HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2016 March 01.

Published in final edited form as:

Nat Genet. 2015 September ; 47(9): 1067–1072. doi:10.1038/ng.3378.

An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers

Kin Chan[1], **Steven A. Roberts**[1,2], **Leszek J. Klimczak**[3], **Joan F. Sterling**[1], **Natalie Saini**[1], **Ewa P. Malc**[4], **Jaegil Kim**[5], **David J. Kwiatkowski**[5,6], **David C. Fargo**[3], **Piotr A. Mieczkowski**[4], **Gad Getz**[5,7], and **Dmitry A. Gordenin**[1]

[1]Genome Integrity & Structural Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC

[2]School of Molecular Biosciences, Washington State University, Pullman, WA

[3]Integrative Bioinformatics Support Group, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC

[4]Department of Genetics, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC

[5]Broad Institute of MIT and Harvard, Cambridge, MA

[6]Brigham and Women's Hospital, Harvard Medical School, Boston, MA

[7]Massachusetts General Hospital, Harvard Medical School, Boston, MA

Corresponding Author: Dmitry A. Gordenin (gordenin@niehs.nih.gov).

Author Contributions

KC and DAG designed the study. KC, SAR, JFS, NS, EPM, and PAM performed the experiments. KC, LJK, JK, DCF, and GG performed statistical analyses. KC, LJK, NS, JK, DJK, DCF, GG, and DAG analyzed the data. SAR, LJK, and PAM contributed reagents, materials, or analysis tools. KC and DAG wrote the paper, with contributions from SAR, NS, DJK, and GG.

Competing Financial Interests

The authors declare no competing financial interests.

URLs

Broad Genome Data Analysis Center (GDAC) Firehose: http://gdac.broadinstitute.org/
Broad GDAC Firehose standard data run (Feb. 15, 2014): http://gdac.broadinstitute.org/runs/stddata__2014_02_15/
dbGaP: https://www.ncbi.nlm.nih.gov/gap
dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP/
GenBank: http://www.ncbi.nlm.nih.gov/genbank/
NCBI Sequence Read Archive (SRA): http://www.ncbi.nlm.nih.gov/Traces/sra/
pLogo: http://plogo.uconn.edu/
R statistical computing package: http://www.r-project.org/
sacCer3 reference genome: http://hgdownload-test.cse.ucsc.edu/goldenPath/sacCer3/bigZips/
UCSC Genome Browser: http://genome.ucsc.edu/

Accession Codes

dbGaP TCGA controlled access Data Portal: phs000178.v8.p7
ySR127 reference genome (GenBank): CP011547–CP011563
Illumina reads (NCBI SRA): SRP056805

Elucidation of mutagenic processes shaping cancer genomes is a fundamental problem whose solution promises insights into new treatment, diagnostic, and prevention strategies[1]. Single-strand DNA-specific APOBEC cytidine deaminase(s) are major source(s) of mutations in several cancer types[2-4]. Previous indirect evidence implicated APOBEC3B as the more likely major mutator deaminase, while APOBEC3A's role is not established[5,6]. Using yeast models enabling controlled generation of long single-strand genomic DNA substrates[7], we show the mutation signatures of APOBEC3A and APOBEC3B are statistically distinguishable. We then apply three complementary approaches to identify cancer samples with mutation signatures resembling either APOBEC. Strikingly, APOBEC3A-like samples have over ten-fold more APOBEC-signature mutations than APOBEC3B-like samples. We propose that APOBEC3A mutagenesis is much stronger because APOBEC3A itself is highly proficient at generating DNA breaks[8-10], whose repair can trigger formation of single-strand hypermutation substrates.

Recently, we and others have shown that some cancers have an abundance of apparently simultaneous, closely-spaced mutations, variously referred to as 'kataegis'[11] and 'mutation clusters'[12]. Many clusters are strand- and nucleotide-coordinated, consisting entirely of mutations at cytosines on one DNA strand, most frequently within 5′-TCW-3′ motifs (mutated cytosine as capital underlined C; W denotes adenine or thymine)[4]. These characteristics are consistent with the mutagenic properties of several APOBEC cytidine deaminases which target 5′-TC-3′ motifs in single-strand DNA (ssDNA)[13-17].

Analyses of cancer mutation datasets have implicated APOBEC3B (A3B) as the leading candidate[5,6], with APOBEC3A (A3A) as another possible mutator[4]. Numerous recent reports have linked high A3B expression to various cancers, reflecting a widely held view that A3B is the likely major mutator[3,4,9,18-22]. On the other hand, there is evidence that A3A could be a mutator in cancers[8-10,23,24]. Consistent with this possibility, breast cancers from carriers of a germline A3B deletion allele, fusing the A3A transcript to A3B's 3′ regulatory sequences, actually tend to have higher TC-signature mutation loads than cancers from non-carriers[25]. Such fusion transcripts are more stable, resulting in higher steady-state levels of A3A enzyme[26].

A more conclusive way to distinguish between possible sources of mutagenesis in cancers is to match mutation signatures extracted from statistical analysis of each cancer with well-defined signature(s) of candidate mutagen(s)[27,28]. Thus, we collected large numbers of mutations induced by either A3A or A3B, in a yeast reporter strain (deleted for uracil glycosylase) that generates chromosomal ssDNA upon temperature shift[7]. Telomere uncapping in the presence of ssDNA-damaging mutagens results in selectable mutation clusters inactivating multiple reporter genes[7,29]. Crucially, resection of the complementary strand precludes excision repair and uracils from cytidine deaminations gave rise to C → T transitions[7]. pLogo analysis[30] of mutations identified by whole genome sequencing (WGS) of yeast revealed almost diametrically opposite motif preferences: A3A favored YTCA, while A3B favored RTCA (Y = pyrimidine, R = purine, see Fig. 1a–f and Supplementary Table 1). This was corroborated by our fold enrichment methodology[12] (see Fig. 1g, 1h). Re-analysis of mutation data from Neuberger and colleagues, generated by expressing A3A or A3B in a conventional yeast system[31], yielded similar results (see Supplementary Fig. 1).

The motif preferences of APOBECs in yeast should be suitable models for the enzymes' preferences in human cells, since the local sequence contexts flanking cytosines in both species' genomes are quite similar, except for depletion of CpG motifs in human[31].

pLogos also showed that mutations at TCA (a component of TCW) were overrepresented for both APOBECs, while TCT was underrepresented. Then TCA enrichment in cancers should exceed TCW enrichment, if TC mutations were caused by either A3A or A3B. We evaluated 15 cohorts of recently published cancer WGS samples[25,32]. Five cancer types[2-4] (six cohorts, see Fig. 2) had high rates of APOBEC-signature mutagenesis: bladder (BLCA), breast (BRCA), head and neck (HNSC), lung adenocarcinoma (LUAD) and squamous cell (LUSC). In BLCA, BRCA, and HNSC, high TCA enrichment for APOBEC mutations was clearly evident. APOBEC mutagenesis was also detectable in LUAD and LUSC, as shown by high TCA enrichment values in C-coordinated mutation clusters, despite high genome-wide mutation loads from non-APOBEC sources. Low-APOBEC mutagenesis cancer types (e.g. multiple myeloma[33], where only a small percentage of samples exhibit significant APOBEC mutagenesis) are included in Supplementary Table 2.

We next examined the relationship between TCW and TCA enrichments on a per-sample basis for the six high-APOBEC cohorts (see Fig. 2). Results for low-APOBEC cohorts are in Supplementary Figure 2. We ordered all samples within each cohort by ascending TCW enrichment and binned into quartiles (see Fig. 2). As TCW enrichment increased, there was a statistically significant trend toward samples with TCA enrichment > TCW enrichment (χ^2 test for trend p-values in Fig. 2). This suggested that A3A, A3B, or both, were mutagenizing cancers with high TCW enrichment. Similar results were obtained when analyzing exomes from BLCA, BRCA, HNSC, LUAD, LUSC, and cervical cancer (CESC) (see Supplementary Fig. 2), bolstering the conclusion that APOBEC(s) preferentially targeting TCA were acting in many cancers. APOBEC-signature mutation load in cancer exomes are statistically correlated with A3B and A3A transcript abundance[4,9]. However, this was not a reliable metric for distinguishing mutagenicity of specific APOBECs within these cancer genomes (see Supplementary Fig. 3), possibly because mRNA abundance in excised tumors need not correlate with mRNA (or protein) abundance at time of mutagenesis.

We next sub-categorized TCA-enriched samples into A3A- and A3B-like subsets by comparing YTCA enrichment vs. RTCA enrichment (see Fig. 3). Samples with non-random ratio of YTCA vs. RTCA mutations (see "Y/RTCA enrichment analysis" in Online Methods) were binned by quartile of TCA enrichment. χ^2 tests for trend (p-values in Fig. 3) indicated significant skewing toward A3A-like signatures as TCA enrichment increased. Results for other cohorts are in Supplementary Figure 4. We estimated the minimal number of TCA mutations attributable to an APOBEC in each sample (see Fig. 3g and Supplementary Table 3), which revealed the overall A3A-like median value (1,480) was over 11-fold greater than the A3B-like median (133). Thus, A3A is a much more prolific mutator than A3B.

To verify these findings, we compared proportions of mutations at each NTCA in cancers vs. each yeast model, using root mean square deviation (RMSD) calculations (see "NTCA proportion analysis" in Online Methods), and generated corresponding pLogos for the

BRCA ICGC cohort (see Fig. 4). Results for five other high-APOBEC WGS cohorts are in Supplementary Figure 5. NTCA and pLogo analyses concurred with Y/RTCA results: lower TCA enrichment quartile samples were usually A3B-like (smaller RMSD vs. A3B model), transitioning to A3A-like samples (smaller RMSD vs. A3A model) in the upper quartiles.

Recent publications reported that A3B germline deletion carriers are at higher risk for breast cancer[34,35], and tumors from these patients have higher APOBEC-signature mutagenesis[25]. Thus, we investigated possible relationships between A3B germline copy number variation and prevalence of A3A- or A3B-like mutation signatures. By all three analyses, A3B deletion samples from the BRCA ICGC cohort were predominantly A3A-like (see Fig. 5a). In contrast, A3B wild-type samples showed a roughly equal split between A3A- (Fig. 5b) and A3B-like (Fig. 5c) signatures. Fisher's exact tests ($p = 0.0024$ by Y/RTCA and $p = 0.0277$ by NTCA analyses) confirmed significant skewing toward A3A-like signatures among A3B deletion samples. Similar results were obtained when the other high-APOBEC cohorts were evaluated (see Supplementary Fig. 6).

Our results (summarized in Fig. 6) strongly suggest that, in general, A3A is the predominant mutagenic deaminase in cancers. In cancers, APOBEC signatures were clearly detectable because abasic sites from uracil excision in ssDNA were not repaired. Instead, they were likely bypassed by error-prone translesion DNA polymerases to create mutations (see ref. 36 and references therein). Our approach relies on the supposition that, with respect to the motif preferences of APOBECs, cytosines in yeast ssDNA are suitable models for cytosines in ssDNA of human cancers. Since the molecular machinery of DNA transactions are not identical between the two species, we do not rule out the possibility that APOBEC motif preferences might be at least somewhat different between yeast and human. As sequencing technologies mature, it should become feasible to put this question to a rigorous test, by analyzing APOBEC motif preferences at thousands of mutated cytosines in human tissue culture models and comparing to our results in yeast.

The finding that A3A-signature mutagenesis is more prominent in cancers might seem surprising, since A3B mRNA abundance tends to be higher than A3A's in cancer samples (see Supplementary Fig. 3). However, A3A is a much more potent inducer of DNA damage, likely via strand breakage as demonstrated by staining for γ-H2AX (a marker for double-strand breaks) and/or comet assay[8-10,23]. This is also consistent with observations that APOBEC-signature mutations and clusters are frequently co-localized with rearrangement breakpoints in cancers[11,12,37]. We propose that A3A-signature mutagenesis is more prominent, at least in part, because A3A itself can trigger homology-directed repair mediated generation of ssDNA substrates (by end resection[38] or break-induced replication[39]) much more readily than A3B can.

As clinical cancer genetics progresses toward genomic analysis of each cancer sample, we have recently integrated sample-specific APOBEC-signature mutation analysis into a standard platform for analysis of large cancer genome datasets[40-42]. Analyses to distinguish between A3A- vs. A3B-like signatures will be incorporated into future pipeline updates, since this might prove important when weighing treatment options, given the substantially higher genotoxic and mutagenic potential of A3A. Moreover, early detection of APOBEC-

signature mutation enrichment, e.g. in cell-free circulating DNA, could have important diagnostic or prognostic value, especially for individuals at higher risk, such as A3B deletion carriers.

When detected in a tumor sample, a high prevalence of APOBEC mutagenesis might be exploited for therapeutic purposes. It has been suggested that hypermutation could enhance the effectiveness of immune stimulation therapy to treat cancer, by generating tumor-specific neoantigens (proteins with new epitopes), that might trigger targeted destruction by the immune system[43,44]. There are two immune therapies for bladder cancers[45,46], which often have high APOBEC enrichment (see Fig. 2) and A3A-like signatures (see Fig. 3 and Supplementary Fig. 5). These clinical observations raise the intriguing possibility that hypermutation in bladder cancers (mainly by A3A) could contribute substantially to the success of immune therapies. Likewise, other A3A-like, high-APOBEC mutagenesis cancers could be promising candidates for similar immune stimulation treatments.

Online Methods

Construction of integrated A3A- and A3B-expressing yeast strains

Human A3A or A3B open read frames (ORFs) with appended 5′ ClaI and 3′ StuI restriction sites were codon optimized for expression in yeast, and purchased from DNA 2.0 as inserts within the pJ201 vector. Each ORF was released from the vector backbone by ClaI and StuI double digestion, and ligated into the multi-cloning site of a tetracycline-regulatable pCM252-derived vector[47], to create plasmids pSR435 (bearing A3A) and pSR440 (A3B) with hph (hygromycin resistance) as the selectable marker instead of *TRP1*. A fragment of each plasmid containing the APOBEC ORF, the tetracycline-regulated promoter, and the hph marker, was amplified by PCR with primers (see Supplementary Table 4 for primer sequences) to add flanks with homology to either side of the *LEU2* gene on Chromosome III.

Purified PCR product was transformed[48] into a yeast host strain descended from CG379[49], with the following genotype: *MATα his7-2 leu2-3,112 trp1-289 cdc13-1 ung1::NAT*. *CAN1*, *URA3*, and *ADE2* were deleted from their native loci and reintroduced into a closely-spaced triple reporter gene array near the de novo telomere on the left arm of Chromosome V[7]. Transformants with an APOBEC-hph cassette stably integrated into the *LEU2* locus target (by homologous recombination) were selected by replica plating onto hygromycin plates, and verified by diagnostic replicas on single-colony isolates, followed by DNA sequencing of the insert.

Mutagenesis by A3A and A3B in yeast

Yeast were inoculated into 5 mL of YPDA media (1% yeast extract, 2% peptone, 2% dextrose, 0.01% adenine sulfate, filter-sterilized) and grown at 23°C for 72 hours. Yeast then were diluted ten-fold into 5 mL of fresh YPDA with 20 μg/mL doxycycline hyclate (Sigma-Aldrich) and shifted to 37°C for 6 hours. Cells then were washed into 5 mL of phosphate-buffered saline and held at 37°C for 42 hours more. Appropriate dilutions were plated onto synthetic complete to verify viability, and onto arginine dropout plates with 60 mg/mL

canavanine sulfate and 20 mg/mL adenine sulfate to identify Can^r Ade^- double mutants, i.e., colonies with mutation clusters.

Whole-genome sequencing of yeast

Yeast colonies with mutation clusters were streaked onto YPDA. A single-colony isolate from each streak was verified for Can, Ura, Ade, and respiratory competency phenotypes by replica plating. Genomic DNA was purified from isolates of interest using a QIAcube robot, per manufacturer's instructions (QIAGEN). 100-nucleotide paired-end reads were obtained from a HiSeq 2000 sequencer (Illumina). Reads were mapped to the ySR127 reference genome and mutations were identified using the fixed ploidy caller in CLC Genomics Workbench 7.5 (QIAGEN). To minimize the possibility of analyzing mutations that were accumulated during routine passaging and culture growth, only unique mutations were included in mutation signature analyses. Illumina reads were uploaded to the NCBI Sequence Read Archive.

Cancer and other yeast sequencing data

Cancer genome and exome datasets were obtained from publications[25,32] or from the dbGaP TCGA controlled access Data Portal. hg19 was the human genome reference for our analyses. Cancer mutation catalogues were filtered to remove calls that overlapped with entries in dbSNP or the UCSC Genome Browser simpleRepeat track. Data from multiple myeloma genomes were from[12]. Additional yeast data were obtained from[31] and re-analyzed, as described in detail below and previously in[4], using the sacCer3 reference genome. Only mutations from the *ung1*Δ background were analyzed, as these were the closest equivalents to our yeast data.

APOBEC mRNA abundance and A3B germline copy number data

APOBEC RNAseq data for 5,868 tumor and 834 normal samples across 17 cancer types (bladder, breast, cervical, colorectal, glioblastoma multiforme, head and neck, kidney chromophobe and renal clear cell, acute myeloid leukemia, lower grade glioma, lung adenocarcinoma and squamous cell carcinoma, ovarian, prostate, melanoma, thyroid, and uterine corpus endometrial) were downloaded from the Broad GDAC Firehose standard data run of Feb. 15, 2014. Segmented copy number (CN) data for 7,191 tumor-normal pairs from these same cancer types were downloaded also. 5,526 samples had both RNAseq and CN data. These data were available for 17 bladder, 95 breast, 25 head and neck, 44 lung adenocarcinoma, and 44 lung squamous cell genomic samples (225 total), which allowed mRNA abundance vs. T<u>C</u>A minimal mutation load correlation, and mutation signature vs. A3B CN, analyses in this study. A3B CN data for the breast cancer ICGC cohort were obtained from[25].

A3B copy number annotation

Examination of the segmented CN data revealed that most A3B germline deletion events were localized between chr22: 39,363,650 and 39,375,350. Some samples had a short deletion within, or multiple discontinuous segmentation events overlapping, this region. This necessitated binning of the region into twelve 1-kb windows and identification of all

segmental copy number variation (CNV) events overlapping any window. Cutoffs for classification were determined by examination of the histogram of inferred A3B CN values (see Supplementary Fig. 6f): A3B CN ≤ 0.7, homozygous deletion (homo.del); 0.7 < A3B CN ≤ 1.69, heterozygous deletion (het.del); 1.69 < A3B CN ≤ 2.29, wild type (WT); and A3B CN > 2.29, amplification (amp). 7,061 samples each had a unique segmental CNV. Among the remaining 130 samples that had more than one segmental CNV, classification was based on the segmental CN farthest removed from the wild-type value of 2. CN call totals were: 99 homo.del (1.38%), 998 het.del (13.88%), 5699 WT (79.25%), and 395 amp (5.49%).

Mutation cluster analysis

Mutation cluster analysis was performed as described previously[4,12]. Mutations spaced ≤ 10 bases apart were treated as a single mutagenic event, since low fidelity translesion DNA synthesis polymerases often synthesize a short tract 3′ of lesion bypass, and mis-incorporate at high frequencies[50,51]. Groups of closely-spaced mutations were identified, such that any pair of adjacent mutations within each group was separated by less than 10 kb. To identify clusters that were unlikely to have formed by random distribution of mutations within a genome, we computed a p-value for each group. Let x = number of bases spanned by a group (from first mutation to last), k = number of mutations in a group, π = number of total mutations divided by number of total bases in a genome, and j = an indexing parameter. Then by the negative binomial distribution[52], the cluster p-value:

$$p=\sum_{j=0}^{x-k}\binom{k+j-2}{j}(1-\pi)^{j}\pi^{k-1}$$

π was computed using all mutations (i.e., including those filtered for dbSNP and simpleRepeat), as this could only increase the p-values. Each group with p-value ≤ 10^{-4} was considered a bona fide mutation cluster. A recursive approach was applied, i.e., all clusters passing p-value filtering were identified, even if such a cluster was a subset within a larger group that did not pass the p-value filter. Clusters composed of only mutations that originated from cytosines along the same DNA strand were classified as C-coordinated. Mutations not found in a cluster were classified as scattered.

Mutation signature analyses

Overall structure of signature analysis involving complementary approaches used to identify, statistically evaluate and compare mutation signatures is outlined in Figure 6 and detailed in sections below.

Enrichment calculations

For all analyses, substitutions at C:G base pairs were treated as mutations at C. Enrichment quantifies how frequently C → G or C → T mutations occur at a specific sequence context compared to C → G or C → T mutations at cytosines overall. C → A substitutions were excluded because such mutations are rare due to abasic site bypass[7,36], and to avoid confounding overlap with frequent G → T substitutions in some cancers[53]. To compute

enrichment for mutations at T$\underline{C}$A, let $Mut_{T\underline{C}A}$ = number of T$\underline{C}$A → T$\underline{G}$A or T$\underline{C}$A → T$\underline{T}$A mutations and Con_{TCA} = number of occurrences of TCA (and reverse complement TGA) contexts within the set of 41-mers centered on each mutation within a sample. Similarly, let $Mut_{\underline{C}}$ = number of C → G or C → T mutations and Con_C = number of cytosines or guanines within the set of 41-mers centered on each mutation within a sample. Then the enrichment for mutations at T$\underline{C}$A:

$$E_{T\underline{C}A}=\frac{Mut_{T\underline{C}A}/Con_{TCA}}{Mut_{\underline{C}}/Con_C}=\frac{Mut_{T\underline{C}A}\times Con_C}{Mut_{\underline{C}}\times Con_{TCA}}$$

Enrichments for the other contexts T$\underline{C}$, T$\underline{C}$W, RT$\underline{C}$A, YT$\underline{C}$A, and each NT$\underline{C}$A, were calculated analogously.

Identification of samples significantly mutated by APOBEC(s)

Statistical overrepresentation of APOBEC mutagenesis within each sample was evaluated by one-sided Fisher's exact test. Taking T$\underline{C}$A as an example, the test computed the p-value for a comparison between the ratio $Mut_{T\underline{C}A}$ / ($Mut_{\underline{C}}$ - $Mut_{T\underline{C}A}$) vs. the ratio Con_{TCA} / (Con_C - Con_{TCA}), based on the prediction that the former ratio exceeds the latter. All samples not matching this prediction were assigned p = 1. Benjamini-Hochberg (BH) p-value correction for multiple testing[54] was applied by the p.adjust() function in the R statistical computing package. Samples with these adjusted q-values < 0.05 were considered significant.

Estimating the number of mutations created by APOBEC(s)

A minimal estimate for the number of T$\underline{C}$A mutations created by APOBEC(s) was computed as:

$$Min_{T\underline{C}A}=Mut_{T\underline{C}A}\times\frac{E_{T\underline{C}A}-1}{E_{T\underline{C}A}}$$

Since enrichment = 1 implies T$\underline{C}$A mutations are neither more nor less frequent (when corrected for motif abundance) than mutations at $\underline{C}$ in general, this minimal estimate reports the number of T$\underline{C}$A mutations in excess of enrichment = 1. It is only this excess which should be attributed to mutagenesis by an APOBEC. Samples with Fisher's exact test q > 0.05 for enrichment at T$\underline{C}$A were assigned a $Min_{T\underline{C}A}$ = 0.

Y/RT$\underline{C}$A enrichment analysis

The χ^2 test for goodness of fit was used to identify samples that had a ratio of YT$\underline{C}$A to RT$\underline{C}$A mutations which differed statistically from random, by comparing observed vs. expected mutation counts. The expected number of YT$\underline{C}$A mutations, given the null hypothesis of random mutagenesis, simply scales with fraction of motifs at YTCA:

$$Exp_{YT\underline{C}A}=Mut_{T\underline{C}A}\times\frac{Con_{YTCA}}{Con_{TCA}}$$

The expected number of RTCA mutations was computed analogously. p-values were corrected by the BH method, with q-values < 0.05 considered significant. Samples within each cohort were filtered first for significant TCA mutagenesis enrichment, then for significant difference from random distribution of YTCA vs. RTCA mutations. Samples passing only the first filter were plotted in the relevant figures as unfilled, gray-bordered circles, while samples passing both filters were plotted in colored circles, and included in χ^2 tests for trend toward A3A-like signatures with increasing TCA enrichment.

NTCA proportion analysis

Similarly, the χ^2 test for goodness of fit was used to identify samples that had a proportion of observed ATCA:CTCA:GTCA:TTCA mutations which differed statistically from random. The expected number of mutations at each NTCA:

$$Exp_{NT\underline{C}A} = Mut_{T\underline{C}A} \times \frac{Con_{NTCA}}{Con_{TCA}}$$

p-values from comparing observed vs. expected mutation counts were corrected by the BH method, with q-values < 0.05 considered significant. Only samples passing filtering for both significant TCA mutagenesis enrichment and non-randomness of NTCA proportion were included in root mean square deviation (RMSD, also called root mean square error) comparisons. RMSD is used commonly to quantify the similarity between two corresponding sets of quantities, e.g. the three-dimensional spatial coordinates of alpha-carbon atoms in one protein structure vs. another[55].

RMSD was used to quantify the difference between the normalized enrichment observed in each sample for mutations at each NTCA vs. the corresponding normalized enrichment values in each yeast model. Taking ATCA as an example, the normalized enrichment:

$$NE_{AT\underline{C}A} = \frac{E_{AT\underline{C}A}}{E_{AT\underline{C}A} + E_{CT\underline{C}A} + E_{GT\underline{C}A} + E_{TT\underline{C}A}}$$

Let yNE_{NTCA} = normalized enrichment for mutations at NTCA observed in a yeast model. Then the RMSD of a cancer sample vs. a yeast model:

$$RMSD = \sqrt{\frac{1}{4}\sum\left(NE_{NT\underline{C}A} - yNE_{NT\underline{C}A}\right)^2}$$

Samples with RMSD vs. A3A < RMSD vs. A3B were considered A3A-like, while those with RMSD vs. A3B < RMSD vs. A3A were A3B-like.

pLogo analysis

pLogos identify nucleotides statistically over- or underrepresented in a ‘foreground’ set of sequences, relative to abundances within a ‘background’ set[30]. pLogos were generated using all C → T substitutions from yeast data and all C → G or C → T substitutions from cancer

samples. Each element within the set of foreground sequences comprised the two bases immediately 5′ of a mutation, the mutated base itself (always C), and one base immediately 3′. The corresponding background was the set of 41-mers each centered on a mutation included in the foreground. The deaminated C was set to position 0. Nucleotides above the horizontal axis were overrepresented, while those below the axis were underrepresented. The height of each nucleotide denotes the magnitude of over- or underrepresentation. Red lines represent cutoffs for p = 0.05. In rare cases, the number of bases in the background set was apparently greater than could be accommodated by the pLogo online tool, so the set of C → G or C → T substitutions was analyzed separately from the G → C or G → A set. As such pairs of pLogos were always very similar, we reported those generated from C → G or C → T substitutions only.

Additional statistical analyses

Additional statistical analyses, including Kolmogorov-Smirnov test, Spearman's correlation, χ^2 test with Yates correction, and χ^2 test for trend, were performed using Graphpad Prism 6 (Graphpad Software).

Code availability

APOBEC mutagenesis pattern was analyzed similarly to the analysis incorporated into the Broad's Institute TCGA GDAC Firehose[42]. R code is available upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Drs. T.A. Kunkel, S.A. Lujan, and D.V. Zaykin for critical reading of the manuscript. This work was supported by NIH Intramural Research Program Project Z1AES103266 to DAG, and NIH grants U24 CA143845 to GG, R01GM052319 to PAM, 1P01CA120964 to DJK, R00ES022633 to SAR, and K99ES024424 to KC.

References

1. Stratton MR. Exploring the Genomes of Cancer Cells: Progress and Promise. Science. 2011; 331:1553–1558. [PubMed: 21436442]
2. Alexandrov LB, et al. Signatures of mutational processes in human cancer. Nature. 2013; 500:415–421. [PubMed: 23945592]
3. Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat Genet. 2013; 45:977–983. [PubMed: 23852168]
4. Roberts SA, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013; 45:970–976. [PubMed: 23852170]
5. Burns M, Leonard B, Harris R. APOBEC3B: Pathological consequences of an innate immune DNA mutator. Biomed J. 2015; 38:102–110. [PubMed: 25566802]
6. Harris RS. Molecular mechanism and clinical impact of APOBEC3B-catalyzed mutagenesis in breast cancer. Breast Cancer Res. 2015; 17:8. [PubMed: 25848704]
7. Chan K, et al. Base Damage within Single-Strand DNA Underlies In Vivo Hypermutability Induced by a Ubiquitous Environmental Agent. PLoS Genet. 2012; 8:e1003149. [PubMed: 23271983]
8. Landry S, Narvaiza I, Linfesty DC, Weitzman MD. APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. EMBO Rep. 2011; 12:444–450. [PubMed: 21460793]

9. Burns MB, et al. APOBEC3B is an enzymatic source of mutation in breast cancer. Nature. 2013; 494:366–370. [PubMed: 23389445]
10. Mussil B, et al. Human APOBEC3A Isoforms Translocate to the Nucleus and Induce DNA Double Strand Breaks Leading to Cell Stress and Death. PLoS ONE. 2013; 8:e73641. [PubMed: 23977391]
11. Nik-Zainal S, et al. Mutational Processes Molding the Genomes of 21 Breast Cancers. Cell. 2012; 149:979–993. [PubMed: 22608084]
12. Roberts SA, et al. Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. Mol Cell. 2012; 46:424–435. [PubMed: 22607975]
13. Bishop KN, et al. Cytidine Deamination of Retroviral DNA by Diverse APOBEC Proteins. Curr Biol. 2004; 14:1392–1396. [PubMed: 15296758]
14. Dang Y, Wang X, Esselman WJ, Zheng Y-H. Identification of APOBEC3DE as Another Antiretroviral Factor from the Human APOBEC Family. J Virol. 2006; 80:10522–10533. [PubMed: 16920826]
15. Harris RS, Petersen-Mahrt SK, Neuberger MS. RNA Editing Enzyme APOBEC1 and Some of Its Homologs Can Act as DNA Mutators. Mol Cell. 2002; 10:1247–1253. [PubMed: 12453430]
16. Henry M, et al. Genetic Editing of HBV DNA by Monodomain Human APOBEC3 Cytidine Deaminases and the Recombinant Nature of APOBEC3G. PLoS ONE. 2009; 4:e4277. [PubMed: 19169351]
17. Yu Q, et al. APOBEC3B and APOBEC3C Are Potent Inhibitors of Simian Immunodeficiency Virus Replication. J Biol Chem. 2004; 279:53379–53386. [PubMed: 15466872]
18. Cescon DW, Haibe-Kains B, Mak TW. APOBEC3B expression in breast cancer reflects cellular proliferation, while a deletion polymorphism is associated with immune activation. Proc Natl Acad Sci USA. 2015; 112:2841–2846. [PubMed: 25730878]
19. Waters CE, Saldivar JC, Amin ZA, Schrock MS, Huebner K. FHIT loss-induced DNA damage creates optimal APOBEC substrates: Insights into APOBEC-mediated mutagenesis. Oncotarget. 2014
20. Sasaki H, et al. APOBEC3B gene overexpression in non-small-cell lung cancer. Biomed Rep. 2014; 2:392–395. [PubMed: 24748981]
21. Leonard B, et al. APOBEC3B Upregulation and Genomic Mutation Patterns in Serous Ovarian Carcinoma. Cancer Res. 2013; 73:7222–7231. [PubMed: 24154874]
22. de Bruin EC, et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. Science. 2014; 346:251–256. [PubMed: 25301630]
23. Caval V, Suspène R, Vartanian J-P, Wain-Hobson S. Orthologous Mammalian APOBEC3A Cytidine Deaminases Hypermutate Nuclear DNA. Mol Biol Evol. 2014; 31:330–340. [PubMed: 24162735]
24. Shee C, et al. Engineered proteins detect spontaneous DNA breakage in human and bacterial cells. eLife. 2013; 2:e01222. [PubMed: 24171103]
25. Nik-Zainal S, et al. Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer. Nat Genet. 2014; 46:487–491. [PubMed: 24728294]
26. Caval V, Suspène R, Shapira M, Vartanian J-P, Wain-Hobson S. A prevalent cancer susceptibility APOBEC3A hybrid allele bearing APOBEC3B 3′UTR enhances chromosomal DNA damage. Nat Commun. 2014; 5:5129. [PubMed: 25298230]
27. Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. Nat Rev Cancer. 2014; 14:786–800. [PubMed: 25568919]
28. Poon S, McPherson J, Tan P, Teh B, Rozen S. Mutation signatures of carcinogen exposure: genome-wide detection and new opportunities for cancer prevention. Genome Med. 2014; 6:24. [PubMed: 25031618]
29. Degtyareva NP, et al. Oxidative stress-induced mutagenesis in single-strand DNA occurs primarily at cytosines and is DNA polymerase zeta-dependent only for adenines and guanines. Nucleic Acids Res. 2013; 41:8995–9005. [PubMed: 23925127]
30. O'Shea JP, et al. pLogo: a probabilistic approach to visualizing sequence motifs. Nat Meth. 2013; 10:1211–1212.

31. Taylor BJ, et al. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. eLife. 2013; 2:e00534. [PubMed: 23599896]
32. Fredriksson NJ, Ny L, Nilsson JA, Larsson E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. Nat Genet. 2014; 46:1258–1263. [PubMed: 25383969]
33. Walker BA, et al. APOBEC family mutational signatures are associated with poor prognosis translocations in multiple myeloma. Nat Commun. 2015; 6:6997. [PubMed: 25904160]
34. Long J, et al. A Common Deletion in the APOBEC3 Genes and Breast Cancer Risk. J Natl Cancer Inst. 2013; 105:573–579. [PubMed: 23411593]
35. Xuan D, et al. APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. Carcinogenesis. 2013; 34:2240–2243. [PubMed: 23715497]
36. Chan K, Resnick MA, Gordenin DA. The choice of nucleotide inserted opposite abasic sites formed within chromosomal DNA reveals the polymerase activities participating in translesion DNA synthesis. DNA Repair. 2013; 12:878–889. [PubMed: 23988736]
37. Drier Y, et al. Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. Genome Res. 2013; 23:228–235. [PubMed: 23124520]
38. Mimitou EP, Symington LS. DNA end resection—Unraveling the tail. DNA Repair. 2011; 10:344–348. [PubMed: 21227759]
39. Sakofsky CJ, et al. Break-Induced Replication Is a Source of Mutation Clusters Underlying Kataegis. Cell Rep. 2014; 7:1640–1648. [PubMed: 24882007]
40. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014; 507:315–322. [PubMed: 24476821]
41. Davis CF, et al. The Somatic Genomic Landscape of Chromophobe Renal Cell Carcinoma. Cancer Cell. 2014; 26:319–330. [PubMed: 25155756]
42. Broad Institute TCGA Genome Data Analysis Center. Analysis of mutagenesis by APOBEC cytidine deaminases (P-MACD). (Broad Institute of MIT and Harvard. 2014
43. Rooney MS, Shukla SA, Wu CJ, Getz G, Hacohen N. Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. Cell. 2015; 160:48–61. [PubMed: 25594174]
44. Snyder A, et al. Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. N Engl J Med. 2014; 371:2189–2199. [PubMed: 25409260]
45. Casey RG, et al. Diagnosis and Management of Urothelial Carcinoma In Situ of the Lower Urinary Tract: A Systematic Review. Eur Urol. 2014; 67:876–888. [PubMed: 25466937]
46. Powles T, et al. MPDL3280A (anti-PD-L1) treatment leads to clinical activity in metastatic bladder cancer. Nature. 2014; 515:558–562. [PubMed: 25428503]
47. Bellí G, Garí E, Piedrafita L, Aldea M, Herrero E. An activator/repressor dual system allows tight tetracycline-regulated gene expression in budding yeast. Nucleic Acids Res. 1998; 26:942–947. [PubMed: 9461451]
48. Storici, F.; Resnick, MA. The Delitto Perfetto Approach to In Vivo Site-Directed Mutagenesis and Chromosome Rearrangements with Synthetic Oligonucleotides in Yeast.. In: Judith, LC.; Paul, M., editors. Methods Enzymol. Vol. 409. Academic Press; 2006. p. 329-345.
49. Morrison A, Bell JB, Kunkel TA, Sugino A. Eukaryotic DNA polymerase amino acid sequence required for 3′--->5′ exonuclease activity. Proc Natl Acad Sci USA. 1991; 88:9473–9477. [PubMed: 1658784]
50. Sakamoto AN, et al. Mutator alleles of yeast DNA polymerase ζ. DNA Repair. 2007; 6:1829–1838. [PubMed: 17715002]
51. Matsuda T, Bebenek K, Masutani C, Hanaoka F, Kunkel TA. Low fidelity DNA synthesis by human DNA polymerase-eta. Nature. 2000; 404:1011–1013. [PubMed: 10801132]
52. Haldane JBS. On a Method of Estimating Frequencies. Biometrika. 1945; 33:222–225. [PubMed: 21006837]
53. Lawrence MS, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499:214–218. [PubMed: 23770567]

54. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Statist Soc B. 1995; 57:289–300.

55. Maiorov VN, Crippen GM. Significance of Root-Mean-Square Deviation in Comparing Three-dimensional Structures of Globular Proteins. J Mol Biol. 1994; 235:625–634. [PubMed: 8289285]

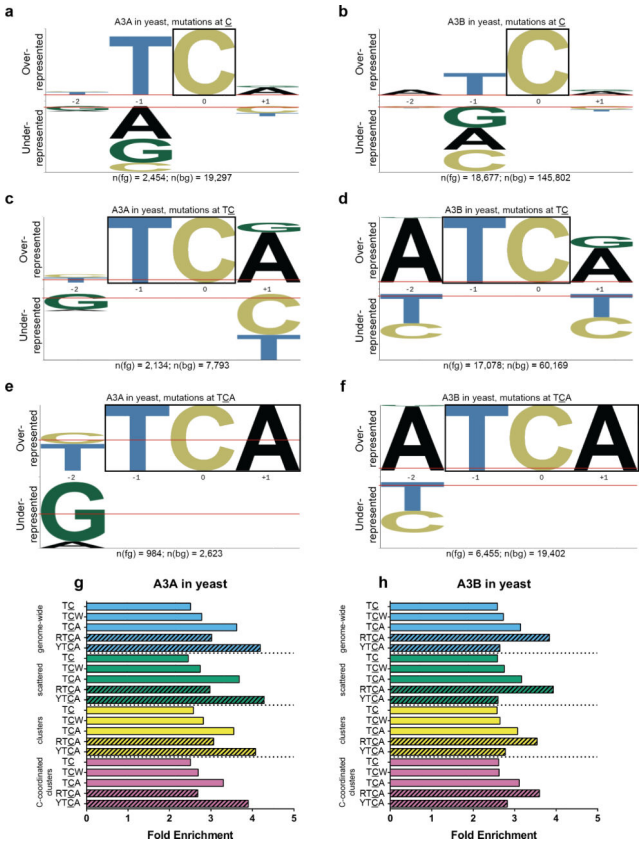

Figure 1.

Analyses of mutations induced by APOBECs in *ung1*Δ yeast. pLogos show overrepresented nucleotides in a motif above the horizontal axis and underrepresented nucleotides below[30]. The size of each letter indicates magnitude of over- or underrepresentation. Fixed positions in each motif are highlighted by a box. n(fg) denotes the number of mutations at C, TC, or TCA, while n(bg) denotes the number of contexts at C, TC, or TCA. (a) All C:G → T:A substitutions induced by (a) A3A or (b) A3B, with C fixed at position 0, indicating overrepresentation of TC. (c) A3A and (d) A3B pLogos with fixed TC, revealing overrepresentation of TCA. (e) A3A and (f) A3B pLogos with fixed TCA, revealing near-diametrically opposite preferences at the –2 nucleotide, two positions 5′ of the deamination site. (g and h) Enrichment values for APOBEC-related motifs among genome-wide, scattered, clustered, and C-coordinated clustered mutations induced by (g) A3A or (h) A3B. Note that the similar enrichment values for the same motif (e.g., TCA) among different mutation categories suggest that the APOBECs targeted their cognate motifs with similar specificity, whether the ssDNA was clustered and presumably persistent (i.e., at uncapped telomeres), or scattered and presumably transient (e.g., transcription intermediates).

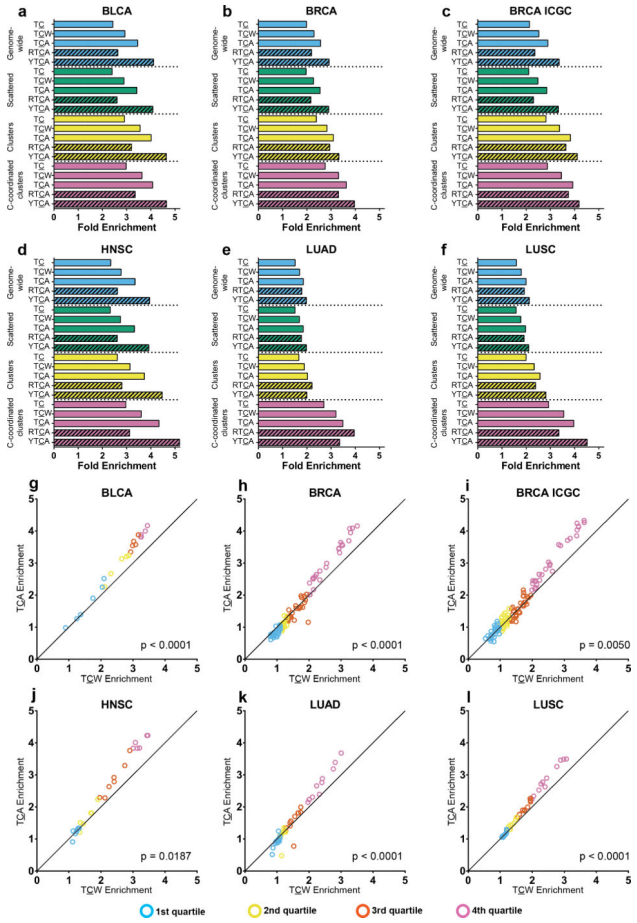

Figure 2.

Enrichment for mutations at various target motifs among all genome samples, and sample-by-sample comparison of genome-wide enrichment at TCA vs. TCW, within six cohorts of highly APOBEC-mutated cancer types. (a–f) Enrichment for mutations at TC, TCW, TCA, RTCA, and YTCA are shown for (a) BLCA, (b) BRCA, (d) HNSC, (e) LUAD, and (f) LUSC cohorts from TCGA, as well as (c) a BRCA cohort from ICGC. High genome-wide non-APOBEC mutation loads obscured the presence of APOBEC mutagenesis in the lung cancers. Nevertheless, APOBEC signature enrichment values in C-coordinated clusters of (e) LUAD and (f) LUSC are similar to those in other cancer types (a–d), confirming that examination of such clusters is the most sensitive means to detect APOBEC mutagenesis. (g–l) Sample-by-sample comparison of enrichment for mutations at TCA vs. TCW for (g) BLCA, (h) BRCA, (i) BRCA ICGC, (j) HNSC, (k) LUAD, and (l) LUSC cohorts. Samples are binned by quartile of TCW enrichment. χ^2 tests for trend (p-values in each panel) confirm that as TCW enrichment increases, there is significant skewing toward samples with TCA enrichment > TCW enrichment.

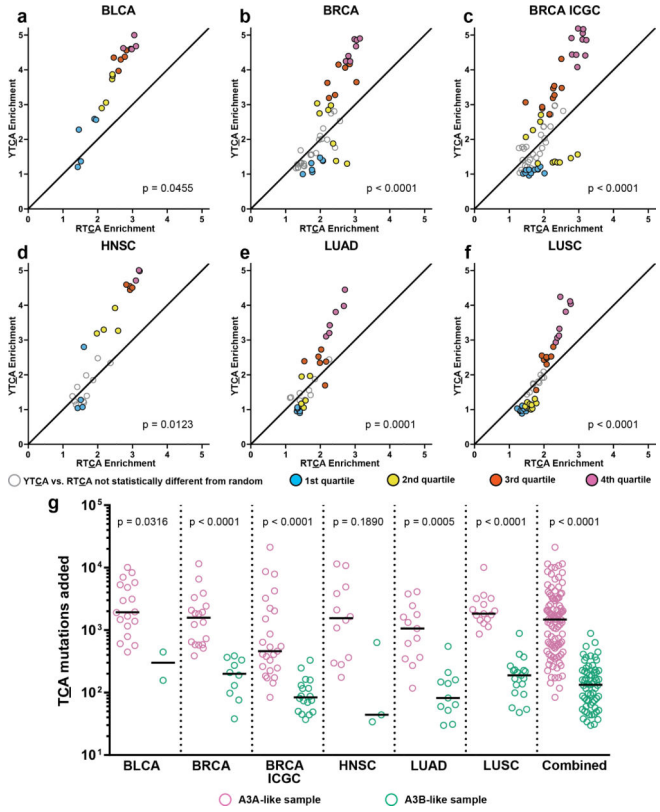

Figure 3.

Y/RTCA analysis and estimated mutation load of highly APOBEC-mutated cohorts. (a–f) Only samples significantly enriched for mutations at TCA ($q < 0.05$) are shown. Samples with YTCA mutation to RTCA mutation ratio statistically different from random ($q < 0.05$) are binned by quartile of TCA enrichment and plotted as filled symbols. Samples with YTCA > RTCA enrichment are considered A3A-like. Those with RTCA > YTCA enrichment are A3B-like. p-values from χ^2 test for trend are shown. Samples significantly enriched at TCA, but with YTCA vs. RTCA ratio not statistically different from random ($q > 0.05$), are plotted as unfilled, gray-bordered symbols and not included in tests for trend. (g) The minimal estimated number of TCA mutations attributable to APOBEC mutagenesis, for A3A- (pink) or A3B-like (green) samples in each cohort, and for all six cohorts combined, are shown along with medians and Kolmogorov-Smirnov p-values. See Online Methods for analytical details.

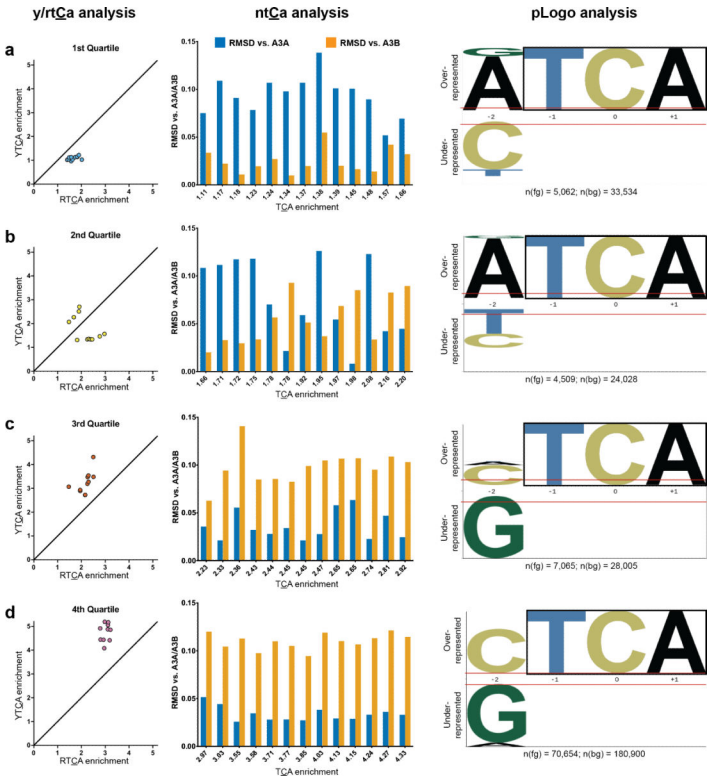

Figure 4.

Three-way comparison of Y/RTCA, NTCA, and pLogo methodologies for identifying samples with A3A- or A3B-like signatures in the BRCA ICGC cohort. Samples are binned by quartile of TCA enrichment, with 1st (lowest) quartile in (a), 2nd quartile in (b), 3rd quartile in (c), and 4th (highest) quartile in (d). All samples in the figure passed statistical filtering for significant TCA-signature enrichment ($q < 0.05$). Samples in Y/RTCA analysis also passed statistical filtering for non-random ratio of YTCA vs. RTCA, by χ^2 test for goodness of fit. Samples in NTCA analysis passed analogous filtering for non-random proportion of four NTCA's. In each RMSD graph (middle panels), samples are arranged by increasing TCA enrichment. All three analyses indicated that lower TCA enrichment samples predominantly have A3B-like signatures, while high enrichment samples in the upper quartiles are all A3A-like. See Online Methods for analytical details.

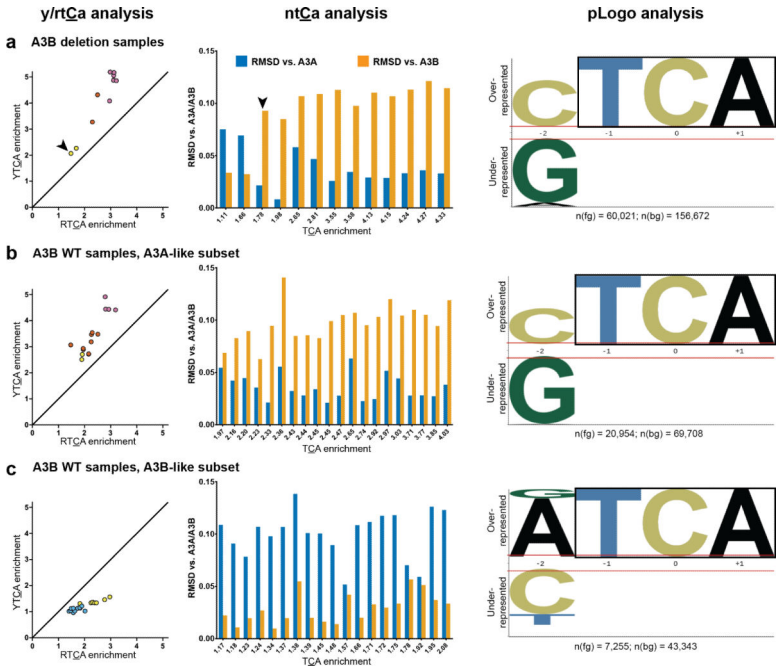

Figure 5.

Relationship between A3B germline copy number and mutation signatures in the BRCA ICGC cohort. Samples passed same filtering criteria as in Figure 4. (a) A3B deletion samples (one homozygous, denoted by arrowhead; remainder heterozygous) skew toward A3A-like signatures. (b) A3A-like and (c) A3B-like subsets of samples with wild-type (WT) A3B copy number.

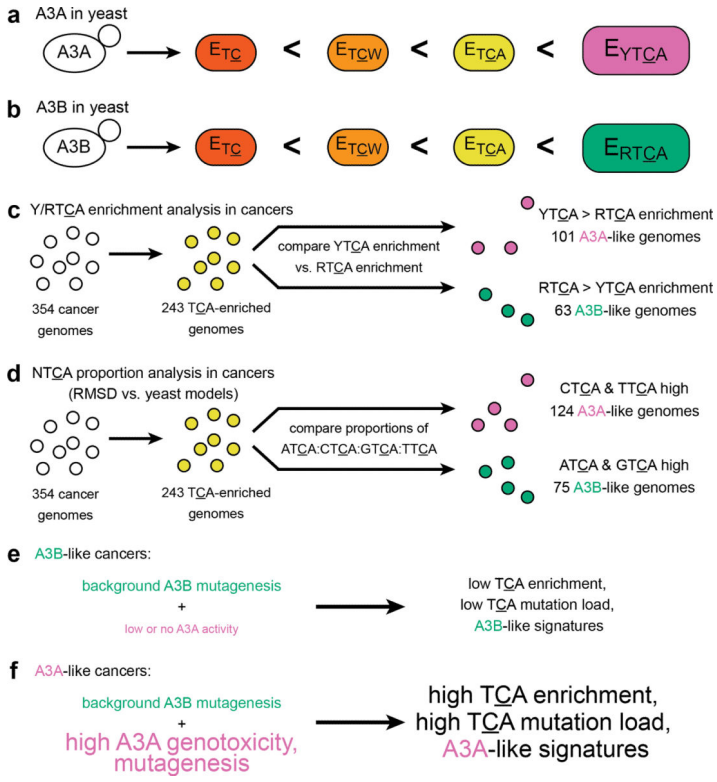

Figure 6.

Summary of data analyses and conclusions. (a and b) Sets of mutations induced by either (a) A3A or (b) A3B in yeast are successively more enriched at TC, TCW, and TCA. (a) For A3A, mutations at YTCA are more enriched than mutations at RTCA. (b) In contrast for A3B, mutations at RTCA are more enriched than mutations at YTCA. (c) By Y/RTCA enrichment analysis, out of 243 cancer genome samples with significant TCA mutagenesis, 101 (41.6%) are A3A-like and 63 (25.9%) are A3B-like. The remaining 79 (32.5%) are indeterminate. (d) By NTCA proportion analysis, 124 cancer samples (51.0%) are A3A-like, 75 (30.9%) are A3B-like, and 44 (18.1%) are indeterminate. (e) In A3B-like cancer samples, background A3B mutagenesis results in low overall TCA enrichment signatures with higher RTCA (especially ATCA) enrichment. (f) In A3A-like cancer samples, background A3B mutagenesis is dwarfed by A3A mutagenic activity, leading to high TCA enrichment signatures with even higher YTCA (especially CTCA) enrichment.