Supplemental Material

Supplemental Table S1

Platform	Group	Number of samples	Number of pooled molecules	Number of short (<=500) mole cules	Number of long (>500) molecules	Proportion of long molecules (% mean±s.d.)	Data source
SMRT	First-trimester	7	7,991,123	6,332,734	1,658,389	19.59±7.8	(Yu et al. 2021)
	Second-trimester	10	18,111,964	14,014,248	4,097,716	24.05±12.7	(Yu et al. 2021)
	Third-trimester	11	19,438,356	13,783,994	5,654,362	31.56±11.3	(Yu et al. 2021)
	Non-pregnant controls	15	16,637,319	11,361,590	5,275,729	27.55±12.8	(Choy et al. 2022)
ONT	First-trimester	7	39,325,009	37,694,202	1,630,807	3.88±1.5	(Yu et al. 2023)
	Second-trimester	8	32,852,897	31,519,876	1,333,021	3.33±1.7	(Yu et al. 2023)
	Third-trimester	16	102,291,511	97,258,581	5,032,930	5.64±2.4	(Yu et al. 2023)
	Non-pregnant controls	5	41,960,061	39,795,239	2,164,822	5.01±0.9	New samples

Supplemental Table S2

	Number of Samples	Number of SMRT circular consensus sequencing				Data source
		Minimum	Median	Mean	Maximum	
Healthy subjects	20	104,678	1,021,412	1,129,888	2,883,101	5 new samples and 15 samples from (Choy et al. 2022)
HBV carriers	19	112,498	286,492	428,050	1,086,432	6 new samples and 13 samples from (Choy et al. 2022)
HCC patients	48	110,550	712,223	934,053	2,752,151	35 new samples and 13 samples from (Choy et al. 2022)

Supplemental Table S3

	Number of Samples	Total number of Pooled molecules	Molecule size (bp) summary			
			Minimum	Median	Mean	Maximum
Wild-type	4	7,691,684	45	181	423	26,276
Dffb$^{-/-}$	5	7,553,210	46	176	344	29,844
Dnase1$^{-/-}$	5	9,360,779	46	196	513	27,109
Dnase1l3$^{-/-}$	5	5,031,720	45	317	712	29,949
Dnase1$^{-/-}$ *Dnase1l3*$^{-/-}$	5	5,315,775	45	377	932	49,722

Supplemental Figure S1

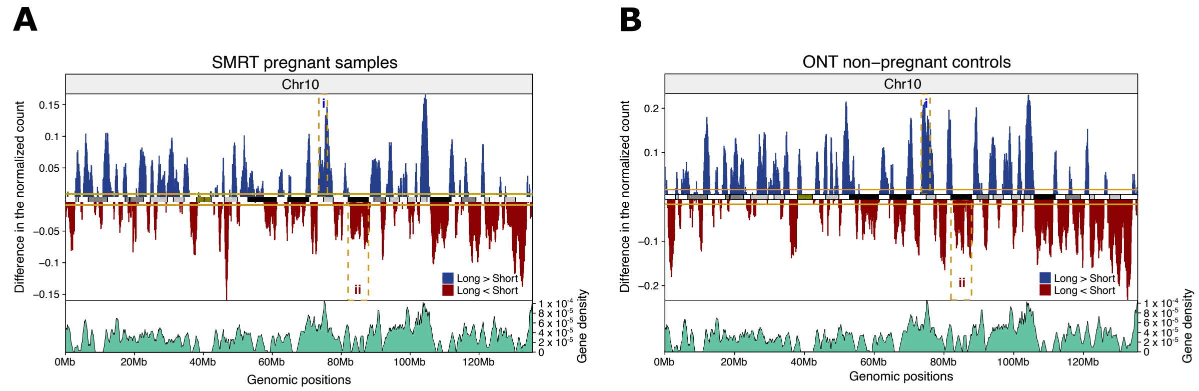

Fig. S1. (A) Comparison of genomic representation on Chromosome 10 between long and short DNA molecules in 28 pregnant samples using SMRT sequencing. Overrepresentation and underrepresentation of long cfDNA molecules with respect to short molecules are indicated in blue and red, respectively. The genomic representation was determined based on 100-kb bins, and was further smoothed by a 1-Mb moving averages sliding window. The horizontal solid lines indicate normalized median differences between long and short molecules. The dashed rectangular boxes indicate one euchromatic (i) and one heterochromatic (ii) region. The track in between overrepresentation and underrepresentation of long molecules shows the chromosome ideogram. The ideogram band colors correspond to cytogenetic bands in UCSC Genome Browser. Darker bands are AT-rich and lighter bands are GC-rich. Centromeric regions are indicated in dark green. The bottom track displays gene densities estimated by number of genes in 100-kb windows. (B) Comparison of genomic representation on Chromosome 10 between long and short DNA molecules in 5 non-pregnant controls using ONT sequencing.

Supplemental Figure S2

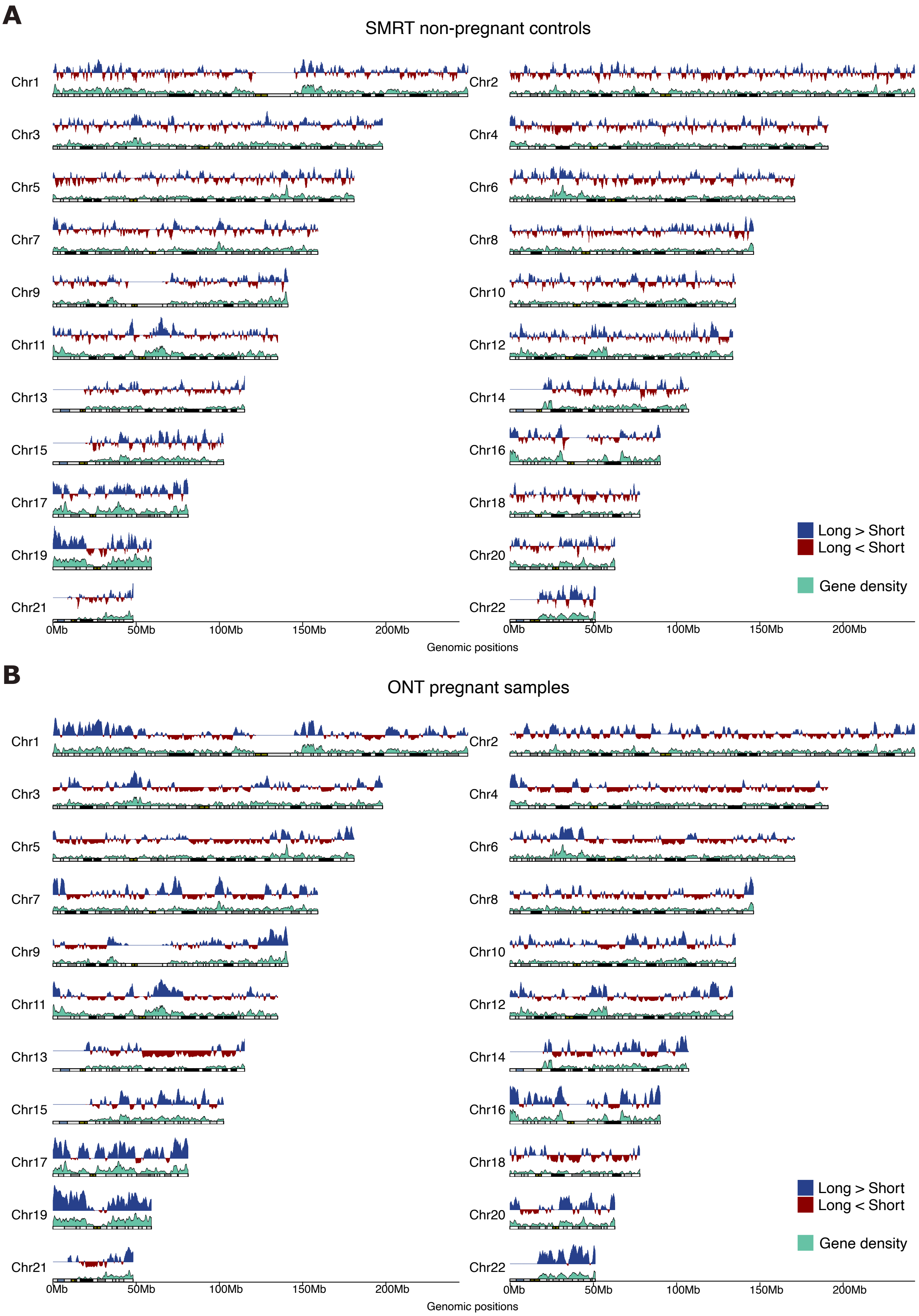

Fig. S2. (A, B) Genome-wide comparison of genomic representation between long and short DNA molecules in 15 non-pregnant controls using SMRT sequencing (A) and in 31 pregnant samples using ONT sequencing (B). The top track shows genomic representation difference between long and short molecules. The second track shows gene densities. The bottom track shows chromosome ideogram. The ideogram band colors correspond to cytogenetic bands in UCSC Genome Browser. Darker bands are AT-rich and lighter bands are GC-rich. Centromeric regions are indicated in dark green.

Supplemental Figure S3

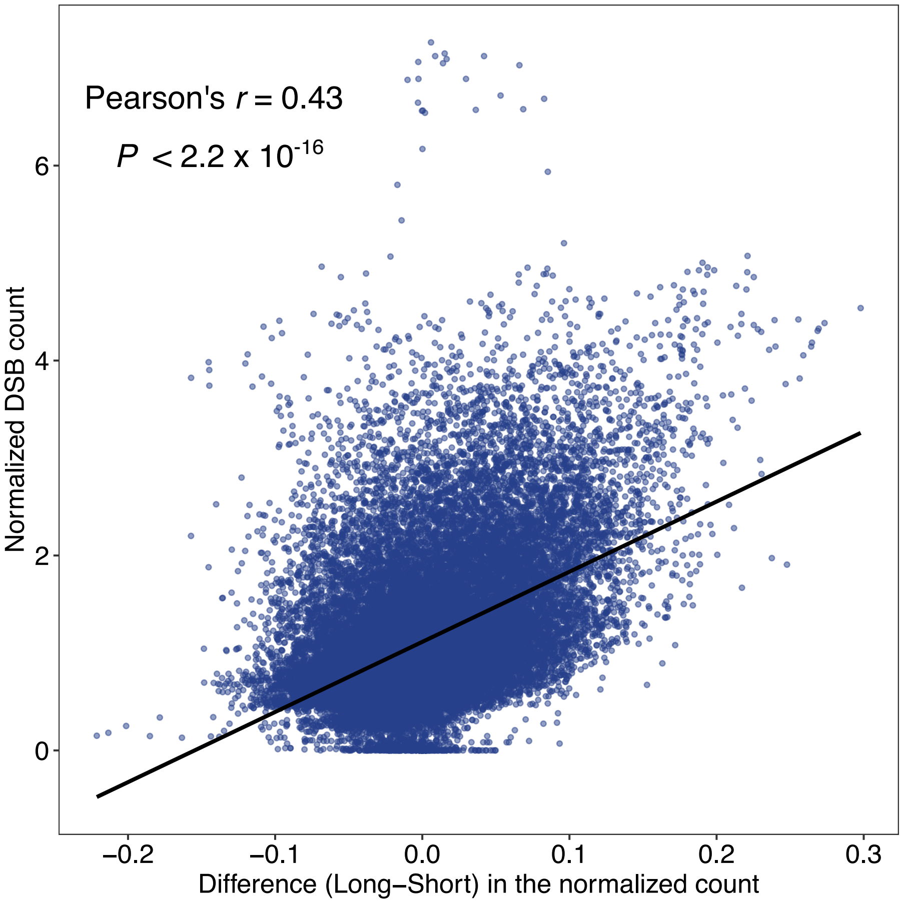

Fig. S3. The correlation between the genomic representation difference and double strand DNA breaks. Each dot represents one 100-kb bin. The X-axis indicates genomic representation differences between long and short molecules of 15 non-pregnant controls from SMRT sequencing. Positive values represent overrepresentation of long molecules and negative values represent underrepresentation of long molecules. The Y-axis indicates normalized DNA double-strand breaks count from a lymphoblastoid cell line sample.

Supplemental Figure S4

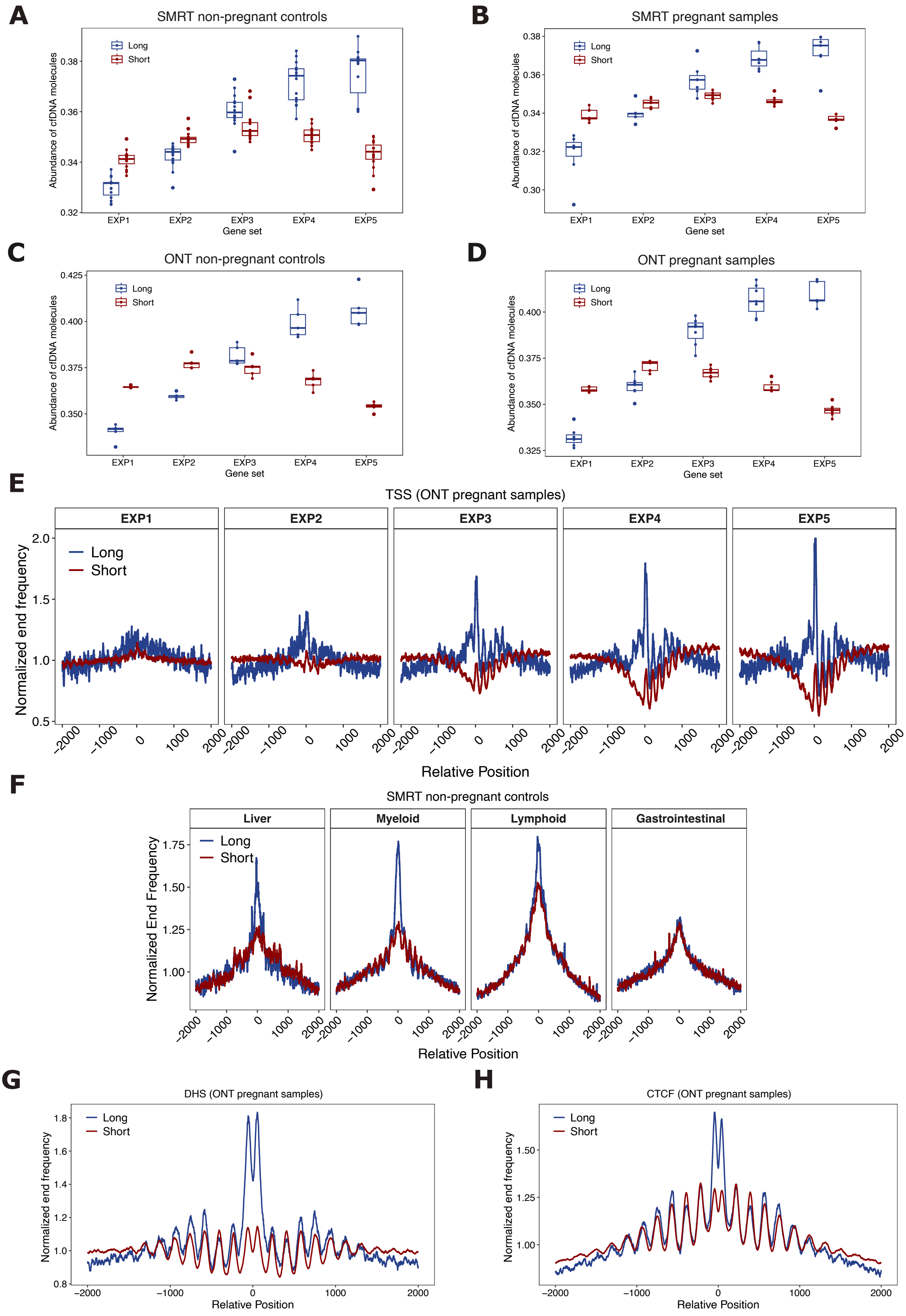

Fig. S4. (A-D) The abundance of long and short molecules on gene bodies of expression-stratified gene groups for non-pregnant controls and first-trimester individuals. (A, B) Non-pregnant controls and first-trimester subjects from SMRT sequencing. (C, D) Non-pregnant controls and first-trimester subjects from ONT sequencing. (E) Normalized end frequencies of ONT long and short molecules from plasma of pregnant individuals at TSSs of expression-stratified gene groups EXP1 to EXP5, corresponding to low to high expression. Transcription start positions are denoted as position 0. All transcription start sites were strand-adjusted so that positive positions are in the direction of transcription. (F) Normalized end frequencies of SMRT long and short molecules from plasma of non-pregnant controls at tissue-specific DHSs. (G, H) Normalized end frequencies of ONT long and short molecules from plasma of pregnant individuals at DHSs (G) and CTCF binding sites (H). DHSs or CTCF binding sites peaks are denoted as position 0; downstream and upstream 2000 bp are shown.

Supplemental Figure S5

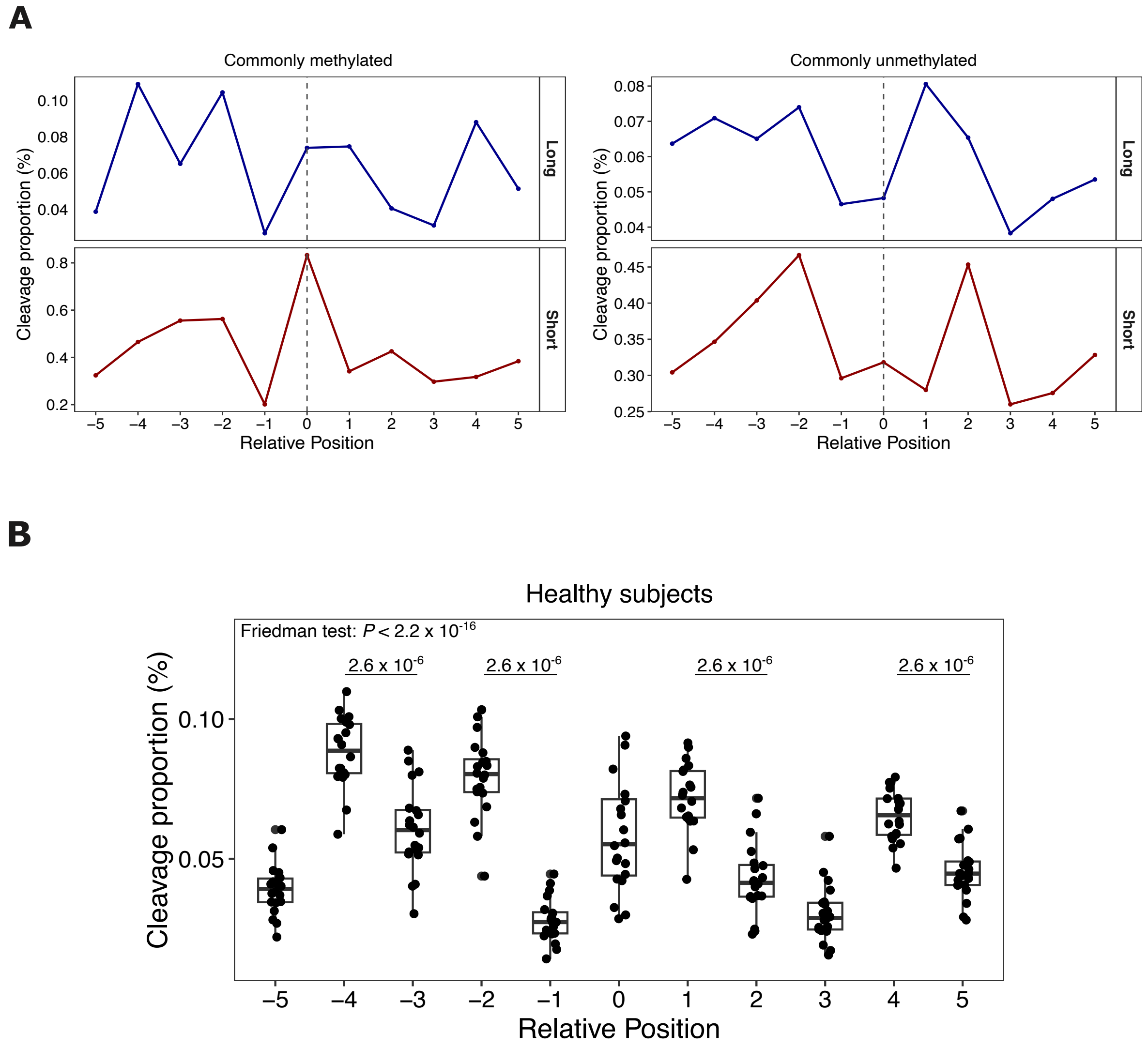

Fig. S5. (A) Cleavage profiles of healthy subjects from SMRT sequencing at commonly methylated (left panel) and commonly unmethylated (right panel) CpGs. A cleavage window of 11 bases is shown. Position 0 and 1 indicate cytosine and guanine, respectively. (B) Cleavage profiles surrounding all autosomal CpGs for long molecules from healthy subjects. Each point represents one sample, and each boxplot represents cleavage proportions from healthy subjects at a specific position. Differences between the cleavage proportions of the 11 positions were assessed by Friedman test. Post hoc paired Wilcoxon signed-rank tests with Benjamini-Hochberg multiple correction were used to compare preferred cleavage positions with respective adjacent downstream positions. Pairwise *P* values are shown above horizontal lines.

Supplemental Figure S6

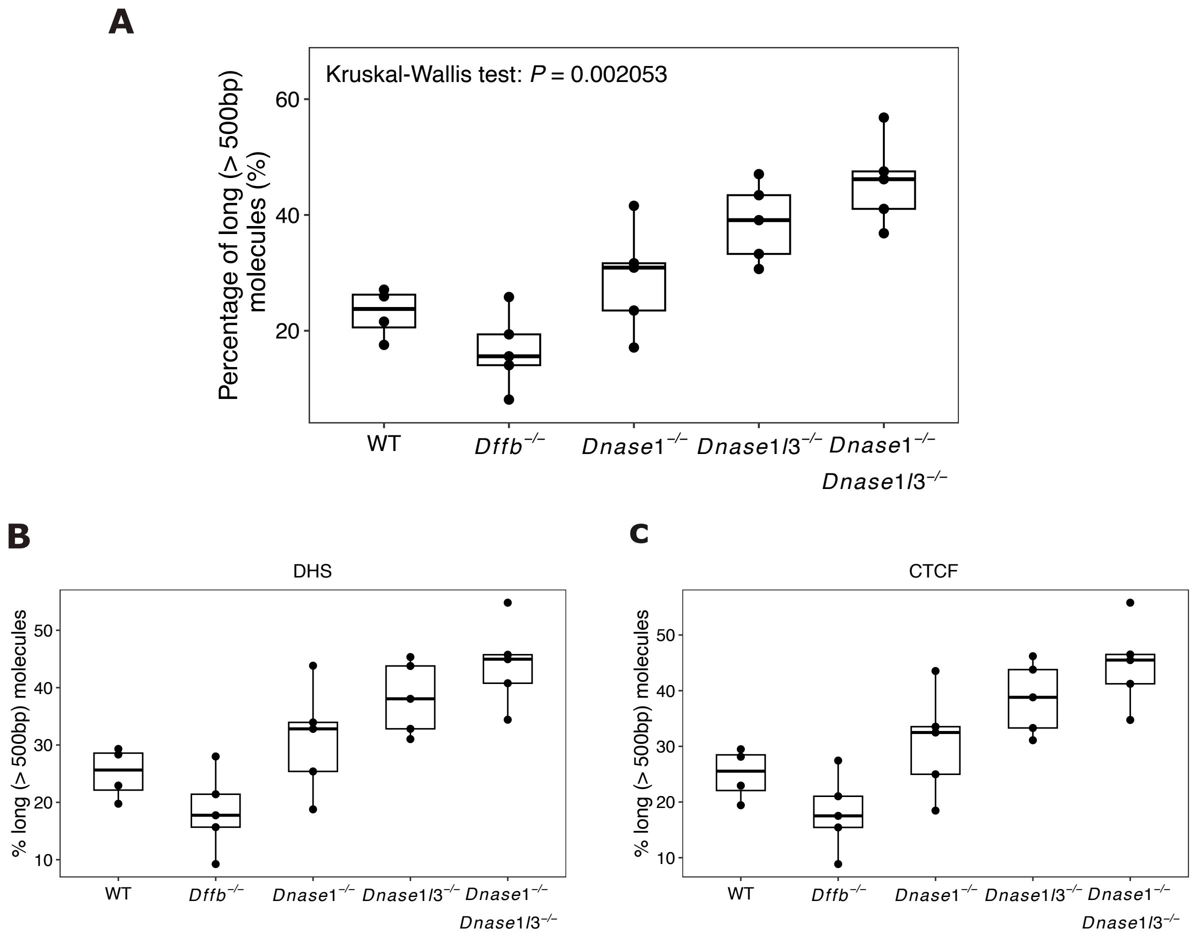

Fig. S6. (A) Boxplot showing the proportions of long molecules in each type of mice. The proportions were compared between different groups of mice using Kruskal-Wallis test. (B, C) Boxplot showing the proportions of long molecules originated from 2000 bp upstream and downstream of DHSs (B) and CTCF binding sites (C).

Supplemental Figure S7

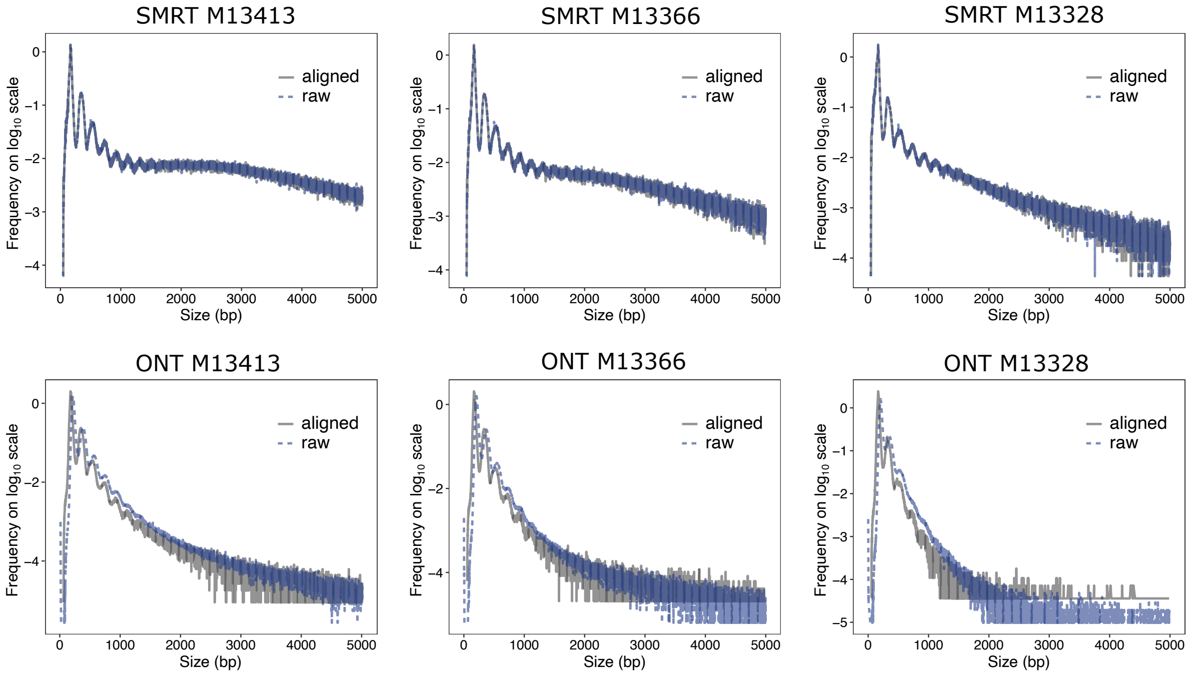

Fig. S7. Comparisons of size distributions between raw and aligned molecules. Size profiles of 6 samples, with 3 from SMRT and 3 from ONT sequencing are shown.

REFERENCES

Choy LYL, Peng W, Jiang P, Cheng SH, Yu SCY, Shang H, Olivia Tse OY, Wong J, Wong VWS, Wong GLH, et al. 2022. Single-molecule sequencing enables long cell-free DNA detection and direct methylation analysis for cancer patients. *Clin Chem* **68**: 1151–1163.

Yu SCY, Deng J, Qiao R, Cheng SH, Peng W, Lau SL, Choy LYL, Leung TY, Wong J, Wong VW-S, et al. 2023. Comparison of single molecule, real-time sequencing and nanopore sequencing for analysis of the size, end-motif, and tissue-of-origin of long cell-free DNA in plasma. *Clin Chem* **69**: 168–179.

Yu SCY, Jiang P, Peng W, Cheng SH, Cheung YTT, Tse OYO, Shang H, Poon LC, Leung TY, Chan KCA, et al. 2021. Single-molecule sequencing reveals a large population of long cell-free DNA molecules in maternal plasma. *Proc Natl Acad Sci U S A* **118**: e2114937118.