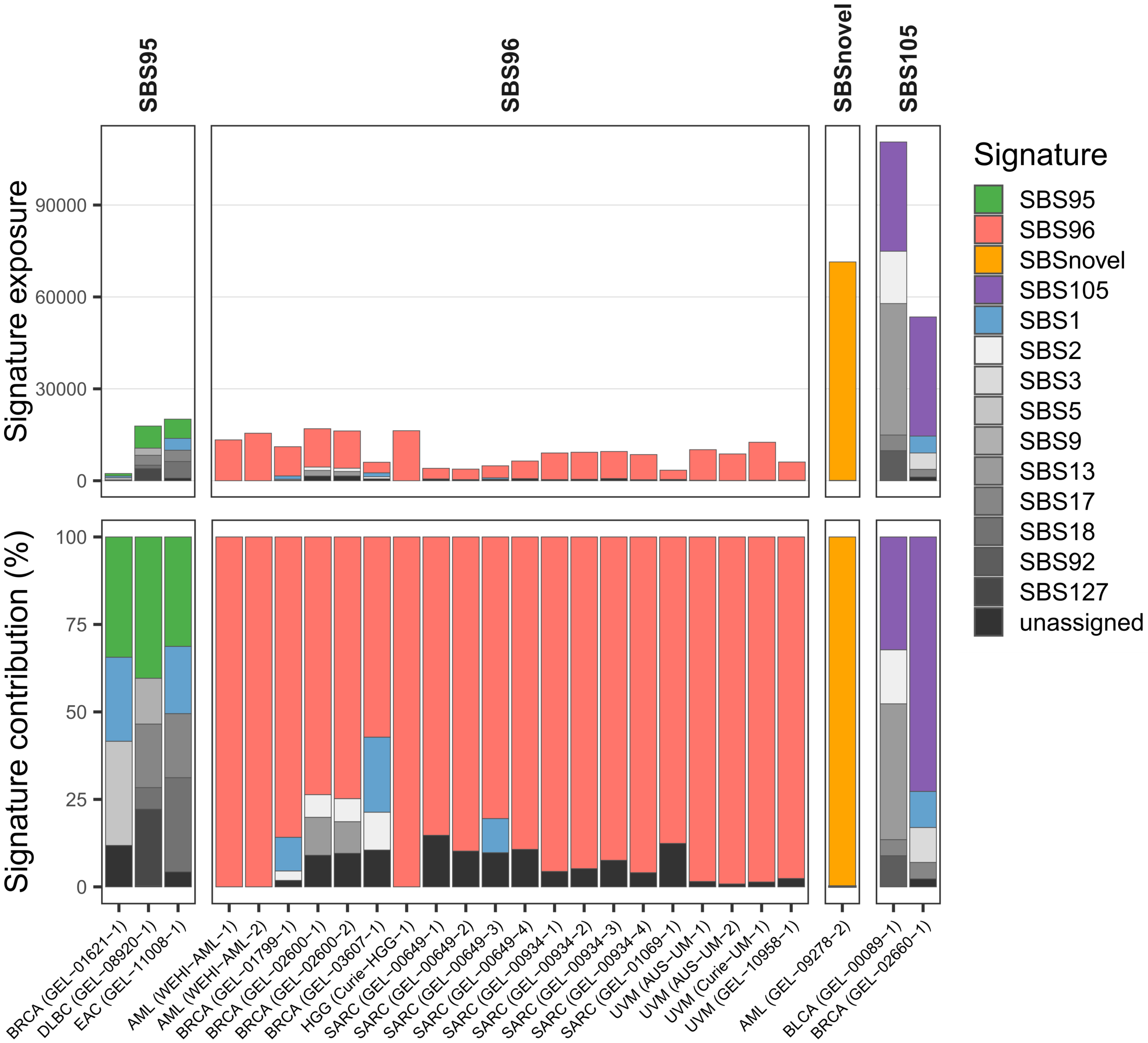
SUPPLEMENTARY INFORMATION

Base-excision repair pathway shapes 5-methylcytosine deamination signatures in pan-cancer genomes

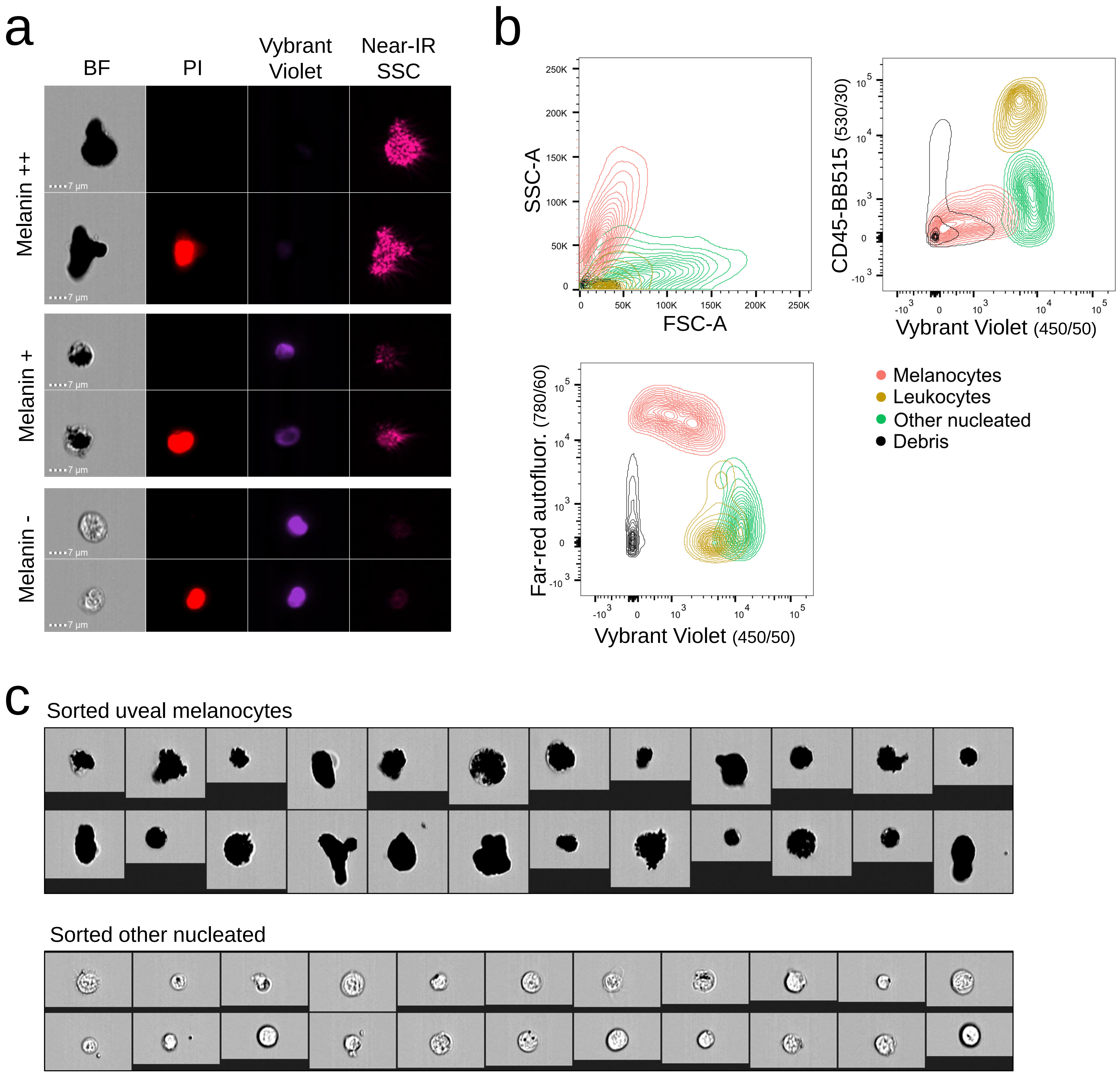
Silveira *et al*.

Supplementary Figure 1.



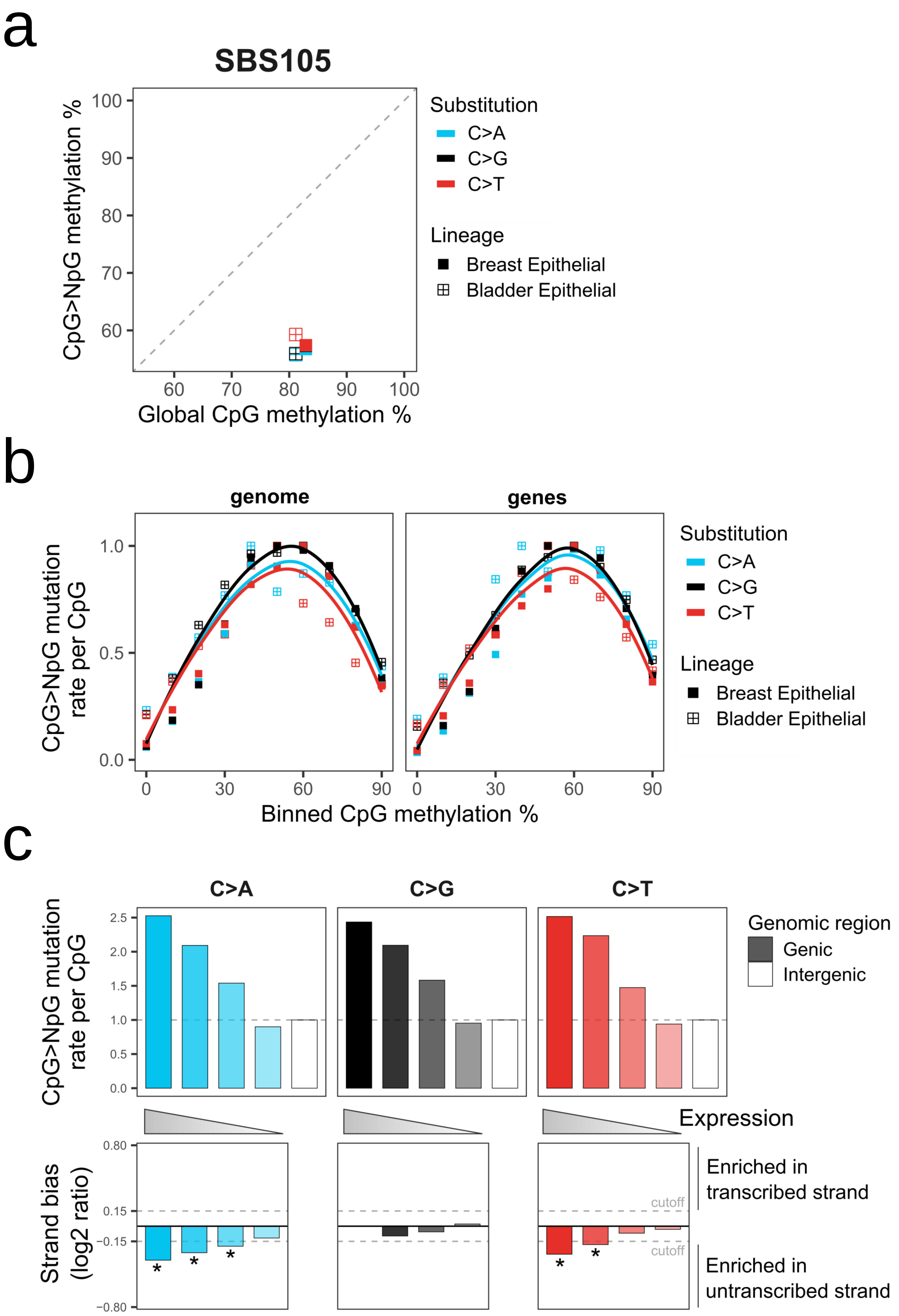
Supplementary Figure 1. Exposure to SBS mutational signatures. Number of substitutions assigned per signature (upper panel) or percent signature exposure contribution (lower panel) in tumor samples harboring rare CpG>NpG signatures. AML, acute myeloid leukemia; BRCA, breast invasive carcinoma; SARC, sarcoma; UVM, uveal melanoma; HHG, high-grade glioma; DLBC, diffuse large B-cell lymphoma; EAC, esophageal adenocarcinoma; BLCA, bladder urothelial carcinoma.

Supplementary Figure 2.



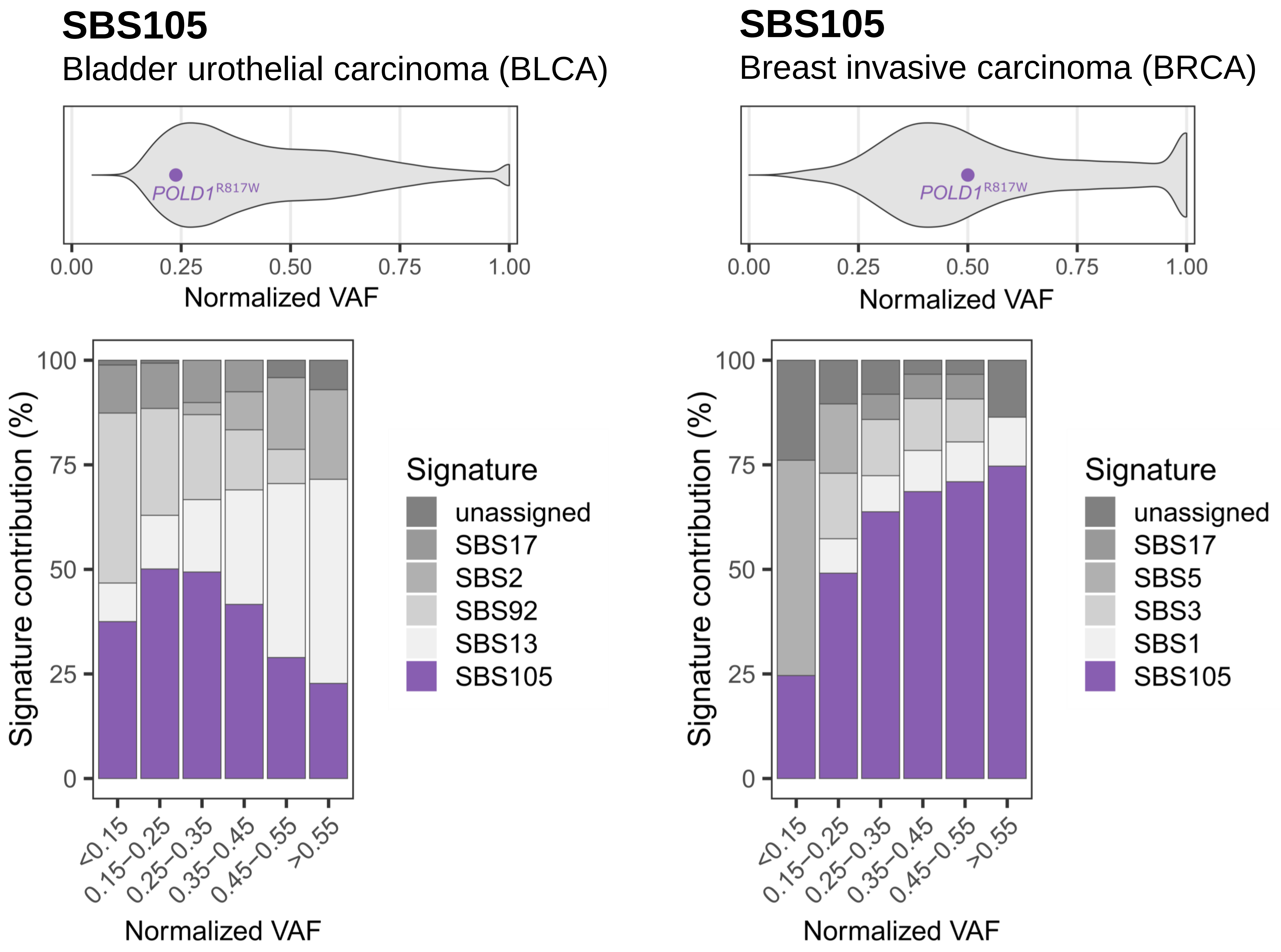
Supplementary Figure 2. Isolation of normal primary uveal melanocytes by melanin content. (a) Imaging flow cytometry of normal human uveal choroid cell suspension. Representative cells with varying levels of melanin, as identified by the brightfield channel (BF), are shown. Propidium iodide (PI) positive cells are shown to illustrate nuclear DNA staining of necrotic cells. Cell membrane-permeable Vybrant Violet nuclear DNA signal brightness is negatively correlated with melanin content due to UV-Violet wavelengths absorption by melanin. **(b)** Flow cytometry analysis of normal uveal choroid cell suspension. Populations were defined as: live leukocytes (PI$^-$/CD45$^+$/Vybrant Violet^{bright}); live melanocytes (PI$^-$/CD45$^-$/Vybrant Violet^{dim}/Far-red autofluorescence^{bright}); other live nucleated cells (PI$^-$/CD45$^-$/Vybrant Violet^{bright}/Far-red autofluorescence^{dim}); debris (Vybrant Violet^{dim}/Far-red autofluorescence^{dim}). **(c)** Imaging flow cytometry of sorted uveal melanocyte or other nucleated cells. Representative images of brightfield channel are shown. The median pixel intensity of the masked cell area was used to estimate the presence of melanin and calculate the purity of sorted melanocytes (>90%).

Supplementary Figure 3.



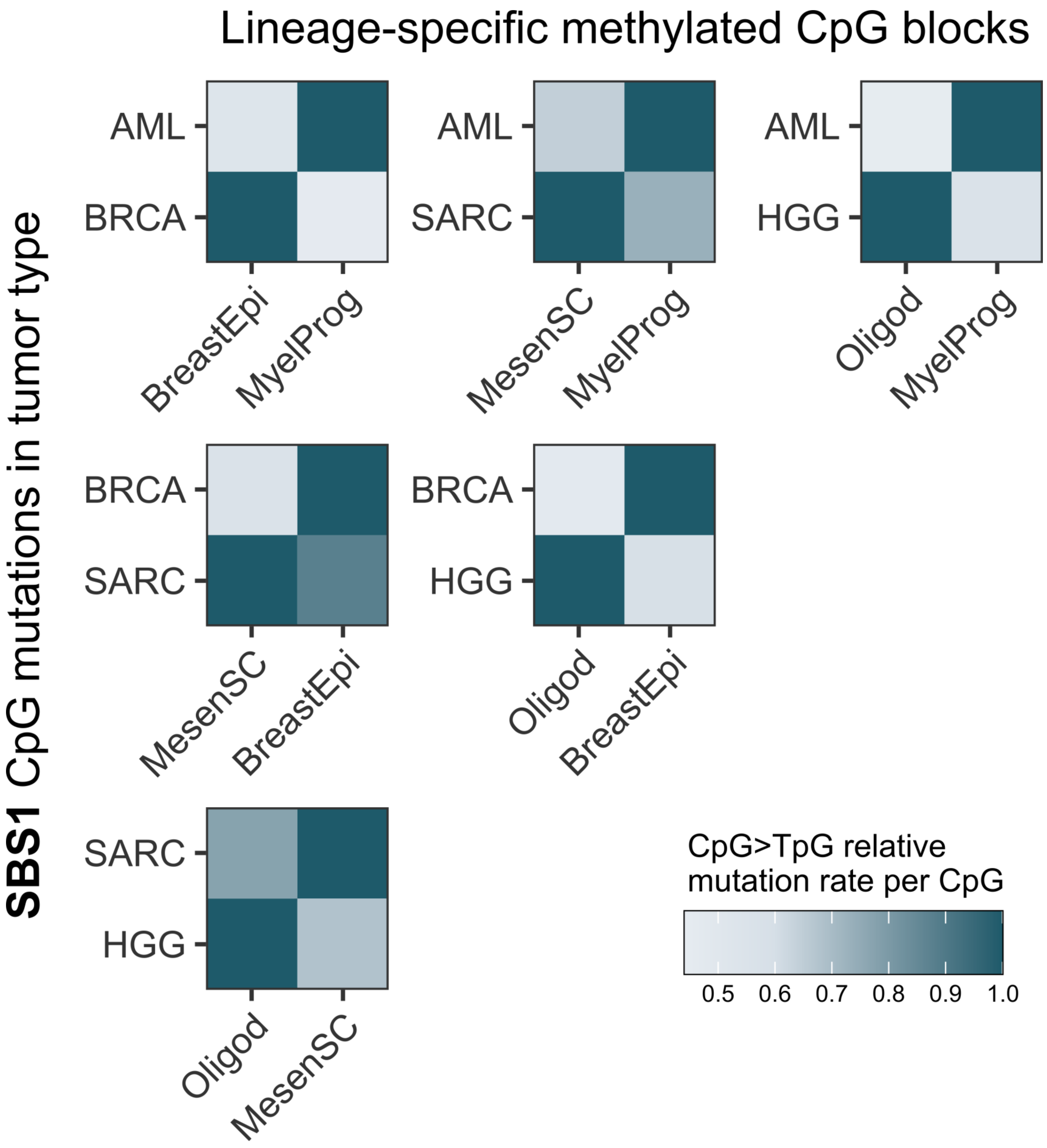
Supplementary Figure 3. Association of SBS105 CpG>NpG substitutions with DNA methylation and transcription strand asymmetry. (a) Scatter plot of DNA methylation percentages in SBS105 CpG>NpG mutated sites versus all CpGs (global), per cell lineage and substitution class. Methylation was interrogated in data from normal human cell types. The dashed line indicates the absence of over- or under-representation of methylation in mutated CpGs. **(b)** Scatter plots of SBS105 CpG>NpG mutation rates per CpG of different tumor types and signatures in 2 kb genomic windows grouped by their mean CpG methylation levels. Mutation rates were normalized by the highest value in each tumor type. The lines indicate data fitting with smoothed conditional means models. **(c)** Bar plots of SBS105 CpG>NpG mutation rates per CpG in genic regions grouped by their expression levels or intergenic regions (upper panel). Transcriptional strand asymmetry of SBS105 CpG>NpG mutations in genic regions grouped by their expression levels (lower panel). Asterisks mark a significant difference in contribution between transcribed and untranscribed strands (see Methods). The dashed line indicates the strand bias cutoff used to assign significance. Genes were grouped based on expression level quartiles.

Supplementary Figure 4.



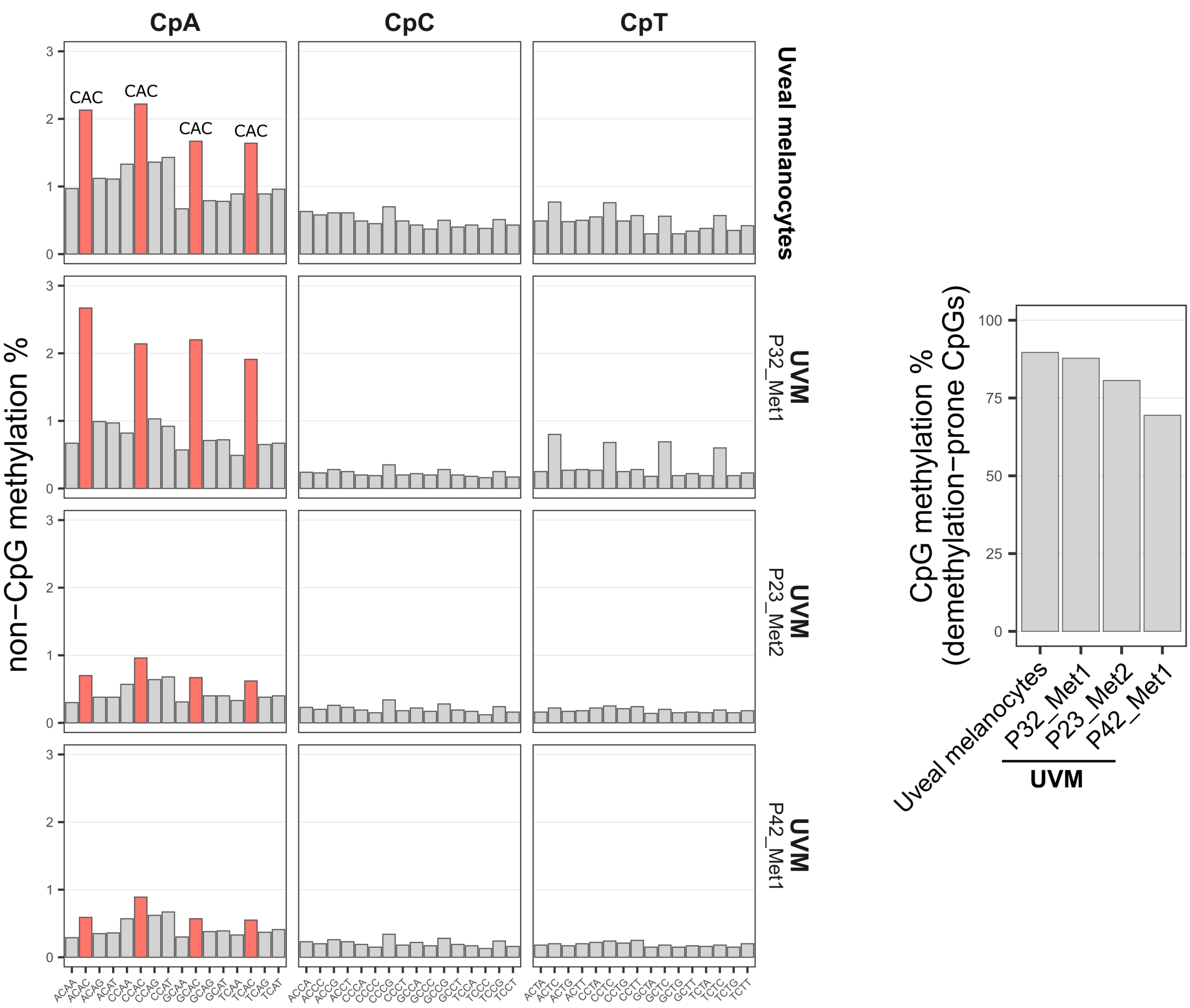
Supplementary Figure 4. Clonality of *POLD1*$^{\text{R817W}}$ mutation and SBS105 signature exposure. Distributions of normalized variant allele frequencies (VAF) of somatic variants in the two SBS105 tumors (upper panel). Percentage contributions of SBS mutational signatures among variants grouped by their normalized VAF (lower panel). In the BLCA sample, both *POLD1*$^{\text{R817W}}$ and SBS105 are subclonal.

Supplementary Figure 5.



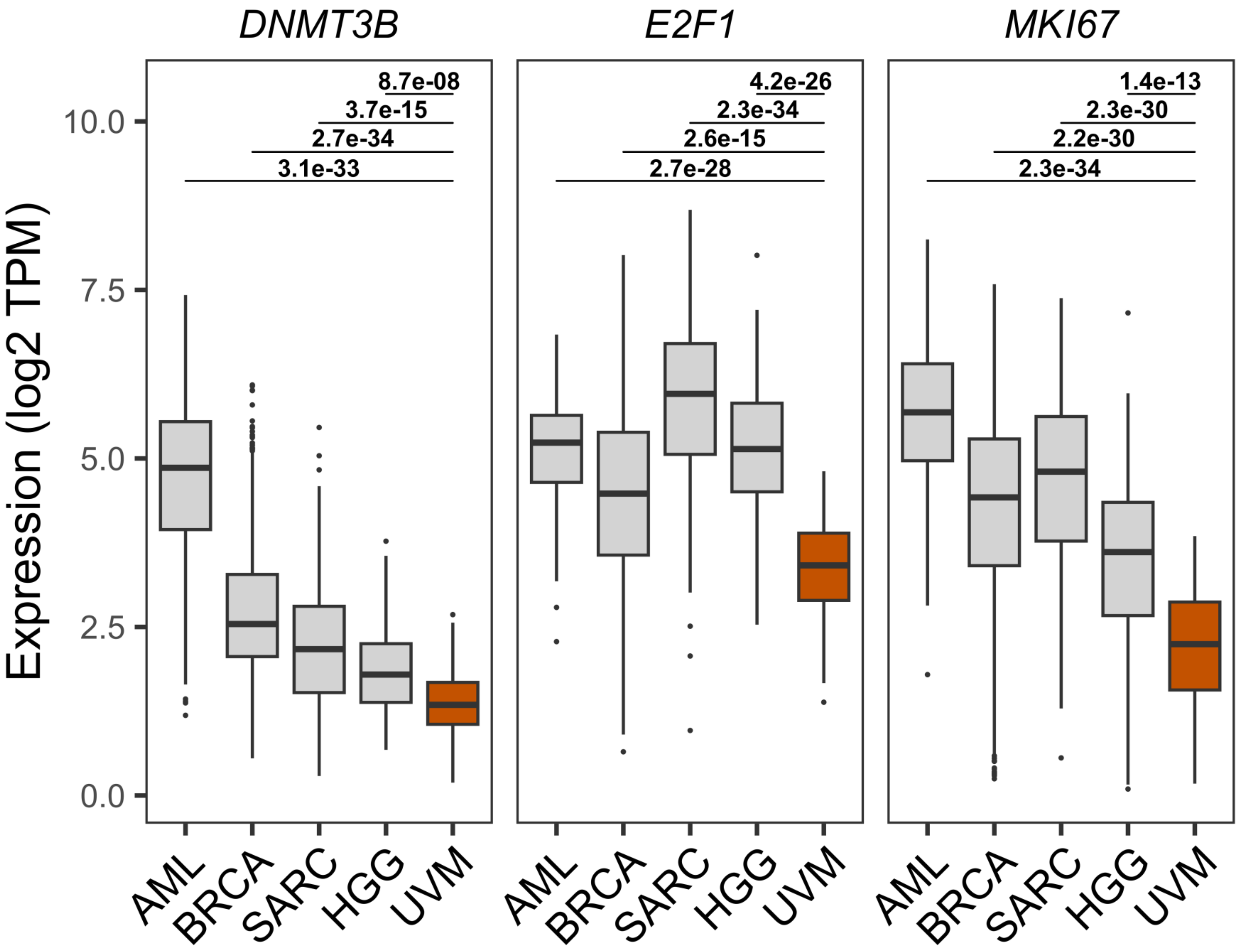
Supplementary Figure 5. SBS1 recapitulates lineage-specific CpG methylation landscapes. Heatmaps of SBS1 relative CpG>TpG mutation rates in CpG blocks differentially methylated between normal cell type pairs. The collections of blocks hypermethylated in each cell type are indicated in the x-axis. Tumor types are indicated in the y-axis. Values are normalized per tumor type. AML, acute myeloid leukemia; BRCA, breast invasive carcinoma; SARC, sarcoma; HGG, high-grade glioma; MyelProg, common myeloid progenitor; BreastEpi, breast luminal epithelium; MesenSC, mesenchymal stem cell; Oligod, oligodendrocyte.

Supplementary Figure 6.



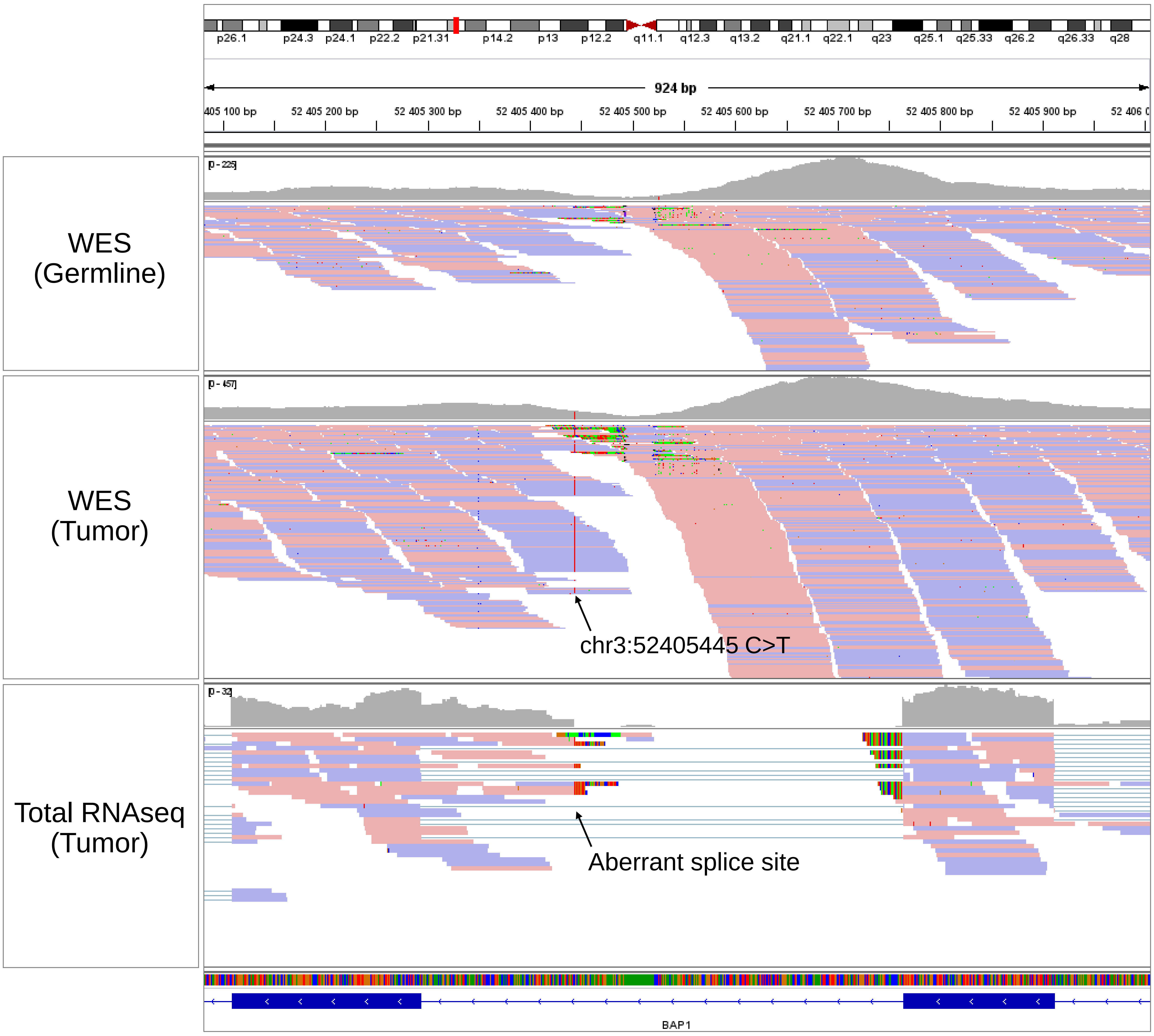
Supplementary Figure 6. Whole-genome DNA methylation landscapes in normal human uveal melanocytes and metastatic uveal melanoma (UVM) samples. Non-CpG methylation percentages by trinucleotide context, illustrating higher methylation levels in CAC contexts in the four samples (left panel). CpG methylation percentages among demethylation-prone CpGs due to cell division (right panel). Samples with lower CAC methylation (P23_Met2 and P42_Met1) show evidence of CpG methylation degradation due to cell division.

Supplementary Figure 7.



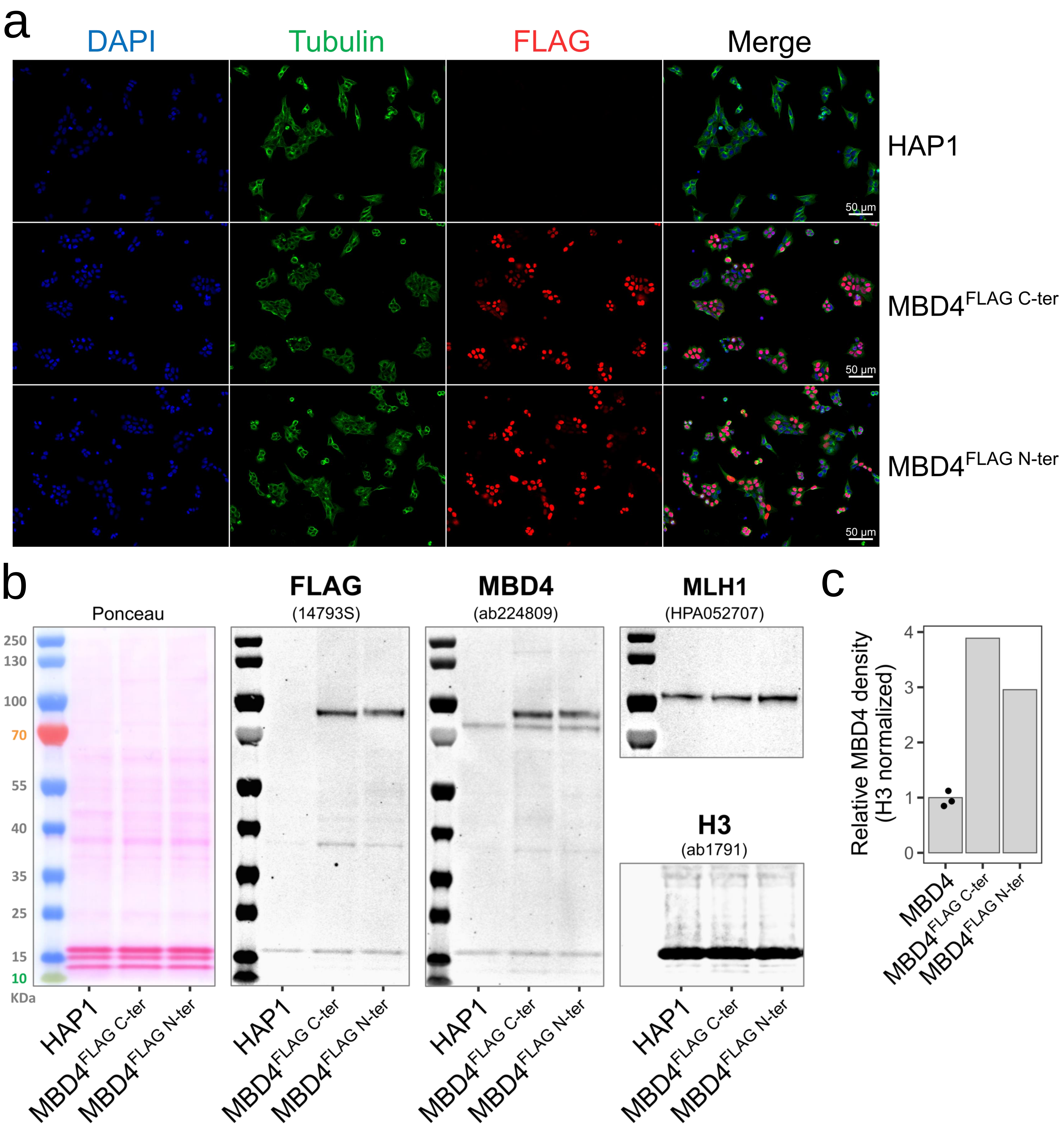
Supplementary Figure 7. Distributions of gene expression in TCGA tumors grouped by tumor type. Distributions of gene expression in TCGA tumors, including AML (n=151), BRCA (n=1,231), SARC (n=265), HGG (n=175) and UVM (n=80). Values are expressed as transcripts per million (TPM). The expression of *E2F1* and *MKI67* represent markers of cell proliferation. Two-sided Wilcoxon test *P*-values without multiple comparisons adjustment are indicated. Boxes indicate the median, 25th and 75th percentiles. Whiskers extend to the largest or lowest value up to 1.5 times the distance between the 25th and 75th percentiles.

Supplementary Figure 8.



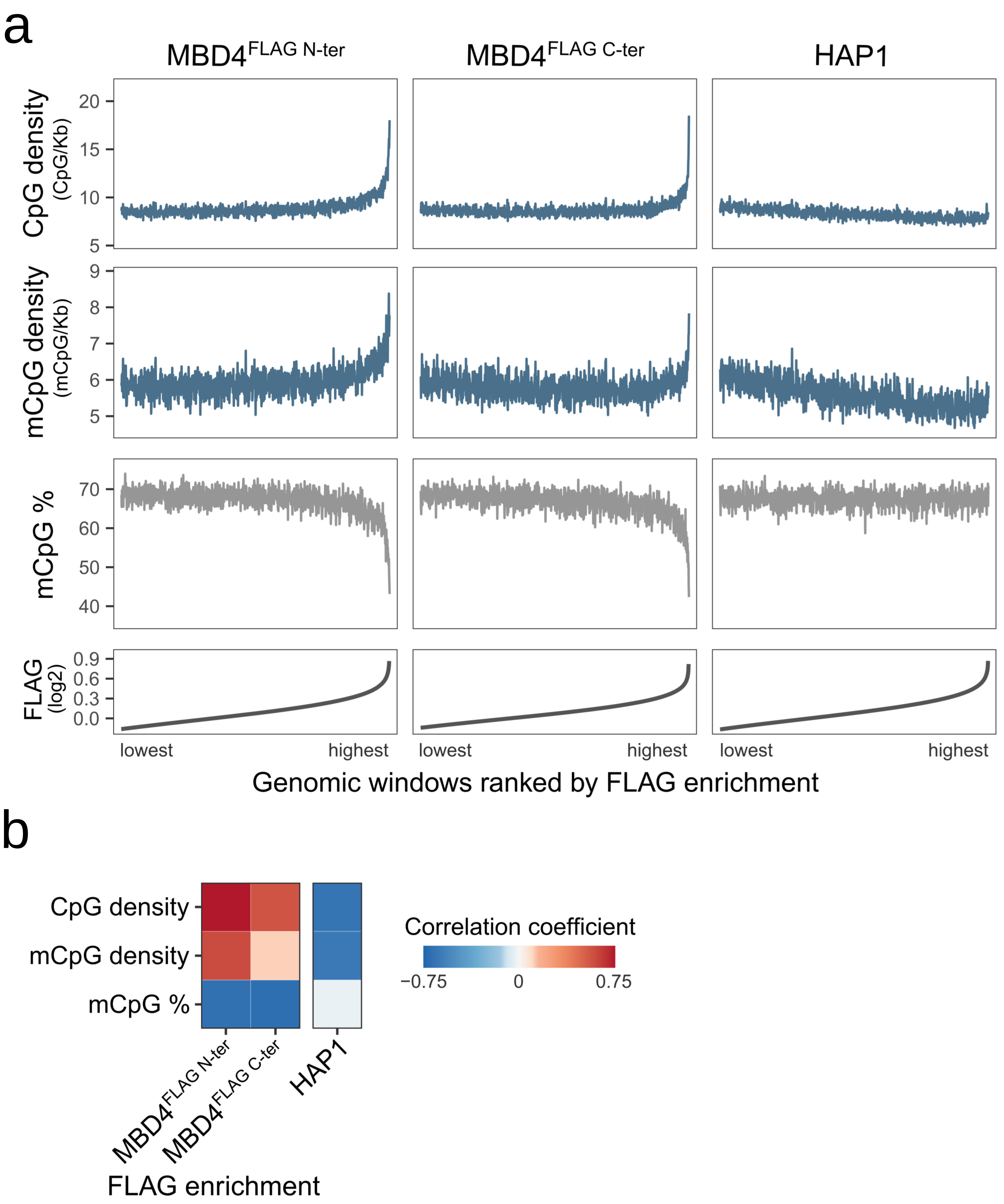
Supplementary Figure 8. Aberrant splicing of *BAP1* caused by a hotspot intronic CpG>TpG mutation. IGV visualization of BAM files mapped to GRCh38 genome assembly of a representative *MBD4*def UVM tumor harboring a somatic hotspot *BAP1* intronic CpG>TpG mutation (chr3:52405445), as shown in upper panels of germline versus tumor whole exome sequencing (WES). The aberrant splice site is indicated in tumor total RNA sequencing (RNAseq), as shown in the lower panel. Read strand information is coded in purple and pink colors.

Supplementary Figure 9.



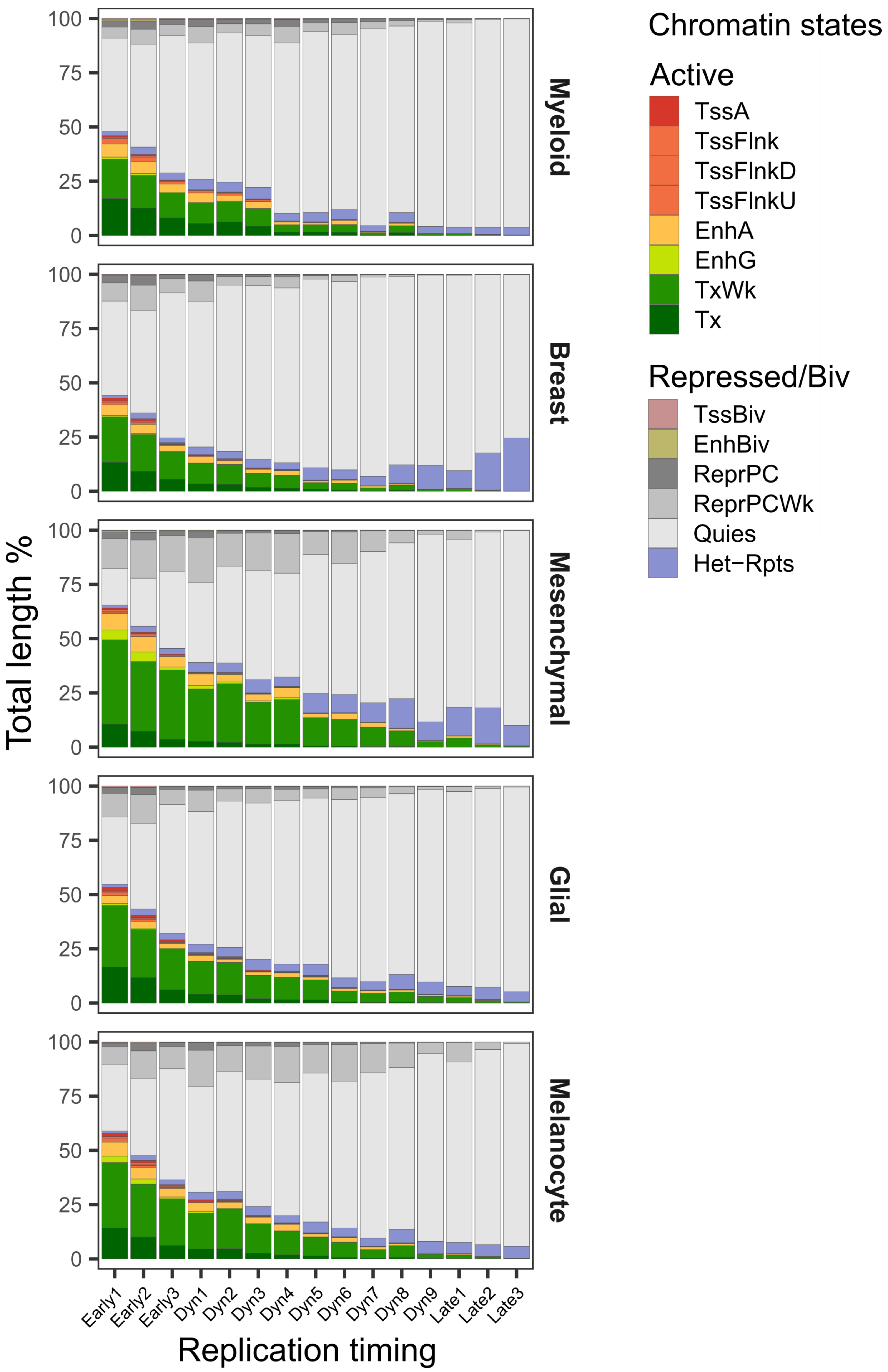
Supplementary Figure 9. Characterization of exogenously expressed FLAG-tagged MBD4 in HAP1 cells. (a) Immunofluorescence microscopy of HAP1 cells either parental or stably overexpressing C- or N-terminally FLAG-tagged MBD4. Cells were stained with anti-tubulin (green) and anti-FLAG (red) antibodies and co-stained with DAPI (blue). Objective magnification of 20x. Scale bars are shown in white. **(b)** Western blotting on nuclear extracts of HAP1 single-cell clones overexpressing C- and N-terminally FLAG-tagged MBD4, and parental HAP1 cells. Total protein stain with ponceau S is shown on the left panel. **(c)** Densitometry of western blotting obtained with anti-MBD4 antibody. Values were normalized by the density of H3 bands in each cell line. Values shown are relative to the mean normalized density of endogenous MBD4 bands. Bars represent the mean (n=1 for FLAG-tagged MBD4 bands; n=3 for endogenous MBD4 bands, representing one band from each of the three cell lines). No replication attempts or biological replicates were performed.

Supplementary Figure 10.



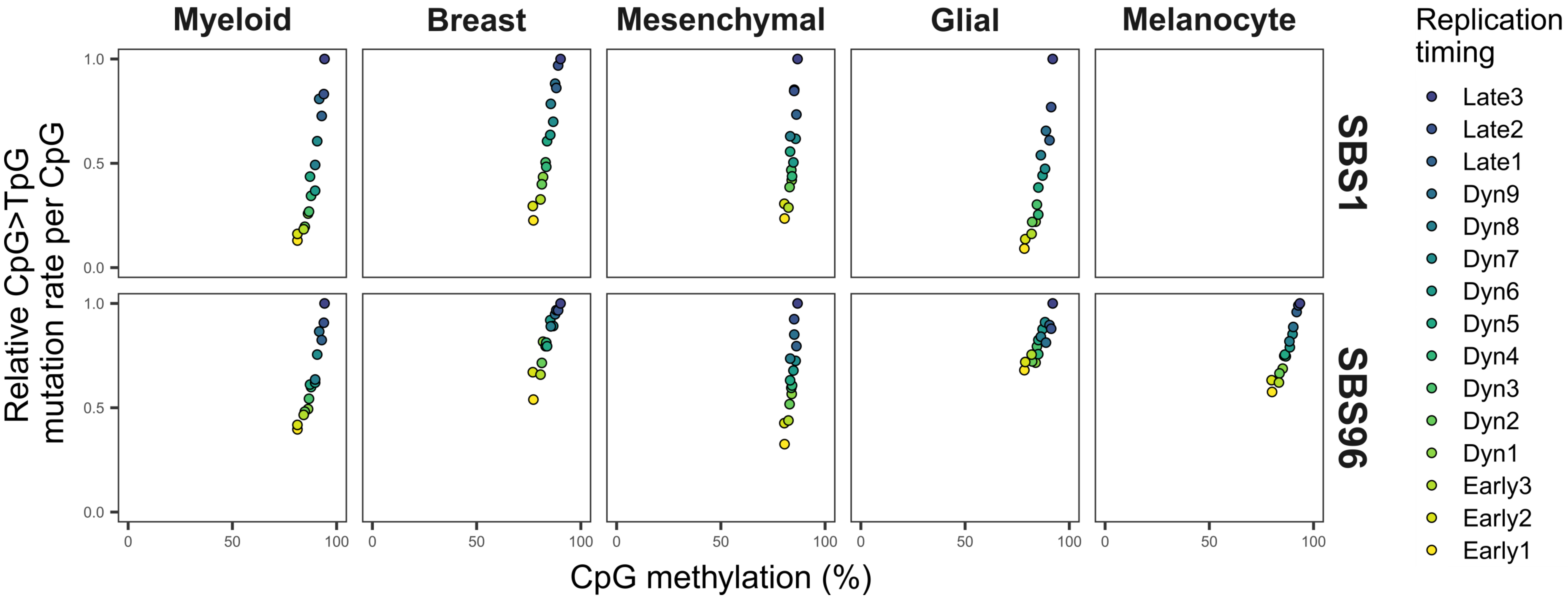
Supplementary Figure 10. Tagged MBD4 signal enrichment correlation with CpG and methylated CpG densities. (a) CpG and methylated mCpG (mCpG) densities, mCpG percentage, and FLAG enrichment over IgG in 4 kb non-overlapping genomic windows, ranked from lowest to highest FLAG signal enrichment obtained on cells overexpressing N- or C-terminally FLAG-tagged MBD4, or on parental HAP1 cells (negative control cell line). CpG methylation was derived from whole-genome DNA methylation data on KBM7 cells. The mean values of every 400 similarly-ranked windows are shown. **(b)** Heatmap of Pearson correlation coefficients between FLAG signal enrichment and CpG density, mCpG density, or mCpG percentage. Coefficients were calculated using the data points shown in the upper panel.

Supplementary Figure 11.



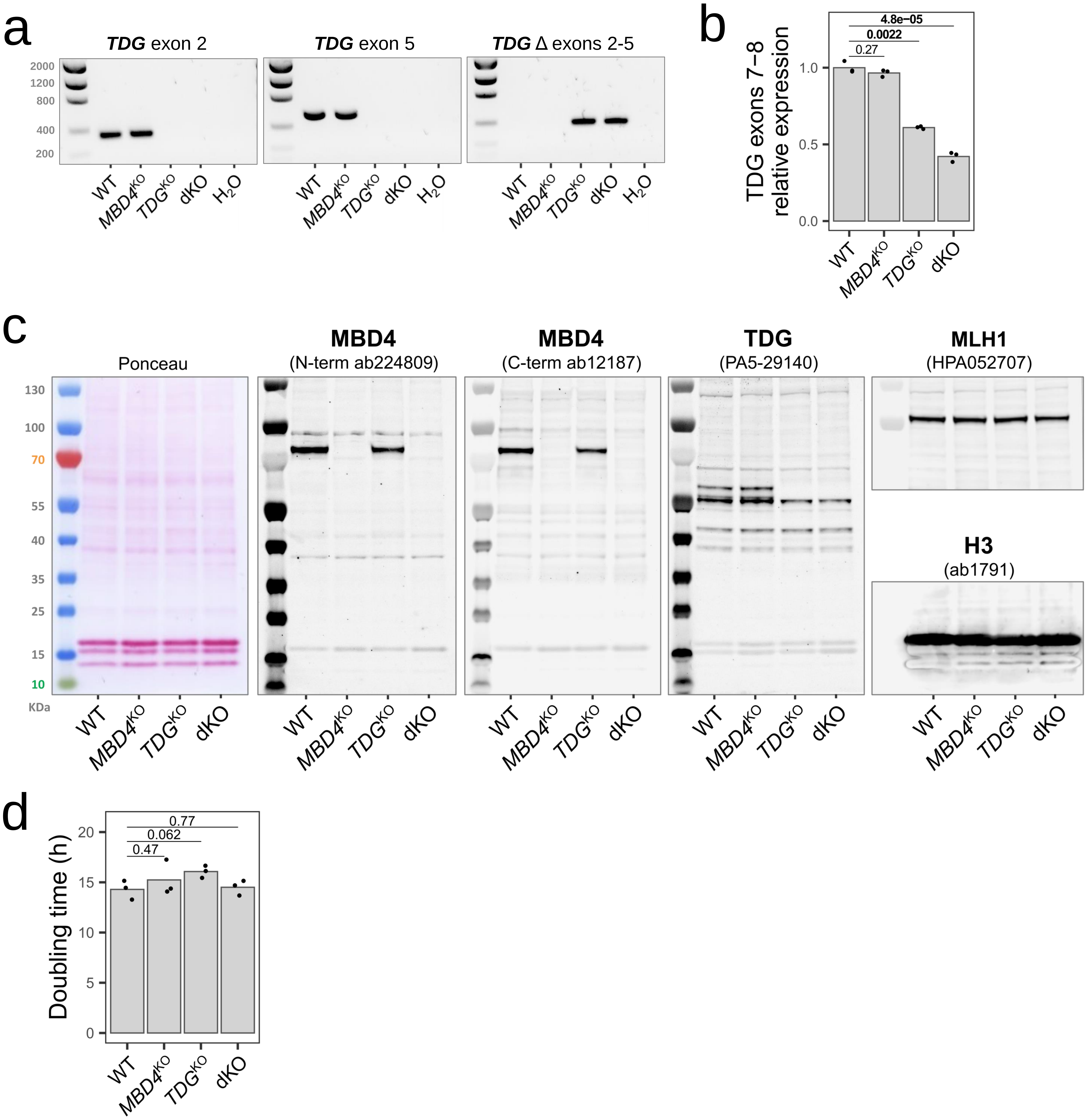
Supplementary Figure 11. Active chromatin is associated with early replicating genomic regions. Barplots of lineage-specific chromatin states, as length percentage contribution among genomic regions annotated by replication timing. Early, Dyn, and Late indicate constitutive early, dynamic, or constitutive late replication timing regions, respectively. TssA, active TSS; TssFlnk, flanking TSS; TssFlnkD, flanking TSS downstream; TssFlnkU, flanking TSS upstream; EnhA, active enhancer; EnhG, genic enhancer; TxWk, weak transcription; Tx, strong transcription; TssBiv, bivalent/poised TSS; EnhBiv, bivalent enhancer; ReprPC, repressed polycomb; ReprPCWk, weak repressed polycomb; Quies, quiescent/low; Het-Rpts, heterochromatin/ZNF genes and repeats.

Supplementary Figure 12.



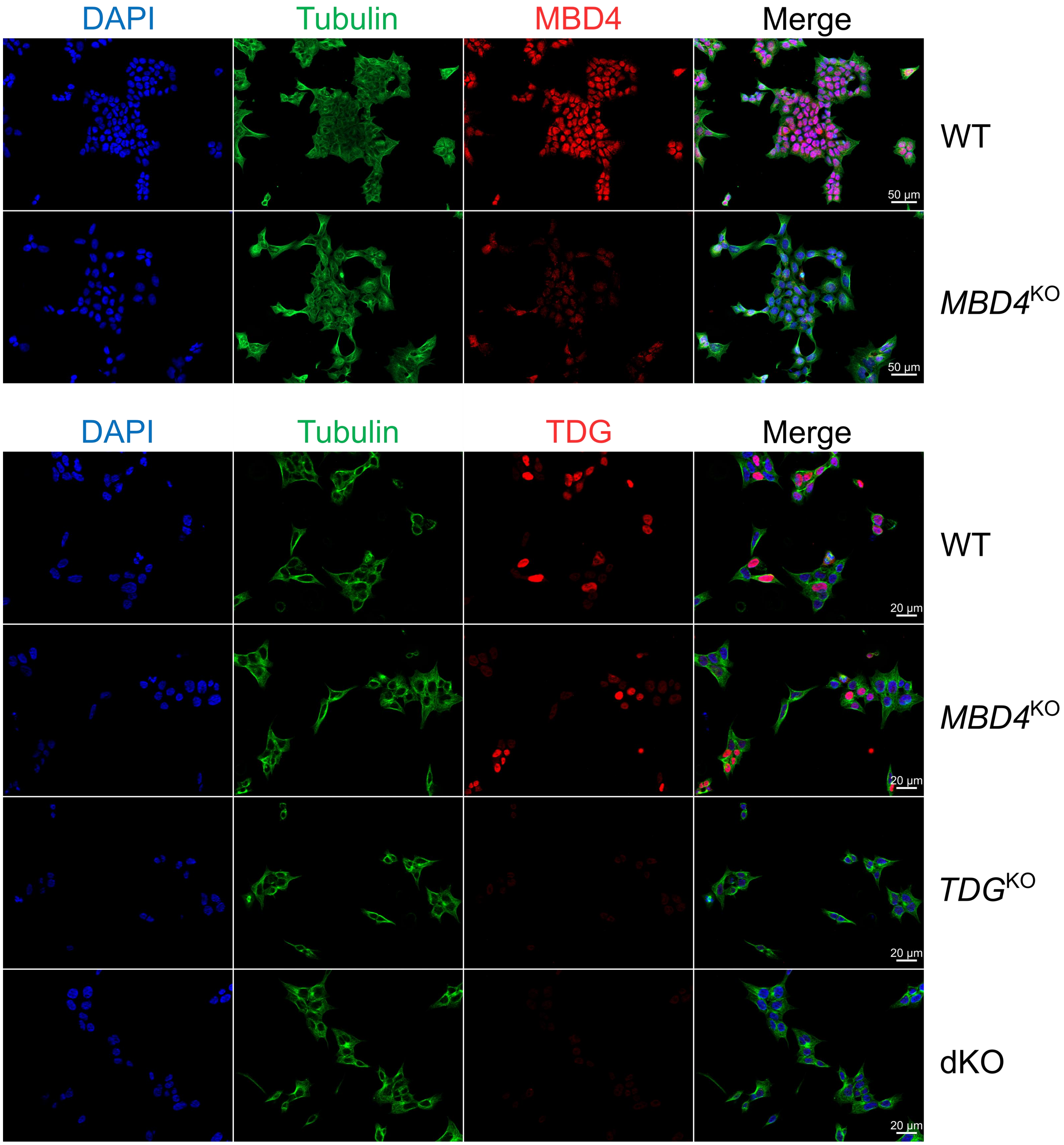
Supplementary Figure 12. Early replicating genomic regions show lower SBS1 and SBS96 mutation rates. Scatter plots of CpG>TpG mutation rates per CpG in pan-tissue replication timing annotations versus mean CpG methylation levels. Tumor mutations and normal epigenomic data are grouped by cell lineage. Mutation rates were normalized by the highest value in each tumor, and the mean of all tumors per lineage is shown. Early, Dyn, and Late indicate constitutive early, dynamic, or constitutive late replication timing regions, respectively.

Supplementary Figure 13.



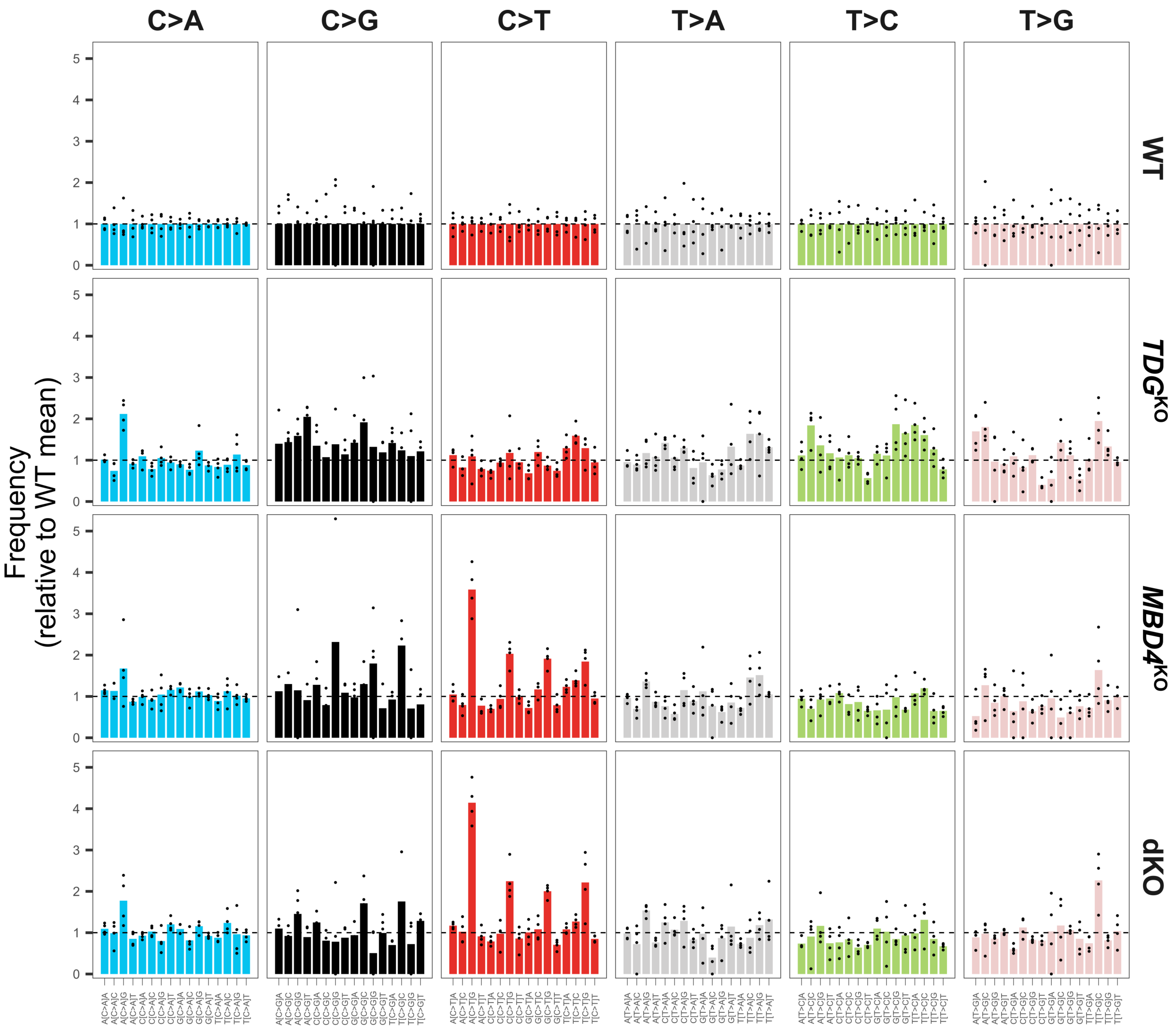
Supplementary Figure 13. Validation of *MBD4* and *TDG* knockout in isogenic HAP1 cell models. (a) Genomic DNA PCR in HAP1 cell models wild-type or knock-out for *MBD4*, *TDG*, or both (dKO). Primers used spanned *TDG* exon 2, exon 5, or the region spanning exons 2-5. **(b)** *TDG* relative expression levels using primers spanning *TDG* exons 7-8, obtained by real-time quantitative reverse transcription PCR with *GAPDH* as endogenous control. Bars represent means of technical replicates (n=3 per genotype, shown as dots). Two-sided unpaired *t*-test *P*-values without multiple comparisons adjustment are indicated. **(c)** Western blotting on nuclear extracts of HAP1 cell models. Antibodies against either the N- or C-terminus of MBD4 are indicated. Total protein stain with ponceau S is shown on the left panel. No replication attempt was performed. **(d)** Doubling times of HAP1 single-cell clones of each genotype used for long-term culturing. Cells that had been cultured for 60 days (D60) were used. Bars represent the means of biological replicates (n=3 per genotype, shown as dots). Two-sided unpaired *t*-test *P*-values without multiple comparisons adjustment are indicated.

Supplementary Figure 14.



Supplementary Figure 14. Validation of *MBD4* and *TDG* knockout in isogenic HAP1 cell models. Immunofluorescence microscopy of HAP1 cell models wild-type or knock-out for *MBD4*, *TDG*, or both (dKO). Cells were stained with anti-tubulin (green), anti-MBD4 (red), or anti-TDG (red) antibodies, and co-stained with DAPI (blue). Objective magnification of 20x. Scale bars are shown in white.

Supplementary Figure 15.



Supplementary Figure 15. Substitution profiles by trinucleotide context of *MBD4* and *TDG* knockout HAP1 cell models. Distributions of substitution frequencies obtained by WGS in HAP1 subclones wild-type or knock-out for *MBD4*, *TDG*, or both (dKO) after 120 days in culture. Values shown are relative to the mean of wild-type subclone, shown as a dashed line. Bars represent the mean of subclones (n=4 per genotype, shown as dots).