Supplemental Material - Table of Contents

TITLE: APOBEC3A drives deaminase mutagenesis in human gastric epithelium

Running title: APOBEC3A is mutagenic in human cell

AUTHORS

Yohan An[1,10], Ji-Hyun Lee[2,10], Joonoh Lim[1,3,10], Jeonghwan Youk[1,11], Seongyeol Park[1,12], Ji-Hyung Park[1,4], Kijong Yi[1,3], Taewoo Kim[1,13], Chang Hyun Nam[1], Won Hee Lee[1,14], Soo A Oh[3], Yoo Jin Bae[5], Thomas M. Klompstra[2], Haeun Lee[1], Jinju Han[1], Junehwak Lee[6], Jung Woo Park[6], Jie-Hyun Kim[5], Hyunki Kim[7], Hugo Snippert[8], Bon-Kyoung Koo[2,9]*, and Young Seok Ju[1,3]*

A) Supplemental Figures

Supplemental Table S5. Summary of somatic mutations (single nucleotide variants (SNVs), indels, and structural variations (SVs)) in each clone.

Supplemental Table S6. Summary of SNV mutational signature burdens in BotSeqS results.

Supplemental Table S7. Summary of the sequencing throughput and the number of C>U RNA editing sites across the hGO_{iA3A} and hGO_{iA3B} lines.

Supplemental Table S8. Summary of the recurrent A3A- and A3B-specific C>U RNA editing sites across the hGO_{iA3A} and hGO_{iA3B} lines.

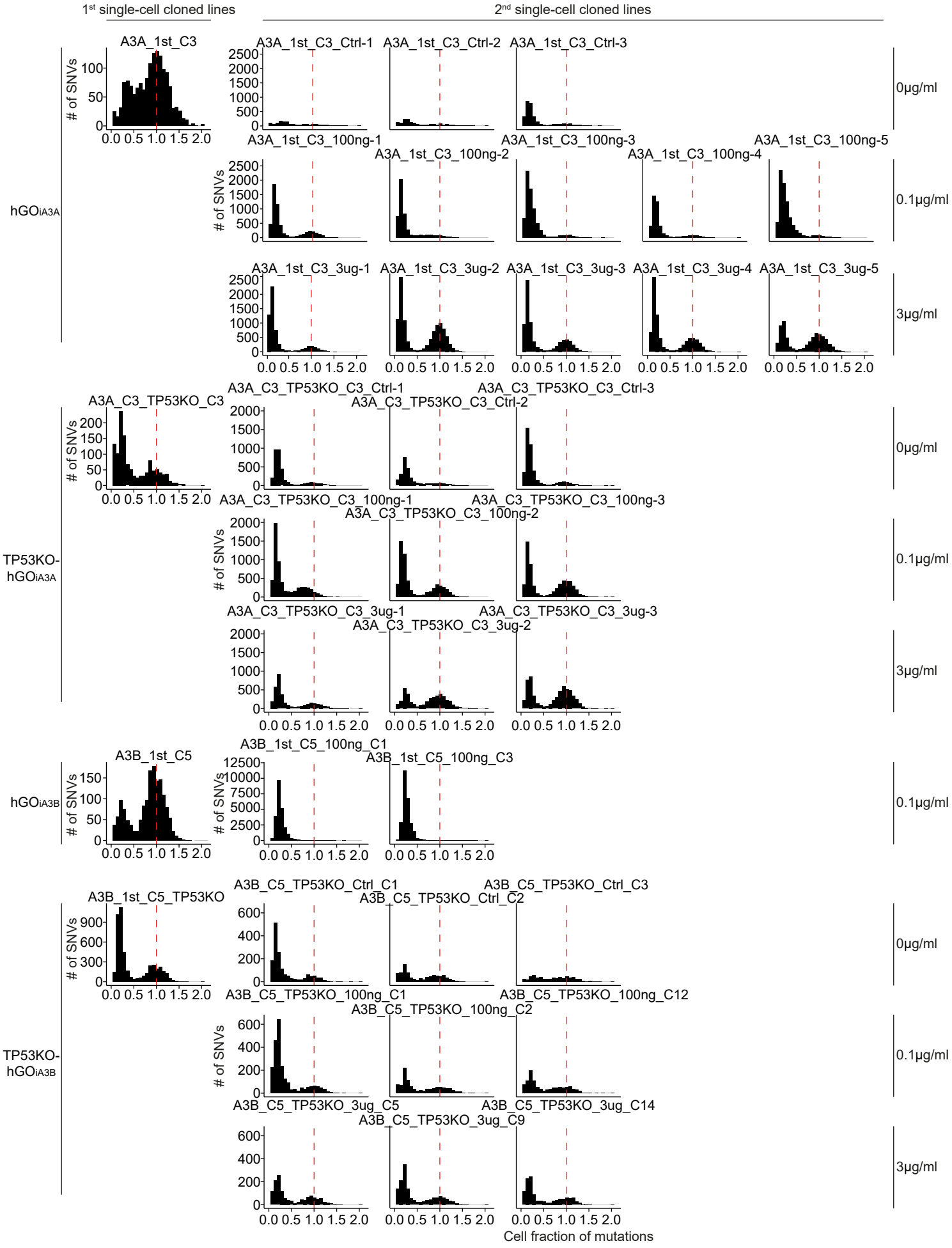
Supplemental Table S9. Summary of the simulation-adjusted number of clustered mutation events in the hGO_{iA3A} clones and public cancer data.

Supplemental Table S10. Summary of negative binomial regression of APOBEC-associated mutations with three genomic features (H3K27me3, local transcription, and replication timing).

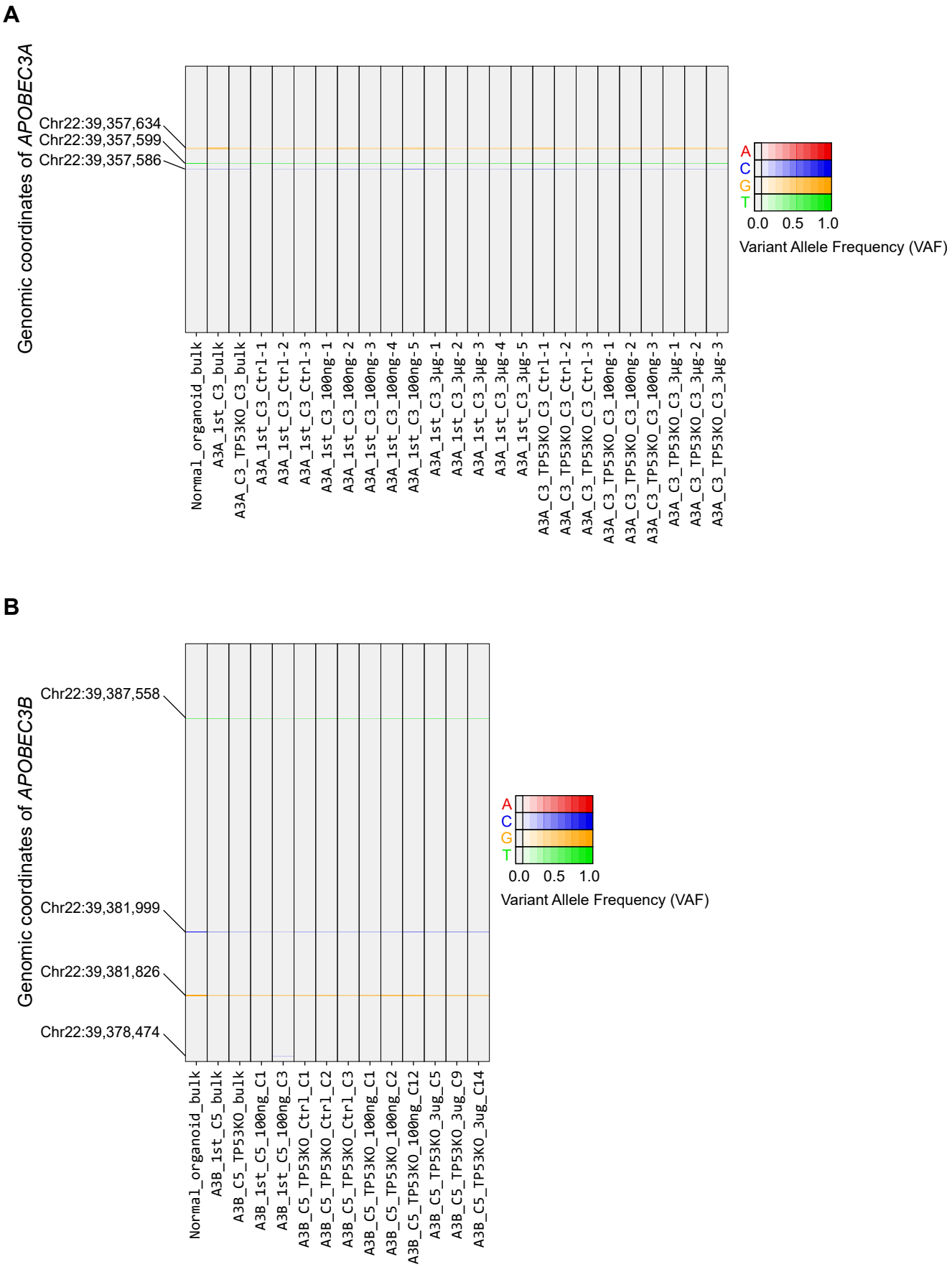
Supplemental Table S11. Culture medium composition for human gastric organoids.

Supplemental Table S12. Electroporation programs for transfection of organoids.

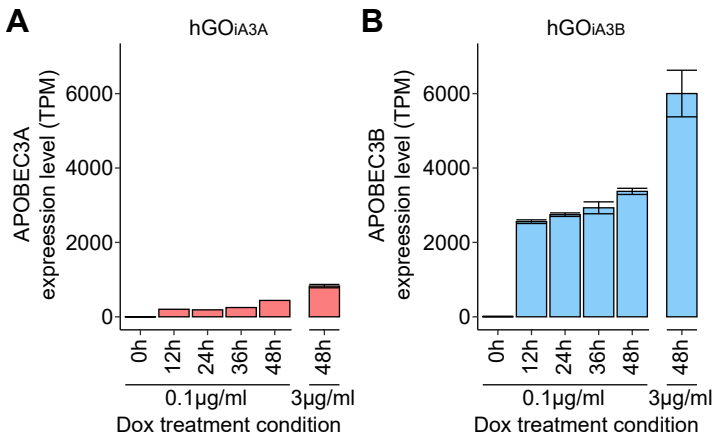
Supplemental Figures



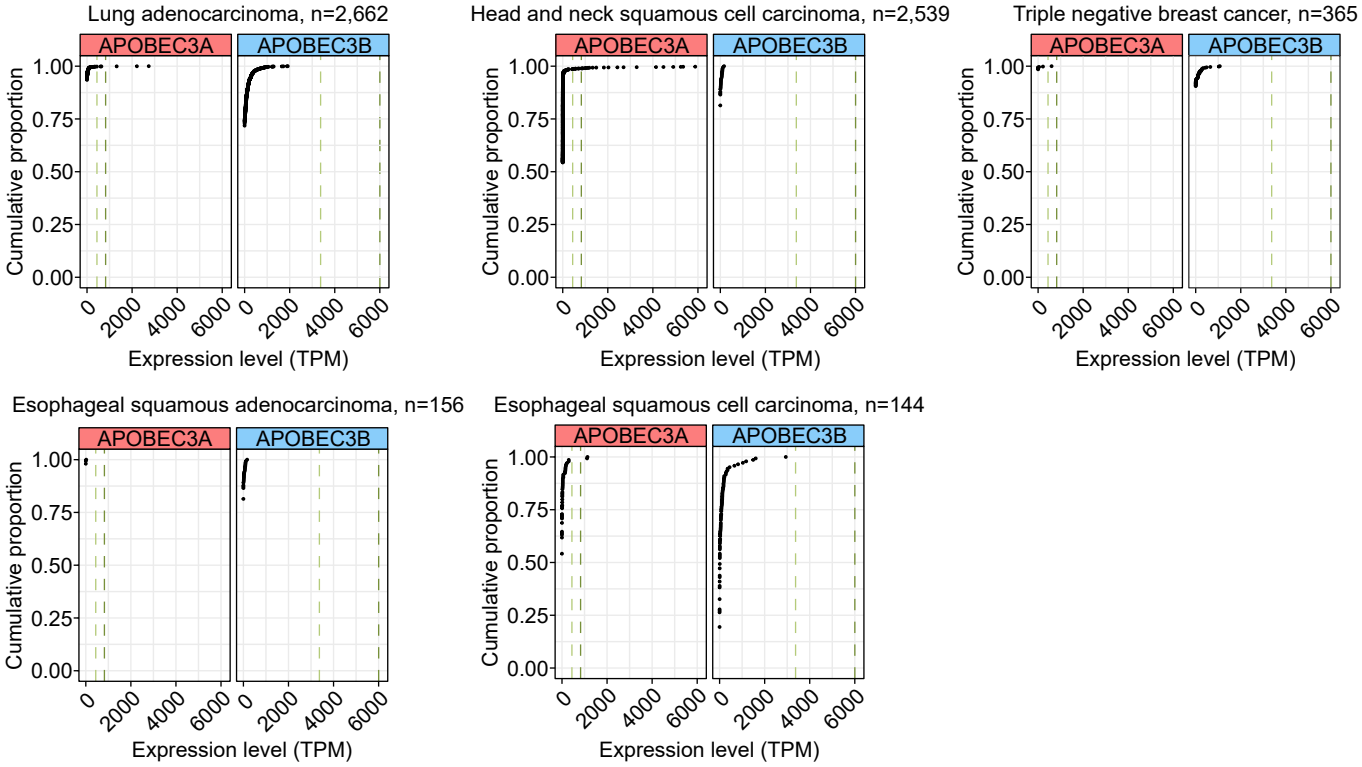
Supplemental Figure S1. Distribution of cell fraction of mutations in individual clones. Each bar represents the cell fraction of SNVs, indicating the proportion of cells wihin a clone that possess the mutation.



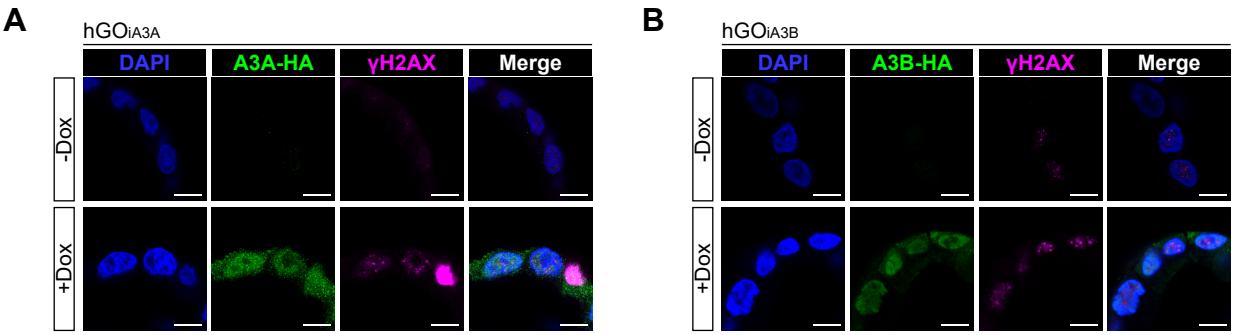
Supplemental Figure S2. Variant allele frequencies (VAFs) against reference APOBEC sequence in each clonal organoid line. (A) hGOiA3A and TP53KO-hGOiA3A **(B)** hGOiA3B and TP53KO-hGOiA3B. All missense mutations annotated with genomic positions originated from either the *A3A* or *A3B* in endogenous copies, except for one missense mutation (Chr22:39,378,474) in A3B_1st_C5_100ng_C3 sample, which resulted from a sequencing error.



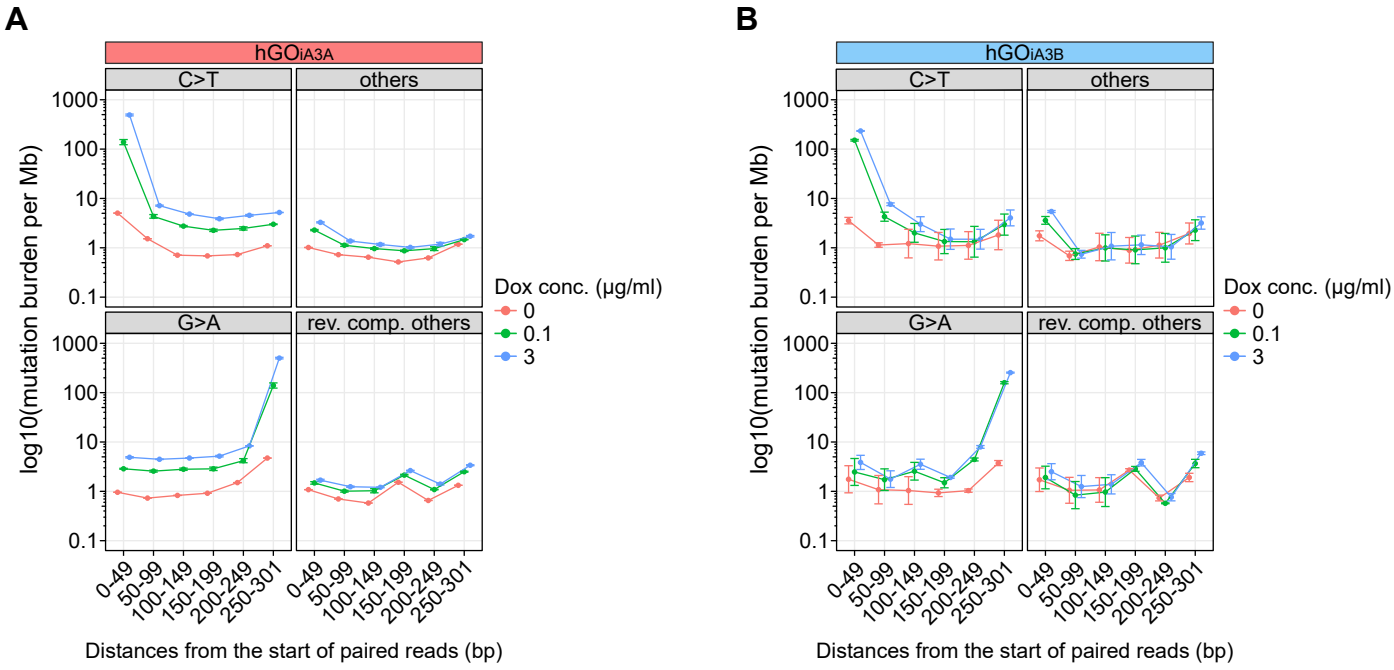
Supplemental Figure S3. Expression levels of A3A or A3B in the corresponding organoid lines under doxycycline treatment. (A) Expression levels of APOBEC3A (A3A) in hGO$_{iA3A}$ line following under each doxycycline condition (n=3 per condition). **(B)** Expression levels of APOBEC3B (A3B) in hGO$_{iA3B}$ line following under each doxycycline condition (n=3 per condition). Data are presented as mean ± 95% confidence intervals.



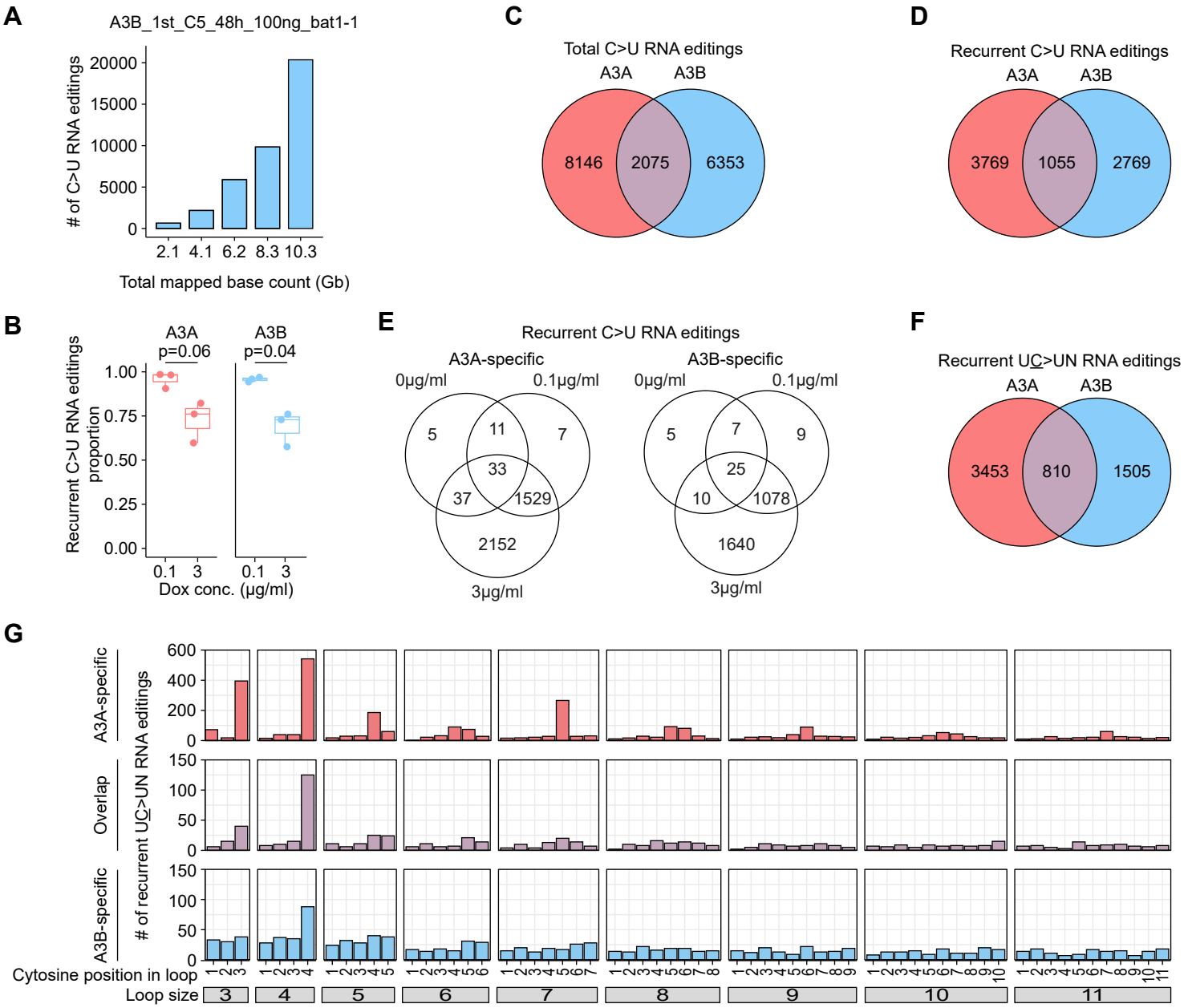
Supplemental Figure S4. Cumulative proportions of expression levels of A3A and A3B in single cancer cells across multiple types of cancer. Green line: average expression levels of A3A and A3B following 0.1 μg/ml doxycycline treatment for 48 hours in each corresponding line, hGO_{iA3A} and hGO_{iA3B}, respectively; dark green line: average expression levels of A3A and A3B following 3 μg/ml doxycycline treatment for 48 hours in each corresponding line, hGO_{iA3A} and hGO_{iA3B}, respectively.



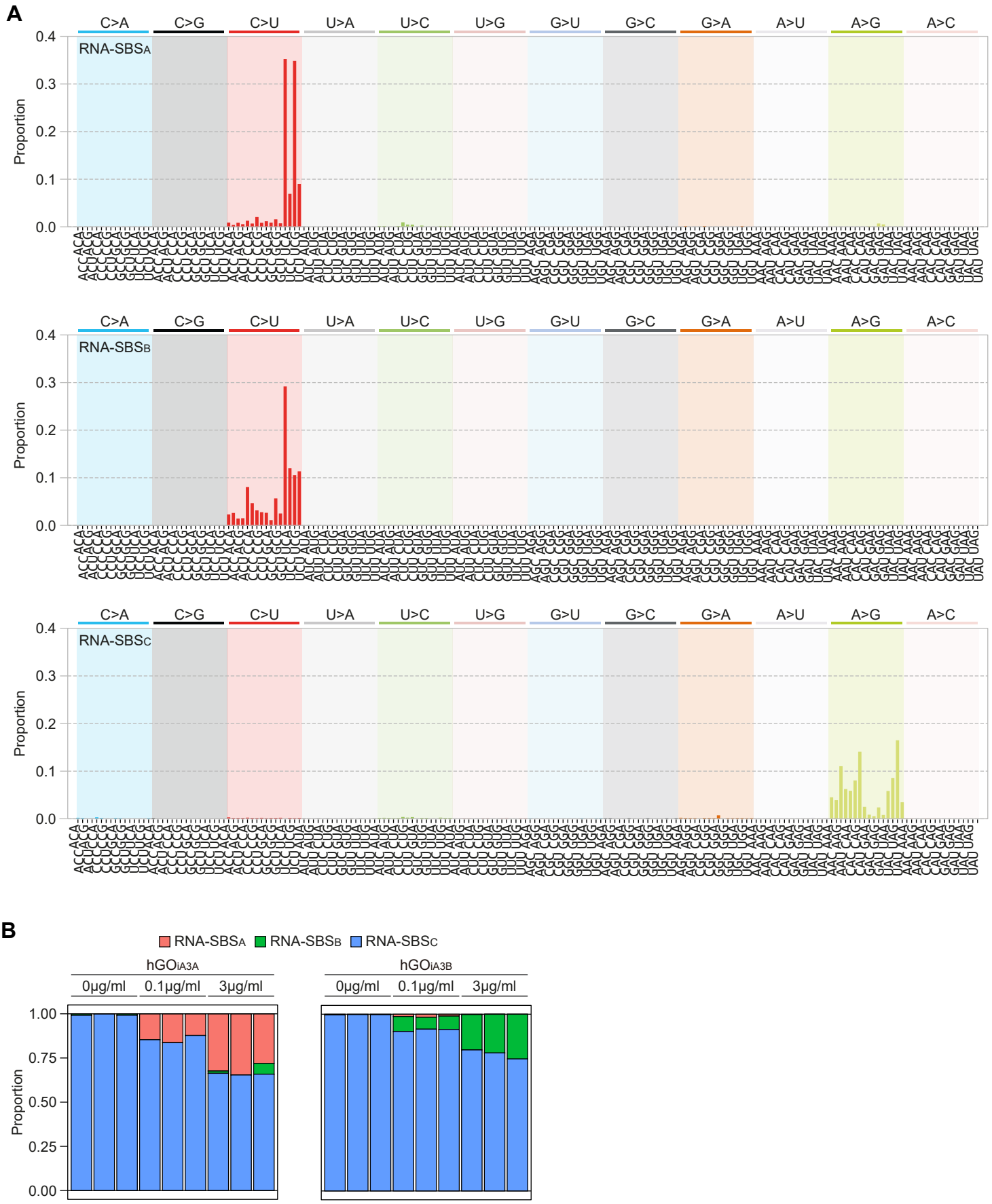
Supplemental Figure S5. High resolution images of immunohistochemistry. (A) hGO$_{iA3A}$ and **(B)** hGO$_{iA3B}$ lines. Scale bars represent 10 μm.



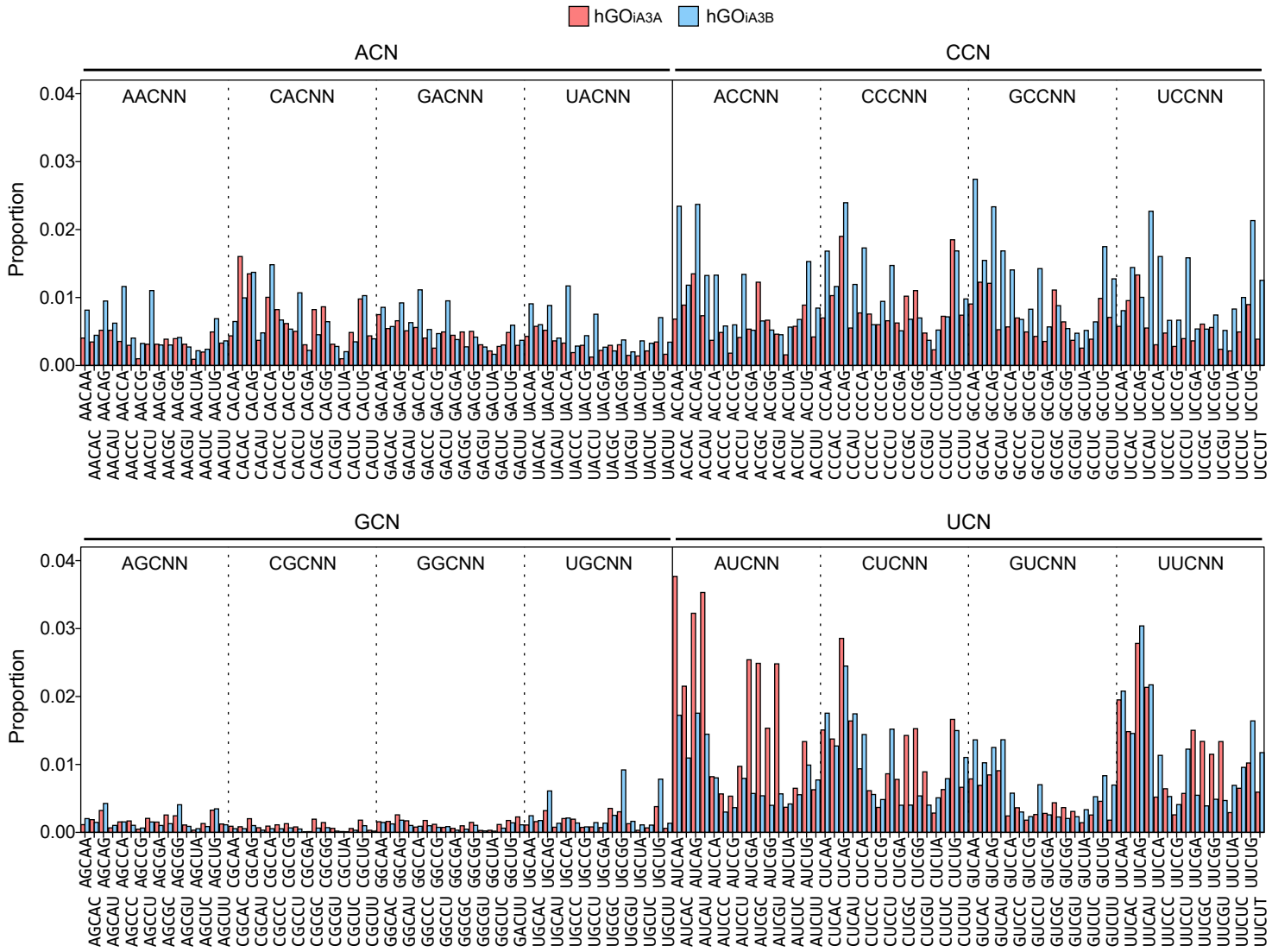
Supplemental Figure S6. Distribution of distances from the start position of paired reads to each single nucleotide variant (SNV) in BotSeqS. (A) hGOiA3A and **(B)** hGOiA3B lines. The 0-position is the 5' head region of each DNA fragment.



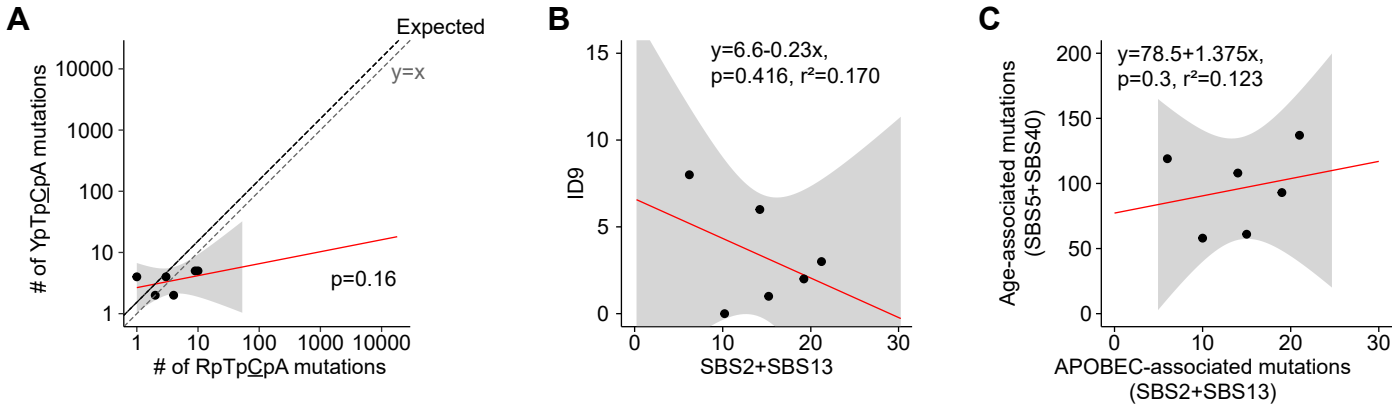
Supplemental Figure S7. Characteristics of APOBEC-associated C>U RNA editing sites in the hGO$_{iA3A}$ and hGO$_{iA3B}$ lines. (A) Correlation between the total number of mapped bases and the number of C>U RNA editing in A3B_1st_C5_48h_100ng_bat1-1 sample. **(B)** Proportions of recurrent C>U RNA editing (≥2 samples, with ≥1 doxycycline-treated sample) following doxycycline treatement. The p-values were determined using a *t*-test. **(C)** Venn diagram showing the total number of C>U RNA editing sites in each models. **(D)** Venn diagram showing the number of recurrent C>U RNA editing sites. **(E)** Venn diagram showing the number of recurrent C>U RNA editing sites depending on the doxycycline treatment condition. **(F)** Venn diagram showing the number of recurrent UC>UN RNA editing sites. **(G)** Distribution of recurrent UC>UN RNA editing sites according to the properties of predicted secondary structures.



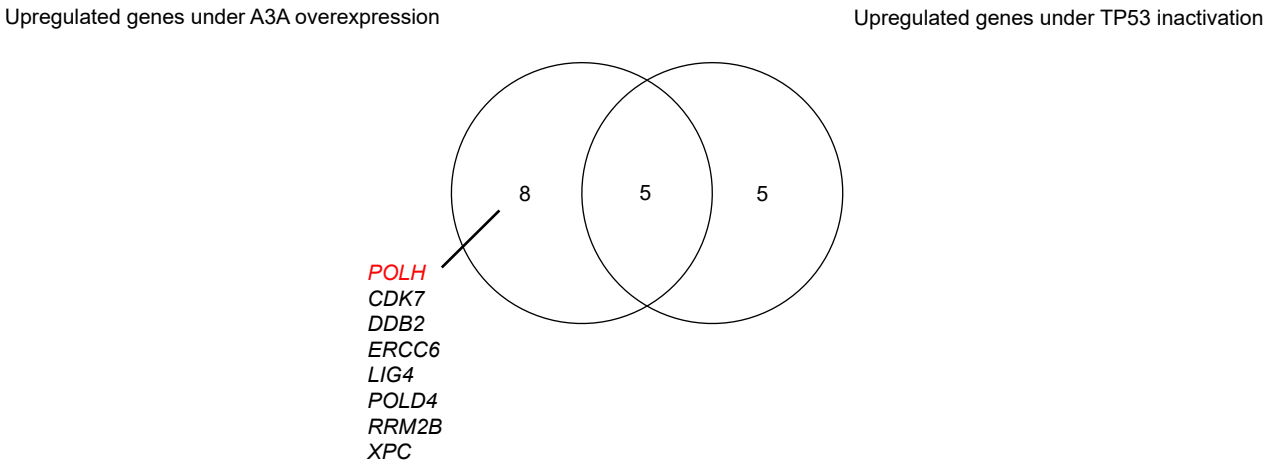
Supplemental Figure S8. *De novo* extracted RNA editing signatures in the hGO$_{iA3A}$ and hGO$_{iA3B}$ lines. (A) Spectra of *de novo* extracted RNA editing signatures from hGO$_{iA3A}$ and hGO$_{iA3B}$ samples following the doxycycline treatment. **(B)** Proportion of contributing RNA editing signatures in each hGO$_{iA3A}$ and hGO$_{iA3B}$ samples following doxycycline treatment conditions.



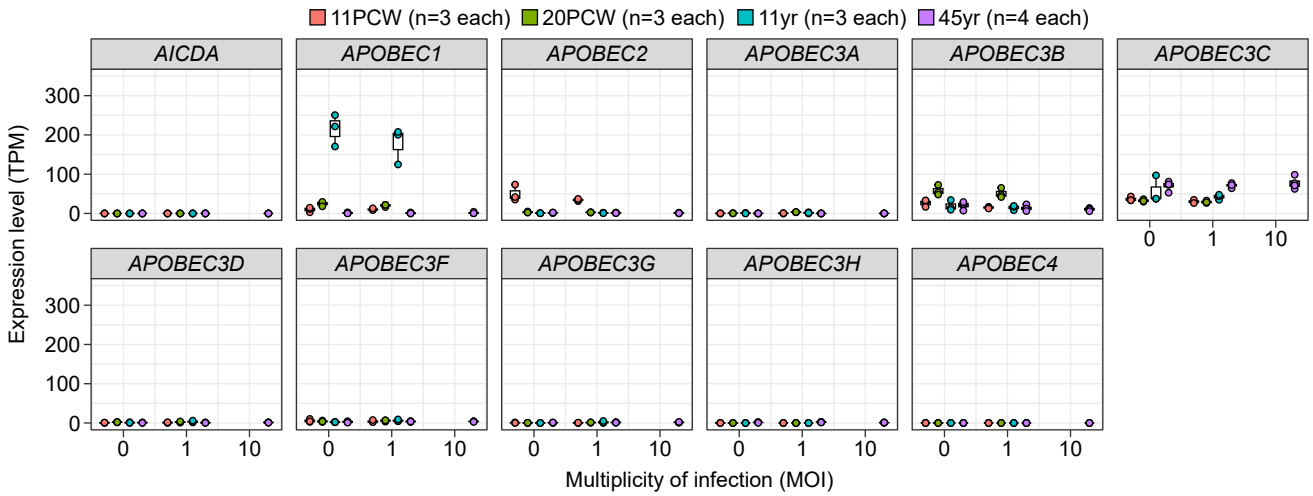
Supplemental Figure S9. Spectra of C>U RNA editing in pentanucleotide contexts from the hGO$_{iA3A}$ and hGO$_{iA3B}$ lines following 3µg/ml doxycycline treatment for 48 hours.



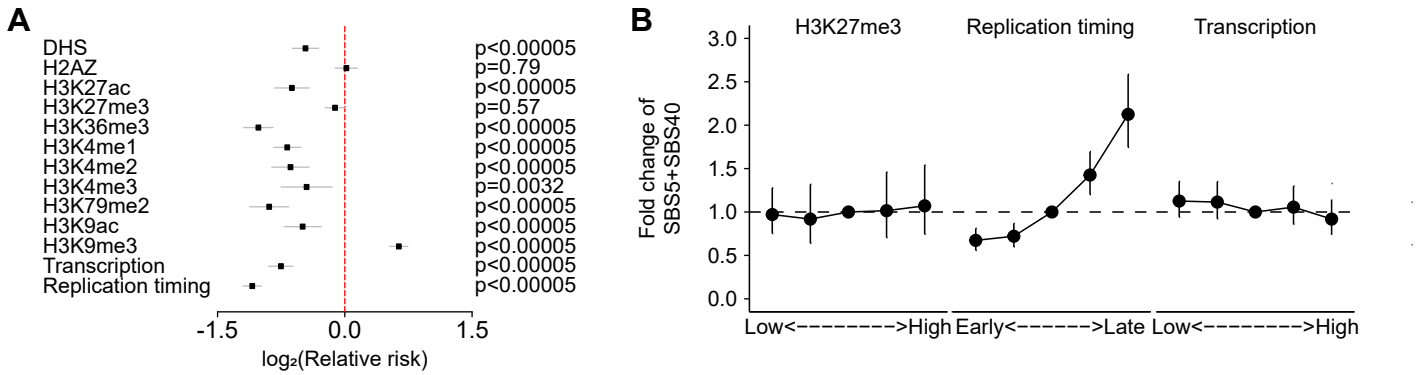
Supplemental Figure S10. Characteristics of A3B-associated mutational signatures in the TP53KO-hGO$_{iA3B}$ clones. (A) Context preference of A3B between YpTpCpA and RpTpCpA context in hGO$_{iA3B}$ lines. All detected C>T and C>G SNVs in TP53KO-hGO$_{iA3B}$ clones were utilized for the analysis. **(B)** Correlation between A3B-associated base substitutions and ID9 contributing indels. **(C)** Associations between A3B-associated (SBS2 and SBS13) and age-associated (SBS5 and SBS40) base substitutions.



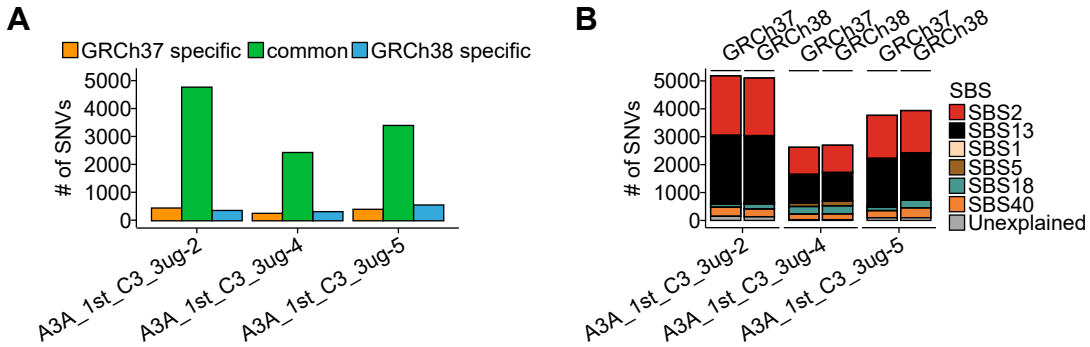
Supplemental Figure S11. Differentially expressed genes contributing DNA repair in the hGO$_{iA3A}$ and TP53KO-hGO$_{iA3A}$ lines following 3µg/ml doxycycline treatment for 48 hours. The numbers in the Venn diagram represent the number of differentially expressed genes belonging to each comparison group.



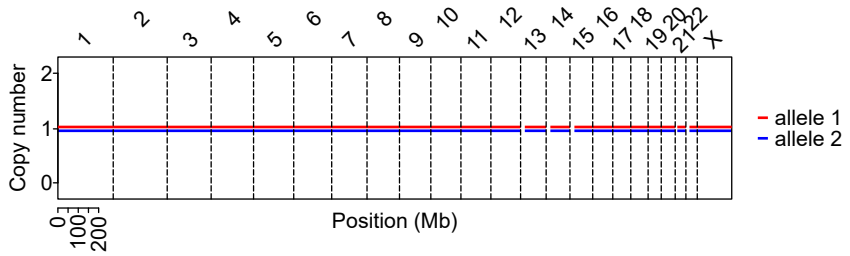
Supplemental Figure S12. Expression levels of 11 APOBEC family genes in human gastric organoids following SARS-CoV-2 infection. PCW, post conception week; yr, year



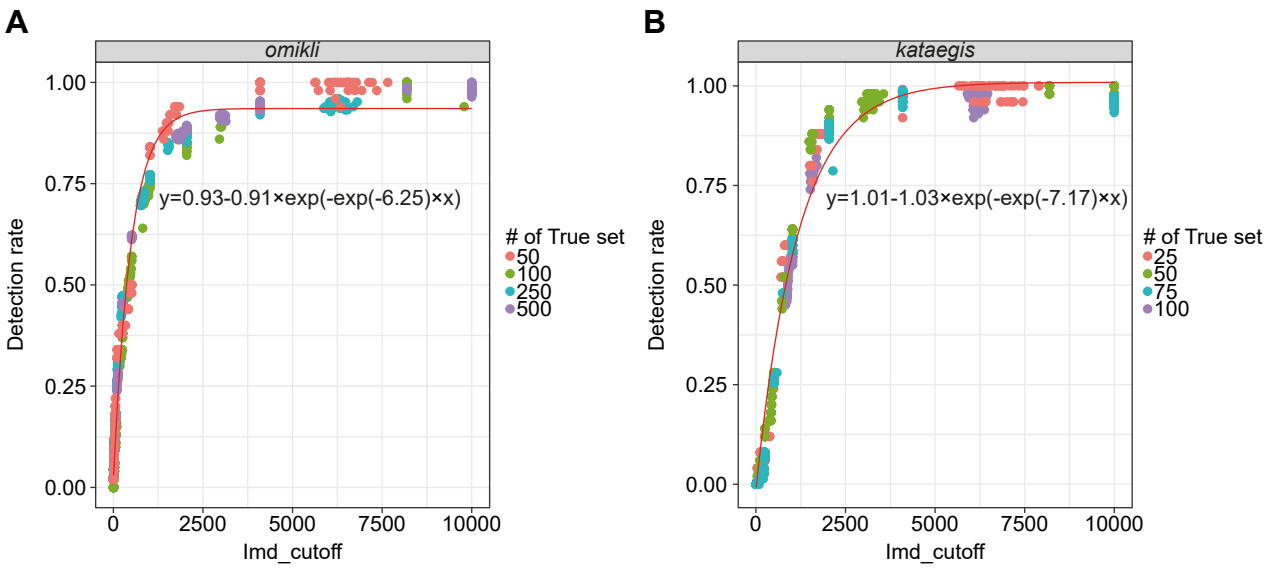
Supplemental Figure S13. Enrichment with SBS5- and SBS40-associated mutations detected from the hGO$_{iA3A}$ clones following doxycycline treatment. (A) Enrichment with epigenetic markers. **(B)** Enrichment with H3K27me3, replication timing and gene expression levels.



Supplemental Figure S14. Impact of human reference genome GRCh37 and GRCh38 on the detection of SNVs. (A) Number of SNVs according to the GRCh37 and GRCh38 reference genome. SNV coordinates derived from GRCh37 were lifted over to GRCh38. **(B)** Mutational signature burdens of SNVs called against the GRCh37 and GRCh38 reference genome.



Supplemental Figure S15. Copy number variations of normal gastric organoids.



Supplemental Figure S16. Detection rates of clustered mutations relative to the total number of mutations. (A) Detection rate of *omikli* according to the total number of mutations. **(B)** Detection rate of *kataegis* according to the total number of mutations. lmd_cutoff, intermutation distance cutoff

Supplemental Methods

Human stomach organoid culture medium

Composition of stomach organoid culture medium was adopted from previous research with minimal modification**.** The complete medium was 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; **Supplemental Table S11**).

Passaging organoid

The medium in each well was replaced with Cell Recovery Solution (Corning, Cat No.354253), and the plates were incubated for 40~60 minutes at 4°C to dissolve the Matrigel. The organoid pellet was isolated by removing the supernatant after centrifugation at 300g for 5 minutes at 4°C. Organoids were then incubated for 5 minutes at 37°C with Accutase (Stemcell Technology, Cat No.07922), followed by pipetting 10-20 times to dissociate them into clusters of 10-15 cells. After diluting the Accutase by adding 1 mL of ADF medium (Gibco), the cell suspension was centrifuged at 300g for 5 minutes at 4°C to isolate the pellet. Organoids were then seeded in 12- or 24-well plates at a ratio of 1:4 to 1:6 in Matrigel.

Construction of doxycycline inducible APOBEC overexpression vector

To create pPB-CMVmin-APOBEC (*A3A* or *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. A Kozak consensus sequences (GCCACC) was added at the 5' end of each A3A or A3B cDNA sequence, and an HA tag sequences (5'-TAC-CCA-TAC-GAT-GTT-CCA-GAT-TAC-GCT-3') was appended to the 3' ends. Subsequently, *XhoI* and *NotI* restriction enzyme cut sites (CTCGAG and GCGGCCGC, respectively) were added to the 5' and 3' ends of the construct. *De novo* gene synthesis from GenScript (Piscataway, NJ) was used for the synthesis of the two constructs. Following synthesis, the A3A and A3B constructs were cloned into the pPB-CMVmin-TRE-IRES-*mCherry* backbone vector (Lee et al. 2022) to make pPB-CMVmin-TRE-APOBEC (*A3A* or *A3B*)-IRES-*mCherry* expression cassettes. Plasmid preparation, including both mini-prep and maxi-prep, was performed using commercial competent cells (Biosearch Technologies, Cat No.60106-1) and mini-prep kit (QIAGEN, Cat No.27104) and maxi-prep kit (QIAGEN, Cat No.12123) according to the manufacturers' protocols.

Transfection of organoids

Organoids were dissociated using the same protocol employed for routine passaging. Previously established protocols were utilized for the transfection (Gaebler et al. 2020; Fujii et al. 2015). A combination of three plasmids were used for electroportation: (1) TRE-APOBEC (*A3A* or *A3B*)-IRES-*mCherry* cassette and (2) CMV-*rtTA*-*HygR* cassette (3) piggyBac 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 programs were adopted from the previous literature (**Supplemental Table S11;** Fujii et al. 2015). Transfected organoids were cultured for seven days with the medium composition described in the established 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 single-cell cloning was conducted according to a previously reported method by using FACSAria II (BD Biosciences) and manually picking single cell originated organoids by pipetting (Youk et al. 2021). 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 in the reported protocol (Fujii et al. 2015). Selection and single-cell cloning steps were conducted using the same method as for hGO_{iA3A} lines.

Western blotting process

Lysate-transferred membranes were blocked in 5% BSA (Sigma-Aldrich, A1470) in TBS-T (Biosesang, TR2007-100-74) for 30 minutes at room temperature. After blocking, membranes were incubated overnight at 40℃ with primary antibodies diluted in the same blocking buffer: Anti-HA.11 Epitope Tag Antibody (1:5000; BioLegend, 951514) or Anti-α-Antin-1 (1:1000; Sigma-Aldrich, A2066). After washing three times with 1x TBS-T on a rocking incubator at room temperature for 5 minutes each, membranes were incubated with HRP-conjugated secondary antibodies (1:2000), including goat anti-mouse IgG-HRP for Anti HA (Santa Cruz, sc-2005) and goat anti-rabbit IgG-HRP for AntiActin (Santa Cruz, sc-2004), diluted in 5% BSA in TBS-T for 1 hour at room temperature. Membranes were then washed three times with 1x TBS-T on a rocking incubator at room temperature for 5 minutes each. Signal was detected using enhanced chemiluminescence (Thermo Scientific, 34580) and

imaged using a Bio-Rad ChemiDoc MP Imaging System. Membranes were stripped for reblotting by incubating them in stripping buffer for 10 minutes after imaging. Reblotting was done by subsequently repeating the whole process from primary antibody incubation.

Counting the ratio between endogenous APOBEC mRNA and overexpressed APOBEC mRNA

To distinguish transcripts originating from endogenous APOBEC genes versus the exogenous construct, we utilize sequence differences between the two sources. For A3A, two heterozygous variants at Chr22:39,357586 (C/T) and 39,357,599 (C/T) were used. In the endogenous A3A allele, reads carried the same base at both positions-either C/C or T/T-while the inserted construct contained C/T bases at the two positions, respectively. For A3B, a homozygous variant at Chr22:39,381,999 (C) was present in the endogenous gene, while the construct carried only T at the same locus.

Procedure of measurement of viability of organoids

Culture mediums in each well in the plate were completely removed. 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 Celltiter-Glo 3D Assay kit (Promega, Cat No.G9681) 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.

Whole-genome sequencing SNVs and indels filtering criteria

Somatic mutations (SNVs and indels) were filtered with the following criteria: (1) depth >5; (2) median mapping quality (MQ) of variant supporting reads ≥25; (3) median MQ of

reference supporting reads ≥25; (4) abs (median MQ of variant supporting reads - median MQ of reference supporting reads) ≤10; (5) variant allele frequencies of position in panel of normal <0.02; (6) the number of variant supporting reads in the corresponding 1st single-cell cloned line= 0; (7) the number of variant supporting read counts ≥3; (8) median base quality of variant supporting reads ≥30; (9) Ratio of variant supporting reads having soft-clipping <0.5; (10) the distance of variants from the near end of read ≥3; (11) variant supporting reads having mixed population of F1R2 and F2R1 reads.

BotSeqS variant filtering criteria

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-supporting reads ≥3 or 90% of reads; (3) distances of mutations from both extreme of each read >5 bp; (4) distances of mutations from the fragment ends >100 bp, taking strand orientation into account (e.g., excluding C>T variants 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.

RNA editing filtering criteria

The RNA editing spectrum was represented as 192 patterns of single-base changes, defined by the substitution type and the immediately adjacent bases on the 5’ and 3’ ends in the canonical mRNA sequence. The number of features is twice that of the DNA mutational signature (Alexandrov et al. 2013) because RNA does not exhibit symmetric base changes in general.

A total of 18 experimental datasets were used as input, divided into two subsets: the A3A and A3B sets. Each subset included three batches from experiments with 0 μg/ml, 0.1 μg/ml, and 3 μg/ml APOBEC3 exposure, totaling nine batches per set. Initially, the signature extraction method from our previous work (Youk et al. 2024) was applied without

modification, and the optimal number of independent RNA editing signatures was determined. As a result, four potential RNA editing signatures were identified, attributable to APOBEC3A, APOBEC3B, ADAR, and other endogenous editing processes.

The APOBEC3A and APOBEC3B signatures aligned well with the C>T editing patterns observed in experiments with varying exposure levels. However, the ADAR signature consistently appeared within these signatures at substantial proportions. This “leakage” phenomenon occurs due to the constitutive activity of ADAR, causing its signal to appear in RNA editing spectra across all experimental conditions. This complicates accurate quantification of RNA editing processes in a sample.

To mitigate the leakage phenomenon, an L1 constraint was imposed on the signature matrix (*W*). We adapted a sparse NMF update algorithm (Le Roux et al. 2015), which originally applied the L1 constraint to the exposure matrix (*H*). We modified its objective function by transposing the data matrix (*V*) as well as *W* and *H*, performing the updates, and then back-transposing to obtain *W*. The matrix *W* was column-normalized so that each column summed to one. The sparsity coefficient (μ) was set to 0.6 after evaluating several values in the range [0.001, 10].

The A3A set yielded well-separated ADAR and APOBEC3A signatures, and the A3B set yielded well-separated ADAR and APOBEC3B signatures. The ADAR signatures learned from the A3A and A3B sets were consistent (cosine similarity = 0.999) and were therefore merged into a single ADAR signature by averaging. Endogenous RNA editing signatures other than ADAR were negligible and contributed little to the overall analysis, so no additional signatures were pursued.

Finally, the three RNA editing signatures were refitted to individual samples to quantify the contributions of each RNA editing process in each sample.

Simulation of for correction of the number of clustered mutation event

Identified clustered mutations from our analysis and 703,858 SNVs from 125 samples in the 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 divided into 33 groups, increasing in steps of 2,000 from 2,000 to 10,000; in steps of 5,000 from 15,000 to 100,000; and in steps of 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 under 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 subtracting the number of clustered mutations from the target number of SNVs. With SigProfilerClusters (Bergstrom et al. 2022), 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 (**Supplemental Fig. S16**) using the “drm” function with the fct=AR.3() option in the drc, R package (Ritz et al. 2019).

Supplemental References

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. 2013. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**: 246–259. http://dx.doi.org/10.1016/j.celrep.2012.12.008.

Bergstrom EN, Kundu M, Tbeileh N, Alexandrov LB. 2022. Examining clustered somatic mutations with SigProfilerClusters. *Bioinformatics* **38**: 3470–3473. https://academic.oup.com/bioinformatics/article-pdf/38/13/3470/49883724/btac335.pdf (Accessed August 1, 2025).

Fujii M, Matano M, Nanki K, Sato T. 2015. Efficient genetic engineering of human intestinal organoids using electroporation. *Nat Protoc* **10**: 1474–1485. http://dx.doi.org/10.1038/nprot.2015.088.

Gaebler A-M, Hennig A, Buczolich K, Weitz J, Welsch T, Stange DE, Pape K. 2020. Universal and Efficient Electroporation Protocol for Genetic Engineering of Gastrointestinal Organoids. *J Vis Exp*. http://dx.doi.org/10.3791/60704.

Lee J-H, Kim S, Han S, Min J, Caldwell B, Bamford A-D, Rocha ASB, Park J, Lee S, Wu S-HS, et al. 2022. p57 imposes the reserve stem cell state of gastric chief cells. *Cell Stem Cell* **29**: 826–839.e9. http://dx.doi.org/10.1016/j.stem.2022.04.001.

Le Roux J, Hershey JR, Weninger F. 2015. Deep NMF for speech separation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE http://ieeexplore.ieee.org/document/7177933/.

Mihara E, Hirai H, Yamamoto H, Tamura-Kawakami K, Matano M, Kikuchi A, Sato T, Takagi J. 2016. Active and water-soluble form of lipidated Wnt protein is maintained by a serum glycoprotein afamin/α-albumin. https://elifesciences.org/articles/11621 (Accessed August 30, 2024).

Ritz C, Jensen SM, Gerhard D, Streibig JC. 2019. *Dose-Response Analysis Using R*. CRC Press https://books.google.com/books/about/Dose_Response_Analysis_Using_R.html?hl=&id=VzWoDwAAQBAJ.

Youk J, Kwon HW, Kim R, Ju YS. 2021. Dissecting single-cell genomes through the clonal organoid technique. *Exp Mol Med* **53**: 1503–1511. http://dx.doi.org/10.1038/s12276-021-00680-1.

Youk J, Kwon HW, Lim J, Kim E, Kim T, Kim R, Park S, Yi K, Nam CH, Jeon S, et al. 2024. Quantitative and qualitative mutational impact of ionizing radiation on normal cells. *Cell Genom* **4**: 100499. http://dx.doi.org/10.1016/j.xgen.2024.100499.