APOBEC3B-driven mutagenesis in breast and other human cancers

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Michael Bradley Burns

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Reuben S. Harris

August 2013

Acknowledgements

I would like to thank my mentor and advisor, Dr. Reuben Harris, for his persistent encouragement and guidance. I would also like to thank my co-authors and lab mates Lela Lackey and Nuri Alpay Temiz; without their help, this work would not have progressed as far as it has. To the members of the Harris lab, past and present, you have my heartfelt thanks for all of your support, both as academic colleagues and as friends. I must also thank my committee members Drs. Michel Sanders, Carol Lange, and Alexandra Sobeck for their critiques, suggestions, and comments as I made my way through graduate school. This work would have been much more difficult to complete had it not been for the hard work and intellectual contributions of my many collaborators. I thank you. I would be remiss to not acknowledging the support of several other people otherwise not mentioned in this text: Dr. Kevin Silverstein for helping me start my bioinformatic analysis of cancer genomes and Darlene Toedter and David Okita for keeping the administrative wheels well-greased. Lastly, I must thank my parents, siblings, and close friends for their unfailing confidence in me.

Dedication

This thesis is dedicated to my lovely wife, Elaine, for her patience and support and to my children, Ezra and Elsa, for being a welcome distraction.

Abstract

Cancer is a disease that results from alteration of the cellular genome. The sources of these changes are multifarious, and in many cases unknown. This thesis focuses on the polynucleotide cytosine (C) deaminase, APOBEC3B, as a newly discovered source of mutation in multiple human cancers. As a deaminase, APOBEC3B converts Cs to uracils (Us) in single-stranded DNA. These U lesions are mutagenic as failure to properly repair them can result in a wide variety of mutation types.

The initial discovery of this mutational phenomenon was described mechanistically using a variety of biochemical, genetic, and cellular assays in breast cancer cell lines. Follow-up work using publicly available next generation sequencing and clinical data indicated that this effect is operating in a large proportion of breast cancers. Expanded bioinformatic analysis that assessed APOBEC3B's potential impact was expanded to include 18 other human cancer types in addition to breast. This work shows that APOBEC3B is likely a significant contributor to the genetic heterogeneity in breast, head & neck, bladder, cervical, and lung (adeno- and squamous cell) carcinomas as evidenced by differential levels of expression in cancer tissues, increased mutation load, mutation clusters (kataegis), and an APOBEC3B mutation signature in tumors expressing high levels of this enzyme. Taken together, this thesis is a body of work describing a previously unappreciated source of genetic heterogeneity in several human cancers.

Table of Contents

List of Tables

List of Figures

CHAPTER 1 – INTRODUCTION

MUTATION IN CANCER

Cancer is a genetic disease. For a malignancy to form and subsequently spread, it needs to override the normal cellular safeguards encoded in the genome. Therefore, to prevent cancer formation and limit its malignant potential, it is vital to understand the different sources of DNA damage that can lead to cancer formation and evolution. There are several hurdles a cancer needs to overcome including: 1) evasion of growth suppression, 2) activating invasion and metastasis, 3) enabling replicative immortality, 4) inducing angiogenesis, 5) resisting cell death, and 6) sustaining proliferative signaling[4]. Each of these "hallmarks of cancer" can be attained through genetic remodeling as the result of mutation. The mutational processes that result in these changes can be classified into the general categories of exogenous and endogenous, where the exogenous sources are those that arise from the environment and the endogenous sources are those that arise from within the cell itself. Exogenous sources of mutation are the simplest to understand, as many of these agents are obvious. Pyrimidine dimers as a result of UV damage are a classic example of an exogenous agent generating lesions that lead to mutation[5,6]. The endogenous processes can be categorized into active and passive sources of mutation depending on whether they directly damage/mutate DNA or if they simply fail repair DNA damage after it has occured. Examples of passive, endogenous source of damage are the mutant alleles of *BRCA1* and *BRCA2* that predispose patients to breast and ovarian cancers[7]. Patients with these mutations develop cancer due to a defect in the normal DNA repair processes. This defect prevents cells from being able to properly correct double-stranded DNA breaks through homologous recombination, thus allowing mutations to accumulate. The focus of this work, however, is an active endogenous source of DNA damage and mutation, namely, the APOBEC family of single-stranded DNA polynucleotide cytosine deaminases.

THE APOBEC FAMILY

The human APOBEC family includes 11 members, the namesake, Apolipoprotein B mRNA Editing enzyme, Catalytic Subunit 1 (*APOBEC1*) on chromosome 6, *APOBEC2* on chromosome 12, *APOBEC3A*, *B*, *C*, *D*, *F*, *G*, *H* all in a tandem head-to-tail array on chromosome 22, *APOBEC4* on chromosome 1, and Activation Induced Cytidine Deaminase (*AICDA*) on chromosome 6. With the exception of *AICDA*, these genes were all named after *APOBEC1* as a result of sequence homology indicating a shared lineage of inter and intra chromosomal gene duplication. However, this nomenclature is misnomeric as *APOBEC1* is the only family member whose gene product carries out the indicated mRNA-editing[8-10]. Most of the APOBEC enzymes are capable of editing cytosine in single-stranded DNA (ssDNA) via a deamination reaction that converts it to uracil (C-to-U). Conversely, APOBEC2 and APOBEC4 have not shown catalytic activity in any of the assays used to date. Uracil can also be generated in cellular DNA as the result of spontaneous hydrolytic deamination, though this is thought to occur at a relatively low frequency *in vivo*[11].

One of the challenges to assessing the potential role of the APOBEC family in cancer is that there are numerous family members that share regions of homology. While

APOBEC-specific monoclonal antibodies (mAb) have been generated, there lack specificity in practice. This makes quantitative determination of the presence or absence of individual family members effectively impossible at the protein level when examining endogenous expression in primary tissues. Fortunately, this is not the case at the level of transcription. While still similar, the mRNA for each of the family members are sufficiently distinct that we have been able to construct and validate a panel of reagents that includes primer pairs specific to each APOBEC that can be used to quantify transcript levels by reverse-transcription and quantitative PCR (RT-qPCR)[2].

The APOBEC family members have been predicted to play a role in cancer since the initial discovery in 2002 that they were able use DNA as a substrate, and not RNA as originally suggested by homology[9]. At that time, due to the aforementioned technological hurdles (*i.e.* the difficulty in determining which APOBEC was which), it was hypothesized that perhaps APOBEC3G or other family members might be driving mutation in cancer. In fact, it has been suggested that APOBEC3G contributes to metastasis in hepatocellular carcinoma, though the research neither proposed nor tested a mechanistic explanation of the observation[12]. Since then, with the use of precise RT-qPCR, APOBEC3G is not currently a suspect in cancer onset or progression as there has not yet been a group to discover abnormal levels of APOBEC3G in human cancer tissue when specific assays are used. There is a chance that the normal level of APOBEC3G expressed in a given tissue may be misregulated at the post-transcriptional level, but again, there has been no evidence presented to support this hypothesis.

Aside from the previous literature on APOBEC3G, two other family members are implicated in carcinogenesis – APOBEC1 in a set of ectopic model systems[13] and AID in murine and a limited set of primary human cancers[14,15]. While rat APOBEC1 clearly has a dramatic carcinogenic effect when expressed constitutively in transgenic mice, no one has yet demonstrated that this "can-do" effect is actually seen in any human cancers, limiting its relevance to human cancer biology. The discovery of the enzyme AID as a causative agent in some B-cell cancers was excellent proof of principle and a starting place from which to follow-up the potential role of the other family members. The AID research mechanistically explained how, in the affected B-cells, the deaminase was already expressed and acting on genomic DNA as part of its natural function.

DEAMINATION AND MUTATION IN THE GENOME

What happens when uracil is present in genomic DNA? As a simple lesion, uracil alone does not appear to have significant toxicity[16]. Figure 1-1 depicts the potential outcomes arising from an initiating DNA cytosine deamination. Once the deamination has taken place, the genomic uracil base-pairs as a T, but since the opposite base is a guanine, the U:G pair is a mismatch and subject to repair from at least three different mechanisms. In the blue, central panel of Figure 1-1, which depicts a standard repair by the base excision pathway (BER), the lesion is initially identified by a uracil DNA glycosylase (UDG). This family of monofunctional glycosylases has several members. SMUG1 is capable of recognizing deoxyuracil (dU) and 5-hydroxymethyluracil in genomic DNA[17]. Additionally, the two UNG isoforms, which are generated by alternative splicing of the mRNA transcribed from the *UNG* gene, each recognize dU and 5-hydroxyuracil in DNA[17,18]. These isoforms, which are identical but for a short 44 amino acid stretch at the N-terminal domain, are referred to as UNG1, the mitochondrial UDG, and UNG2, the nuclear isoform[18]. Kemmerich and colleagues, in an elegant series of experiments, clarified the distinct mechanisms by which UNG2 and SMUG1 recognize U:G mismatches and "pass the baton" to the downstream repair pathway members[17]. Their research, suggests that UNG2 is likely the only UDG to play a major role in the repair of APOBEC3-induced lesions in human cells. Once the dU is recognized, the aberrant base is flipped in the active site of UNG2 and the N-glycosidic bond tethering the uracil in place is cleaved[19,20]. The resulting abasic (or alternatively apurinic or apyrimidinic) site (AP site) is a target for the endonuclease APEX1, alternatively known as APE1, which nicks the phosphodiester backbone at the 5'side of the lesion[21]. Once the DNA has been nicked on the damaged strand, the subsequent steps can be completed using long-patch or short-patch mechanisms. In long-patch repair (not shown in Figure 1-1), 2-12 of the original nucleotides are displaced by the activity of pol δ or pol ε as they extend from the 3'-terminal hydroxyl. The displaced flap is then trimmed by the flap endonuclease FEN1 and the nick sealed by ligase I (LIGI), recruited by proliferating cell nuclear antigen (PCNA)[21-23]. In short-patch BER, pol β trims the 5'-terminal deoxyribose phosphate (dRP) with its lyase activity and incorporates the correct (dC) nucleotide opposite the G. The nick at the 3' end of the newly added base is sealed by LIGIII/XRCC1[21,22].

Another mechanism of DNA repair that could potentially remove the U lesion is the mismatch repair (MMR) pathway, as depicted in the green panel of Figure 1-1. Unlike BER, the MMR pathway is not efficient at correcting mismatches in DNA that is chromatinized[21,24]. MMR is active primarily during DNA replication where it works to correct misincorporation of nucleotides that have resulted in mismatches[24]. The specific component of the MMR pathway in humans that might detect a U:G mismatch is the MSH2-MSH6 heterodimer, which can also detect small indels. The alternative heterodimer, MSH2-MSH3, detects larger indels and loops and is not likely to contribute to repair of U:G mismatches[21,24]. The MSH2-MSH6 dimer acts as a sliding clamp, moving along the DNA until it comes across a recognizable defect. Once it does, each subunit exchanges an ADP for an ATP and recruits MLH1-PMS2, a dimer with endonuclease activity, to nick the DNA at the site of damage[21,25]. Once the nick has been made, replication factor C (RFC) loads PCNA, which in turn activates EXOI (a 3'-5' exonuclease) to resect the damaged region[26]. Following removal of the mismatch and several bases surrounding the region, MLH1-PMS2 recruits POL δ to fill in the gap, which is then sealed by LIGI[22]. While it is entirely likely that MMR will correctly repair the targeted lesion, it is conceivable that the process contributes to mutation in the genome. In the event that there are deaminated bases on opposite strands of the DNA, correct repair of one uracil lesion (involving all of the above steps, including resection of bases on the target strand) may permanently alter the DNA sequence nearby if the second uracil is used as a template for DNA repair synthesis (as POL δ is not capable of recognizing uracil as a lesion)[27]. Additionally, a recent report has characterized an alternative form of MMR that utilizes mutation-prone translesion polymerases[28].

In the event that the uracil lesion, either by random chance or as a function of high levels of deamination having over-taxed the repair machinery, has escaped proper repair and is replicated over (base pairing as a T rather than a C), the resulting daughter cells will have inherited asymmetric genomes. One daughter will have a normal genome (from the original G-containing strand) and the other will have processed the lesion into a mutation as a C-to-T transition (Figure 1-1, brown panel).

If the standard BER pathway is unable to complete the repair process, for instance being interrupted after having successfully excised the uracil, leaving an abasic site, genome replication can still occur, though it is accomplished by way of the translesion synthesis polymerases (Yellow panel in Figure 1-1). This particular class of polymerases is able to successfully, though in an error-prone manner, synthesize DNA across from a non-instructional abasic site, inserting G (the correct nucleotide) or an A, C, or T, though A is the most commonly used nucleotide (this phenomenon is called the “A rule”)[29].

The final potential outcome of genomic DNA cytosine deamination is double-stranded DNA (dsDNA) breaks. This is illustrated in the purple panel of Figure 1-1, in which the deamination events have occurred on opposite strands of the DNA and in close proximity to one another. The lesions themselves are not break-inducing, but as mentioned in the previous paragraphs, the repair processes rely on the ability to nick the DNA to remove and repair the lesion. There is potential for the DNA double strand to break if a replication fork were to attempt to replicate across a nick. Alternatively, if the repair machinery were to operate on the proximal but opposite strands and nick the DNA within the same window of time, there is the potential for the double helix to break, necessitating repair by homologous recombination (HR) or the more mutation-prone non-homologous end joining (NHEJ) pathways. An additional point to be made about dsDNA breaks is that they result in long stretches of ssDNA, a suitable substrate for APOBEC, at the break-point. Nik-Zainal and colleagues looked specifically at these breakpoints in a small set of breast cancers and discovered an APOBEC-like mutation signature surrounding the breaks[30]. They termed these mutation clusters "kataegis" after the Greek for thunderstorm; this nomenclature makes sense when one examines the rainfall plots these researchers used to identify the mutation clusters. It is still unclear if the breaks are the result of deamination and/or if the deamination occurs after the break occurs.

APOBEC AND THE MUTATOR HYPOTHESIS

The APOBEC family members have the potential to generate the mutations in genomic DNA that contribute to cancer onset and progression, but only if several specific criteria are met. These criteria form the foundation of this thesis. The implicated APOBEC must:

1. be expressed in the cancer;
2. be catalytically active;
3. have access to genomic DNA; and
4. result in detectable changes to genomic DNA.

Point 1 is the most obvious of the criteria, though it bears mentioning that it is possible that transient expression may also play a physiological role, though this is a difficult hypothesis to test as it would require assessment of APOBEC levels in tissue at multiple time points as the cancer progressed. In addition, it is important to clarify that the expression must be in the cancerous cells themselves and not in stromal cells or other normal infiltrate that is assayed at the same time as the bulk of the primary tumor. Cell culture models are an excellent means of answering this particular point as they have been cultured clonally and only contain cancerous cells. It bears mentioning that this is a mutational phenomenon that, while an intrinsic, endogenous source of mutation, it differs from other common endogenous sources, such as defects in MMR in hereditary nonpolyposis colorectal cancer (HNPCC)[31,32], as its levels can potentially be modulated. This active source can be removed by simply blocking expression or inhibiting the deaminase activity. Conversely, under some circumstances, cells with high expression/high mutation rates may be selected for as they are the most likely to have diversified their genomes, allowing them to adapt to new cellular microenvironments or escape detrimental therapeutics. The same cannot be said for cancers, such as HNPCC, as it would be mechanistically difficult to re-establish functional MMR. In that case, the mutational driver is under the control of a binary switch – on or off – and it is weighted to remain in the off position, once there.

Point 2, that the enzyme must retain catalytically active, while critical, is also fortuitous. As indicated above, appropriate mAbs are not available for most of the APOBEC3 proteins, leaving any study that relies solely on transcript measurements subject to the criticism that mRNA levels do not necessitate that the protein is expressed, stable, or active. However, since the DNA deamination activity of the APOBEC family members can be assayed directly using DNA oligonucleotides that contain a single C as a substrate there is an alternative to more generic detection methods. Tagging the oligonucleotide with a fluorophore or a FRET pair facilitates quantification of activity[1]. This addresses two points simultaneously: 1) it demonstrates that the enzyme is not inactivated post-transcriptionally and 2) it provides an alternative to the use of mAbs for protein detection while still indicating the presence of protein.

Point 3 highlights the simple fact that without access to the genome, an APOBEC, no matter how well expressed or active, will not be able to produce resultant mutations in the DNA. There are several different ways that an enzyme might be exposed to genomic DNA. Subcellular localization, *i.e.* constitutive compartmentalization to the cytoplasm or nucleus) is often controlled via nuclear localization or export signals (NLS and NES, respectively) embedded within the proteins themselves. Alternate mechanisms such as post-translational modifications are another means of controlling subcellular localization. Several previous lab members have characterized the constitutive localization as well as the cytoplasm-nucleus shuttling of various APOBEC family members[1,33]. As pertains to this particular hypothesis, one would not expect an APOBEC that is confined to the cytoplasm to generate mutation in the genomic DNA unless there was a mechanism by which it could enter the nucleus (*e.g.* a phosphorylation event that changes the subcellular localization). An additional means of accessing the genome might be during cell division when the nuclear membrane breaks down to allow for chromosome segregation[1,33]. A complication with this particular hypothesis is that one would have to explain how

APOBEC is able to function on the condensed chromosomes. This criterion is particularly relevant, as a recent report has put forth the idea that perhaps APOBEC3A might be causing mutations in breast and other cancers[34]. While this is a logical stance to take with the data that the researchers had at the time, additional work looking at the sub-cellular localization of endogenous APOBEC3A using newly developed, specific monoclonal antibodies have made that hypothesis unlikely[35]. Ectopic forced overexpression of APOBEC3A results in cell-wide expression. However, when expressed from the endogenous locus, the enzyme remains exclusively cytoplasmic[1,33,35].

In order for the APOBEC-driven mutator hypothesis to remain tenable, APOBEC presence, activity, and genome access must correlate with a detectable change to the cellular genome. If there is no change, the finding may still be intellectually useful, but not likely clinically relevant. The reader will no doubt note that the phrase "detectable changes" to describe the critical output indicated in point 4. This definition includes lesions as well as fixed mutations and large-scale chromosomal aberrations (as discussed and demonstrated with Figure 1-1). There are myriad ways to assay these changes. The most direct is to look specifically for an increase in genomic uracil. The hypothesis is that elevated levels of the suspected APOBEC3 will correlate with increased steady-state levels of genomic uracil. Uracil measurement should be performed using isogenic controls. This is due to the transient nature of the lesion. Genomic uracil is subject to the repair processes outlined above; meaning that variation in the activity of any of the different pathway components has the potential to confound any experiment looking at just APOBEC3 levels (*i.e.* there are numerous uncontrolled variables if the system is not isogenic).

We developed a technique to facilitate accurate quantitation of genomic uracil levels. This protocol relies upon the efficiency of recombinant UDG to remove uracil from bulk DNA and the quantitative power of HPLC-MS/MS to separate the excised uracil from background contaminants and detect it.

Observation of genetic changes (*i.e.* mutations) as a direct result of APOBEC3 activity is complicated by the random nature of the mutagenesis. There is currently no way to accurately predict which genes or specific genomic regions might be mutated. Without a specific target, simple direct sequencing is likely to fail to detect any differences. This challenge can be bypassed by using selection and/or mutation enrichment techniques. For instance, 3D-PCR (differential DNA denaturation-PCR) is particularly useful in this regard as it enriches pools of DNA for sequences that have decreased C:G content[36,37]. This is accomplished by performing PCR targeting a genomic region while varying the temperature of the denaturation step. The principle here is that DNA templates that have low C:G content will denature at lower temperature, and thus amplify successfully, whereas high C:G content templates will not. This effectively and selectively amplifies mutated sequences that then can be cloned into bacterial vectors and directly sequenced to assess mutation load.

Other techniques have used an artificially incorporated genetic marker into the genome of cells either expressing APOBEC3B or its catalytic mutant. Shinohara and colleagues used a lentivirus to stably integrate an *EGFP* gene into the genome of HEK293 cells and then over-expressed APOBEC3B using a catalytically dead mutant

APOBEC3B as a control[38]. Rather than sequencing the genome of these cells, instead, they used an array to capture millions of copies of the integrated *EGFP* sequence and selectively sequence them. They found that exogenously introduced, over-expressed APOBEC3A, APOBEC3B, and AID are able to mutate the cellular genome. This was a clever way to avoid the problem of random mutation being lost among the background when sequencing the entire genome.

A prediction of the APOBEC mutator hypothesis is that tumors that are affected by this process will bear the marks of the mutator. Again, this is a tricky phenomenon to assess directly using model systems for the reasons indicated above, but it is assumed that for long established tumors from patients, that the correlation would be clear between mutator expression, activity, and mutation signature. Several different research groups have emphasized that there is, in fact, an "APOBEC signature" in the genomes of several different types of human cancer[30,34,39,40]. None of these researchers have definitively identified a single APOBEC that they believe might actually be the mutational agent; rather they refer to the entire APOBEC3 subfamily, though based on mutation signature they do exclude APOBEC3G and AID as potential candidates This work focuses primarily on breast cancer, as it is a prevalent malignancy with many of the hallmarks of a potential mutator cancer, but it also looks at the role of APOBEC mutagenesis in several other forms of cancer as well.

FIGURES & TABLES

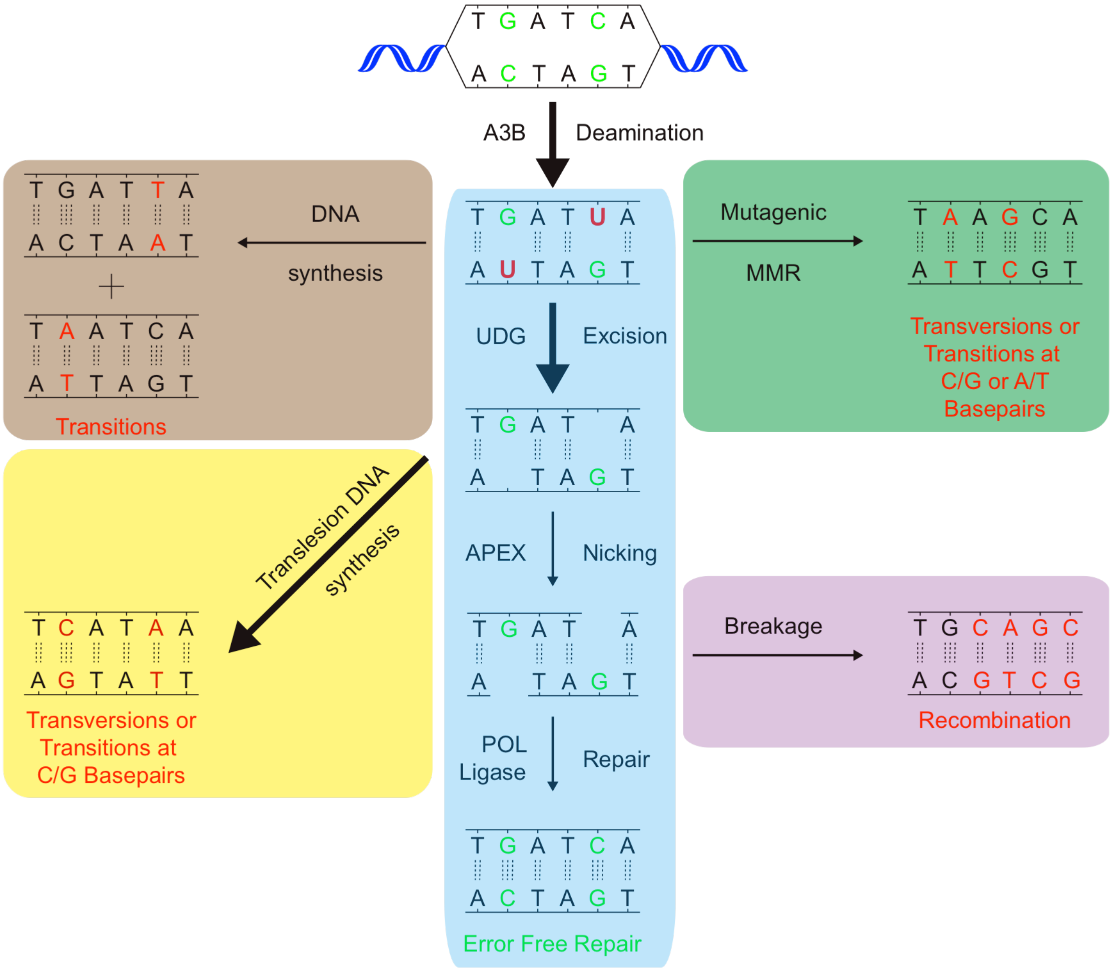


Figure 1-1. Schematic outlining the potential outcomes of genome deamination.

CHAPTER 2 – APOBEC3B IS AN ENZYMATIC SOURCE OF MUTATION IN BREAST CANCER

This chapter is adapted, with permission, from: Burns & Lackey, *et al.*, (2013) *Nature* vol. 494, pp. 366-371.

Authors: Michael B. Burns*, Lela Lackey*, Michael A. Carpenter, Anurag Rathore, Allison M. Land, Brandon Leonard, Eric W. Refsland, Delshanee Kotandeniya, Natalia Tretyakova, Jason B. Nikas, Douglas Yee, Nuri A. Temiz, Duncan E. Donohue, Rebecca M. McDougle, William L. Brown, Emily K. Law, and Reuben S. Harris

**these authors contributed equally to the completion of this work*

INTRODUCTION

Multiple 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[30,41-48]. Most of these mutations occur at hydrolytically disfavoured[11] non-methylated cytosines throughout the genome, and are sometimes clustered[30]. 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.

RESULTS & DISCUSSION

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[13,49]. 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[9] and expression studies[2] (Fig. 2-5).

To identify the contributing APOBEC3 protein, we quantified mRNA levels for each of the 11 family members in breast cancer cell lines (Fig. 2-6). 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 more than 3 s.d. relative to controls in 28 out of 38 lines, with levels exceeding tenfold in 12 out of 38 lines (Fig. 2-1A and Table 2-1). 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 (Fig. 2-7, Tables 2-2 - 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 (Fig. 2-8).

Epitope-tagged APOBEC3B (A3B) localizes to the nucleus of several transfected cell types[33]. 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, 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. 2-1B and Fig. 2-9). Corroborating data were obtained for haemagglutinin (HA)-tagged proteins (Fig. 2-9). 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. 2-1C and Figs. 2-10 and 2-11). 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[50], endogenous A3B showed a strong preference for editing cytosines in the TC dinucleotide context (Fig. 2-1D and Fig. 2-10). 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 (Fig. 2-12). Both A3B–HA and A3A–HA could elicit measurable TC-to-TU activity in lysates from transfected HEK293T cells (Fig. 2-13). However, because *APOBEC3A* mRNA is myeloid lineage-specific[36] and non-detectable in breast cancer cell lines (Figs. 2-5 and 2-

6), 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. 2-2A and Fig. 2-14). Genomic uracil loads decreased by 30% in HCC1569 cells expressing shA3B and by 70% in MDA-MB-453 cells, where knockdown was stronger (Fig. 2-2B and Fig. 2-14). 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[36] 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 *TK* gene of herpes simplex virus type, 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. 2-2C and Fig. 2-14).

Third, differential DNA denaturation PCR (3D-PCR)[36,37] 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. 2-2D and Fig. 2-14). These amplicons were cloned and sequenced, and more C-to-T transition mutations were observed in A3B^{high} compared with A3B^{low} samples (Fig. 2-2D and Fig. 2-14). *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[4], 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 (Fig. 2-15). A3A overexpression caused rapid S-phase arrest, cytotoxicity and g-H2AX focus formation, as reported previously[51] (Fig. 2-3A-C and Fig. 2-15). 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. 2-3A, B and Fig. 2-15). A3B induction also caused g-H2AX focus formation, DNA fragmentation, as evidenced by visible comets, and C-to-T mutations (Fig. 2-3C-E). A3B catalytic activity, as evidenced by the glutamate mutants, was required for the induction of these cancer phenotypes.

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 ($P = 0.0003$) (Fig. 2-16). 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. 2-4A; $P = 7.1 \times 10^{-7}$ and $P = 2 \times 10^{-5}$, respectively; patient information in Table 2-10). 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 mean levels in mammoplasty samples (Fig. 2-4A; $P = 0.002$). 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[2]($P = 0.591$). Similar results were obtained by quantifying RNA-sequencing data for independent matched tumour and normal pairs[52], 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. 2-4B).

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 (Fig. 2-17). 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[6], liver[53] and breast[30,47,52] tumours. Consistent with a spontaneous origin, the C-to-T frequency is low in liver tumours (~20%) and mutational events appear random (Fig. 2-4C, 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. 2-4C, 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. 2-4C, D and Fig. 2-17). The availability of both high-throughput RNA sequencing (RNAseq) and somatic mutation data[52] 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. 2-4E-G). Importantly, tumours expressing high A3B levels have twice as many mutations (Fig. 2-4E and F and Fig. 2-18). 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 Fig. 2-19 with similarities to AID-dependent antibody diversification mechanism[3]). 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. 2-4G and Fig. 2-20). *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[54], 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[55]. A3B inhibition may decrease the rate of tumour evolution and stabilize the targets of existing therapeutics.

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 2-2 through 2-9. For instance, the commonly used Affymetrix Genechip Human Genome Array U133A has 11 probes intended for *A3B* detection (Table 2-2 & 2-4). 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 2-2 though 2-9).

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[2,36,56,57]) (Figs. 2-5 & 2-6). 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. 2-7A; 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. 2-4A). 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 2-2). 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 2-2). In comparison, probe sets with low identity to *A3B* showed no significant correlation (*e.g.*, *A3C*; Table 2-2). As shown in Fig. 2-7B, 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[9].

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[58], http://dbCRID.biolead.org), which might have been found by classical cytogenetic techniques[59] or, more recently, by deep sequencing[30,47]. 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. 2-4B). 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.

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[2] (Table 2-11). Knockdown and control shRNA constructs were obtained from Open Biosystems. Microscopy, cellular fractionation and deaminase activity assays were done as described[33,36]. 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 line[36]. 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 g-H2AX. Recombinant A3B195-382-mycHis was purified and used for deamination kinetics as described[60] 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[6,30,47,52,53]. Potential mechanistic overlap with hydrolytic deamination of 5-methyl-cytosines was avoided by excluding CpG dinucleotides from mutational preference calculations.

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[2]. Tissue RNA was from 100 mg flash-frozen tissue disrupted by a 2-h water bath sonication in 1 ml of Qiazol Lysis Reagent (RNeasy, Qiagen). Cell RNA was made using Qiashredder (RNeasy, Qiagen). qPCR was performed on a Roche Lightcycler 480 instrument. The housekeeping gene *TBP* was used for normalization. Statistical analyses for matched tissues were done using the Wilcoxson signed-rank test, and unmatched sets with the Mann–Whitney U-test (Graphpad Prism). Primer and probe sequences are listed in Table 2-11.

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 pD-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 mg ml^{-1}).

Cell fractionation and DNA deaminase activity assays

Subcellular activity analysis and dinucleotide preferences were measured as described[60,61]. Briefly, cellular fractionation was performed by syringe treatment of 10^7 cells in 0.5 ml of hypotonic buffer[61]. 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[36]. Dilutions were incubated 2 h at 37 °C with a DNA oligonucleotide 5′-(6-FAM)-AAA-TT$\underline{\text{C}}$-TAA-TAG-ATA-ATG-TGA-(TAMRA). Fluorescence was measured on SynergyMx plate reader (BioTek). Local dinucleotide preferences in extracts were analysed similarly using 5′-AC, CC, GC or TC at the NN position of 5′-(6-FAM)-ATA-A$\underline{\text{NN}}$-AAA-TAG-ATA-AT-(TAMRA).

Genomic uracil quantifications

Genomic DNA was prepared from shA3B-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 mm, Thermo Scientific) was maintained at 40 °C and a flow rate of 15 ml 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^+] - 70.08 [M-CONH]$^+$ and *m*/*z* 96.08 [$M\text{-}NH_2$]$^+$ for uracil, whereas the internal standard ([^{15}N-2, ^{13}C-4]-uracil) was monitored by the transitions *m*/*z* 119.08 [M^+H^+] - *m*/*z* 74.08 [M-CONH]$^+$ and *m*/*z* 101.08 [$M\text{-}NH_2$]$^+$, respectively. Internal standards were used for quantification.

TK fluctuations

TK-neo was introduced into MDA-MB-453 and HCC1569 cells as described[36]. 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[62].

3D-PCR and sequencing

DNA was collected from Ugi-expressing[63] T-REx-293 clones or HCC1569 cells transduced with shA3B or shControl lentiviruses. 3D-PCR was done using Taq (Denville Scientific) as described[36]. 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 mg 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 mg ml^{-1} propidium iodide and 40 mg 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 mg 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 g-H2AX foci was adapted[64]. Fixed cells were incubated overnight in 0.2% Triton X-100, 1% BSA in PBS (blocking buffer) with 1:100 rabbit anti-g-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 mg ml^{-1} of Dox before fixation with 4% paraformaldehyde and incubation with 1:50 anti-g-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[65], 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[6,30,47,52,53]. 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.

Microarray comparisons

Affymetrix GeneChip microarray data were reported previously by others. Tripathi *et al.*[66] (GEO ID GSE9574) and Graham *et al.*[67] (GEO ID GSE20437) reported data for 15 and 7 reduction mammoplasty samples, respectively. Tabchy *et al.*[68] (GEO ID GSE20271) reported data for 178 stage I-III breast cancers (procured at 6 sites worldwide), and Lasham *et al.*[69] (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:

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

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. 2-7B). Having performed the same normalization method to all *APOBEC3* genes, we were able to obtain expression data for the tumor versus normal samples (Table 2-2). As previously described[70-74], 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 2-2, the probe sets corresponding to *A3A*, *A3B*, and *A3(F,G)* are deemed to have significant differential expression according to all three methods.

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., D.K. and N.T. assisted with the HPLC-ESI+MS/MS set up. 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. J. Hultquist and R. Vogel provided statistical assistance, T. Hwang bioinformatic assistance, V. Polunovsky hTERT-HMECs, V. Simon shRNA, S. Kaufmann, C. Lange and D. Largaespada consultation, and the Masonic Cancer Center Breast Cancer Research Fund 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. Operational support in the Harris laboratory was provided in part by the Children's Cancer Research Fund and a seed grant from the University of Minnesota Clinical and Translational Science Institute (supported by NIH 1UL1RR033183).

FIGURES & TABLES

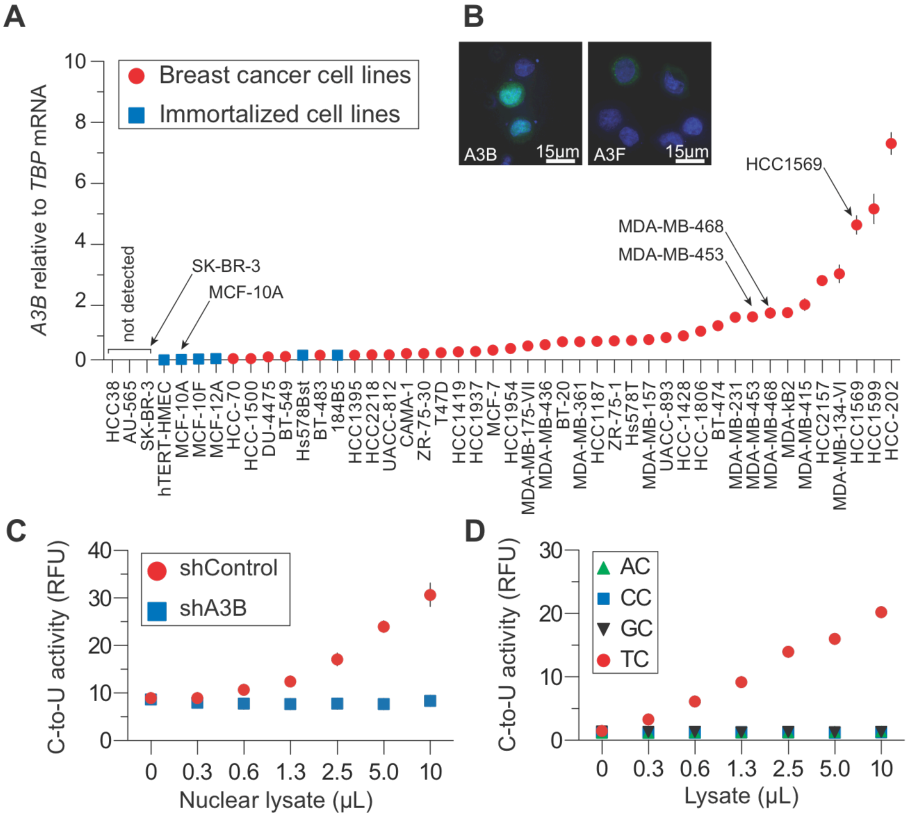


Figure 2-1. ***APOBEC3B*** **upregulation and activity in breast cancer cell lines.**

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).

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).

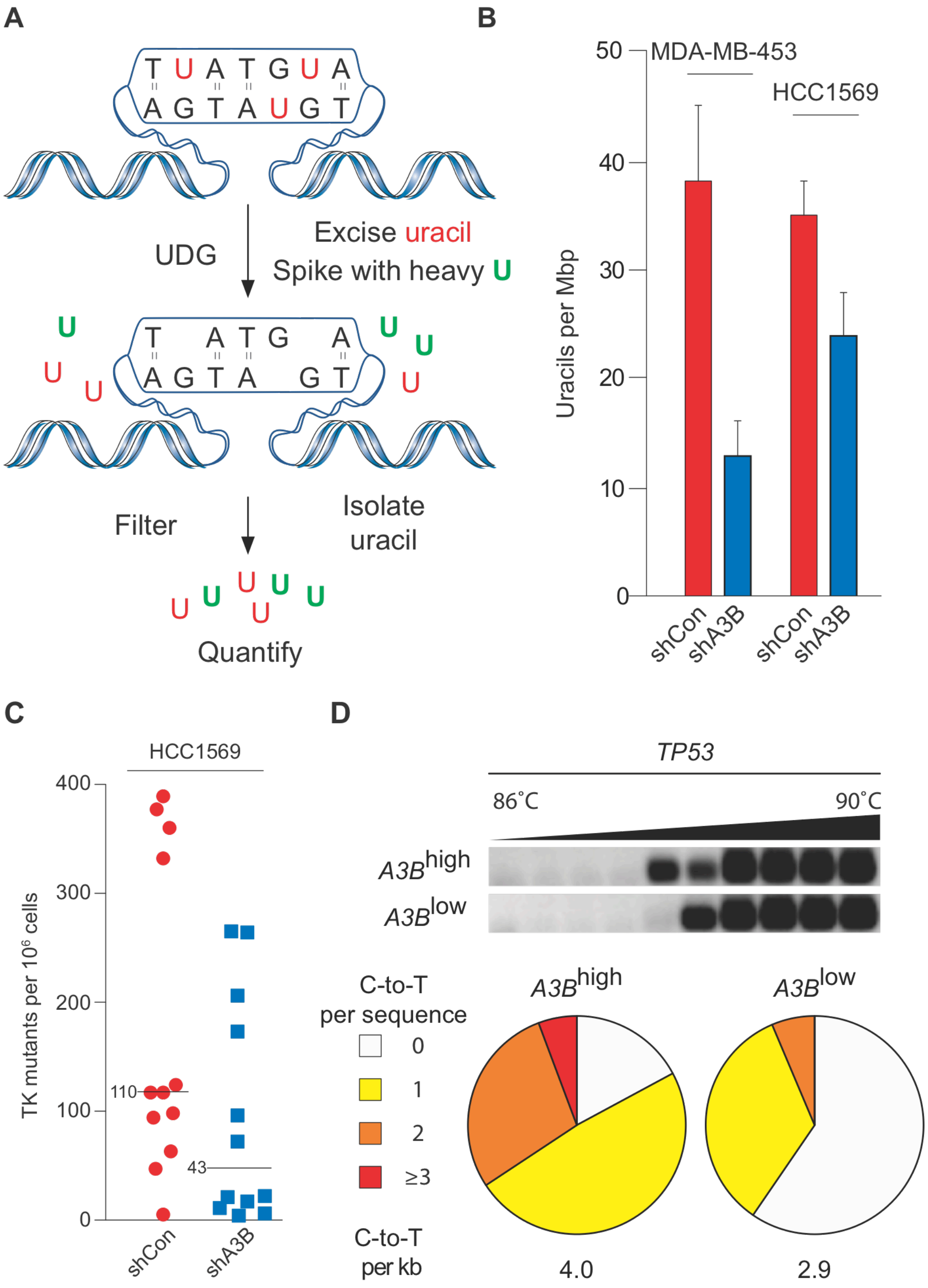

A
T U A T G U A
A G T A U G T
UDG
Excise uracil
Spike with heavy U
T A T G A
A G T A G T
U U U U U U
Filter
Isolate uracil
U U U U U U
Quantify
B
Uracils per Mbp
0
10
20
30
40
50
MDA-MB-453
HCC1569
shCon
shA3B
shCon
shA3B
C
HCC1569
TK mutants per 10⁶ cells
0
100
200
300
400
110
43
shCon
shA3B
D
TP53
86˚C
90˚C
A3Bhigh
A3Blow
C-to-T per sequence
0
1
2
≥3
A3Bhigh
A3Blow
C-to-T per kb
4.0
2.9

Figure 2-2. APOBEC3B-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 plots 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 or shA3B (*n* ≥35 sequences per condition). See Fig. 2-14 for further data.

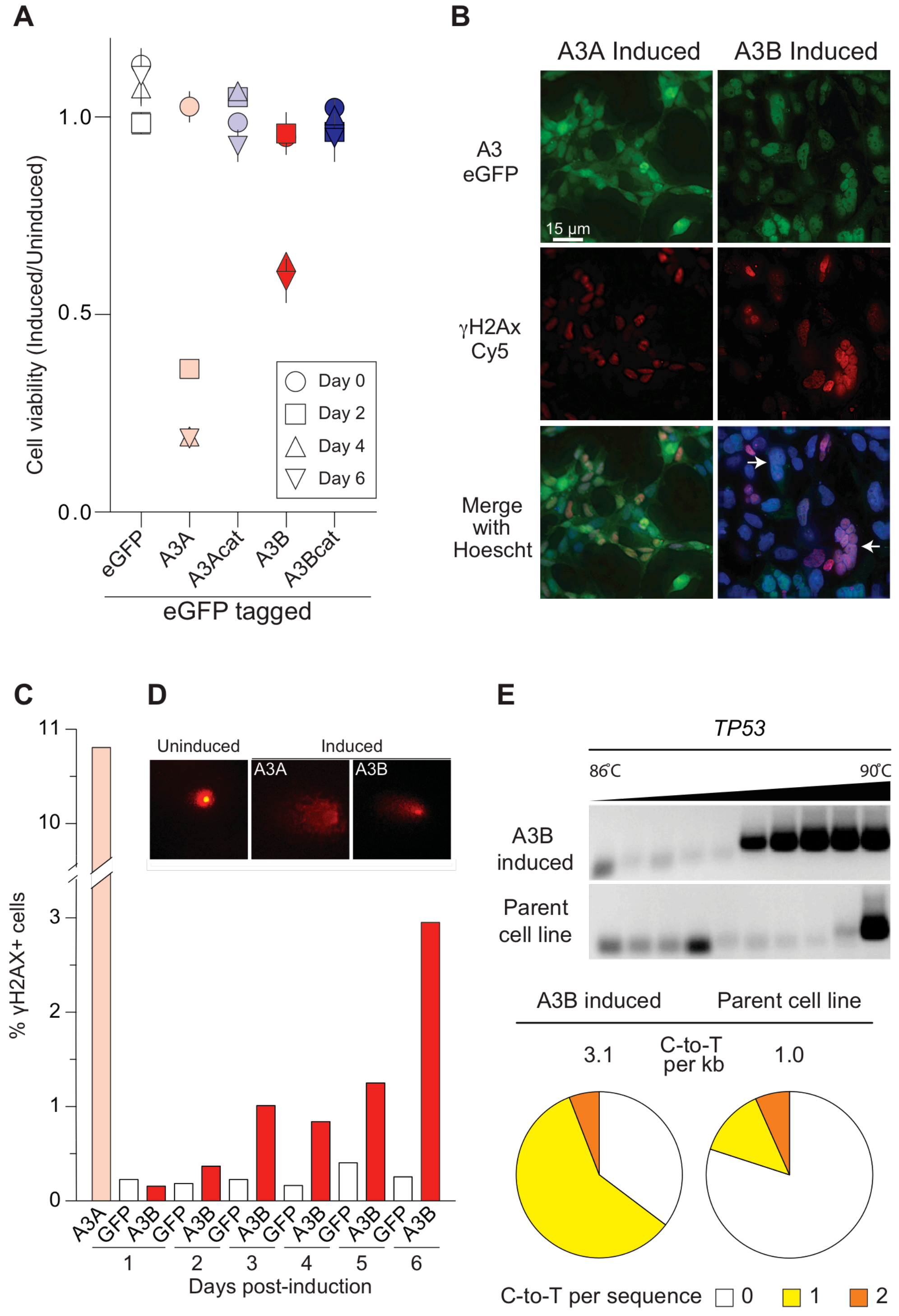

A
Cell viability (Induced/Uninduced)
1.0
0.5
0.0
eGFP
A3A
A3Acat
A3B
A3Bcat
eGFP tagged
○ Day 0
□ Day 2
△ Day 4
▽ Day 6
B
A3A Induced
A3B Induced
A3
eGFP
15 µm
γH2Ax
Cy5
Merge
with
Hoescht
C
% γH2AX+ cells
11
10
3
2
1
0
A3A
GFP
A3B
GFP
A3B
GFP
A3B
GFP
A3B
GFP
A3B
GFP
A3B
1
2
3
4
5
6
Days post-induction
D
Uninduced
Induced
A3A
A3B
E
TP53
86°C
90°C
A3B
induced
Parent
cell line
A3B induced
Parent cell line
3.1
C-to-T
per kb
1.0
C-to-T per sequence
□ 0
🟨 1
🟧 2

Figure 2-3. Cancer phenotypes triggered by inducing APOBEC3B overexpression.

A, Cell viability at indicated times after induction (mean and s.d. for *n* = 3 per condition).

B, **C**, Representative fields of cells imaged for g-H2AX and A3A–eGFP (1 day) or A3B–eGFP (3 days) after induction, and g-H2AX quantification. Abnormal, multinuclear clusters are typical of induced A3B–eGFP (white arrows).

D, Representative images of A3-induced DNA comets (400x magnification).

E, C-to-T mutations in *TP53* detected by sequencing 3D-PCR products 4 days after induction (*n* > 12 sequences per condition).

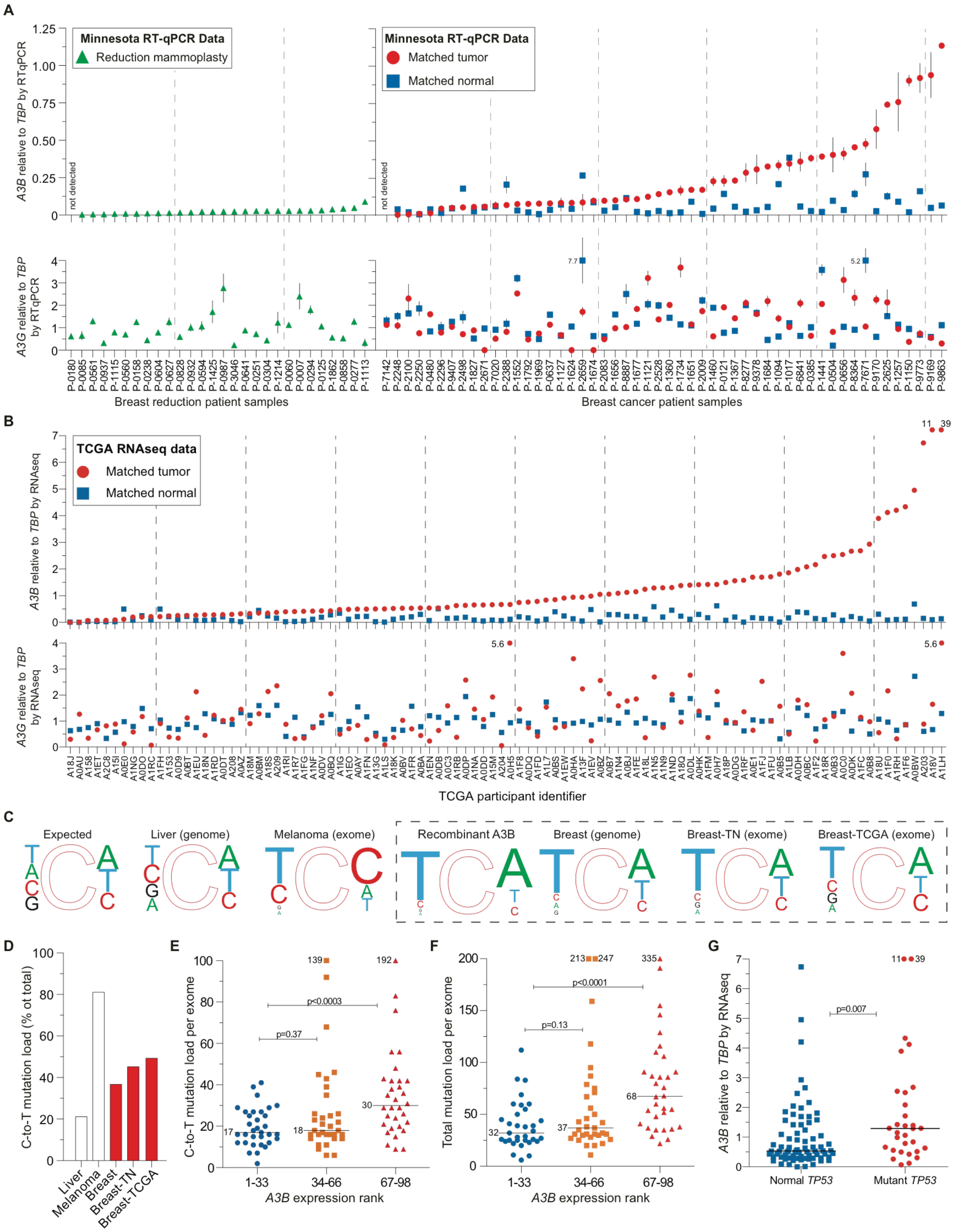

A
Minnesota RT-qPCR Data
Reduction mammoplasty
Minnesota RT-qPCR Data
Matched tumor
Matched normal
A3B relative to TBP by RTqPCR
1.25
1.00
0.75
0.50
0.25
0
not detected
not detected
A3G relative to TBP by RTqPCR
4
3
2
1
0
7.7
5.2
P-0180 P-0085 P-0561 P-0937 P-1115 P-0560 P-0158 P-0238 P-0604 P-0627 P-0828 P-0932 P-0594 P-1425 P-0987 P-3046 P-0641 P-0251 P-0304 P-1214 P-0060 P-0007 P-0294 P-0125 P-1862 P-0858 P-0277 P-1113
Breast reduction patient samples
P-7142 P-2248 P-2100 P-2250 P-0480 P-2296 P-9407 P-2498 P-1827 P-2671 P-7020 P-2388 P-1552 P-1792 P-1969 P-0637 P-1127 P-1624 P-2659 P-1674 P-2083 P-1656 P-8887 P-1677 P-1121 P-2528 P-1360 P-1734 P-1651 P-2009 P-1460 P-0121 P-1367 P-8277 P-9378 P-1684 P-1094 P-1017 P-6841 P-0385 P-1441 P-0504 P-0656 P-8364 P-7671 P-9170 P-2625 P-1257 P-1150 P-9773 P-9169 P-9863
Breast cancer patient samples
B
TCGA RNAseq data
Matched tumor
Matched normal
A3B relative to TBP by RNAseq
7
6
5
4
3
2
1
0
11
39
A3G relative to TBP by RNAseq
4
3
2
1
0
5.6
5.6
A18J A0AU A158 A1ET A2C8 A15I A0E0 A1NG A0DO A1RC A1FH A153 A0D9 A0BT A1EU A18N A1RD A0DT A208 A0AZ A18M A0BM A18S A209 A1RI A1R7 A1FG A1NF A0DV A0BQ A1IG A1EO A0AY A1FN A13G A1LS A18K A0BV A1FR A0BA A1EN A0DB A0C3 A1RB A0DP A1NA A0DD A15M A204 A0H5 A1F8 A0DQ A1FD A1L7 A0BS A1EW A0HA A13F A1EV A0BZ A0B7 A1N4 A0BJ A1FE A18L A1N5 A1N9 A1ND A18Q A0DL A0HK A1FM A0H7 A18P A0DG A1RF A0E1 A1FJ A1FU A0B5 A1LB A0DH A0BC A1F2 A18R A0B3 A0C0 A0DK A1FC A0B8 A18U A1F0 A1RH A1F6 A0BW A203 A18V A1LH
TCGA participant identifier
C
Expected
T A C G C A T C
Liver (genome)
T C G A C A T C
Melanoma (exome)
T C G A C C A T
Recombinant A3B
T C A C A T C
Breast (genome)
T C A G C A T C
Breast-TN (exome)
T C G A C A T C
Breast-TCGA (exome)
T C G A C A T C
D
C-to-T mutation load (% ot total)
100
80
60
40
20
0
Liver
Melanoma
Breast
Breast-TN
Breast-TCGA
E
C-to-T mutation load per exome
100
80
60
40
20
0
139
192
p<0.0003
p=0.37
17
18
30
1-33
34-66
67-98
A3B expression rank
F
Total mutation load per exome
200
150
100
50
0
213
247
335
p<0.0001
p=0.13
32
37
68
1-33
34-66
67-98
A3B expression rank
G
A3B relative to TBP by RNAseq
7
6
5
4
3
2
1
0
11
39
p=0.007
Normal TP53
Mutant TP53

Figure 2-4. ***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; off-scale values are indicated numerically).

B, RNA-seq data for *APOBEC3B* and *APOBEC3G* in the indicated samples (off-scale values are indicated numerically). TCGA, The Cancer Genome Atlas.

C, Local sequence contexts for all genomic cytosines (expected), cytosines deaminated by recombinant A3B (Fig. 2-17), and observed C-to-T transitions in the indicated cancers. Font size is proportional to each nucleotide frequency. TN, Triple-Negative.

D, The percentage of C-to-T mutations in the indicated tumours.

E, **F**, C-to-T and total mutation counts for tumours in **B** grouped into lower, middle and upper thirds based on *APOBEC3B* levels (medians are labelled; *P* values from Mann–Whitney U test; off-scale values are indicated numerically).

G, Relationship between *APOBEC3B* level (RNAseq data) and *TP53* status for tumours in **B** (*P* values from Mann–Whitney U test; off-scale values are indicated numerically).

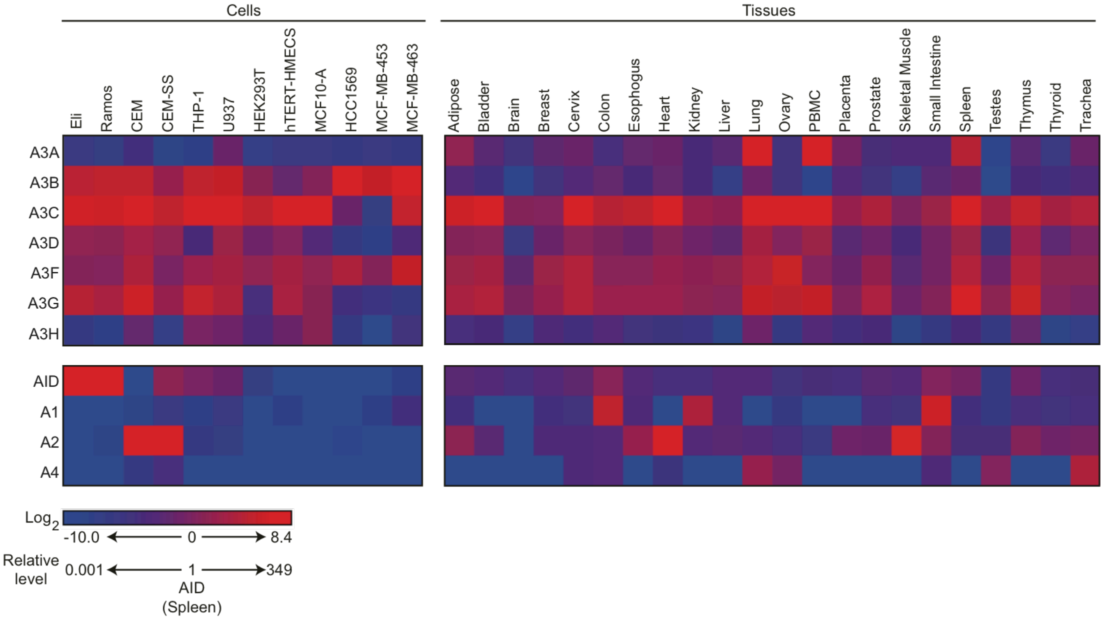


Figure 2-5. Expression profiles for *APOBEC* family members in human cell lines and tissues.

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 [2]. 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.

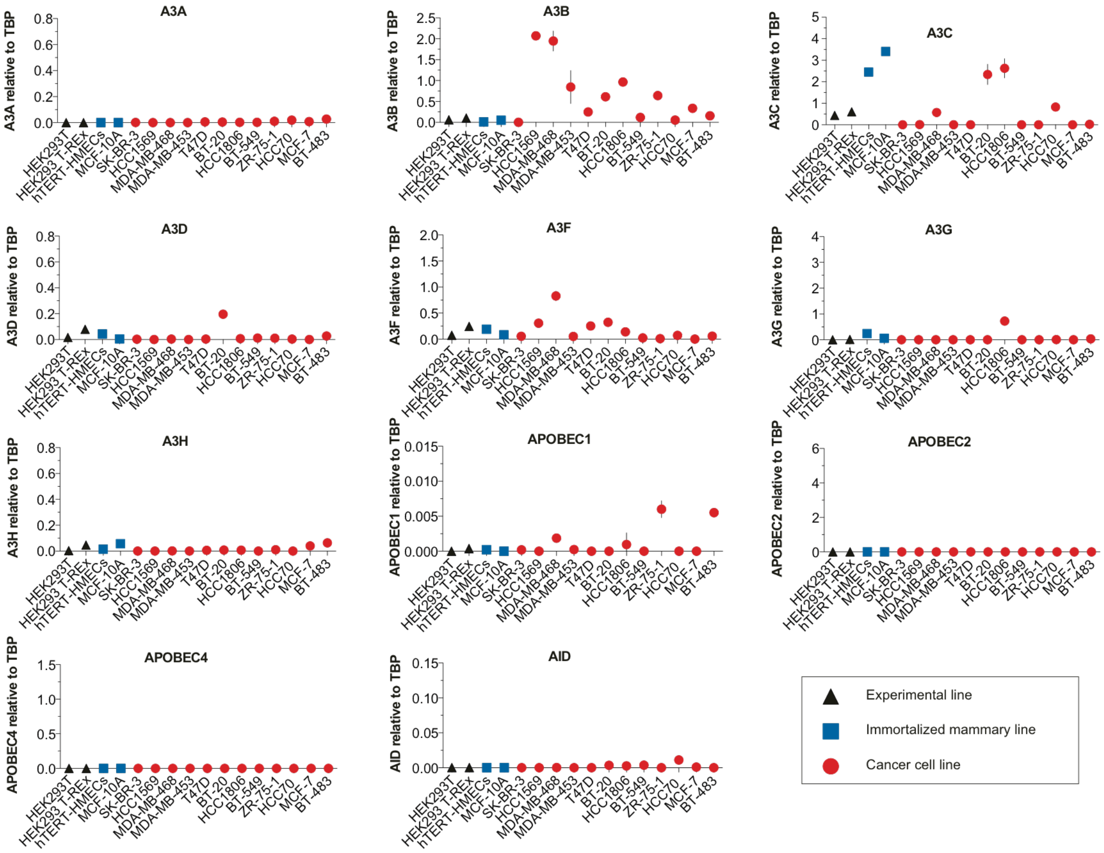


Figure 2-6. Full expression profiles for *APOBEC* family members in a panel of representative cell lines.

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. 2-1A in the context of the full panel of normal and breast cancer cell lines.

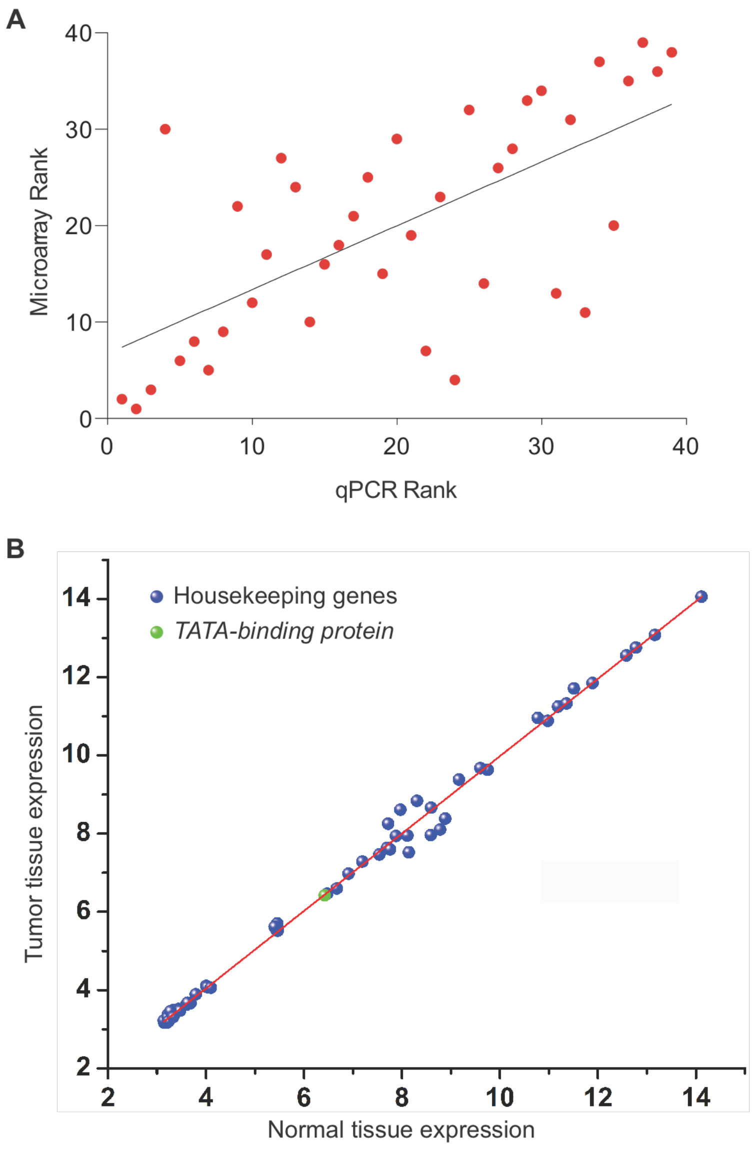
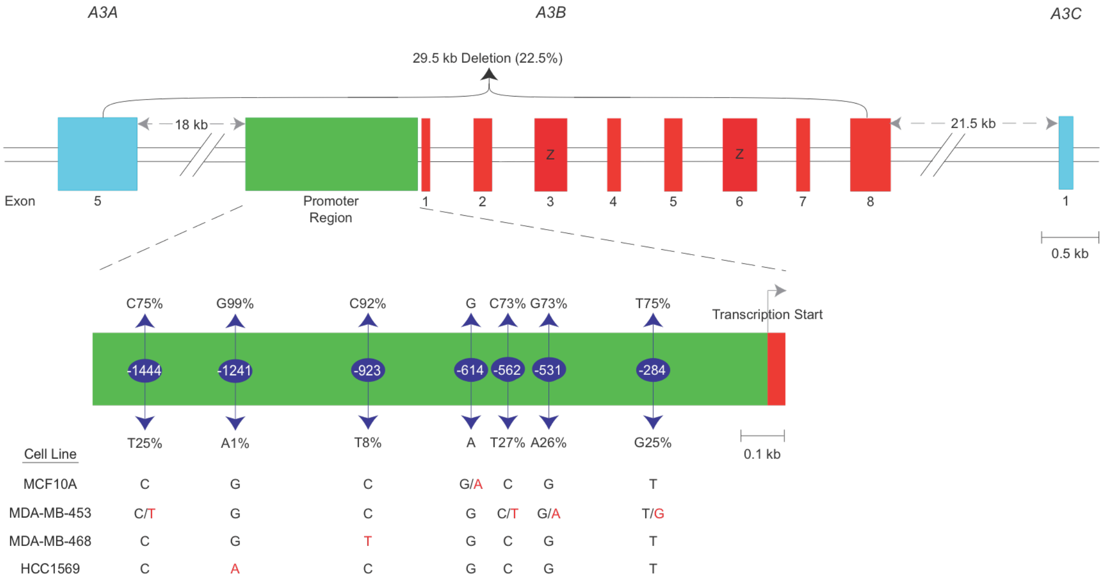


Figure 2-7. Microarray information.

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).



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

Figure 2-8. *APOBEC3B* 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.

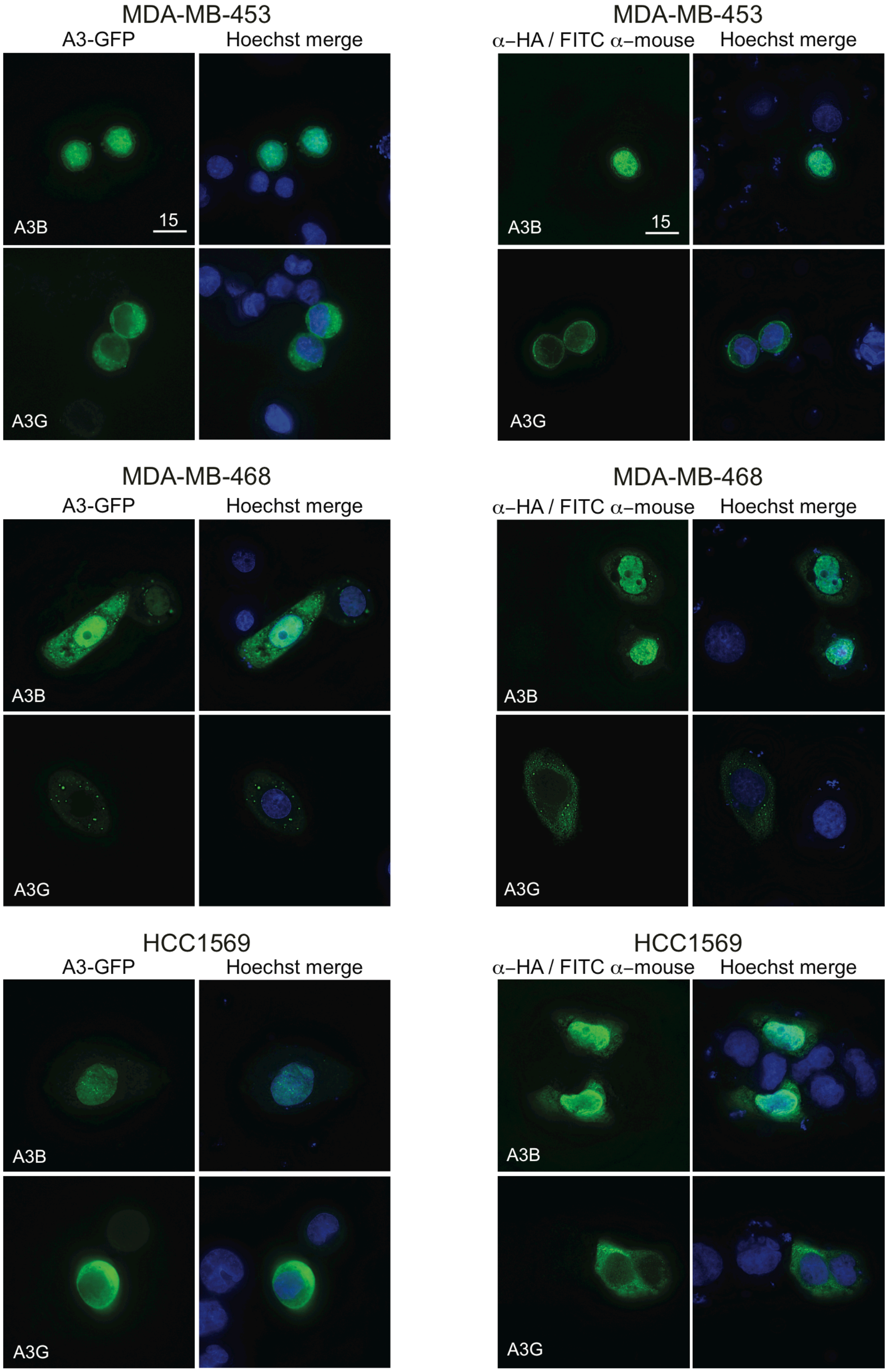

MDA-MB-453
A3-GFP
Hoechst merge
A3B
15
A3G
MDA-MB-453
α–HA / FITC α–mouse
Hoechst merge
A3B
15
A3G
MDA-MB-468
A3-GFP
Hoechst merge
A3B
A3G
MDA-MB-468
α–HA / FITC α–mouse
Hoechst merge
A3B
A3G
HCC1569
A3-GFP
Hoechst merge
A3B
A3G
HCC1569
α–HA / FITC α–mouse
Hoechst merge
A3B
A3G

Figure 2-9. 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.

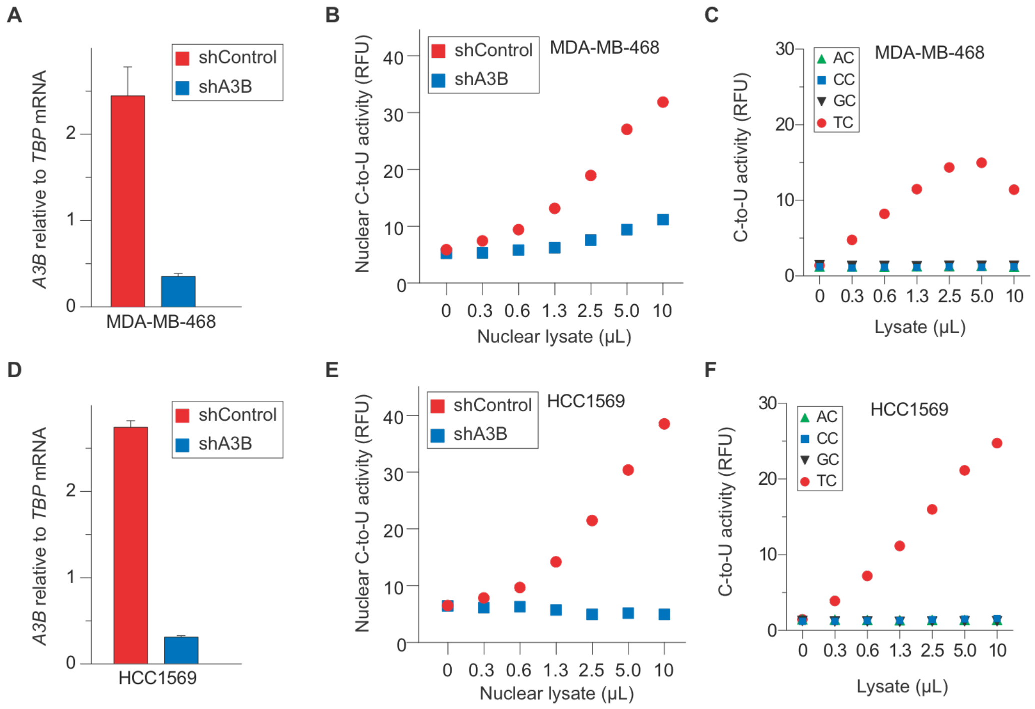


Figure 2-10. DNA cytosine deaminase activity of endogenous APOBEC3B in breast cancer cell line nuclear extracts.

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).

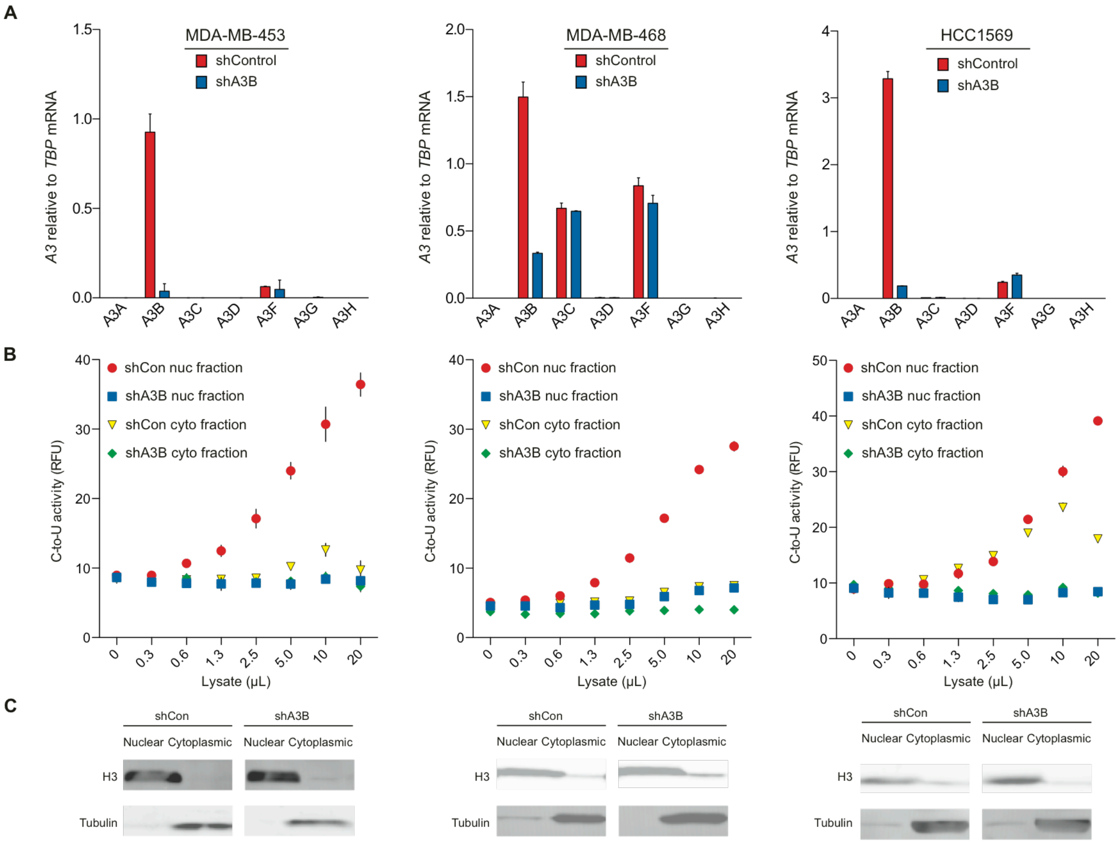

A
MDA-MB-453
shControl
shA3B
A3 relative to TBP mRNA
1.5
1.0
0.5
0.0
A3A
A3B
A3C
A3D
A3F
A3G
A3H
MDA-MB-468
shControl
shA3B
2.0
1.5
1.0
0.5
0.0
HCC1569
shControl
shA3B
4
3
2
1
0
B
shCon nuc fraction
shA3B nuc fraction
shCon cyto fraction
shA3B cyto fraction
C-to-U activity (RFU)
40
30
20
10
0
50
Lysate (μL)
0
0.3
0.6
1.3
2.5
5.0
10
20
C
shCon
shA3B
Nuclear
Cytoplasmic
H3
Tubulin

Figure 2-11. APOBEC3B 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.

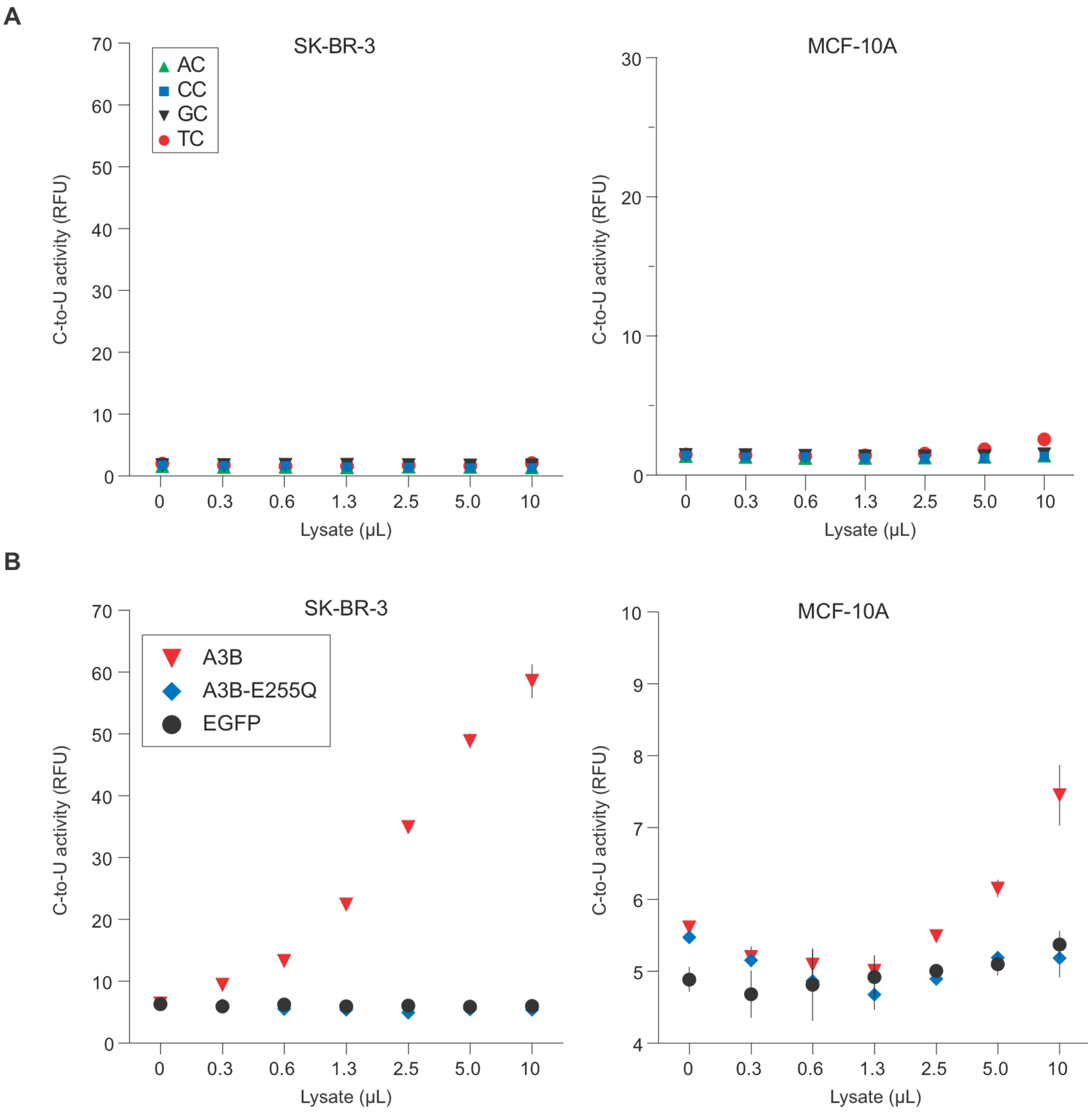

A
SK-BR-3
MCF-10A
▲ AC
■ CC
▼ GC
● TC
C-to-U activity (RFU)
70
60
50
40
30
20
10
0
30
20
10
0
0
0.3
0.6
1.3
2.5
5.0
10
Lysate (μL)
B
SK-BR-3
MCF-10A
▼ A3B
◆ A3B-E255Q
● EGFP
C-to-U activity (RFU)
70
60
50
40
30
20
10
0
10
9
8
7
6
5
4
0
0.3
0.6
1.3
2.5
5.0
10
Lysate (μL)

Figure 2-12. DNA deaminase activity in APOBEC3B-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).

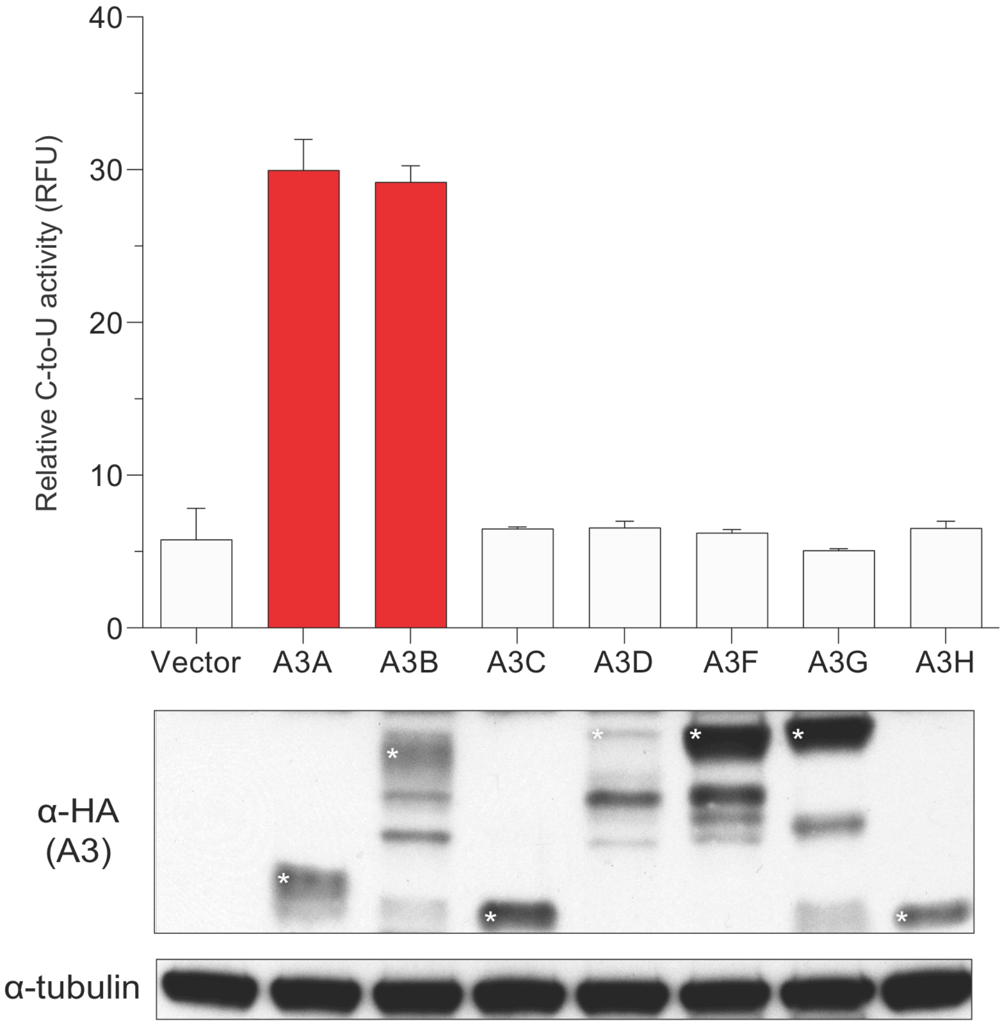


Figure 2-13. Deaminase activity of HEK293T cell extracts with individual over-expressed APOBEC3 proteins.

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.

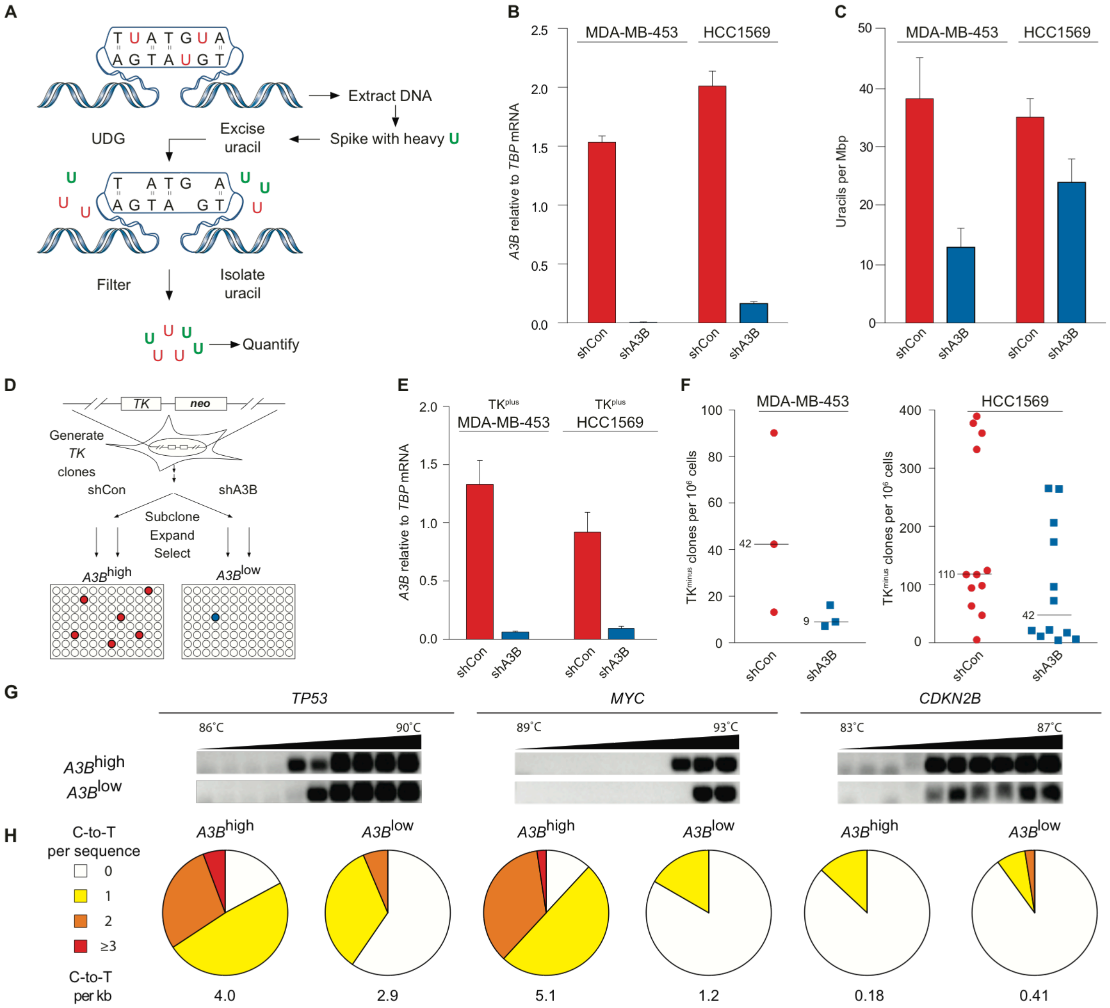

A
T U A T G U A
A G T A U G T
Extract DNA
Spike with heavy U
UDG
Excise uracil
T A T G A
A G T A G T
U U U U U U
Filter
Isolate uracil
U U U U U U
Quantify
B
MDA-MB-453
HCC1569
A3B relative to TBP mRNA
2.5
2.0
1.5
1.0
0.5
0.0
shCon
shA3B
shCon
shA3B
C
MDA-MB-453
HCC1569
Uracils per Mbp
50
40
30
20
10
0
shCon
shA3B
shCon
shA3B
D
TK
neo
Generate TK clones
shCon
shA3B
Subclone
Expand
Select
A3Bhigh
A3Blow
E
TKplus MDA-MB-453
TKplus HCC1569
A3B relative to TBP mRNA
2.0
1.5
1.0
0.5
0.0
shCon
shA3B
shCon
shA3B
F
MDA-MB-453
HCC1569
TKminus clones per 10⁶ cells
100
80
60
40
20
0
42
9
400
300
200
100
0
110
42
shCon
shA3B
shCon
shA3B
G
TP53
MYC
CDKN2B
86˚C
90˚C
89˚C
93˚C
83˚C
87˚C
A3Bhigh
A3Blow
H
C-to-T per sequence
0
1
2
≥3
C-to-T per kb
A3Bhigh
A3Blow
A3Bhigh
A3Blow
A3Bhigh
A3Blow
4.0
2.9
5.1
1.2
0.18
0.41

Figure 2-14. APOBEC3B-dependent uracil lesions and mutations in breast cancer genomic DNA (data from Fig. 2-2 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).

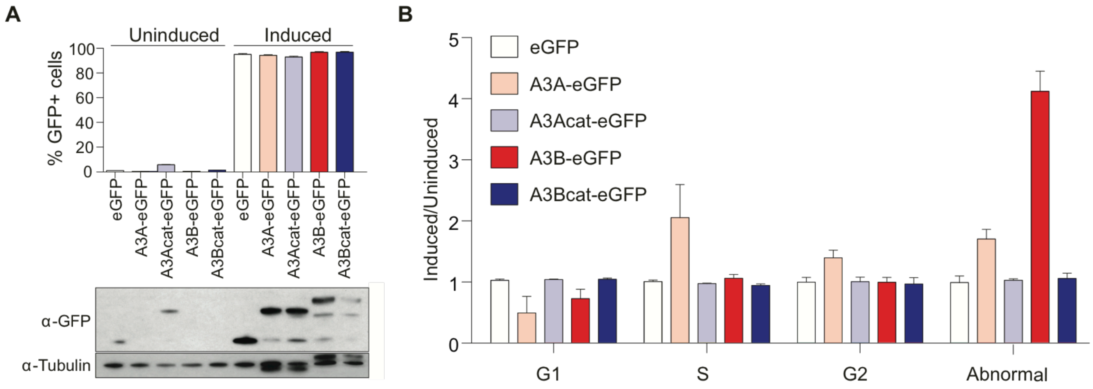


Figure 2-15. Experimental system and cell cycle data for APOBEC3A and APOBEC3B induction.

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).

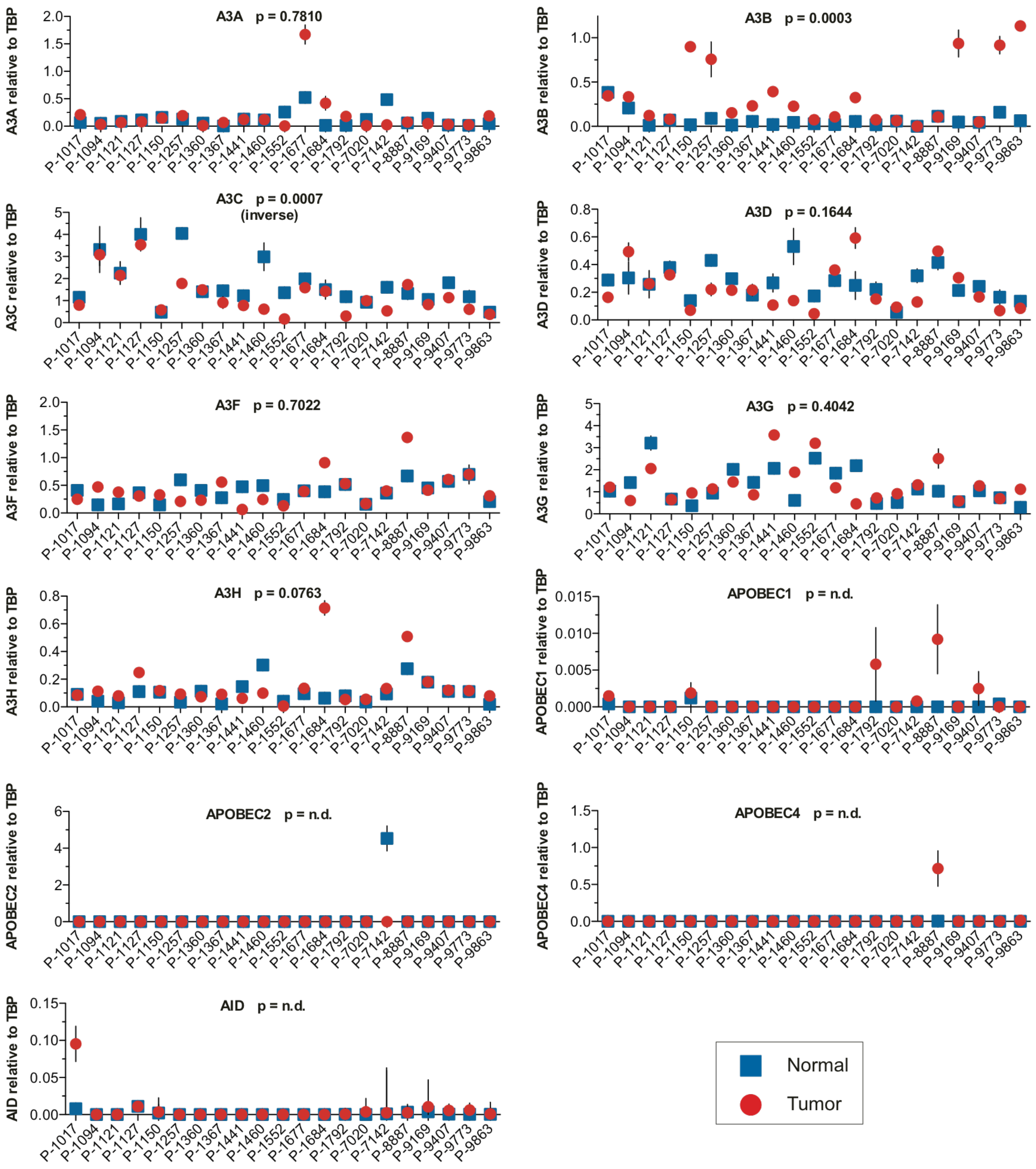

A3A p = 0.7810
A3A relative to TBP
A3B p = 0.0003
A3B relative to TBP
A3C p = 0.0007
(inverse)
A3C relative to TBP
A3D p = 0.1644
A3D relative to TBP
A3F p = 0.7022
A3F relative to TBP
A3G p = 0.4042
A3G relative to TBP
A3H p = 0.0763
A3H relative to TBP
APOBEC1 p = n.d.
APOBEC1 relative to TBP
APOBEC2 p = n.d.
APOBEC2 relative to TBP
APOBEC4 p = n.d.
APOBEC4 relative to TBP
AID p = n.d.
AID relative to TBP
P-1017
P-1094
P-1121
P-1127
P-1150
P-1257
P-1360
P-1367
P-1441
P-1460
P-1552
P-1677
P-1684
P-1792
P-7020
P-7142
P-8887
P-9169
P-9407
P-9773
P-9863
■ Normal
● Tumor

Figure 2-16. 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. 2-4A.

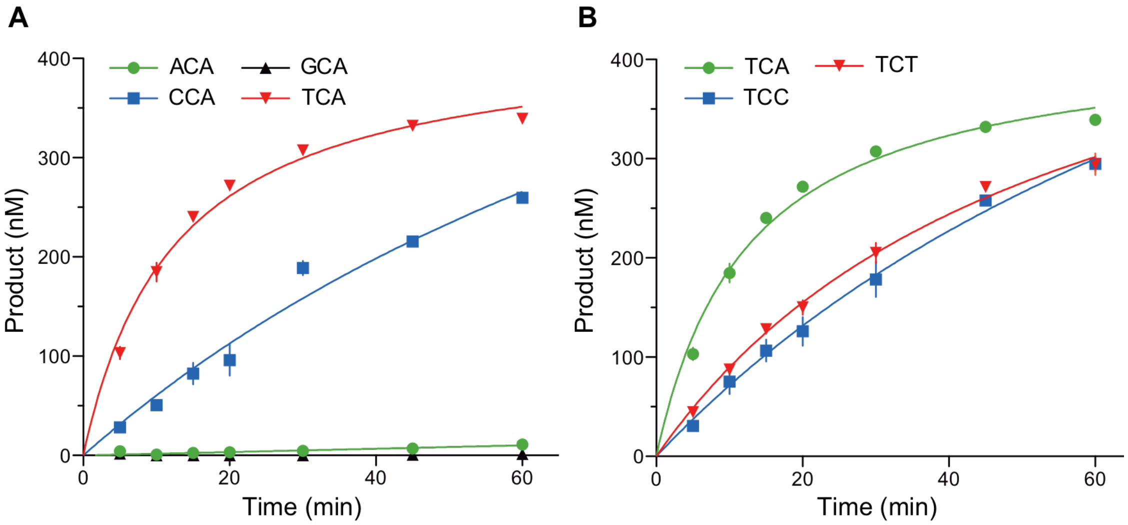


Figure 2-17. APOBEC3B catalytic domain local deamination preferences.

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.

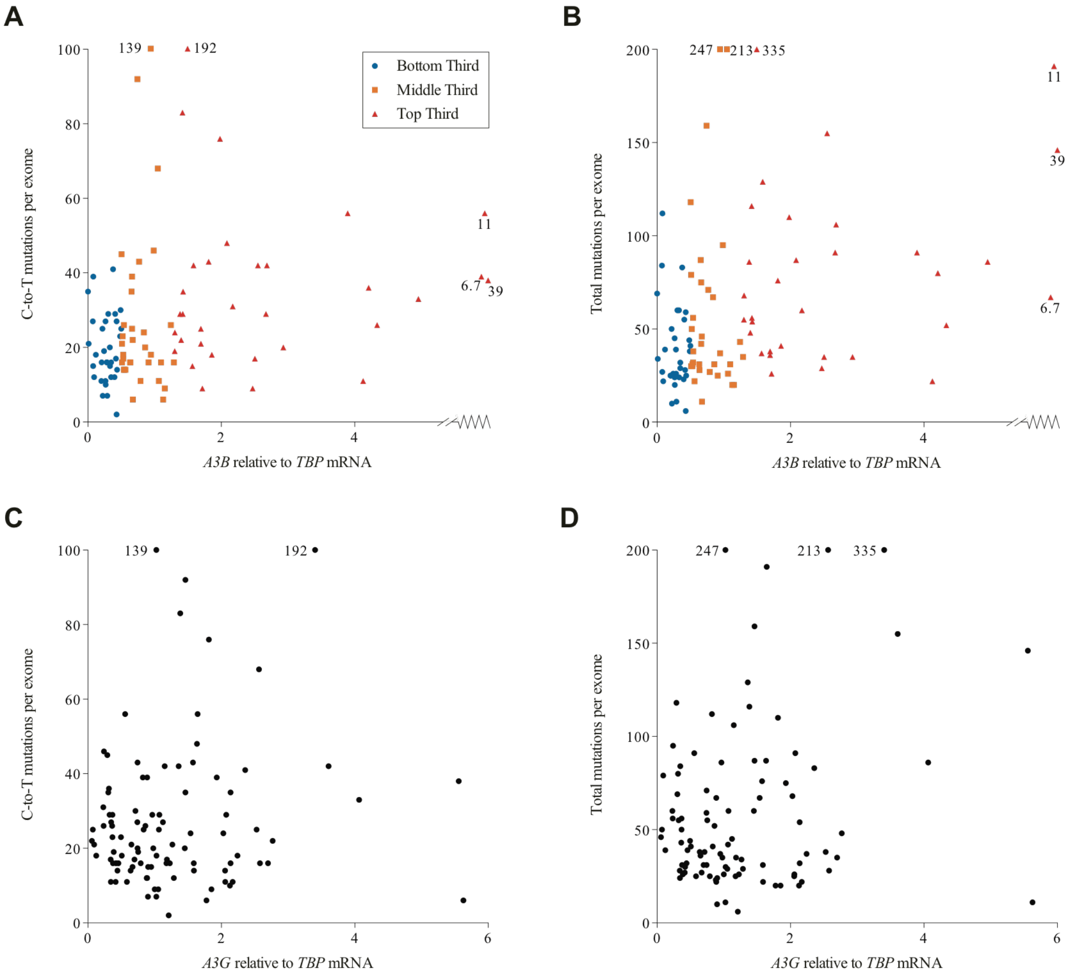


Figure 2-18. Breast cancer mutation load and gene expression correlations.

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. 2-4E & 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. 2-4B.

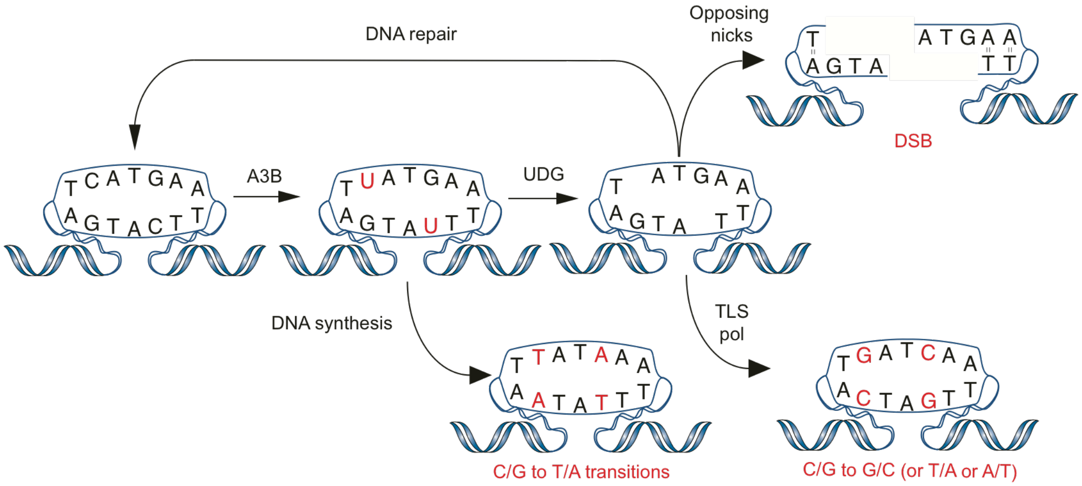


Figure 2-19. DNA deamination model for APOBEC3B in cancer.

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 [3].

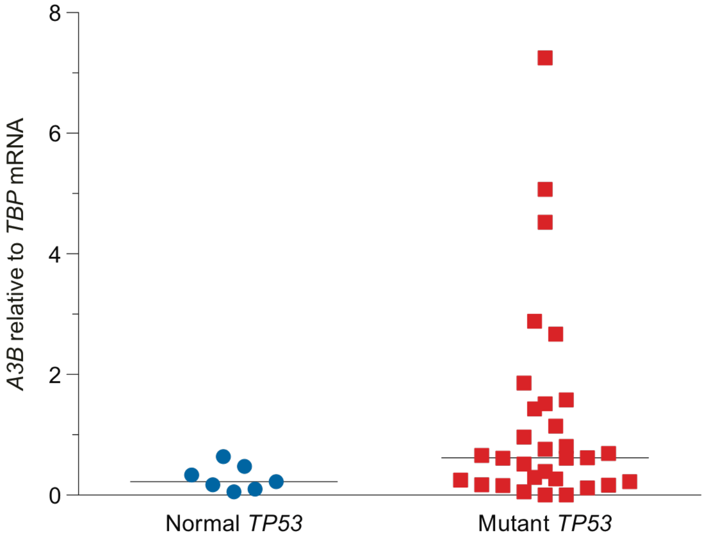


Figure 2-20. *APOBEC3B* up-regulation and *TP53* inactivation in the ATCC breast cancer cell line panel.

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 Table 2-1).

Table 2-1. 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.

Table 2-2. 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 2-3. 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 2-4. 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 2-5. 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 2-6. 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 2-7. 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 2-8. 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 2-9. 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.

Table 2-10. Breast cancer patient information.

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. 2-4**. [#] 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.

Table 2-11. 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	gaccctttggtccttcgac	RSH3221	gcacagccccaggagaag	UPL1	cctggagc
APOBEC3C	NM_014508	RSH3085	agcgcttcagaaaagagtgg	RSH3086	aagtttcgttccgatcgttg	UPL155	ttgccttc
APOBEC3D	NM_152426	RSH2749	acccaaacgtcagtcgaatc	RSH2750	cacatttctgcgtggttctc	UPL51	ggcaggag
APOBEC3F	NM_145298	RSH2751	ccgtttggacgcaaagat	RSH2752	ccaggtgatctggaaacactt	UPL27	gctgcctg
APOBEC3G	NM_021822	RSH2753	ccgaggacccgaaggttac	RSH2754	tccaacagtgctgaaattcg	UPL79	ccaggagg
APOBEC3H	NM_181773	RSH2757	agctgtggccagaagcac	RSH2758	cggaatgtttcggctgtt	UPL21	tggctctg
AID	NM_020661	RSH3066	gactttggttatcttcgcaataaga	RSH3067	aggtcccagtccgagatgta	UPL69	ggaggaag
APOBEC1	NM_001644	RSH3068	gggaccttgttaacagtggagt	RSH3069	ccaggtgggtagttgacaaaa	UPL67	tgctggag
APOBEC2	NM_006789	RSH3070	aagtagggcaactgggcttt	RSH3071	ggctgtacatgtcattgctgtc	UPL74	ctgctgcc
APOBEC4	NM_203454	RSH3072	ttctaacacctggaatgtgatcc	RSH3073	tttactgtcttctagctgcaaacc	UPL80	cctggaga
Reference Gene							
TBP	NM_003194	RSH3231	cccatgactcccatgacc	RSH3232	tttacaaccaagattcactgtgg	UPL51	ggcaggag

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

CHAPTER 3 – APOBEC3B-DRIVEN MUTAGENESIS IN MULTIPLE HUMAN CANCERS

This chapter is adapted with permission from: Burns, Temiz, and Harris, (2013) *Nature Genetics* vol. 45, pp. 997-983.

INTRODUCTION

Thousands of somatic mutations accrue in most human cancers and causes are largely unknown. We recently showed that the DNA cytosine deaminase APOBEC3B may account 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, totaling over 4,800 exomes and 1,000,000 somatic mutations. Remarkably, *APOBEC3B* is upregulated and its preferred target sequence is frequently mutated and clustered in at least 6 distinct cancers: bladder, cervix, lung (adeno- and squamous cell), head/neck, and breast. Interpreted in light of prior genetic, cellular, and biochemical studies, the most parsimonious conclusion based on 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 revealed the existence of hundreds to thousands of mutations in most cancers[30,41-48,75]. 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 C-to-T transition bias in skin cancers can be explained by a mechanism in which UV-induced lesions, cyclobutane pyrimidine dimers (C*C, C*T, T*C, or T*T), are bypassed by DNA polymerase-catalyzed insertion of two adenine bases opposite each unrepaired lesion[76]. A second round of DNA replication or excision and repair of the pyrimidine dimer results in C-to-T transitions. Notably, the nature of this type of DNA damage dictates that each resulting C-to-T transition occurs in a dipyrimidine context, with each mutated cytosine invariably flanked on the 5' or the 3' side by a cytosine or thymine. 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.

Non-random mutation patterns are also observed in other types of cancer, such as C/G base pairs being more frequently mutated than A/T pairs[30,41-48,75] and the occurrence of strand-coordinated clusters of cytosine mutations[30,40,77]. Spontaneous hydrolytic deamination of cytosine to uracil (C-to-U) may explain a subset of these events, but not the majority because most occur outside of potentially methylatable CpG dinucleotide motifs (*i.e.*, sites most prone to spontaneous deamination) and the occurrence of these mutations in clusters is highly non-random. Another possible source of these mutations is enzyme-catalyzed C-to-U deamination by one or more of the nine active DNA cytosine deaminases encoded by the human genome. Such a mechanism was originally hypothesized when the DNA deaminase activity of these enzymes was discovered[9], and was recently highlighted with demonstrations of clustered mutations in breast, head/neck, and other cancers[30,40,77]. These clusters have been named kataegis, as their sporadic but concentrated nature bears likeness to rain showers[30]. Although enzymatic deamination has been implicated in this phenomenon, the actual enzyme responsible has not been determined.

Enzyme-catalyzed DNA C-to-U 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[3,78]. AID also catalyzes uracil lesions in antibody gene switch regions that lead to DNA breaks and juxtaposition of the expressed, and often mutated, variable region next to a new constant region (*i.e.*, isotype switch recombination)[3,78]. 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[8,10]. A well-studied example is the cDNA replication intermediate of HIV-1, which during reverse transcription is vulnerable to enzymatic deamination by at least 3 different APOBEC3 proteins[79,80]. 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[81-84]. More distantly related proteins, APOBEC2 and APOBEC4, have yet to elicit enzymatic activity. In total, nine of eleven APOBEC family members have demonstrated DNA deaminase activity in a variety of biochemical and biological assay systems[9,85-89].

However, a possible drawback of encoding nine active DNA deaminases could be chromosomal DNA damage and, ultimately, mutations that lead to cancer[9]. AID has been linked to B-cell tumorigenesis through off-target chromosomal deamination as well as triggering translocations between the expressed heavy chain locus and various oncogenes[90]. Transgenic expression of AID causes tumor formation in mice[91], as does transgenic expression of APOBEC1[13]. Most recently, we showed that APOBEC3B is upregulated in breast tumors and correlated with a doubling of both C-to-T and overall base substitution mutation loads[1]. Since AID and APOBEC1 are expressed tissue specifically and there is no reason to suspect developmental confinement of APOBEC3B, we hypothesized that APOBEC3B may be a general mutagenic factor impacting the genesis and evolution of many different cancers. This hypothesis is supported by studies indicating high *APOBEC3B* expression in many different cancer cell lines[1,2,92], in contrast to relatively low expression in 21 normal human tissues spanning all major organs[1,2,93]. This DNA mutator hypothesis is additionally supported by the fact that APOBEC3B is the only deaminase family member with constitutive nuclear localization[1,33].

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 carcinomas, 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[1] 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 the 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 eleven APOBEC family members across a panel of 19 distinct tumor types, including breast cancer as a positive control[1] (Table 3-1 and Fig. 3-5). The expression values for each target mRNA were normalized to those of the constitutive housekeeping gene *TATA-binding protein* (*TBP*) to enable quantitative comparisons between RNAseq and RT-qPCR data sets and to provide controls for the few instances where RNAseq values for normal tissues were not available publicly (Methods).

Several cancers showed *APOBEC3B* expression levels comparable to those in corresponding normal tissues (Figs. 3-1 and 3-5, Tables 3-1 and 3-2). Prostate and renal clear cell carcinomas showed statistically significant upregulation of *APOBEC3B* in the tumors, albeit with median expression values that are only a fraction of *TBP*. In contrast, 6 different cancers showed evidence for strong APOBEC3B upregulation in the majority of tumors of the breast, uterus, bladder, head & neck, and lung (adeno- and squamous cell carcinomas) ($p<0.0001$ 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. Remarkably, a total of 10 cancers showed a median level of *APOBEC3B* upregulation greater than that of the intended positive control, breast cancer. This was particularly striking for bladder, head/neck, both lung, and cervical cancers.

The second major prediction of the APOBEC mutator hypothesis is chromosomal DNA C-to-U deamination, which should result in strong biases toward mutations at C/G base pairs. Such mutational events may be either transitions or transversions because genomic uracils can directly template the insertion of adenines during DNA replication and, if converted to abasic sites by uracil DNA glycosylase, the lesions become non-instructional and error-prone polymerases may insert 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 C/G base pairs (*i.e.*, C/G-to-T/A, C/G-to-G/C, and C/G-to-A/T mutations; see Discussion for model).

Interestingly, the fraction of mutations at C/G 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. 3-2A). The massive bias in skin cancers is largely attributable to error-prone DNA synthesis (A insertion) opposite cyclobutane pyrimidine dimers caused by UV light[76]. However, the biases observed in urogenital carcinomas such as bladder and cervical cancers are probably not due to UV but more likely to an alternative mutagenic source such as enzymatic DNA deamination. Indeed, the top 5 tumor types with C/G dominated mutation spectra are among the top 6 tumors in terms of *APOBEC3B* expression (compare Fig. 3-1 and Fig. 3-2A). A possible mechanistic relationship is further supported by a positive correlation between overall proportion of mutations occurring at C/G base pairs and median *APOBEC3B* levels (p=0.0031, r=0.64 by Spearman's correlation; Fig. 3-2B). The positive correlation is remarkable 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 in skin cancer), undefined mutational sources (glioma with the 6th highest C/G mutation bias and lowest *APOBEC3B* levels), and differential DNA repair capabilities among the distinct tumor types (discussed further below).

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' adenines or guanines, APOBEC3G prefers 5' cytosines, and other family members prefer 5' thymines[50,94,95]. We recently showed that recombinant APOBEC3B prefers 5' thymines and strongly disfavors 5' purines; on the 3' side, it prefers adenines or guanines, and disfavors pyrimidines[1] (Fig. 3-3A). Therefore, the third and possibly most important prediction of the APOBEC mutator hypothesis is that cancers impacted by enzymatic deamination should show non-random nucleotide distributions immediately 5' and 3' of mutated cytosines, 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 sequence signature analysis on all available cytosine mutation data from the upper 50% of APOBEC3B-expressing tumors for each tumor type (this cut-off was chosen to minimize the impact of unrelated mutational mechanisms). These mutation data were first compiled and subjected to a hierarchical cluster analysis to group tumors with similar cytosine mutation signatures (Fig. 3-3A). Short Euclidean distances (*i.e.*, smaller measures) between the mutation signatures of different tumors indicate a high degree of concordance, *i.e.* similar mutational patterns (Table 3-3 lists calculated values). Bladder and cervical cancers, two of the top *APOBEC3B*-expressing cancers, had cytosine mutation signatures remarkably similar to each other and to that of recombinant APOBEC3B. This is visually evidenced by strong mutation biases at 5’T$\underline{C}$A motifs, which match the enzyme’s optimal *in vitro* substrate. The two lung cancers, breast cancer, and head/neck cancer also had cytosine mutation signatures that strongly resembled the preference of recombinant APOBEC3B (Fig. 3-3A and Table 3-3). Several cancers had cytosine mutation signatures with an intermediate relatedness to recombinant APOBEC3B (renal papillary, thyroid, ovarian, renal clear cell, GBM, and skin). In further contrast, the seven remaining cancers had the largest separation from recombinant APOBEC3B ranging from uterine to colon cancer (Fig. 3-3A and Table 3-3).

We next separated each composite mutation distribution into the 16 individual local trinucleotide contexts to further resolve cytosine-focused mutational mechanisms that may be influencing each cancer. Bladder, cervical, lung squamous, lung adeno-, head/neck, and breast carcinomas all shared strong 5’T$\underline{C}$N mutation signatures, with 5’T$\underline{C}$A being strongest of the four possibilities (boxed in Fig. 3-3B). 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 C-to-T transitions predominating within dipyrimidine contexts (middle dashed boxes in Fig. 3-3B). Only two other obvious cytosine-focused mutation patterns were evident. C-to-T mutations at 5’CG contexts dominated at least seven types of cancer, consistent with a 5’CG targeted mechanism such as spontaneous deamination of methyl-cytosine (lower dashed boxes in Fig. 3-3B). Finally, uterine, low-grade glioma, rectal, and colon cancers had an inordinate number to C-to-A transversions in 5’(YCT contexts) consistent with at least one additional distinct cytosine-focused mutational mechanism (*e.g.*, POLE proofreading domain variants have been implicated in a subset of colorectal tumors[96]).

A fourth prediction of a general mutator hypothesis is that tumor mutation loads ought to 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 ensure the inclusion of all data, yet simultaneously minimize the impact of uncontrollable variables such as other mutational mechanisms, jackpot effects, bottlenecks, tumor ages, *etc*. As recently reviewed[97], mutation loads vary 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 (AML to skin cancer in Fig. 3-4A). However, despite this incredible 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. 3-4B). This result is consistent with the possibility that APOBEC3B may be a general endogenous mutagen that contributes to most human cancers albeit, as outlined above, clearly much more to a subset of cancers. A dominant role for APOBEC3B in a subset of cancers is further evidenced 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 (Figs. 3-6 and 3-7).

A final prediction of a general APOBEC mutator hypothesis is that impacted cancers should bear evidence for strand-coordinated clusters of cytosine mutations[30,40,77]. As proposed[40], clusters can be defined as 2 or more mutation events within a 10 kbp window. By this criterion, every cancer showed evidence for 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 dramatically and can result in the inclusion of false-positives (see Roberts *et al.*[40] and 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 3-1). Interestingly, the number of mutations grouped into kataegis was a relatively small percentage of the total number of cytosine mutations for each cancer (maximally 7.9%). However, the sheer existence of clustered cytosine mutation in nearly every cancer provides further evidence for APOBEC involvement. For most cancers this is likely to be APOBEC3B because average number of kataegis per tumor correlates positively with median *APOBEC3B* expression levels (p=0.017 and r=0.54 by Spearman correlation; Fig.

3-4C). The 6 cancer types with cytosine mutation signatures that grouped most closely with recombinant APOBEC3B, bladder, cervix, lung (adeno- and squamous cell carcinomas), head/neck, and breast, all showed strong evidence for kataegis with a mean 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 analyses of Nik-Zainal *et al.*[30]). Interestingly, low-grade gliomas and uterine carcinomas are clear outliers in this analysis, consistent with the close hierarchical clustering of their cytosine mutation signatures (distant from recombinant APOBEC3B) and strongly suggesting another distinct mutational mechanism.

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[30,41-48,75,77]. The 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, lung adenocarcinoma, head/neck, and breast, clearly stood out, with elevated *APOBEC3B* expression in the majority of tumors, strong overall C/G mutation biases, cytosine mutation contexts that closely resemble the deamination signature of recombinant APOBEC3B, and evidence for 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 these cytosine mutations in these types of cancers. These data are consistent with a general mutator hypothesis, in which APOBEC3B mutagenesis has the capacity to broadly shape the mutation 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 C-to-U lesions, which may be processed into a variety of mutagenic outcomes[1] (Fig. 3-8). In most instances, uracil lesions are repaired faithfully by canonical base excision repair. However, in some instances, uracil lesions may template the insertion of adenines during DNA synthesis, which may result in C-to-T transitions (G-to-A on the opposing strand). In other instances, genomic uracils may be converted to abasic sites by uracil DNA glycosylase. These lesions are noninstructional such that DNA polymerases, in particular translesion DNA polymerases, may place any base opposite, with an A leading to a transition and a C or T 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 undoubtedly intermediates in the formation of cytosine mutation clusters (kataegis)[30,40,77] and larger-scale chromosomal aberrations such as translocations.

The significant positive correlations between *APOBEC3B* expression levels and the percentage of mutations at C/G pairs, the overall mutation loads, and the number of kataegis events combine to suggest that most cancers are impacted 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 rank and clear evidence for kataegis, but it also has a strong dipyrimidine-focused C-to-T mutation pattern that could easily eclipse an APOBEC3B deamination signature. APOBEC3B may help explain melanomas that occur with minimal UV exposure[5]. Several other cancers such as uterine, rectal, stomach, and ovarian also have significant *APOBEC3B* upregulation and evidence for kataegis, which combine to suggest direct involvement, 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[90], roles for other APOBECs are unlikely to be as great as those of APOBEC3B for the following reasons: i) no reported enzymatic activity (APOBEC2 and APOBEC4), ii) tissue-restricted expression profiles (AID, APOBEC3A, APOBEC1, APOBEC2, and APOBEC4)[1,2,36,93,98-101], iii) localization to the cytoplasmic compartment (APOBEC3A, APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H)[33,35,86,102], and iv) in two instances, a completely different intrinsic preference for bases surrounding the target cytosine (AID and APOBEC3G prefer 5'R$\underline{C}$ and 5'C$\underline{C}$, respectively)[1,50,94,95]. Thus, taken together with the comprehensive analyses presented here of expression data (Fig. 3-

1), C/G mutation frequencies (Fig. 3-2), local cytosine mutation signatures (Fig. 3-3), overall mutation loads (Fig. 3-4), and kataegis (Fig. 3-4C and Table 3-1), all available data converge upon the conclusion that APOBEC3B is a major source of mutation in multiple human cancers. This knowledge provides foundations for future studies focused on each cancer type and sub-type to further delineate the impact of this potent DNA mutator on each cancer genome and on associated therapeutic responses and patient outcomes.

METHODS

Data Analyses

A description of tumor types, tumor *APOBEC3B* expression data, and tumor exome mutation data is provided in Table 3-1. Information for the corresponding normal tissues is provided in Table 3-2. Somatic mutations and RNAseq expression data were retrieved from the Cancer Genome Atlas Data Matrix on January 3rd, 2013. Gene expression data were mined from RNAseqV2 datasets for all cancers (normalized expression values) with the exception of LAML and STAD, which were from RNAseq datasets (RPKM values). Additional normal sample RNAseqV2 data were downloaded from TCGA on April 4th, 2013 to include recently released normal sample information for READ and COAD. *APOBEC3B* expression values were normalized to the expression of *TBP* for each patient sample. Comparisons between the normal RNAseq-derived gene expression values and the tumor expression values were performed using the Mann-Whitney U test to determine significance. All RT-qPCR values for normal tissues were reported previously based on data from pooled normal samples[1,2], with the exception of salivary gland, stomach, skin, and rectal tissues, which are unique to this report. The primary tissue RNA was generated using published methods[2] and total RNA obtained commercially (salivary gland RNA for head/neck and stomach RNA were obtained from Clontech and skin and rectal RNA were obtained from USBiological). Each *A3B* relative to *TBP* value from RTqPCR was multiplied by an experimentally derived factor of 2 to facilitate direct comparisons with RNAseq values (unpublished data).

Mutation data were taken from maf files downloaded from TCGA Somatic Mutation database (http://tcga-data.nci.nih.gov/tcga). Insertions/deletions and adjacent multiple mutations (di- and trinucleotide variations) were removed and the remaining single nucleotide variations (SNVs) were converted to hg19 coordinates (Table 3-4). Non-mutations with respect to the reference genome (*e.g.*, C-to-C) were eliminated and duplicate entries were removed unless they were reported for different patient samples. Comparisons between mutation 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 C 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 C-to-T, C-to-G, and C-to-A mutations were counted and placed in a table and normalized to one to reflect the fraction of each mutation type. This 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 "complete" option (http://www.r-project.org). The Euclidean distance is the ordinary distance between two data points on a 2D plot (Table 3-3 lists all calculated Euclidean distances).

A kataegis event is defined as two or more mutations within a 10,000 nucleotide genomic DNA window. The probability of each event occurring by chance is then calculated following the work of Gordenin and colleagues[40]. 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 utilizing 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 patients and exome size). The significant kataegis events with p-values less than 10^{-4} for each cancer are reported in Table 3-1. "Gordenin significance" indicates that a given cluster of mutations has met the above criteria and attained significance. This approach minimizes false positive cluster-calls resulting by random chance.

Contributions: All authors contributed to the study designs, data analyses, and manuscript preparation. MBB and NAT analyzed data from TCGA. NAT performed mutation and cluster analysis. The Cancer Genome Atlas (TCGA) Network generated the RNAseq and somatic mutation data and provided open access, and Harris lab members and S. Kaufmann gave 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, Minnesota Ovarian Cancer Alliance, and National Institutes for Health (R01 AI064046 and P01 GM091743).

FIGURES & TABLES

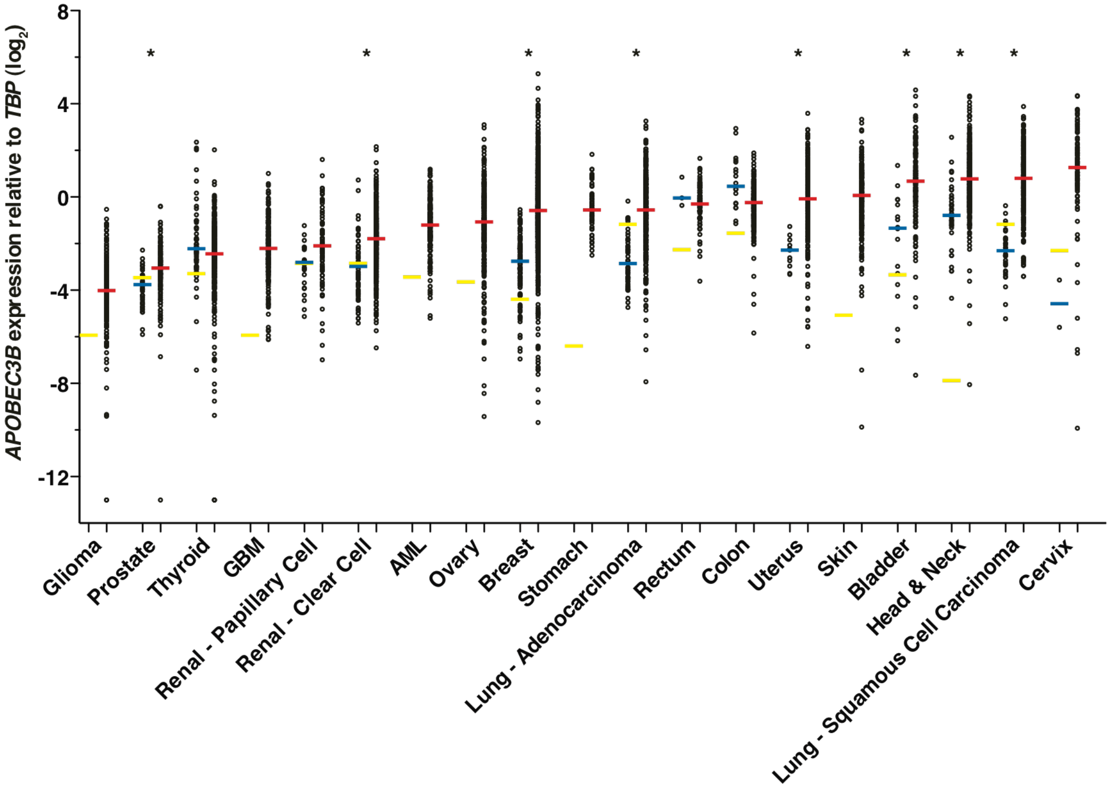


Figure 3-1. ***APOBEC3B*** **is upregulated in numerous cancer types.**

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 lines indicate the median *APOBEC3B*/*TBP* value for each cancer type (**Table 3-1**), the median value for each set of normal tissue RNAseq data (**Table 3-2**), and individual RT-qPCR 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.

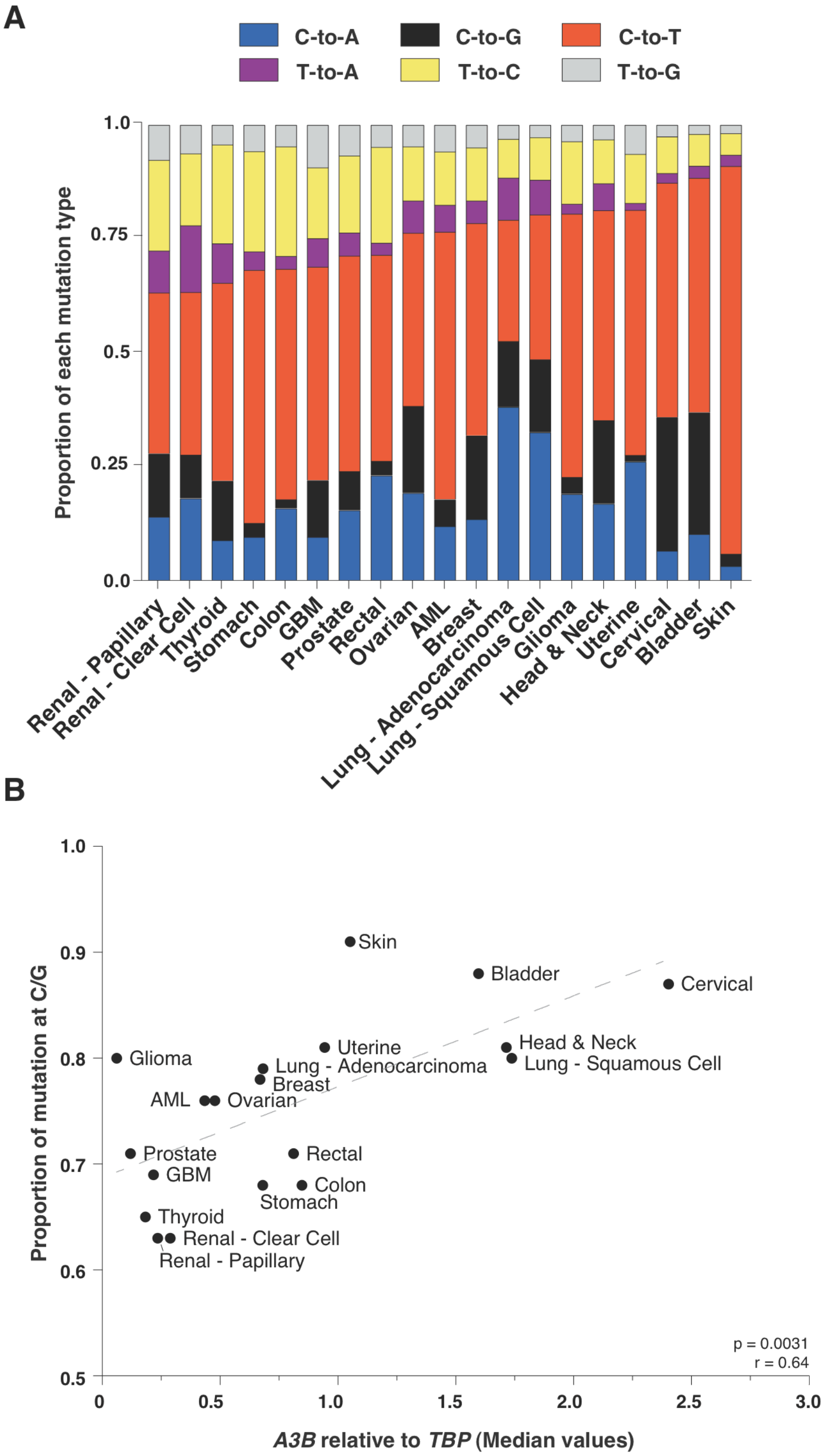

A
C-to-A
C-to-G
C-to-T
T-to-A
T-to-C
T-to-G
Proportion of each mutation type
1.0
0.75
0.5
0.25
0.0
Renal - Papillary
Renal - Clear Cell
Thyroid
Stomach
Colon
GBM
Prostate
Rectal
Ovarian
AML
Breast
Lung - Adenocarcinoma
Lung - Squamous Cell
Glioma
Head & Neck
Uterine
Cervical
Bladder
Skin
B
Proportion of mutation at C/G
1.0
0.9
0.8
0.7
0.6
0.5
Skin
Bladder
Cervical
Uterine
Head & Neck
Lung - Squamous Cell
Glioma
Lung - Adenocarcinoma
Breast
AML
Ovarian
Prostate
Rectal
GBM
Colon
Stomach
Thyroid
Renal - Clear Cell
Renal - Papillary
p = 0.0031
r = 0.64
0
0.5
1.0
1.5
2.0
2.5
3.0
A3B relative to TBP (Median values)

Figure 3-2. Mutation types and signatures in 19 human cancers.

A, Stacked bar graph summarizing the 6 types of base substitution mutations as proportions of the total mutations per cancer.

B, Median *APOBEC3B* relative to *TBP* expression levels plotted against the proportion of mutations at C/G base pairs (Spearman p = 0.0031, r = 0.64). Dashed grey line is the best-fit for visualization.

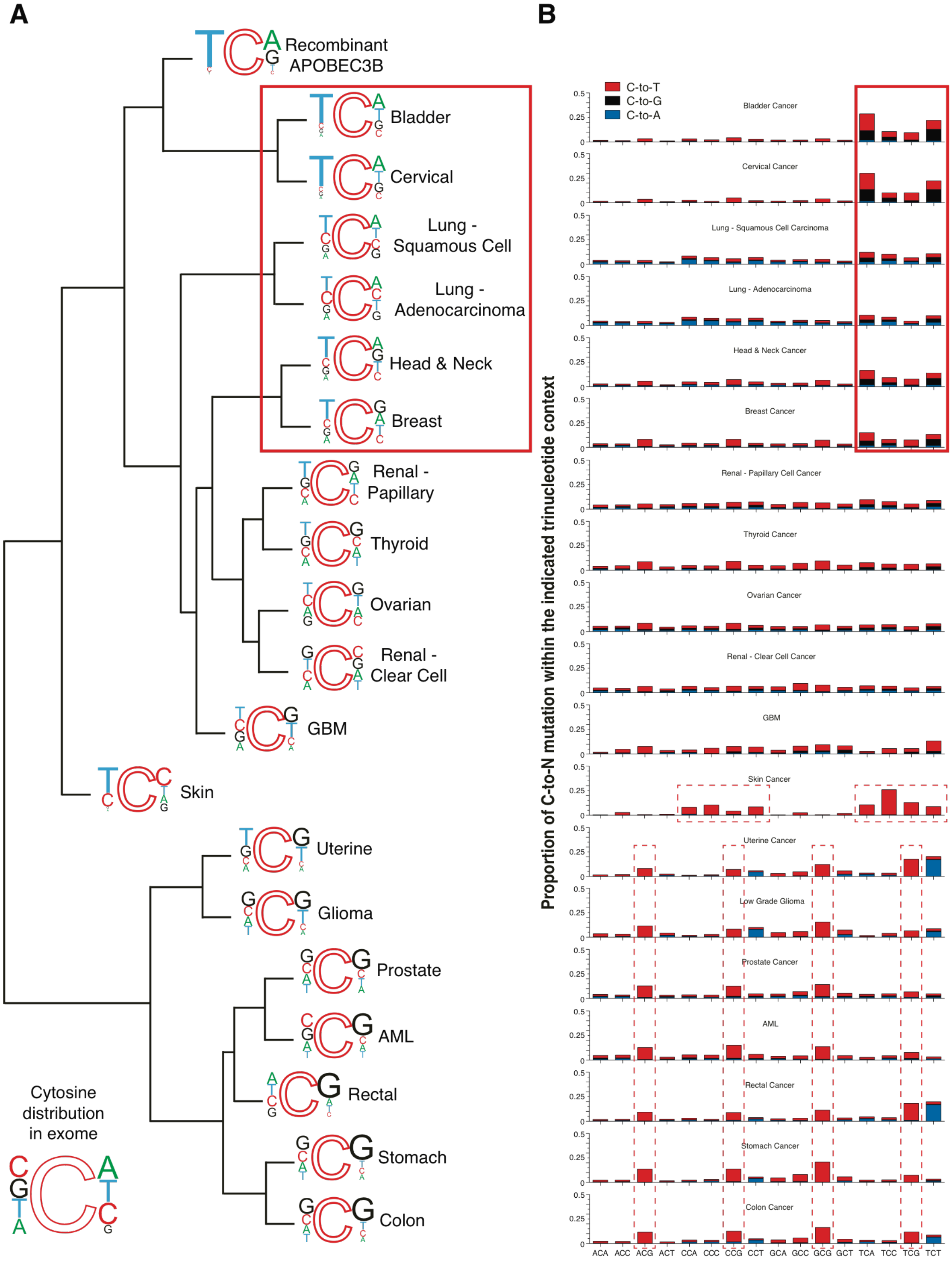

A
B
Recombinant APOBEC3B
Bladder
Cervical
Lung - Squamous Cell
Lung - Adenocarcinoma
Head & Neck
Breast
Renal - Papillary
Thyroid
Ovarian
Renal - Clear Cell
GBM
Skin
Uterine
Glioma
Prostate
AML
Rectal
Stomach
Colon
Cytosine distribution in exome
C-to-T
C-to-G
C-to-A
Proportion of C-to-N mutation within the indicated trinucleotide context
Bladder Cancer
Cervical Cancer
Lung - Squamous Cell Carcinoma
Lung - Adenocarcinoma
Head & Neck Cancer
Breast Cancer
Renal - Papillary Cell Cancer
Thyroid Cancer
Ovarian Cancer
Renal - Clear Cell Cancer
GBM
Skin Cancer
Uterine Cancer
Low Grade Glioma
Prostate Cancer
AML
Rectal Cancer
Stomach Cancer
Colon Cancer
0.5
0.25
0
ACA ACC ACG ACT CCA CCC CCG CCT GCA GCC GCG GCT TCA TCC TCG TCT

Figure 3-3. Cytosine mutation spectra for 19 cancers.

A, Dendrogram with weblogos indicating the relationship among cancer types determined by the trinucleotide contexts of mutations occurring at C nucleotides for the top 50% *APOBEC3B* expressing samples within each cancer type. Font size of the bases at the 5' and 3' positions are proportional to their observed occurrence in exome mutation datasets. The preferred mutation context for recombinant APOBEC3B from Burns and Lackey, *et al.* [1] is included in the hierarchical clustering in order to determine how closely each cancers' actual mutation spectrum matches the preferred motif for APOBEC3B *in vitro*. The pattern expected if the mutations were to occur at random C 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 [5'-N<u>C</u>N-to-N(<u>T/G/A</u>)N]. Bar color indicates each mutation type: red: C-to-T, black: C-to-G, and blue: C-to-A. The top 6 cancer types (highlighted by solid line box) show clear biases toward mutations within 5'TCN motifs, at frequencies that resemble the preferences of recombinant APOBEC3B *in vitro* [1]. Skin cancer and the bottom 7 cancers (highlighted by dashed line boxes) have obviously different cytosine mutation spectra.

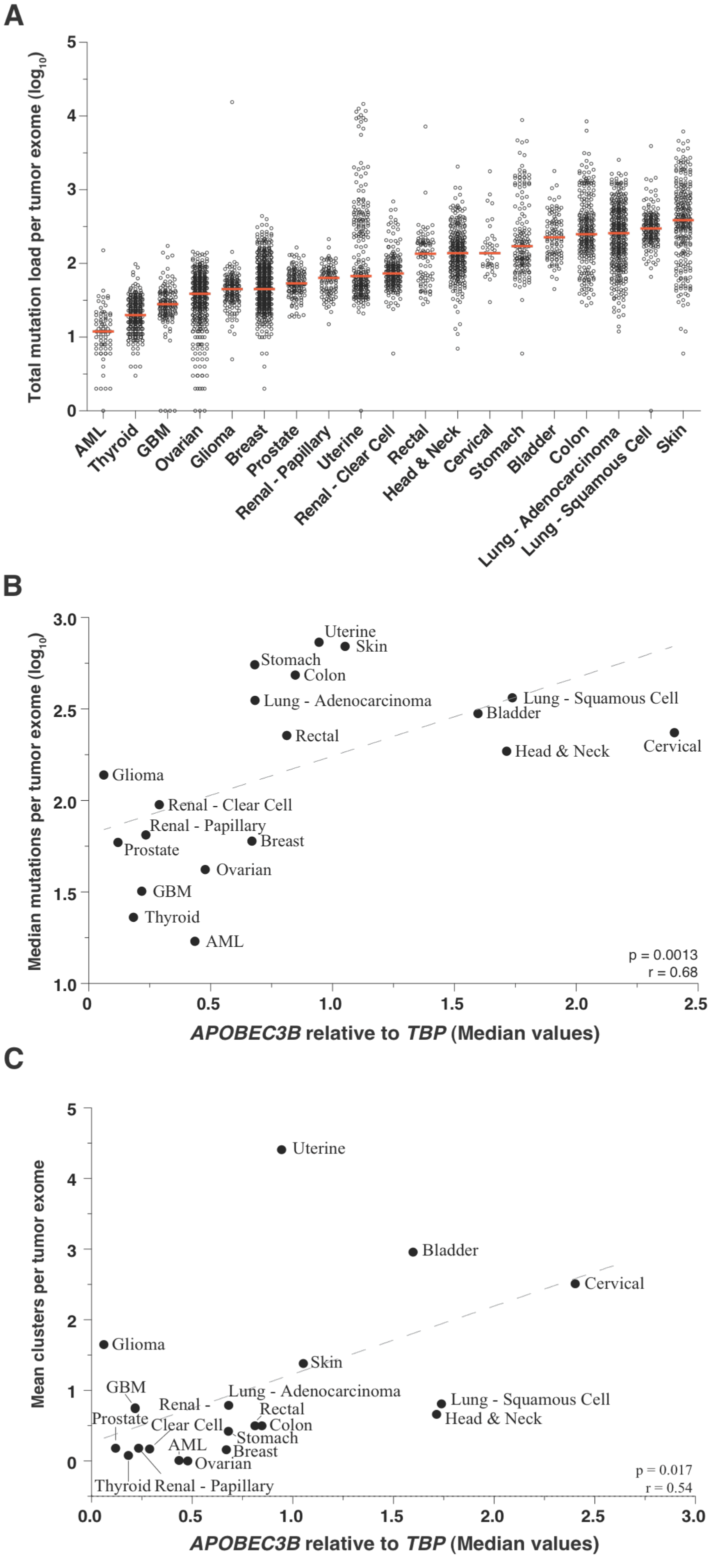

A
Total mutation load per tumor exome ($\log_{10}$)
0
1
2
3
4
5
AML
Thyroid
GBM
Ovarian
Glioma
Breast
Prostate
Renal - Papillary
Uterine
Renal - Clear Cell
Rectal
Head & Neck
Cervical
Stomach
Bladder
Colon
Lung - Adenocarcinoma
Lung - Squamous Cell
Skin
B
Median mutations per tumor exome ($\log_{10}$)
1.0
1.5
2.0
2.5
3.0
0
0.5
1.0
1.5
2.0
2.5
APOBEC3B relative to *TBP* (Median values)
Uterine
Skin
Stomach
Colon
Lung - Adenocarcinoma
Lung - Squamous Cell
Bladder
Rectal
Head & Neck
Cervical
Glioma
Renal - Clear Cell
Renal - Papillary
Prostate
Breast
Ovarian
GBM
Thyroid
AML
p = 0.0013
r = 0.68
C
Mean clusters per tumor exome
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
APOBEC3B relative to *TBP* (Median values)
Uterine
Bladder
Cervical
Glioma
Skin
GBM
Lung - Adenocarcinoma
Lung - Squamous Cell
Head & Neck
Renal -
Clear Cell
Rectal
Colon
Prostate
Stomach
AML
Breast
Ovarian
Thyroid
Renal - Papillary
p = 0.017
r = 0.54

Figure 3-4. ***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 line shows the median mutation load for each cancer type.

B, Median mutation loads per tumor exome for each cancer type plotted against the median *APOBEC3B* relative to *TBP* expression values (Spearman p = 0.0013, r = 0.68). Dashed grey 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* relative to *TBP* expression values (Spearman p = 0.0017, r = 0.54). Dashed grey line is the best-fit for visualization.

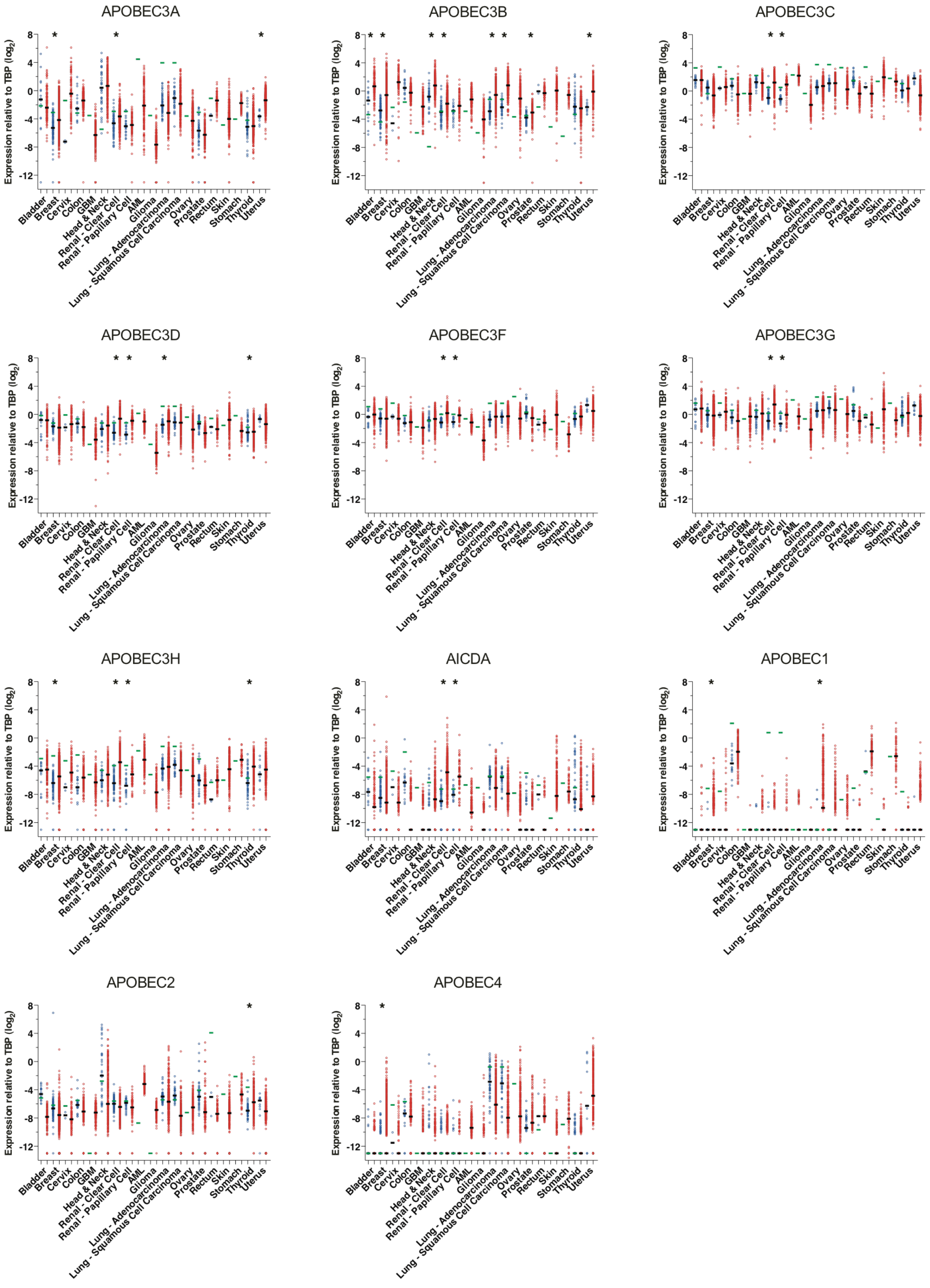
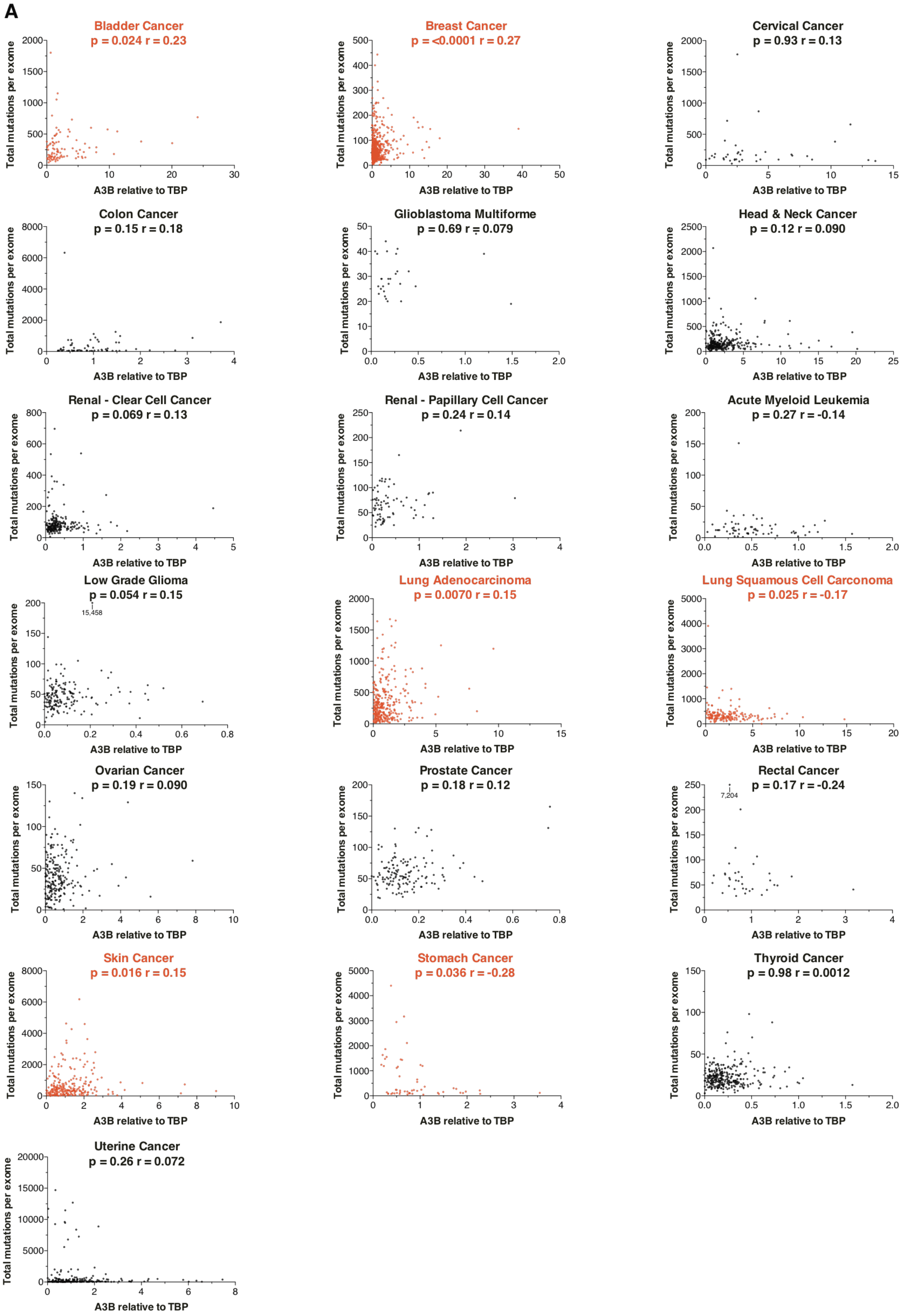

APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
AICDA
APOBEC1
APOBEC2
APOBEC4
Expression relative to TBP (log$_2$)
8
4
0
-4
-8
-12
*
Bladder
Breast
Cervix
Colon
GBM
Head & Neck
Renal - Clear Cell
Renal - Papillary Cell
AML
Glioma
Lung - Adenocarcinoma
Lung - Squamous Cell Carcinoma
Ovary
Prostate
Rectum
Skin
Stomach
Thyroid
Uterus

Figure 3-5. ***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 3-1** and **Table 3-1**). 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. 3-1** 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 [1,2]. 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 [1]. 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

Figure 3-5 (cont.) 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**).


A
Bladder Cancer
p = 0.024 r = 0.23
Breast Cancer
p = <0.0001 r = 0.27
Cervical Cancer
p = 0.93 r = 0.13
Colon Cancer
p = 0.15 r = 0.18
Glioblastoma Multiforme
p = 0.69 r = 0.079
Head & Neck Cancer
p = 0.12 r = 0.090
Renal - Clear Cell Cancer
p = 0.069 r = 0.13
Renal - Papillary Cell Cancer
p = 0.24 r = 0.14
Acute Myeloid Leukemia
p = 0.27 r = -0.14
Low Grade Glioma
p = 0.054 r = 0.15
15,458
Lung Adenocarcinoma
p = 0.0070 r = 0.15
Lung Squamous Cell Carconoma
p = 0.025 r = -0.17
Ovarian Cancer
p = 0.19 r = 0.090
Prostate Cancer
p = 0.18 r = 0.12
Rectal Cancer
p = 0.17 r = -0.24
7,204
Skin Cancer
p = 0.016 r = 0.15
Stomach Cancer
p = 0.036 r = -0.28
Thyroid Cancer
p = 0.98 r = 0.0012
Uterine Cancer
p = 0.26 r = 0.072
Total mutations per exome
A3B relative to TBP

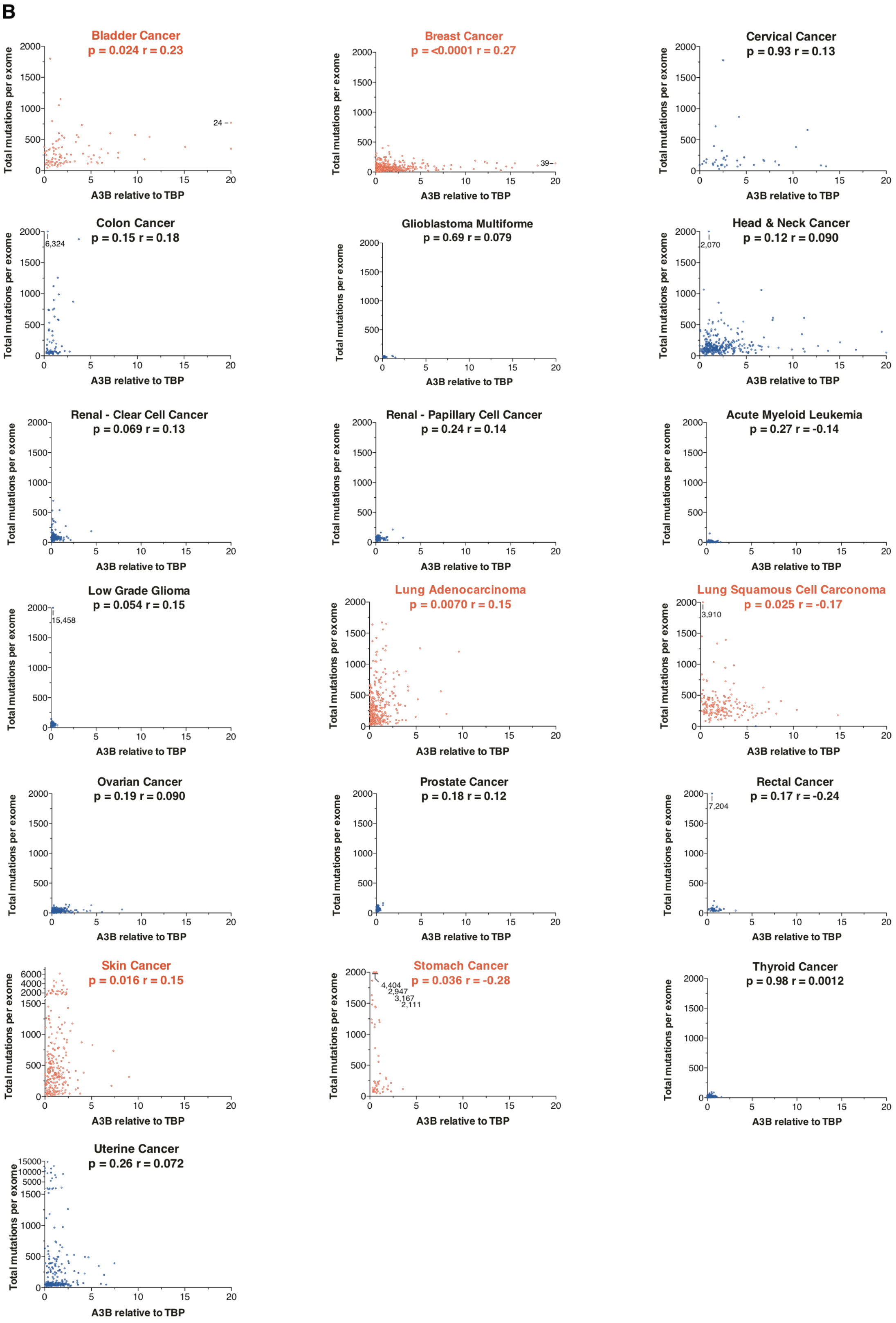
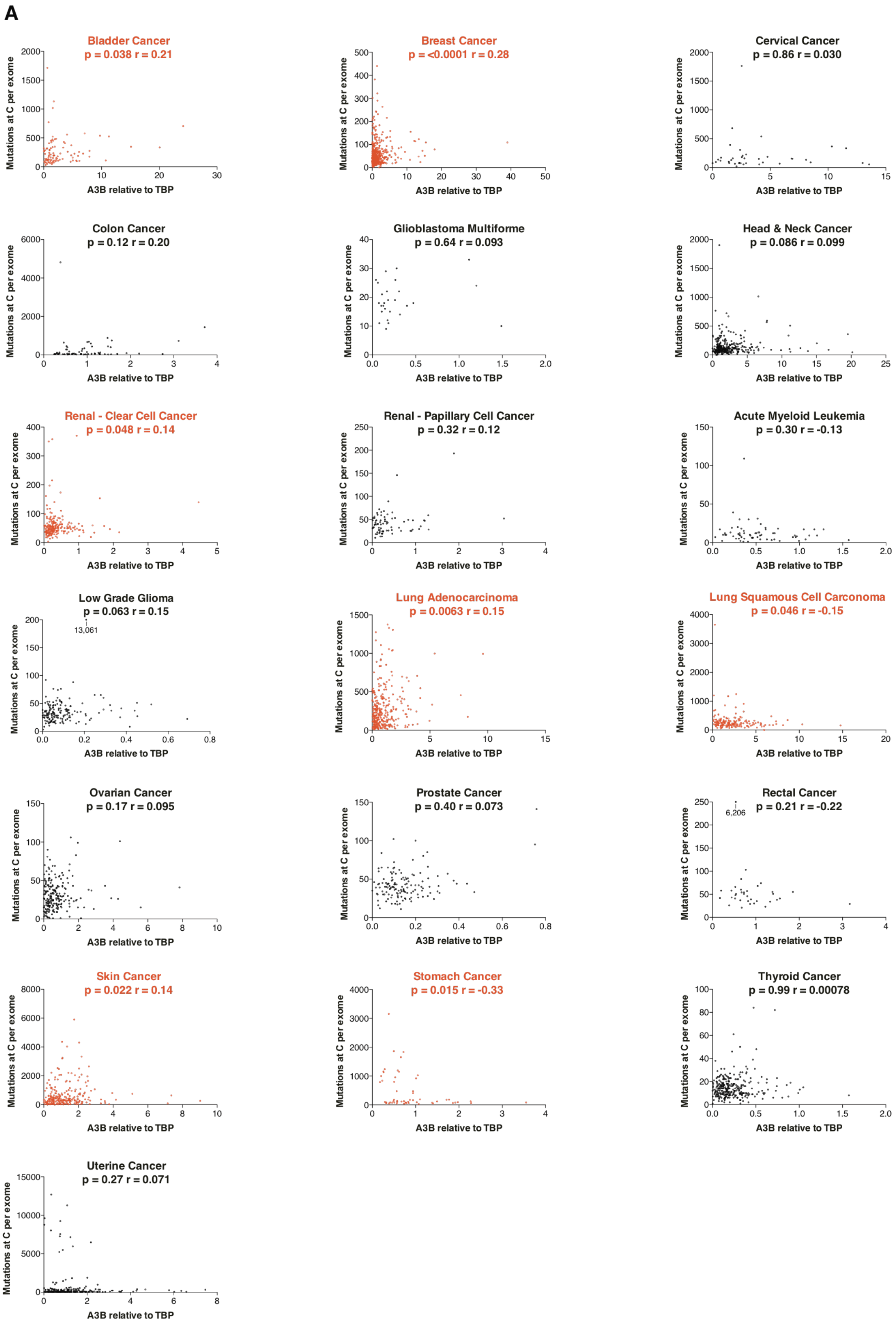

B
Bladder Cancer
p = 0.024 r = 0.23
24
Breast Cancer
p = <0.0001 r = 0.27
39
Cervical Cancer
p = 0.93 r = 0.13
Colon Cancer
p = 0.15 r = 0.18
6,324
Glioblastoma Multiforme
p = 0.69 r = 0.079
Head & Neck Cancer
p = 0.12 r = 0.090
2,070
Renal - Clear Cell Cancer
p = 0.069 r = 0.13
Renal - Papillary Cell Cancer
p = 0.24 r = 0.14
Acute Myeloid Leukemia
p = 0.27 r = -0.14
Low Grade Glioma
p = 0.054 r = 0.15
15,458
Lung Adenocarcinoma
p = 0.0070 r = 0.15
Lung Squamous Cell Carconoma
p = 0.025 r = -0.17
3,910
Ovarian Cancer
p = 0.19 r = 0.090
Prostate Cancer
p = 0.18 r = 0.12
Rectal Cancer
p = 0.17 r = -0.24
7,204
Skin Cancer
p = 0.016 r = 0.15
6000
4000
2000
Stomach Cancer
p = 0.036 r = -0.28
4,404
2,947
3,167
2,111
Thyroid Cancer
p = 0.98 r = 0.0012
Uterine Cancer
p = 0.26 r = 0.072
15000
10000
5000
Total mutations per exome
0
500
1000
1500
2000
A3B relative to TBP
0
5
10
15
20

Figure 3-6. 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.


A
Bladder Cancer
p = 0.038 r = 0.21
Breast Cancer
p = <0.0001 r = 0.28
Cervical Cancer
p = 0.86 r = 0.030
Colon Cancer
p = 0.12 r = 0.20
Glioblastoma Multiforme
p = 0.64 r = 0.093
Head & Neck Cancer
p = 0.086 r = 0.099
Renal - Clear Cell Cancer
p = 0.048 r = 0.14
Renal - Papillary Cell Cancer
p = 0.32 r = 0.12
Acute Myeloid Leukemia
p = 0.30 r = -0.13
Low Grade Glioma
p = 0.063 r = 0.15
13,061
Lung Adenocarcinoma
p = 0.0063 r = 0.15
Lung Squamous Cell Carconoma
p = 0.046 r = -0.15
Ovarian Cancer
p = 0.17 r = 0.095
Prostate Cancer
p = 0.40 r = 0.073
Rectal Cancer
p = 0.21 r = -0.22
6,206
Skin Cancer
p = 0.022 r = 0.14
Stomach Cancer
p = 0.015 r = -0.33
Thyroid Cancer
p = 0.99 r = 0.00078
Uterine Cancer
p = 0.27 r = 0.071
Mutations at C per exome
A3B relative to TBP

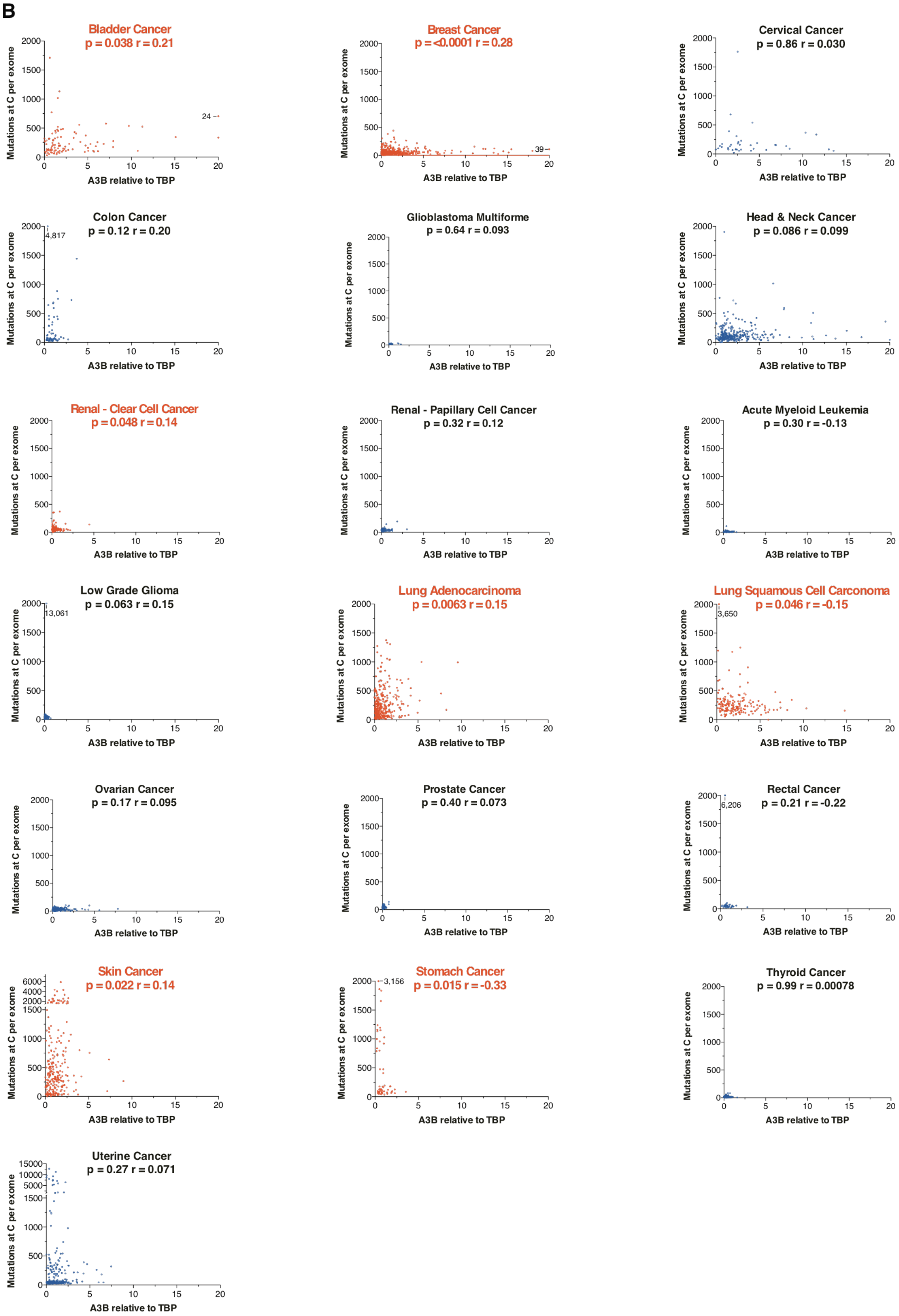

B
Bladder Cancer
p = 0.038 r = 0.21
Mutations at C per exome
2000
1500
1000
500
0
24
A3B relative to TBP
0
5
10
15
20
Breast Cancer
p = <0.0001 r = 0.28
Mutations at C per exome
2000
1500
1000
500
0
39
A3B relative to TBP
0
5
10
15
20
Cervical Cancer
p = 0.86 r = 0.030
Mutations at C per exome
2000
1500
1000
500
0
A3B relative to TBP
0
5
10
15
20
Colon Cancer
p = 0.12 r = 0.20
Mutations at C per exome
2000
1500
1000
500
0
4,817
A3B relative to TBP
0
5
10
15
20
Glioblastoma Multiforme
p = 0.64 r = 0.093
Mutations at C per exome
2000
1500
1000
500
0
A3B relative to TBP
0
5
10
15
20
Head & Neck Cancer
p = 0.086 r = 0.099
Mutations at C per exome
2000
1500
1000
500
0
A3B relative to TBP
0
5
10
15
20
Renal - Clear Cell Cancer
p = 0.048 r = 0.14
Mutations at C per exome
2000
1500
1000
500
0
A3B relative to TBP
0
5
10
15
20
Renal - Papillary Cell Cancer
p = 0.32 r = 0.12
Mutations at C per exome
2000
1500
1000
500
0
A3B relative to TBP
0
5
10
15
20
Acute Myeloid Leukemia
p = 0.30 r = -0.13
Mutations at C per exome
2000
1500
1000
500
0
A3B relative to TBP
0
5
10
15
20
Low Grade Glioma
p = 0.063 r = 0.15
Mutations at C per exome
2000
1500
1000
500
0
13,061
A3B relative to TBP
0
5
10
15
20
Lung Adenocarcinoma
p = 0.0063 r = 0.15
Mutations at C per exome
2000
1500
1000
500
0
A3B relative to TBP
0
5
10
15
20
Lung Squamous Cell Carconoma
p = 0.046 r = -0.15
Mutations at C per exome
2000
1500
1000
500
0
3,650
A3B relative to TBP
0
5
10
15
20
Ovarian Cancer
p = 0.17 r = 0.095
Mutations at C per exome
2000
1500
1000
500
0
A3B relative to TBP
0
5
10
15
20
Prostate Cancer
p = 0.40 r = 0.073
Mutations at C per exome
2000
1500
1000
500
0
A3B relative to TBP
0
5
10
15
20
Rectal Cancer
p = 0.21 r = -0.22
Mutations at C per exome
2000
1500
1000
500
0
6,206
A3B relative to TBP
0
5
10
15
20
Skin Cancer
p = 0.022 r = 0.14
Mutations at C per exome
6000
4000
2000
1500
1000
500
0
A3B relative to TBP
0
5
10
15
20
Stomach Cancer
p = 0.015 r = -0.33
Mutations at C per exome
2000
1500
1000
500
0
3,156
A3B relative to TBP
0
5
10
15
20
Thyroid Cancer
p = 0.99 r = 0.00078
Mutations at C per exome
2000
1500
1000
500
0
A3B relative to TBP
0
5
10
15
20
Uterine Cancer
p = 0.27 r = 0.071
Mutations at C per exome
15000
10000
5000
1500
1000
500
0
A3B relative to TBP
0
5
10
15
20

Figure 3-7. 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.

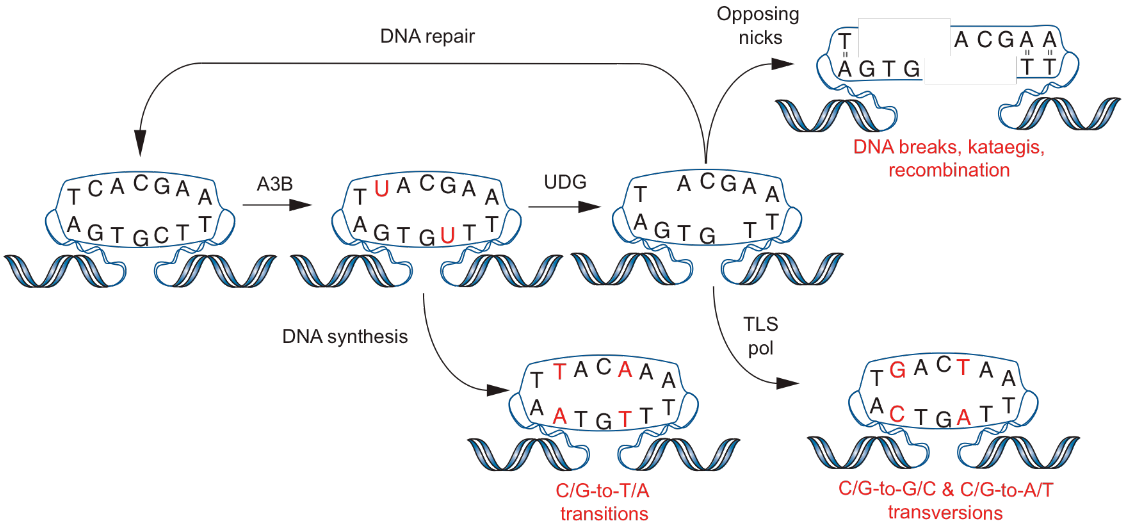

DNA repair
Opposing nicks
T C A C G A A
A G T G C T T
A3B
T U A C G A A
A G T G U T T
UDG
T A C G A A
A G T G T T
T A C G A A
A G T G T T
DNA breaks, kataegis, recombination
DNA synthesis
T T A C A A A
A A T G T T T
C/G-to-T/A transitions
TLS pol
T G A C T A A
A C T G A T T
C/G-to-G/C & C/G-to-A/T transversions

Figure 3-8. 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 Burns and Lackey, *et al.* [1], see also Figs. 1-1 and 2-19.

Table 3-1. Summary statistics for the 19 different tumor types in this study.

		A3B expression data[1]			Exome mutation data[2]				Clustered mutation data[3]		
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	0.06	170	5 - 15458	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	0.08	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 cystadenocarcinoma	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 - 8849	172	551	66	0.42	0.32
Lung adenocarcinoma	LUAD	355	0.0041 - 9.6	0.68	392	12 - 2547	259	355	310	0.79	0.73
Rectum adenocarcinoma	READ	72	0.082 - 3.2	0.81	88	28 - 7204	136	227	44	0.50	1.2
Colon adenocarcinoma	COAD	192	0.017 - 3.7	0.85	266	27 - 8459	250	487	133	0.50	0.39
Uterine corpus endometrioid carcinoma	UCEC	370	0.012 - 12	0.94	248	1 - 14687	68	722	1093	4.4	2.9
Skin cutaneous melanoma	SKCM	267	0.0011 - 10	1.1	255	6 - 6174	389	697	353	1.4	0.68
Bladder urotheilal carconoma	BLCA	122	0.0050 - 24	1.6	99	45 - 1802	226	291	293	3.0	3.5
Head & neck squamous cell carcinoma	HNSC	303	0.0038 - 20	1.7	306	7 - 2070	138	180	203	0.66	1.4
Lung squamous cell carcinoma	LUSC	259	0.094 - 15	1.7	177	1 - 3910	299	363	144	0.81	0.77
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	97	0.0010 - 20	2.4	39	30 - 1779	138	233	98	2.5	3.4

[1] A3B expression values relative to those of the housekeeping gene TBP by RNAseq.
[2] Somatic mutations in each exome, spanning aproximately 38 Mb of the human genome.
[3] Kataegis events from exome mutation data are defined as ≥2 cytosine mutations within 10kb intervals which meet Gordenin significance (see Methods).

Table 3-2. 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.

Table 3-3. 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	-

Table 3-4. 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%

CHAPTER 4 – CONCLUSIONS AND FUTURE QUESTIONS

APOBEC3B CONTRIBUTES TO BREAST CANCER MUTAGENESIS

For the APOBECs, as a class of DNA-editing enzymes resident in human tissue, the obvious implication is that their misregulation might be involved in the generating the genetic diversity seen in cancer. The first major question I faced as I began my studies on the APOBEC family and cancer was, which one(s)? Which of the family members might be responsible for driving genetic heterogeneity in cancer? Up to that point, there were clear signs that AID was likely a contributing factor in B-cell cancer[14] and that APOBEC1 might theoretically play a role in carcinogenesis[13], though, to date, no one has shown that it actually *is* implicated in any human malignancies. To determine a potential culprit(s) I quantified the full repertoire of *APOBEC* family mRNA species in a pilot set of human breast cancer tissues and cell lines. It became immediately clear that APOBEC3B was preferentially and specifically upregulated in a large proportion of the samples tested. This allowed subsequent efforts to be focused on elucidating the molecular mechanism by which this enzyme might operate in breast cancer.

Additional work revealed that among the APOBEC3 family members, APOBEC3B was the only enzyme that localizes to the cell nucleus[1,33,35]. Additionally, it retains deamination activity, increases the steady-state level of uracil in the cell's genome, and correlates with increased mutation, as determined by selection and enrichment techniques (*TK*-fluctuation assay and 3D-PCR/sequencing)[1]. These findings indicated that in a large proportion of breast cancer cell lines, APOBEC3B is driving mutations that diversify the genetic landscape.

Second, we incorporated publicly accessible gene expression and mutation data. The cancer genome atlas (TCGA) was the primary source of this information[52]. Up to this point, APOBEC3B's mechanistic role in cancer was clear, but we didn't have a genome-wide view of what this looked like in actual patient samples. Using data from TCGA, I was able to discover the correlation between APOBEC3B expression and mutation. This, thus far, is the only work on this topic that has so closely tied APOBEC3B together with patient samples at so many different levels, from the basic mechanism up through to patient genomes.

APOBEC3B AS A MUTATION SOURCE IN OTHER HUMAN CANCERS

At this point, it is clear that APOBEC3B contributes to breast cancer genome mutagenesis. The next question was to see whether there are other cancers that might be subject to the same process. To answer this question, I expanded my original analysis of TCGA breast cancer datasets to include 18 other cancer types[103]. The datasets we used included mutations in the exome and *APOBEC* mRNA expression levels. I was able to use to mutation coordinates to generate the trinucleotide context for each of the mutations present in each tumor sample. Mutations occurring at Cs were of particular interest, for obvious reasons, and were the focus of the much of the work. Each cancer was then assigned a mutation signature as a result of these contexts. This expanded the analysis to assess three different criteria: mutation load, APOBEC expression, and mutation signature. One reviewer was curious to see us assess mutation clusters using TCGA exome mutation data. Each cluster was defined as 2 or more mutations within a 10kb window. This determination was accomplished using a correction method that accounted for the amount of DNA sequence available within each sequence window[40]. This step was critical as it eliminated false positives that would contaminate the dataset as the result of random chance. Again, starting with only exomic mutation and APOBEC expression, we then had yet another facet of the data to use to examine the APOBEC mutator hypothesis in these cancers.

Overall, the finding from this examination of 19 different forms of cancer showed that *APOBEC3B* mRNA was upregulated in a set of cancers and that in a sub-set of these cancers, mutation load, APOBEC3B mutation signature, and mutation clusters all correlated. Of the 19 cancers, 6 appear to be subject to APOBEC3B mutation: head & neck, bladder, cervical, breast, and lung (both squamous cell and adenocarcinoma). This work has highlighted a previously unappreciated mutator and the specific cancer types in which it operates. Although further research is needed, these mechanistic and correlative findings will undoubtedly spur further studies on APOBEC3B-drive mutagenesis and ultimately how it might be leveraged to benefit cancer patients.

FUTURE QUESTIONS

What is driving APOBEC3B upregulation?

It is abundantly clear that there are several different cancers that express high levels of APOBEC3B relative to normal tissues. There is, thus, some mechanism at work specifically in the cancer to increase APOBEC3B expression, but what that factor(s) is is still unclear. As was mentioned in chapter 2, the APOBEC3B locus is not a mutation hotspot, nor are there SNPs present in the promoter region that might explain this upregulation (see Fig. 2-8). Copy-number analysis, likewise, does not explain this upregulation (see http://dbCRID.biolead.org).

The original discovery of APOBEC3B, at that time termed phorbolin-2, depended on increased expression of the gene by stimulation of keratinocytes with phorbol 12-myristate 13-acetate (PMA)[104,105]. This same effect is seen in breast cell lines as well as cell lines from other tissues[34]. Researchers have also reported increases in APOBEC3B expression by treating cell lines with interferon alpha[34]. The pathways that these compounds activate (either directly or via signaling cross-talk) are multiple and not clearly tied to any elements in the proximal promoter region of APOBEC3B (unpublished findings).

These findings imply that the upregulation of APOBEC3B is not likely a cis-effect arising from an alteration in the gene itself or the upstream regulatory region. Instead, it is likely that there is a misregulated or mutated signaling pathway that is driving APOBEC3B expression. In an attempt to leverage publicly available expression data to answer this question, my colleagues and I mined the RNAseq data from TCGA to determine which genes, if any, correlated either positively or negatively with APOBEC3B expression. The hope was that we might discover a member or members of a signaling pathway that might explain APOBEC3B's upregulation. Unfortunately, while the analysis was straightforward to perform, the results do not indicate any single pathway whose misregulation at the transcriptional level would explain why APOBEC3B levels are elevated.

While there is a wealth of RNAseq data (over 4500 tumors), this data only provides information limited to looking at pathways that themselves are regulated at the level of transcription. Thus, these data are not capable of detecting changes that occur post-translationally (*e.g.* phosphorylation events, *etc.*). TCGA does have some data that addresses this point. In addition to the RNAseq and exomic mutation datasets, there are also results available from reverse-phase protein arrays (RPPA). These RPPA data allow for determination of the levels of different proteins, and in many cases the levels of post-translationally modified forms of proteins. Unlike RNAseq, however, there is a limited number of proteins that are assayed as the technique requires that each target have an associated well-validated, specific monoclonal antibody. A correlational examination of this more-limited dataset was, likewise, unsuccessful, as no single pathway was consistently associated with elevated levels of APOBEC3B.

Does APOBEC3B expression impact patients?

This question is actually two-fold. There is a question of cancer incidence and there is a question of cancer progression. The two are not necessarily linked. The argument could be made that APOBEC3B upregulation in a pre-cancerous lesion could provide the mutational impetus to form a malignant tumor. An alternative hypothesis might be that APOBEC3B expression acts to somehow prevent cancer from forming by, perhaps restricting endogenous retroelements or restricting a hypothetical oncogenic virus[106-112]. There is an APOBEC3B deletion polymorphism circulating in the human population with deletion allele frequencies ranging from approximately 1% to 93%, dependent upon the biogeographical ancestry of the population examined[55,113]. One group used a small (<50 patients) Japanese cohort to assess breast cancer incidence and the APOBEC3B deletion polymorphism. Their results were not statistically significant, but trended toward an inverse correlation between APOBEC3B and breast cancer incidence[114]. Two other groups used much larger cohorts to assess the relationship between the deletion allele and breast cancer incidence. Unlike the Japanese study that collected data on the deletion allele frequency from normal healthy patients recruited into their study, these groups relied upon data from the 1000 genomes project to determine the frequency of the deletion allele within their cohorts[113-116]. These two larger studies determined that there was a significant increase in the APOBEC3B deletion allele among women who had breast cancer[115,116]. These findings argue that APOBEC3B is somehow a protective factor, reducing the incidence of breast cancer in the populations studied. This work is an excellent starting place to examine the role of APOBEC3B in carcinogenesis, but further work needs to be done. The two major critiques of this work are 1) reliance on the 1000 genomes project as a baseline and 2) lack of mechanistic explanation for the effect. The 1000 genomes project, if nothing else, provides an excellent insight into the genetic diversity among human populations[117], but there are concerns that even with the large number of genomes and geographical locations represented, the project may not have adequate sampling power to be used for rigorous analyses of all populations[118]. As for explanatory mechanism, at this point all we have is speculation.

CONCLUSION

Disease heterogeneity is one of the major factors continuing to hamper cancer treatment. The lofty, yet nebulous, goal of "curing cancer" is laudable, but a bit naïve in that there are hundreds, if not thousands, of distinct forms of cancer that plague the human population. A more appropriate goal would be to discover cure**s** for cancer**s**. Obviously, the tissue of origin plays a major part in differentiating these cancers from one another, but even within the same tissue there remains dramatic variability[39,103]. In some cases this variability can be explained by the events that initially gave rise to the malignancy. Clinicians have protocols in place to attempt to assess the source of the cancer, when they are able, because this information can be used to tailor the therapies that are used to treat the disease[119]. The logical conclusion for basic scientists is to focus efforts on uncovering as many different sources of variability that give rise to and drive cancer. These efforts are currently underway in the form of TCGA[52], the Global Cancer Genomics Consortium (GCGC)[120], the International Cancer Genome Consortium

(ICGC)[121], and many others. These groups are each using high-throughput techniques to determine what is actually happening in human cancer at the level of the genome and couple that to clinical information. This cataloging is the first step toward a fuller understanding of the mutation processes at work shaping cancers. This thesis describes using some of these data (in conjunction with laboratory experimentation) to discover a new mutational process.

The next steps will be to continue working to uncover what is controlling APOBEC3B expression, what role it plays in cancer incidence and progression in the implicated cancers, and critically, what can be done to improve patient outcome. The most direct method would be to inhibit the enzyme's deaminase activity using a small molecule. Alternative methods are easy to envisage as well. For instance, one could target the as yet unknown pathway that is driving APOBEC3B expression, again highlighting the need for further research on that specific topic. Conversely, as has been done for MMR-deficient HNPCC and *BRCA1/2*-mutant cancers, DNA repair pathways could be inhibited to attempt to kill cancers cell by way of synthetic lethality[119]. In either scenario, one that involves blocking APOBEC3B activity or expression or a strategy that uses the enzyme's mutagenic capacity against itself, the hope is that this basic research will be translated into successful clinical results that improve and extend patient lives.

REFERENCES

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 Refsland, E. W. *et al.* Quantitative profiling of the full APOBEC3 mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction. *Nucleic acids research* **38**, 4274-4284, doi:10.1093/nar/gkq174 (2010).

3 Di Noia, J. M. & Neuberger, M. S. Molecular mechanisms of antibody somatic hypermutation. *Annual review of biochemistry* **76**, 1-22, doi:10.1146/annurev.biochem.76.061705.090740 (2007).

4 Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646-674, doi:10.1016/j.cell.2011.02.013 (2011).

5 Berger, M. F. *et al.* Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature* **485**, 502-506, doi:10.1038/nature11071 (2012).

6 Wei, X. *et al.* Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nature genetics* **43**, 442-446, doi:10.1038/ng.810 (2011).

7 King, M. C., Marks, J. H., Mandell, J. B. & New York Breast Cancer Study, G. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* **302**, 643-646, doi:10.1126/science.1088759 (2003).

8 Conticello, S. G. The AID/APOBEC family of nucleic acid mutators. *Genome biology* **9**, 229, doi:10.1186/gb-2008-9-6-229 (2008).

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

10 LaRue, R. S. *et al.* Guidelines for naming nonprimate APOBEC3 genes and proteins. *Journal of virology* **83**, 494-497, doi:10.1128/JVI.01976-08 (2009).

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

12 Ding, Q. *et al.* APOBEC3G promotes liver metastasis in an orthotopic mouse model of colorectal cancer and predicts human hepatic metastasis. *The Journal of clinical investigation* **121**, 4526-4536, doi:10.1172/JCI45008 (2011).

13 Yamanaka, S. *et al.* Apolipoprotein B mRNA-editing protein induces hepatocellular carcinoma and dysplasia in transgenic animals. *Proceedings of the National Academy of Sciences of the United States of America* **92**, 8483-8487 (1995).

14 Ramiro, A. R. *et al.* AID is required for c-myc/IgH chromosome translocations in vivo. *Cell* **118**, 431-438, doi:10.1016/j.cell.2004.08.006 (2004).

15 Shinmura, K. *et al.* Aberrant expression and mutation-inducing activity of AID in human lung cancer. *Annals of surgical oncology* **18**, 2084-2092, doi:10.1245/s10434-011-1568-8 (2011).

16 Luo, Y., Walla, M. & Wyatt, M. D. Uracil incorporation into genomic DNA does not predict toxicity caused by chemotherapeutic inhibition of thymidylate synthase. *DNA repair* **7**, 162-169, doi:10.1016/j.dnarep.2007.09.001 (2008).

17 Kemmerich, K., Dingler, F. A., Rada, C. & Neuberger, M. S. Germline ablation of SMUG1 DNA glycosylase causes loss of 5-hydroxymethyluracil- and UNG-backup uracil-excision activities and increases cancer predisposition of Ung-/-Msh2-/- mice. *Nucleic acids research* **40**, 6016-6025, doi:10.1093/nar/gks259 (2012).

18 Nilsen, H. *et al.* Nuclear and mitochondrial uracil-DNA glycosylases are generated by alternative splicing and transcription from different positions in the UNG gene. *Nucleic acids research* **25**, 750-755 (1997).

19 Lindahl, T. An N-glycosidase from Escherichia coli that releases free uracil from DNA containing deaminated cytosine residues. *Proceedings of the National Academy of Sciences of the United States of America* **71**, 3649-3653 (1974).

20 Stivers, J. T. & Jiang, Y. L. A mechanistic perspective on the chemistry of DNA repair glycosylases. *Chemical reviews* **103**, 2729-2759, doi:10.1021/cr010219b (2003).

21 Germann, M. W., Johnson, C. N. & Spring, A. M. Recognition of damaged DNA: structure and dynamic markers. *Medicinal research reviews* **32**, 659-683, doi:10.1002/med.20226 (2012).

22 Tomkinson, A. E. *et al.* Completion of base excision repair by mammalian DNA ligases. *Progress in nucleic acid research and molecular biology* **68**, 151-164 (2001).

23 Dianov, G. L. & Hubscher, U. Mammalian base excision repair: the forgotten archangel. *Nucleic acids research* **41**, 3483-3490, doi:10.1093/nar/gkt076 (2013).

24 Hsieh, P. & Yamane, K. DNA mismatch repair: molecular mechanism, cancer, and ageing. *Mechanisms of ageing and development* **129**, 391-407, doi:10.1016/j.mad.2008.02.012 (2008).

25 Drotschmann, K. *et al.* DNA binding properties of the yeast Msh2-Msh6 and Mlh1-Pms1 heterodimers. *Biological chemistry* **383**, 969-975, doi:10.1515/BC.2002.103 (2002).

26 Majka, J. & Burgers, P. M. The PCNA-RFC families of DNA clamps and clamp loaders. *Progress in nucleic acid research and molecular biology* **78**, 227-260, doi:10.1016/S0079-6603(04)78006-X (2004).

27 Wardle, J. *et al.* Uracil recognition by replicative DNA polymerases is limited to the archaea, not occurring with bacteria and eukarya. *Nucleic acids research* **36**, 705-711, doi:10.1093/nar/gkm1023 (2008).

28 Pena-Diaz, J. *et al.* Noncanonical mismatch repair as a source of genomic instability in human cells. *Molecular cell* **47**, 669-680, doi:10.1016/j.molcel.2012.07.006 (2012).

29 Strauss, B. S. The 'A rule' of mutagen specificity: a consequence of DNA polymerase bypass of non-instructional lesions? *BioEssays : news and reviews in molecular, cellular and developmental biology* **13**, 79-84, doi:10.1002/bies.950130206 (1991).

30 Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979-993, doi:10.1016/j.cell.2012.04.024 (2012).

31 Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027-1038 (1993).

32 Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **77**, 1 p following 166 (1994).

33 Lackey, L. *et al.* APOBEC3B and AID have similar nuclear import mechanisms. *Journal of molecular biology* **419**, 301-314, doi:10.1016/j.jmb.2012.03.011 (2012).

34 Taylor, B. J. *et al.* DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *eLife* **2**, e00534, doi:10.7554/eLife.00534 (2013).

35 Land, A. M. *et al.* Endogenous APOBEC3A DNA Cytosine Deaminase Is Cytoplasmic and Nongenotoxic. *The Journal of biological chemistry* **288**, 17253-17260, doi:10.1074/jbc.M113.458661 (2013).

36 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 structural & molecular biology* **17**, 222-229, doi:10.1038/nsmb.1744 (2010).

37 Suspene, R. *et al.* Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 4858-4863, doi:10.1073/pnas.1009687108 (2011).

38 Shinohara, M. *et al.* APOBEC3B can impair genomic stability by inducing base substitutions in genomic DNA in human cells. *Scientific reports* **2**, 806, doi:10.1038/srep00806 (2012).

39 Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214-218, doi:10.1038/nature12213 (2013).

40 Roberts, S. A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Molecular cell* **46**, 424-435, doi:10.1016/j.molcel.2012.03.030 (2012).

41 Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214-220, doi:10.1038/nature09744 (2011).

42 Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-158, doi:10.1038/nature05610 (2007).

43 Jones, S. *et al.* Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* **330**, 228-231, doi:10.1126/science.1196333 (2010).

44 Kumar, A. *et al.* Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 17087-17092, doi:10.1073/pnas.1108745108 (2011).

45 Parsons, D. W. *et al.* The genetic landscape of the childhood cancer medulloblastoma. *Science* **331**, 435-439, doi:10.1126/science.1198056 (2011).

46 Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-274, doi:10.1126/science.1133427 (2006).

47 Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400-404, doi:10.1038/nature11017 (2012).

48 Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-1160, doi:10.1126/science.1208130 (2011).

49 Pavri, R. & Nussenzweig, M. C. AID targeting in antibody diversity. *Advances in immunology* **110**, 1-26, doi:10.1016/B978-0-12-387663-8.00005-3 (2011).

50 Albin, J. S. & Harris, R. S. Interactions of host APOBEC3 restriction factors with HIV-1 in vivo: implications for therapeutics. *Expert reviews in molecular medicine* **12**, e4, doi:10.1017/S1462399409001343 (2010).

51 Landry, S., Narvaiza, I., Linfesty, D. C. & Weitzman, M. D. APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. *EMBO reports* **12**, 444-450, doi:10.1038/embor.2011.46 (2011).

52 Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

53 Zhang, J. *et al.* International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. *Database : the journal of biological databases and curation* **2011**, bar026, doi:10.1093/database/bar026 (2011).

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

55 Kidd, J. M., Newman, T. L., Tuzun, E., Kaul, R. & Eichler, E. E. Population stratification of a common APOBEC gene deletion polymorphism. *PLoS genetics* **3**, e63, doi:10.1371/journal.pgen.0030063 (2007).

56 Chen, H. *et al.* APOBEC3A is a potent inhibitor of adeno-associated virus and retrotransposons. *Current biology : CB* **16**, 480-485, doi:10.1016/j.cub.2006.01.031 (2006).

57 Peng, G. *et al.* Myeloid differentiation and susceptibility to HIV-1 are linked to APOBEC3 expression. *Blood* **110**, 393-400, doi:10.1182/blood-2006-10-051763 (2007).

58 Kong, F. *et al.* dbCRID: a database of chromosomal rearrangements in human diseases. *Nucleic acids research* **39**, D895-900, doi:10.1093/nar/gkq1038 (2011).

59 Edwards, P. A. Fusion genes and chromosome translocations in the common epithelial cancers. *The Journal of pathology* **220**, 244-254, doi:10.1002/path.2632 (2010).

60 Carpenter, M. A. *et al.* Methylcytosine and normal cytosine deamination by the foreign DNA restriction enzyme APOBEC3A. *The Journal of biological chemistry* **287**, 34801-34808, doi:10.1074/jbc.M112.385161 (2012).

61 Shlyakhtenko, L. S. *et al.* Atomic force microscopy studies provide direct evidence for dimerization of the HIV restriction factor APOBEC3G. *The Journal of biological chemistry* **286**, 3387-3395, doi:10.1074/jbc.M110.195685 (2011).

62 Lea, D. E. & Coulson, C. A. The distribution of the numbers of mutants in bacterial populations. *Journal of Genetics* **49**, 264-285, doi:10.1007/BF02986080 (1949).

63 Di Noia, J. & Neuberger, M. S. Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. *Nature* **419**, 43-48, doi:10.1038/nature00981 (2002).

64 Huang, X. & Darzynkiewicz, Z. Cytometric assessment of histone H2AX phosphorylation: a reporter of DNA damage. *Methods in molecular biology* **314**, 73-80, doi:10.1385/1-59259-973-7:073 (2006).

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

66 Tripathi, A. *et al.* Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients. *International journal of cancer. Journal international du cancer* **122**, 1557-1566, doi:10.1002/ijc.23267 (2008).

67 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, doi:10.1038/sj.bjc.6605576 (2010).

68 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. *Clinical cancer research : an official journal of the American Association for Cancer Research* **16**, 5351-5361, doi:10.1158/1078-0432.CCR-10-1265 (2010).

69 Lasham, A. *et al.* YB-1, the E2F pathway, and regulation of tumor cell growth. *Journal of the National Cancer Institute* **104**, 133-146, doi:10.1093/jnci/djr512 (2012).

70 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 informatics* **10**, 233-247, doi:10.4137/CIN.S8104 (2011).

71 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. *Computer methods and programs in biomedicine* **104**, e133-147, doi:10.1016/j.cmpb.2011.03.004 (2011).

72 Nikas, J. B. & Low, W. C. ROC-supervised principal component analysis in connection with the diagnosis of diseases. *American journal of translational research* **3**, 180-196 (2011).

73 Nikas, J. B. & Low, W. C. Linear Discriminant Functions in Connection with the micro-RNA Diagnosis of Colon Cancer. *Cancer informatics* **11**, 1-14, doi:10.4137/CIN.S8779 (2012).

74 Nikas, J. B., Low, W. C. & Burgio, P. A. Prognosis of treatment response (pathological complete response) in breast cancer. *Biomarker insights* **7**, 59-70, doi:10.4137/BMI.S9387 (2012).

75 Stephens, P. *et al.* A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nature genetics* **37**, 590-592, doi:10.1038/ng1571 (2005).

76 Makridakis, N. M. & Reichardt, J. K. Translesion DNA polymerases and cancer. *Frontiers in genetics* **3**, 174, doi:10.3389/fgene.2012.00174 (2012).

77 Drier, Y. *et al.* Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome research* **23**, 228-235, doi:10.1101/gr.141382.112 (2013).

78 Longerich, S., Basu, U., Alt, F. & Storb, U. AID in somatic hypermutation and class switch recombination. *Current opinion in immunology* **18**, 164-174, doi:10.1016/j.coi.2006.01.008 (2006).

79 Harris, R. S., Hultquist, J. F. & Evans, D. T. The restriction factors of human immunodeficiency virus. *The Journal of biological chemistry* **287**, 40875-40883, doi:10.1074/jbc.R112.416925 (2012).

80 Malim, M. H. APOBEC proteins and intrinsic resistance to HIV-1 infection. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **364**, 675-687, doi:10.1098/rstb.2008.0185 (2009).

81 Bishop, K. N., Holmes, R. K., Sheehy, A. M. & Malim, M. H. APOBEC-mediated editing of viral RNA. *Science* **305**, 645, doi:10.1126/science.1100658 (2004).

82 Blanc, V. & Davidson, N. O. C-to-U RNA editing: mechanisms leading to genetic diversity. *The Journal of biological chemistry* **278**, 1395-1398, doi:10.1074/jbc.R200024200 (2003).

83 Ikeda, T. *et al.* Intrinsic restriction activity by apolipoprotein B mRNA editing enzyme APOBEC1 against the mobility of autonomous retrotransposons. *Nucleic acids research* **39**, 5538-5554, doi:10.1093/nar/gkr124 (2011).

84 Petit, V. *et al.* Murine APOBEC1 is a powerful mutator of retroviral and cellular RNA in vitro and in vivo. *Journal of molecular biology* **385**, 65-78, doi:10.1016/j.jmb.2008.10.043 (2009).

85 Chelico, L., Pham, P., Calabrese, P. & Goodman, M. F. APOBEC3G DNA deaminase acts processively 3' --> 5' on single-stranded DNA. *Nature structural & molecular biology* **13**, 392-399, doi:10.1038/nsmb1086 (2006).

86 Hultquist, J. F. *et al.* Human and rhesus APOBEC3D, APOBEC3F, APOBEC3G, and APOBEC3H demonstrate a conserved capacity to restrict Vif-deficient HIV-1. *Journal of virology* **85**, 11220-11234, doi:10.1128/JVI.05238-11 (2011).

87 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, doi:10.1038/nature00862 (2002).

88 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). *The Journal of biological chemistry* **278**, 19583-19586, doi:10.1074/jbc.C300114200 (2003).

89 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, doi:10.1038/nature01760 (2003).

90 Robbiani, D. F. & Nussenzweig, M. C. Chromosome translocation, B cell lymphoma, and activation-induced cytidine deaminase. *Annual review of pathology* **8**, 79-103, doi:10.1146/annurev-pathol-020712-164004 (2013).

91 Okazaki, I. M. *et al.* Constitutive expression of AID leads to tumorigenesis. *The Journal of experimental medicine* **197**, 1173-1181, doi:10.1084/jem.20030275 (2003).

92 Jarmuz, A. *et al.* An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics* **79**, 285-296, doi:10.1006/geno.2002.6718 (2002).

93 Koning, F. A. *et al.* Defining APOBEC3 expression patterns in human tissues and hematopoietic cell subsets. *Journal of virology* **83**, 9474-9485, doi:10.1128/JVI.01089-09 (2009).

94 Kohli, R. M. *et al.* Local sequence targeting in the AID/APOBEC family differentially impacts retroviral restriction and antibody diversification. *The Journal of biological chemistry* **285**, 40956-40964, doi:10.1074/jbc.M110.177402 (2010).

95 Wang, M., Rada, C. & Neuberger, M. S. Altering the spectrum of immunoglobulin V gene somatic hypermutation by modifying the active site of AID. *The Journal of experimental medicine* **207**, 141-153, doi:10.1084/jem.20092238 (2010).

96 Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature genetics* **45**, 136-144, doi:10.1038/ng.2503 (2013).

97 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

98 Fujino, T., Navaratnam, N. & Scott, J. Human apolipoprotein B RNA editing deaminase gene (APOBEC1). *Genomics* **47**, 266-275, doi:10.1006/geno.1997.5110 (1998).

99 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. *The Journal of biological chemistry* **274**, 18470-18476 (1999).

100 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).

101 Sato, Y. *et al.* Deficiency in APOBEC2 leads to a shift in muscle fiber type, diminished body mass, and myopathy. *The Journal of biological chemistry* **285**, 7111-7118, doi:10.1074/jbc.M109.052977 (2010).

102 Rada, C., Jarvis, J. M. & Milstein, C. AID-GFP chimeric protein increases hypermutation of Ig genes with no evidence of nuclear localization. *Proceedings*

of the National Academy of Sciences of the United States of America **99**, 7003-7008, doi:10.1073/pnas.092160999 (2002).

103 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).

104 Madsen, P. *et al.* Psoriasis upregulated phorbolin-1 shares structural but not functional similarity to the mRNA-editing protein apobec-1. *The Journal of investigative dermatology* **113**, 162-169, doi:10.1046/j.1523-1747.1999.00682.x (1999).

105 Rasmussen, H. H. & Celis, J. E. Evidence for an altered protein kinase C (PKC) signaling pathway in psoriasis. *The Journal of investigative dermatology* **101**, 560-566 (1993).

106 Bogerd, H. P., Wiegand, H. L., Doehle, B. P. & Cullen, B. R. The intrinsic antiretroviral factor APOBEC3B contains two enzymatically active cytidine deaminase domains. *Virology* **364**, 486-493, doi:10.1016/j.virol.2007.03.019 (2007).

107 Doehle, B. P., Schafer, A. & Cullen, B. R. Human APOBEC3B is a potent inhibitor of HIV-1 infectivity and is resistant to HIV-1 Vif. *Virology* **339**, 281-288, doi:10.1016/j.virol.2005.06.005 (2005).

108 McDougle, R. M., Hultquist, J. F., Stabell, A. C., Sawyer, S. L. & Harris, R. S. D316 is critical for the enzymatic activity and HIV-1 restriction potential of human and rhesus APOBEC3B. *Virology* **441**, 31-39, doi:10.1016/j.virol.2013.03.003 (2013).

109 Muckenfuss, H. *et al.* APOBEC3 proteins inhibit human LINE-1 retrotransposition. *The Journal of biological chemistry* **281**, 22161-22172, doi:10.1074/jbc.M601716200 (2006).

110 Pak, V., Heidecker, G., Pathak, V. K. & Derse, D. The role of amino-terminal sequences in cellular localization and antiviral activity of APOBEC3B. *Journal of virology* **85**, 8538-8547, doi:10.1128/JVI.02645-10 (2011).

111 Stenglein, M. D. & Harris, R. S. APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. *The Journal of biological chemistry* **281**, 16837-16841, doi:10.1074/jbc.M602367200 (2006).

112 Yu, Q. *et al.* APOBEC3B and APOBEC3C are potent inhibitors of simian immunodeficiency virus replication. *The Journal of biological chemistry* **279**, 53379-53386, doi:10.1074/jbc.M408802200 (2004).

113 Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

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

115 Long, J. *et al.* A common deletion in the APOBEC3 genes and breast cancer risk. *Journal of the National Cancer Institute* **105**, 573-579, doi:10.1093/jnci/djt018 (2013).

116 Xuan, D. *et al.* APOBEC3 deletion polymorphism is Associated with Breast Cancer Risk among women of European Ancestry. *Carcinogenesis*, doi:10.1093/carcin/bgt185 (2013).

117 Kuehn, B. M. 1000 Genomes Project finds substantial genetic variation among populations. *JAMA : the journal of the American Medical Association* **308**, 2322, 2325, doi:10.1001/jama.2012.88674 (2012).

118 Lu, D. & Xu, S. Principal component analysis reveals the 1000 Genomes Project does not sufficiently cover the human genetic diversity in Asia. *Frontiers in genetics* **4**, 127, doi:10.3389/fgene.2013.00127 (2013).

119 Lord, C. J. & Ashworth, A. The DNA damage response and cancer therapy. *Nature* **481**, 287-294, doi:10.1038/nature10760 (2012).

120 Global Cancer Genomics, C. The Global Cancer Genomics Consortium: interfacing genomics and cancer medicine. *Cancer research* **72**, 3720-3724, doi:10.1158/0008-5472.CAN-12-1054 (2012).

121 International Cancer Genome, C. *et al.* International network of cancer genome projects. *Nature* **464**, 993-998, doi:10.1038/nature08987 (2010).