaneous deamination), and the occurrence of these mutations in clusters is highly nonrandom. Another possible source of these mutations is enzyme-catalyzed cytosine-to-uracil deamination by one or more of the nine active DNA cytidine deaminases encoded by the human genome. Such a mechanism was originally hypothesized when the DNA deaminase activity of these enzymes was discovered[14] and was recently highlighted with demonstrations of clustered mutations in breast, head and neck, and other cancers[9,12,13]. These mutational clusters have been named kataegis, as their sporadic but concentrated nature bears a likeness to rain showers[9]. Although enzymatic deamination has been implicated in this phenomenon, the actual enzyme responsible has not been determined.

Enzyme-catalyzed DNA cytosine-to-uracil deamination is central to both adaptive and innate immune responses. B lymphocytes use activation-induced deaminase (AID) to create antibody diversity by inflicting uracil lesions in the variable regions of expressed immunoglobulin genes, which are ultimately processed into all six types of base-substitution mutations[15,16] (somatic hypermutation). AID also creates uracil lesions in antibody gene switch regions that lead to DNA breaks and the juxtaposition of the expressed and often mutated variable region next to a new constant region (isotype switch recombination)[15,16]. In humans, seven related enzymes—APOBEC3A, APOBEC3B, APOBEC3C, APOBEC3D, APOBEC3F, APOBEC3G and APOBEC3H—combine to provide innate immunity to a variety of DNA-based parasitic elements[17,18]. A well-studied example is the cDNA replication intermediate of HIV-1, which during reverse transcription is vulnerable to enzymatic deamination by at least three different APOBEC3 proteins[19,20]. APOBEC1 also has a similar capacity for viral cDNA deamination, and it is the only family member known to have a biological role in cellular mRNA editing[21–24]. The more distantly related proteins APOBEC2 and APOBEC4 have yet to elicit enzymatic activity. In total, 9 of the 11 APOBEC family members have demonstrated DNA deaminase activity in a variety of biochemical and biological assay systems[14,25–29].

[1]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, Minneapolis, Minnesota, USA. [2]Masonic Cancer Center, University of Minnesota, Minneapolis, Minnesota, USA. [3]Institute for Molecular Virology, University of Minnesota, Minneapolis, Minnesota, USA. [4]Center for Genome Engineering, University of Minnesota, Minneapolis, Minnesota, USA. Correspondence should be addressed to R.S.H. (rsh@umn.edu).

Received 17 March; accepted 20 June; published online 14 July 2013; doi:10.1038/ng.2701

Table 1 Summary statistics for the 19 different tumor types in this study

		APOBEC3B expression data[a]			Exome mutation data[b]				Clustered mutation data[c]		
Tumor type	TCGA ID	*n*	Range	Median	*n*	Range	Median	Average	Total number of clusters	Mean per tumor	Percentage of total mutations
Low-grade glioma	LGG	174	0–0.69	0.062	170	5–15,458	45	138	280	1.6	5.1
Prostate adenocarcinoma	PRAD	140	0–0.76	0.12	150	19–165	54	59	27	0.18	1.1
Thyroid carcinoma	THCA	384	0–4.1	0.18	326	3–98	20	22	25	0.077	1.2
Glioblastoma multiforme	GBM	169	0.014–2.0	0.22	167	1–173	28	34	114	0.68	7.9
Kidney renal papillary cell carcinoma	KIRP	76	0.0079–3.0	0.24	100	15–214	64	69	18	0.18	1.0
Kidney renal clear-cell carcinoma	KIRC	480	0.011–4.5	0.29	244	6–696	73	92	42	0.17	0.67
Acute myeloid leukemia	LAML	179	0.027–2.3	0.44	74	1–151	12	17	1	0.010	0.21
Ovarian serous cystadeno-carcinoma	OV	266	0.0015–8.6	0.48	469	1–145	39	55	1	0.0021	0.010
Breast invasive carcinoma	BRCA	849	0.0012–39	0.67	777	2–443	45	59	122	0.16	0.86
Stomach adenocarcinoma	STAD	57	0.18–3.6	0.68	156	6–8,849	172	551	66	0.42	0.32
Lung adenocarcinoma	LUAD	355	0.0041–9.6	0.68	392	12–2,547	259	355	310	0.79	0.73
Rectum adenocarcinoma	READ	72	0.082–3.2	0.81	88	28–7,204	136	227	44	0.50	1.2
Colon adenocarcinoma	COAD	192	0.017–3.7	0.85	266	27–8,459	250	487	133	0.50	0.39
Uterine corpus endometrioid carcinoma	UCEC	370	0.012–12	0.94	248	1–14,687	68	722	1093	4.4	2.9
Skin cutaneous melanoma	SKCM	267	0.0011–10	1.1	255	6–6,174	389	697	353	1.4	0.68
Bladder urotheilal carcinoma	BLCA	122	0.0050–24	1.6	99	45–1,802	226	291	293	3.0	3.5
Head and neck squamous cell carcinoma	HNSC	303	0.0038–20	1.7	306	7–2,070	138	180	203	0.66	1.4
Lung squamous cell carcinoma	LUSC	259	0.094–15	1.7	177	1–3,910	299	363	144	0.81	0.77
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	97	0.0010–20	2.4	39	30–1,779	138	233	98	2.5	3.4

[a]*APOBEC3B* expression levels relative to those of the housekeeping gene *TBP* determined by RNA-seq. [b]Somatic mutations in each exome, spanning approximately 38 Mb of the human genome. [c]Kataegis events from exome mutation data are defined as ≥2 cytosine mutations within 10-kb intervals that meet Gordenin significance (Online Methods).

However, a possible drawback of encoding nine active DNA deaminases could be chromosomal DNA damage and, ultimately, mutations that lead to cancer[14]. AID has been linked to B cell tumorigenesis through off-target chromosomal deamination as well as the triggering of translocations between the expressed heavy chain locus and various oncogenes[30]. Transgenic expression of AID causes tumor formation in mice[31], as does transgenic expression of APOBEC1 (ref. 32). Most recently, we showed that APOBEC3B is upregulated in breast tumors and is correlated with a doubling of both cytosine-to-thymine and overall base-substitution mutation loads[33]. Because AID and APOBEC1 are expressed in a tissue-specific manner and there is no reason to suspect developmental confinement of APOBEC3B, we hypothesized that APOBEC3B may be a general mutagenic factor affecting the genesis and evolution of many different cancers. This hypothesis is supported by studies indicating that *APOBEC3B* is expressed in many different cancer cell lines[33–35], in contrast to its relatively low expression in 21 normal human tissues spanning all major organs[33,35,36]. This DNA mutator hypothesis is additionally supported by the fact that APOBEC3B is the only deaminase family member with constitutive nuclear localization[33,37].

Here we test this mutator hypothesis by performing a global analysis of all available DNA deaminase family member expression data and exomic mutation data from 19 different cancers representing over 4,800 tumors and 1,000,000 somatic mutations. Mutation frequencies, local sequence contexts and distributions, including kataegis events, were analyzed systematically for each tumor and cancer type. In addition, we calculated the hierarchical distances between the deamination signature of recombinant APOBEC3B derived from biochemical experiments[33] and the observed frequencies of cytosine mutation spectra in all 19 cancer types. Taken together, these analyses converge upon APOBEC3B as the most likely cause of a large fraction of both dispersed and clustered cytosine mutations in six distinct cancers.

RESULTS

As a first test of the hypothesis that APOBEC3B is a general endogenous cancer mutagen, we performed a comprehensive analysis of the expression profiles of all 11 APOBEC family members across a panel of 19 distinct tumor types, including breast cancer as a positive control[33] (**Table 1** and **Supplementary Fig. 1**). The expression values for each target mRNA were normalized to those of the constitutive housekeeping gene *TBP* (encoding TATA-binding protein) to enable quantitative comparisons between RNA sequencing (RNA-seq) and quantitative RT-PCR (qRT-PCR) data sets and to provide controls in the few instances where RNA-seq values for normal tissues were not available publicly (Online Methods).

Several cancers showed *APOBEC3B* expression levels comparable to those in corresponding normal tissues (**Fig. 1**, **Table 1**, **Supplementary Fig. 1** and **Supplementary Table 1**). Prostate and renal clear-cell carcinomas showed statistically significant upregulation of *APOBEC3B* in the tumors, albeit with median expression values that were only a fraction of the *TBP* levels. In contrast, six different cancers showed evidence of strong *APOBEC3B* upregulation in the majority of tumors of the breast, uterus, bladder, head and neck, and lung (adenocarcinoma and squamous cell carcinoma) ($P < 0.0001$ by Mann-Whitney U test). Other cancers, such as cervical and skin, also showed high *APOBEC3B* levels, but a lack of data for corresponding normal tissues precluded statistical analysis. A total of ten cancers showed a median level of *APOBEC3B* upregulation greater than that of the intended positive control, breast cancer.

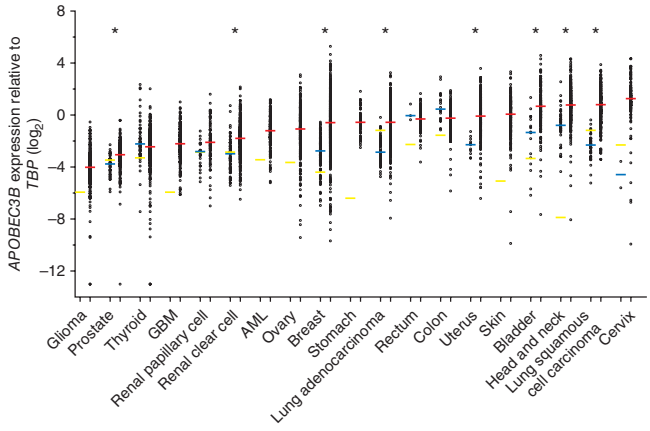


Figure 1 *APOBEC3B* is upregulated in numerous cancer types. Each data point represents one tumor or normal sample, and the *y* axis is log transformed for better data visualization. Red, blue and yellow horizontal bars indicate median *APOBEC3B* levels relative to *TBP* levels for each cancer type (**Table 1**), the median values for each set of RNA-seq data from normal tissues (**Supplementary Table 1**) and individual qRT-PCR data points, respectively. Asterisks indicate significant upregulation of *APOBEC3B* in the indicated tumor type relative to the corresponding normal tissues ($P < 0.0001$ by Mann-Whitney *U* test). *P* values for negative or insignificant associations are not shown.

This finding was particularly notable for bladder, head and neck, both lung carcinomas and cervical cancers.

The second major prediction of the APOBEC mutator hypothesis is the occurrence of chromosomal DNA cytosine-to-uracil deamination, which should result in strong biases toward mutations at CG base pairs. Such mutational events may be either transitions or transversions because genomic uracil bases can either directly template the insertion of adenine bases during DNA replication, or, if converted to abasic sites by uracil DNA glycosylase, the lesions become non-instructional, and error-prone polymerases may insert an adenine, thymine or cytosine opposite the abasic site (most often adenine, following the A rule). In both scenarios, an additional round of DNA synthesis or repair can yield either transitions or transversions at CG base pairs (including CG-to-TA, CG-to-GC and CG-to-AT base-pair mutations).

Notably, the fraction of mutations at CG base pairs ranges considerably, from a low of 60% in renal cancers to a high of approximately 90% in skin, bladder and cervical cancers (**Fig. 2a**). The massive bias in skin cancers is largely attributable to error-prone DNA synthesis (adenine insertion) opposite cyclobutane pyrimidine dimers caused by UV light[11]. However, the biases observed in urogenital carcinomas, such as bladder and cervical cancers, are probably not due to UV light but more likely to an alternative mutagenic source such as enzymatic DNA deamination. Indeed, the top five tumor types with CG base pair–dominated mutation spectra were among the top six tumors in terms of *APOBEC3B* expression (compare **Figs. 1** and **2a**). A possible mechanistic relationship is further supported by a positive correlation between the overall proportion of mutations occurring at CG base pairs and median *APOBEC3B* levels ($P = 0.0031$, $r = 0.64$ by Spearman's correlation; **Fig. 2b**). The positive correlation is notable given the fact that all available data were included in the analysis and multiple variables could have undermined a positive correlation, such as known mutational sources (UV light in skin cancer), undefined mutational sources (in glioma with the sixth highest CG base-pair mutation bias and lowest *APOBEC3B* levels) and differential DNA repair capabilities among the distinct tumor types.

DNA deaminases such as APOBEC3B are strongly influenced by the bases adjacent to the target cytosine, particularly at the immediate 5′ position. For instance, AID prefers 5′ adenine or guanine bases, APOBEC3G prefers 5′ cytosine bases and other family members prefer 5′ thymine bases[38–40]. We recently showed that recombinant APOBEC3B prefers 5′ thymine bases and strongly disfavors 5′ purines, whereas, on the 3′ side, it prefers adenine or guanine bases and disfavors pyrimidines[33] (**Fig. 3a**). Therefore, the third prediction of the APOBEC mutator hypothesis is that cancers affected by enzymatic deamination should show nonrandom nucleotide distributions immediately 5′ and 3′ of mutated cytosine bases and that these signatures can then be used with expression information, additional mutation data and existing literature and biochemical constraints to identify the enzyme responsible.

We therefore performed a global analysis of sequence signatures for all available cytosine mutation data from the top 50% of *APOBEC3B*-expressing tumors for each tumor type (this expression cutoff was chosen to minimize the impact of unrelated mutational mechanisms). These mutation data were first compiled and subjected to hierarchical cluster analysis to group tumors with similar cytosine mutation signatures (**Fig. 3a**). Short Euclidean distances (smaller measures) between the mutation signatures of different tumors indicated a high degree of concordance, thereby implying similar mutational patterns (**Supplementary Table 2** lists the calculated values). Bladder and cervical cancers, two of the top *APOBEC3B*-expressing cancers, had cytosine mutation signatures notably similar to each other and to that of recombinant APOBEC3B protein. This relationship is illustrated by strong mutation biases at TCA motifs (with the targeted base underlined), which match the enzyme's optimal *in vitro* substrate. The two lung carcinomas, breast, and head and neck cancers also had cytosine mutation signatures that strongly resembled the preference of recombinant APOBEC3B protein (**Fig. 3a** and **Supplementary Table 2**). Several cancers had cytosine mutation signatures with an intermediate relatedness

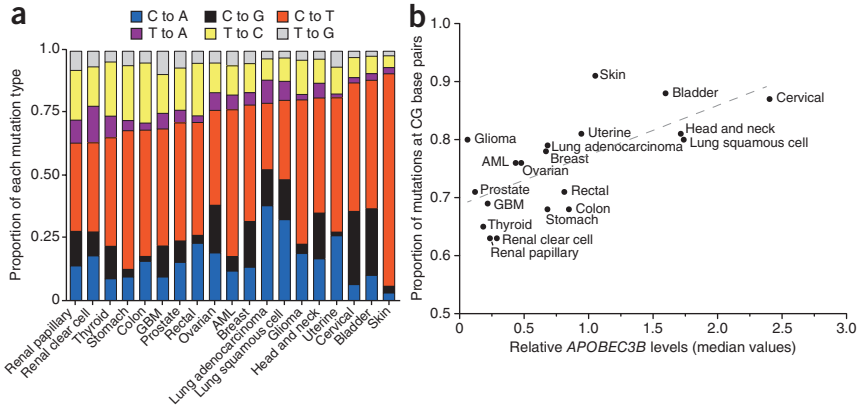


Figure 2 Mutation types and signatures in 19 human cancers. (**a**) Stacked bar graph summarizing the six types of base-substitution mutations as proportions of the total mutations per cancer type. (**b**) Median *APOBEC3B* levels relative to *TBP* levels plotted against the proportion of mutations at CG base pairs (Spearman's $P = 0.0031$, $r = 0.64$). The dashed gray line is the best fit for visualization.

to the motif for recombinant APOBEC3B (renal papillary, thyroid, ovarian, renal clear-cell, glioblastoma multiforme (GBM) and skin cancers). In further contrast, the seven remaining cancers, ranging from uterine to colon, had cytosine mutations with the largest separation from the motif for recombinant APOBEC3B (**Fig. 3a** and **Supplementary Table 2**).

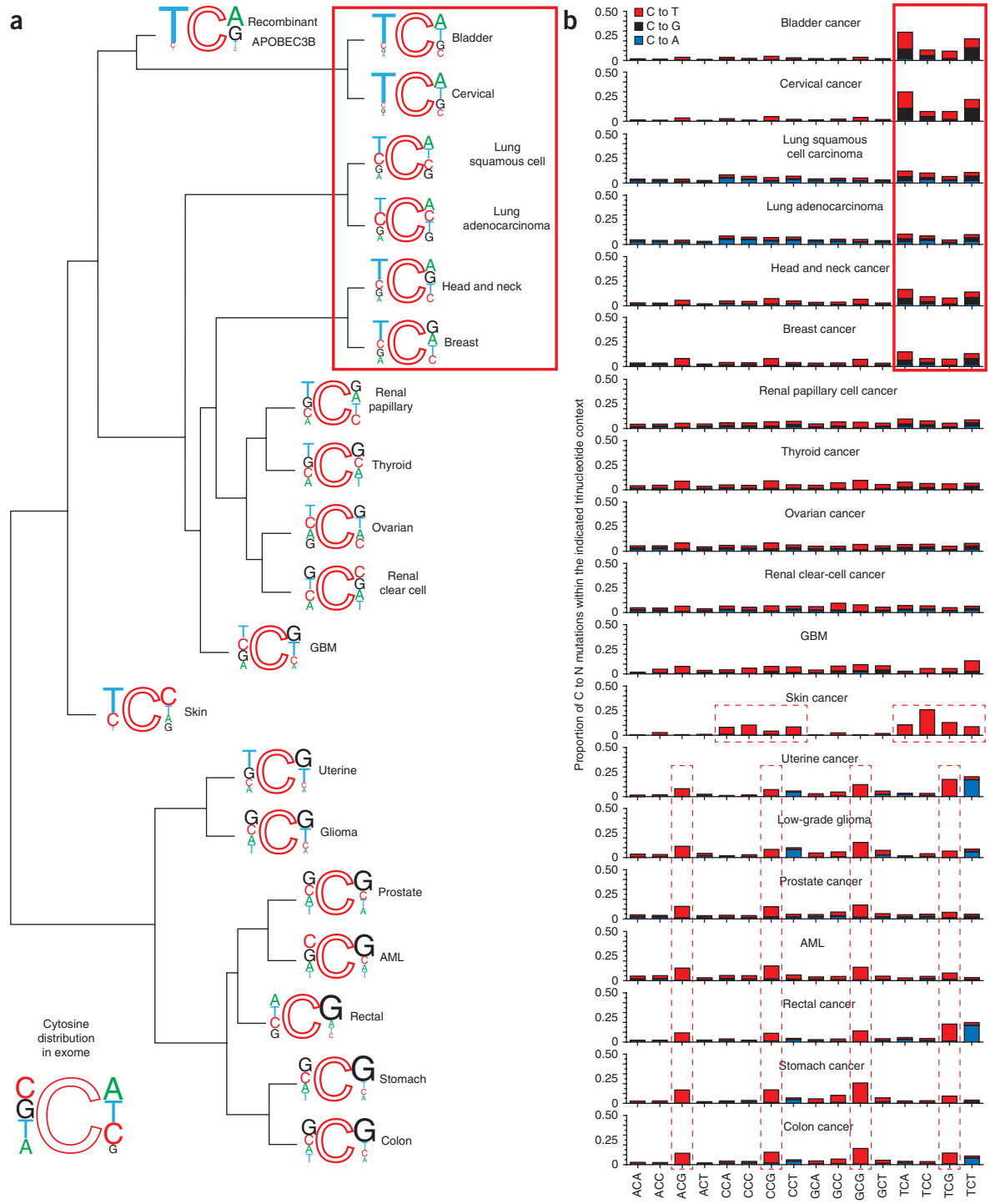


Figure 3 Cytosine mutation spectra for 19 cancers. (**a**) Dendrogram with web logos indicating the relationship among cancer types determined by the trinucleotide contexts of mutations occurring at cytosine nucleotides for the top 50% of *APOBEC3B*-expressing samples in each cancer type. Font size of the bases at the 5′ and 3′ positions are proportional to their observed occurrence in exome mutation data sets. The preferred mutation context for recombinant APOBEC3B from ref. 33 is included in hierarchical clustering to determine how closely each cancer's actual mutation spectrum matches the preferred motif for APOBEC3B *in vitro*. The pattern expected if the mutations were to occur at random cytosine bases in the exome is included as an inset at the bottom left. (**b**) Stacked bars indicate the observed proportion of cytosine mutations at each unique trinucleotide (NCN to NTN, NGN or NAN).The top six cancer types (highlighted by the solid box) show clear biases toward mutations within TCN motifs, at frequencies that resemble the preference of recombinant APOBEC3B *in vitro*[33]. Skin cancer and the bottom seven cancers (highlighted by dashed boxes) have obviously different cytosine mutation spectra.

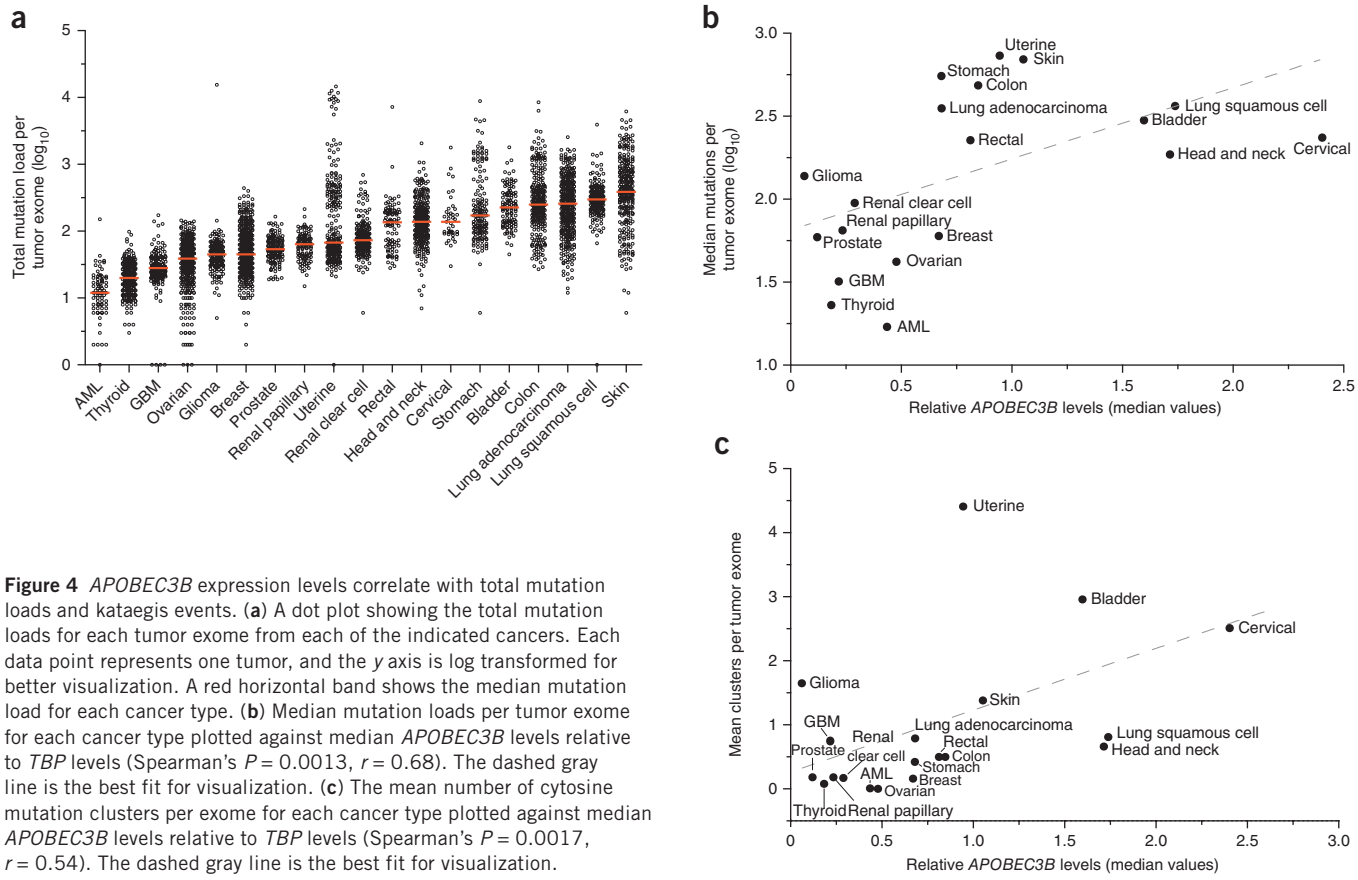


Figure 4 *APOBEC3B* expression levels correlate with total mutation loads and kataegis events. (**a**) A dot plot showing the total mutation loads for each tumor exome from each of the indicated cancers. Each data point represents one tumor, and the *y* axis is log transformed for better visualization. A red horizontal band shows the median mutation load for each cancer type. (**b**) Median mutation loads per tumor exome for each cancer type plotted against median *APOBEC3B* levels relative to *TBP* levels (Spearman's $P = 0.0013$, $r = 0.68$). The dashed gray line is the best fit for visualization. (**c**) The mean number of cytosine mutation clusters per exome for each cancer type plotted against median *APOBEC3B* levels relative to *TBP* levels (Spearman's $P = 0.0017$, $r = 0.54$). The dashed gray line is the best fit for visualization.

We next separated each composite mutation distribution into the 16 individual local trinucleotide contexts to further resolve cytosine-focused mutational mechanisms that might influence each cancer. Bladder, cervical, lung squamous cell carcinoma, lung adenocarcinoma, head and neck, and breast cancers all shared strong bias for TCN mutation signatures (where N is any base), with the strongest bias for TCA of the four possibilities (**Fig. 3b**). A background of other mutations was apparent in the two types of lung cancer, possibly associated with tobacco carcinogens or other mutational mechanisms. The next most obvious signature occurred in skin cancer, as expected, with cytosine-to-thymine transitions predominating in dipyrimidine contexts (**Fig. 3b**). Only two other obvious cytosine-focused mutation patterns were evident. Cytosine-to-thymine mutations in CG contexts dominated at least seven types of cancer, consistent with a CG-targeted mechanism such as spontaneous deamination of methylcytosine (**Fig. 3b**). Finally, uterine, low-grade glioma, rectal and colon cancers had an inordinate number of cytosine-to-adenine transversions in YCT contexts, where Y is either C or T, consistent with at least one additional distinct cytosine-focused mutational mechanism (for example, POLE proofreading domain variants have been implicated in a subset of colorectal tumors[41]).

A fourth prediction of a general mutator hypothesis is that tumor mutation loads correlate with *APOBEC3B* expression levels. To test this possibility on a global level, we used median mutation loads for each tumor type and median *APOBEC3B* expression values. Median values were chosen to ensure the inclusion of all data, yet simultaneously minimized the impact of uncontrollable variables, such as other mutational mechanisms, jackpot effects, bottlenecks and durations of tumor existence. As recently reviewed in ref. 42, mutation loads varied considerably within each tumor type and between the different cancers, with more than a full log difference from the bottom to the top of this range (acute myeloid leukemia to skin cancer; **Fig. 4a**). However, despite this wide variation, a strong positive correlation was found between median mutation loads and *APOBEC3B* expression levels ($P = 0.0013$, $r = 0.68$ by Spearman's correlation; **Fig. 4b**). This result is consistent with the possibility that APOBEC3B may be a general endogenous mutagen that contributes to multiple human cancers, albeit, as outlined above, it clearly contributes much more to a particular subset of cancers. A dominant role for APOBEC3B in a subset of cancers is further supported by significant correlations between mutation loads and *APOBEC3B* expression levels when these analyses were performed for each cancer type on a tumor-by-tumor basis (**Supplementary Figs. 2** and **3**).

A final prediction of a general APOBEC mutator hypothesis is that affected cancers should bear evidence of strand-coordinated clusters of cytosine mutations[9,12,13]. As proposed in ref. 12, clusters can be defined as two or more mutation events within a 10-kb window. By this criterion, every cancer showed evidence of cytosine mutation clustering, with a large range between different cancer types (0.016 to 38 cytosine mutation clusters per tumor). However, it is necessary to apply an additional calculation to take into consideration the sequence length of each cluster, which also varies substantially and can result in the inclusion of false positives (see ref. 12 and Online Methods). This additional filter yielded a much smaller number of likely kataegis events, ranging from 0.002 clusters per ovarian carcinoma to 4.4 clusters per uterine tumor (**Table 1**). Notably, the number of mutations grouped into kataegis events was a relatively small percentage of the total number of cytosine mutations for each cancer (at most 7.9%). However, the mere existence of clustered cytosine mutations in nearly every cancer provides further evidence

for APOBEC involvement. For most cancers, this is likely to be APOBEC3B, as the average number of kataegis events per tumor correlated positively with median *APOBEC3B* expression levels ($P = 0.017$, $r = 0.54$ by Spearman's correlation; **Fig. 4c**). The six cancer types with cytosine mutation signatures that grouped most closely with that of recombinant APOBEC3B—bladder, cervix, lung (adenocarcinoma and squamous cell carcinoma), head and neck, and breast—all showed strong evidence of kataegis, with means of 3.0, 2.5, 0.79, 0.81, 0.66 and 0.16 clusters per tumor, respectively. It is notable that breast cancer is at the low end of this range, but 50-fold higher frequencies would be expected if full genomic sequences had been available (concordant with the analyses in ref. 9). Notably, low-grade gliomas and uterine carcinomas were clear outliers in this analysis, consistent with the close hierarchical clustering of their cytosine mutation signatures (distant from that of recombinant APOBEC3B) and strongly suggesting a distinct mutational mechanism in these cancers.

DISCUSSION

We performed an unbiased analysis of all available DNA deaminase expression profiles and cytosine mutation patterns in 19 different cancer types to try to explain the origin of the cytosine-biased mutation spectra and clustering observed in many different cancers[1–10,13]. Observed cytosine mutation patterns were compared using a hierarchical clustering method to group cancers with similar mutation patterns. Six distinct cancer types—bladder, cervical, lung squamous cell carcinoma, lung adenocarcinoma, head and neck, and breast—clearly stood out, with elevated *APOBEC3B* expression in the majority of tumors, strong overall CG base-pair mutation biases, cytosine mutation contexts that closely resemble the deamination signature of recombinant APOBEC3B and evidence of kataegis events. The most parsimonious explanation for this convergence of independent data sets is that APOBEC3B-dependent genomic DNA deamination is the direct cause of most of the cytosine mutations in these types of cancer. These data are consistent with a general mutator hypothesis in which APOBEC3B mutagenesis has the capacity to broadly shape the mutational landscapes of at least six distinct tumor types and possibly also those of several others, albeit to lesser extents.

The large data sets analyzed here support a model in which upregulated levels of APOBEC3B cause genomic cytosine-to-uracil lesions, which may be processed into a variety of mutagenic outcomes[33] (**Supplementary Fig. 4**). In most cases, uracil lesions are repaired faithfully by canonical base-excision repair. However, in some instances, uracil lesions may template the insertion of adenine bases during DNA synthesis, which might result in cytosine-to-thymine transitions (guanine-to-adenine transitions on the opposing strand). In other cases, genomic uracil bases may be converted to abasic sites by uracil DNA glycosylase. These lesions are non-instructional, such that DNA polymerases, in particular, translesion DNA polymerases, may place any base opposite, with an adenine leading to a transition and a cytosine or thymine leading to a transversion. In addition, uracil lesions that are processed into nicks through the concerted action of a uracil DNA glycosylase and an abasic site endonuclease can result in single- or double-stranded DNA breaks, which are substrates for recombination repair and are undoubtedly intermediates in the formation of cytosine mutation clusters (kataegis)[9,12,13] and larger-scale chromosomal aberrations such as translocations.

The significant positive correlations between *APOBEC3B* expression levels and the percentage of mutations at CG base pairs, the overall mutation loads and the number of kataegis events combine to suggest that most cancers are affected by APOBEC3B-dependent mutagenesis, but unambiguous determinations were not possible for several cancers for a variety of reasons. Skin cancer, for example, has the fifth highest *APOBEC3B* expression level and clear evidence of kataegis, but it also has a strong dipyrimidine-focused cytosine-to-thymine mutation pattern that could easily eclipse an APOBEC3B deamination signature. APOBEC3B may help explain melanomas that occur with minimal UV exposure[43]. Several other cancers, including uterine, rectal, stomach and ovarian, also have significant *APOBEC3B* upregulation and evidence of kataegis, which combine to suggest direct involvement of APOBEC3B, but the trinucleotide cytosine mutation motifs were too distantly related to that of the recombinant enzyme to enable unambiguous associations. Therefore, additional large data sets such as high-depth full-genome sequences will be required to distinguish an APOBEC3B-dependent mechanism unambiguously from the multiple other mechanisms contributing to these tumor types.

We note that we have not completely excluded the possibility of other DNA deaminase family members contributing to mutation in cancer, but, apart from AID in B cell cancers[30], roles for other APOBECs are unlikely to be as great as those of APOBEC3B, namely because other APOBECs (i) have no reported enzymatic activity (APOBEC2 and APOBEC4), (ii) have tissue-restricted expression profiles (AID, APOBEC3A, APOBEC1, APOBEC2 and APOBEC4)[33,35,36,44–48], (iii) are localized to the cytoplasmic compartment (APOBEC3A, APOBEC3D, APOBEC3F, APOBEC3G and APOBEC3H)[29,37,49,50] and (iv) have a completely different intrinsic preference for bases surrounding the target cytosine than APOBEC3B (AID and APOBEC3G prefer 5′ RC and 5′ CC, respectively, where R is either A or G)[33,38–40]. Thus, taken together with the comprehensive analyses presented here of expression data (**Fig. 1**), CG base-pair mutation frequencies (**Fig. 2**), local cytosine mutation signatures (**Fig. 3**), overall mutation loads (**Fig. 4**) and kataegis (**Fig. 4c** and **Table 1**), all available data converge on the conclusion that APOBEC3B is a major source of mutation in multiple human cancers. This knowledge provides a foundation for future studies focused on each cancer type and subtype to further delineate the impact of this potent DNA mutator on each cancer genome and on associated therapeutic responses and patient outcomes.

URLs. TCGA database, http://tcga-data.nci.nih.gov/tcga/; R software, http://www.r-project.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank The Cancer Genome Atlas (TCGA) Network for generating the RNA-seq and somatic mutation data and for providing open access, and we thank the Harris laboratory members and S. Kaufmann for comments. M.B.B. was supported by a Department of Defense Breast Cancer Research Program Predoctoral Fellowship (BC101124). This work was supported by grants from the Jimmy V Foundation, the Minnesota Ovarian Cancer Alliance and the US National Institutes of Health (R01 AI064046 and P01 GM091743).

AUTHOR CONTRIBUTIONS

All authors contributed to the study design, data analysis and manuscript preparation. M.B.B. and N.A.T. analyzed data from TCGA. N.A.T. performed mutation and cluster analysis.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Stephens, P. *et al.* A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat. Genet.* **37**, 590–592 (2005).
2. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
3. Jones, S. *et al.* Frequent mutations of chromatin remodeling gene *ARID1A* in ovarian clear cell carcinoma. *Science* **330**, 228–231 (2010).
4. Sjöblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
5. Kumar, A. *et al.* Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proc. Natl. Acad. Sci. USA* **108**, 17087–17092 (2011).
6. Parsons, D.W. *et al.* The genetic landscape of the childhood cancer medulloblastoma. *Science* **331**, 435–439 (2011).
7. Berger, M.F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
8. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160 (2011).
9. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
10. Stephens, P.J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404 (2012).
11. Makridakis, N.M. & Reichardt, J.K. Translesion DNA polymerases and cancer. *Front. Genet.* **3**, 174 (2012).
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. Drier, Y. *et al.* Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res.* **23**, 228–235 (2013).
14. 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).
15. Di Noia, J.M. & Neuberger, M.S. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* **76**, 1–22 (2007).
16. Longerich, S., Basu, U., Alt, F. & Storb, U. AID in somatic hypermutation and class switch recombination. *Curr. Opin. Immunol.* **18**, 164–174 (2006).
17. Conticello, S.G. The AID/APOBEC family of nucleic acid mutators. *Genome Biol.* **9**, 229 (2008).
18. LaRue, R.S. *et al.* Guidelines for naming nonprimate APOBEC3 genes and proteins. *J. Virol.* **83**, 494–497 (2009).
19. Malim, M.H. APOBEC proteins and intrinsic resistance to HIV-1 infection. *Phil. Trans. R. Soc. Lond. B* **364**, 675–687 (2009).
20. Harris, R.S., Hultquist, J.F. & Evans, D.T. The restriction factors of human immunodeficiency virus. *J. Biol. Chem.* **287**, 40875–40883 (2012).
21. Blanc, V. & Davidson, N.O. C-to-U RNA editing: mechanisms leading to genetic diversity. *J. Biol. Chem.* **278**, 1395–1398 (2003).
22. Bishop, K.N., Holmes, R.K., Sheehy, A.M. & Malim, M.H. APOBEC-mediated editing of viral RNA. *Science* **305**, 645 (2004).
23. Petit, V. *et al.* Murine APOBEC1 is a powerful mutator of retroviral and cellular RNA *in vitro* and *in vivo*. *J. Mol. Biol.* **385**, 65–78 (2009).
24. Ikeda, T. *et al.* Intrinsic restriction activity by apolipoprotein B mRNA editing enzyme APOBEC1 against the mobility of autonomous retrotransposons. *Nucleic Acids Res.* **39**, 5538–5554 (2011).
25. Petersen-Mahrt, S.K., Harris, R.S. & Neuberger, M.S. AID mutates *E. coli* suggesting a DNA deamination mechanism for antibody diversification. *Nature* **418**, 99–103 (2002).
26. Petersen-Mahrt, S.K. & Neuberger, M.S. *In vitro* deamination of cytosine to uracil in single-stranded DNA by apolipoprotein B editing complex catalytic subunit 1 (APOBEC1). *J. Biol. Chem.* **278**, 19583–19586 (2003).
27. Pham, P., Bransteitter, R., Petruska, J. & Goodman, M.F. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* **424**, 103–107 (2003).
28. Chelico, L., Pham, P., Calabrese, P. & Goodman, M.F. APOBEC3G DNA deaminase acts processively 3′ → 5′ on single-stranded DNA. *Nat. Struct. Mol. Biol.* **13**, 392–399 (2006).
29. Hultquist, J.F. *et al.* Human and rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H demonstrate a conserved capacity to restrict Vif-deficient HIV-1. *J. Virol.* **85**, 11220–11234 (2011).
30. Robbiani, D.F. & Nussenzweig, M.C. Chromosome translocation, B cell lymphoma, and activation-induced cytidine deaminase. *Annu. Rev. Pathol.* **8**, 79–103 (2013).
31. Okazaki, I.M. *et al.* Constitutive expression of AID leads to tumorigenesis. *J. Exp. Med.* **197**, 1173–1181 (2003).
32. Yamanaka, S. *et al.* Apolipoprotein B mRNA-editing protein induces hepatocellular carcinoma and dysplasia in transgenic animals. *Proc. Natl. Acad. Sci. USA* **92**, 8483–8487 (1995).
33. Burns, M.B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494**, 366–370 (2013).
34. Jarmuz, A. *et al.* An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics* **79**, 285–296 (2002).
35. 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).
36. Koning, F.A. *et al.* Defining APOBEC3 expression patterns in human tissues and hematopoietic cell subsets. *J. Virol.* **83**, 9474–9485 (2009).
37. Lackey, L. *et al.* APOBEC3B and AID have similar nuclear import mechanisms. *J. Mol. Biol.* **419**, 301–314 (2012).
38. Kohli, R.M. *et al.* Local sequence targeting in the AID/APOBEC family differentially impacts retroviral restriction and antibody diversification. *J. Biol. Chem.* **285**, 40956–40964 (2010).
39. Wang, M., Rada, C. & Neuberger, M.S. Altering the spectrum of immunoglobulin V gene somatic hypermutation by modifying the active site of AID. *J. Exp. Med.* **207**, 141–153 (2010).
40. 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).
41. Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat. Genet.* **45**, 136–144 (2013).
42. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
43. Berger, M.F. *et al.* Melanoma genome sequencing reveals frequent *PREX2* mutations. *Nature* **485**, 502–506 (2012).
44. Fujino, T., Navaratnam, N. & Scott, J. Human apolipoprotein B RNA editing deaminase gene (*APOBEC1*). *Genomics* **47**, 266–275 (1998).
45. Muramatsu, M. *et al.* Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. *J. Biol. Chem.* **274**, 18470–18476 (1999).
46. 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).
47. Sato, Y. *et al.* Deficiency in APOBEC2 leads to a shift in muscle fiber type, diminished body mass, and myopathy. *J. Biol. Chem.* **285**, 7111–7118 (2010).
48. Rogozin, I.B., Basu, M.K., Jordan, I.K., Pavlov, Y.I. & Koonin, E.V. APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases predicted by computational analysis. *Cell Cycle* **4**, 1281–1285 (2005).
49. Rada, C., Jarvis, J.M. & Milstein, C. AID-GFP chimeric protein increases hypermutation of Ig genes with no evidence of nuclear localization. *Proc. Natl. Acad. Sci. USA* **99**, 7003–7008 (2002).
50. Land, A.M. *et al.* Endogenous APOBEC3A DNA cytosine deaminase is cytoplasmic and non-genotoxic. *J. Biol. Chem.* **288**, 17253–17260 (2013).

ONLINE METHODS

Data analyses. A description of tumor types, tumor *APOBEC3B* expression data and tumor exome mutation data is provided in **Table 1**. Information for the corresponding normal tissues is provided in **Supplementary Table 1**. Somatic mutations and RNA-seq expression data were retrieved from The Cancer Genome Atlas (TCGA) Data Matrix on 3 January 2013. Gene expression data were mined from RNAseqV2 data sets for all cancers (normalized expression values), with the exception of LAML and STAD, which were from RNA-seq data sets (reads per kilobase of transcript per million mapped reads (RPKM) values). Additional RNAseqV2 data for normal samples were downloaded from TCGA on 4 April 2013 to include recently released normal sample information for READ and COAD. *APOBEC3B* expression values were normalized to the expression of *TBP* for each tumor sample. Comparisons between the normal RNA-seq–derived gene expression values and the tumor expression values were performed using the Mann-Whitney *U* test to determine significance. All qRT-PCR values for normal tissues were reported previously based on data from pooled normal samples[33,35], with the exception of salivary gland, stomach, skin and rectal tissues, which are unique to this report. Primary-tissue RNA was generated using published methods[35], and total RNA was obtained commercially (salivary gland RNA for head and neck cancer and stomach RNA were obtained from Clontech, and skin and rectal RNA were obtained from US Biological). Each *APOBEC3B* expression value relative to the *TBP* value from qRT-PCR was multiplied by an experimentally derived factor of 2 to facilitate direct comparisons with RNA-seq values (B. Leonard, S.N. Hart, M.B.B., M.A. Carpenter, N.A.T. *et al.*, unpublished data).

Mutation data were taken from maf files downloaded from the TCGA database. Insertions-deletions and adjacent multiple mutations (di- and trinucleotide variations) were removed, and the remaining single-nucleotide variations (SNVs) were converted to hg19 coordinates (**Supplementary Table 3**). Non-mutations with respect to the reference genome (for example, cytosine-to-cytosine changes) were eliminated, and duplicate entries were removed unless they were reported for different tumor samples. Comparisons between mutations and gene expression were calculated using Spearman's rank correlation.

Trinucleotides with cytosines in the center position were used to calculate the sequence context dependence of mutations. There are a total of 16 unique trinucleotides containing cytosine in the center position. The corresponding 16 reverse complements were also included in the analysis, but, for simplicity, discussion was focused on the cytosine-containing strand. For each unique trinucleotide, the observed cytosine-to-thymine, cytosine-to-guanine and cytosine-to-adenine mutations were counted and placed in a table and were then internally normalized to 1 to reflect the fraction of each mutation type (for a given cancer, each mutation type was normalized as a proportion of the total mutations). The resulting table reflects the global mutation profile of cytosines for each cancer. These data were then used to hierarchically cluster the cancer mutation signatures. This was done using the hclust function of R using Euclidean distance and the 'complete' option. The Euclidean distance is the ordinary distance between two data points on a two-dimensional plot (**Supplementary Table 2** lists all calculated Euclidean distances).

A kataegis event was defined as two or more mutations within a 10,000-nucleotide genomic DNA window. The probability of each event occurring by chance was then calculated according to the work of Gordenin and colleagues[12]. Briefly, the *P* value of observing a given number of mutations within a given number of base pairs was calculated using a negative binomial distribution of the genomic size of each event, the number of mutations in each event and the base probability of finding a random mutation in the exome (number of mutations in each cancer type divided by the number of subjects and exome size). Significant kataegis events with *P* values less than 1×10^{-4} for each cancer are reported in **Table 1**. 'Gordenin significance' indicates that a given cluster of mutations met the above criteria and attained significance. This approach minimizes false positive cluster calls resulting by random chance.

doi:10.1038/ng.2701

Supplementary online materials for:

Evidence for APOBEC3B mutagenesis in multiple human cancers

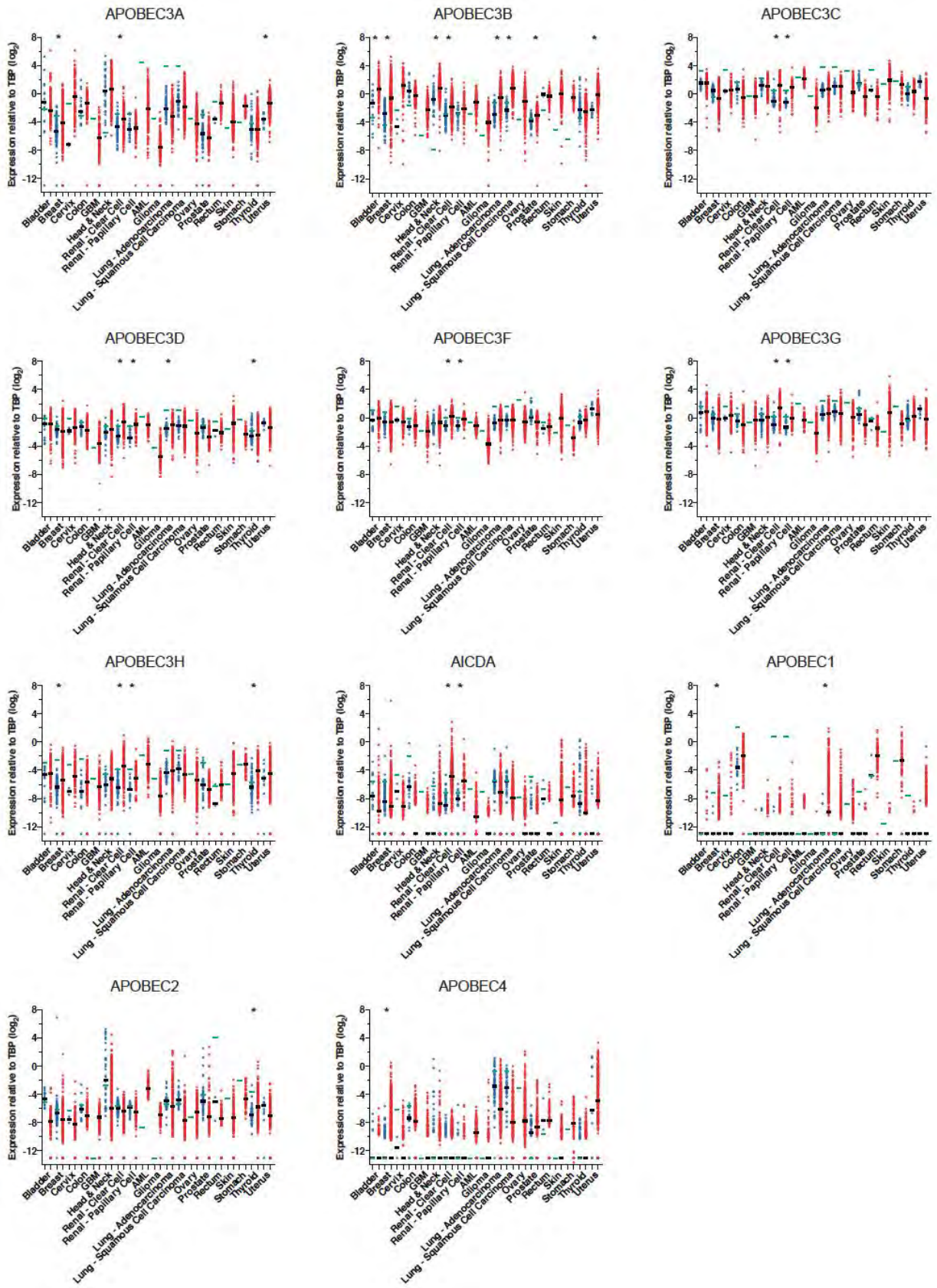
Michael B. Burns[1-4], Nuri A. Temiz[1-2] & Reuben S. Harris[1-4,#]

[1]Biochemistry, Molecular Biology and Biophysics Department, [2]Masonic Cancer Center, [3]Institute for Molecular Virology, [4]Center for Genome Engineering, University of Minnesota, Minneapolis, MN 55455, USA

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

This section contains Supplementary Figures 1-4 and Supplementary Tables 1-3.

Nature Genetics: doi:10.1038/ng.2701

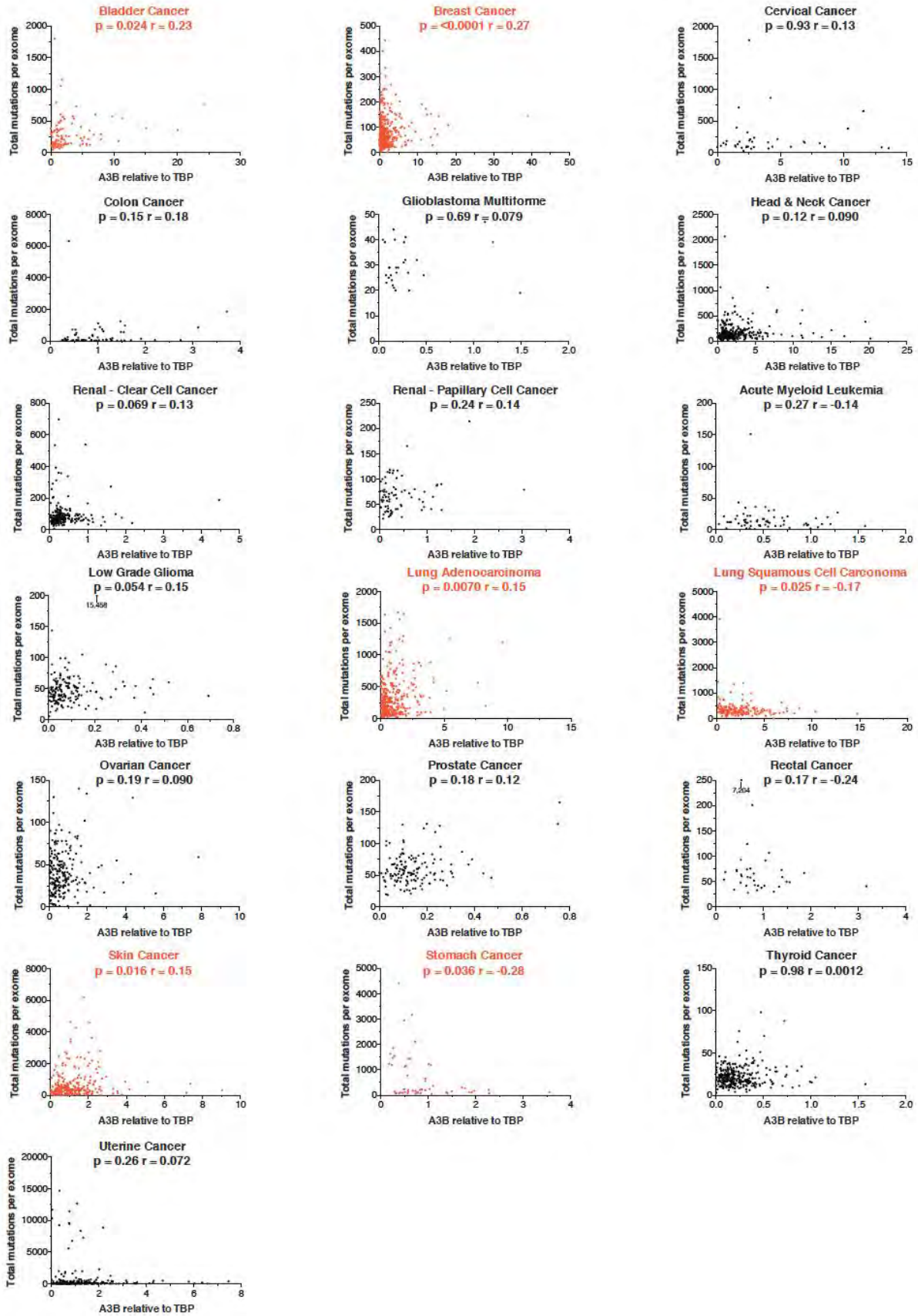


Supplementary Fig. 1. *APOBEC* family member mRNA expression levels for all 19 cancers analyzed here. See next page for full legend.

Nature Genetics: doi:10.1038/ng.2701

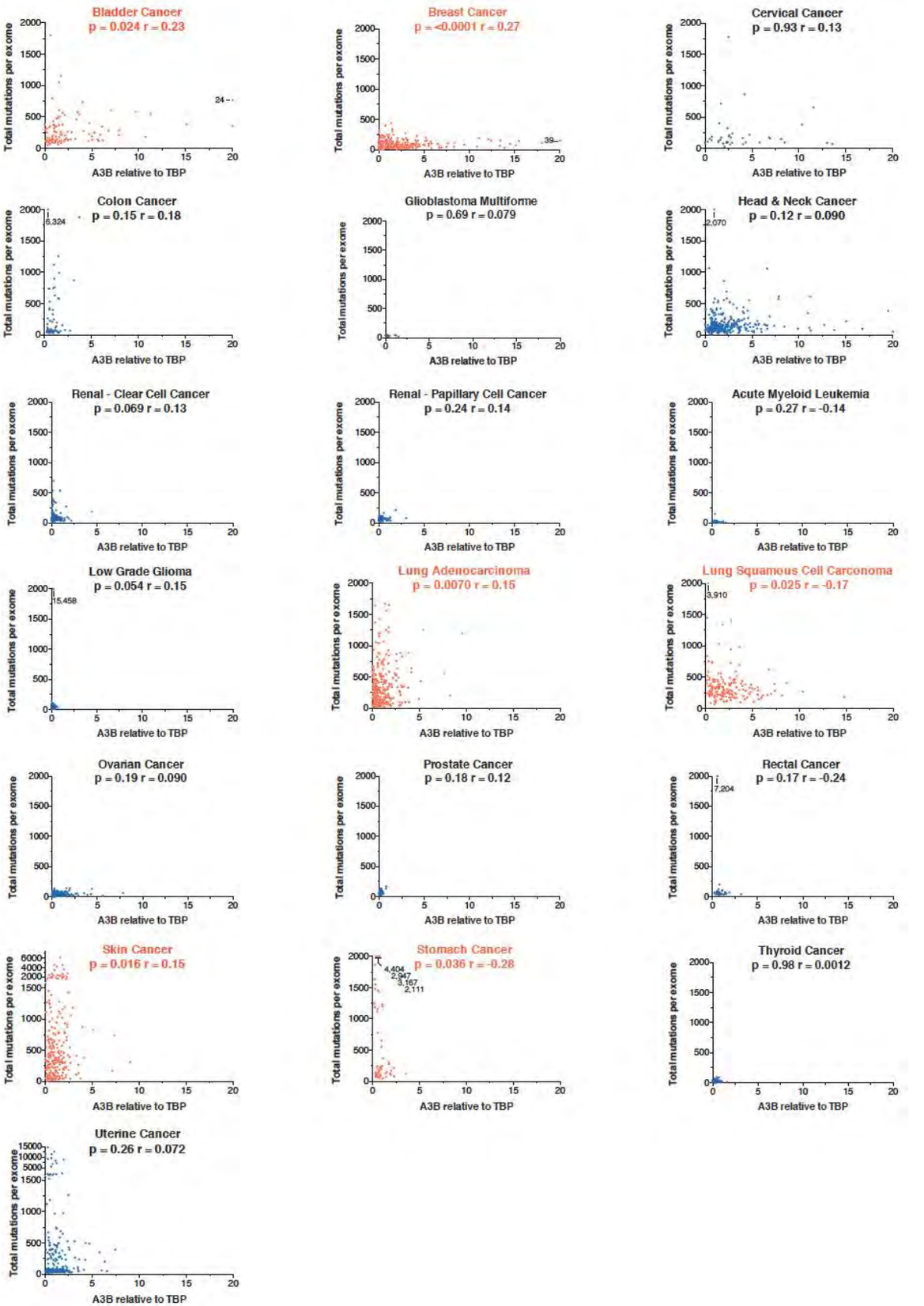
Supplementary Fig. 1. *APOBEC* family member mRNA expression levels for all 19 cancers analyzed here. RNAseq and RT-qPCR data for expression of the indicated *APOBEC* family member genes relative to the housekeeping gene, *TBP*. Each data point represents one tumor (red symbol) or normal (blue symbol) sample, and the Y-axis is log-transformed for better data visualization. Black horizontal lines indicate the median *APOBEC*/*TBP* value for each cancer or normal data set (**Table 1** and **Supplementary Table 1**). Green horizontal lines indicate the *APOBEC*/*TBP* value determined by RT-qPCR. Asterisks indicate significant upregulation of the indicated gene in the tumor relative to the corresponding normal tissues ($p<0.0001$ by Mann-Whitney U-test). APOBEC3B expression data are reproduced from **Fig. 1** for comparison with other family members. The positive expression correlations in the two types of renal tumors for nearly all *APOBEC* family members cannot be explained at this time. The positive association of *APOBEC3A* in breast and bladder cancer may be due to infiltrating macrophages, as this mRNA is only expressed in myeloid lineage cell types and is not present in breast cancer cell lines (refs. 33 & 35). The positive correlations for *APOBEC3D* in lung adenocarcinoma and thyroid cancers barely reach significance. The positive correlations for *APOBEC3H*, *APOBEC1*, and *APOBEC4* in breast cancer were not observed previously by RT-qPCR in tumors with patient-matched normal tissues as controls (ref. 33). The positive correlations for *APOBEC3H* and *APOBEC2* in thyroid cancer and *APOBEC1* in lung adenocarcinoma are not explainable at this time and could be interesting subjects for further work. P-values for negative or insignificant associations are not indicated in this figure. Overall, although these data indicate that *APOBEC3B* is the most abundantly upregulated *APOBEC* family member across the many different cancers, these data are only one line of evidence suggesting a role in cancer and they must be interpreted in alongside other analyses presented here and in prior literature, which impose strong biochemical, genetic, and cellular constraints on what is and is not possible or plausible (see **Results** and **Discussion**).

Nature Genetics: doi:10.1038/ng.2701



Supplementary Fig. 2a. Correlations between total mutation loads and *APOBEC3B* expression levels.
See page 6 for full legend.

Nature Genetics: doi:10.1038/ng.2701



Supplementary Fig. 2b. Correlations between total mutation loads and *APOBEC3B* expression levels. See page 6 for full legend.

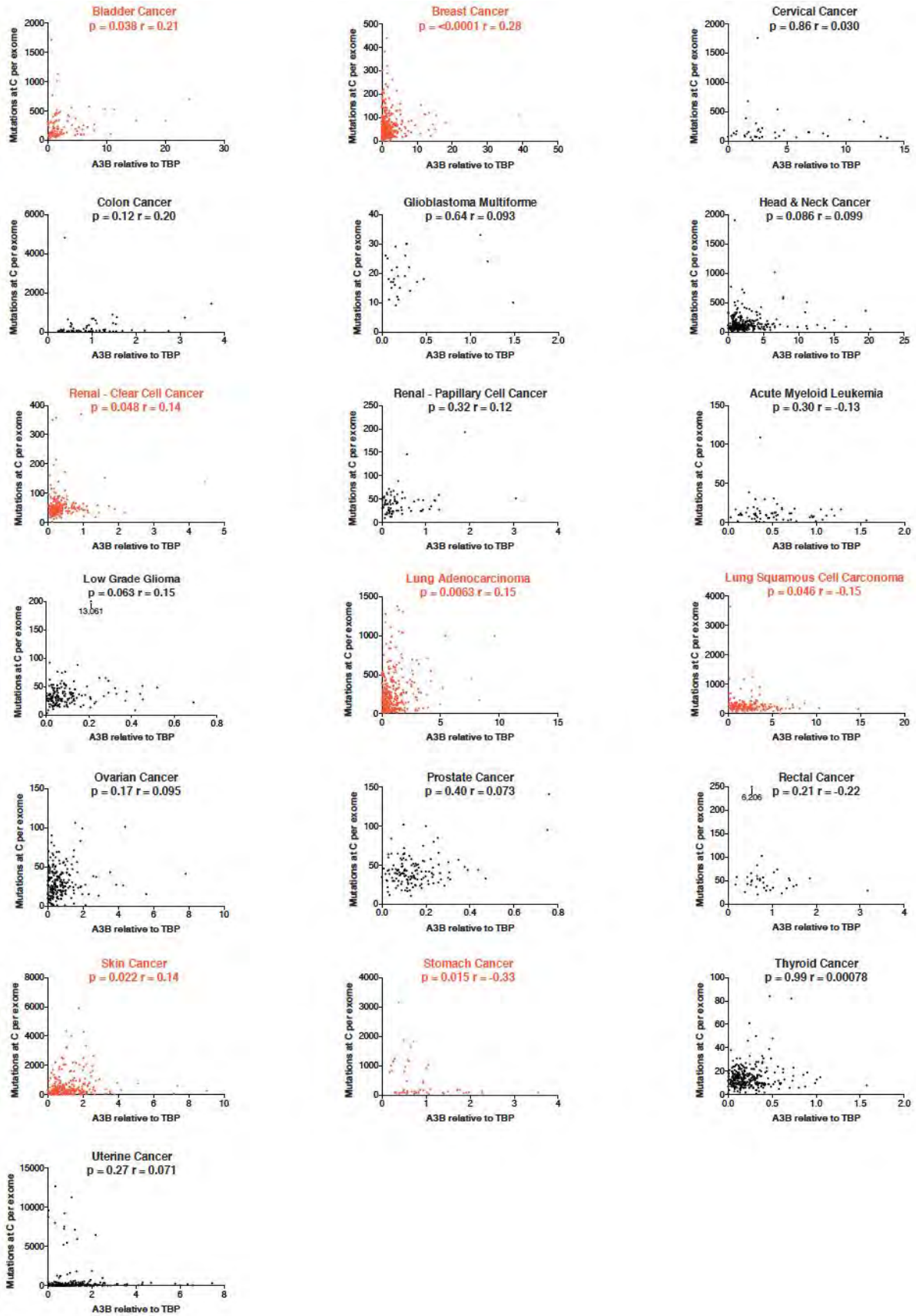
Nature Genetics: doi:10.1038/ng.2701

Supplementary Fig. 2. Correlations between total mutation loads and *APOBEC3B* expression levels.

(**a**) Total exonic mutation loads plotted against *APOBEC3B*/*TBP* expression levels for each of the 19 tumor types analyzed here. P and r-values are from Spearman's correlation. Data sets with p-values less than or equal to 0.05 are highlighted in red. The high variability in mutation loads amongst each tumor type is due to the stochastic nature of the underlying mutational processes, different tumor ages, differential repair capacities, selection bottlenecks, chemotherapeutic drug exposures, *etc.*

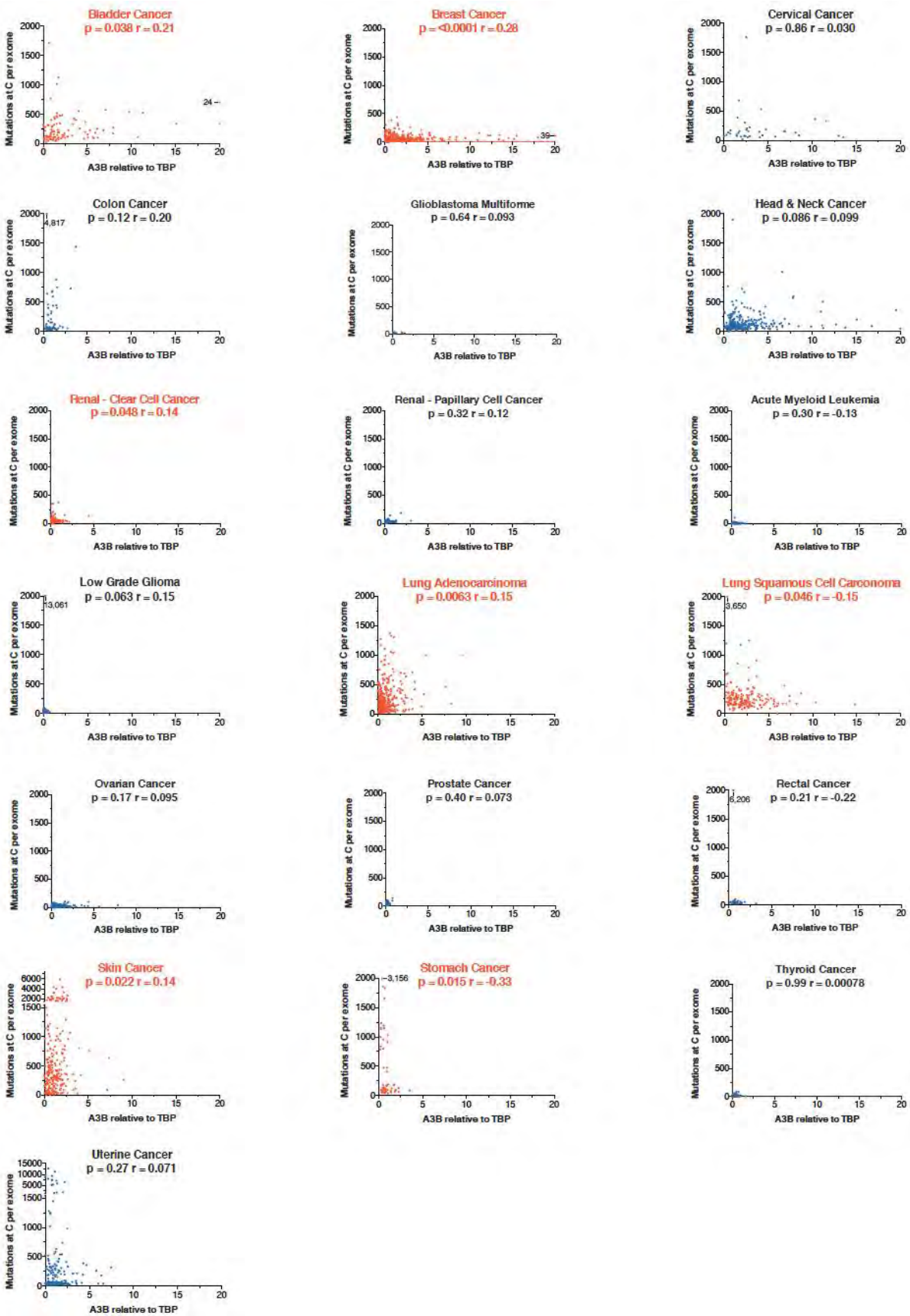
(**b**) The same data as in panel (a) but projected onto fixed axes to facilitate comparison between tumor types.

Nature Genetics: doi:10.1038/ng.2701



Supplementary Fig. 3a. Correlations between C/G-specific mutation counts and *APOBEC3B* expression levels. See page 9 for full legend.

Nature Genetics: doi:10.1038/ng.2701



Supplementary Fig. 3b. Correlations between C/G-specific mutation counts and *APOBEC3B* expression levels. See page 9 for full legend.

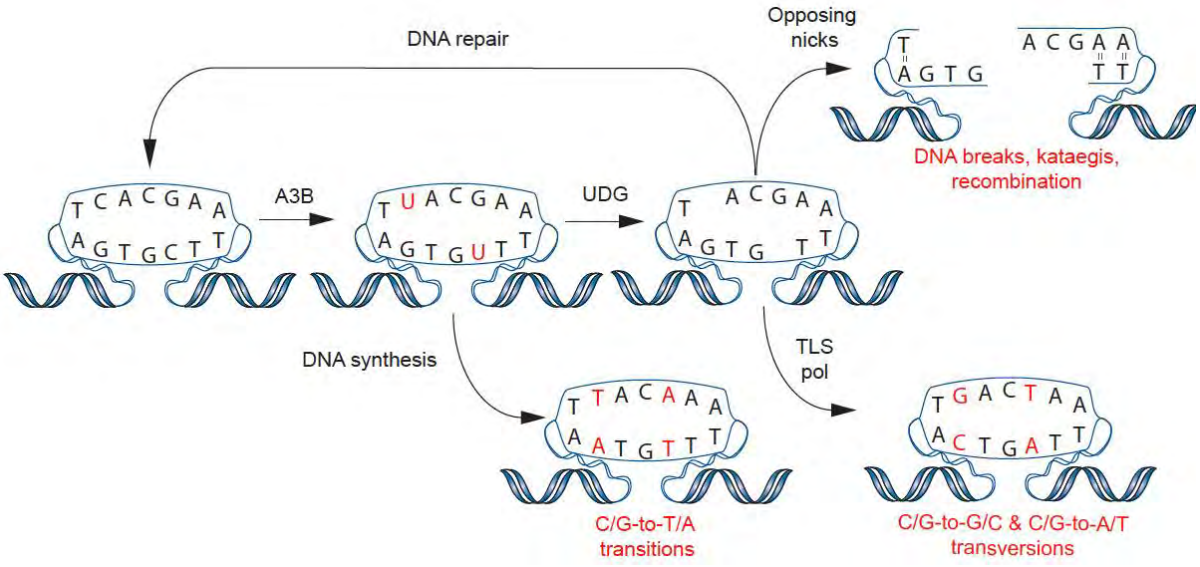
Nature Genetics: doi:10.1038/ng.2701

Supplementary Fig. 3. Correlations between C/G-specific mutation counts and *APOBEC3B* expression levels.

(**a**) Exonic C/G mutation counts plotted against *APOBEC3B*/*TBP* expression levels for each of the 19 tumor types analyzed here. P and r-values are from Spearman's correlation. Data sets with p-values less than or equal to 0.05 are highlighted in red. The high variability in mutation loads among each tumor type is due to the stochastic nature of the underlying mutational processes, different tumor ages, differential repair capacities, selection bottlenecks, chemotherapeutic drug exposures, *etc.*

(**b**) The same data as in panel (a) but projected onto fixed axes to facilitate comparison between tumor types.

Nature Genetics: doi:10.1038/ng.2701



Supplementary Fig. 4. Model for APOBEC3B-induced mutagenesis in cancer. APOBEC3B deaminates genomic cytosines in preferred contexts resulting in uracils. DNA repair by uracil DNA glycosylase (UDG) and canonical base excision repair may correct many lesions. C-to-T transitions may result from DNA synthesis templated directly by genomic uracils or from DNA synthesis to bypass abasic sites (following established 'A-rule', not shown). C-to-G and C-to-A transversions may result during bypass of template abasic sites by a translesion synthesis DNA polymerase (TLS pol). Abasic sites may be further processed by a base excision repair endonuclease (APEX, not shown) into nicks, which can lead to single- and double-stranded DNA breaks, to exposed single-stranded DNA and kataegis events, as well as to recombination and larger-scale genomic aberrations such as translocations. Model adapted from Ref. 33.

Nature Genetics: doi:10.1038/ng.2701

Supplementary Table 1. Summary statistics for the normal control samples in this study.

		***A3B* expression in normal controls[1]**			***A3B* expression in normal controls[2]**
Tumor Type	**TCGA ID**	**n**	**Range**	**Median**	**Mean of 3 measurements**
Low Grade Glioma	LGG	n.a	n.a.	n.a.	0.016
Prostate adenocarcinoma	PRAD	44	0.017 - 0.21	0.41	0.090
Thyroid carcinoma	THCA	58	0.0058 - 5.1	1.0	0.10
Glioblastoma multiforme	GBM	n.a	n.a.	n.a.	0.016
Kidney renal papillary cell carcinoma	KIRP	25	0.029 - 0.43	0.10	0.14
Kidney renal clear cell carcinoma	KIRC	71	0.024 - 1.7	0.25	0.14
Acute myeloid leukemia	LAML	n.a	n.a.	n.a.	0.092
Ovarian serous cystadenocarcinoma	OV	n.a	n.a.	n.a.	0.080
Breast invasive carcinoma	BRCA	107	0.0081 - 0.69	0.15	0.048
Stomach adenocarcinoma	STAD	n.a	n.a.	n.a.	0.012
Lung adenocarcinoma	LUAD	57	0.037 - 0.89	0.16	0.44
Rectum adenocarcinoma	READ	3	0.78 - 1.8	0.54	0.21
Colon adenocarcinoma	COAD	18	0.46 - 7.7	2.0	0.34
Uterine corpus endometrioid carcinoma	UCEC	11	0.10 - 0.42	0.10	n.a.
Skin cutaneous melanoma	SKCM	n.a	n.a.	n.a.	0.030
Bladder urotheilal carconoma	BLCA	16	0.014 - 2.6	0.66	0.10
Head & neck squamous cell carcinoma	HNSC	37	0.049 - 5.9	1.0	0.0042
Lung squamous cell carcinoma	LUSC	35	0.027 - 0.77	0.16	0.44
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	2	0.021 - 0.085	0.099	0.20

[1]*A3B* expression values relative to those of the housekeeping gene *TBP*, determined by RNAseq.

[2]*A3B* expression values relative to those of the housekeeping gene *TBP*, determined by qPCR.

Supplementary Table 2. Euclidean distances between each tumor type and the signature of recombinant APOBEC3B (recA3B).

	recA3B	BLCA	BRCA	CESC	COAD	GBM	HNSC	KIRC	KIRP	LAML	LGG	LUAD	LUSC	OV	PRAD	READ	SKCM	STAD	THCA	UCEC
recA3B	-	0.180	0.178	0.190	0.328	0.241	0.162	0.213	0.179	0.299	0.302	0.179	0.154	0.202	0.271	0.311	0.320	0.337	0.220	0.278
BLCA	0.180	-	0.123	0.040	0.317	0.221	0.102	0.219	0.160	0.288	0.299	0.202	0.173	0.203	0.258	0.295	0.293	0.316	0.182	0.300
BRCA	0.178	0.123	-	0.135	0.211	0.132	0.036	0.115	0.064	0.175	0.197	0.134	0.111	0.092	0.140	0.189	0.295	0.210	0.078	0.217
CESC	0.190	0.040	0.135	-	0.322	0.236	0.116	0.235	0.176	0.296	0.308	0.221	0.189	0.218	0.266	0.299	0.312	0.319	0.193	0.309
COAD	0.328	0.317	0.211	0.322	-	0.217	0.233	0.186	0.203	0.100	0.093	0.260	0.256	0.193	0.099	0.112	0.394	0.057	0.167	0.171
GBM	0.241	0.221	0.132	0.236	0.217	-	0.147	0.139	0.133	0.167	0.205	0.167	0.165	0.110	0.142	0.195	0.312	0.214	0.128	0.239
HNSC	0.162	0.102	0.036	0.116	0.233	0.147	-	0.126	0.071	0.199	0.215	0.125	0.097	0.108	0.165	0.215	0.289	0.235	0.097	0.230
KIRC	0.213	0.219	0.115	0.235	0.186	0.139	0.126	-	0.067	0.151	0.159	0.108	0.109	0.060	0.118	0.197	0.312	0.202	0.086	0.199
KIRP	0.179	0.160	0.064	0.176	0.203	0.133	0.071	0.067	-	0.169	0.177	0.110	0.096	0.065	0.133	0.203	0.288	0.214	0.065	0.203
LAML	0.299	0.288	0.175	0.296	0.100	0.167	0.199	0.151	0.169	-	0.138	0.223	0.220	0.148	0.059	0.098	0.363	0.087	0.127	0.207
LGG	0.302	0.299	0.197	0.308	0.093	0.205	0.215	0.159	0.177	0.138	-	0.225	0.225	0.166	0.124	0.160	0.374	0.131	0.159	0.140
LUAD	0.179	0.202	0.134	0.221	0.260	0.167	0.125	0.108	0.110	0.223	0.225	-	0.045	0.102	0.192	0.253	0.322	0.273	0.154	0.243
LUSC	0.154	0.173	0.111	0.189	0.256	0.165	0.097	0.109	0.096	0.220	0.225	0.045	-	0.099	0.187	0.246	0.316	0.267	0.141	0.241
OV	0.202	0.203	0.092	0.218	0.193	0.110	0.108	0.060	0.065	0.148	0.166	0.102	0.099	-	0.113	0.183	0.311	0.199	0.082	0.202
PRAD	0.271	0.258	0.140	0.266	0.099	0.142	0.165	0.118	0.133	0.059	0.124	0.192	0.187	0.113	-	0.099	0.349	0.096	0.097	0.187
READ	0.311	0.295	0.189	0.299	0.112	0.195	0.215	0.197	0.203	0.098	0.160	0.253	0.246	0.183	0.099	-	0.382	0.080	0.165	0.192
SKCM	0.320	0.293	0.295	0.312	0.394	0.312	0.289	0.312	0.288	0.363	0.374	0.322	0.316	0.311	0.349	0.382	-	0.398	0.291	0.368
STAD	0.337	0.316	0.210	0.319	0.057	0.214	0.235	0.202	0.214	0.087	0.131	0.273	0.267	0.199	0.096	0.080	0.398	-	0.171	0.202
THCA	0.220	0.182	0.078	0.193	0.167	0.128	0.097	0.086	0.065	0.127	0.159	0.154	0.141	0.082	0.097	0.165	0.291	0.171	-	0.202
UCEC	0.278	0.300	0.217	0.309	0.171	0.239	0.230	0.199	0.203	0.207	0.140	0.243	0.241	0.202	0.187	0.192	0.368	0.202	0.202	-

Supplementary Table 3. Description of the mutation subset analysed in this study.

Tumor Type	TCGA ID	Number of tumors	Total mutations	Filtered mutations	Percent of mutations filtered (non-SNP)
Low Grade Glioma	LGG	170	24650	1213	5%
Prostate adenocarcinoma	PRAD	150	9784	881	9%
Thyroid carcinoma	THCA	326	12143	4826	40%
Glioblastoma multiforme	GBM	167	5862	146	2%
Kidney renal papillary cell carcinoma	KIRP	100	8068	1167	14%
Kidney renal clear cell carcinoma	KIRC	244	33280	10811	32%
Acute myeloid leukemia	LAML	74	1368	137	10%
Ovarian serous cystadenocarcinoma	OV	469	28049	2227	8%
Breast invasive carcinoma	BRCA	777	52160	6290	12%
Stomach adenocarcinoma	STAD	156	100913	14899	15%
Lung adenocarcinoma	LUAD	392	152307	13269	9%
Rectum adenocarcinoma	READ	88	21199	1181	6%
Colon adenocarcinoma	COAD	266	148114	18503	12%
Uterine corpus endometrioid carcinoma	UCEC	248	184829	5719	3%
Skin cutaneous melanoma	SKCM	255	186839	9207	5%
Bladder urotheilal carconoma	BLCA	99	30801	1948	6%
Head & neck squamous cell carcinoma	HNSC	306	63508	8282	13%
Lung squamous cell carcinoma	LUSC	177	65306	967	1%
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	39	10021	936	9%