

Supplementary information:

High-throughput telomere length measurement at nucleotide resolution using the PacBio high fidelity sequencing platform

Cheng Yong Tham^{1,†}, LaiFong Poon^{2,†}, TingDong Yan^{2,3}, Javier Yu Peng Koh², Muhammad Khairul Ramlee², Vania Swee Imm Teoh¹, Suihan Zhang⁴, Yi Cai^{2,5}, Zebin Hong⁸, Gina S. Lee⁶, Jin Liu⁷, Hai Wei Song⁸, William Ying Khee Hwang^{2,9,10}, Bin Tean Teh^{2,8,11}, Patrick Tan^{1,2,12,13}, Lifeng Xu⁴, Angela S. Koh⁶, Motomi Osato^{1,14*} and Shang Li^{2,15,*}

[†]These authors contributed equally

*Co-corresponding authors

¹Cancer Science Institute of Singapore, National University of Singapore, 14 Medical Drive, Singapore 117599

²Cancer and Stem Cell Biology Program, Duke-NUS Medical School, 8 College Road, Singapore 169857

³School of Life Sciences, Shanghai University, 99 Shangda Road, Shanghai, China 200444

⁴Department of Microbiology and Molecular Genetics, University of California, Davis, CA USA 95616

⁵Key Laboratory of Molecular Target & Clinical Pharmacology, School of Pharmaceutical Science, Guangzhou Medical University, Guangzhou, China 511436

⁶National Heart Centre Singapore, Duke-NUS Medical School, 5 Hospital Drive, Singapore 169609

⁷Centre for Quantitative Medicine, Duke-NUS Medical School, 8 College Road, Singapore 169857

⁸Institute of Molecular and Cell Biology, Agency for Science, Technology and Research, (A*STAR), 61 Biopolis Drive, Singapore 138673

⁹Department of Haematology, Singapore General Hospital, 1 Hospital Drive, Singapore 169608

¹⁰Hematopoietic Stem Cell and Cellular Therapy Laboratory, Division of Medical Sciences, National Cancer Centre Singapore, 11 Hospital Drive, Singapore 169610

¹¹Laboratory of Cancer Epigenome, Division of Medical Science, National Cancer Centre Singapore, 11 Hospital Drive, Singapore 169610

¹²SingHealth/Duke-NUS Institute of Precision Medicine, National Heart Centre Singapore, Singapore 168752

¹³Epigenetic and Epitranscriptomic Regulation Domain, Genome Institute of Singapore, Agency for Science, Technology and Research, 60 Biopolis Drive, Singapore 138672

¹⁴International Research Center for Medical Sciences, Kumamoto University, 2-2-1 Honjo, Chuo-ku, Kumamoto, Japan 860-0811

¹⁵Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore, 2 Medical Drive, Singapore 117597

Corresponding authors

Motomi Osato: csimo@nus.edu.sg; motomi.osato@gmail.com

Shang Li: shang.li@duke-nus.edu.sg

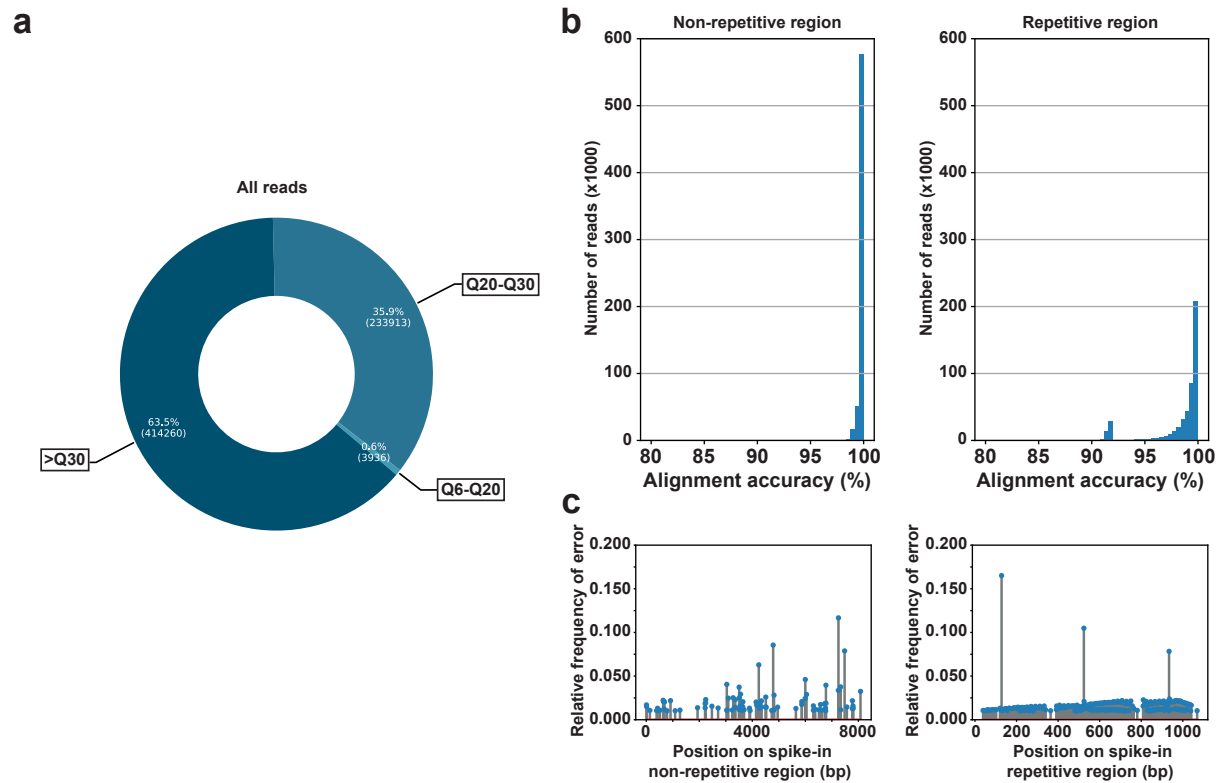


Figure S1: PacBio HiFi sequencing of DNA vector with telomeric repeats. **a.** Pie chart showing the accuracy of PacBio HiFi sequencing reads of the spike-in pWY82 DNA vector that contains the *Arabidopsis* telomeric sequence 5'-TTTAGGG-3'. **b.** Alignment accuracy of the non-repetitive and repetitive regions of the spike-in pWY82 DNA vector as calculated by HS-BLASTN. **c.** Relative frequency of sequencing error per base for the non-repetitive and repetitive regions of the spike-in pWY82 DNA vector. Only sequencing errors with a relative frequency greater than or equal to 0.01 are shown.

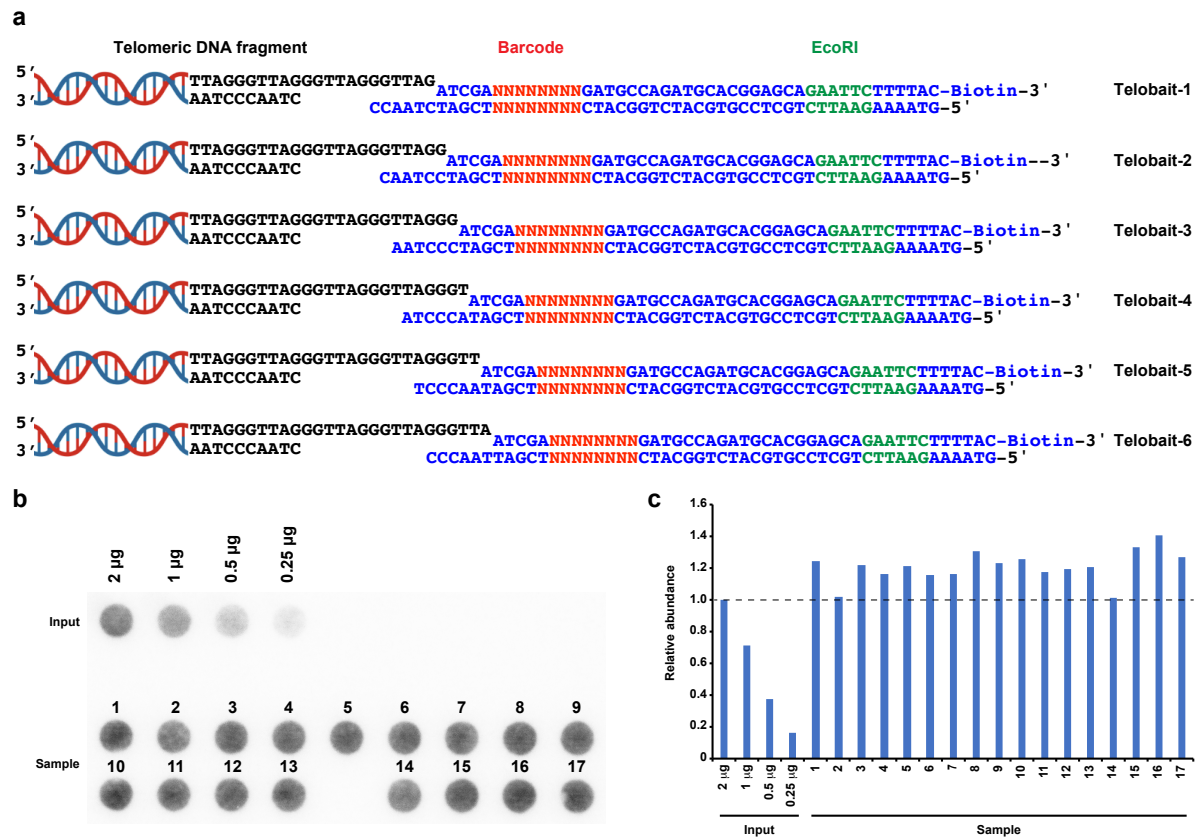


Figure S2: Telomere enrichment using telobaits. **a.** Schematic illustration of the design of telobaits for the enrichment of telomere-containing genomic DNA. Six telobaits were designed, each containing a single telomere repeat with all the six possible ends that can anneal to the single-stranded telomeric G-rich 3' overhangs. Each telobait also contains a unique barcode (Red), an EcoRI restriction endonuclease recognition site (green), and 3' biotin labelling. **b.** Dotblot showing the efficiency of telomere-containing genomic DNA enrichment using telobaits. The signal from a total input of 2 µg, 1 µg, 0.5 µg and 0.25 µg of genomic DNA extracted from HCT116 cells is shown on the first row. About 10 µg of genomic DNA obtained from HCT116 cells was enriched using telobaits without a barcode at a molar ratio of telobait:genomic DNA=10,000:1 (sample 1); or 12 different telobaits with unique barcodes at a molar ratio of telobait:genomic DNA=10,000:1 (sample 2 – 13); telobait without a barcode at a molar ratio telobait:genomic DNA=1,000:1 (sample 14); telobait:genomic DNA=5,000:1 (sample 15), telobait:genomic DNA=20,000:1 (sample 16); telobait:genomic DNA=50,000:1 (sample 17). **c.** Relative abundance of telomere repeat signal in **b** was quantified using ImageJ.

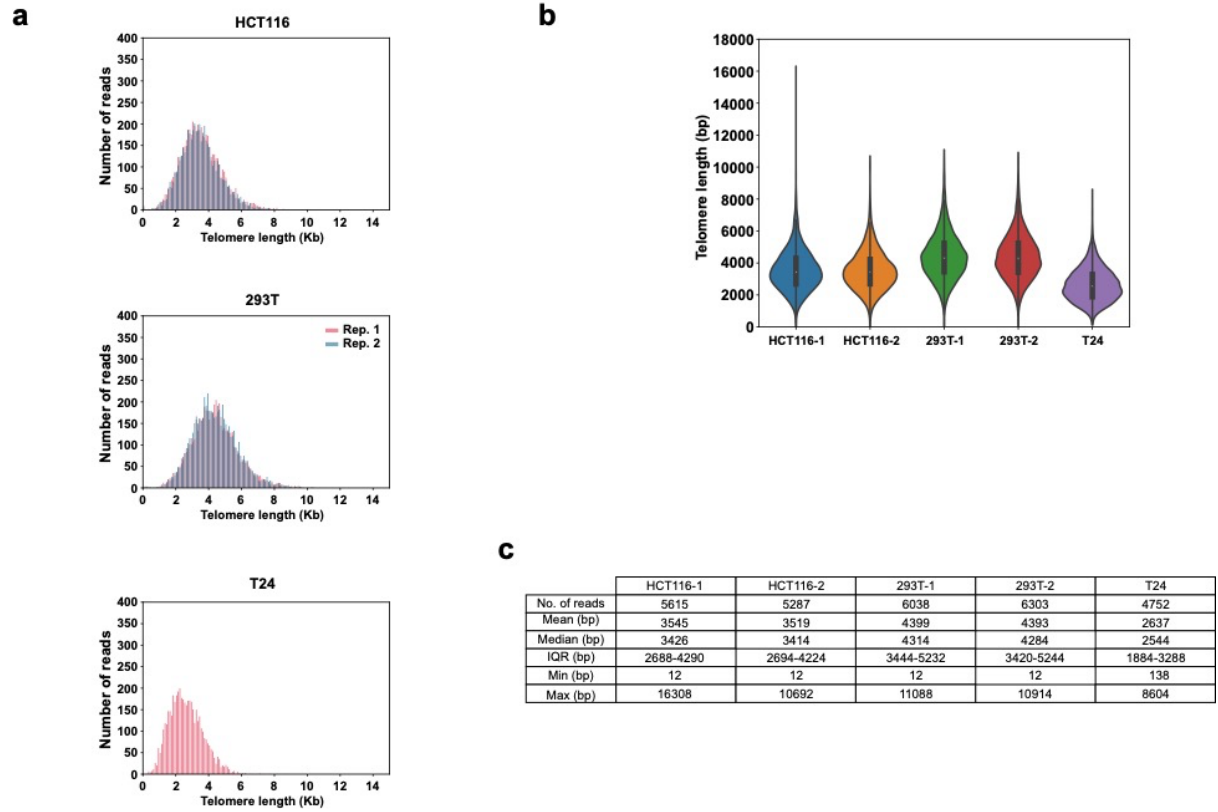


Figure S4: Telomere length measurement in culture cells using PacBio HiFi sequencing. **a. and b.** Histograms and violin plots of telomere length distribution in HCT116 (two biological replicates), 293T (two biological replicates) and T24 cells derived from an independent PacBio HiFi sequencing run. **c.** Table showing the characteristics of telomere length distribution in HCT116, 293T, and T24 cells as shown in violin plots in **b.** Within each violin, the white center circle denotes the median value, the bounds of box represent the 25th to 75th percentile values, the whiskers represent adjacent values within 1.5 interquartile range, the ends of the whiskers depict the minimum and maximum values within the range, and the violin shape reflect the kernel density plot of the entire dataset.

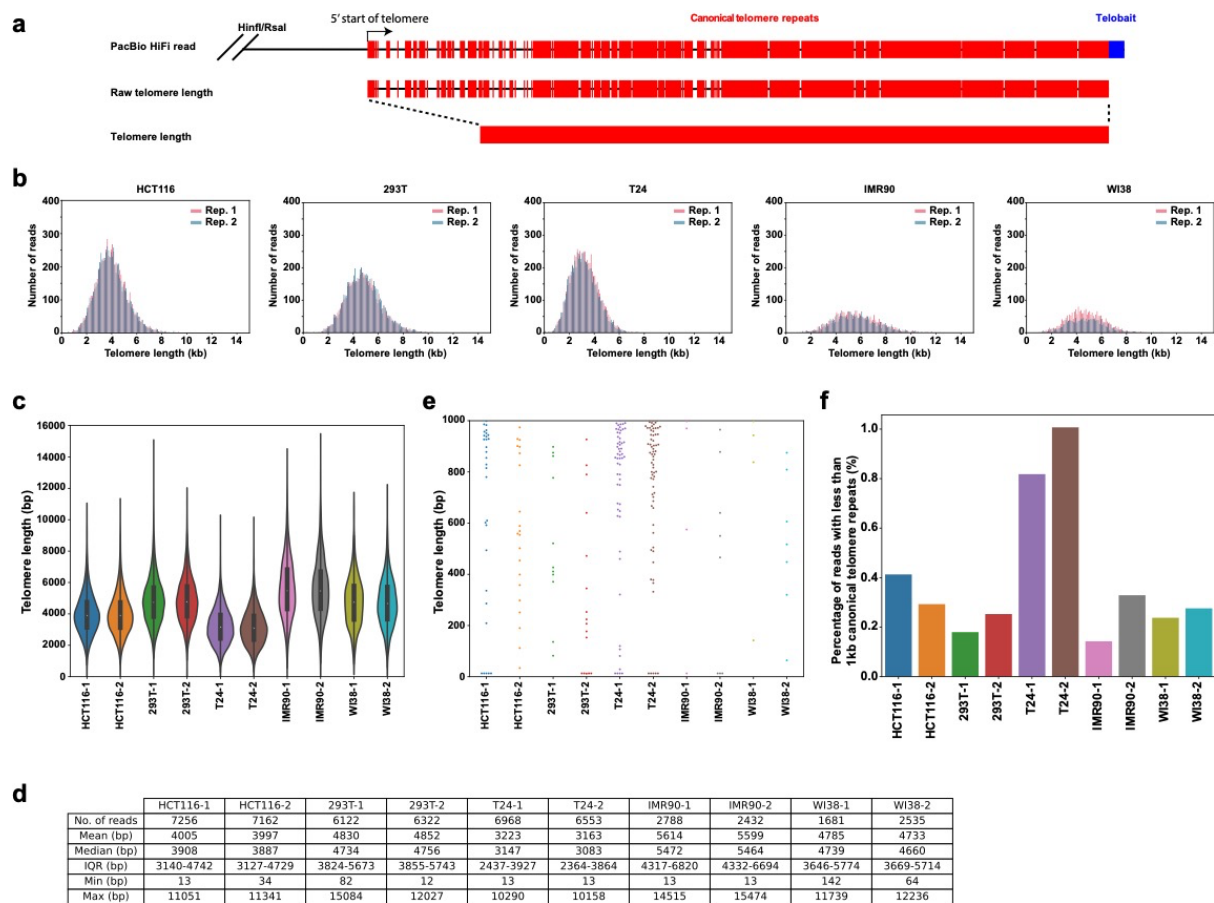


Figure S5: Raw telomere length measurement in culture cells using PacBio HiFi sequencing data as presented in Figure 2. **a.** Schematic illustration for the estimation of raw telomere length or telomere length with or without the inclusion of heterogenous TVSSs in sequencing reads, respectively. **b. and c.** Histograms and violin plots of raw telomere length distribution in HCT116, 293T, T24, IMR90 (PD47) and WI38 (PD40) cells obtained from two biological replicates (Rep.) in a single sequencing run. PD: population doubling. Within each violin, the white center circle denotes the median value, the bounds of box represent the 25th to 75th percentile values, the whiskers represent adjacent values within 1.5 interquartile range, the ends of the whiskers depict the minimum and maximum values within the range, and the violin shape reflect the kernel density plot of the entire dataset. **d.** Table showing the characteristics of raw telomere length distribution in HCT116, 293T, T24, IMR90 (PD47) and WI38 (PD40) cells obtained from two biological replicates as shown in the violin plot in **c.** **e.** Beeswarm plot showing the over-representation of extremely short telomeres (<1000 bp raw telomere length) in sequencing reads, specifically from T24 cells. **f.** Bar graph showing the over-representation of extremely short telomeres (<1000 bp raw telomere length) in sequencing reads, specifically in T24 cells after normalization.

a

	Run 1-1	Run 1-2	Run 2-1	Run 2-2
HCT116	3539 bp	3509 bp	3512 bp	3498 bp
293T	4395 bp	4389 bp	4286 bp	4299 bp
T24	2634 bp		2737 bp	2672 bp

b

HCT116 (mean: 3519.0, SD: 1231.8, Total reads: 25,320)

Sample size	1,000	2,000	5,000	10,000	15,000	20,000
Mean of means	3519.1	3518.6	3519.0	3519.1	3518.8	3518.9
SE of mean	39.1	27.4	17.5	12.2	9.9	8.7
1.96*SE of mean	76.6	53.7	34.3	23.9	19.4	17.1
95% CI	3442.4 - 3595.6	3465.3 - 3572.7	3484.7 - 3553.3	3495.1 - 3542.9	3499.6 - 3538.4	3501.9 - 3536.1

c

293T (mean: 4347.0, SD: 1404.5, Total reads: 24,785)

Sample size	1,000	2,000	5,000	10,000	15,000	20,000
Mean of means	4346.9	4347.1	4347.3	4347.0	4347.1	4347.1
SE of mean	44.6	31.4	19.8	14.1	11.4	9.9
1.96*SE of mean	87.4	61.5	38.8	27.6	22.3	19.4
95% CI	4259.6 - 4434.4	4285.5 - 4408.5	4308.2 - 4385.8	4319.4 - 4374.6	4324.7 - 4369.3	4327.6 - 4366.4

Figure S6: Accuracy and consistency of telomere length measurement using PacBio HiFi sequencing. **a.** Estimated mean telomere length in biological replicates of HCT116, 293T and T24 cells derived from two independent PacBio HiFi sequencing runs. **b.** and **c.** The predicted standard error (SE) and 95% confidence interval (CI) of mean calculated by random sampling at different sequencing depths using the pooled telomere-containing reads from HCT116 and 293T cells in two independent sequencing runs, respectively. The bootstrap method was used where telomeric reads were resampled from the actual sample with replacement for 10,000 iterations for the different sample sizes.

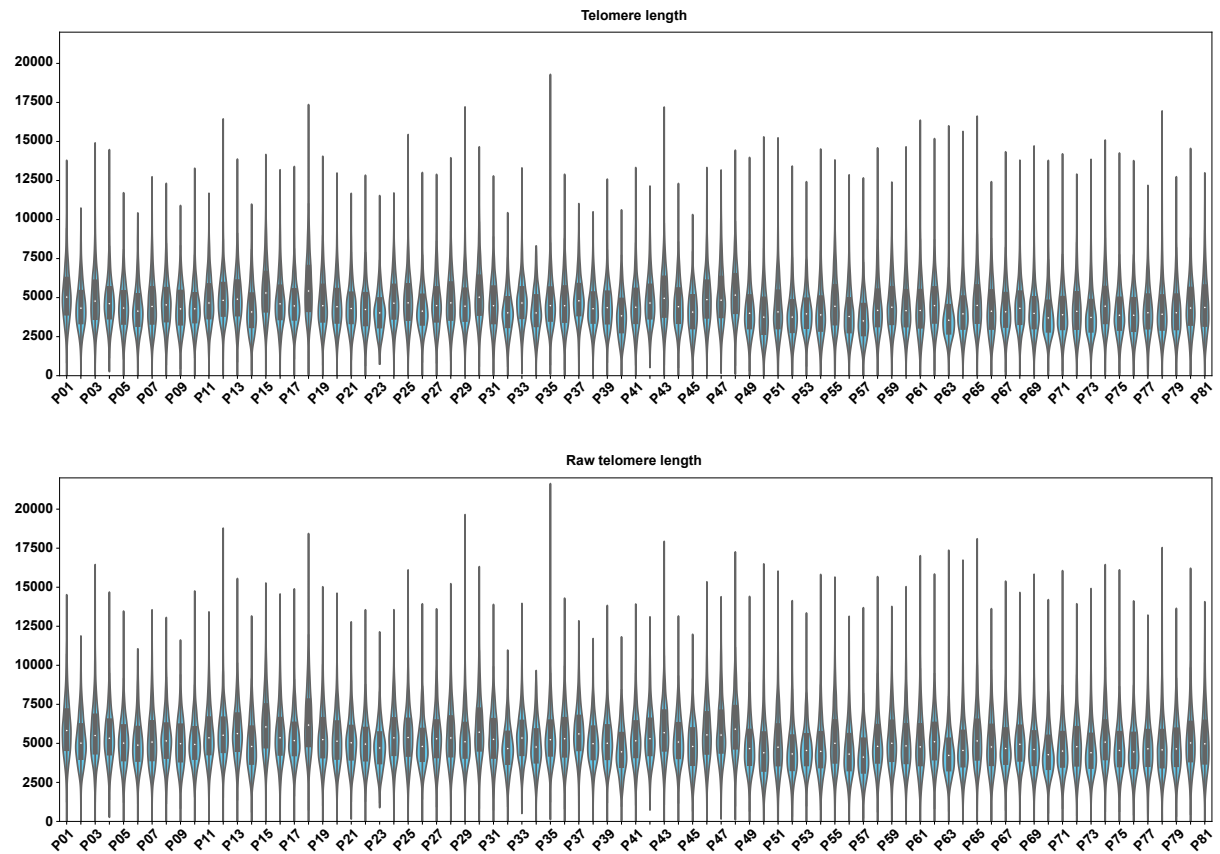


Figure S7: Violin plots showing telomere length and raw telomere length distribution in 81 patient PBL samples from Singapore cohort. Please refer to the Source Data 2 file for the n value of each sample. Within each violin, the white center circle denotes the median value, the bounds of box represent the 25th to 75th percentile values, the whiskers represent adjacent values within 1.5 interquartile range, the ends of the whiskers depict the minimum and maximum values within the range, and the violin shape reflect the kernel density plot of the entire dataset.

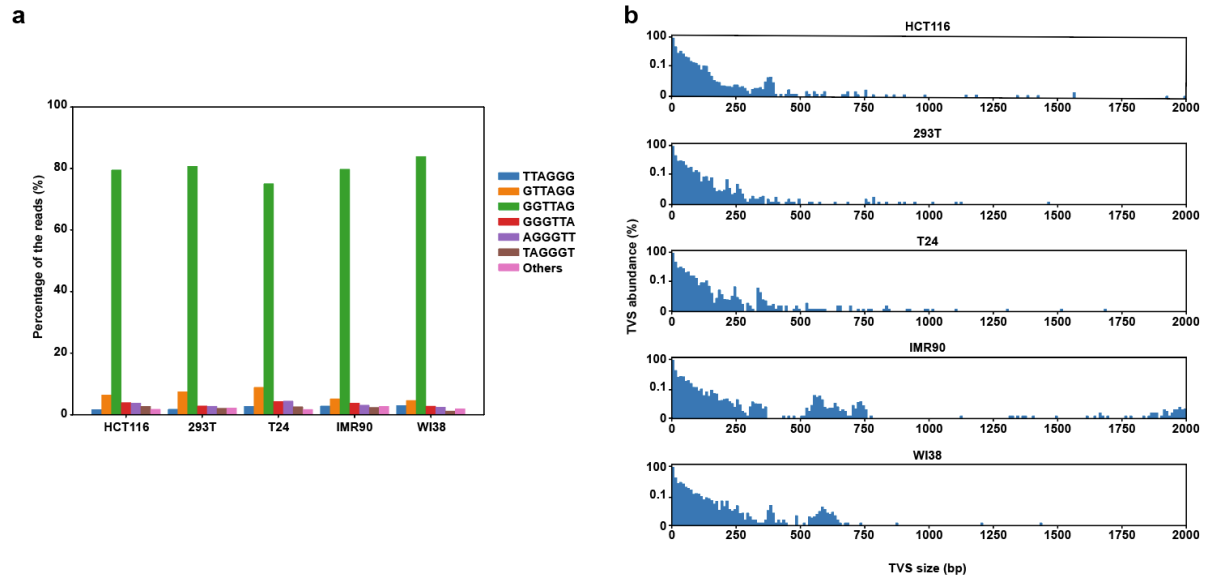
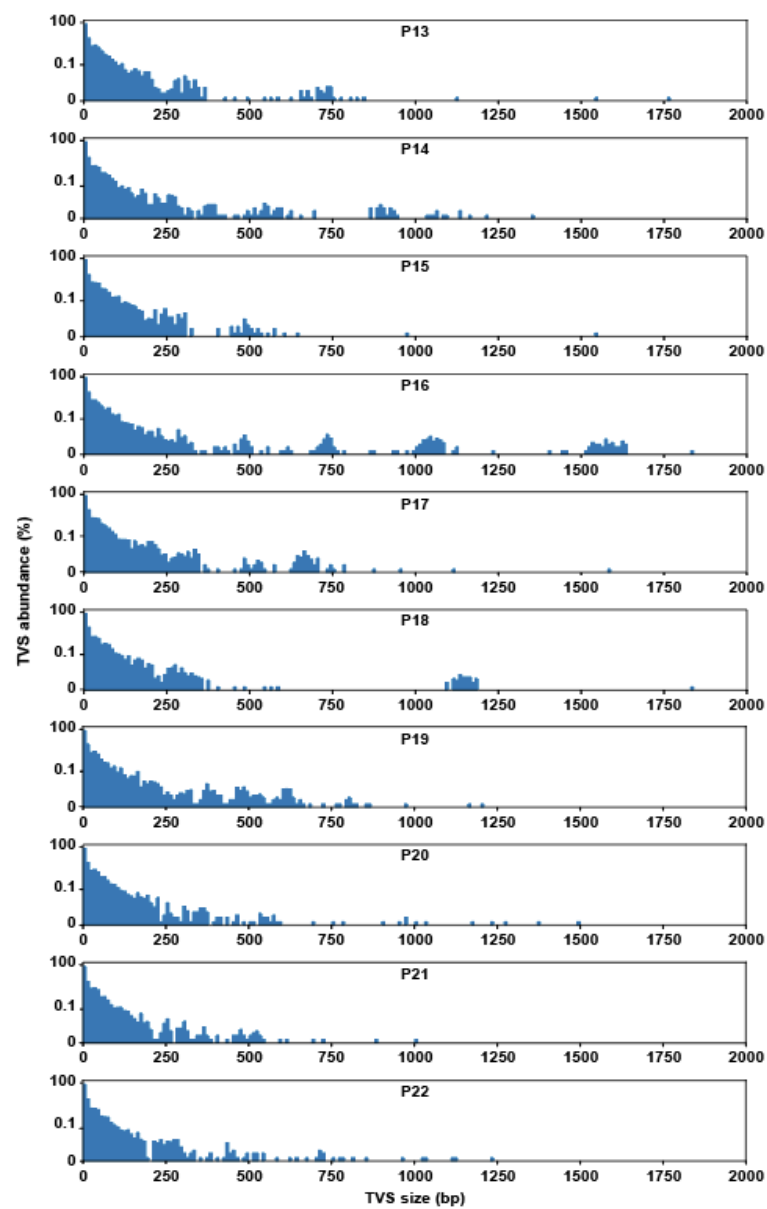
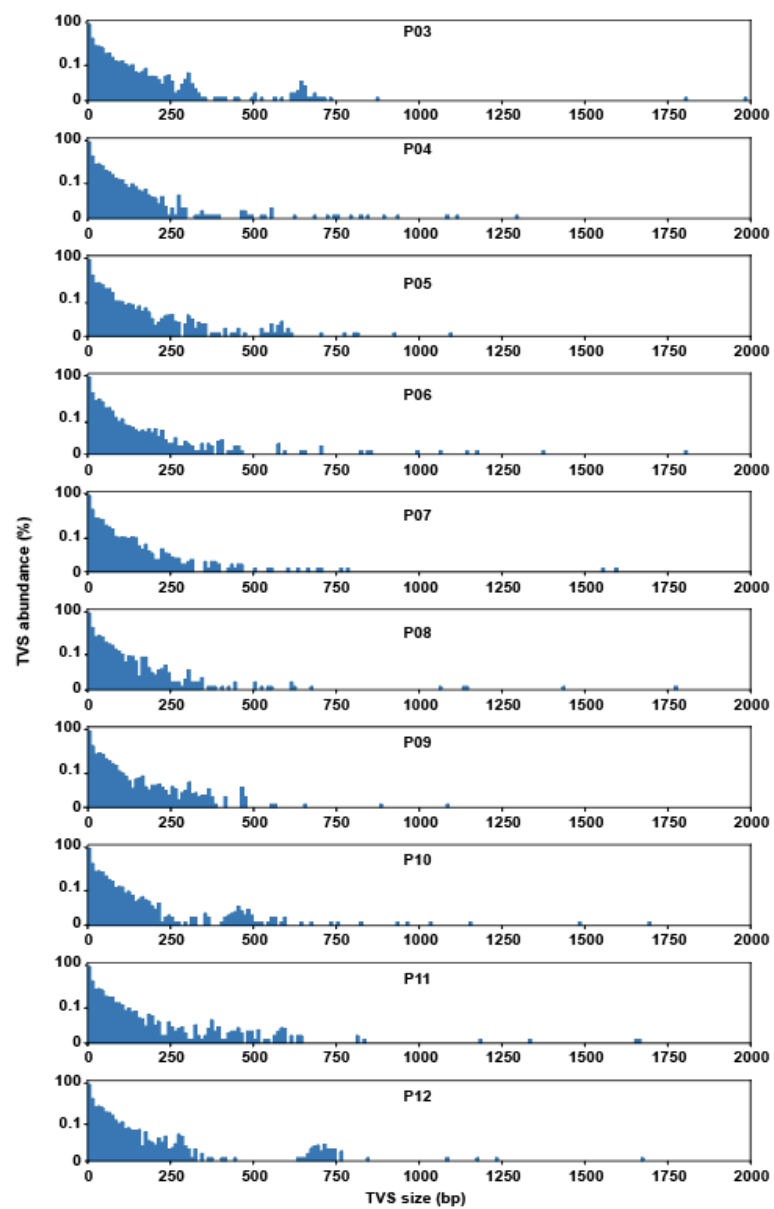
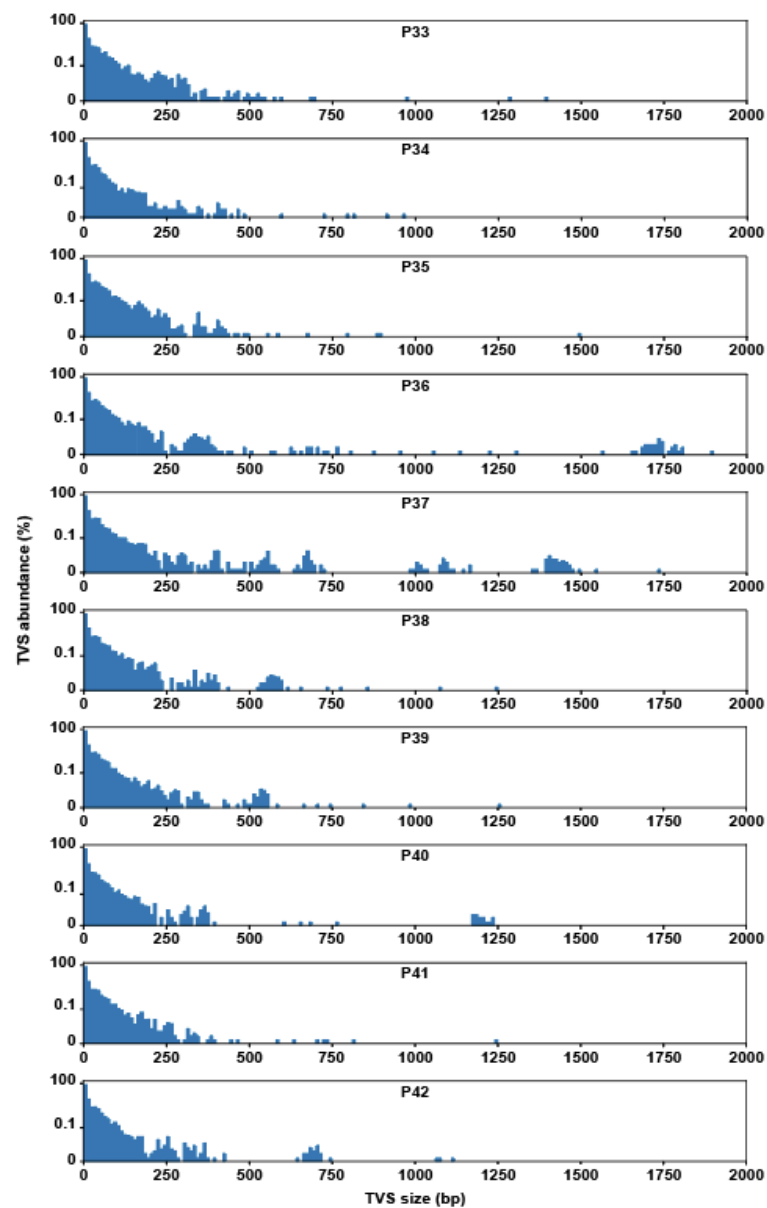
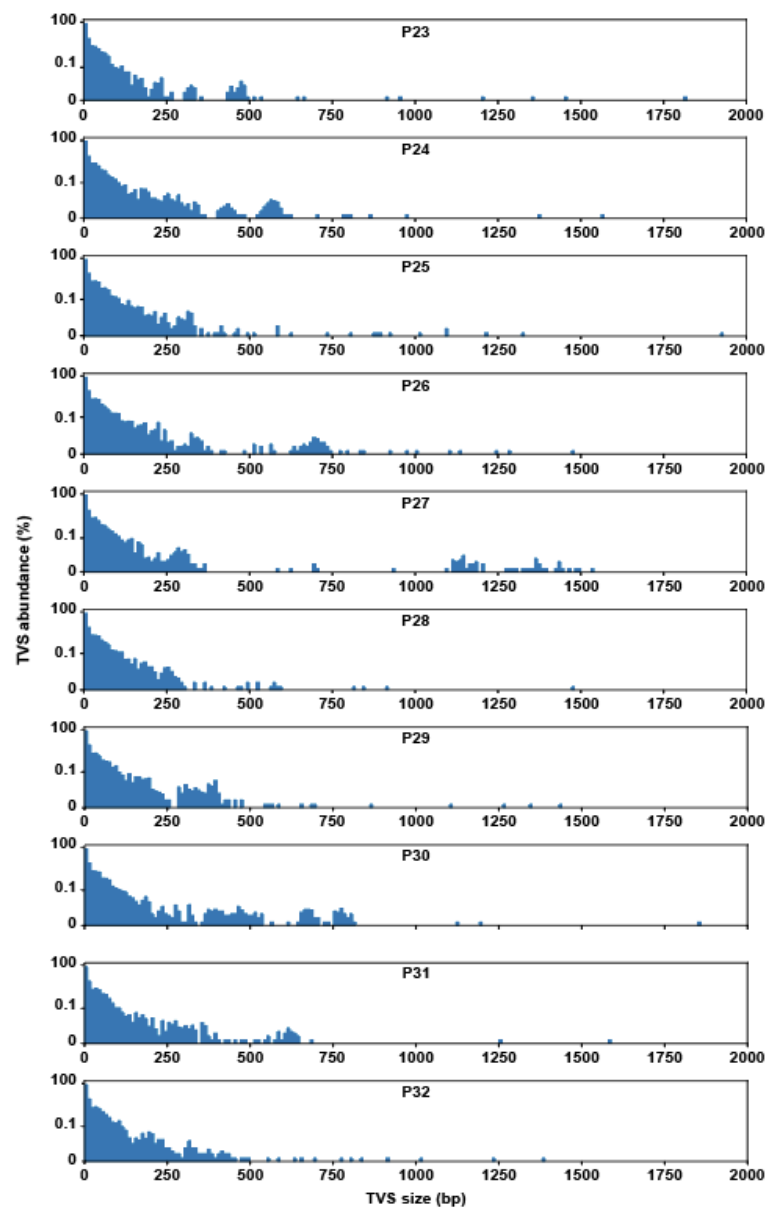
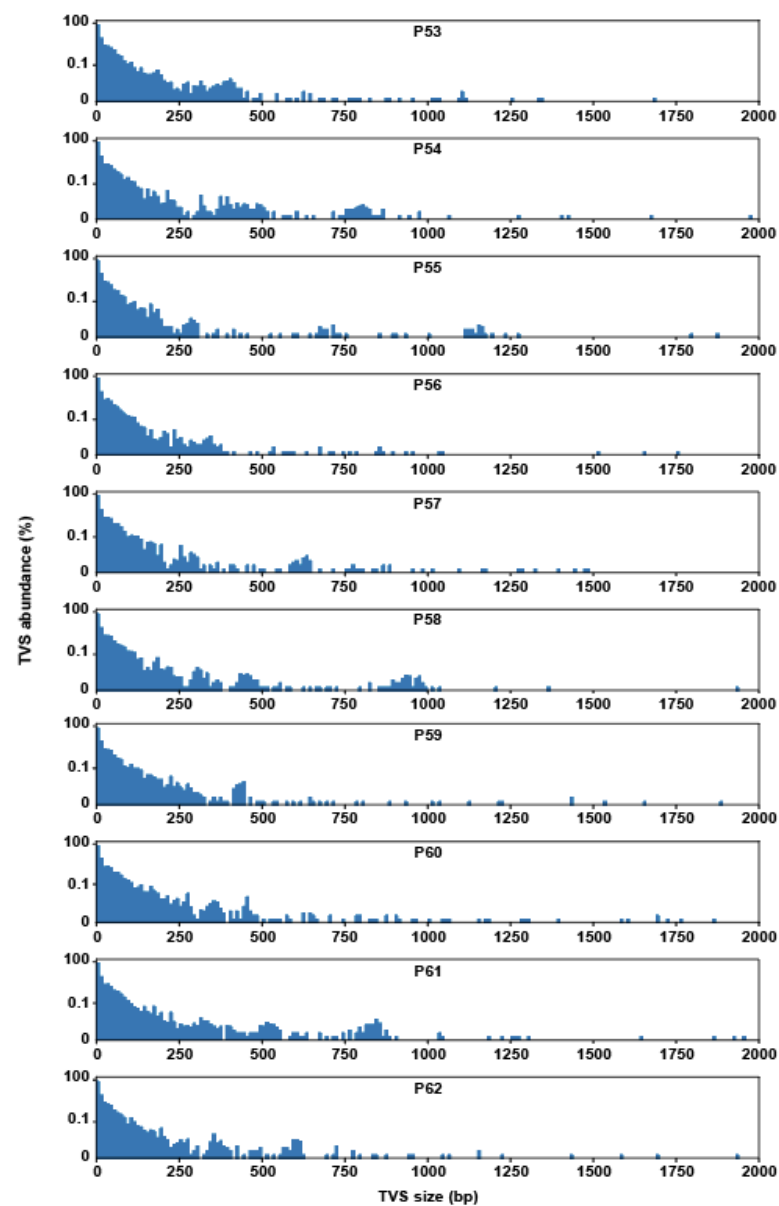
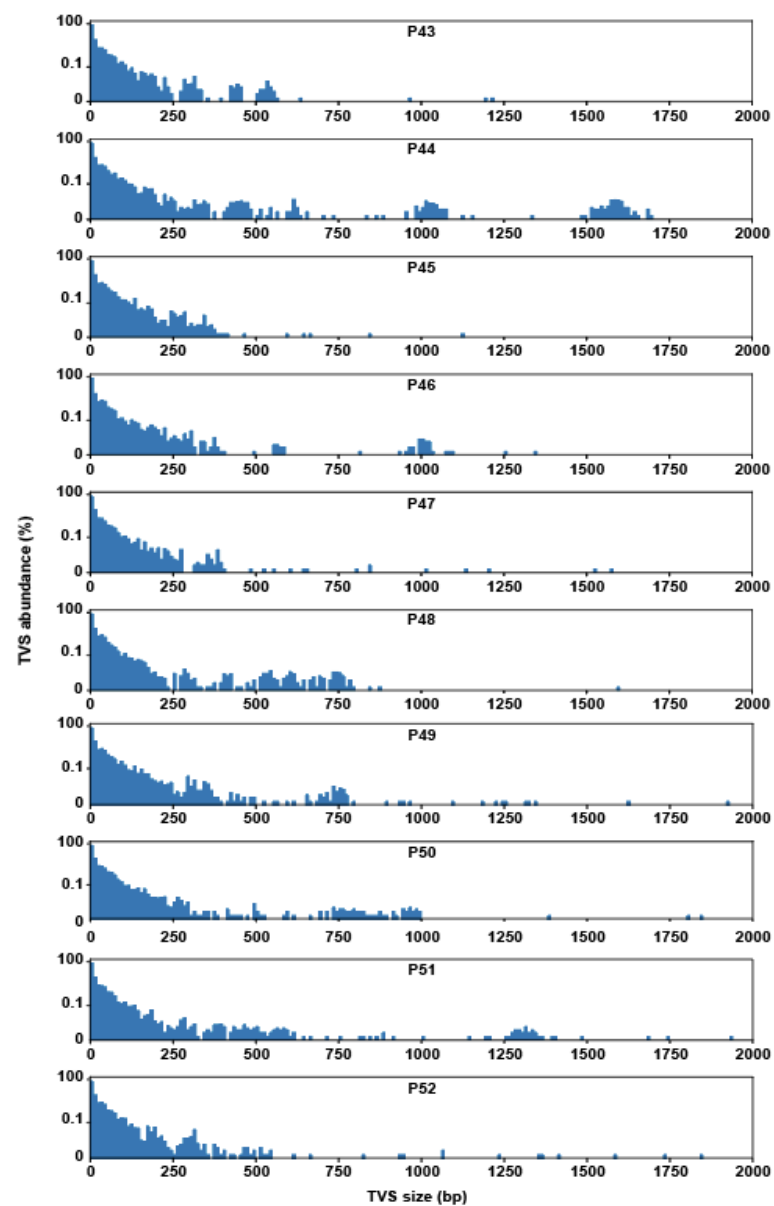
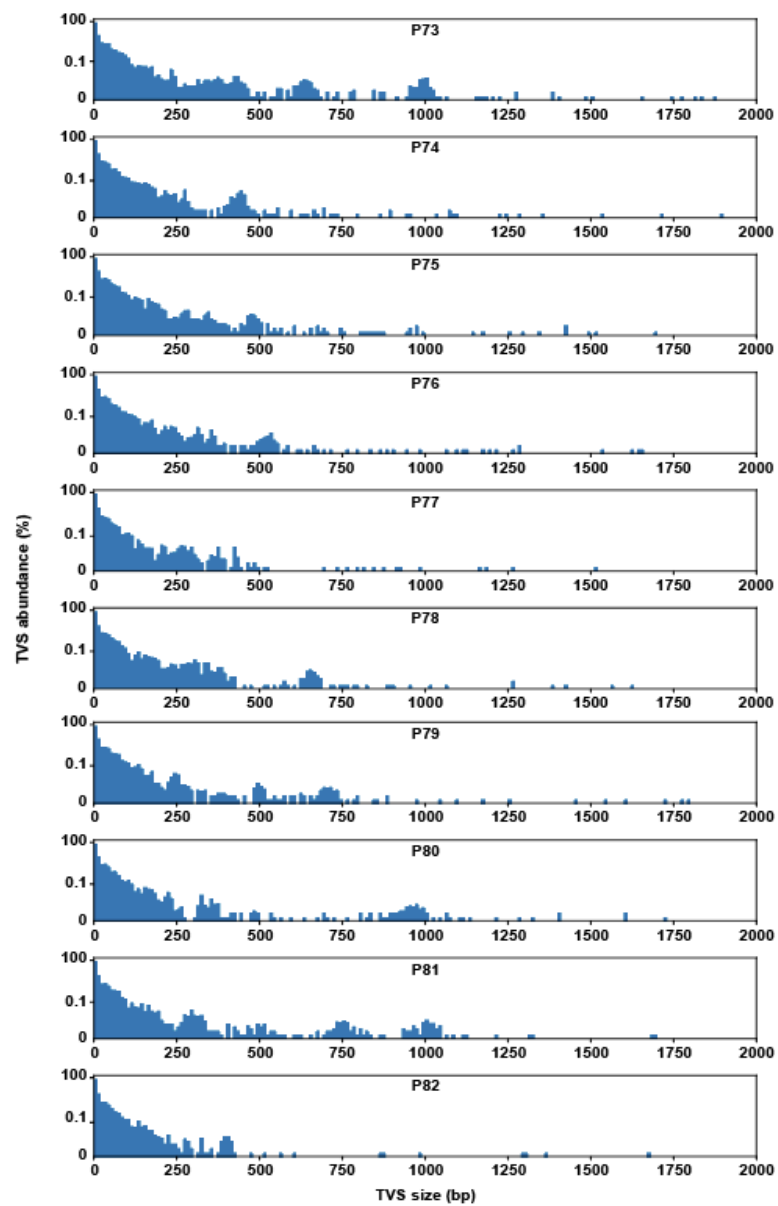
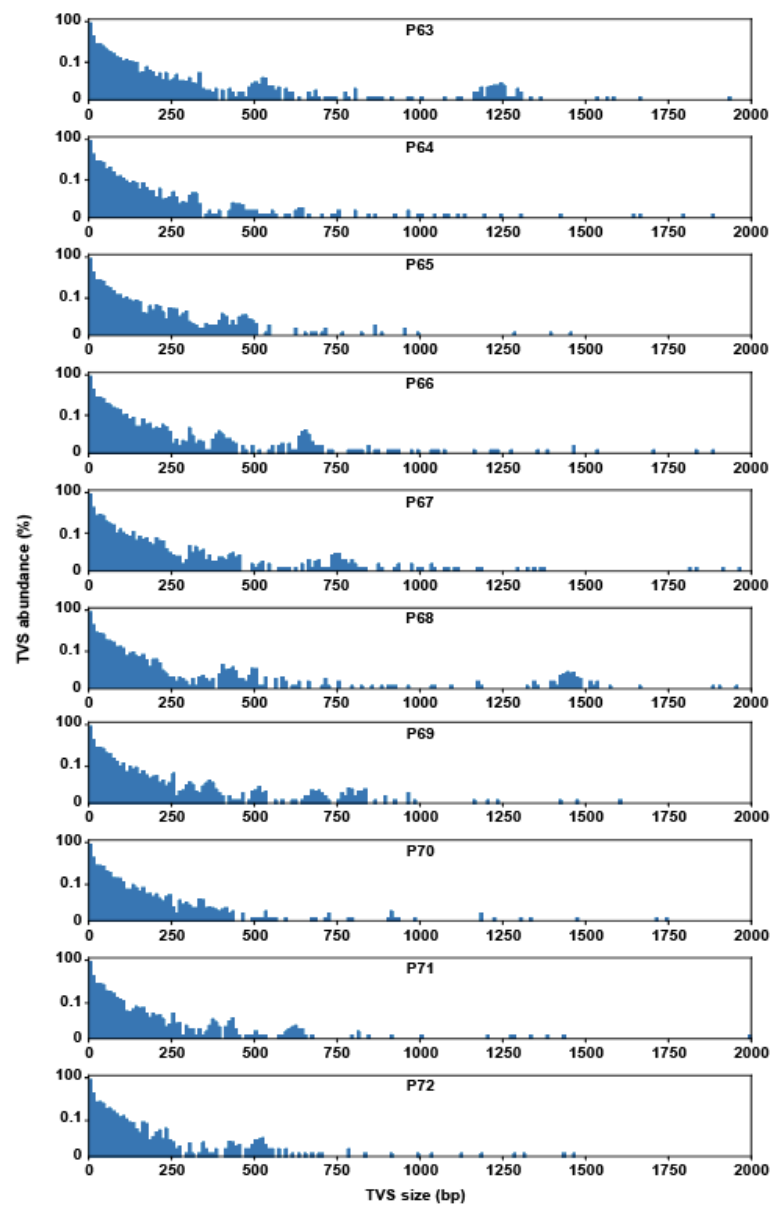


Figure S8: Proportion of different G strand end sequences and TVS size abundance distributions in culture cell lines. a. Bar graph showing the percentage of different G strand end sequences across culture cell lines. Like patient PBL samples, the most common G strand end observed is 5'-GGTTAG-3'. **b.** Histograms showing the unique TVS size abundance profile of each culture cell line. The Y-axes are presented in logarithmic scale. Only TVSs with sizes less than or equal to 2kb are shown.









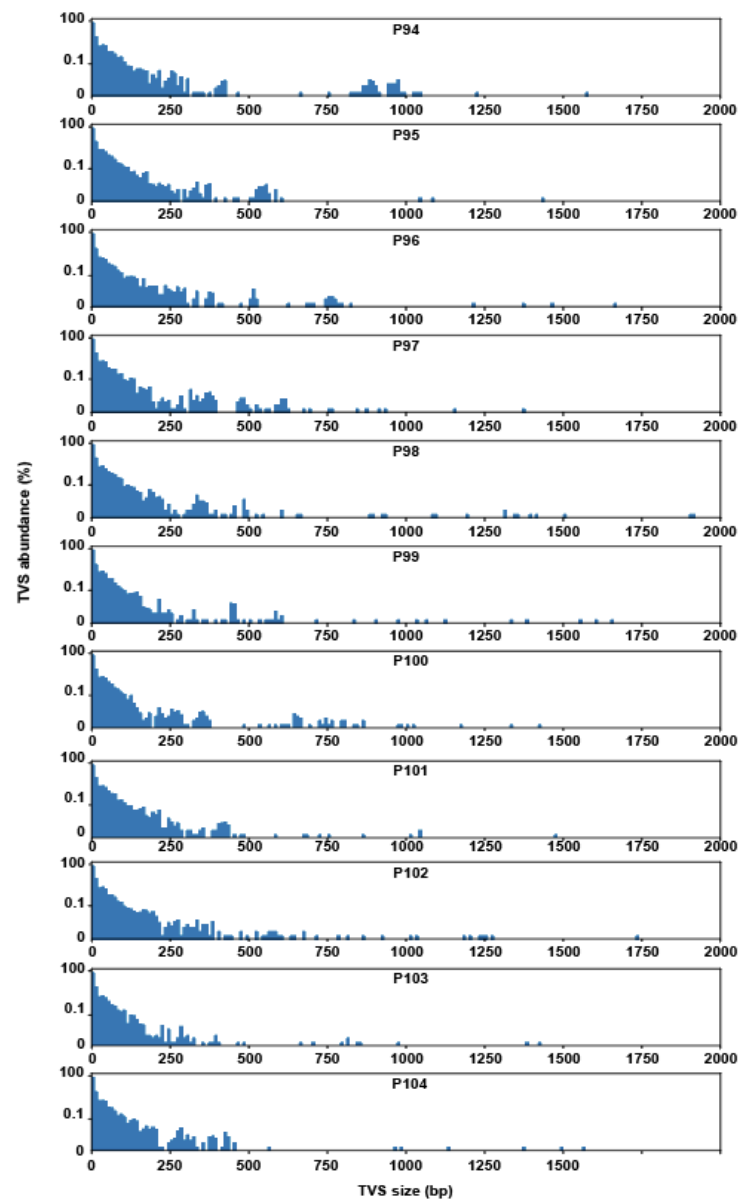
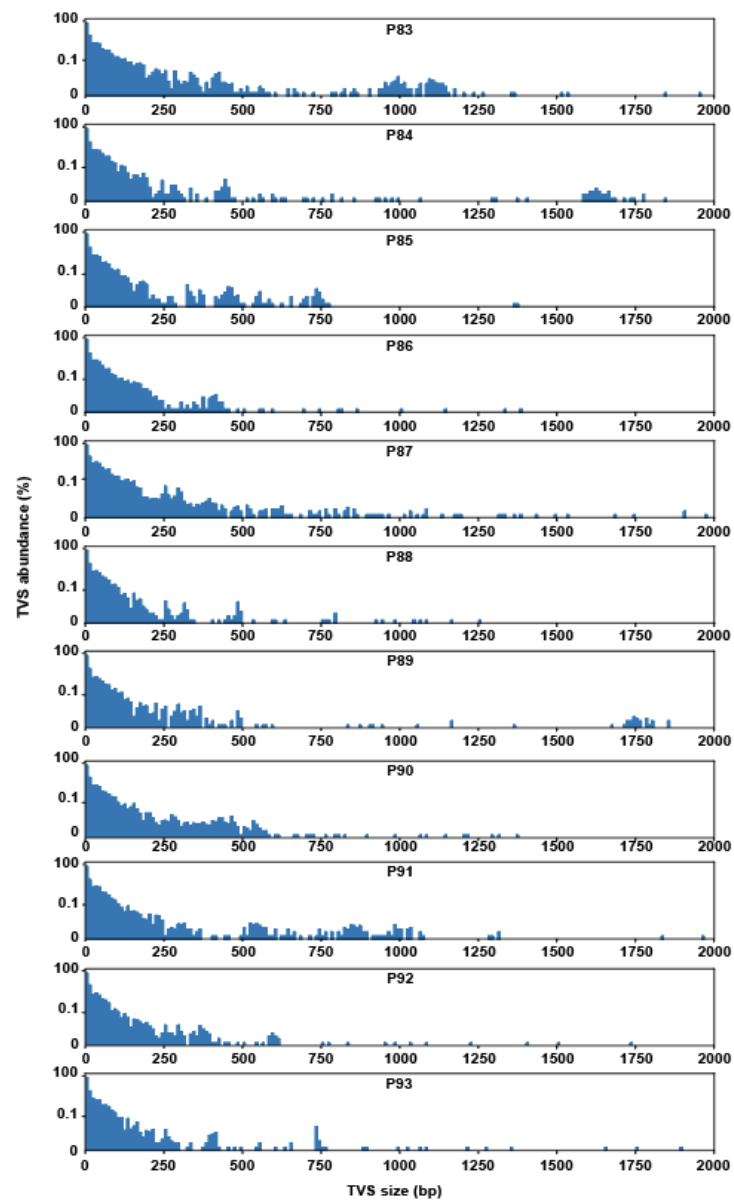
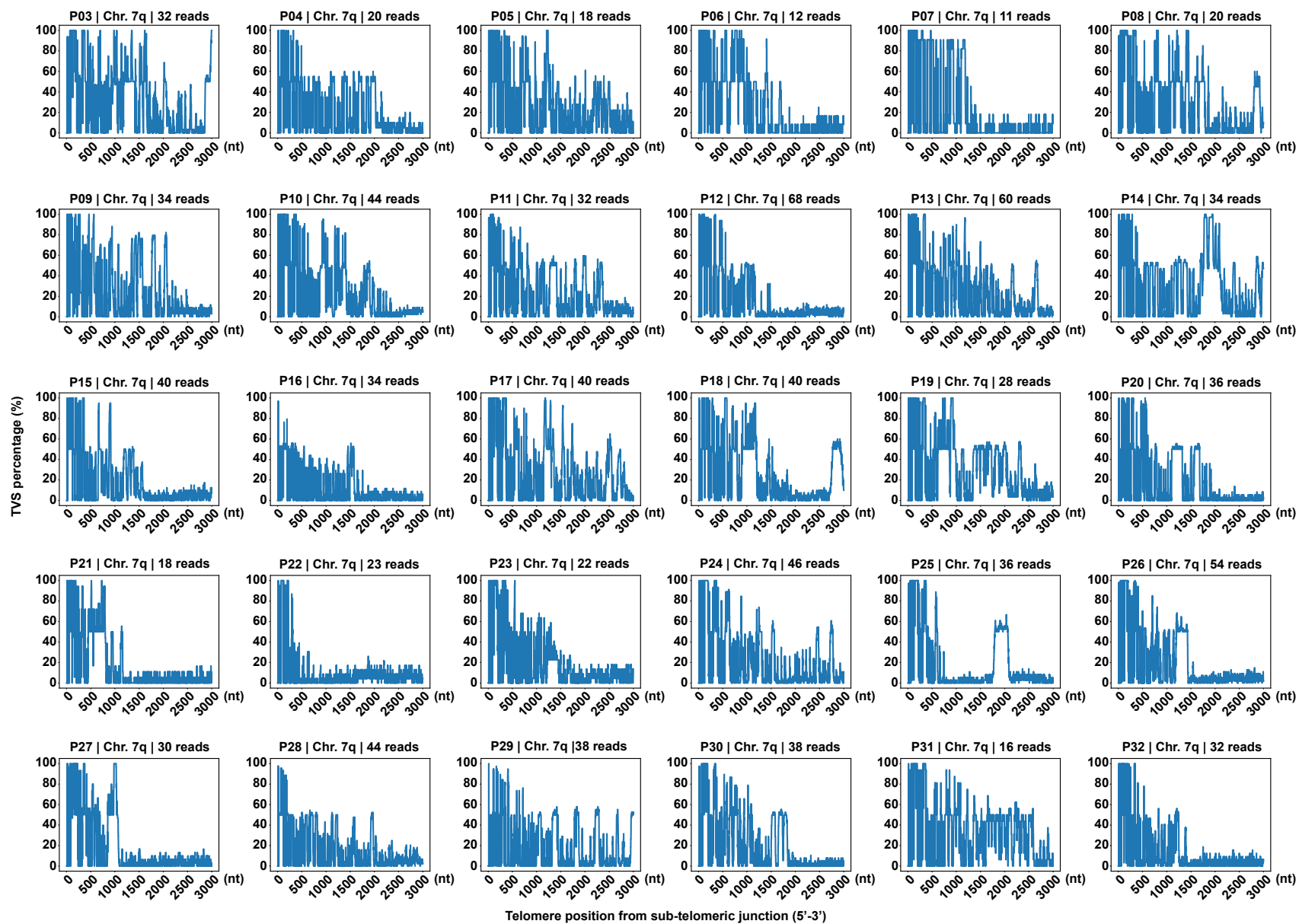
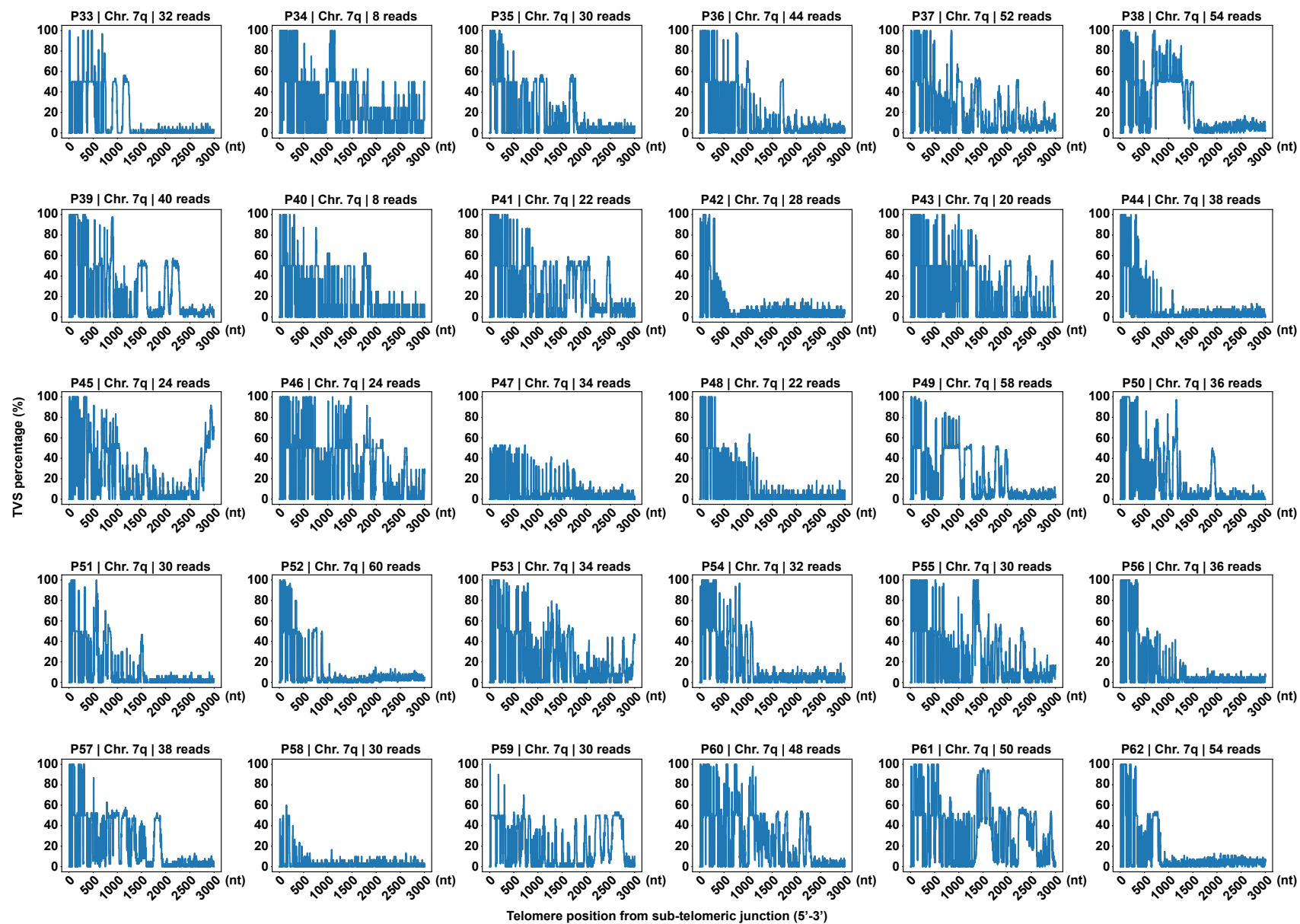
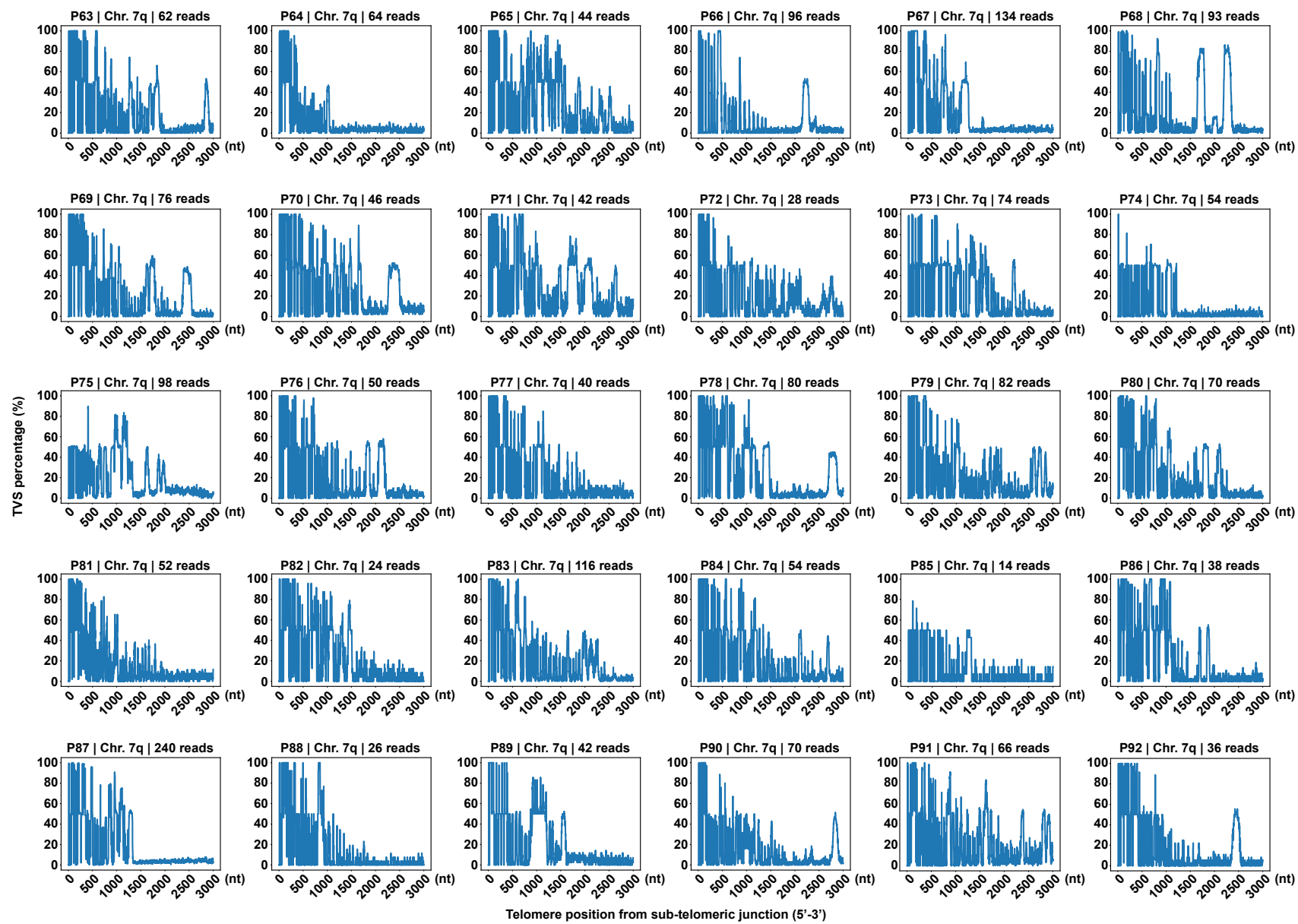


Figure S9: The unique distribution of TVS size abundance profile in patient samples (P03-P104).







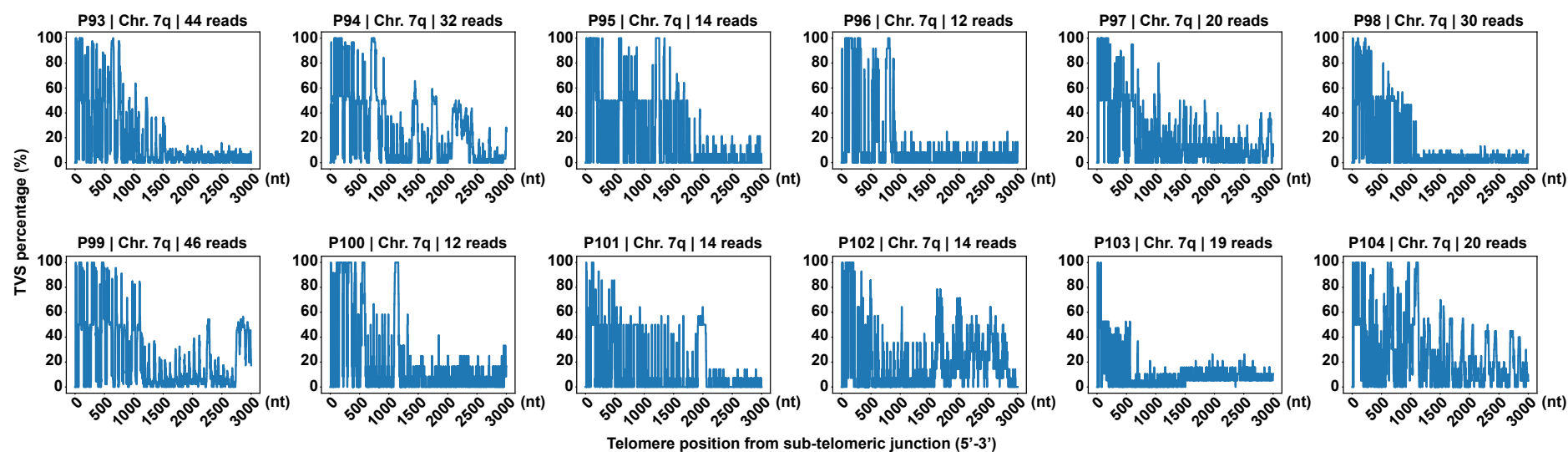


Figure S10: TVS signatures of individual patient samples (P03 to P104) for Chr. 7q. Longer and more abundant TVSs are presented near the sub-telomeric ends of the telomere repeat regions in the patient samples as shown by high TVS percentages.

TVS signature correlation for Chr. 7q

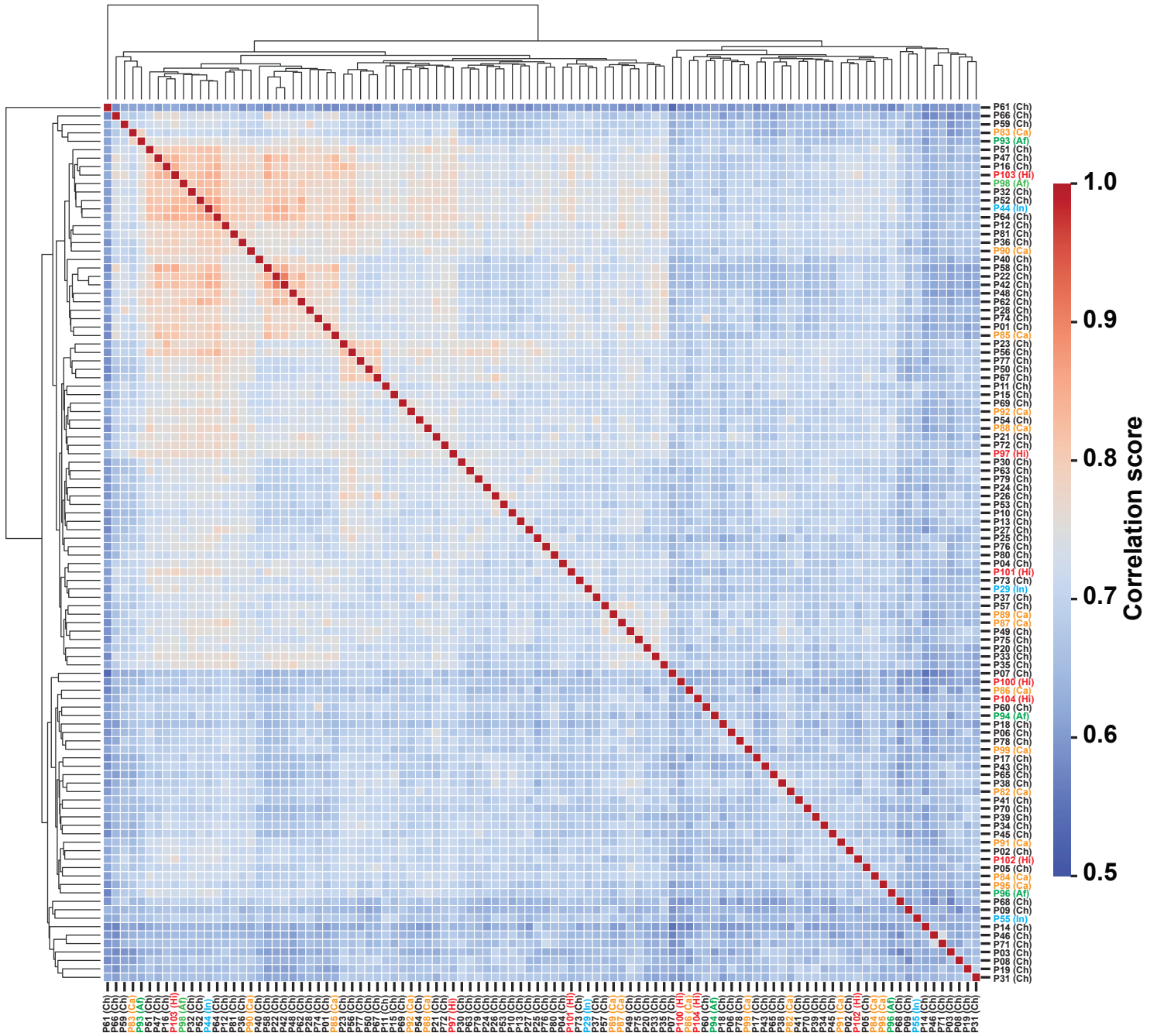


Figure S11: The dendrogram of TVS signature correlation for Chr. 7q from patient samples (P01 to P104). The TVS signatures were pairwise compared by Euclidean distance and were found to be unique to each patient. The ethnicity of each patient sample is color-coded. Af=African American; Ch=Chinese; Ca=Caucasian; Hi=Hispanic, In=Indian.

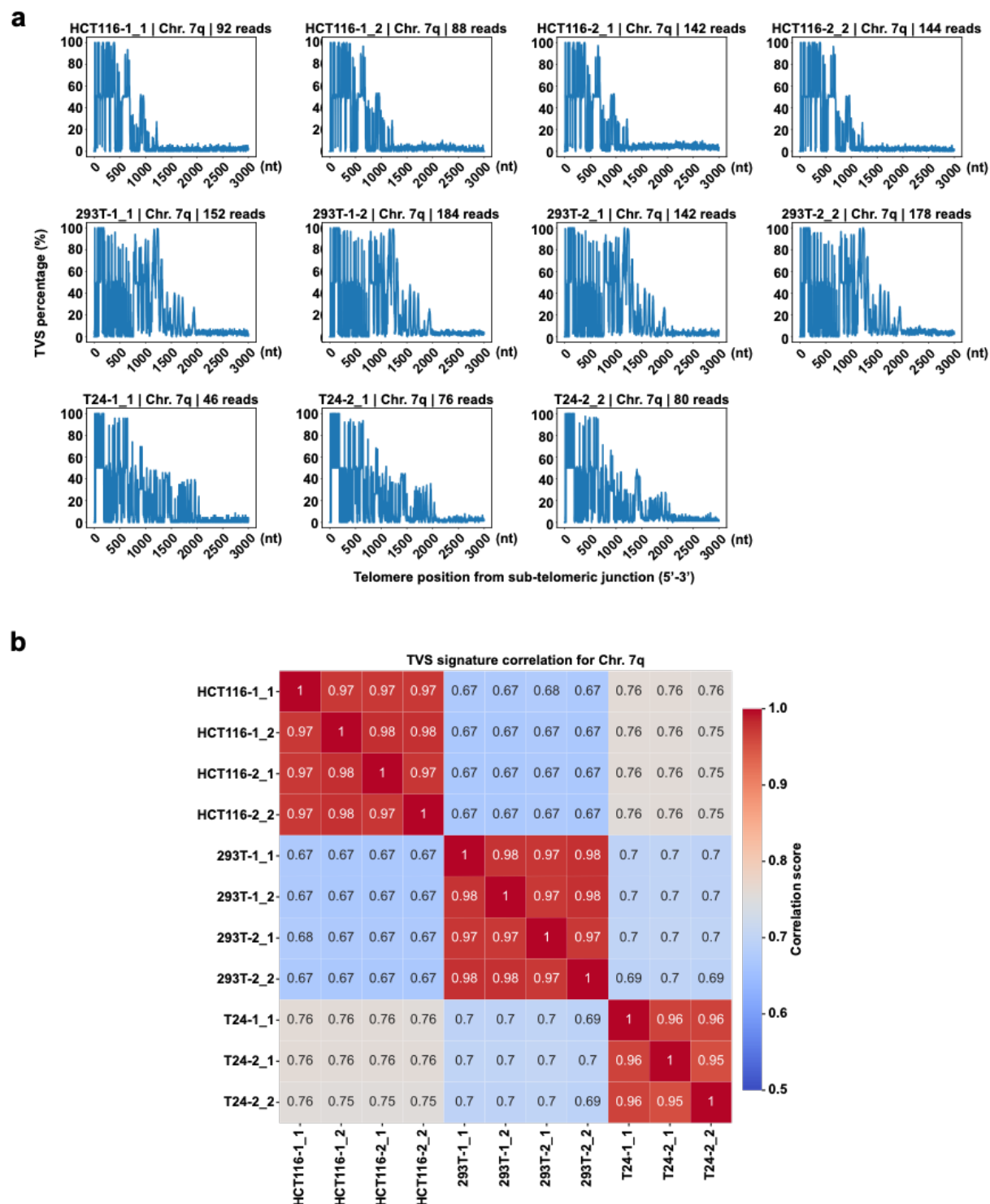


Figure S12: The consistency of TVS signature correlation of Chr. 7q from biological replicates of cell line samples. a. TVS signature plots of biological replicates of cell line samples from independent sequencing runs. **b.** The TVS signature correlation of Chr. 7q from biological replicates of cell line samples. Source data are provided in Source Data 2.

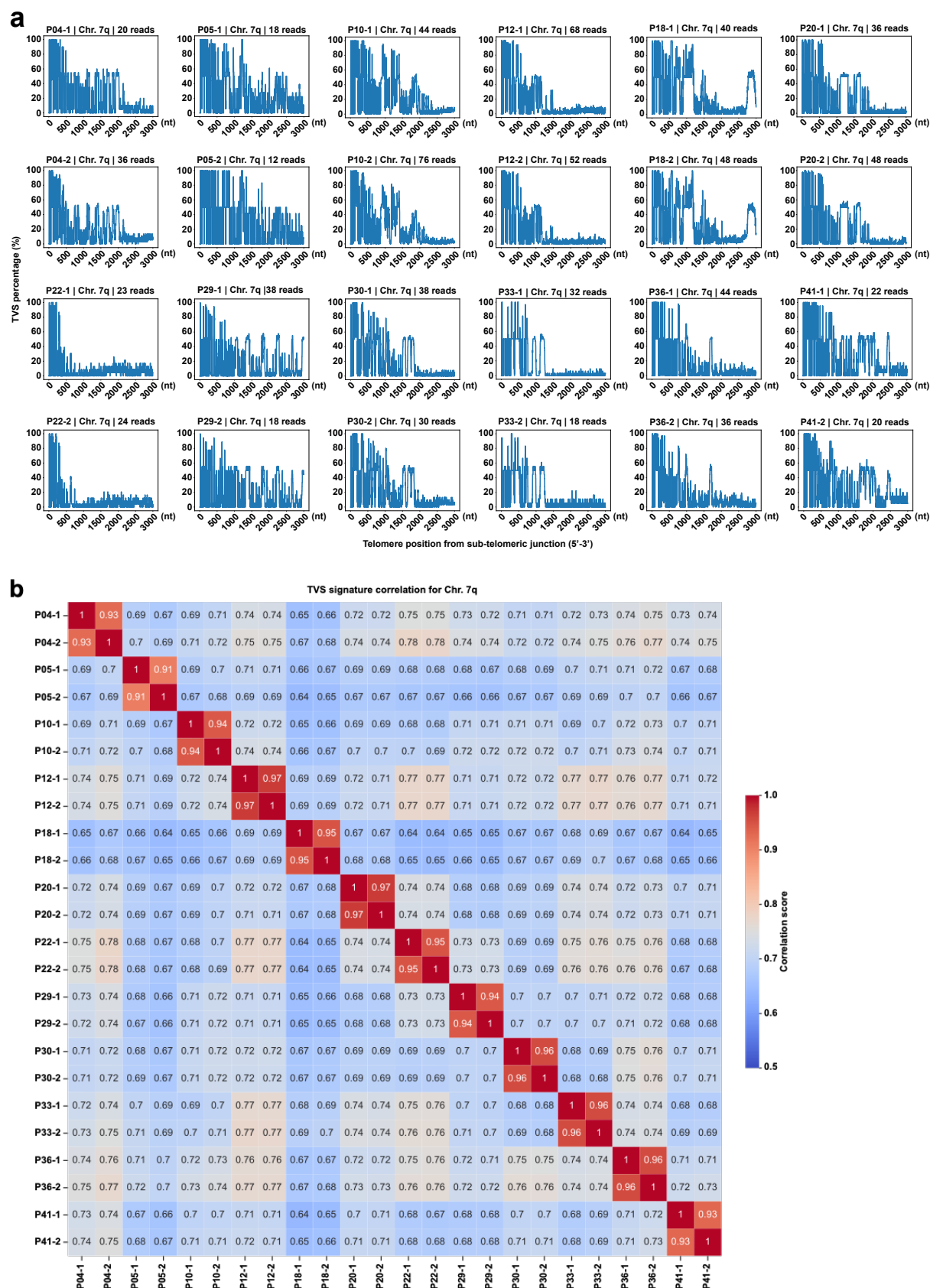


Figure S13: The TVS signature correlation of Chr. 7q from biological replicates of patient samples. A. TVS signature plots of biological replicates of patient samples from independent sequencing runs. **B.** The TVS signature correlation of Chr. 7q from biological replicates of patient samples. Source data are provided in Source Data 2.

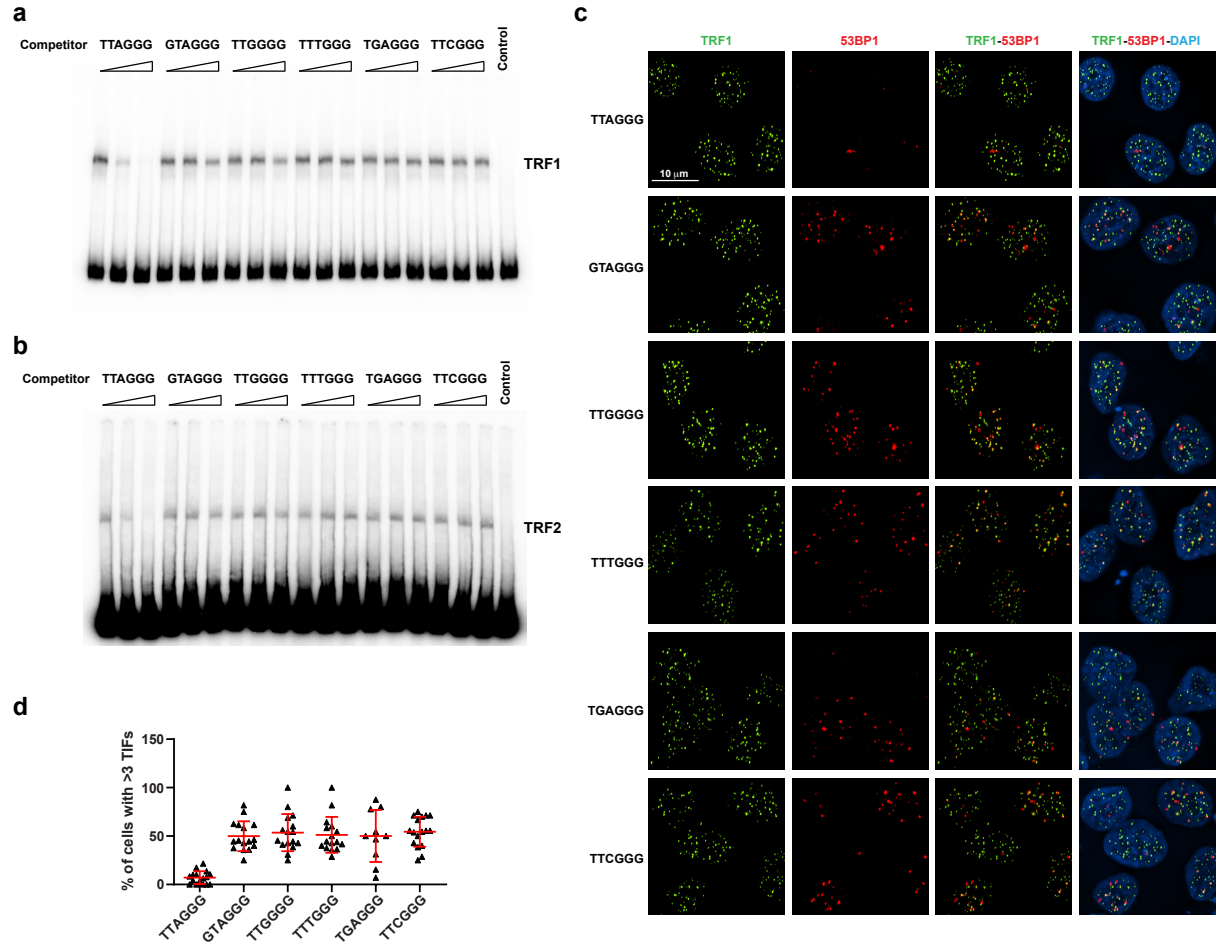


Figure S14: TVSSs disrupt the binding of shelterin complex at telomere. **a.** Gel mobility shift assay⁴² of *in vitro*-translated TRF1 and ³²P-labeled DNA fragment containing canonical telomere repeats sequence in the presence of competitive DNA fragments with either canonical telomere repeats or TVSSs. Representative results from one of the two independent experiments are shown. **b.** Gel mobility shift assay of *in vitro*-translated TRF2 and ³²P-labeled DNA fragment containing canonical telomere repeats sequence in the presence of competitive DNA fragments with either canonical telomere repeats or TVSSs. Representative results from one of the two independent experiments are shown. **c.** Telomere dysfunction induced foci (TIFs) in the WA18 (hTR^{-/-}) embryonic stem cells transduced with lentivirus co-expressing a puromycin resistance cassette and hTRs encoding either canonical telomere repeat or TVSSs as indicated. Lentivirus transduced cells were subjected to puromycin selection for 48 hours (except for the “TGAGGG” variant), and then immunostaining using anti-TRF1 and anti-53BP1 antibodies to identify TIFs. Due to the toxicity of “TGAGGG”, the immunostaining was done 3 days after lentivirus transduction without puromycin selection. **d.** Quantification of TIFs in WA18 hTR^{-/-} cells expressing different hTRs. Each dot on the plot represents value obtained from one image, and bars represent mean values ± SD. About 16 images (8-15 cells per image) from two independent experiments were taken randomly and examined. In sum, >120 cells per line were analyzed for TIFs. All image files were randomly assigned coded names to allow blinded scoring for spots co-localization. The lower number of TIF in cells expressing “TGAGGG” is likely due to no enrichment of cells using puromycin selection. Therefore, a large proportion of cells likely did not uptake the lentivirus expressing the mutant hTR containing “TGAGGG” variant sequence.

[illegible]

Figure S15: The sequence of a typical telomere-containing PacBio HiFi sequencing read of Chr. 7q obtained from P14. The first 1kb of canonical telomere repeats next to the sub-telomeric region is highlighted in red. The telobait sequence and the barcode within are highlighted in green and blue, respectively.

[illegible]

Figure S16: The sequence of a typical telomere-containing PacBio HiFi sequencing read of Chr. 7q obtained from P15. The first 1kb of canonical telomere repeats next to the sub-telomeric region is highlighted in red. The telobait sequence and the barcode within are highlighted in green and blue, respectively.

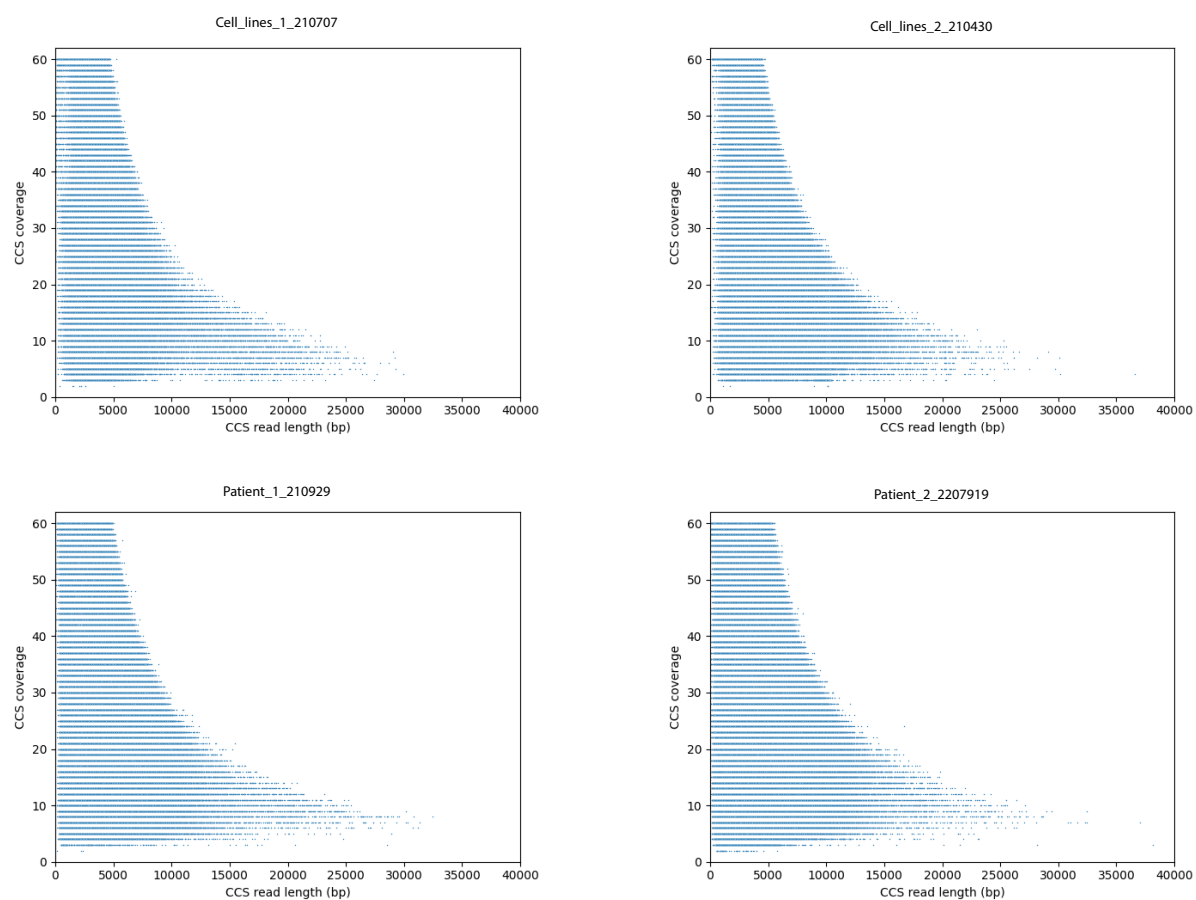


Figure S17: CCS read coverage and length for all four PacBio Hi-Fi sequencing runs presented in this paper.

Un-cropped immunostaining raw data for Figure S14c

