nature genetics

An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers

Steven A Roberts[1], Michael S Lawrence[2], Leszek J Klimczak[3], Sara A Grimm[3], David Fargo[3], Petar Stojanov[2], Adam Kiezun[2], Gregory V Kryukov[2,4], Scott L Carter[2], Gordon Saksena[2], Shawn Harris[5], Ruchir R Shah[5], Michael A Resnick[1], Gad Getz[2,6–8] & Dmitry A Gordenin[1,8]

Recent studies indicate that a subclass of APOBEC cytidine deaminases, which convert cytosine to uracil during RNA editing and retrovirus or retrotransposon restriction, may induce mutation clusters in human tumors. We show here that throughout cancer genomes APOBEC-mediated mutagenesis is pervasive and correlates with APOBEC mRNA levels. Mutation clusters in whole-genome and exome data sets conformed to the stringent criteria indicative of an APOBEC mutation pattern. Applying these criteria to 954,247 mutations in 2,680 exomes from 14 cancer types, mostly from The Cancer Genome Atlas (TCGA), showed a significant presence of the APOBEC mutation pattern in bladder, cervical, breast, head and neck, and lung cancers, reaching 68% of all mutations in some samples. Within breast cancer, the HER2-enriched subtype was clearly enriched for tumors with the APOBEC mutation pattern, suggesting that this type of mutagenesis is functionally linked with cancer development. The APOBEC mutation pattern also extended to cancer-associated genes, implying that ubiquitous APOBEC-mediated mutagenesis is carcinogenic.

Genome instability triggers the development of many types of cancers[1,2]. Radiation and chemical damage are traditionally invoked as culprits in theories of carcinogenic mutagenesis[3]. However, normal enzymatic activities can also be a source of DNA damage and mutation. Cytidine deaminases, which convert cytosine bases to uracil, likely contribute to DNA damage[4]. Activation-induced cytidine deaminase (AID), a key enzyme in adaptive immunity, not only initiates the hypermutation and class-switch recombination of immunoglobulin genes but also can mutate chromosomal DNA at a limited number of secondary targets, some of which have been implicated in carcinogenesis[5]. In addition to AID, the human genome encodes several homologous APOBEC (apolipoprotein B mRNA-editing enzyme, catalytic polypeptide-like) cytidine deaminases that function in innate immunity as well as in RNA editing[6]. Previous human cell culture studies showed that a subclass of APOBECs with mutational specificity for TC motifs (with the mutated base underlined) is capable of inducing mutations in chromosomal and mitochondrial DNA and therefore could have a role in carcinogenesis[7–9]. (APOBEC without the gene-specifying suffix is used hereafter to designate a subclass of cytidine deaminases with TC motif specificity. Note that, on the basis of motif specificity, APOBEC3G and AID do not fall into this subclass.) Supporting a role for APOBECs in cancer, a mutation signature consistent with APOBEC editing was found in individual cancer-related genes[10,11]. Recently, clustered mutations (termed kataegis in ref. 12) identified through next-generation sequencing suggested that APOBECs can induce base substitutions in tumor genomes[12,13]. Clustered mutations showed even higher preference for a more stringent TCW motif (where W corresponds to either A or T). Tightly linked strand-coordinated clustered mutations (clusters whose mutations all occur at one type of nucleotide) were often colocalized with rearrangement breakpoints, suggesting that this mutagenesis results from aberrant DNA double-strand break (DSB) repair that produces single-stranded DNA (ssDNA), an ideal substrate for the APOBEC enzymes[6]. The frequency of base substitutions in the APOBEC motif was higher for clustered mutations identified in whole genome–sequenced breast cancers[12], as well as in multiple myeloma, prostate, and head and neck cancers[13]. Notably, non-clustered substitutions in the TCW motif occurred near rearrangement breakpoints more frequently than expected by random chance across the genomes of several cancer types[14]. Analysis of breast cancer sequencing and expression data suggested that it is specifically APOBEC3B that causes mutations in this cancer type[9].

Despite the indication that APOBEC-mediated mutagenesis may have a role in cancer, it was unclear how strong of a mutagenic factor APOBEC enzymes are, whether APOBEC mutagenesis is a ubiquitous characteristic of many cancer types and cases, and whether it is associated with any specific tumor characteristics. Here we have developed an analysis to evaluate the strength of the APOBEC mutation pattern in individual samples from multiple whole-genome and exome mutation data sets, such as TCGA. We found that the APOBEC mutation pattern is prominent and even prevailing in many samples

[1]Laboratory of Molecular Genetics, National Institute of Environmental Health Sciences, Durham, North Carolina, USA. [2]The Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [3]Integrative Bioinformatics, National Institute of Environmental Health Sciences, Durham, North Carolina, USA. [4]Harvard Medical School, Boston, Massachusetts, USA. [5]SRA International, Inc., Durham, North Carolina, USA. [6]Massachusetts General Hospital Cancer Center, Boston, Massachusetts, USA. [7]Department of Pathology, Massachusetts General Hospital, Boston, Massachusetts, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to D.A.G. (gordenin@niehs.nih.gov) or G.G. (gadgetz@broadinstitute.org).

Received 28 January; accepted 20 June; published online 14 July 2013; doi:10.1038/ng.2702

from several types of cancer, in contrast to other cancer types where it is barely detectable, and that it correlates with APOBEC mRNA levels and extends into a subset of genes considered by multiple criteria to be cancer drivers.

RESULTS

A pattern for detecting APOBEC mutagenesis

Our approach to the statistical exploration of complex mutation spectra in multiple cancer samples involved the formulation of a single hypothesis surrounding a diagnostic mutation pattern, which uses knowledge obtained in previous experiments as well as in data analyses and minimizes overlap with other known sequence-specific mutagenesis mechanisms. The first step in defining a measure for a pattern of APOBEC-mediated mutagenesis in a cancer sample was to find mutations that occurred in the motif most likely to be an APOBEC-specific target. We chose the TCW motif instead of the less stringent TC motif because of its demonstrated prevalence in mutagenesis caused by some APOBECs in model systems as well as in mutation clusters found in cancers (refs. 9,12,13 and references therein). The more stringent TCW motif also eliminated potential overlap with sequence-specific mutagenesis in highly mutable CpG sequences that would occasionally be preceded by a thymine. Second, we proposed that APOBEC-induced mutagenesis would involve primarily cytosine-to-guanine and/or cytosine-to-thymine substitutions, with rare cytosine-to-adenine changes. This substitution pattern is based on the tendency of translesion synthesis to misincorporate cytosine or adenine bases across from abasic sites (resulting in cytosine-to-guanine and cytosine-to-thymine mutations) that are generated frequently by the activity of uracil DNA glycosylase[15–17] toward the products of both spontaneous and APOBEC-induced cytosine deamination, as well as copying a cytosine deamination–derived uracil (resulting in cytosine-to-thymine changes). Thus, in the present analysis, we defined cytosine-to-thymine and cytosine-to-guanine substitutions in TCW motifs as APOBEC signature mutations. To identify samples that experienced APOBEC-mediated mutagenesis, we further defined an APOBEC mutagenesis pattern within a sample as a statistically significant enrichment of the frequency of APOBEC signature mutations compared to that expected with random mutagenesis (Online Methods). Enrichment for APOBEC signature mutations (TCW to TTW or TGW and the complementary WGA to WAA or WCA changes) among all similar mutations of cytosine or guanine bases (cytosine to thymine or guanine and guanine to adenine or cytosine) was calculated relative to the frequency of the APOBEC mutation motif (TCW or WGA) in the 41-nucleotide regions centered on the mutated bases. We used only the nucleotides immediately surrounding the mutations in this calculation because APOBEC enzymes are thought to scan a limited area of ssDNA to deaminate a cytosine in a preferred motif[18,19]. This approach does not exclude any given area of the genome in general, but rather uses the areas within each sample where mutagenesis has happened, and then evaluates whether the mutagenesis in this sample was enriched for APOBEC signature mutations. To test the accuracy of our analysis, we compared our measure of the APOBEC mutation pattern (fold enrichment) to a previously reported measure of APOBEC-mediated mutagenesis obtained via a very different approach involving the mathematical decomposition and extraction of multiple mutation signatures from 21 breast cancer samples[12]. The results showed a very high level of correlation between the measures (**Supplementary Fig. 1** and **Supplementary Table 1**), supporting the applicability of our method. Moreover, our analysis remained robust even when applied to samples containing small numbers of mutations. The fold enrichment of APOBEC-mediated mutations in a subset of mutations representing exomes from the aforementioned 21 breast cancers (~2% of total mutations in the whole genome) correlated strongly with values obtained from the entire genome (**Supplementary Fig. 2**), suggesting that our analysis may be effectively applied to mutations identified through exome sequencing and would thereby substantially increase the number of cancer samples that are available for analysis.

An APOBEC mutagenesis pattern in mutation clusters

We evaluated the APOBEC mutation pattern in a large number of whole-genome and exome mutation data sets accumulated in TCGA as well as in several publications[20,21]. APOBECs are highly specific for ssDNA and are capable of simultaneously making multiple mutations if an ssDNA region persists[17,19]. Such mutations are strand coordinated, as changes in multiple cytosines occur on the same DNA strand. We and others have detected this APOBEC mutation pattern in cytosine and complementary guanine strand-coordinated clusters from a limited number of whole genome–sequenced tumors[12,13]. These clusters often colocalize with rearrangement breakpoints[12,13] (**Fig. 1a** and **Supplementary Fig. 3**), which agrees with mutagenesis occurring in ssDNA regions that are either prone to breakage and/or are formed during a DSB repair process. Clustered cytosine or guanine mutations identified previously[13] as well as in additional analysis of whole genome–sequenced colorectal adenocarcinomas[22] presented here showed a strong APOBEC mutation pattern (with the highest enrichment observed for the TCW motif and strong preference for cytosine-to-thymine and cytosine-to-guanine changes in this motif; **Fig. 1a**, **Supplementary Fig. 3** and **Supplementary Table 2**).

We next addressed whether an APOBEC mutation pattern is common among different cancer samples and types. We accumulated lists of cancer-specific mutations from the whole-exome sequencing of 2,680 tumors, mostly by the TCGA Research Network (**Supplementary Table 3**). Although exome sequencing substantially increases the number of samples available for analysis, its general specificity for protein-coding regions results in only ~1% of total genomic DNA being assessed. In identifying clusters from exome sequencing, we therefore estimated the total mutation load in a given tumor sample under the assumption that exome mutations constitute 1% of mutations in the entire genome, using this value to identify clusters with our previously described algorithm[13]. This method found 498 total clusters in the 2,680 sequenced exomes from 14 different cancer types. In total, 218 cytosine- or guanine-coordinated clusters were identified, occurring in every cancer type analyzed except acute myeloid leukemia (LAML) (**Supplementary Fig. 4**). Similar to results obtained by whole-genome analysis (compare **Figs. 1a** and **1b**), these clusters showed a robust APOBEC mutation pattern, whereas other known mutagenic motifs involving cytosine or guanine nucleotides were depleted. In contrast, the APOBEC mutation pattern was barely detectable or was undetectable in non-coordinated clustered cytosine and guanine mutations and scattered mutations, respectively (**Supplementary Fig. 5**). The enrichment of APOBEC signature mutations in cytosine- or guanine-coordinated clusters was more pronounced in clusters with >2 mutations (**Supplementary Fig. 4**) because clusters with only 2 mutations had a higher chance of occurring independently through non-APOBEC-dependent mechanisms.

The APOBEC mutagenesis pattern across 2,680 cancer exomes

The strength of the APOBEC mutation pattern in cytosine- or guanine-coordinated clusters from our analysis of exome mutations (**Fig. 1b**) was comparable to that of clusters found in whole-genome mutation lists (**Fig. 1a**), suggesting that exome-wide mutation data

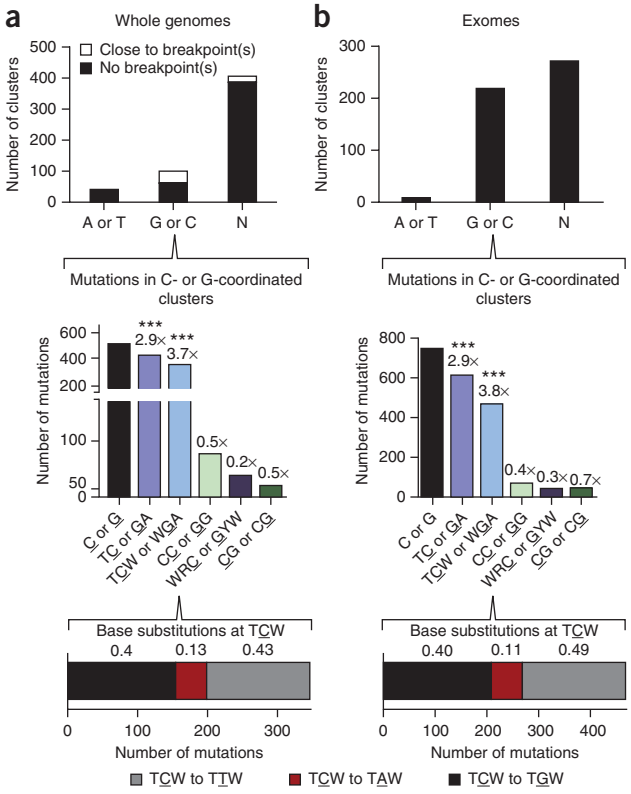

Figure 1 APOBEC mutation pattern in clusters. (**a**,**b**) Analysis of all clusters identified in whole-genome data sets, including 23 multiple myeloma[47], 2 head and neck squamous cell carcinoma[21], 7 prostate carcinoma[48] and 9 colorectal adenocarcinoma[22] data sets (**a**), and in 2,680 exomes from 14 different cancer types from TCGA as well as from other published sources[20,21] (Online Methods) (**b**). Top, colocalization of clusters with breakpoints was identified as described in ref. 13. Clusters identified as being close to breakpoints had at least one mutation within 20 kb of a breakpoint. N, any nucleotide. Middle, fold enrichment (shown above bars) of mutation motifs (with the mutated base underlined) was calculated for all three possible changes of cytosine (or guanine) as described in the Online Methods. ***Bonferroni-corrected q value < 0.0001, as determined by a one-tailed Fisher's exact test comparing the ratio of the number of cytosine mutations at TCW motifs and the number of cytosine mutations not in the sequence TCW (R is A or G, Y is T or C) to the analogous ratio for all cytosines within a sample fraction of the genome. Bottom, the numbers and fractions (above appropriate sections of the bars) of three different base substitutions of cytosine (or guanine).

may be sufficient to detect the APOBEC mutation pattern among all mutations in a sample's exome. Indeed, the APOBEC mutation pattern was clearly present throughout many exomes, indicating that APOBEC enzymes were probably a major source of mutagenesis in these samples (**Fig. 2a** and **Supplementary Table 4**). Samples showing the APOBEC mutation pattern occurred primarily in six cancer types, whereas the other eight cancer types were depleted of this pattern, despite high general mutation rates in many samples ($P < 0.0001$, two-sided χ^2 comparison of the number of samples in each cancer type showing fold enrichment of APOBEC signature mutations greater than the median fold enrichment for all samples; $n = 2,680$). Bladder (BLCA), cervical (CESC), head and neck (HNSC), breast (BRCA) and lung cancers (LUAD and LUSC) were enriched in samples with a high level of APOBEC mutagenesis or having greater odds ratios compared to those for the total range of APOBEC-mediated mutagenesis in exomes (**Fig. 2** and **Supplementary Fig. 6a**,**b**). Motif-specific functional selection is unlikely to have caused the observed over-representation of the APOBEC mutation pattern, as corresponding calculations of fold enrichment for silent and noncoding mutations in each sample produced similar results (**Supplementary Fig. 7**). Across all tumors analyzed, high fold enrichment of APOBEC mutations correlated strongly with decreased Fisher's q value as well as with an increase in the fraction of total mutations in a tumor that had the APOBEC signature (**Supplementary Fig. 8**). In individual tumors with a strong APOBEC pattern, the number of APOBEC signature mutations was often large, making the APOBEC enzyme the predominant source of mutations in the sample (**Fig. 2b**). Notably, some samples contained over a thousand APOBEC signature mutations, constituting up to 68% of mutations in the exome.

In cancer types where an APOBEC mutation pattern was not noticeable in the exome data, the pattern was detectable in clusters of strand-coordinated cytosine (or guanine) mutations from whole-genome data. Whole-genome data contained about 100-fold more mutations than exome data, facilitating the detection of clusters. We previously reported such clusters to be enriched for the APOBEC mutation pattern when considering the mutations in whole genome–sequenced prostate carcinomas[13], and we show here the same pattern in nine whole-genome colorectal cancer mutation data sets[22] (**Supplementary Fig. 3**). In each of these data sets, many of the cytosine- or guanine-coordinated clusters colocalized with chromosomal rearrangement breakpoints, a phenomenon that supports the involvement of ssDNA (the exclusive substrate of APOBEC enzymes) in cluster formation[23,24]. Neither cancer type, however, showed a detectable presence of the APOBEC mutation pattern in exome data. Thus, the APOBEC mutagenesis pattern seems to be ubiquitous at a background level in all types of cancer but is more prominent in particular types.

APOBEC-mediated mutagenesis correlates with APOBEC mRNA

Several cancer type–specific factors, including the availability of ssDNA substrate and the expression level of APOBEC enzymes, could contribute to the extent of APOBEC-mediated mutagenesis. Recently, a tumor-specific increase in the transcription of *APOBEC3B*, determined by quantitative PCR (qPCR) and microarray analysis, as well as by RNA sequencing (RNA-seq), in breast cancer samples was shown to correlate with an increased number of cytosine-to-thymine transitions[9]. Cytosine-to-thymine mutations are a relaxed measure of total deamination, which includes both the APOBEC signature in TCW defined in our analysis as well as mutations stemming from other processes. We used RNA-seq expression data to address whether the expression of any of the eight APOBEC enzymes known to have biochemical deamination activity toward DNA correlated with the observed extent of APOBEC-mediated mutagenesis. Consistent with the previous report in breast cancer, *APOBEC3B* expression was frequently higher in tumor samples compared to matched normal samples; however, median expression of *APOBEC3H* and *APOBEC3A* (**Supplementary Figs. 9** and **10**) was also higher by approximately twofold in tumors. Of the 483 breast cancers analyzed for both APOBEC-mediated mutagenesis and APOBEC expression, *APOBEC3B* as well as *APOBEC3A* mRNA levels correlated strongly with the total number of cytosine-to-thymine mutations per exome (**Supplementary Fig. 11a**; Spearman's $r = 0.233$, Bonferroni-corrected $q < 0.001$ and Spearman's $r = 0.1998$, Bonferroni-corrected $q < 0.001$, respectively). Notably, when transcript levels were compared to the number of mutations conforming to our stringent definition of APOBEC-mediated mutagenesis (TCW to TTW or TGW), the strength of the association was higher for both enzymes (**Fig. 3a**; Spearman's $r = 0.3150$, Bonferroni-corrected $q < 0.001$ and

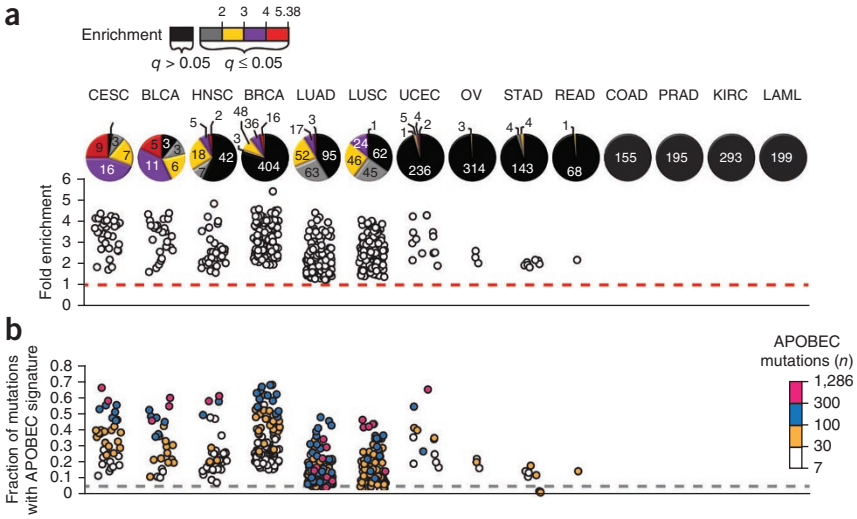

Figure 2 Presence of an APOBEC mutation pattern in exome data sets from different cancer types. (**a**,**b**) Fold enrichment (**a**) and mutation load (**b**) of the APOBEC mutation pattern were determined in each of 2,680 whole exome–sequenced tumors representing 14 cancer types. Samples were categorized by the statistical significance of the APOBEC mutation pattern and the magnitude of enrichment. The significance of the APOBEC mutation pattern was calculated by one-sided Fisher's exact test comparing the ratio of the number of C-to-T or C-to-G substitutions and complementary G-to-A or G-to-C substitutions that occur in and out of the APOBEC target motif (TCW or WGA) to an analogous ratio for all cytosines or guanines that reside inside and outside of the TCW or WGA motif within a sample fraction of the genome (Benjamini-Hochberg–corrected q value < 0.05). The number of tumor samples in each category is presented in each pie chart in **a**. Samples with q value > 0.05 are represented in black. These samples are excluded from the scatter graphs in **a**,**b**. Color scales indicate the magnitude of enrichment in **a** and the number of APOBEC signature mutations in **b** for samples with $q < 0.05$. Dashed lines indicate effects expected with random mutagenesis. Cancer types are abbreviated as in TCGA: cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), bladder urothelial carcinoma (BLCA), head and neck squamous cell carcinoma (HNSC), breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), uterine corpus endometrioid carcinoma (UCEC), ovarian serous cystadenocarcinoma (OV), stomach adenocarcinoma (STAD), rectum adenocarcinoma (READ), colon adenocarcinoma (COAD), prostate adenocarcinoma (PRAD), kidney renal clear-cell carcinoma (KIRC) and acute myeloid leukemia (LAML).

Spearman's $r = 0.3088$, Bonferroni-corrected $q < 0.001$ for *APOBEC3B* and *APOBEC3A*, respectively). Extending this analysis to all 2,048 tumors with available RNA-seq data across cancer types, expression of *APOBEC3B* again most strongly correlated with the number of TCW-to-TTW and TCW-to-TGW mutations per exome (**Fig. 3a** and **Supplementary Table 4**; Spearman's $r = 0.2953$, Bonferroni-corrected $q < 0.001$), with *APOBEC1*, *APOBEC3A*, *APOBEC3F* and *APOBEC3G* levels also associated, but to lesser extents (**Supplementary Fig. 11b**). Within individual cancer types, only *APOBEC3A* in breast cancer and *APOBEC3B* in breast cancer and lung adenocarcinomas showed a positive correlation between their expression and APOBEC-mediated mutagenesis (**Supplementary Fig. 11b**). However, in bladder and lung squamous cell cancers—the remaining 2 cancer types with available RNA-seq data and high APOBEC-mediated mutagenesis—median *APOBEC3B* expression was elevated by ~3-fold compared to the median of *APOBEC3B* expression in all samples (Bonferroni-corrected Mann-Whitney $q < 0.001$) (**Fig. 3b**). Thus, the APOBEC3B enzyme is probably the major candidate inducing the APOBEC mutation pattern across cancer types.

HER2E breast cancers are enriched for APOBEC mutagenesis

Several cancer types showed high levels of the APOBEC mutation pattern as well as a wide variation among individual samples, which could reflect different biological pathways leading to carcinogenesis. The greatest range of variation was observed in breast cancer, which is often divided into subtypes on the basis of differences in biomedical characteristics (see ref. 25 and references therein). To determine whether the APOBEC mutagenesis pattern is associated with specific breast cancer subtypes, we divided the samples on the basis of their PAM50 classification as presented in ref. 25. The PAM50 algorithm uses the mRNA levels of 50 differentially expressed genes to classify breast cancers into specific subtypes[26]. Four subtypes—luminal A (Lum A), luminal B (Lum B), basal-like and HER2 enriched (HER2E)—were well represented in our data set. Each subtype contained samples with a prominent APOBEC mutation pattern and a correspondingly large number of APOBEC signature mutations. However, such samples were unevenly distributed among the subtypes, occurring much more frequently in the HER2-enriched class (**Fig. 4** and **Supplementary Fig. 12**).

Unlike in breast cancer as a whole (**Fig. 3a**), no correlation between the number of APOBEC signature mutations and APOBEC mRNA levels was observed in the HER2-enriched subtype (**Supplementary Fig. 13a**). This finding could result from consistently high *APOBEC3B* expression in HER2-enriched samples (~3-fold greater than the median expression across all cancer types), which reduces the power of correlation analysis (**Supplementary Fig. 13b**). Notably, basal-like and luminal B cancers also had median *APOBEC3B* expression levels comparable to that of HER2-enriched cancers but showed significantly less APOBEC mutagenesis, suggesting that additional factors are likely as important as APOBEC expression.

The HER2-enriched subtype is reportedly associated not only with the amplification of the *ERBB2* gene locus but also with a high level of copy number variation (CNV) across the genome[25]. This feature, as well as frequent colocalization of APOBEC signature mutations with chromosome rearrangements, suggested that a direct connection might exist between enrichment with the APOBEC mutagenesis signature and the number of segmental CNVs (CNVs originating from breakage). As shown in cell culture experiments[27], increased APOBEC-induced deamination can lead to higher levels of breakage, which in turn could result in greater numbers of CNVs. Alternatively, increased breakage could provide more ssDNA substrate for APOBEC deamination. However, comparison of the number of segmental CNV breakpoints with the fold enrichment of APOBEC signature mutations in 449 breast cancer samples did not show any correlation (**Supplementary Fig. 14**). Although the underlying reason for the enrichment of the APOBEC mutagenesis pattern in the HER2-enriched subtype remains unclear, the association of this mutagenesis with a specific breast cancer subtype suggests that physiological aspects of this subtype are probably relevant.

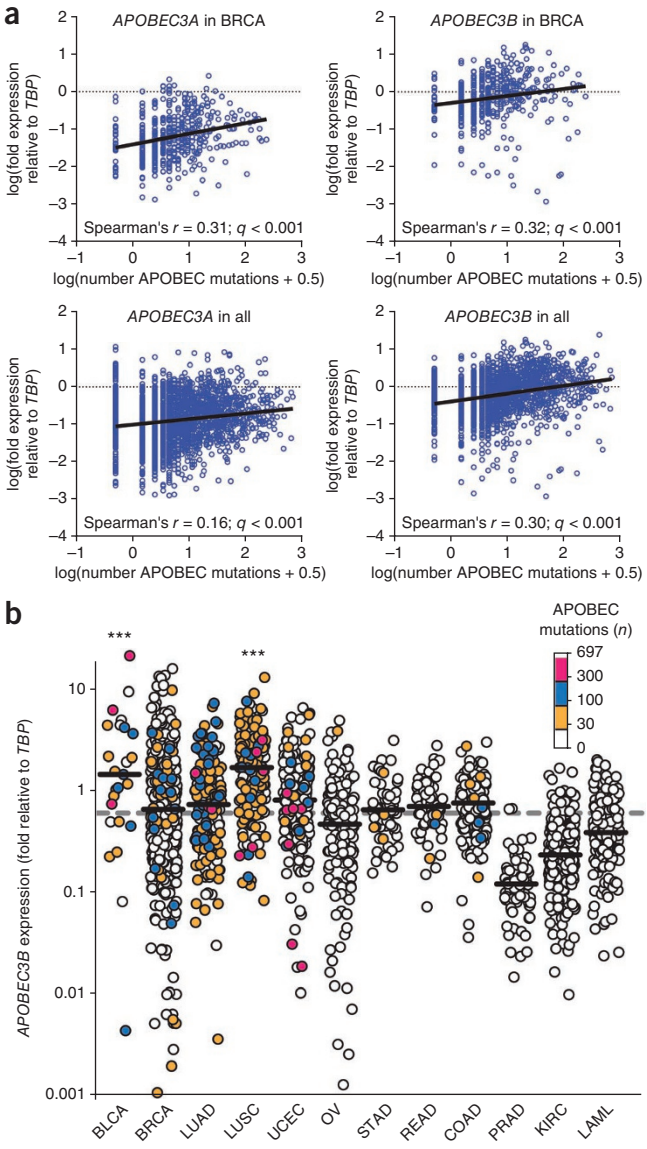

Figure 3 APOBEC mRNA levels positively correlate with the number of APOBEC signature mutations. RNA-seq–derived mRNA levels for each APOBEC family member with documented deaminase activity on DNA were standardized relative to the levels of *TBP* (encoding TATA-binding protein). APOBEC mutations refers to the number of TCW-to-TTW and TCW-to-TGW changes. (**a**) Expression (relative to *TBP*) of *APOBEC3A* and *APOBEC3B* was compared to the total number of APOBEC mutations in each exome (blue circles) in 483 breast cancers (BRCA) and in all 2,048 tumor samples (all) with available RNA-seq data by non-parametric Spearman's correlation. Graphs show log-transformed values with mutation values augmented by 0.5 to allow depiction of exomes with no observed APOBEC signature mutations. Spearman's coefficients and corresponding *q* values (two sided; corrected for multiple-testing error by the Bonferroni method) are indicated. Black lines represent linear regressions. Correlation data for other APOBECs and individual cancer types are shown in **Supplementary Figure 11**. (**b**) *APOBEC3B* transcription relative to *TBP* in 2,048 tumor samples separated by cancer type. Horizontal bars indicate the median expression levels in the cancer types. The dashed gray line indicates the median *APOBEC3B* expression level in all cancers analyzed. ***Elevated *APOBEC3B* expression in a cancer type ($q < 0.001$ by pairwise two-sided Mann-Whitney comparison of a specific cancer type to the overall distribution, corrected for multiple analyses by the Bonferroni method). Color scales indicate the number of APOBEC signature mutations in each individual exome. Individual cancer types are abbreviated as in **Figure 2**.

The APOBEC mutagenesis pattern includes cancer driver genes

An APOBEC-mediated mutagenesis pattern present in a sample or group of samples indicates that the level of this mutagenesis is significantly higher than expected if all base substitutions in cytosine (or guanine) bases have occurred randomly. However, because of the sequence specificity of APOBEC-catalyzed deamination and its tight association with ssDNA, the fraction of the genome where carcinogenic mutations can occur may escape the bulk of APOBEC-mediated mutagenesis. We therefore examined the overlap of APOBEC signature mutations with mutations that are potentially cancer drivers. Three approaches were used to identify driver mutations. First, a stringent list of probable cancer driver mutations was assembled using the online software package CRAVAT[28,29]. On the basis of multiple parameters, including the occurrence of a mutation in the Catalogue of Somatic Mutations in Cancer (COSMIC) database[30], this software calculated a probability that a given missense mutation drives cancer. Mutations having false discovery rate (FDR)-corrected *q* values of 0.05 or less were selected as likely drivers. We subsequently employed two additional 'less stringent' criteria to identify potentially carcinogenic mutations, selecting for (i) mutations that were present in the COSMIC database (as indicated by CRAVAT) and (ii) mutations that affected a subset of genes from the Cancer Gene Census—genes in which missense or nonsense mutations are considered causative in cancer[31]. Both of these less stringent definitions of driver mutations extended the spectrum of changes beyond missense mutations to include nonsense and synonymous mutations as potentially carcinogenic alterations[31,32].

Using any of these three criteria, APOBEC signature mutations occurred at a higher frequency among carcinogenic mutations in the group of samples with high APOBEC presence compared to samples in which the APOBEC mutation pattern was not detected

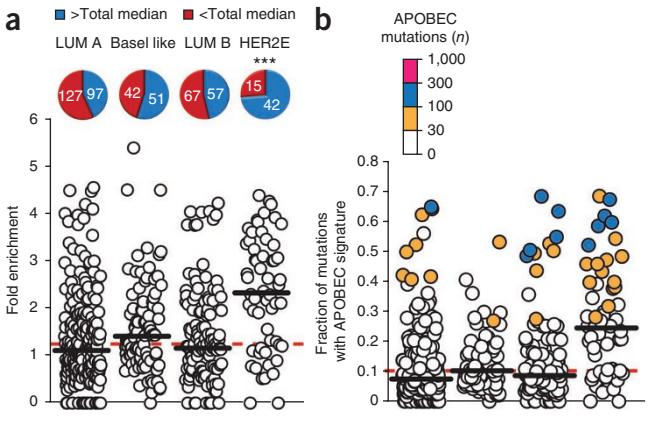

Figure 4 APOBEC mutation pattern in exome data sets from four breast cancer subtypes. (**a**,**b**) Fold enrichment (**a**) and mutation load (**b**) of the APOBEC mutation pattern were determined in each of 507 whole exome–sequenced BRCA tumors. The number of samples above (blue) and below or equal to (red) the median for all 507 exomes (dashed red lines) was determined for each cancer subtype. Horizontal black bars indicate the medians in the subtypes. ***Significant enrichment of a cancer type in samples containing a high presence of the APOBEC mutation pattern ($q < 0.001$ by pairwise two-sided χ^2 comparison of a specific cancer type to the overall distribution, corrected for analysis of multiple subtypes by the Bonferroni method). Color scales indicate the magnitude of enrichment in **a** and the number of APOBEC signature mutations in **b**. Cancer types are abbreviated as luminal A (Lum A), basal like, luminal B (Lum B) and HER2 enriched (HER2E).

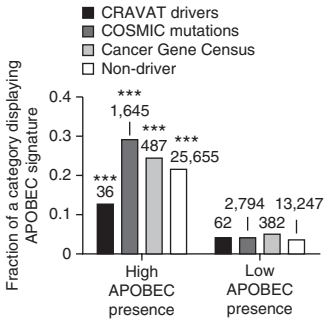

Figure 5 APOBEC signature mutations in potential cancer drivers. The fraction of potential cancer-driving mutations that have an APOBEC signature was determined for samples with high (*q* value for the enrichment of the APOBEC mutation pattern ≤ 0.05; **Fig. 2**) and low (*q* value > 0.05) presence of an exome-wide APOBEC mutation pattern. Mutations were designated as potential cancer drivers by one of three criteria: (i) Benjamini-Hochberg–corrected *q* value < 0.05 after CRAVAT analysis, (ii) listing within the COSMIC database and (iii) effect on a subset of genes in the Cancer Gene Census, whose alteration by missense or nonsense mutations can contribute to cancer. ***$P < 0.0001$ in a two-sided χ^2 test comparing the number of APOBEC and non-APOBEC signature mutations in potential cancer drivers in samples with high and low presence of the APOBEC mutation pattern for a given criterion defining a driver. Corresponding analysis for non-driver mutations is provided for comparison. The specific mutated genes are presented in **Supplementary Table 5**.

(**Fig. 5**). This implies that APOBEC signature mutations themselves can contribute to carcinogenesis in samples with a strong APOBEC mutation pattern. Further supporting this carcinogenic potential, many of the APOBEC signature mutations that are also driver mutations in CRAVAT occurred in genes that are highly mutated in the COSMIC database and are present in the Cancer Gene Census (**Supplementary Table 5**).

DISCUSSION

Determining the mutagenic factors that underlie the mix of mutations in tumors is important for a general understanding of carcinogenesis. However, this analysis is daunting, as it often requires the testing of numerous poorly defined hypotheses. Here we have developed a single detailed hypothesis—that APOBEC cytidine deaminases are a significant source of mutagenesis in human cancer genomes. This hypothesis is based on knowledge of the sequence- and single-strand specificity of APOBEC enzymes, their capacity to generate strand-coordinated mutation clusters in model systems and the extensive correlation between experimentally determined APOBEC mutagenesis patterns and the patterns of mutations in strand-coordinated clusters found in cancers. Although we cannot formally exclude the possibility that another mutagenic factor might closely mimic both the motif and mutagenic specificities of the APOBEC mutation pattern, there is yet no indication that such a factor exists. Furthermore, our observed correlation between the APOBEC mutagenesis pattern and APOBEC expression in cancer samples provides strong support for our hypothesis. Additional support could be sought by analyzing correlations with the germline genotypes of patients, as soon as such information would be available.

Our TCGA-based analysis indicates a widespread APOBEC mutagenesis pattern and suggests that this pattern is associated with biological mechanisms underlying carcinogenesis. With our approach, we establish a resource for identifying this pattern in the rapidly growing TCGA database as well as in other databases of genome- or exome-wide mutations in humans. In addition, the predominance of APOBEC signature mutations across tumors of multiple cancer types underscores the importance of validating the specific APOBEC proteins responsible for mutagenesis and evaluating the presence of this mutagenesis in other types and subtypes of cancers, the stage(s) of cancer development that are most prone to APOBEC mutagenesis and the relative impact of this mutagenesis on genome changes that lead to cancer.

Multiple mechanisms could facilitate APOBEC-mediated mutagenesis. Environmental and physiological factors might trigger and/or support mutagenesis by (i) affecting the cellular abundance or activity of APOBEC proteins, (ii) altering access to nuclear DNA and (iii) increasing the amount and/or persistence of ssDNA substrates for APOBEC-mediated cytidine deamination. Our study and previous analyses suggest that the level of *APOBEC3B* transcription affects APOBEC-mediated mutagenesis. How higher *APOBEC3B* transcript levels are established remains unclear. Among the factors that could increase the amount of APOBEC protein(s) is the presence of the viral and retrotransposable elements that these enzymes restrict[6,33]. Such factors can stimulate APOBEC expression through a complex network of innate immunity signaling, involving components such as Toll-like receptors, interferons, interleukins and even the 'usual suspect' in carcinogenesis, the p53 protein[34–37]. Infection with several viruses[38] as well as retrotransposition[39] are associated with carcinogenesis; however, the mechanisms of this association are far from clear. A potential relationship between APOBEC-mediated mutagenesis and viral infection is appealing, as cervical and head and neck cancers, which are highly associated with human papillomavirus (HPV) infection, showed strong enrichment of APOBEC-mediated mutagenesis.

Despite a positive correlation between *APOBEC3B* expression and APOBEC-mediated mutagenesis, the extent of the association is relatively small (Spearman's $r = 0.30$). Thus, other factors probably contribute more prominently to APOBEC-mediated mutagenesis. Factors that could increase the abundance and persistence of ssDNA include DNA-damaging agents[40,41] as well as defects in DNA transactions that impede break repair[42,43] and replication integrity[44,45]. Our work in yeast demonstrated that proliferation in the presence of an alkylation agent leads to the formation of ssDNA at DSB sites and dysfunctional forks and subsequently results in mutation clusters[13]. Notably, a high level of APOBEC-mediated deamination may in itself lead to DNA breakage[27], which could generate a ssDNA substrate for APOBEC-mediated hypermutation. It is generally acknowledged that carcinogenesis requires the accumulation of multiple genetic changes[46]. As discussed in ref. 13, simultaneous mutations in scattered stretches of ssDNA formed at DSBs, replication forks and other cell contexts would be excellent substrates for APOBEC-mediated mutagenesis, which in turn might produce multiple changes without excessive genome-wide mutation and provide a means to accumulate multiple carcinogenic mutations in a single or a few generations.

URLs. TCGA data portal, https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp; database of Genotypes and Phenotypes (dbGaP), http://www.ncbi.nlm.nih.gov/gap; 21 breast cancer genomes, ftp://ftp.sanger.ac.uk/pub/cancer/Nik-ZainalEtAl; 9 colorectal adenocarcinoma genomes, http://www.broadinstitute.org/~lawrence/crc/CRC9.genomic.v3.maf; CRAVAT, http://www.cravat.us/; COSMIC database, http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/; Cancer Gene Census, http://cancer.sanger.ac.uk/cancergenome/projects/census/.

METHODS

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

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

ACKNOWLEDGMENTS

We would like to thank J. Taylor, P. Wade and D. Zaykin for helpful discussions and critical reading of the manuscript. The results published here are in part based on data generated by the TCGA project established by the National Cancer Institute and the National Human Genome Research Institute (database of Genotypes and Phenotypes (dbGaP) accession phs000178.v8.p7). The work was supported in part by the Intramural Research Program of the US National Institutes of Health, the National Institute of Environmental Health Sciences (project ES065073 to M.A.R.; contract GS-23F-9806H and order HHSN273201000086U to R.R.S.) and by the National Human Genome Research Institute (grant U54HG003067 to G.G.).

AUTHOR CONTRIBUTIONS

S.A.R., G.G. and D.A.G. designed the study. S.A.R., M.S.L., L.J.K., S.A.G., D.F., P.S., A.K., G.V.K., S.L.C., G.S., S.H., R.R.S., M.A.R., G.G. and D.A.G. contributed to data analysis. S.A.R. and D.A.G. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).
2. Loeb, L.A. Mutator phenotype may be required for multistage carcinogenesis. *Cancer Res.* **51**, 3075–3079 (1991).
3. Luch, A. Nature and nurture—lessons from chemical carcinogenesis. *Nat. Rev. Cancer* **5**, 113–125 (2005).
4. Conticello, S.G. Creative deaminases, self-inflicted damage, and genome evolution. *Ann. NY Acad. Sci.* **1267**, 79–85 (2012).
5. Pavri, R. & Nussenzweig, M.C. AID targeting in antibody diversity. *Adv. Immunol.* **110**, 1–26 (2011).
6. Smith, H.C., Bennett, R.P., Kizilyer, A., McDougall, W.M. & Prohaska, K.M. Functions and regulation of the APOBEC family of proteins. *Semin. Cell Dev. Biol.* **23**, 258–268 (2012).
7. Suspène, R. *et al.* Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. *Proc. Natl. Acad. Sci. USA* **108**, 4858–4863 (2011).
8. Shinohara, M. *et al.* APOBEC3B can impair genomic stability by inducing base substitutions in genomic DNA in human cells. *Sci. Rep.* **2**, 806 (2012).
9. Burns, M.B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494**, 366–370 (2013).
10. Stephens, P. *et al.* A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer. *Nat. Genet.* **37**, 590–592 (2005).
11. Beale, R.C. *et al.* Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra *in vivo*. *J. Mol. Biol.* **337**, 585–596 (2004).
12. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
13. Roberts, S.A. *et al.* Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. *Mol. Cell* **46**, 424–435 (2012).
14. Drier, Y. *et al.* Somatic rearrangements across cancer reveal classes of samples with distinct patterns of DNA breakage and rearrangement-induced hypermutability. *Genome Res.* **23**, 228–235 (2013).
15. Gibbs, P.E. & Lawrence, C.W. Novel mutagenic properties of abasic sites in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **251**, 229–236 (1995).
16. Simonelli, V., Narciso, L., Dogliotti, E. & Fortini, P. Base excision repair intermediates are mutagenic in mammalian cells. *Nucleic Acids Res.* **33**, 4404–4411 (2005).
17. Chan, K. *et al.* Base damage within single-strand DNA underlies *in vivo* hypermutability induced by a ubiquitous environmental agent. *PLoS Genet.* **8**, e1003149 (2012).
18. Senavirathne, G. *et al.* Single-stranded DNA scanning and deamination by APOBEC3G cytidine deaminase at single molecule resolution. *J. Biol. Chem.* **287**, 15826–15835 (2012).
19. Chelico, L., Pham, P. & Goodman, M.F. Mechanisms of APOBEC3G-catalyzed processive deamination of deoxycytidine on single-stranded DNA. *Nat. Struct. Mol. Biol.* **16**, 454–455, author reply 455–456 (2009).
20. Barbieri, C.E. *et al.* Exome sequencing identifies recurrent *SPOP*, *FOXA1* and *MED12* mutations in prostate cancer. *Nat. Genet.* **44**, 685–689 (2012).
21. Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157–1160 (2011).
22. Bass, A.J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent *VTI1A-TCF7L2* fusion. *Nat. Genet.* **43**, 964–968 (2011).
23. Shammas, M.A. *et al.* Dysfunctional homologous recombination mediates genomic instability and progression in myeloma. *Blood* **113**, 2290–2297 (2009).
24. Liu, P., Carvalho, C.M., Hastings, P.J. & Lupski, J.R. Mechanisms for recurrent and complex human genomic rearrangements. *Curr. Opin. Genet. Dev.* **22**, 211–220 (2012).
25. TCGA. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
26. Parker, J.S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).
27. Landry, S., Narvaiza, I., Linfesty, D.C. & Weitzman, M.D. APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. *EMBO Rep.* **12**, 444–450 (2011).
28. Carter, H. *et al.* Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.* **69**, 6660–6667 (2009).
29. Douville, C. *et al.* CRAVAT: Cancer-Related Analysis of VAriants Toolkit. *Bioinformatics* **29**, 647–648 (2013).
30. Forbes, S.A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc. Hum. Genet.* Chapter 10, Unit 10.11 (2008).
31. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
32. Lampson, B.L. *et al.* Rare codons regulate *KRas* oncogenesis. *Curr. Biol.* **23**, 70–75 (2012).
33. Schumacher, A.J., Nissley, D.V. & Harris, R.S. APOBEC3G hypermutates genomic DNA and inhibits Ty1 retrotransposition in yeast. *Proc. Natl. Acad. Sci. USA* **102**, 9854–9859 (2005).
34. Einav, U. *et al.* Gene expression analysis reveals a strong signature of an interferon-induced pathway in childhood lymphoblastic leukemia as well as in breast and ovarian cancer. *Oncogene* **24**, 6367–6375 (2005).
35. Refsland, E.W. *et al.* Quantitative profiling of the full APOBEC3 mRNA repertoire in lymphocytes and tissues: implications for HIV-1 restriction. *Nucleic Acids Res.* **38**, 4274–4284 (2010).
36. Menendez, D., Shatz, M. & Resnick, M.A. Interactions between the tumor suppressor p53 and immune responses. *Curr. Opin. Oncol.* **25**, 85–92 (2013).
37. Zhou, L. *et al.* Activation of toll-like receptor-3 induces interferon-λ expression in human neuronal cells. *Neuroscience* **159**, 629–637 (2009).
38. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Biological agents. Volume 100 B. A review of human carcinogens. *IARC Monogr. Eval. Carcinog. Risks Hum.* **100**, 1–441 (2012).
39. Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971 (2012).
40. Lopes, M., Foiani, M. & Sogo, J.M. Multiple mechanisms control chromosome integrity after replication fork uncoupling and restart at irreparable UV lesions. *Mol. Cell* **21**, 15–27 (2006).
41. Pagès, V. & Fuchs, R.P. Uncoupling of leading- and lagging-strand DNA replication during lesion bypass *in vivo*. *Science* **300**, 1300–1303 (2003).
42. Bouwman, P. *et al.* 53BP1 loss rescues BRCA1 deficiency and is associated with triple-negative and BRCA-mutated breast cancers. *Nat. Struct. Mol. Biol.* **17**, 688–695 (2010).
43. Bunting, S.F. *et al.* 53BP1 inhibits homologous recombination in Brca1-deficient cells by blocking resection of DNA breaks. *Cell* **141**, 243–254 (2010).
44. Bando, M. *et al.* Csm3, Tof1, and Mrc1 form a heterotrimeric mediator complex that associates with DNA replication forks. *J. Biol. Chem.* **284**, 34355–34365 (2009).
45. Katou, Y. *et al.* S-phase checkpoint proteins Tof1 and Mrc1 form a stable replication-pausing complex. *Nature* **424**, 1078–1083 (2003).
46. Yates, L.R. & Campbell, P.J. Evolution of the cancer genome. *Nat. Rev. Genet.* **13**, 795–806 (2012).
47. Chapman, M.A. *et al.* Initial genome sequencing and analysis of multiple myeloma. *Nature* **471**, 467–472 (2011).
48. Berger, M.F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).

ONLINE METHODS

Genome and exome data sets. Genome and exome data sets were obtained from publications[20,21] or from the TCGA data portal (see URLs; Controlled Data Access HTTP Directory). The catalog of base substitutions identified by whole-genome sequencing in 21 breast cancers was downloaded from the website provided in ref. 12 (see URLs).

Hyperlinks to TCGA data sets and references to published mutation lists are provided in **Supplementary Table 3**.

Cluster analysis. Clusters and colocalization between clusters and rearrangement breakpoints in whole-genome data sets were identified as described in ref. 13. Analysis of mutation clustering in exomes was conducted similarly to that in whole-genome data sets. Briefly, we first filtered out mutations identical to variants in dbSNP. These SNPs generally constituted a small percentage (0.9–12.1%) of all exome mutations for a given cancer type. However, LUSC, KIRC, PRAD and STAD samples contained somewhat higher numbers of mutations identical to variants in dbSNP (19.5–25.1%). Notably, each prefiltered mutation was included in the total number of mutations in the genome, which would thereby only increase the *P* values of clusters. We next identified groups of closely spaced mutations (with at most 10 nucleotides between neighbors), which we categorized as complex. Complex mutations are likely to arise from a mutagenesis event triggered by translesion synthesis across a single DNA lesion[49,50]. Each complex mutation was counted as a single mutation event. Then, all groups of at least two mutations in which neighboring changes were separated by 10 kb or less were identified. The *P* value for each group was calculated under the assumption that all mutations were distributed randomly across the genome. The total number of mutations in the genome was estimated as 100-fold greater than the number of mutations in the exome, including those identical to variants in dbSNP.

A cluster *P* value was defined as the probability of observing $k-1$ mutations in $x-1$ or fewer base pairs and was calculated using a negative binomial distribution as follows:

$$\text{cluster } P \text{ value} = \sum_{j=0}^{x-k} \left(\frac{(k-1)+j-1}{j} \right) (1-\pi)^j \pi^{k-1} \quad (1)$$

where x denotes the size of the mutation cluster (size is defined as the number of nucleotides in the region starting at the position of the first mutation and ending at the position of the last mutation of a cluster), k denotes the number of mutations observed in a cluster and π denotes the probability of finding a mutation at any random location in the genome calculated as

$$\pi = \frac{n}{G} \quad (2)$$

where n denotes the total number of mutations in a genome and G denotes the total genome size (number of nucleotides).

Groups of mutations were identified as clusters if the calculated *P* value was no greater than 1×10^{-4}. A recursive algorithm was used, such that all clusters that met the *P*-value criterion were identified, even if they were part of a larger group that fit the spacing criterion but did not meet the probability cutoff. Individual mutations and clusters with *P* values no greater than 1×10^{-4} were classified as follows: clusters in which all mutations resulted from a change of the same kind of nucleotide were defined as strand-coordinated clusters, and clusters containing mutations of at least two different kinds of bases were called non-coordinated clusters. Mutations that did not belong to a cluster were classified as scattered, and the other category was named clustered.

Detecting an APOBEC mutation pattern. *Enrichment*. The numeric value of enrichment *E* characterizing the strength of mutagenesis at the TCW motif in mutation clusters was calculated as

$$E = \frac{\text{mutations}_{\underline{\text{TC}}\text{W}} \times \text{context}_{\underline{\text{C}}\text{ (or G)}}}{\text{mutations}_{\underline{\text{C}}\text{ (or G)}} \times \text{context}_{\text{TC W}}}$$

where $\text{mutations}_{\underline{\text{TC}}\text{W}}$ is the number of mutated cytosines (and guanines) falling in a TCW (or WGA) motif, $\text{mutations}_{\underline{\text{C}}\text{ (or G)}}$ is the total number of mutated cytosines (or guanines), $\text{context}_{\underline{\text{TC}}\text{W}}$ is the total number of TCW (or WGA) motifs within the 41-nucleotides region centered on the mutated cytosines (and guanines) and $\text{context}_{\underline{\text{C}}\text{ (or G)}}$ is the total number of cytosines (or guanines) within the area 41-nucleotides region centered on the mutated cytosines (or guanines).

In determining the presence of the APOBEC mutagenesis pattern, enrichment was calculated as above, except that only specific base substitutions (TCW to TTW or TGW, WGA to WAA or WCA, C to T or G, and G to A or C) were included.

Fisher's exact test. Statistical evaluation of the over-representation of APOBEC signature mutations in each sample was performed using a one-sided Fisher's exact test comparing the ratio of the number of cytosine-to-thymine or cytosine-to-guanine substitutions and guanine-to-adenine or guanine-to-cytosine substitutions that occurred in and out of the APOBEC target motif (TCW or WGA) to an analogous ratio for all cytosines and guanines that reside inside and outside of the TCW or WGA motif within a sample fraction of the genome. *P* values calculated for multiple samples or multiple comparisons were corrected using the Benjamini-Hochberg method or the Bonferroni method as indicated[51]. Only corrected *q* values of <0.05 were considered significant.

Determining the number of breakpoints associated with segmental CNVs. The number of breakpoints associated with segmental CNVs was determined on the basis of TCGA Affymetrix Genome-Wide Human SNP Array 6.0 analysis of 449 breast cancer (BRCA) samples. Breakpoints were identified as pairs of adjacent segments on the same chromosome with a difference in copy ratio of >0.1. Any segments with fewer than five probes were removed from analysis as being likely due to technical noise.

Defining cancer driver mutations. The online software package CRAVAT[28,29] (see URLs) was used to identify potential cancer-driving mutations among missense mutations. For acute myeloid leukemia (LAML), breast (BRCA), colorectal (COAD), ovarian (OV), rectal (READ), stomach (STAD) and uterine endometrial (UCEC) cancers, the matched tissue-specific passenger mutation profile provided in the CRAVAT package was used. For all other cancer types for which a tissue-specific profile was unavailable, a generic profile was used. CRAVAT outputs included a CHASM score, a *P* value indicating the likelihood of a mutation being a driver and a Benjamini-Hochberg (FDR) *q* value to correct for multiple-hypothesis testing. In our analysis, potential cancer drivers were identified as those mutations with a Benjamini-Hochberg *q* value no greater than 0.05. In addition to CRAVAT analysis, two other metrics to identify driver mutations were considered: the occurrence of mutations in the COSMIC database and alteration of genes listed in the Cancer Gene Census[31], a curated list of genes whose alteration has been shown to be causative in at least some cancers. For the latter metric, only genes in the Cancer Gene Census where missense and nonsense mutations are known to be involved in carcinogenesis were used to identify potential drivers.

Analysis of controlled-access data. The complete list of analyzed mutations used to create all figures and make conclusions in this paper will be submitted as a TCGA substudy and will be available through controlled access to dbGaP study phs000178.v8.p7. The file will be in TCGA MAF format. In addition to the information from the original TCGA MAFs (**Supplementary Table 3**), the file will contain the results of mutation cluster analysis, sequence context of mutations and CRAVAT analysis. Before the acceptance of the substudy by TCGA, the file will be available to investigators after they acquire access to controlled TCGA data levels in coordination with D.A.G.

49. Harfe, B.D. & Jinks-Robertson, S. DNA polymerase ζ introduces multiple mutations when bypassing spontaneous DNA damage in *Saccharomyces cerevisiae*. *Mol. Cell* **6**, 1491–1499 (2000).
50. Sakamoto, A.N. *et al.* Mutator alleles of yeast DNA polymerase ζ. *DNA Repair (Amst.)* **6**, 1829–1838 (2007).
51. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc., B* **57**, 289–300 (1995).