APOBEC3A, not APOBEC3B, drives deaminase mutagenesis in human gastric epithelium

Authors

Yohan An[1,*], Ji-Hyun Lee[2,*], Joonoh Lim[1,3,*], Jeonghwan Youk[1,4], Seongyeol Park[1], Ji-Hyung Park[1,5], Kijong Yi[1,3], Taewoo Kim[1], Chang Hyun Nam[1], Won Hee Lee[1], Soo A Oh[3], Yoo Jin Bae[6], Junehwak Lee[7], Jung Woo Park[7], Jie-Hyun Kim[6], Hyunki Kim[8], Hugo Snippert[9], Bon-Kyoung Koo[2,10,$], and Young Seok Ju[1,$]

Affiliations

1 Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea
2 Center for Genome Engineering, Institute for Basic Science, 55, Expo-ro, Yuseong-gu, Daejeon, 34126, Republic of Korea
3 Inocras Inc., 6330 Nancy Ridge Drive, San Diego, CA 92121, United States
4 Department of Internal Medicine, Seoul National University Hospital, Seoul, Republic of Korea
5 Department of Neurology, Seoul National University Bundang Hospital, 172, Dolma-ro, Bundang-gu, Seongnam 13605, Korea
6 Department of Internal Medicine, Division of Gastroenterology, Gangnam Severance Hospital, Yonsei University College of Medicine, 20, Eonju-ro 63-gil, Gangnam-gu, Seoul 06229, Korea
7 Korea Institute of Science and Technology Information, Daejeon 34141, Republic of Korea
8 Department of Pathology, Yonsei University College of Medicine, 50-1 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea
9 Oncode Institute, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, the Netherlands
10 Department of Life Sciences, Pohang University of Science and Technology (POSTECH), 77, Chengam-ro, Nam-gu, Pohang 37673, Republic of Korea
*Co-first authors with equal contribution
$Co-corresponding authors

Corresponding authors

Bon-Kyoung Koo, PhD.

koobk@ibs.re.kr

Director

Center for Genome Engineering, Institute for Basic Science, 55, Expo-ro, Yuseong-gu, Daejeon, 34126, Republic of Korea

Young Seok Ju, MD, PhD.

ysju@kaist.ac.kr

Associate Professor

Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea

Abstract

Cancer genomes frequently carry APOBEC (apolipoprotein B mRNA editing catalytic polypeptide-like)-associated DNA mutations, suggesting APOBEC enzymes as innate mutagens during cancer initiation and/or evolution. However, the pure mutagenic impacts of the specific enzymes among this family that are responsible for APOBEC-associated mutagenesis remain unclear, particularly the comparative mutagenic activities of *APOBEC3A* and *APOBEC3B*. Here, we investigated the mutagenic contributions of these enzymes through whole-genome sequencing of human normal gastric organoid lines carrying doxycycline-inducible *APOBEC3A* or *APOBEC3B* cassettes. Our findings demonstrated that transcriptional *APOBEC3A* upregulation led to the acquisition of a massive number of genomic mutations in a few cell cycles. By contrast, *APOBEC3B* upregulation did not generate a substantial number of mutations in gastric epithelium. *APOBEC3B*-associated mutagenesis remained insignificant even after a combined inactivation of *TP53*. Based on the spectrum of acquired mutations after *APOBEC3A* upregulation, we further analyzed *APOBEC3A*-associated mutational signatures, encompassing indels mainly composed of 1bp deletions, characteristics of clustered mutations, and selective pressures operative on cells carrying the mutations. Our observations provide a clear foundation for understanding the mutational impact of APOBEC enzymes in human cells.

Introduction

Large-scale cancer genome studies have revealed various mutational processes in human somatic cells[1]. In ~75% of human cancer types, including breast, bladder, and stomach cancers, APOBEC-associated mutations are frequently detected in the genome[1]. The APOBEC-associated mutational activity is not confined to cancer cells but is also observed in non-neoplastic normal cells, particularly within the epithelium of the bladder, bronchus, and small intestine[2–4].

APOBEC enzymes are well known for their role in innate immunity, where they destabilize pathogens by inducing DNA and RNA deamination[5]. Recently, they have gained wide recognition as mutagens that also affect host genomes. In particular, APOBEC-associated mutations in DNA are characterized by predominantly C>T and C>G base substitutions enriched in Tp<u>C</u>pN context (with the mutated cytosine is underlined), and highly consistent across different cancer types[1]. This mutation pattern has been summarized and cataloged as mutational signatures SBS2 and SBS13 in the COSMIC database[6]. Given the similarity in the characteristic sequence-context dependency, their mutagenic origin was associated with APOBEC activity[7–9]. Of the 11 APOBEC family genes in the human genome, *APOBEC3A* (**A3A**) and *APOBEC3B* (**A3B**) have been highlighted as potential contributors to this mutagenic activity[9]. However, how each of them contributes to the mutagenesis remains to be elucidated[10].

Complete elucidation of mutagenic activities of A3A and A3B has been challenging, especially because APOBEC-associated mutagenesis is not constitutively active but episodic in the somatic lineages over a lifetime[11]. Indeed, the transient activity of APOBEC enzymes during cancer development and/or clonal evolution, and even in non-malignant tissues, has been well documented[4,11–13]. The potential influence of other cancer driver genes (e.g., the loss of *TP53* function in cancer cells) also complicates the investigation of pure mutational impact of APOBEC-enzymes on cancer genomes[14–18]. Understanding APOBEC-associated mutagenic activity in non-human model systems[10], such as *Mus musculus*[19,20] or *Saccharomyces cerevisiae*[21], may not directly translate to human cells due to differences in genomic backgrounds.

In this regard, we explored the mutagenic activity of A3A and A3B by establishing human non-neoplastic gastric organoid lines with doxycycline-inducible APOBEC activity. Combining single-cell clonalization and whole-genome sequencing (**WGS**)[22–24], we comprehensively investigated the pure mutagenic impact of A3A and A3B at single-cell resolution.

Results

Doxycycline-inducible *APOBEC3A* and *APOBEC3B* in normal gastric organoids

To regulate the transcription of *A3A* and *A3B* in non-neoplastic human cells, we established genome-engineered gastric organoid lines, referred to as hGO_{iA3A} and hGO_{iA3B}, respectively. In these lines, doxycycline treatment simultaneously triggers the upregulation of (1) fluorescence and (2) either the A3A or A3B enzyme (**Fig. 1a**). To this end, we utilized *PiggyBac* transposon system[25,26] to integrate two independent constructs into the genome of gastric organoids: (1) a cassette expressing rtTA and hygromycin B resistance protein (CMV-rtTA-HygR) and (2) a cassette expressing APOBEC enzymes and fluorescence protein (TRE-APOBEC (A3A or A3B)-IRES-mCherry). Successfully engineered organoids were selected through hygromycin treatment and subsequently clonalized into lines. WGS confirmed clonality (originating from a single founder cell; **Extended Data Fig. 1**), genomic insertion positions of the two constructs (**Supplementary Table 1**), the number of integrated cassettes (**Supplementary Table 2**), and APOBEC gene sequence conservation of the inserted constructs in the hGO_{iA3A} and hGO_{iA3B} lines (data not shown).

Upon doxycycline treatment, upregulation of mCherry fluorescence and APOBEC transcripts were clearly observed within 48 hours (**Fig. 1b**). Here, we treated 0.1 µg/ml and 3 µg/ml of doxycycline for low- and high-level induction, respectively. Bulk RNA-seq analysis revealed that gene expression levels of A3A or A3B ranged from 400 to 4,000 transcripts per millions (**TPMs**), representing a several hundred-fold increase compared to background levels (**Fig. 1c**). The range of expression levels is comparable to the transcriptional levels in single cells carrying APOBEC expression bursts (**Extended Data Fig. 2**).

Following APOBEC induction, the cell viability of the lines was substantially compromised, suggesting a negative impact of APOBEC enzymes on mitosis (**Fig. 1d**). Differentially expressed genes after APOBEC induction included genes related to cell cycle (e.g., *CDKN1A*[27] and *CDC20*[28]) and immune response (e.g., *IL8*[29] and *CCL20*[30]) in both the hGO_{iA3A} and hGO_{iA3B} models (**Fig. 1e**).

We also observed the nuclear localization of A3A and A3B enzymes following protein translation (**Fig. 1f**). The increased number of □-H2AX foci after APOBEC upregulation confirmed that the APOBEC enzymes induced double-strand DNA breaks (**DSBs**) in the cells (**Fig. 1g**), as previously observed[31,32].

APOBEC3A induces genomic mutations

To assess the mutagenic impacts of A3A and A3B, we investigated acquired mutations after APOBEC induction in the hGO_{iA3A} and hGO_{iA3B} lines (**Fig. 1a**). Interestingly, a substantial number of base substitutions, whose spectrum is concordant with a linear combination of SBS2 and SBS13, were detected in the hGO_{iA3A} lines (**Fig. 2a,b**). On average, the hGO_{iA3A} lines with 0.1 µg/ml and 3 µg/ml doxycycline exhibited 267 and 2,448 base substitutions genome-wide, respectively (0.1 µg/ml: 95% CI -107-505, 3 µg/ml: 95% CI 1,052-3,708) (**Fig. 2a and Extended Data Fig. 3a; Supplementary Table 3**). By contrast, the hGO_{iA3B} lines did not show a notable number of APOBEC-associated mutations (0.1 µg/ml: 95% CI 0-1) (**Fig. 2a and Extended Data Fig. 3b; Supplementary Table 3**).

Despite the minimal mutational impact on the organoids, we believe that the A3B enzyme possesses DNA deamination activity. In our *in vitro* experiments, where cell-free DNA fragments were directly exposed to A3B enzymes, we observed intensive cytosine deamination changes in ~99% of unmethylated cytosines, regardless of surrounding sequence adjacent to the deaminated cytosines (**Extended Data Fig. 4**). Given A3B-induced DNA damage in the nucleus (**Fig. 1g**), this active deamination activity suggests either that the DNA repair processes in organoids can effectively correct A3B-induced deaminated cytosines (uracils) almost completely, leaving minimal DNA mutations, or that A3B enzymes may lose the cytosine deamination activity in the nuclear environment.

Then, we evaluated the possibility of detection bias in our clonalization system, which may underestimate the mutational impact of APOBEC enzymes. Here, if APOBEC induction substantially compromises proliferation potential, the damaged cells cannot be genome-sequenced. To directly detect mutations without the clonalization step, we applied a duplex-DNA sequencing technique to directly capture mutations in APOBEC-activated lines, even in single cells unable to proliferate further[33].

In the population of cells exposed to A3A, the duplex-DNA sequencing technique revealed a higher mutational burden of APOBEC-associated base substitutions (0.1 µg/ml: 95% CI 4,590-6,038, 3 µg/ml: 95% CI 9,316-10,077; **Fig. 2a and Extended Data Fig. 5a**; **Supplementary Table 4**). Our observation confirms that (1) A3A is a strong mutator enzyme, and (2) its mutational impact could be greater than what is observed in the clonal lines. Interestingly, APOBEC-associated base substitutions were still minimal in the population of cells exposed to A3B (0.1 µg/ml: 95% CI 136-352, 3 µg/ml: 95% CI 79-133; **Fig. 2a and Extended Data Fig. 5b**; **Supplementary Table 4**), suggesting that the insignificant A3B-associated mutations were not stemming from negative selection during clonal expansion.

To understand whether a combined inactivation of tumor suppressor genes is necessary for the acquisition of APOBEC-associated mutations, particularly A3B, we further knocked out *TP53* in the hGO_{iA3A} and hGO_{iA3B} lines (referred to as TP53KO-hGO_{iA3A} and TP53KO-hGO_{iA3B} lines, respectively; **Fig. 1a**).

From the TP53KO-hGO_{iA3A} lines, we observed a similar burden of APOBEC-associated mutations to what was observed in the hGO_{iA3A} lines (0.1 µg/ml: 95% CI 785-1,917, 3 µg/ml: CI 152-3,091; **Fig. 2a and Extended Data Fig. 3a**; **Supplementary Table 3**), suggesting that the loss of function of *TP53* does not accelerate A3A-associated mutagenesis. Similarly, the TP53KO-hGO_{iA3B} lines also showed results consistent with the hGO_{iA3B} lines, with a lack of APOBEC-associated mutations (0.1 µg/ml: 95% CI 11-19, 3 µg/ml: 95% CI 2-12; **Fig. 2a and Extended Data Fig. 3b**; **Supplementary Table 3**).

Despite the increased occurrence of DSBs in nuclear DNA following APOBEC exposure (**Fig. 1g**), we did not observe an increase in structural variations (**SVs**; **Supplementary Table 3**). The few APOBEC-induced SVs were also evident in the TP53KO-hGO_{iA3A} and TP53KO-hGO_{iA3B} lines. We speculate that APOBEC-induced DSBs were mostly repaired correctly within the cells.

Both APOBEC3A and APOBEC3B induce C>U RNA editing

Despite the distinct mutagenic impact on DNA, both A3A and A3B demonstrated deaminase activity on RNA, leading to C>U RNA editing events in the clones. In both the hGO_{iA3A} and hGO_{iA3B} lines, we observed an approximately 8-fold and 30-fold increase in C>U RNA editing following treatment with 0.1 µg/ml and 3 µg/ml doxycycline for 48 hours, respectively, compared to baseline levels (**Fig. 2c,e**). The spectra of C>U RNA-editing deviated from SBS2 for DNA mutations and varied between A3A and A3B (**Fig. 2d,f**). Specifically: (1) both A3A and A3B exhibited reduced specificity for the Up<u>C</u>pU context (equivalent to Tp<u>C</u>pT in DNA); (2) A3A displayed enhanced specificity for the Up<u>C</u>pG context (equivalent to Tp<u>C</u>pG in DNA); and (3) A3B showed decreased specificity for the Up<u>C</u>pN context (equivalent to Tp<u>C</u>pN in DNA). Further *de novo* extraction of RNA editing signatures revealed distinct context-specific characteristics of the A3A and A3B-associated RNA-editing (**Extended Data Fig. 6a,b**).

Characterization of APOBEC-associated mutations

Using the mutational profiles from hGO_{iA3A} lines, we further examined the characteristics of A3A-associated mutational signatures. A3A-associated mutations in the TpCpA context were 2.7 times more abundant following pyrimidine bases (YpTpCpA) compared to purine bases (RpTpCpA) (**Fig. 3a**). This preference closely mirrors the enrichment level observed in cancer tissues carrying APOBEC-induced hypermutations (2.5 times), indicating that A3A is likely a key driver of the hypermutations.

Further, in line with previous reports from cancer[21,34], indels attributable to the ID9 signature (**Fig. 3b**) showed a positive correlation with the number of A3A-associated base substitutions in hGO_{iA3A} clones. It was estimated that one ID9 indel event occurs for every 333 SBS2/13 base substitutions (**Fig. 3c**).

Interestingly, base substitutions attributable clock-like mutational signatures (SBS5 and SBS40) also demonstrated a positive correlation with A3A-associated base substitutions (SBS2 and SBS13; **Fig. 3d**). This suggests that A3A-induced genomic damage indirectly promotes error-prone DNA repair processes across the genome. However, this association was absent in clones carrying *TP53* truncating mutations (**Fig. 3d**). We speculate that the difference stems from differential DNA repair pathways according to the activities of TP53[35]. Notably, the transcription of DNA polymerase eta (*POLH*), which is involved in translesion synthesis (TLS)[36,37] responsible for SBS5 and SBS40 mutations[17], was exclusively observed in clones without *TP53* knockout (**Fig. 3e**).

In cancer genomes, APOBEC-induced localized hypermutation events are frequently observed[38,39]. In our clones, approximately 5% of the 29,650 acquired single-nucleotide variants were clustered within 1 kbp, a frequency 100 times higher than random chance (**Fig. 4a**). Such clustered mutation events can be classified into *omikli* and *kataegis* (usually composed of 2-3 and >3 mutations, respectively) based on statistical models, not by absolute distance[40]. A3A-associated clustered mutation events were distributed across 615 *omikli* and 109 *kataegis* events. The observed *kataegis* events ranged from 4 to 22 base substitutions (median = 5). Contrary to the ~36% of *kataegis* events in cancers reportedly being observed with nearby SVs[39,40], rearrangement-associated hypermutations were rare in our data (2.8%; 3 out of 109 events). Similarly, the *kataegis* in our clones were independent of other known *kataegis*-shaping genomic events, such as anaplastic DNA bridge[41] and/or extrachromosomal DNA (ecDNA)[40]. Therefore, we believe that the clustered mutations found in our clones were driven by pure activity of A3A.

The relative frequencies of *omikli* and *kataegis* remained consistent at 2.4% (95% CI: 1.7-3.1%) and 0.4% (95% CI: 0.22-0.58%) of all A3A-associated mutation events, respectively,

with each isolated and clustered mutations counted as a single event (**Fig. 4b**; **Supplementary Table 5**). These ratios were more or less constant in clones, regardless of the A3A expression levels and/or *TP53* mutational status. Interestingly, the frequencies of *omikli* and *kataegis* were comparable to the SV-unrelated *omikli* and *kataegis* frequencies in cancers (**Fig. 4b**; **Supplementary Table 5**).

Within *kataegis* regions, cytosine alterations did not always occur in the TpCpN context. We observed that ~7% of cytosine substitutions occurred in non-TpCpN context (39 out of 556 mutations; **Fig. 4c**), approximately 2-fold higher than expectation based on SBS2 and SBS13 signatures (7.00% vs. 3.33%; chi-square test, $p<0.005$). Our data indicated these non-TpCpN mutations arose simultaneously with the *kataegis* for two reasons: (1) Non-TpCpN mutations were 267-fold more abundant than observed in the background regions (**Fig. 4d**), making contamination by *kataegis*-unrelated mutations unlikely; and (2) Non-TpCpN mutations within *kataegis* regions were completely phased with other classical TpCpN mutations on the same chromosome (27 out of 27), where 50% was expected for random contaminating mutations. This suggests that DNA repair mechanisms for cytosine deamination in non-TpCpN contexts are less accurate during the clustered mutagenesis compared to isolated deamination events.

Of the 615 *omikli* and 106 *kataegis* events (excluding 3 *kataegis* events associated with rearrangements*)*, approximately 5% (35 *omikli* and 5 *kataegis* events) exhibited strand-switching of the mutated cytosines between parental and daughter strands during replication (**Fig. 4e,f**). In one *kataegis* event composed of 9 base substitutions (in clone A3A_1st_C3_3µg-5), we observed five strand-switching events (**Fig. 4g**). Phasing analysis suggested that clustered mutations across both DNA strands originated from single mutation events (26 out of all 26 informative events). Additionally, the mutational spectrum for the mutations in the minorly contributing strand was predominantly composed of mainly C>T or C>G mutations in the TpCpN context (62%; 13 out of 21 mutations; **Fig. 4c**), implying APOBEC-associated mutations. Besides, the rate of TpCpN mutations in the minor strand within strand-switching *kataegis* regions was 124-fold higher compared to outside of the strand-switching *kataegis* region (**Fig. 4h**). Although the underlying mechanism is unclear, strand-switching of the A3A enzyme should be possible when generating clustered mutations during replication.

Epigenetic contexts associated with A3A-associated mutations

Mutational processes are often influenced by the epigenetic contexts of the genome[42]. Of the 14 genomic features examined, three features (replication timing, local transcription level, and H3K27me3) showed potential associations with local A3A-associated mutational burdens (**Fig. 5a**). Of the three features, the replication timing and local gene transcription level demonstrated consistent trends correlating with A3A-associated mutation rates (**Supplementary Table 6**).

For the replication timing, the latest-replicating regions showed a 1.26-fold higher rate of A3A-associated mutation compared to the earliest-replicating regions (**Fig. 5b**). This may be attributed to the DNA repair mechanisms, including base excision repair, which are particularly active in early-replicating regions consisting of open chromatin[43,44]. Further, the A3A-associated mutation rate in the lagging strand of DNA replication was 1.26-fold higher compared to the leading strand (**Fig. 5c**), which presumably originated from more frequent exposure of single-stranded DNA (ssDNA) induced by Okazaki fragments in the lagging strand[45].

A similar pattern was observed in gene transcription. Mutation rates in genic regions were correlated with expression levels (**Fig. 5d**), showing a 1.37-fold higher mutation rate in actively transcribed genomic regions compared to silent genic regions. Of note, the lagging strand and highly transcribed genes tend to be more frequently single-stranded than the leading strand and silent genes[46,47], which may make them more susceptible to A3A-induced DNA damage, respectively. We believe that these two associations represent that DNA regions with frequent ssDNA exposure have a higher chance of being damaged by A3A. In line with the observation, the non-transcribed genic strand was mutated 1.13-fold more frequently than the transcribed strand (**Fig. 5e**).

Compared with non-coding sequences, protein-coding sequences showed much lower mutation rates, at 0.79-fold the genome average, suggesting a selective pressure against mutations that could alter amino-acid-changing mutations (**Fig. 5f**). Interestingly, in TP53-inactivated clones (TP53KO-hGO$_{iA3A}$ clones), mutation rates in protein-coding regions increased to 0.834-fold of the genome average, potentially due to reduced negative selection pressures in the absence of functional *TP53*.

Discussion

Our study clearly demonstrated the qualitative and quantitative mutational impact of A3A and A3B in human non-neoplastic cells. We did not observe a substantial number of A3B-associated mutations. Moreover, significant increases in A3B-associated mutations were absent even after *TP53* inactivation. Previous studies across various model systems have yielded inconclusive results regarding the mutagenic impact of A3B[10]. Our observations align with the previous report indicating inactivation of A3B gene in human cancer cell lines leads to a minimal reduction in APOBEC-associated mutations[17]. Meanwhile, one study reported that hundreds of A3B-associated mutations were comparable to the number of A3A-associated mutations in the HAP1 human cancer cell line[18]. In line with this, A3B demonstrated mutagenic potential in non-human models, such as in *Mus musculus*[20] and *Saccharomyces cerevisiae* where a DNA repair gene (*Ung1*, involved in the excision of deaminated cytosines) was inactivated[21]. These contradicting reports imply that there may be cellular conditions required for A3B mutagenic activity, which will be an intriguing avenue for future research.

When A3A was induced by 3 µg/ml of doxycycline in hGO_{iA3A} lines, transcription was activated for ~2-3 days, reaching peak gene expression levels of ~800 TPM. Under these conditions, we detected ~2,500 A3A-induced base substitutions in the proliferative clones, suggesting that ~1,000 base substitutions could be acquired in a day. If mutations in non-proliferative cells are included, the total impact could be several times greater. Our model supports the idea that a massive number of mutations can be acquired by a transcriptional burst of APOBEC3A enzyme[11].

Although A3B exhibited limited mutagenicity in our system, this could be due to the cell-type specificity. The clones we studied were derived from human stomach tissue. Our study warrants future studies with similar approaches but using organoids from other tissues, such as the small intestine, bladder, breast, both healthy and the neoplastic lineages.

Figure Legends

Fig. 1. Overview of conditional APOBEC (APOBEC3A or APOBEC3B) overexpression models with human gastric organoids.

(a) Experimental design of the study. Schematic illustration of genetic engineering, clonalization, and sequencing (whole-genome sequencing, duplex sequencing, and bulk RNA-seq) is shown. Dox, doxycycline.
(b) Changes of morphology and mCherry fluorescence during 48 hours with 0.1 µg/ml doxycycline treatment. (top) hGO_{iA3A} lines and (bottom) hGO_{iA3B} lines. Scale bars represent 1mm.
(c) Expression levels of APOBEC3A (A3A) or APOBEC3B (A3B) in each line following 0.1 µg/ml and 3 µg/ml doxycycline treatment for 48 hours. (left) hGO_{iA3A} lines and (right) hGO_{iA3B} lines. Data are presented as mean ± 95% confidence intervals. Statistical significance was determined using a two-tailed t-test: *$p < 0.05$, **$p < 0.005$, ***$p < 0.0005$, ****$p < 0.00005$.
(d) Changes of viability of each line following 0.1 µg/ml and 3 µg/ml doxycycline treatment for 48 hours. (left) hGO_{iA3A} lines and (right) hGO_{iA3B} lines. Data are presented as mean ± 95% confidence intervals.
(e) Differentially expressed genes upon APOBEC (A3A or A3B) overexpression following 3 µg/ml doxycycline treatment for 48 hours.
(f) Subcellular localization of overexpressed A3A and A3B following 3 µg/ml doxycycline treatment for 48 hours in each line. (top) hGO_{iA3A} line and (bottom) hGO_{iA3B} line. Scale bars represent 50µm.
(g) □-H2AX foci following 3 µg/ml doxycycline treatment for 48 hours in each line. (left) hGO_{iA3A} line and (right) hGO_{iA3B} line. Scale bars represent 50µm.

Fig. 2. Mutational impact of APOBEC-associated single base substitutions in DNA and RNA.

(a) The number of APOBEC-associated single nucleotide variants (SNVs) after overexpression of APOBEC (A3A or A3B) measured by whole-genome sequencing of the clones and duplex DNA sequencing. (left) A3A and (right) A3B.
(b) Mutational burdens and spectra of APOBEC-associated SNVs in each experimental condition. (left) A3A and (right) A3B.
(c-f) RNA editing after overexpression of A3A and A3B in bulk RNA-seq. Box-plots show the number of C>U RNA editing counts by **(c)** A3A, and **(e)** A3B. Spectra of RNA editing events in **(d)** hGO_{iA3A} lines, and **(f)** hGO_{iA3B} lines.

Fig. 3. Characteristics of A3A-associated mutational signatures.

(a) Context preference of A3A between YpTpCpA and RpTpCpA context in hGO_{iA3A} lines and cancer. Dashed black line, expected; orange line, hGO_{iA3A}; dark brown line, cancer
(b) Mutational spectra of the ID9 signature (a reference signature in COSMIC).
(c) Correlation between A3A-associated base substitutions and ID9 contributing indels.
(d) Associations between A3A-associated (SBS2 and SBS13) and age-associated (SBS5 and SBS40) base substitutions among hGO_{iA3A} lines and TP53KO-hGO_{iA3A} lines.
(e) Gene expression changes of *POLH* (translesion DNA polymerase) after A3A induction in hGO_{iA3A} and TP53KO-hGO_{iA3A} lines.

Fig. 4. Landscape of APOBEC3A-associated clustered mutation.

(a) Observed-to-expected ratio of the intermutational distances between substitutions in each clone. Data are presented as mean ± standard error.
(b) Frequency of clustered mutation events (*omikli* and *kataegis*) among A3A-associated base substitutions in clones and cancer genomes.
(c) Proportion of classic (in TpCpN context) and non-classical mutations (in non-TpCpN and NTN) in SBS2, SBS13, and *kataegis* regions in the clones.
(d) Comparison of expected and observed SNV counts in the non-TpCpN context within *kataegis* regions.
(e) Schematic representation of strand-switching *kataegis*.
(f) Frequency of strand-switching *omikli* and *kataegis.*
*(***g)** A complex strand-switching *kataegis* with five switches found in clone A3A_1st_C3_3µg-5.
(h) Enrichments of minor strand mutations in the TpCpN context in strand-switching *kataegis* events.

Fig. 5. Genomic and epigenomic distribution of APOBEC3A-associated mutations.

(a) Correlation between epigenetic markers and A3A-associated substitutions.
(b) Fold change of mutation rates of A3A-associated SNVs according to the replication timing of the genomic region. Data are presented as mean ± 95% confidence intervals.
(c) Mutation rates depending on the DNA strands during replication (leading and lagging strands). Statistical significance was determined using chi-square test: ****$p<0.00005$.
(d) Fold change of mutation rates of A3A-associated SNVs according to the local transcription levels in the genic region. Data are presented as mean ± 95% confidence intervals.

(e) Mutation rates depending on the DNA strands during transcription (transcribed and untranscribed strands). Statistical significance was determined using a chi-square test: ***$p<0.0005$.

(f) Mutation rates within sub-genic regions (5`UTR, intron, protein coding sequences (CDS), and 3`UTR) in (left) hGO_{iA3A} clones and (right) TP53KO-hGO_{iA3A} clones. Red dashed line, average genome-wide mutation rate.

Methods

Human normal gastric samples

Normal gastric tissues were obtained via endoscopic biopsy from a female undergoing routine screening. The protocol for this study was approved by the institutional review board of Yonsei University Gangnam Severance Hospital (3-2018-0207) and KAIST (KH2022-211).

Confirmation of non-neoplasticity of normal gastric organoids with whole-genome sequencing

Whole-genome sequencing analysis confirmed the absence of potent cancer driver mutations and copy number variations of DNA (**Extended Data Fig. 7**) in the normal gastric organoids.

Human stomach organoid culture

Organoid culture methods and media compositions were adopted from previous research with slight modification[48]. Wnt3A and R-spondin-1 conditioned media was produced with HEK293 cell line producing Afamin-Wnt3a[49] and Cultrex HA-R-Spondin 1-Fc 293T cell line (Trevigen). Complete medium composed of Wnt3A conditioned medium, R-spondin-1 conditioned medium, Advanced DMEM/F-12 (Gibco, Cat No.12634028)), HEPES (1M) (Gibco, Cat No.15630-080), Penicillin/streptomycin (10,000 U/mL) (Gibco, Cat No.15140122), GlutaMax (Gibco, Cat No.35050061), Human EGF Recombinant Protein (Invitrogen, Cat No.PHG0311), hNoggin (Peprotech, Cat No.120-10C), hFGF10 (Peprotech, Cat No.3100-26), B27 supplement (50x) serum free (Gibco, Cat No.17504044), N-acetylcysteine (Sigma-Aldrich, Cat No.A9165), Gastrin (Sigma-Aldrich, Cat No.G9145), Y27632 (Sigma-Aldrich, Cat No.Y0503), TGF-b R kinase inhibitor IV (Biogems, Cat No.3014193) (**Supplementary Table 7**).

Tissues were dissociated into 10~15 clustered cells by pipetting with 1000µl tips after 30 minutes of incubation in 4°C in TrypLE (Gibco, Cat No.12604013). Tissues were washed twice with PBS, and the pellet was isolated by removing the supernatant after centrifugation at 300g for 5 minutes at 4°C. The resuspended pellets with cold Matrigel (Corning, Cat No.BDL356231) were seeded on the 12- or 24-well plates (Merck) in dorm shape. After incubation of plates in the humidified incubator (5% CO2, 37℃) for 10 minutes to polymerize Matrigel, prewarmed 0.5-1 ml of culture medium was added to the well. Organoids were cultured in the humidified incubator. The medium was changed every two to three days.

Organoid passaging was conducted for about two weeks. Medium in the well was replaced by cell recovery solution (Corning, Cat No.354253), and the plates were incubated for 40~60 minutes at 4°C to dissolve Matrigel. Organoid pellet was isolated by removing supernatants after centrifugation at 300g for 5 min at 4°C. Organoids followed by incubation for 5 minutes at 37°C with resuspension with Accutase (Stemcell Technology, Cat No.07922) were dissociated into 10-15 cell clusters by pipetting 10-20 times. After diluting Accutase by adding 1 ml of ADF media (Gibco), the pellet was isolated by centrifugation at 300g for 5 minutes at 4°C. Organoids were seeded in a 12- or 24-well plate at a ratio of 1:4 to 1:6 within Matrigel.

Preparation of vectors for transfection

Two vector systems: (1) CMV-rtTA-HygR vector (Addgene, Cat No.102423) / and (2) CRISPR-*Cas9* vectors containing gRNA sequence for TP53 (Addgene, Cat No.121917) were purchased from Addgene. Basically, mini-prep and maxi-prep were conducted with commercial competent cells (Biosearch Technologies, Cat No.6010) and mini-prep kit (QIAGEN, Cat No.27104) and maxi-prep kit (QIAGEN, Cat No.12123) according to the manufacturers' protocols. To create pPB-CMVmin-A3A/A3B-IRES-mCherry vectors, APOBEC constructs were designed following these steps. NCBI reference sequences for *APOBEC3A* (NM_145699.4) and *APOBEC3B* (NM_004900.4) were utilized for the design. Kozak consensus sequences (GCCACC) were added at the 5' end of the A3A/A3B cDNA sequences, while HA tag sequences (5'-TAC-CCA-TAC-GAT-GTT-CCA-GAT-TAC-GCT-3') were appended to the 3' ends of the A3A/B cDNA sequences. Subsequently, *XhoI* and *NotI* restriction enzyme cut sites (CTCGAG and GCGGCCGC) were added to the 5' and 3' ends of the construct, respectively. *De novo* gene synthesis from GenScript (Piscataway, NJ) was used for the synthesis of the two constructs. Following the acquisition of the A3A and A3B constructs, the constructs were cloned into the backbone vector (pPB-CMVmin-TRE-IRES-mCherry[50]) to make pPB-CMVmin-A3A/B-IRES-mCherry cassette.

Transfection of organoids

Organoids were prepared with the same dissociation protocol of passaging organoids. Transfection methods were adopted from previously reported protocols[51,52]. A mixture of three types of vectors were utilized for transfection: having (1) TRE-APOBEC (A3A or A3B)-IRES-mCherry cassette and (2) CMV-rtTA-HygR cassette (3) pPiggyBac transposase cassette. For hGO_{iA3A} lines, organoids were suspended in 90µl of Opti-MEM (Gibco, Cat No.31985062), and mixed with about 30 µg of each vector. Electroporation was conducted using a previously described program from the literature[52] (**Supplementary Table 8**).

Transfected organoids were cultured for seven days with the medium composition described in the previously reported protocol. Organoids were incubated with the medium containing 1 µg/ml hygromycin (InvivoGen, Cat No.ant-hg-1) for seven days after splitting. To isolate the transfected organoids having insertions of two vector cassettes for conditional overexpression, organoids were incubated with doxycycline (Sigma-Aldrich., Cat No.D9891-1G) in a 3 µg/ml containing medium for 12-16 hours. mCherry-positive organoids were manually isolated by pipetting under a fluorescent microscope, and clonalization was conducted according to a previously reported method by using FACSAriaII(BD Biosciences) and manually picking single cell originated organoids by pipetting[53]. Isolated organoids were dissociated in the same way to passage. Then, organoids were then filtered through a 40 µm strainer (Falcon, Cat No.352340). Using FACSDiva software, pure single cells were isolated. After seeding the organoids, the single-cell-originated organoids were isolated. For hGO_{iA3B} lines, organoids were suspended in 100 µl of BTXpress buffer (BTX), and 10 µg of vector mixtures were mixed. The following steps were carried out the reported protocol[52]. Selection and clonalization steps were conducted using the same method as for hGO_{iA3A} lines.

Doxycycline treatment

A doxycycline stock solution was prepared by dissolving doxycycline in the DMSO. The stock solution was added to the medium to achieve final concentrations of 0.1 µg/ml and 3 µg/ml. Before treatment, dissociated 10k viable cells were seeded in 24-well the plates. After 7 days, doxycycline solution was added to the medium, and the organoids were incubated with doxycycline.

Viability assay of organoids

According to the manufacturer's protocol, the viability of organoids was measured using a Celltiter-Glo 3D Assay kit (Promega, Cat No.G9681). Each well was washed twice with 1 ml of PBS by repeatedly removing the medium and adding PBS. After washing, PBS was replaced with fresh medium. An equivalent volume of reaction solution was added to each well. Matrigel was dissolved in the reaction solution by incubating at 4℃ for 30 minutes with shaking. The amounts of ATP in each well were measured with a luminometer (BERTHOLD Technologies GmbH, Cat No.TriStar LB 942). The viability percentage was calculated by dividing luminescence after doxycycline treatment by the average luminescence of untreated control groups.

Capturing fluorescent images

Fluorescents of mCherry after doxycycline treatment were monitored under the fluorescent microscope (LEICA, DMi8; excitation = 541-551 nm; emission = 565-605 nm). Fluorescent images were captured using Las X programs provided by the manufacturer, and brightness/contrast were adjusted using the same Las X program.

Immunostaining and imaging of organoids

Whole-mount staining of human gastric organoids was performed as previously described[54]. Briefly, organoids were fixed with 4 % paraformaldehyde (PFA) after depolymerizing the 3D matrix using an ice-cold cell recovery solution (Corning). After washing the cells with 0.1% PBS-Tween-20 and blocking with organoid washing buffer (0.1 % Triton X-100, 0.02 % SDS, 0.2 % bovine serum albumin (BSA) in 1X PBS), immunolabeling was performed with mouse anti-HA tag antibody (Santa Cruz, Cat No.sc-7392; 1:50) and rabbit anti-phospho-histone H2A.X antibody (CST, Cat No.2577S ; 1:400) to detect inserted HA-tagged APOBEC proteins and double-strand breaks, respectively. Goat anti-mouse Alexa Fluor 488 (Thermo Fisher Scientific, Cat No.A-11001; 1:500) and donkey anti-rabbit Alexa Fluor 647 (Thermo Fisher Scientific, Cat No.A32795; 1:500) were used as secondary antibodies, and nuclei were counter-stained with DAPI (Sigma-aldrich., Cat No.D9542; 1:1000). After washing the samples, FUnGI clearing solution was added to the organoids. The samples with the clearing solution were directly mounted between two coverslips with a 0.25 mm-deep iSpacer (SunJin Lab, Cat No.IS213), and the imaging was performed at least 1 hour after slide preparation.

Imaging was performed with a Leica Stellaris 8 Confocal Microscope at the Research Solution Center (RSC) at the Institute for Basic Science (IBS).Alexa Fluor 488 (excitation, 479 nm; emission, 482–554 nm), Alexa Fluor 647 (excitation, 580nm; emission, 663-716 nm), and DAPI (excitation, 405 nm; emission, 425-489 nm) signals were obtained. Either 40X or 63X objectives with a digital zoom factor of 1 to 7-fold were used. The X/Y resolution was set to 1024 x 1024 pixels. Images were processed and analyzed using ImageJ.

Standard whole genome sequencing and alignment

Organoids were isolated by removing matrigel with a cell recovery solution (Corning), similar to the method used for passaging organoids. According to the manufacturer’s protocol, DNA from clonalized organoids was extracted using DNeasy Blood & Tissue Kit (QIAGEN, Cat No.69506). Libraries were constructed with Truseq DNA PCR-Free Library Prep Kits (Illumina, Cat No.20015963) according to the manufacturer’s protocols. Whole genome

sequencing was conducted by the NovaSeq 6000. Organoid clones were whole-genome sequenced with a mean 30-fold depth of coverage. Adapter sequence in the FASTA files was removed by Cutadapt software (https://cutadapt.readthedocs.io/en/stable/)[55]. Sequencing reads were aligned to the human reference genome, GRCh37, with bwa-mem v0.7.17 (https://github.com/lh3/bwa)[56]. Further processing, including sorting, marking duplication, and indel realignment, was conducted with samtools v1.9 (https://www.htslib.org/)[57] , PICARD v2.1.0 (https://broadinstitute.github.io/picard/)[58] , and GATK tools v3.8.0 (https://gatk.broadinstitute.org/hc/en-us/categories/360002369672-Tool-Index)[58].

BotSeqS library construction

DNA for the BotSeqS library was extracted with All-Prep DNA/RNA Mini kit (Qiagen, Cat No.80204) according to the manufacturer's protocol. Constructed libraries with Truseq DNA PCR-Free Library Prep Kits (Illumina) were utilized for subsequent steps. The construction of BotSeqS libraries was based on the previous literature with slight modifications[33]. Briefly, the quantification of DNA libraries was conducted with KAPA Library Quantification Kit ILLUMINA® Platforms (Roche, Cat No.KK48247) according to the manufacturer's protocol. An effective library concentration was calculated considering the size of added DNA length during library construction. PCR amplification of libraries equivalent to 4 pg of DNA was performed with the KAPA HiFi Uracil+ kit (Roche, Cat No.KK2802) and primers having a Y-adaptor sequence with a phosphorothioate bond(*) at the 3' end from IDT (Coralville, IA):

Forward: 5'-AATGATACGGCGACCACCGAG*A-3'

Reverse: 5'-CAAGCAGAAGACGGCATACGA*G-3'

PCR programs were adopted from the protocol of KAPA Library Amplification Kits (Roche), and PCR was conducted with 20 cycles. Library clean-up was performed with SPRIselect beads (BECKMAN COULTER life Sciences) according to the manufacturer's protocol. The libraries were sequenced as paired-end sequencing (2x151 bps) using NovaSeq 6000.

***In vitro* deamination of DNA with synthesized APOBEC3B**

The NEBNext Enzymatic Methyl-seq (EM-seq) Kit (NEB, Cat No.E7120S) and the manufacturer's protocol were utilized for the construction of the DNA library with slight modification. Firstly, the DNA library was constructed with DNA extracted from a doxycycline treatment-free hGO_{iA3B} line. DNA libraries equivalent to 10 pg of input DNA were prepared following the process used for BotSeqS library construction. Oxidation of DNA by TET2 and

denaturation steps were performed according to the manufacturer's instructions. For deamination steps, synthesized APOBEC3B from EUROPROTEIN INC. (North Brunswick Township, NJ) was used instead of the APOBEC protein included in the EM-seq kit. Additionally, the reaction time was modified from 3 hours to 30 and 60 seconds. Library amplification steps followed the BotSeqS library construction.

Whole transcriptome sequencing library construction

RNA was isolated during the extraction of DNA for BotSeqS libraries with AllPrep DNA/RNA Mini kit according to the manufacturer's instructions. RIN value of raw RNA was measured with Agilent 4200 TapeStation (Agilent, Cat No.G2991BA). Libraries were constructed with NEBNext Ultra II Directional DNA Library Prep Kit for Illumina (NEB, Cat No.E7760) and QiAseq FastSelect -rRNA HMR kit (Qiagen, Cat No.334388) according to manufacturers' instructions. Clean-up process was conducted with SPRIselect beads according to the manufacturer's protocol. Libraries were sequenced with Novaseq 6000 with paired-end sequencing.

Confirming non-neoplasticity in primary human gastric organoid

The germline mutation calling mode in GATK v4.0.12.0 (https://gatk.broadinstitute.org/hc/en-us/categories/360002369672-Tool-Index)[58] was utilized for the identification of mutation in the organoid lines. The primary call set was first filtered by in-house scripts using the pysam module in Python (https://pysam.readthedocs.io/en/latest/index.html)[57] with the following conditions: (1) the number of reference reads ≥ 5; (2) the number of variant reads ≥ 3; (3) median miss-matched bases in variant reads < 5; (4) median mismatched bases in variant reads < 5 (5) median mapping quality of reference reads > 20 (6) median mapping quality of variant reads > 20 (7) median distance from the ends of reads > 4. The impact of variants in the protein-coding region was annotated with ANNOVAR (https://annovar.openbioinformatics.org/en/latest/)[59]. Non-silent mutations located in the protein-coding and splicing regions were isolated. The functional impact in cancer driver genes (the list can be downloaded at https://cancer.sanger.ac.uk/cosmic) was evaluated with PolyPhen-2[60] and SIFT[61] scores annotated by ANNOVAR, AlphaMissence (https://alphamissense.hegelab.org/)[62] , and MutPred-Indel

(http://mutpred2.mutdb.org/mutpredindel/about.html)[63]. Considering the role of cancer driver genes, the impact of variants was finally assessed.

Calling copy number variations

Copy number variations were accessed by Sequenza (https://sequenza-utils.readthedocs.io/en/latest/#)[64]. Copy number variations in segments smaller than 1 Mbp were considered false positives. After removing the copy number variations in short segments, final copy number variations were investigated by a second run of the Sequenza.

Detection of somatic mutations

To detect single nucleotide variants and indels in clonal organoids, GATK Mutect2 v4.1.9 (https://gatk.broadinstitute.org/hc/en-us/categories/360002369672-Tool-Index)[58] and Strelka2 v2.9.2 (https://github.com/Illumina/strelka)[65] were utilized. In each call set, false-positive calls were filtered out with in-house python scripts annotating information within BAM files with the pysam, a python module. The filtered call set was merged, and the union call set was utilized for analysis. To exclude mutations generated during the culture of mother clones, commonly discovered mutations were removed from each call set.

In the case of BotSeqS, VarScan2 v2.3.9 (https://varscan.sourceforge.net/)[66] was used to increase the sensitivity of calls. Similarly, in-house Python scripts were utilized to filter out false-positive calls. Specifically, F1R2 and F2R1 reads from same DNA was grouped and filtered with the following criteria: (1) total depth of each type of read ≥ 3; (2) the number of variant reads ≥ 3 or 90% of reads; (3) distances of mutations from each end of each read > 5 bp; (4) distances of mutations from the end of reads > 100 bp considering the DNA strand (e.g., excluding C>T near the end of F1R2 reads, but not G>A); (5) median mapping quality of variant reads ≥ 20; (6) median base quality of variant reads ≥ 30; (7) the number of variants in WGS of HEK293T <3; (8) the number of variants in WGS of control < 1. Unlike with standard whole genome sequencing, C>T mutations were frequently observed at the end of the reads, and these were considered false positives and removed. Additionally, rare

variants from HEK293T cells, which were used to generate conditioned medium, were observed. Thus, variants observed in HEK293T cell line BAM files at least three times were removed from the call set.

Calculation of mutation rate in BotSeqS

Unlike standard WGS, the length of the effective covered region was calculated for each BAM file. Firstly, each read was evaluated with the following criteria, considering DNA strands, using in-house scripts to isolate effective DNA fragments: (1) median mapping quality > 20 (2) total depth of each type of reads ≥ 3. Among covered regions, only those where both F1R2 and F2R1 were aligned together were counted. Since mutations close to the end of reads (within 100 bp from the ends of reads) were dropped during the variant filtering process, (151*2-110)/(151*2)) were multiplied by the total length of the covered region. Additionally, considering ploidy, the length of the covered region was multiplied by two to compare mutation rates with standard WGS results.

Detection of deamination during *in vitro* denomination with APOBEC3B

The calling pipeline for BotSeqS results was utilized with slight modifications. Among the criterias, the filtering criteria for distances from the ends of reads were neglected.Initially, all of the filtered mutations were collected. Among the mutations in grouped DNA, only those in DNA fragments where only one strand of DNA was mapped were counted.

To calculate the genome-wide mutation rates, the BotSeqS pipeline was also used, with the exception of counting cytosine or guanine considering the strand of DNA. Since original methylated CpG sites were preserved during the library construction, total counts and mutation counts in all of CpG contexts were excluded. To calculate mutation rates in DNA fragments with at least one C>T variant, ranges of DNA fragments overlapping with the variants were isolated while considering the DNA strands using BEDtools (https://bedtools.readthedocs.io/en/latest/)[67]. After this step, cytosine or guanine outside of CpG contexts were counted.

Analysis of mutational signatures

Mutational signature analysis of single base substitutions (SBS) and small indels were carried out using non-negative least squares method. The mutational signature was represented by 96 patterns of SBS and 83 patterns of small indels[1]. Pre-learned catalogs of mutational signatures in the COSMIC[68] were used to fit individual samples with a known set of signatures for each tissue type. SBS2 and SBS13 (known APOBEC signatures) were included for all cases.

Analysis of mutation rates depending on epigenetic marker

Risk ratio of mutation rates in non-signal regions and signal-detected regions depending on each epigenetic marker was analyzed based on previous researches[69,70]. Genome-wide signals of each marker, including replication timing, across eight cell types were downloaded from Roadmap Epigenomics Consortium. Fold-enrichment signal was averaged across the cell types, and the regions showing fold-enrichment signal lower than 1 identified as bin 0, non-signal regions. The other regions were classified as the signal-detected regions.

To analyze fold-enrichment of mutation rates, in case of replication timing and H3K27me3 signals, the signal-detected regions were further divided into four groups, considering the equal distribution of length of each group. In case of expression level, expression level with TPM unit in each base calculated from merged whole-transcriptome sequencing data of doxycycline treated organoids for 48 hours under 3 µg/ml doxycycline was utilized. First of all, non-genic regions and genic regions were classified based on gene information from the “hg19_refGene.txt” file. Only genic regions were included in the analysis. Genic regions having 0 of TPM were defined as bin0 regions. TPM>0 regions were classified into 4 groups (cutoff; bin1=0.05, bin2=1.73, and bin3&bin4=9.68) based on quartiles of TPM. Considering interference among H3K27me3, replication timing, and transcription, glm.nb() function in MASS, R package (https://cran.r-project.org/web/packages/MASS/index.html)[71], was utilized in enrichment analysis as described in previous research[70].

Analysis of mutation rates depending on genomic location

SNVs located in previously described mappable regions were utilized throughout the analysis. Gene information included in “hg19_refGene.txt” file from ANNOVAR tools were utilized to match the additional information of position of SNVs. Classification of sub-genic (5`/3`UTR, protein coding sequence (CDS), and intron) and transcription strand were also based on the gene information. All of the merged genic regions were used as standard for discrimination of genic and intergenic regions. Instead of that, genic regions that don't

overlap with other genic regions and CDS regions located between intron regions were utilized for comparing mutation rates in the sub-genic region as previous research reported[72].

Calling structural variations

Structural variations were called using Delly v0.7.6 (https://github.com/dellytools/delly)[73]. Raw calls were filtered using in-house scripts. The final call set was manually reviewed by Integrative Genomics Viewer (https://igv.org/)[74].

Calculating RNA expression level

Bulk RNA-seq reads were aligned to GRCh37 using STAR2 v2.6.1d (https://github.com/hbctraining/Intro-to-rnaseq-hpc-O2)[75]. TPM and reads count were calculated with RSEM v1.3.1 (https://github.com/deweylab/RSEM)[76]. Differential expression gene analysis was conducted using the DEseq2 package in R (https://bioconductor.org/packages/release/bioc/html/DESeq2.html)[77].

Calling RNA editings

VarScan2 was utilized to call RNA editing. WGS of mother clones was utilized as paired normal. False positives that did not meet the following criteria were filtered out using an in-house Python script with the pysam module: (1) excluding HLA regions; (2) DNA depth ≥ 5; (3) number of variant reads in DNA = 0; (4) RNA depth ≥ 4; (5) number of variant reads ≥ 3; (6) median number of NM tags in the variant reads < 4; (7) percentage of variant reads with NM tags greater than 3 < 0.5; (8) distance from the end of reads ≥ 5; (9) variants in only one strand; (10) median base quality of variants > 20; (11) percentage of variant reads with clips < 0.5; (12) median mapping quality of variant reads ≥ 30; (13) single-base variant. RNA editings that did not meet all criteria but were commonly observed with all conditions fulfilled in more than two samples were rescued.

To compare the RNA editings, the number of RNA editings was normalized with total sequencing output size. The target output size was determined by the sample having the lowest sequencing output. RNA editings were filtered again by adjusting the ratio of

sequencing output to the total depth and variant reads. Variants that met the following two criteria were used in the analysis: (1) corrected depth of RNA ≥ 4 (2) corrected the number of variants.

Analysis of RNA editing signatures

RNA editing signatures were obtained by a modified version of the mutational signature extraction method used in[78,79]. Briefly, this method uses non-negative matrix factorization (NMF) to disentangle an individual RNA editing spectrum based on a notion of mixed spectra[80].

RNA editing spectrum was represented as 192 patterns of single-base change with immediately adjacent bases on the 5' and 3'-end considered. The number of features is doubled of DNA mutational signature[79], because RNA does not necessarily have symmetric base change due to general non-strandedness.

Total 18 experimental data were used as input, split into two subsets: A3A and A3B sets. Each of the A3A and A3B sets included three batches from experiments with 0µg/ml, 0.1 µg/ml, and 3 µg/ml APOBEC3 exposure, totaling 9 batches for each set. Initially, the signature extraction method was used with no additional procedures from the previous method[78], and then the optimal number of independent RNA editing signatures was determined. We observed that there were 4 potential candidates for RNA editing signature that are attributable to APOBEC3A, APOBEC3B, ADAR, and the other endogenous editing processes.

APOBEC3A and APOBEC3B aligned well with editing patterns of C>T found in experiments with varying exposure amounts. However, the ADAR signature co-appeared in these signatures at significant proportion. This "leakage" phenomenon occurs due to the constant activity of the ADAR editing process, causing its signal to always appear in RNA editing spectra regardless of the experimental settings. This can be problematic when trying to accurately quantify the attributes of each RNA editing process in a given sample.

To mitigate the potential effect of the leakage phenomenon, an L1 constraint was imposed on the signature matrix (W)[78]. For that, the NMF update algorithm was adopted from a work on sparse NMF[81]. It presents a method that puts L1 constraint on the exposure matrix (H). Therefore, we tweaked its objective function by transposing the data matrix (V) as well as W and H and back-transposing them to get W. Afterwards, the matrix W was normalized so that

each column vector sums to one. The sparsity coefficient (μ) was set to 0.6, after trying out several values in a range [0.001, 10].

The A3A set resulted in ADAR and APOBEC3A that were now isolated from each other. The A3B set led to ADAR and APOBEC3B signatures that were also well-separated. ADAR signatures learned from the A3A and A3B set were consistent with cosine similarity 0.999 and were merged into a single ADAR signature by taking average. Endogenously occurring RNA editing other than ADAR was negligible in amount and did not add much value to overall analysis. Thus, we did not seek further signatures.

Lastly, the three RNA editing signatures learned from the data were refitted to individual samples to evaluate the attributes of the RNA editing processes in each sample.

Calling clustered mutation

SigProfilerClusters (https://github.com/AlexandrovLab/SigProfilerClusters)[82], a python module, was utilized to identify clustered mutations. FlexMix , R package (https://cran.r-project.org/web/packages/flexmix/index.html)[83], was utilized to classify the identified clustered mutations into *omikli* and *kataegis*. Since the tools determine the standard intermutation distance for clustering through simulations that randomly distribute SNVs, the total number of SNVs influence the detection rate of clustered mutation events. To correct for the number of clustered mutation events, simulations were conducted.

Identified clustered mutations from our analysis and 703,858 SNVs from 125 samples in PCAWG database, where the contribution of age-related mutational signatures (SBS1, SBS5, and SBS40) is over 0.8, were utilized for simulation. The target number of SNVs was set to 33 (increased by 2,000 from 2,000 to 10,000; increased by 5,000 from 15,000 to 100,000; increased by 10,000 from 110,000 to 200,000). From the observed 615 *omikli* (1,369 SNVs) and 109 *kataegis* (559 SNVs) events, the clustered mutation events were randomly sampled based on the following conditions for 50 iterations: 50, 100, 250, 500 *omikli*, and 25, 50, 75, 100 *kataegis*. Background mutations were selected to match the number remaining after excluding the number of clustered mutations from the target number of SNVs. With the same SigProfilerClusters module, clustered mutations, *omikli and kataegis,* were identified, and the detection rate was calculated. By combining data from the two types of clustered mutations, the relationship between the number of mutations and the detection rate of clustered mutations was estimated using the “drm” function with the fct=AR.3() option in the drc, R package (https://cran.r-project.org/web/packages/drc/index.html)[84] (**Extended Data Fig. 8a,b**).

To compare clustered mutation events, 146 samples from the PCAWG database were selected based on the following criteria: (1) nine cancer types showing high prevalence of APOBEC mutational signatures; (2) samples with fewer than 500 SBS2+SBS13 were excluded in both our samples and PCAWG cancer samples to avoid bias.

Analysis of 4bp context mutations

SNVs found in 10 hGO_{iA3A} samples (excluding control samples) were utilized in the analysis. In case of PCAWG samples, among 567 samples in nine cancer types showing high prevalence of APOBEC mutational signatures, only those with APOBEC-associated mutational signatures (SBS2+SBS13) over 5,000 were included in the analysis.

Analysis of public scRNA data of breast cancer cells

Following tutorial code in the Seurat, R package (https://satijalab.org/seurat/)[85] except for the cell filtering step, expression levels of APOBEC3A and APOBEC3B in breast cancer cells were identified. To investigate cancer cells showing exceptionally high expression levels, all of the cells that meet following criteria were included in the analysis: (1) Identified as "epithelial" in the previous publication[86], and (2) identified as "Cancer cells" with ScType tools (https://github.com/IanevskiAleksandr/sc-type)[87].

Publicly available datasets

Publicly available whole-genome sequencing data were utilized to demonstrate copy number variation in the normal human gastric organoid. WGS of blood from HC05 sample was utilized as the unpaired normal sample (https://ega-archive.org/studies/EGAS00001006213)[69]. To compare the 4bp-context preference and frequency of clustered mutation, SNV calls from the ICGC/TCGA Pan-Cancer Analysis of Whole-Genome (PCAWG) Consortium were utilized for the analysis. The call set data is available for download at https://dcc.icgc.org/releases/PCAWG. To compare levels of transcriptional bursts of A3A and A3B *in vivo*, single-cell RNA-seq data of breast cancer generated with Smart-seq platform was utilized (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE118390)[86].

Quantification and statistical analysis

All statistical analyses were performed with R version 4.1.3 (R Core Team, Vienna, Austria). A two-tailed one-sample t-test was used to evaluate p-values for comparing APOBEC-

associated SNVs and expression levels between the groups. Linear regressions were conducted using the basic 'lm' function in R to analyze the association among APOBEC-associated SNVs, ID9, SBS5, and SBS40. A chi-square test was utilized to evaluate p-values for comparing replication strand bias and transcription strand bias. A 95% confidence interval was used to determine the statistical range of continuous data.

Data availability

Whole-genome sequencing, EM-seq, BotSeqS, and bulk RNA-seq raw data produced in this study were deposited in the Korean Nucleotide Archive (KoNA; https://www.kobic.re.kr/kona/). The project accession ID is KAP240815 (https://kbds.re.kr/reviewer/0f8ec3f3-c994-4b38-8e5f-979568069628-KAD2462524; https://kbds.re.kr/reviewer/192fda81-ca6d-47c6-a546-a2b560e221c1-KAP240815). All data are provided for review purposes upon request by reviewers.

Code availability

Essential in-house scripts used in this study are available on Zenodo (DOI: 10.5281/zenodo.12771074; https://zenodo.org/records/12771074?preview=1&token=eyJhbGciOiJIUzUxMiJ9.eyJpZCI6IjdiNDQ2YzU4LTIxNDMtNDViOS1hYTQ5LTE1MTY4OTEzMmQ2NyIsImRhdGEiOnt9LCJyYW5kb20iOiIwZTg1YmFiYjAzNDRmMzgwOTQ1ZGI2MGI0NDYxMmZkZiJ9.pxeRUXdTXwcoJrUMkx2ZGlMZNnOyNai2-mcrqyiOIbJkB10lgSOhcvc5Z6mrNKc4TSicAm7NrSgqNops57_SWg).

Materials Availability

Organoids established in this study will be available under a material transfer agreement. To do so, please contact the lead author (ysju@kaist.ac.kr).

Acknowledgements

The authors thank Youngwon Cho (Epithelial Biology Center, Vanderbilt University Medical Center) for helping with the cell viability test, Myungsuk Choi (KAIST) for technical help for organoids, and Mia Petljak (New York University) for her valuable advice and feedback on the manuscript. This work was supported by the Young Investigator Grants from the Human Frontier Science Program (RGY0071/2018 to Y.S.J., B.-K.K., and H.S.); the National Research Foundation of Korea (Leading Researcher Program NRF-2020R1A3B2078973 to

Y.S.J.); the Korea Bio Data Station (N24NM016-24 to J.L and J.W.P); and the Institute for Basic Science (IBS-R021-D1 to B.-K.K).

Author contributions

Y.S.J., B.-K.K., and H.S. designed the experiments. J.-H.K. provided human gastric epithelium sample from biopsy. Y.J.B. established organoids from primary human gastric tissues. J.L. and J.W.P. helped basic bioinformatic works (alignment and calling somatic variants). J-H.L. and J.Y. conducted vector cloning. Y.A. and T.K. conducted the transformation of organoids, with J-H.L. and J.Y. providing advice. Y.A. performed organoid culture, clonal expansion, and DNA/RNA extraction. B.-K.K. and H.K. provided training in organoids culture technologies. Y.A. performed cell viability analysis. S.A. conducted *in vitro* deamination experiments and most of the library construction including WGS and RNA-seq. Y.A., J.-H.P., and K.Y. conducted library construction for duplex sequencing analysis. Y.A. conducted most bioinformatic analyses, including alignment, mutation calling, duplex sequencing analysis, and analysis of expression levels with bulk RNA-seq data, with S.P., J.Y., and Y.S.J. providing advice. Y.A. and Jo.L. conducted quantitative and statistical data analysis. Y.A. and Jo.L. conducted mutational signature analysis including RNA editing. W.H.L. helped with the construction of pipeline for analysis of duplex sequencing. Y.A. and C.H.N. conducted epigenomic analysis of variants. Y.A. and Y.S.J. wrote the manuscript with contributions from all the authors. Y.S.J. supervised the overall study.

Conflict of interest

Y.S.J. is a genomic co-founder of Inocras Inc..

References

1. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
2. Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75–82 (2020).
3. Wang, Y. *et al.* APOBEC mutagenesis is a common process in normal human small intestine. *Nat. Genet.* **55**, 246–254 (2023).
4. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
5. Vieira, V. C. & Soares, M. A. The role of cytidine deaminases on innate immune responses against human viral infections. *Biomed Res. Int.* **2013**, 683095 (2013).
6. Sondka, Z. *et al.* COSMIC: a curated database of somatic variants and clinical data for

cancer. *Nucleic Acids Res.* **52**, D1210–D1217 (2024).
7. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
8. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
9. Roberts, S. A. *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* **45**, 970–976 (2013).
10. Dananberg, A., Striepen, J., Rozowsky, J. S. & Petljak, M. APOBEC Mutagenesis in Cancer Development and Susceptibility. *Cancers* **16**, (2024).
11. Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282–1294.e20 (2019).
12. Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N. Engl. J. Med.* **376**, 2109–2121 (2017).
13. Lee, J.-K. *et al.* Clonal History and Genetic Predictors of Transformation Into Small-Cell Carcinomas From Lung Adenocarcinomas. *J. Clin. Oncol.* **35**, 3065–3074 (2017).
14. Jakobsdottir, G. M., Brewer, D. S., Cooper, C., Green, C. & Wedge, D. C. APOBEC3 mutational signatures are associated with extensive and diverse genomic instability across multiple tumour types. *BMC Biol.* **20**, 117 (2022).
15. Chang, J. *et al.* Genomic alterations driving precancerous to cancerous lesions in esophageal cancer development. *Cancer Cell* **41**, 2038–2050.e5 (2023).
16. Moody, S. *et al.* Mutational signatures in esophageal squamous cell carcinoma from eight countries with varying incidence. *Nat. Genet.* **53**, 1553–1563 (2021).
17. Petljak, M. *et al.* Mechanisms of APOBEC3 mutagenesis in human cancer cells. *Nature* **607**, 799–807 (2022).
18. Carpenter, M. A. *et al.* Mutational impact of APOBEC3A and APOBEC3B in a human cell line and comparisons to breast cancer. *PLoS Genet.* **19**, e1011043 (2023).
19. Law, E. K. *et al.* APOBEC3A catalyzes mutation and drives carcinogenesis in vivo. *J. Exp. Med.* **217**, (2020).
20. Durfee, C. *et al.* Human APOBEC3B promotes tumor development in vivo including signature mutations and metastases. *Cell Rep Med* **4**, 101211 (2023).
21. Chan, K. *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.* **47**, 1067–1072 (2015).
22. Jager, M. *et al.* Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. *Genome Res.* **29**, 1067–1077 (2019).
23. Youk, J. *et al.* Mutational impact and signature of ionizing radiation. *bioRxiv* 2021.01.12.426324 (2021) doi:10.1101/2021.01.12.426324.
24. Pleguezuelos-Manzano, C. *et al.* Mutational signature in colorectal cancer caused by

genotoxic pks E. coli. *Nature* **580**, 269–273 (2020).

25. Woodard, L. E. & Wilson, M. H. piggyBac-ing models and new therapeutic strategies. *Trends Biotechnol.* **33**, 525–533 (2015).
26. Wilson, M. H., Coates, C. J. & George, A. L., Jr. PiggyBac transposon-mediated gene transfer in human cells. *Mol. Ther.* **15**, 139–145 (2007).
27. Cazzalini, O., Scovassi, A. I., Savio, M., Stivala, L. A. & Prosperi, E. Multiple roles of the cell cycle inhibitor p21(CDKN1A) in the DNA damage response. *Mutat. Res.* **704**, 12–20 (2010).
28. Amador, V., Ge, S., Santamaría, P. G., Guardavaccaro, D. & Pagano, M. APC/C(Cdc20) controls the ubiquitin-mediated degradation of p21 in prometaphase. *Mol. Cell* **27**, 462–473 (2007).
29. Matsushima, K., Yang, D. & Oppenheim, J. J. Interleukin-8: An evolving chemokine. *Cytokine* **153**, 155828 (2022).
30. Wu, Y.-Y. *et al.* Upregulation of CCL20 and recruitment of CCR6+ gastric infiltrating lymphocytes in Helicobacter pylori gastritis. *Infect. Immun.* **75**, 4357–4363 (2007).
31. Burns, M. B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494**, 366–370 (2013).
32. Green, A. M. *et al.* APOBEC3A damages the cellular genome during DNA replication. *Cell Cycle* **15**, 998–1008 (2016).
33. Hoang, M. L. *et al.* Genome-wide quantification of rare somatic mutations in normal human tissues using massively parallel sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 9846–9851 (2016).
34. DeWeerd, R. A. *et al.* Prospectively defined patterns of APOBEC3A mutagenesis are prevalent in human cancers. *Cell Rep.* **38**, 110555 (2022).
35. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
36. Choi, J.-H. & Pfeifer, G. P. The role of DNA polymerase eta in UV mutational spectra. *DNA Repair* **4**, 211–220 (2005).
37. Delbos, F. *et al.* Contribution of DNA polymerase eta to immunoglobulin gene hypermutation in the mouse. *J. Exp. Med.* **201**, 1191–1196 (2005).
38. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
39. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
40. Bergstrom, E. N. *et al.* Mapping clustered mutations in cancer reveals APOBEC3 mutagenesis of ecDNA. *Nature* **602**, 510–517 (2022).
41. Maciejowski, J. *et al.* APOBEC3-dependent kataegis and TREX1-driven chromothripsis during telomere crisis. *Nat. Genet.* **52**, 884–890 (2020).

42. Otlu, B. *et al.* Topography of mutational signatures in human cancer. *Cell Rep.* **42**, 112930 (2023).
43. Amouroux, R., Campalans, A., Epe, B. & Radicella, J. P. Oxidative stress triggers the preferential assembly of base excision repair complexes on open chromatin regions. *Nucleic Acids Res.* **38**, 2878–2890 (2010).
44. Rhind, N. & Gilbert, D. M. DNA replication timing. *Cold Spring Harb. Perspect. Biol.* **5**, a010132 (2013).
45. Wu, T., Lyu, R., You, Q. & He, C. Kethoxal-assisted single-stranded DNA sequencing captures global transcription dynamics and enhancer activity in situ. *Nat. Methods* **17**, 515–523 (2020).
46. Okazaki, R., Okazaki, T., Sakabe, K., Sugimoto, K. & Sugino, A. Mechanism of DNA chain growth. I. Possible discontinuity and unusual secondary structure of newly synthesized chains. *Proc. Natl. Acad. Sci. U. S. A.* **59**, 598–605 (1968).
47. Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A. & Kornberg, R. D. Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. *Science* **292**, 1876–1882 (2001).
48. Bartfeld, S. *et al.* In vitro expansion of human gastric epithelial stem cells and their responses to bacterial infection. *Gastroenterology* **148**, 126–136.e6 (2015).
49. Mihara, E. *et al.* Active and water-soluble form of lipidated Wnt protein is maintained by a serum glycoprotein afamin/α-albumin. (2016) doi:10.7554/eLife.11621.
50. Lee, J.-H. *et al.* p57 imposes the reserve stem cell state of gastric chief cells. *Cell Stem Cell* **29**, 826–839.e9 (2022).
51. Gaebler, A.-M. *et al.* Universal and Efficient Electroporation Protocol for Genetic Engineering of Gastrointestinal Organoids. *J. Vis. Exp.* (2020) doi:10.3791/60704.
52. Fujii, M., Matano, M., Nanki, K. & Sato, T. Efficient genetic engineering of human intestinal organoids using electroporation. *Nat. Protoc.* **10**, 1474–1485 (2015).
53. Youk, J., Kwon, H. W., Kim, R. & Ju, Y. S. Dissecting single-cell genomes through the clonal organoid technique. *Exp. Mol. Med.* **53**, 1503–1511 (2021).
54. van Ineveld, R. L., Ariese, H. C. R., Wehrens, E. J., Dekkers, J. F. & Rios, A. C. Single-Cell Resolution Three-Dimensional Imaging of Intact Organoids. *J. Vis. Exp.* (2020) doi:10.3791/60709.
55. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
56. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
57. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

58. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
59. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
60. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* **Chapter 7**, Unit7.20 (2013).
61. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
62. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
63. Pagel, K. A. *et al.* Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. *PLoS Comput. Biol.* **15**, e1007112 (2019).
64. Favero, F. *et al.* Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* **26**, 64–70 (2015).
65. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
66. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
67. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
68. Sondka, Z. *et al.* COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res.* **52**, D1210–D1217 (2024).
69. Nam, C. H. *et al.* Widespread somatic L1 retrotransposition in normal colorectal epithelium. *Nature* **617**, 540–547 (2023).
70. Supek, F. & Lehner, B. Clustered Mutation Signatures Reveal that Error-Prone DNA Repair Targets Mutations to Active Genes. *Cell* **170**, 534–547.e23 (2017).
71. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer Science & Business Media, 2003).
72. Frigola, J. *et al.* Reduced mutation rate in exons due to differential mismatch repair. *Nat. Genet.* **49**, 1684–1692 (2017).
73. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
74. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
75. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

76. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
77. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
78. Youk, J. *et al.* Quantitative and qualitative mutational impact of ionizing radiation on normal cells. *Cell Genom* **4**, 100499 (2024).
79. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep.* **3**, 246–259 (2013).
80. Cichocki, A., Zdunek, R. & Amari, S. New algorithms for non-negative matrix factorization in applications to blind source separation. in *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings* (IEEE, 2006). doi:10.1109/icassp.2006.1661352.
81. Le Roux, J., Weninger, F.J. and Hershey, J.R., 2015. Sparse NMF–half-baked or well done?. Mitsubishi Electric Research Labs (MERL), Cambridge, MA, USA, Tech. Rep., no. TR2015-023, 11, pp.13-15
82. Bergstrom, E. N., Kundu, M., Tbeileh, N. & Alexandrov, L. B. Examining clustered somatic mutations with SigProfilerClusters. *bioRxiv* 2022.02.11.480117 (2022) doi:10.1101/2022.02.11.480117.
83. Leisch, F. FlexMix: A general framework for finite mixture models and latent class regression inR. *J. Stat. Softw.* **11**, (2004).
84. Ritz, C., Jensen, S. M., Gerhard, D. & Streibig, J. C. *Dose-Response Analysis Using R*. (CRC Press, 2019).
85. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
86. Karaayvaz, M. *et al.* Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.* **9**, 3588 (2018).
87. Ianevski, A., Giri, A. K. & Aittokallio, T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat. Commun.* **13**, 1246 (2022).

Extended Data Figures Legends

Extended Data Fig. 1 Distribution of cell fraction of mutations in each clone in this study.

Extended Data Fig. 2 Cumulative probabilities of expression levels of A3A and A3B in single breast cancer cells. Green line; average expression levels of A3A and A3B following 0.1 µg/ml doxycycline for 48 hours in each corresponding line, hGO_{iA3A} and hGO_{iA3B}, respectively; dark green line; average expression levels of A3A and A3B following 3 µg/ml doxycycline for 48 hours in each corresponding line, hGO_{iA3A} and hGO_{iA3B}, respectively.

Extended Data Fig. 3 Comparison of the number of APOBEC-associated single nucleotide variants (SNVs) in the clone sequencing.

(a) The number of A3A-associated SNVs (SBS2+SBS13) in hGO_{iA3A} and TP53KO-hGO_{iA3A} clones under each condition. Statistical significance was determined using a two-tailed t-test: *$p < 0.05$, **$p < 0.005$, ***$p < 0.0005$,****$p < 0.00005$.
(b) The number of A3B-associated SNVs (SBS2+SBS13) in hGO_{iA3B} and TP53KO-hGO_{iA3B} clones under each condition. Statistical significance was determined using a two-tailed t-test: n.s., non-significant.

Extended Data Fig. 4 Deamination rate of unmethylated cytosines in A3B-exposed cell-free DNA fragments. CTRL, negative control, 30s, DNA exposed for 30 seconds, 60s, DNA exposed for 60 seconds.

Extended Data Fig. 5 Comparison of the number of APOBEC-associated single nucleotide variants (SNVs) in the Bottleneck Sequencing System (BotSeqS).

(a) The corrected number of A3A-associated SNVs (SBS2+SBS13) in BotSeqS results for hGO_{iA3A} lines under each condition. Statistical significance was determined using a two-tailed t-test: *$p < 0.05$.
(b) The corrected number of A3B-associated SNVs (SBS2+SBS13) in BotSeqS results for hGO_{iA3B} lines under each condition. Statistical significance was determined using a two-tailed t-test: n.s., non-significant.

Extended Data Fig. 6 RNA editing signatures in the hGO_{iA3A} and hGO_{iA3B} clones.

(a) Decomposed spectra of RNA editing signatures.
(b) Proportion of each decomposed RNA editing signature in each hGO_{iA3A} and hGO_{iA3B} clone and doxycycline treatment condition.

Extended Data Fig. 7 Copy number variations of normal gastric organoids.

Extended Data Fig. 8 Detection rate of clustered mutation depending on total number of mutations.

(a) Detection ratio of *omikli* depending on the total number of mutations.
(b) Detection ratio of *kataegis* depending on the total number of mutations. imd_cutoff, cutoff value of inter-mutation distance for classifying clustered mutations

Fig. 1

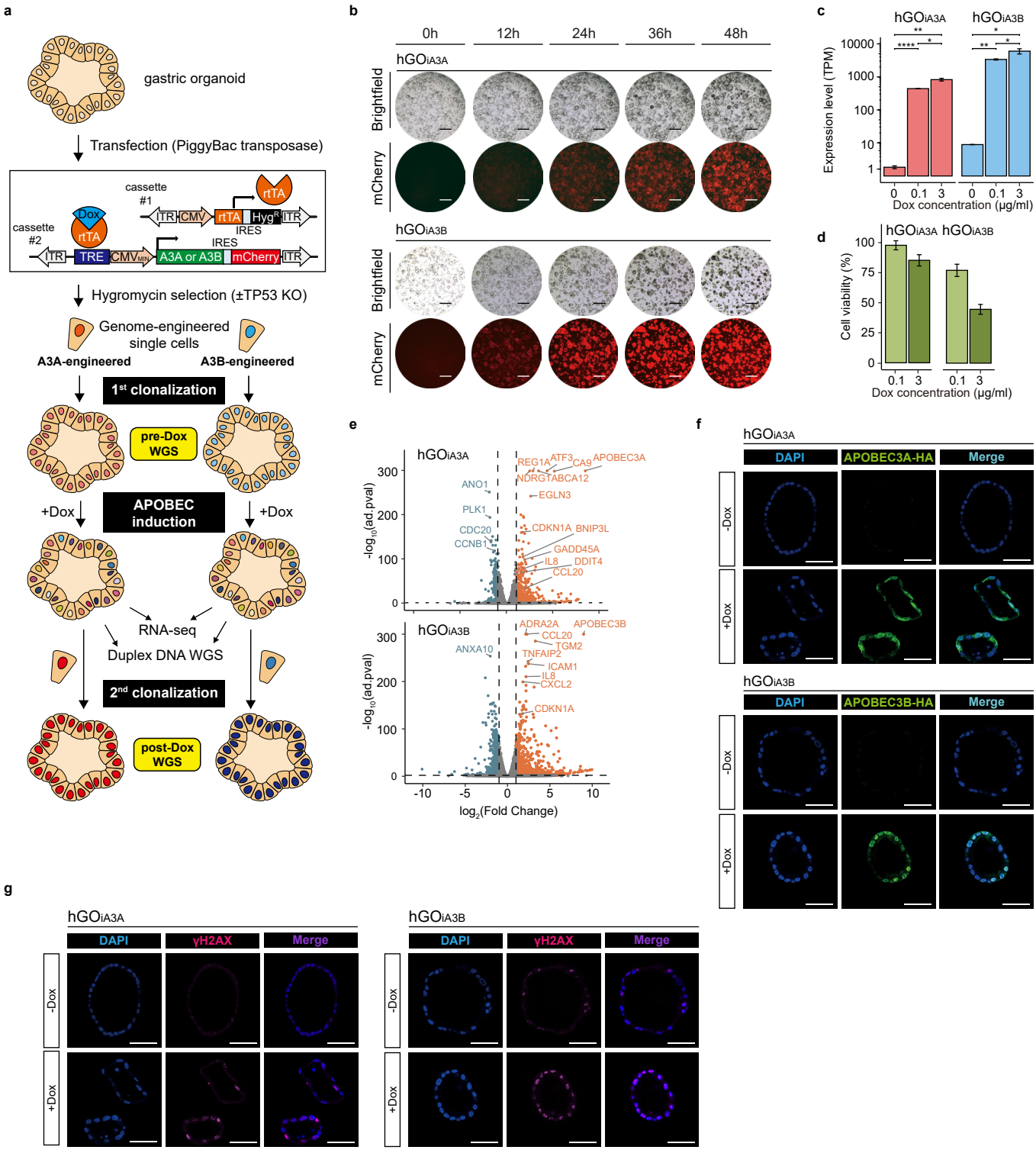

a
gastric organoid
Transfection (PiggyBac transposase)
cassette #1
cassette #2
Dox
rtTA
ITR
CMV
Hyg^R
IRES
TRE
CMV_MIN
A3A or A3B
mCherry
Hygromycin selection (±TP53 KO)
Genome-engineered single cells
A3A-engineered
A3B-engineered
1st clonalization
pre-Dox WGS
+Dox
APOBEC induction
RNA-seq
Duplex DNA WGS
2nd clonalization
post-Dox WGS
b
0h
12h
24h
36h
48h
hGOiA3A
hGOiA3B
Brightfield
mCherry
c
Expression level (TPM)
Dox concentration (μg/ml)
d
Cell viability (%)
e
-log10(ad.pval)
log2(Fold Change)
REG1A
ATF3
CA9
APOBEC3A
NDRG1
ABCA12
EGLN3
ANO1
PLK1
CDC20
CCNB1
CDKN1A
BNIP3L
GADD45A
IL8
DDIT4
CCL20
ADRA2A
APOBEC3B
TGM2
ANXA10
TNFAIP2
ICAM1
CXCL2
f
DAPI
APOBEC3A-HA
APOBEC3B-HA
Merge
-Dox
+Dox
g
γH2AX

Fig. 2

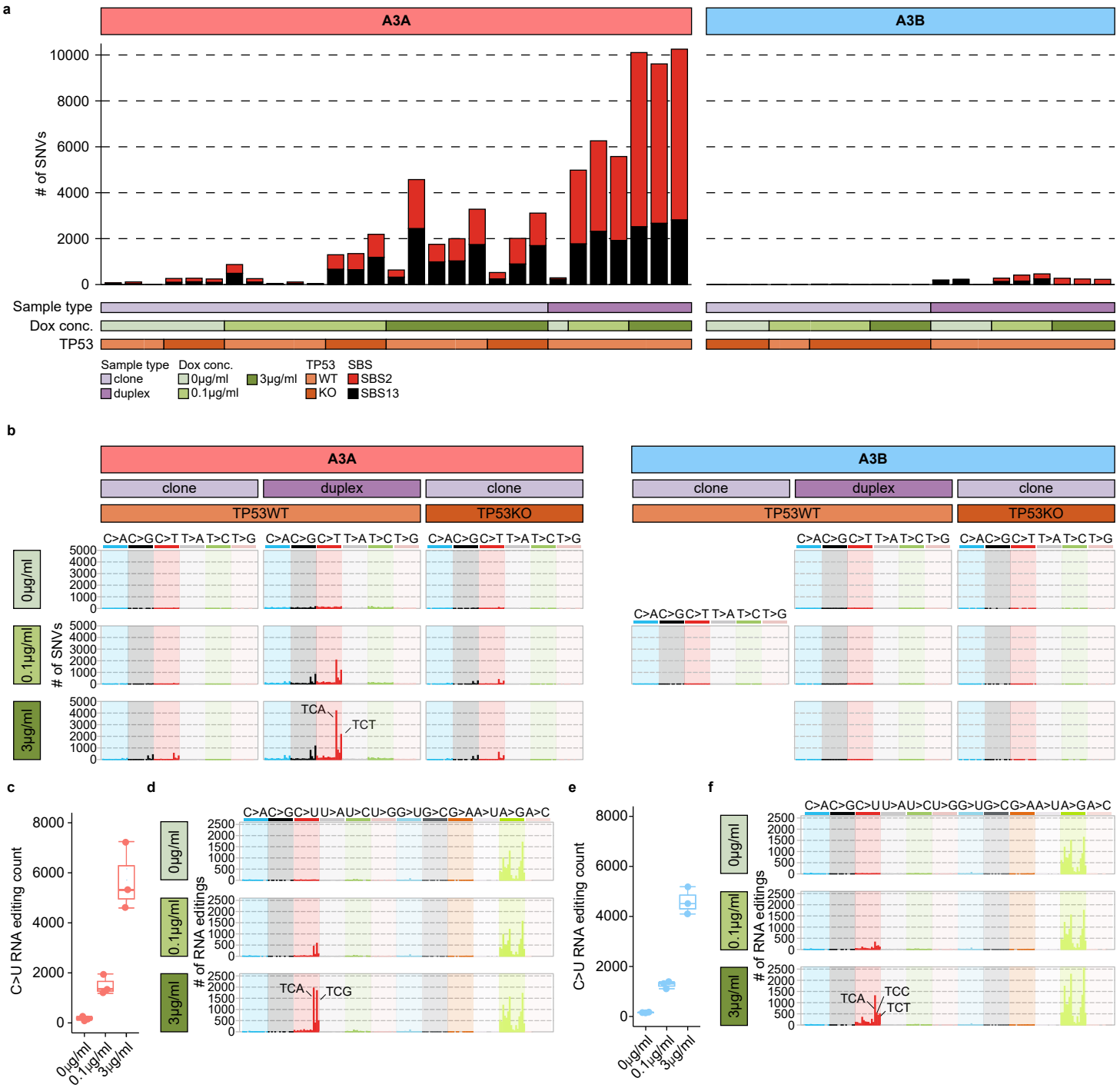

a
A3A
A3B
# of SNVs
Sample type
Dox conc.
TP53
Sample type
clone
duplex
Dox conc.
0μg/ml
0.1μg/ml
3μg/ml
TP53
WT
KO
SBS
SBS2
SBS13
b
A3A
A3B
clone
duplex
clone
TP53WT
TP53KO
C>AC>GC>T T>A T>C T>G
0μg/ml
0.1μg/ml
3μg/ml
# of SNVs
TCA
TCT
c
C>U RNA editing count
0μg/ml
0.1μg/ml
3μg/ml
d
C>AC>GC>UU>AU>CU>GG>UG>CG>AA>UA>GA>C
# of RNA editings
TCA
TCG
e
C>U RNA editing count
f
TCA
TCC
TCT

Fig. 3

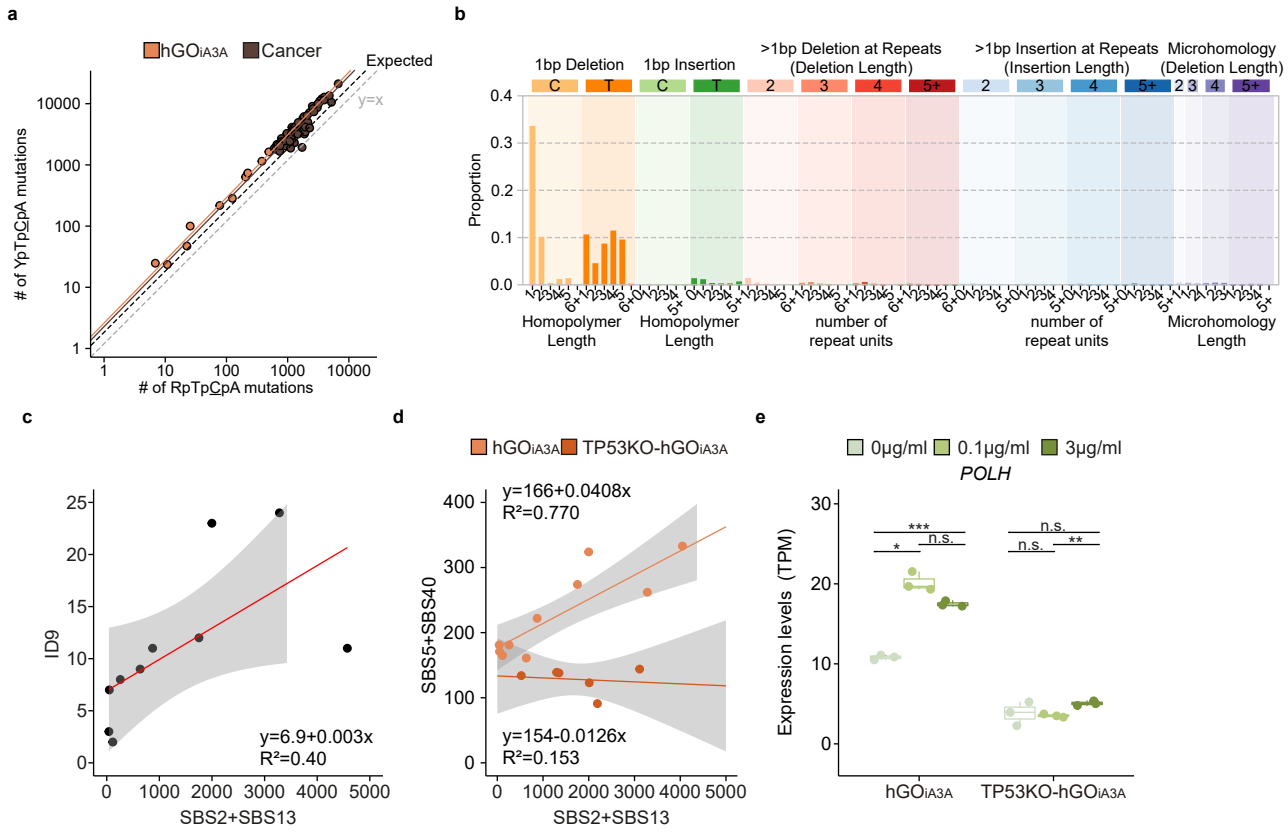

a
hGOiA3A
Cancer
Expected
y=x
# of YpTpCpA mutations
# of RpTpCpA mutations
b
1bp Deletion
1bp Insertion
>1bp Deletion at Repeats (Deletion Length)
>1bp Insertion at Repeats (Insertion Length)
Microhomology (Deletion Length)
Proportion
Homopolymer Length
number of repeat units
Microhomology Length
c
ID9
SBS2+SBS13
y=6.9+0.003x
R²=0.40
d
hGOiA3A
TP53KO-hGOiA3A
SBS5+SBS40
y=166+0.0408x
R²=0.770
y=154-0.0126x
R²=0.153
SBS2+SBS13
e
0μg/ml
0.1μg/ml
3μg/ml
POLH
Expression levels (TPM)
hGOiA3A
TP53KO-hGOiA3A
n.s.
***
**
*

Fig. 4

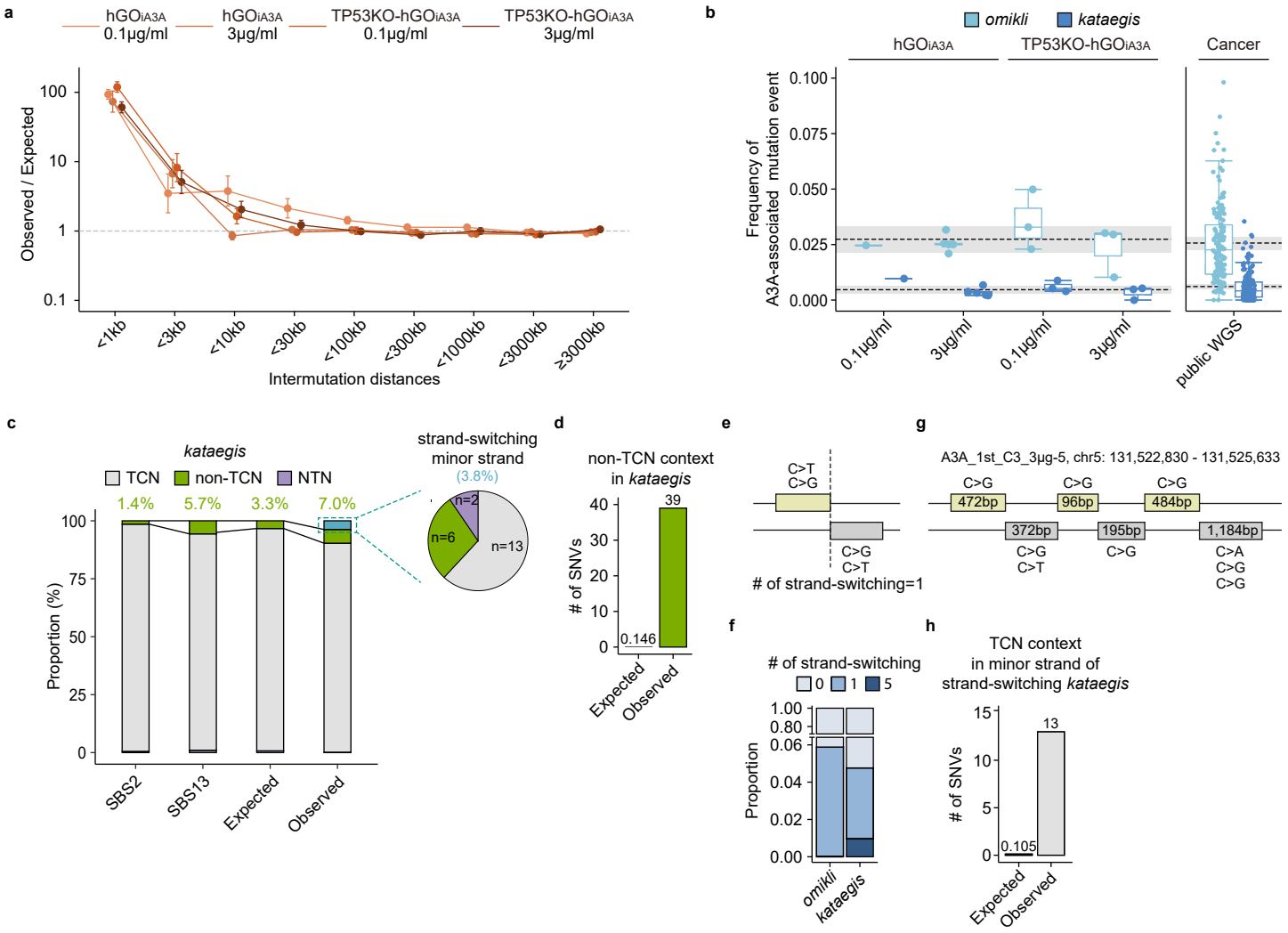
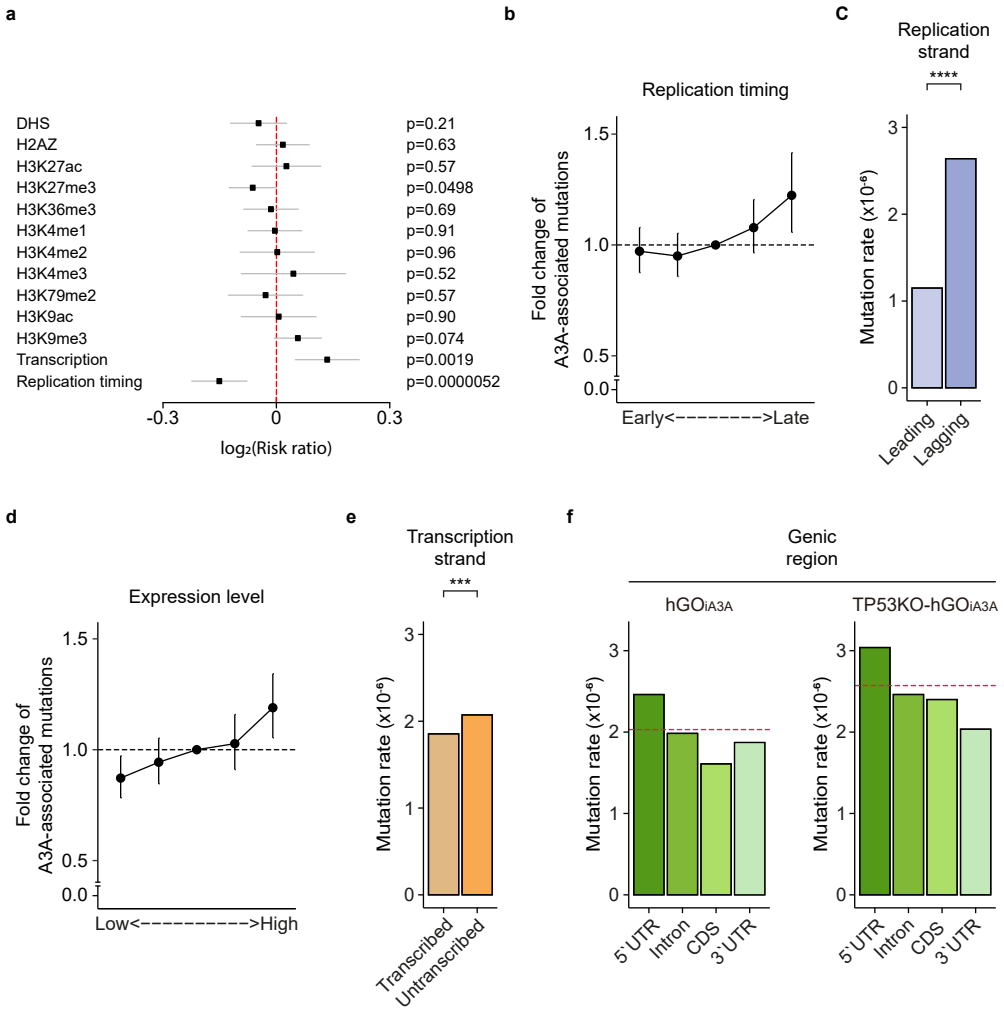
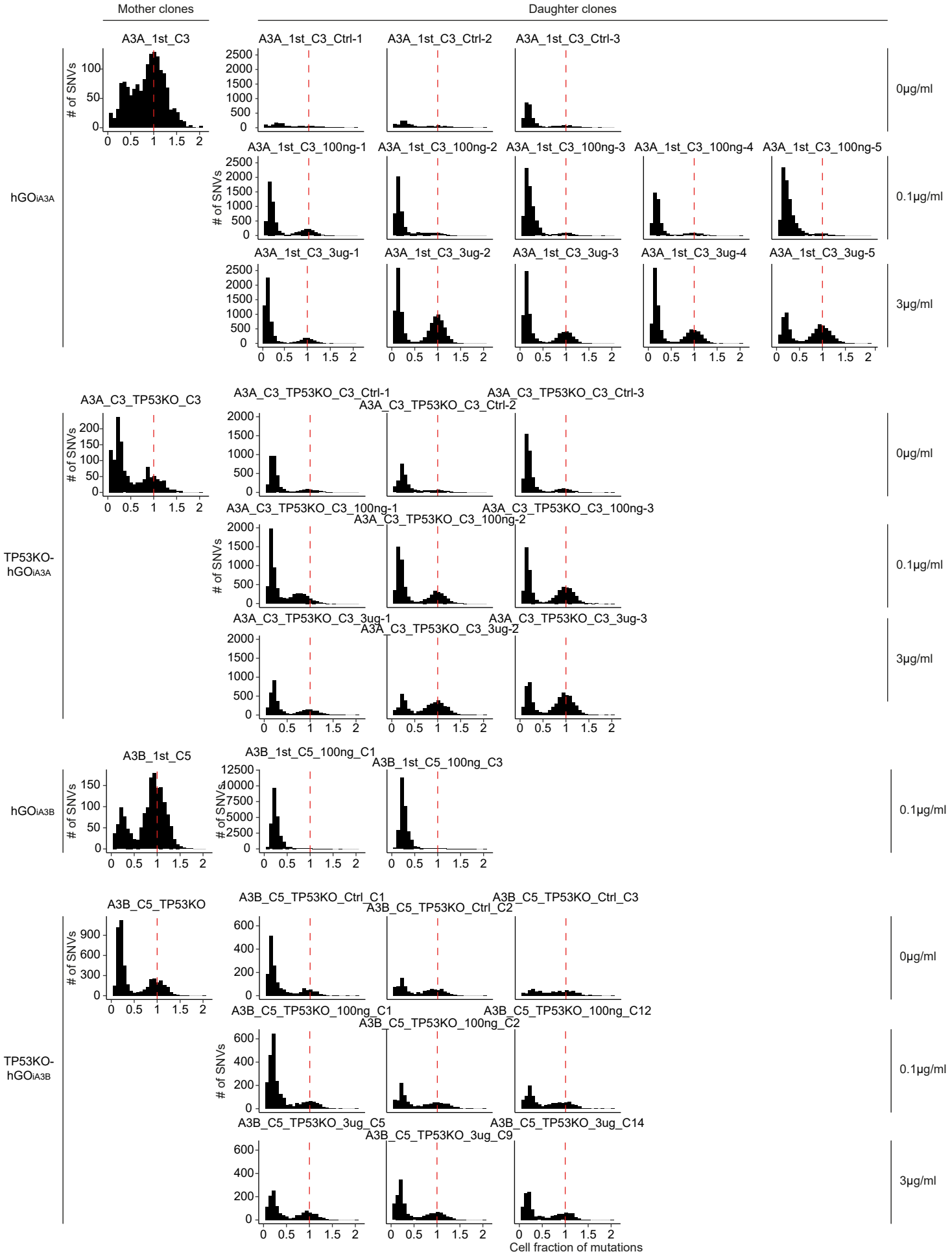
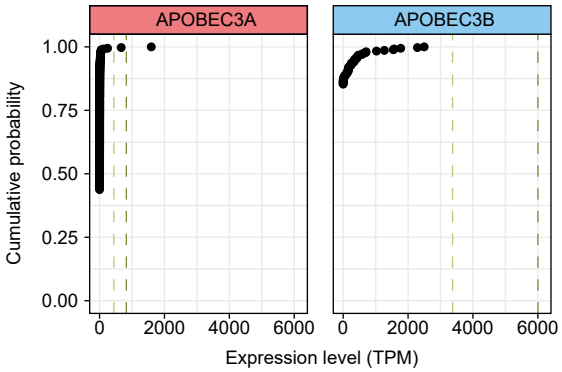

a
hGOiA3A 0.1μg/ml
hGOiA3A 3μg/ml
TP53KO-hGOiA3A 0.1μg/ml
TP53KO-hGOiA3A 3μg/ml
Observed / Expected
Intermutation distances
<1kb
<3kb
<10kb
<30kb
<100kb
<300kb
<1000kb
<3000kb
≥3000kb
b
omikli
kataegis
hGOiA3A
TP53KO-hGOiA3A
Cancer
Frequency of A3A-associated mutation event
0.1μg/ml
3μg/ml
public WGS
c
kataegis
TCN
non-TCN
NTN
1.4%
5.7%
3.3%
7.0%
Proportion (%)
SBS2
SBS13
Expected
Observed
strand-switching minor strand
(3.8%)
n=2
n=6
n=13
d
non-TCN context in kataegis
# of SNVs
39
0.146
Expected
Observed
e
C>T
C>G
C>G
C>T
# of strand-switching=1
g
A3A_1st_C3_3μg-5, chr5: 131,522,830 - 131,525,633
C>G
472bp
96bp
484bp
372bp
195bp
1,184bp
C>G
C>T
C>A
C>G
C>G
f
# of strand-switching
0
1
5
Proportion
omikli
kataegis
h
TCN context in minor strand of strand-switching kataegis
# of SNVs
13
0.105
Expected
Observed

Fig. 5

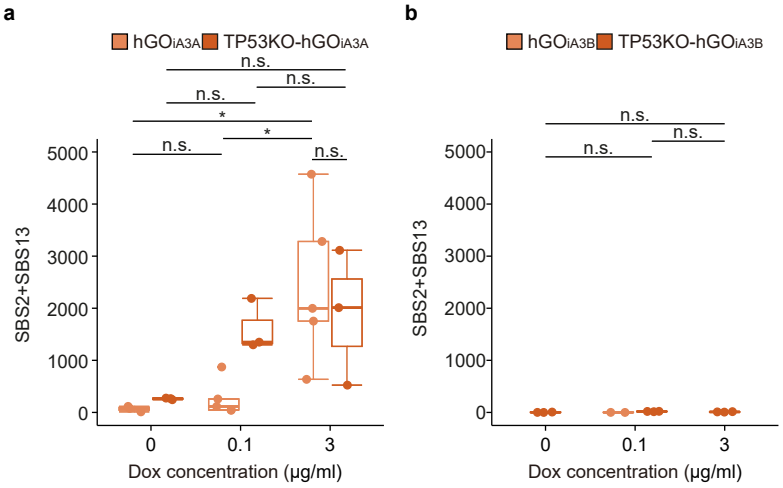

a
DHS p=0.21
H2AZ p=0.63
H3K27ac p=0.57
H3K27me3 p=0.0498
H3K36me3 p=0.69
H3K4me1 p=0.91
H3K4me2 p=0.96
H3K4me3 p=0.52
H3K79me2 p=0.57
H3K9ac p=0.90
H3K9me3 p=0.074
Transcription p=0.0019
Replication timing p=0.0000052
-0.3
0
0.3
log2(Risk ratio)
b
Replication timing
Fold change of A3A-associated mutations
1.5
1.0
0.5
0.0
Early<--------->Late
C
Replication strand
****
Mutation rate (x10-6)
3
2
1
0
Leading
Lagging
d
Expression level
Fold change of A3A-associated mutations
1.5
1.0
0.5
0.0
Low<---------->High
e
Transcription strand
***
Mutation rate (x10-6)
3
2
1
0
Transcribed
Untranscribed
f
Genic region
hGOiA3A
TP53KO-hGOiA3A
Mutation rate (x10-6)
3
2
1
0
5'UTR
Intron
CDS
3'UTR



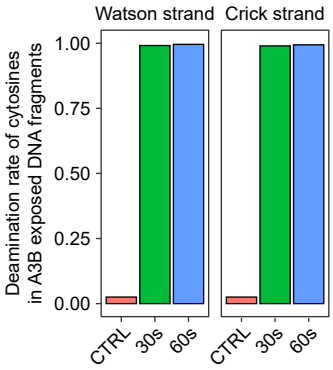
Extended Data Fig. 1 Distribution of cell fraction of mutations in each clone in this study.



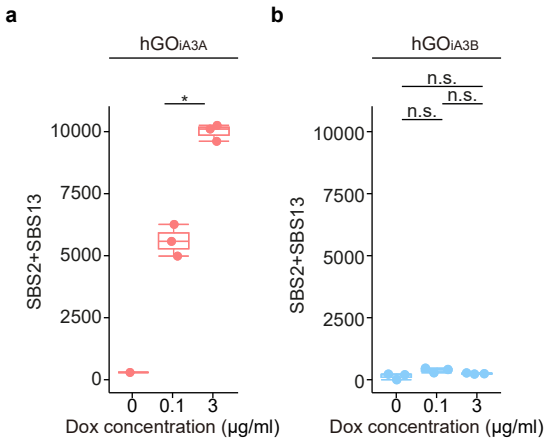
Extended Data Fig. 2 Cumulative probabilities of expression levels of A3A and A3B in single breast cancer cells. Green line; average expression levels of A3A and A3B following 0.1 μg/ml doxycycline for 48 hours in each corresponding line, hGOiA3A and hGOiA3B, respectively; dark green line; average expression levels of A3A and A3B following 3 μg/ml doxycycline for 48 hours in each corresponding line, hGOiA3A and hGOiA3B, respectively.



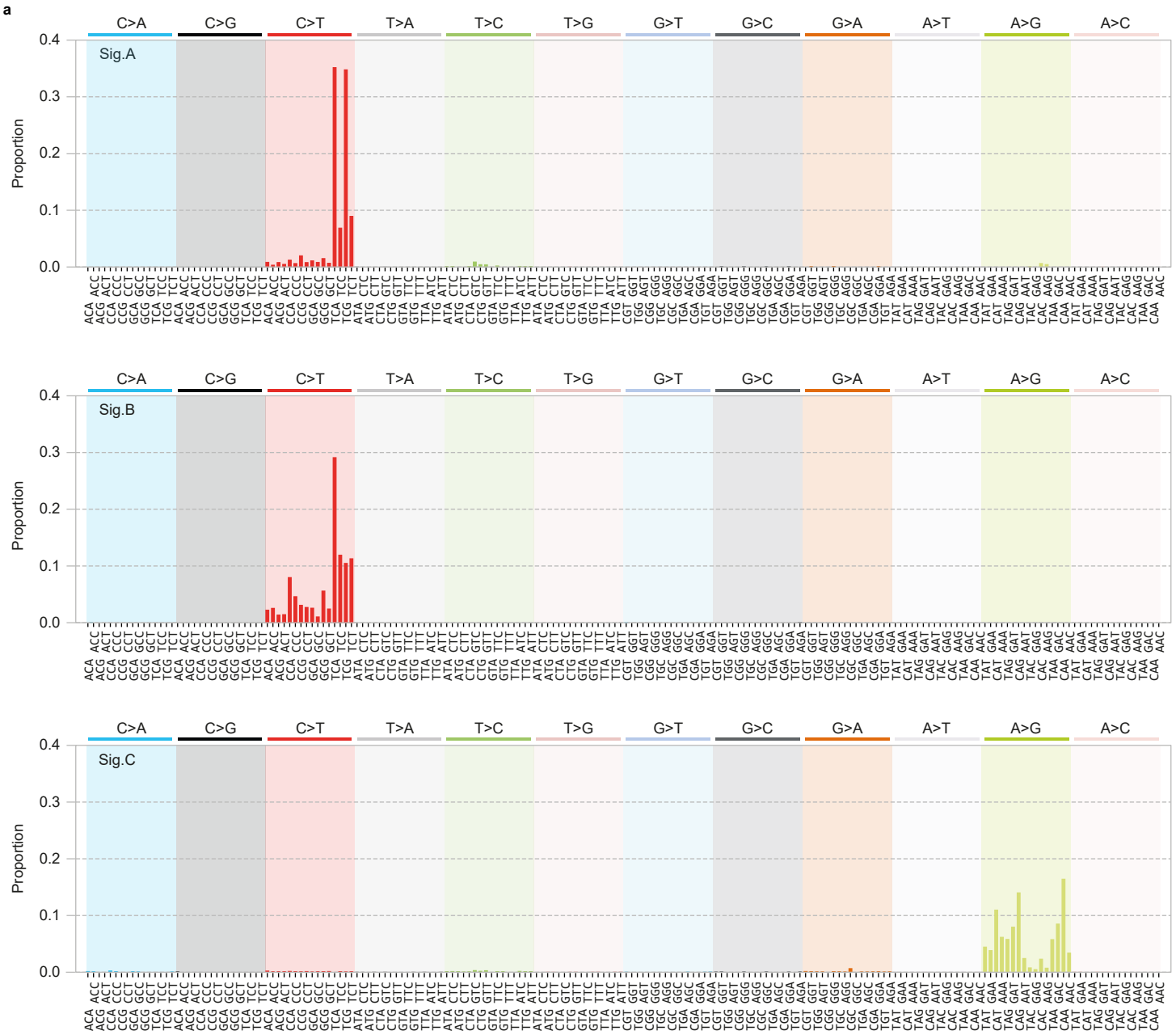
Extended Data Fig. 3 Comparison of the number of APOBEC-associated single nucleotide variants (SNVs) in the clone sequencing. (a) The number of A3A-associated SNVs (SBS2+SBS13) in hGOiA3A and TP53KO-hGOiA3A clones under each condition. Statistical significance was determined using a two-tailed t-test: *p < 0.05, **p < 0.005, ***p < 0.0005,****p < 0.00005. (b) The number of A3B-associated SNVs (SBS2+SBS13) in hGOiA3B and TP53KO-hGOiA3B clones under each condition. Statistical significance was determined using a two-tailed t-test: n.s., non-significant.

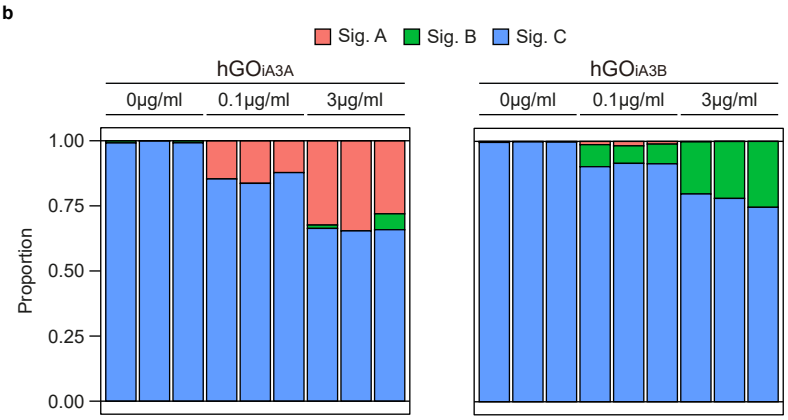


Extended Data Fig. 4 Deamination rate of unmethylated cytosines in A3B-exposed cell-free DNA fragments. CTRL, negative control, 30s, DNA exposed for 30 seconds, 60s, DNA exposed for 60 seconds.

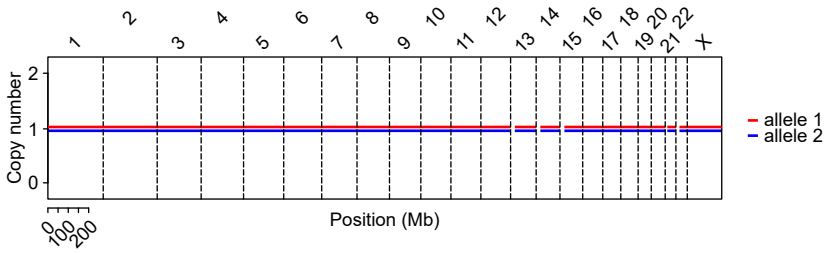


Extended Data Fig. 5 Comparison of the number of APOBEC-associated single nucleotide variants (SNVs) in the Bottleneck Sequencing System (BotSeqS). (a) The corrected number of A3A-associated SNVs (SBS2+SBS13) in BotSeqS results for hGOiA3A lines under each condition. Statistical significance was determined using a two-tailed t-test: *$p < 0.05$. (b) The corrected number of A3B-associated SNVs (SBS2+SBS13) in BotSeqS results for hGOiA3B lines under each condition. Statistical significance was determined using a two-tailed t-test: n.s., non-significant.

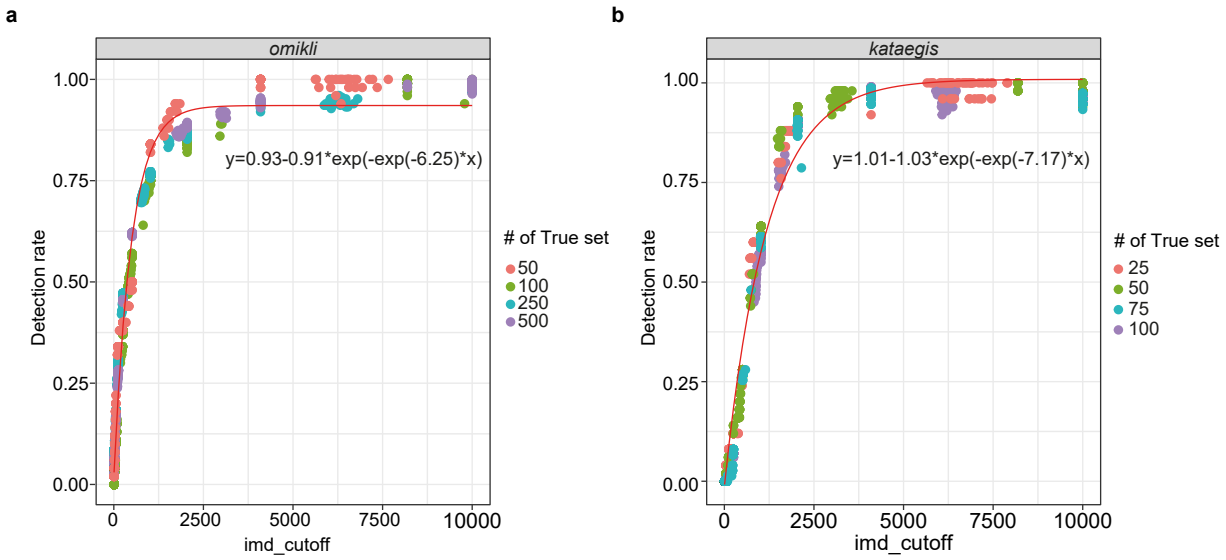




Extended Data Fig. 6 RNA editing signatures in the hGOiA3A and hGOiA3B clones. (a) Decomposed spectra of RNA editing signatures. (b) Proportion of each decomposed RNA editing signature in each hGOiA3A and hGOiA3B clone and doxycycline treatment condition.



Extended Data Fig. 7 Copy number variations of normal gastric organoids.



Extended Data Fig. 8 Detection rate of clustered mutation depending on total number of mutations. (a) Detection ratio of *omikli* depending on the total number of mutations. (b) Detection ratio of *kataegis* depending on the total number of mutations. imd_cutoff, cutoff value of inter-mutation distance for classifying clustered mutations