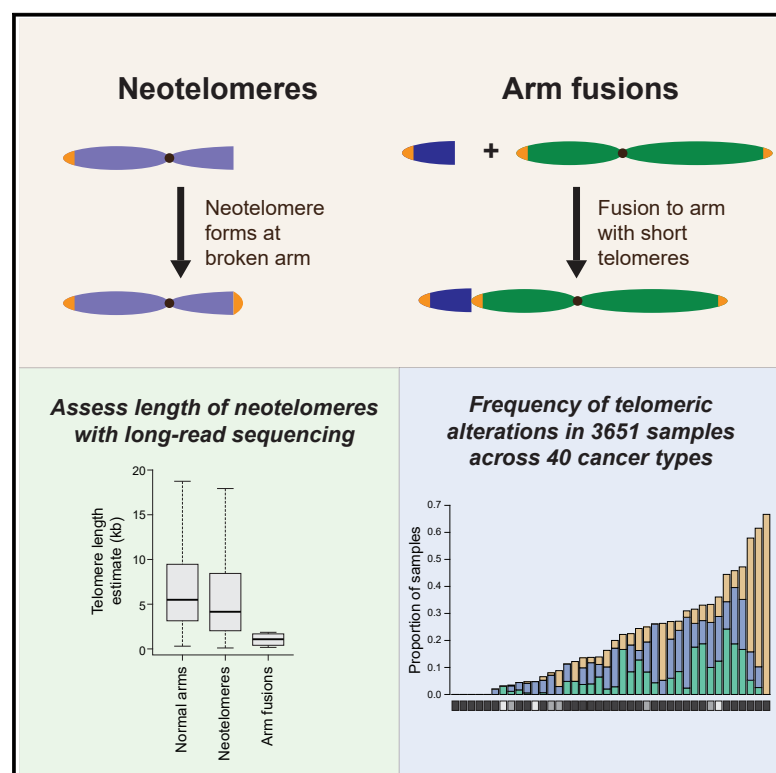


# Neotelomeres and telomere-spanning chromosomal arm fusions in cancer genomes revealed by long-read sequencing

## Graphical abstract



## Authors

Kar-Tong Tan, Michael K. Slevin, Mitchell L. Leibowitz, Max Garrity-Janger, Jidong Shan, Heng Li, Matthew Meyerson

## Correspondence

hli@jimmy.harvard.edu (H.L.), matthew\_meyerson@dfci.harvard.edu (M.M.)

## In brief

Long-read genome sequencing describes the existence and structures of neotelomeres and telomere-spanning chromosome arm fusions in cancer samples. Short-read sequences of 3,651 cancer samples were used to infer the frequency of these telomeric alterations across 40 cancer types.

## Highlights

- Long-read cancer DNA sequences define structure of neotelomeres and telomere fusions
- Neotelomeres have similar telomere length as normal chromosomal arms
- Short telomeric repeats are found at sites of chromosomal arm fusions
- Frequency of neotelomere and arm fusion events varies across 40 cancer types



## Resource

# Neotelomeres and telomere-spanning chromosomal arm fusions in cancer genomes revealed by long-read sequencing

Kar-Tong Tan,<sup>1,2,3</sup> Michael K. Slevin,<sup>1</sup> Mitchell L. Leibowitz,<sup>1,2,3</sup> Max Garrity-Janger,<sup>1,2,3</sup> Jidong Shan,<sup>4</sup> Heng Li,<sup>1,3,\*</sup> and Matthew Meyerson<sup>1,2,3,5,\*</sup>

<sup>1</sup>Dana-Farber Cancer Institute, Boston, MA 02215, USA

<sup>2</sup>Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

<sup>3</sup>Harvard Medical School, Boston, MA 02215, USA

<sup>4</sup>Department of Genetics, Albert Einstein College of Medicine, Bronx, NY, USA

<sup>5</sup>Lead contact

\*Correspondence: [hli@jimmy.harvard.edu](mailto:hli@jimmy.harvard.edu) (H.L.), [matthew\\_meyerson@dfci.harvard.edu](mailto:matthew_meyerson@dfci.harvard.edu) (M.M.)

<https://doi.org/10.1016/j.xgen.2024.100588>

## SUMMARY

Alterations in the structure and location of telomeres are pivotal in cancer genome evolution. Here, we applied both long-read and short-read genome sequencing to assess telomere repeat-containing structures in cancers and cancer cell lines. Using long-read genome sequences that span telomeric repeats, we defined four types of telomere repeat variations in cancer cells: neotelomeres where telomere addition heals chromosome breaks, chromosomal arm fusions spanning telomere repeats, fusions of neotelomeres, and peri-centromeric fusions with adjoined telomere and centromere repeats. These results provide a framework for the systematic study of telomeric repeats in cancer genomes, which could serve as a model for understanding the somatic evolution of other repetitive genomic elements.

## INTRODUCTION

Cancer is driven by alterations to the genome. The development of massively parallel short-read sequencing over the past 15 years has enabled the detailed characterization of somatic and germline variants in tens of thousands of cancer genomes.<sup>1–7</sup> The recent development of linked-read genome sequencing of long molecules with barcoded short-reads then facilitated the characterization of more complex structural variations and of genomic alterations at the haplotype level in cancer.<sup>8–12</sup>

Despite these advances in genome technology, the identification and characterization of somatic alterations at repetitive elements, which constitute roughly one-half of the human genome,<sup>13–15</sup> remain significant challenges. Repetitive elements and duplicated sequences in the human genome are typically 100 to 8,000 bp in size,<sup>15</sup> although centromeres are much longer arrays of repetitive elements.

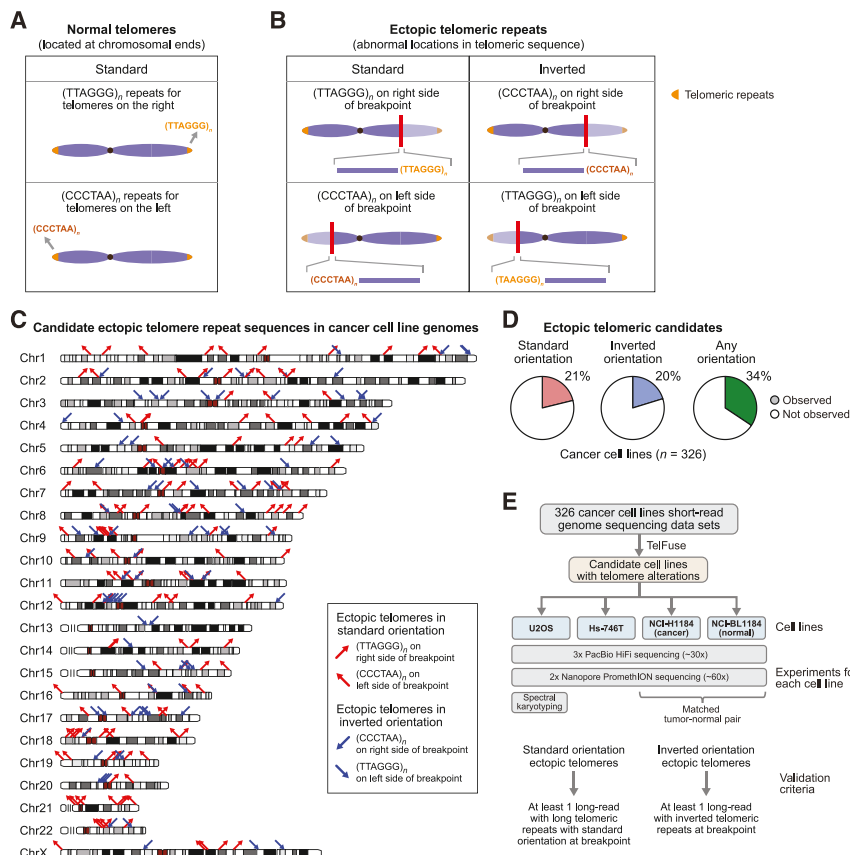
Telomeres cannot be readily resolved by short-read sequencing methodologies. Human telomeres, which act as protective caps on the ends of chromosomes are composed of approximately 2–10 kb (TTAGGG)<sub>n</sub> tandem repeats.<sup>16,17</sup> Somatic integration of telomeric sequences into non-telomeric DNA in tumor samples has been observed,<sup>18</sup> although the origin and structures of these sequences remain unclear. Short-read sequencing, typically 2 × 150 bp paired reads, is unable to fully span the 2- to 10-kb-long highly repetitive telomeres.

The study of telomere structure is important in cancer genomics because telomere maintenance is crucial in cancer pathogenesis. Cancer cell immortality requires a mechanism to activate telomerase or otherwise maintain telomeres and is a key hallmark of cancer.<sup>19,20</sup> Telomerase, the enzyme that adds telomeric repeats to the ends of chromosomes, has been estimated to undergo reactivation in as many as 90% of human cancers and was shown experimentally to be critical for malignant transformation.<sup>21–26</sup> The reactivation of telomerase activity in cancer is driven in part by promoter mutations, amplifications and translocations in the telomerase catalytic subunit gene, *TERT*,<sup>27–30</sup> and also by amplification of the RNA component of telomerase, *TERC*, in cancer.<sup>30,31</sup> In some cancer types, genetic inactivation of the *ATRX* and *DAXX* genes are also associated with telomere elongation, independent of telomerase, by the alternative lengthening of telomeres (ALT) pathway.<sup>32,33</sup>

Long-read genome sequencing now enables the analysis of somatic alterations in highly repetitive regions, such as telomeric repeats, with greater precision and detail. Recently, the first telomere-to-telomere human genome was assembled using PacBio high-fidelity (HiFi) sequencing, with an average read length of 13.5 kb,<sup>34</sup> as well as ultra-long-read nanopore sequencing, which can generate reads of more than 100 kb.<sup>35</sup>

Here, we explored the structure of previously unresolved telomeric events in cancer genomes. We applied TelFuse, a computational method to profile ectopic telomeric repeat sites, to identify candidate telomeric alterations in genomes of 326 cancer cell





**Figure 1. Classes of ectopic telomeric repeats found in cancer cell genomes**

(A) Schematic of sequences and positions of normal telomeres at chromosomal termini.

(B) Schematic of ectopic telomeric repeats. Standard orientation: (TTAGGG)<sub>n</sub> on the right side of a breakpoint and (CCCTAA)<sub>n</sub> on the left side of the breakpoint in the 5' to 3' direction (same as normal telomere in Figure 1A). Inverted orientation: (CCCTAA)<sub>n</sub> on the right side of a breakpoint and (TTAGGG)<sub>n</sub> on the left side of the breakpoint in the 5' to 3' direction. Note that the terminal segment shown in faded color has been deleted or rearranged and is not part of the neotelomere-containing derivative chromosome.

(C) Genome-wide localization of ectopic telomeric repeats in cancer cell line genomes (n = 326) identified using short-read genome sequencing. Red: ectopic telomeric sequences in the standard orientation. Blue: ectopic telomeric sequences in the inverted orientation. Position of telomeric repeats relative to the breakpoint is indicated by arrows oriented in different directions.

(D) Percentage of cancer cell lines in the CCLE with ectopic telomeric sequences in either orientation. Total sample number as indicated.

(E) Flow-chart of long-read genome sequencing and cytogenetic analyses in cancer cell lines, with the indicated validation criteria.

lines and 3,651 cancer samples. Then, we resolved the structures of putative telomeric variants in cancer-derived cell lines using long-read genome sequencing and complementary methods. This analysis defined neotelomeres (a telomere at a new site agnostic to the mechanism by which it arose), telomere-spanning chromosomal arm fusion events, and complex telomeric alterations that were not previously resolvable with short-read genome sequencing. Our study provides a framework that can be applied to the examination of other highly repetitive sequences that are likely to be of biological significance in disease, including centromere arrays, transposable element insertions, and microsatellite repeats.

## RESULTS

### Identification of ectopic telomeric repeat sequences

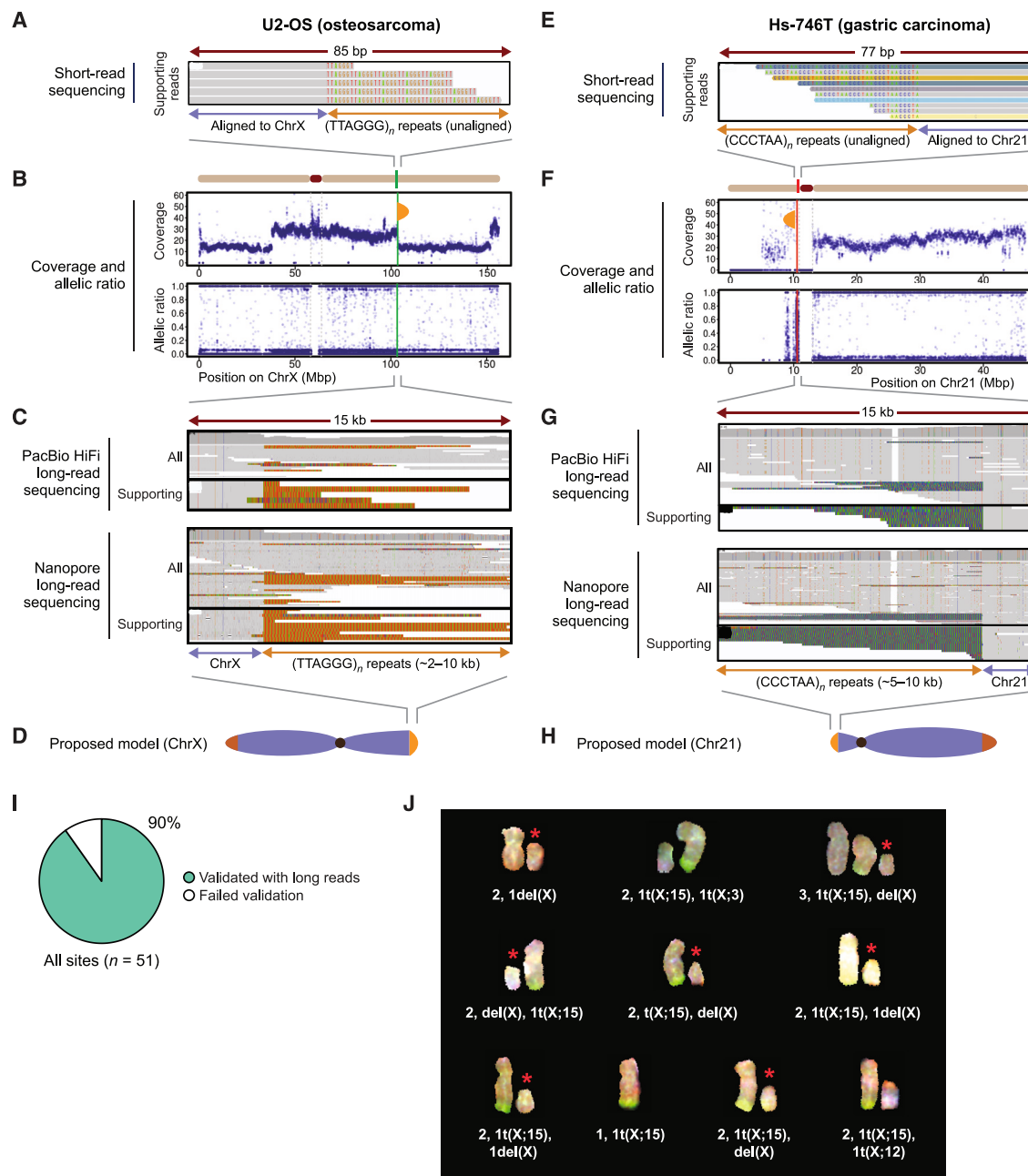
Telomeric repeat arrays within cancer genomes can be found at their original position at chromosomal termini (Figure 1A) or at new positions within the genome (Figure 1B). Telomeric repeats at new genomic locations may be in the same orientation as the original telomeric repeat, with reference to the adjacent chromosomal sequence (i.e., standard orientation) or in an inverted orientation (Figure 1B). Significantly, telomeric repeats oriented in different directions may represent different chromosome structures and may originate via highly distinct biological processes.

We developed an analytic method, TelFuse, to identify ectopic telomeric repeats within the cancer genome, and to estimate

telomere length of each chromosomal arm with long-read sequencing respectively (Figure S1A). TelFuse identifies ectopic telomeric repeat sequences (TTAGGG)<sub>n</sub> or (CCCTAA)<sub>n</sub> that are absent from the germline and can be mapped to intra-chromosomal regions (i.e., ≥ 500 kb from chromosomal ends) (STAR Methods, Figures S1A and S1B). TelFuse begins by identifying read pairs that contain at least two perfect consecutive telomeric repeats (≥ 12 base pairs of telomere sequence) with adjacent sequences that map to intra-chromosomal sites. Paired read sequences that are fully aligned to the reference genome are removed, eliminating telomeric repeats in the reference, which include ancient chromosome fusion events.<sup>36,37</sup> To ensure the specificity of our calls, we also developed a series of filters (Figures S1A and S1B, STAR Methods) to remove spurious sites caused by artifacts induced during the mapping process (Figures S1A and S1B), assessed by a variety of quality control metrics (STAR Methods). Those sites that pass all filters and are at least 500 kb from the GRCh38 reference genome chromosome terminus, a sufficient distance to avoid sub-telomere sequences,<sup>38–40</sup> are considered candidate sites of ectopic telomere sequence.

### Frequency and genome-wide distribution of candidate ectopic telomere repeat sequences in cancer cell lines inferred from short-read genome sequencing

To assess the landscape of ectopic telomere repeats in cancer, we began by applying TelFuse to whole genome sequencing datasets from 326 cancer cell line DNA specimens from the Cancer Cell Line Encyclopedia (CCLE).<sup>7</sup> We detected a total of



**Figure 2. Neotelomeres in cancer genomes revealed by long-read genome sequencing**

(A–H) Genomic analysis of telomere repeat alterations in the standard orientation that were detected (A–D) in U2-OS cells at chrX:103,320,553, and (E–H) in Hs-746T cells at chr21:10,547,397. (A) Integrative Genomics Viewer (IGV) screenshots of short-read genome sequencing data. Ectopic telomeric repeats (TTAGGG)<sub>n</sub> are shown in color. (B) Sequencing coverage and allelic ratios of chromosome X. Orange semi-oval: site of the neotelomeric event. (C) IGV screenshots depicting long telomeric repeat sequences (TTAGGG)<sub>n</sub> with PacBio HiFi and Nanopore long-read sequencing at the site shown in (A). (D) Schematic of neotelomere location on chromosome Xq. (E) IGV screenshots of short-read genome sequencing data. Ectopic telomeric repeats (CCCTAA)<sub>n</sub> are shown in color. (F) Sequencing coverage and allelic ratios of chromosome 21. Orange semi-oval: site of the neotelomeric event. (G) IGV screenshots depicting long telomeric repeat sequences (CCCTAA)<sub>n</sub> with PacBio HiFi and Nanopore long-read sequencing at the site shown in (E). (H) Schematic of neotelomere location on chromosome 21p.

(legend continued on next page)

240 candidate ectopic intra-chromosomal telomeric repeat sequence sites in 34% of cell lines (112/326) (Figures 1C and 1D and Tables S1 and S2). Analysis of telomere repeat orientation identified 149 candidate sites with telomeric repeat sequences in the standard orientation, and 91 candidate sites with telomeric repeat sequences in the reverse orientation (Figures 1C and 1D). An additional 42 candidate sites with telomeric repeat sequences within softclipped sequences (termed non-perfect ectopic telomeric repeat), but not on the first 12 base-pairs, were also detected (Table S3).

### Validation of putative ectopic telomeres by long-read sequencing

We performed high-depth long-read sequencing of selected cell line genomes to validate and extend the analysis of short-read sequences. We selected the U2-OS osteosarcoma cancer cell line, with 55 candidate telomeric repeat sites from short-read sequencing (46 in the standard orientation and 9 in the inverse orientation), the Hs-746T gastric carcinoma cell line with 6 candidate events (1 standard and 5 inverse orientation), and the NCI-H1184 small cell lung cancer cell line with 6 candidate events (5 standard and 1 inverted orientation) (Figures S1C and S1E), together with its matched normal sample (NCI-BL1184). Notably, the U2-OS cell line was found to be highly rearranged, with ectopic telomeric sites found near regions with changes in sequencing coverage and allelic ratios (Figure S2).

We then performed PacBio HiFi and Oxford Nanopore long-read genome sequencing on cell line genomes (Figure 1E). We achieved a median genomic coverage of 49 $\times$ , 62 $\times$ , 65 $\times$ , and 73 $\times$  for the U2OS, Hs-746T, NCI-BL1184 and NCI-H1184 cell lines, respectively, with Nanopore long-read genome sequencing (Figure S3, Tables S4 and S5), median coverage of 19 $\times$ , 20 $\times$ , 19 $\times$ , and 23 $\times$  for the same four cell lines using high quality PacBio HiFi reads, and median coverage of 29 $\times$ , 31 $\times$ , 33 $\times$ , and 36 $\times$  when all PacBio reads were considered (Figure S3, Tables S4 and S5). Nanopore sequencing data had a median read length of 6–13 kb (N50: 18–21 kb), while the PacBio HiFi data had a median read length of 15–17 kb (N50: 16–19 kb) (Figure S3, Tables S4 and S5).

### Benchmarking of TelFuse through simulation

To assess the precision and sensitivity of calls made by TelFuse, we performed simulation experiments. We introduced artificial neotelomeric and chromosomal arm fusion sites into two short-read whole genome sequencing datasets from NA12878 and NCI-BL1184 cell lines<sup>41</sup> at a variant allelic frequency of 30% using the BAMSurgeon software.<sup>42,43</sup> On these synthetically generated datasets, we observed that precision for neotelomere identification was 1.000 in the NA12878 and NCI-BL1184 datasets (Figure S4). Precision of 0.999 and 0.997 was observed for chromosomal arm fusion events in the NA12878 and NCI-BL1184 datasets, respectively (Figure S4). For neotelomeric

sites, we observed a sensitivity of 0.900 using a cutoff of three or more reads in the NA12878 dataset, which was sequenced to a coverage of 96.7 $\times$  (Figure S4). A lower sensitivity of 0.651 was observed in the NCI-BL1184 dataset, which was sequenced only to a coverage of 31.2 $\times$  (Figure S4). Nonetheless, sensitivity in this dataset could be significantly improved to 0.806, while retaining similar precision (0.998) by decreasing the reads cutoff from three or more reads to two or more reads (Figure S4). We observed similar results on the sensitivity of TelFuse for chromosomal arm fusion sites (NA12878: 0.902, BL1184: 0.649) (Figure S4). The precision and sensitivity we observed for TelFuse were similar to some of the best performing structural variant callers<sup>43</sup> and indicate that TelFuse is able to identify neotelomeric and chromosomal arm fusion sites with high precision and good sensitivity on these simulated datasets.

### Neotelomeres in cancer revealed by long-read genome sequencing

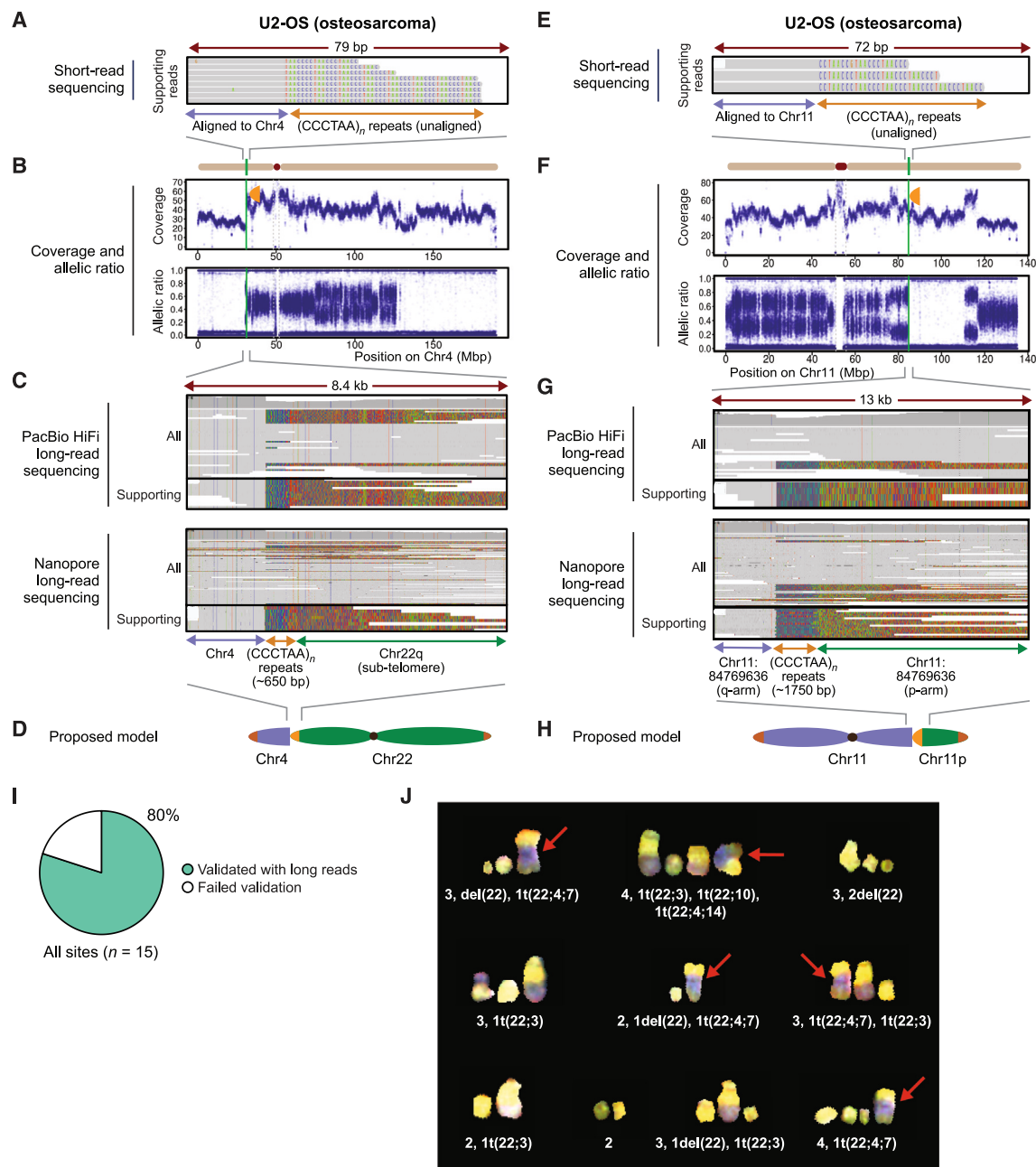
Long-read sequencing analyses demonstrated that ectopic telomere repeat sequences identified by TelFuse in the standard orientation were long and unbounded, consistent with neotelomere addition. For example, a candidate ectopic telomere repeat sequence site, adjacent to sequence from chrX:103,320,553 in U2-OS cells, contains at least seven tandem (TTAGGG)<sub>n</sub> repeats in short-read sequencing data (Figure 2A). Sequencing coverage at this position is decreased, corresponding with the telomeric repeat location (Figure 2B). Analysis of this region by long-read sequencing, both PacBio HiFi and Oxford Nanopore, showed long telomeric repeats of approximately 3–10 kb in the standard orientation (Figure 2C). These data support a model where breakage of the chrXq arm was capped by generation of novel telomeric sequence representing a neotelomere (Figure 2D).

Another example of a neotelomere is seen in the Hs-746T cell line, within chromosome arm 21p at chr21:10,547,397. Short-read sequencing showed at least six tandem (CCCTAA)<sub>n</sub> repeats (Figure 2E). At this location, fluctuation in both sequencing coverage and allelic ratios could be observed (Figure 2F). Analysis of PacBio HiFi and Nanopore long-read genome sequencing data again revealed long telomeric repeats (approximately 5–10 kb) in the standard orientation with reference to the break point at this site (Figure 2G), lending support to the existence of a neotelomere, which had likely formed following breakage of the chr21p arm (Figure 2H). Similar observations were made at other ectopic neotelomeric sites, such as chr7:24,302,169 in U2-OS cells (Figures S5A–S5D), and chr1:214,460,753 in NCI-H1184 cells (Figures S5E–S5H).

In total, 46 sites of ectopic telomere sequence showed long telomere repeats suggestive of neotelomeres based on long-read sequencing, out of 51 sites predicted by short-read sequence analysis to contain standard orientation telomere repeat sequences (Figures 2I and Table S6). No telomeric long

(I) Percentage of ectopic telomeric repeat sites in the standard orientation, found by short-read genome sequencing using TelFuse, that were validated by long-read genome sequencing.

(J) Spectral karyogram of chrX in ten U2-OS single cells assessed by spectral karyotyping with corresponding karyotype labels. First label: total # of X chromosomes and their derivatives observed in given cell. Second label: karyotypes of the aberrant X chromosomes or derivatives. Asterisk (\*): truncated X chromosome. See also Figure S5.



**Figure 3. Chromosomal arm fusions in cancer genomes revealed by long-read genome sequencing**

(A–H) Genomic analysis of telomere repeat alterations in the inverted orientation that were detected in U2-OS cells (A–D) at the site chr4:30,909,846, and (E–H) at the site chr11:84,769,636. (A) Integrative Genomics Viewer (IGV) screenshots of short-read genome sequencing data. Ectopic telomeric repeats (CCCTAA)<sub>n</sub> are shown in color. (B) Sequencing coverage and allelic ratios of chromosome 4. Orange semi-oval: site of the ectopic telomere repeat sequence. (C) IGV screenshots of PacBio HiFi and Nanopore long-read sequencing data at the site shown in (A). Ectopic telomeric repeats in the inverted orientation contained approximately 650 bp of (CCCTAA)<sub>n</sub> telomeric repeat sequences followed by chr22q sub-telomeric sequences, indicative of a chromosomal arm fusion event of chr22q to the site at chr4:30,909,846. (D) Schematic of telomere-spanning fusion event between chromosomes 22q-ter and 4p. (E) IGV screenshots of short-read genome sequencing data. Ectopic telomeric repeats (CCCTAA)<sub>n</sub> are shown in color. (F) Sequencing coverage and allelic ratios of chromosome 11. Orange semi-oval: site of the ectopic telomere repeat sequence. (G) IGV screenshots of PacBio HiFi and Nanopore long-read sequencing at the site shown in (E). Approximately 1,750 bp of (CCCTAA)<sub>n</sub> telomeric repeat sequences are found corresponding to chr11p (chr11:43,002,345), suggestive of a complex event consistent with the formation of a neotelomere on chr11p, followed by a chromosomal arm fusion event of this neotelomere to the site on chr11q (chr11:84,769,636). (H) Schematic telomere-spanning fusion event between chromosome arms 11q (with a predicted neotelomere) and 11p.

(legend continued on next page)

reads could be found at the other five sites. Our results confirm that telomeric repeats in the standard orientation, identified with short-read sequencing, are likely to be neotelomeres with long telomeric repeats.

To assess the relationship between neotelomeres and chromosomal alterations, we performed spectral karyotyping of U2-OS cells (representative cell shown in Figure S6A). Integrative analysis of sequencing coverage, allelic ratios and long-read data inferred two copies of chromosome X in U2-OS cells, one complete copy and one truncated chromosome X. Concordant with a neotelomere detected by long-read genome sequencing data (Figures 2A–2D), a shorter chromosome X with q-arm deletion was observed by spectral karyotyping in 7 of 10 cells assessed (Figure 2J), together with a full-length chromosome X in 10 of 10 cells karyotyped. Thus, spectral karyotyping analysis confirms that neotelomeres identified by long-read sequencing can be correlated with chromosomal truncations observed by cytogenetics.

We also observed a significant level of chromosomal heterogeneity in U2-OS cells (Figures S6B, S6C, and Table S7), including slight variations in chromosome number (number of chromosomes = 76–80) (Table S7) and heterogeneity in translocation events between cells, concordant with a prior study.<sup>27</sup> Specifically, while a t(4; 22) translocation could be observed in 10 of 10 cells assessed (Figure S6C), a t(15; 19) translocation was only observed in 6 of 10 cells assessed. Similar cellular heterogeneity might explain why long-read sequencing was unable to validate 5 of the 51 candidate sites that were detected in the cell populations sequenced by CCLE, despite the high precision of TelFuse.

### Telomere repeat-spanning chromosomal arm fusions in cancer resolved by long-read genome sequencing

We next explored sites with ectopic telomeric repeat sequences in the inverted orientation with respect to the breakpoint. Long-read sequencing revealed that these sites largely represent chromosomal arm fusion events. At one candidate site at position chr4:30,909,846, we observed eight inverted telomeric repeats (CCCTAA)<sub>n</sub> (approximately 48 bp) using short-read sequencing data (Figure 3A). This site includes an additional “C” nucleotide in the first four nucleotides and was, therefore, classified as a non-perfect telomeric repeat site (Table S3). Significant changes in sequencing coverage and allelic ratios were also observed in support of the fusion event (Figure 3B). Analyzing this region with both PacBio HiFi and Nanopore long-read genome sequencing, we observed approximately 650 bp of inverted (CCCTAA)<sub>n</sub> repeats after the breakpoint (Figure 3C), followed by 5–8 kb of sequences on chr22q sub-telomeres (Figure 3C). Individual long-reads that span the telomere insertion suggest that the inverted (CCCTAA)<sub>n</sub> repeat sequences formed via fusion of the chr22q arm with its short telomere to an intra-chromosomal site (Figure 3D).

We also observed more complex fusion events, including evidence for the formation of a neotelomere followed by a subse-

quent chromosomal fusion. At chr11:84,769,636, five inverted ectopic telomeric repeats (CCCTAA)<sub>n</sub> (approximately 30 bp) were detected at the breakpoint with short-read sequencing (Figure 3E). A drastic change in allelic ratios was observed despite minimal changes in copy number estimated from sequencing coverage (Figure 3F), suggesting changes to one of the parental chromosomes despite no overall changes in chromosomal number. Using both PacBio HiFi and Nanopore long-read sequencing data, we observed approximately 1,750 bp of inverted (CCCTAA)<sub>n</sub> telomeric repeats at this site (Figure 3G). Surprisingly, we could further observe more than 5 kb of sequences corresponding with an intra-chromosomal site on the chr11p arm, suggesting that the neotelomere was the consequence of multiple steps. It may have first formed on the centromeric side of the chr11p breakpoint (chr11p:43,002,345), which then subsequently fused to the breakpoint on chr11q at position 84,769,636 (Figure 3H).

To assess telomere-spanning chromosomal fusions in other samples, we examined long-read genome sequencing data of Hs-746T and NCI-H1184 cells. Inverted ectopic telomeric repeats identified from TelFuse were confirmed as sites of chromosomal arm fusion events with long-read data in the Hs-746T sample (Figure S7) at the sites chr11:79,325,679 and chr1:244,201,717, but not for the single candidate site in the NCI-H1184 sample (Table S6). Overall, across 15 inverted telomeric repeat sites predicted by TelFuse in these cell lines, 12 events (80%) were validated as chromosomal arm fusion events using long-read genome sequencing (Figures 3I and Table S6).

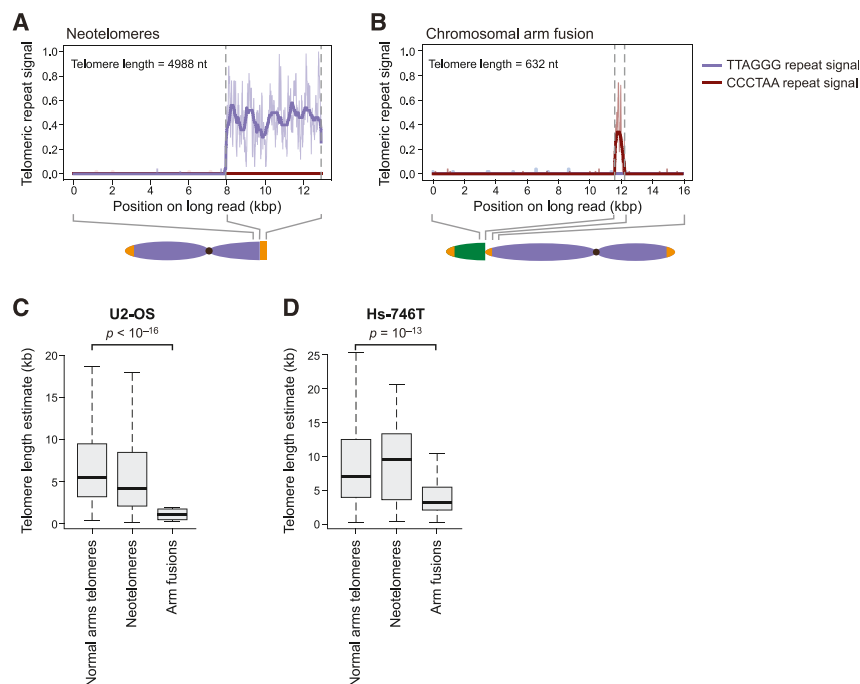
We investigated chromosomal arm fusion events for concordance with spectral karyotyping results of the U2-OS cells. Consistent with the t(4; 22) fusion seen in long-read sequencing (Figure 3A–3D), a fusion between chromosome 22 and chromosome 4 was observed by spectral karyotyping in 5 of 10 cells assessed (Figure 3J). These results suggest that telomere-spanning chromosomal arm fusion events detected by long-read sequencing are concordant with chromosomal scale observations.

### Length distribution of neotelomeres matches that of normal telomeres

To assess telomere length, we developed an approach (TelSize) to estimate the length of telomeric repeats in long-read sequences (STAR Methods) that accounts for noise in telomeric long-reads which are interspersed with errors and/or *bona fide* deviations from the standard “TTAGGG” repeat motif (Figure S8A). Using TelSize, we estimate the length of telomeres on each chromosomal arm and at intra-chromosomal telomeric sites. As the sub-telomeric region of the GRCh38 reference genome has not been fully assembled, we first assessed the reliability of assigning telomeric long reads to their respective arms for the CHM13 cell line for which the genome has been fully assembled (Figure S8B). TelSize was used to generate telomere

(I) Percentage of new telomeric sites in the inverted orientation that were predicted by TelFuse from short-read genome sequencing, and then validated by long-read genome sequencing as telomere-spanning chromosome arm fusion events.

(J) Spectral karyogram of chromosome 22 for which a chromosomal arm fusion was detected with chromosome 4. Ten U2-OS single cells assessed are as indicated. The fusion event between chromosome 22 (yellow) and chromosome 4 (blue) is indicated by a red arrow. See also Figure S7.



**Figure 4. Neotelomeres have similar telomere length distribution as normal telomeres, while telomeric repeats at sites with chromosomal arm fusions are short**

(A and B) Telomeric repeat signal observed at a representative Nanopore read with (A) a neotelomere in U2-OS DNA at chrX:103,320,553 and (B) a chromosomal arm fusion event in U2-OS DNA at chr4:30,909,846. The length of telomeric repeats on each long-read was estimated from these telomeric repeat signal profiles. Aggregated telomeric length of all long-reads at the normal chromosomal arms (p- and q-arms), neotelomeres, and chromosomal arm fusion events in the (C) U2-OS and (D) Hs-746T cell lines. The  $p$  values indicated in the plots were calculated using the two-sided Wilcoxon rank-sum test. Number of telomeric reads for U2-OS cells are:  $n_{\text{normal}} = 1.106$ ,  $n_{\text{neotelomere}} = 531$ ,  $n_{\text{fusion}} = 74$ , and as follows for Hs-746T:  $n_{\text{normal}} = 1.843$ ,  $n_{\text{neotelomere}} = 15$ ,  $n_{\text{fusion}} = 86$ . See also [Figures S10 and S11](#).

length estimates for all cell lines with long-read sequencing data ([Figure S9](#)).

We then assessed telomere length at each neotelomere, at each natural telomere, and at each chromosomal arm fusion event. For example, at a site of neotelomere addition at position chrX:103,320,553 in DNA from U2-OS cells that was described earlier ([Figures 2A–2D](#)), TelSize predicts a telomere length of at least 4,988 bp from a single nanopore read ([Figure 4A](#)). In a site of chromosome arm fusion between positions chr4:30,909,846 and the chr22 telomeric end ([Figures 3A–3D](#)) in DNA from U2-OS cells, TelSize predicts a telomere length of 632 bp from a single nanopore read ([Figure 4B](#)), with intra-chromosomal and sub-telomeric sequences flanking these sites. Most neotelomeres identified were multi-kilobasepair long with an average telomere length of approximately 5 kb in both U2-OS and Hs-746T cancer cell lines ([Figures 4C, 4D, S10A, S10B, S11A, and S11B](#)). Chromosomal fusion events are largely only a few hundred base pairs long in U2-OS but longer in the small number of examples in Hs-746T ([Figures 4C, 4D, S10C, S10D, S11C, and S11D](#)). Overall, we see that telomeres at chromosomal arm fusion events are shorter than neotelomere events or normal functional telomeres (U2-OS  $p < 10^{-16}$ ; Hs-746T  $p < 10^{-13}$ ) ([Figures S11E and S11F](#)).

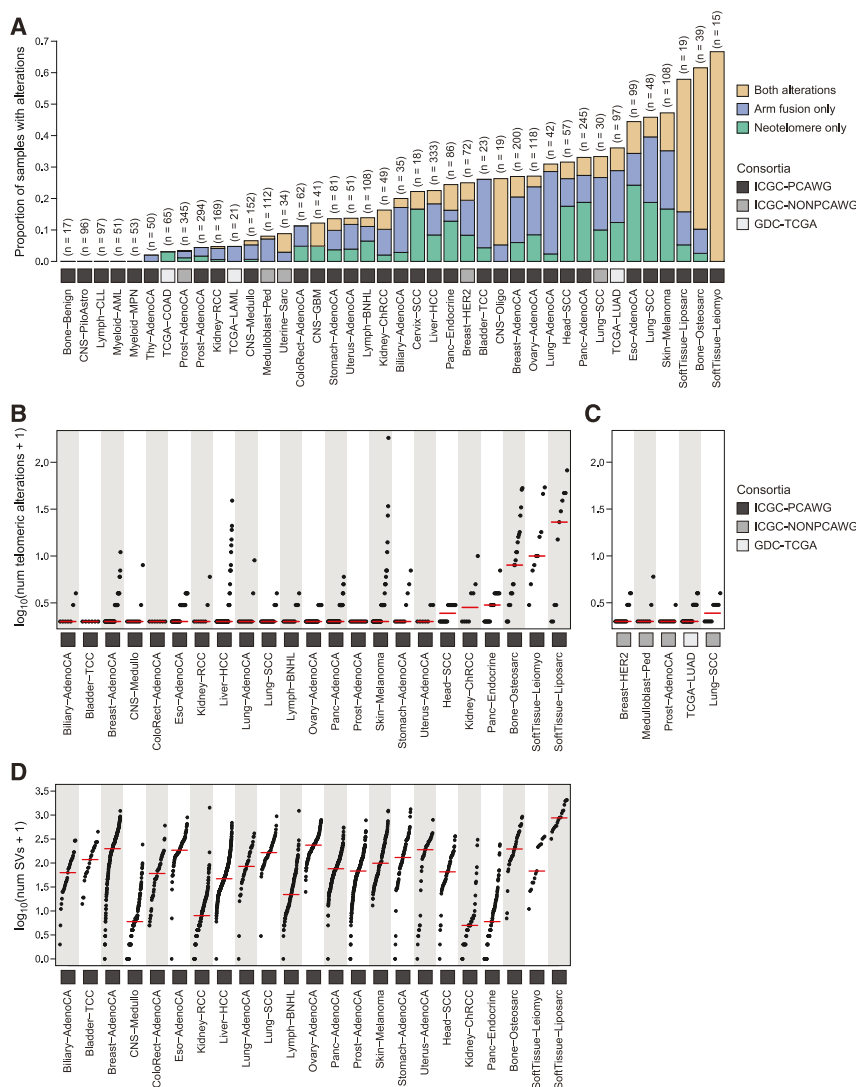
### Somatically altered ectopic telomere repeat sequences in lung adenocarcinoma genomes

Next, we sought to determine whether neotelomere events (standard telomeric repeat orientation) and telomere-spanning chromosome fusion events (inverted telomeric repeat orientation) could be observed as somatic genome alterations in primary human cancers. We applied TelFuse to 95 pairs of lung adenocarcinoma tumor/normal genome sequences from The

Cancer Genome Atlas (TCGA) (TCGA-LUAD) ([Table S8](#)). This analysis identified 34 sites with ectopic telomere sequences in the standard orientation and 46 sites with ectopic telomere sequences in the inverted orientation ([Tables S9 and S10](#)). Putative sites of ectopic telomeric repeat sequences could be seen across the genome on almost all chromosome arms, without a particular distribution at this resolution of sample numbers and events ([Figure S12A](#)). These ectopic telomere sequences, in both the standard and inverted orientations, could be in either the centromeric or counter-centromeric direction ([Figure S12A](#)).

Among the standard orientation ectopic telomere repeats in the TCGA-LUAD sequence data, 31 of the 34 sites were confirmed as somatic alterations and, therefore, as putative somatically generated neotelomeres by comparing the lung adenocarcinoma DNA sequence with the matched normal sequence. In addition, 41 of 46 of the inverted orientation repeats were confirmed as somatic alterations that are likely to represent telomere-spanning chromosomal arm fusions ([Figure S12B](#)). Together, among the set of 80 potential neotelomeres and chromosomal arm fusion events detected in the TCGA-LUAD tumor samples, we found that 72 of the 80 events (90%) were only detected in the tumor sample ([Figure S12B and Table S9](#)), with 8 of the 80 events (10%) detected in both the tumor and normal samples. This finding suggests that a large majority of calls (90%) made in tumor samples by TelFuse are somatic, even though no matched normal samples were assessed in our initial analysis.

Examples of a neotelomere addition in TCGA-44-4112 ([Figure S12C](#)) and a chromosomal arm fusion event in TCGA-49-4507 ([Figure S12D](#)) are shown. Similar observations were also made at other sites ([Figures S13A–S13D](#)) in primary lung adenocarcinoma genomes. Altogether, we observed ectopic telomeric repeats in the standard orientation and inverted orientation in 26% and 31% of the TCGA-LUAD cohort, respectively ([Figure S13E](#)).



**Figure 5. Pan-cancer analysis of somatic neotelomeres and chromosomal arm fusion events**

(A) Proportion of tumor samples with somatic neotelomeres, chromosomal arm fusions, or either type of alteration across each tumor type.

(B and C) Number of somatic ectopic telomeric events in each tumor type. Each dot in the plot represents a single tumor sample. Note that only samples with at least one ectopic telomeric event are presented in this plot. The red line represents the median value.

(D) Number of structural variants detected in each tumor type in PCAWG data. Each dot represents the number of somatic structural variants detected in a tumor sample. The red line represents the median value. Note that data are missing for some tumor types as structural variant calls were unavailable.

6987 at low allelic frequencies in both tumor and the adjacent normal sample (Figure S14E), which may point to contamination of the normal sample or to somatic mosaicism. Together, these results indicate that ectopic telomeric repeats might be frequent germline variants.

### Pan-cancer analysis of neotelomeric and chromosomal arm fusion events

To establish the frequency of neotelomeres and chromosomal arm fusion events in a wider range of cancer types, we applied TelFuse to 3651 pairs of cancer-normal samples encompassing 40 cancer types (2,875 Pan-Cancer Analysis of Whole Genomes [PCAWG], 593 ICGC-non-PCAWG and 183 TCGA samples).<sup>3,44,45</sup> A total of 1,728 somatic ectopic telomeric sites were identified in these

samples (1,209 neotelomere and 519 arm fusion events) (Tables S11, S12, S13, and S14). Our analysis revealed a high fraction of cancers (>50%) containing neotelomeres and chromosomal arm fusion events in cancer types, including leiomyosarcoma, osteosarcoma, liposarcoma, and melanoma (Figure 5A, Tables 15, S16, and S17). Conversely, neotelomeric and chromosomal arm fusion events were rare in blood cancers such as myeloproliferative neoplasms, chronic lymphocytic leukemia, and acute myeloid leukemia (Figures 5A and Tables 15, S16, and S17). Some cancer types (leiomyosarcoma, osteosarcoma, and liposarcoma) with the highest fraction of samples with telomeric alterations, also have high numbers of telomeric alterations per sample (Figures 5B and 5C).

To establish the frequency of somatic and germline alterations in the non-cancer samples in the same cohort, we compared the cancer genomes with the matched normal genomes (Figure S15 and Tables S18, S19, S20, and S21). With this analysis, we showed that putative somatic ectopic neotelomeric variants in normal control samples (0.3% in the PCAWG cohort) are

### Germline variations leading to ectopic telomeric repeat insertions

We also observed right likely germline examples of ectopic telomere repeat sequence alterations across four different individuals in the TCGA-LUAD cohort (Figure S12B and Table S9). Two ectopic telomeric sites were found on the chr12q arm in both blood and tumor samples of TCGA-44-6778 at the sites chr12:54,480,142 and chr12:54,494,011 and were noted to contain a 14-kb deletion, coupled to an insertion of 6× CCCTAA repeat sequences (Figure S14A). In both blood and tumor samples of the same individual at the sites chr12:25,085,740 and chr12:25,085,754 on chr12p, an insertion of 7× CCCTAA repeats was observed in tandem with duplication of a neighboring 14-bp region (Figure S14B). A similar germline deletion event of 13 bp, coupled with the insertion of telomeric repeat sequences, was found in TCGA-62-A470 at chr4:184,711,090 (Figure S14C), while a duplication of 19 bp was coupled to a telomeric repeat insertion at chr6:170,186,789 in TCGA-44-5643 (Figure S14D). Ectopic telomeric repeats could also be observed in TCGA-55-

significantly fewer than somatic variants in cancer samples (19.6% in the PCAWG cohort) (Figure S16). At the same time, only 0.78% of normal samples (23/2,916) were found to contain germline ectopic telomeric sites. Thus, the detected somatic ectopic telomeric sites are 59× more common in cancer genomes relative to putative somatic sites in normal genomes in the PCAWG cohort. Further, somatic ectopic telomeric sites are 25× more common in the cancer genomes compared with germline ectopic telomeric sites. Thus, under the stringent assumption that ectopic telomeric sites in normal samples are all false positives, sites detected in the tumor samples have a false discovery rate of less than 2%. These results indicate that somatic ectopic intra-chromosomal telomeric repeats are extremely uncommon in normal cells and tissues, and that our method has relatively low false-positive rates.

We next sought to assess the potential relationship between the number of genome-wide structural variants and the number of telomeric alterations (Figure S17). On a per tumor-type level, we did not observe an obvious relationship between the number of ectopic telomeric events and the number of structural variants (Figures 5B–5D). However, when analyzing these alterations on a per sample level, a weak correlation is present between the number of structural variants and number of telomeric alterations (Figure S17A).

To assess the relationship between the overall copy number and the presence of ectopic telomeres, we integrated copy number calls from the PCAWG consortium with ectopic telomeric repeat calls generated in this study. Both ploidy and presence of arm fusions are correlated with neotelomere frequency; ploidy status has a stronger correlation with the number of neotelomeres (gradient = 0.5409,  $p < 0.001$ ,  $n = 2704$ ) (Figure S17B) than does the number of arm fusions (gradient = 0.1993,  $p < 0.001$ ,  $n = 2704$ ) (Figure S17B). The mean number of ectopic telomeres was significantly higher in samples with whole genome duplications (WGDs), with 5.2×, 3.4×, and 4.3× as many neotelomeres, arm fusions, and total ectopic telomeric sites, respectively, in samples with WGDs versus those without WGDs (Figure S17C).

We next sought to assess whether the formation of neotelomeres and chromosomal arm fusion may be related to the telomere maintenance mechanisms involved. Samples were categorized by *ATRX/DAXX* mutations (indicative of ALT positivity) and *TERT* promoter mutations as per a prior study.<sup>18</sup> Samples with *ATRX/DAXX* mutations showed increased neotelomeres, arm fusion, and total telomeric events compared with unmutated samples (Figure S17D). However, no significant differences were found between samples with and without *TERT* promoter mutations. These results, therefore, indicate that the occurrence of ectopic telomere events may be related to ALT status.

This previous analysis only assessed the mutation status of the *TERT* promoter, but *TERT* overexpression in cancers can be driven by other factors as well.<sup>30</sup> Therefore, we generated a more comprehensive analysis of *TERT*-positive cancers by analyzing RNA sequencing (RNA-seq) expression data generated by the PCAWG consortium.<sup>3</sup> Samples without *TERT* expression had a greater mean number of ectopic telomeric events, which is probably related to the greater number of events typically identified in ALT<sup>+</sup> but *TERT*<sup>−</sup> samples (Figure S18A). We

then assessed the proportion of samples with *TERT* expression based on the presence or absence of ectopic telomeric events. *TERT* expression with reads per kilobase per million mapped reads (RPKM) of more than one was detected in 12.8% of samples with neotelomeres observed, compared with approximately 7.9% of total samples ( $p = 0.0602$ ) (Figure S18B). Comparing the distribution of *TERT* RNA expression, we see that samples with ectopic telomeres tended to have a more right-skewed distribution associated with higher *TERT* expression (Kolmogorov-Smirnov  $p = 5.43 \times 10^{-6}$ ) (Figure S18C). Using a less stringent RNA-seq expression cutoff of an RPKM of more than 0, we also observe *TERT* expression in 106 of 125 samples (84.8%) with neotelomeres, and 130 of 151 sample (86.1%) with chromosomal arm fusion events, and 704 of 795 sample (88.6%) with no telomere events detected. Given the observation of *TERT* expression in many cancer samples with neotelomeres and chromosomal arm fusion events, these data suggest that our results extend to telomerase-positive cancers.

We next assessed the clonality of telomeric events in tumors. Across 1,307 events evaluated, we see that ectopic telomere events have a mean variant allele frequency (VAF) of 22.4% and a median VAF of 19.4% (Figure S18D and Table S22). Correcting for tumor purity and ploidy, a mean VAF of 26.2% and a median VAF of 20.6% was observed (Figure S18E and Table S22), indicating that these events are present in approximately 40%–50% of cells in the average tumor sample. In addition, to assess if the clonality of the events differs between ALT-positive and telomerase positive samples, we compared samples with *ATRX/DAXX* mutations (i.e., ALT<sup>+</sup>) versus those with *TERT* promoter mutations (i.e., telomerase positive). Significantly, we see that the VAF values are quite similar between ALT<sup>+</sup>, telomerase-positive samples, and samples without either class of mutations (median VAF = 18.9%, 20.4%, and 19.0%, respectively) (Figure S18F). Correcting for tumor and ploidy, a lower VAF value was observed for ALT<sup>+</sup> samples (median VAF = 11.5%) as compared with samples without either class of mutations ( $p < 10^{-16}$ , median VAF = 22.9%) (Figure S18G). However, we did not observe a statistically significant difference between ALT<sup>+</sup> samples (median VAF = 11.5%) and samples with *TERT* promoter mutations ( $p = 0.443$ , median VAF = 23.2%) possibly due to the small number of samples with *TERT* promoter mutations ( $n = 14$ ). This result, therefore, indicates that the ectopic telomeric events that we have found in our study are true events at high levels of cellular clonality.

### Sequence signatures at neotelomeric and chromosomal arm fusion sites

To ascertain potential processes involved in the formation of neotelomeric and chromosomal arm fusion events, we also further examined sequence composition and structure around neotelomeres and chromosomal arm fusion events (Figures S19 and S20). These analyses revealed an approximately 2-fold enrichment of G and a approximately 2-fold depletion of C nucleotides in the region retained (<5 nt upstream of breakpoint) after formation of neotelomeres (Figure S19A). At the same time, an enrichment of C and depletion of G nucleotides was identified at chromosomal arm fusion sites around the retained region of the breakpoint (Figure S19B). Further

evaluation of these sites suggests that these signals were likely observed due to the presence of microhomologies between these sites and telomeric repeat sequences for neotelomeric sites (mean = 1.72 nt) (Figures S19C and S19D) and chromosomal arm fusion sites (mean = 1.63 nt) (Figures S19E and S19F).

It was previously reported that a telomere seed sequence was sufficient for direct addition of telomeric repeats at a telomere healed end.<sup>46</sup> Notably, complementarity between the first two to four nucleotides of the RNA component of the telomerase RNA template sequence (3'-CAAUCCCAAUC-5' in human) with a single-stranded DNA was found to be sufficient for telomerase to directly elongate and synthesize telomeric repeats to generate a neotelomere.<sup>46</sup> Thus, if neotelomeres in our study were generated directly by telomerase under this model, then (TAGGGT)<sub>n</sub>, (AGGGTT)<sub>n</sub>, and (GGGTTA)<sub>n</sub> sequences would be directly observed at the start of these neotelomeric sites (Figures S20A and S20B). While these repeats were observed immediately after the presume DNA break site with two to four nucleotides of microhomology, respectively (Figures S20C, S20D, and Table S23), these sites that match the telomere seed sequence only constitute approximately 10% of all neotelomeric sites (Figure S20E and Table S23). Our results in cancer samples are, therefore, not consistent with the model for neotelomere synthesis by telomerase as put forth by Morin.<sup>46</sup> That said, it is important to note that our results here do not preclude other models for neotelomere synthesis by telomerase (e.g., a model in which strict complementarity in the first two to four nucleotides between telomerase template RNA and the DNA break is not strictly required).

### Neotelomeres and chromosomal arm fusion events disrupt protein coding genes and are highly prevalent in cancer cell lines

In addition to allowing chromosome fusions to occur and capping truncated chromosomes, the insertion of telomeric DNA might also disrupt genes, including tumor suppressors. To assess this possibility, we evaluated ectopic telomere sites for overlap with protein coding genes. Among sites that we detected, 47% of sites were found to colocalize to a protein coding gene in cancer cell lines (112/240) or primary lung adenocarcinomas (34/72) (Tables S2 and S9). Notable genes with insertion events include *PTPN2*, a gene related to immunotherapy response,<sup>47</sup> and the *NRDC*, *FOXN3*, and *RUNX3* genes (Figure S21). The disruption of protein coding genes is also observed in primary lung adenocarcinoma samples (Figures S22A–S22D). However, we did not observe significant changes in the expression of the putatively disrupted genes in the PCAWG cohort (Figure S22E). Our results, therefore, indicate that the formation of neotelomeres and telomere-spanning chromosomal arm fusions may represent a mechanism for gene disruption, in addition to their roles in defining gross chromosomal structure, but do not reveal a functional role for such disruption.

## DISCUSSION

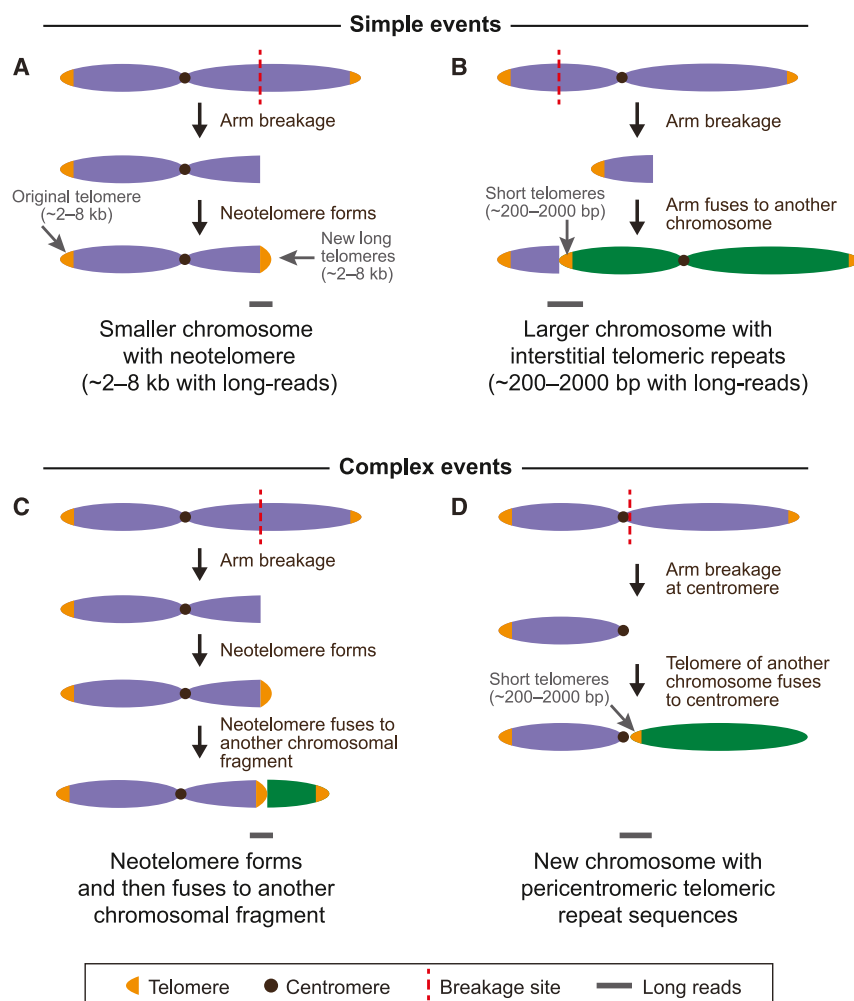
While alterations in telomere sequences are key events in cancer genome evolution, the precise nucleotide-level structure of these alterations has been hitherto inaccessible because of the

inability of short-read sequence data to resolve longer repetitive sequences. Here, using long-read sequencing technologies, we delineated four types of alterations in telomere repeat sequences. First, we provide evidence that cancer cell line and primary cancer genomes contain long (several kilobase) additions of telomere repeat sequences to intra-chromosomal sites, in the standard telomere orientation (Figure 6A). Second, we identify telomeric repeat sequences of varying length that bridge the end of one chromosome to an intra-chromosomal site on a different chromosome (Figure 6B). These telomeric repeats are consistent with karyotyping analyses that have observed the attachment of chromosomal fragments to the ends of existing chromosomes,<sup>48–53</sup> which are key events in cancer genome evolution. Third, we observe more complex alterations where the formation of a neotelomere is followed by the fusion of the neotelomere to a second intra-chromosomal location (Figure 6C). Fourth, we observe fusions that link centromeric to telomeric sequence repeats (Figure 6D). The implications of several of these alterations are described below.

A previous study, analyzing short-read genome sequencing of patients' cancer samples from the PCAWG project, was able to identify a number of intra-chromosomal telomeric repeat insertion sites.<sup>18</sup> In comparison with this previous study, our work shows that telomere length at these repeat insertion sites can be estimated using long-read sequencing, and the underlying sequence structure can be analyzed in the context of adjacent sequences compared with free telomeric ends. This technical advance allowed us to differentiate intra-chromosomal telomeric repeat sites based on the orientation of the telomeric repeat sequences. Our study provides support for the existence of neotelomeres and chromosomal arm fusion events in cancer genomes.

A recent study by Muyas et al.<sup>54</sup> also profiled somatic telomere fusion events across more than 30 cancer types, leading to the identification of novel patterns of telomere fusions that the authors linked to ALT pathway activity. The ectopic telomeric alterations assessed in our study are distinct from the telomere fusion events assessed by Muyas et al.<sup>54</sup> Specifically, the study by Muyas et al. studied (a) inward telomere fusions of (TTAGGG)<sub>n</sub> - (CCCTAA)<sub>n</sub> and (b) outward telomere fusions of (CCCTAA)<sub>n</sub> - (TTAGGG)<sub>n</sub>. Conversely, our study focused on the study of (i) neotelomeres, which were identified by ectopic telomeric repeats in the standard orientation at intra-chromosomal sites [i.e., (intra-chromosomal sequence) - (TTAGGG)<sub>n</sub>] and (ii) chromosomal arm fusions that were identified by ectopic telomeric repeats in the inverted orientation at intra-chromosomal sites [i.e., (intra-chromosomal sequence) - (CCCTAA)<sub>n</sub>]. In addition, the alterations in the Muyas study are telomere-to-telomere, while in our study, the alterations are between telomeres and non-repetitive autosomal sequences.

Although our study strongly supports the presence of new telomeres in human cancer cell lines and samples, the precise mechanism behind their formation remains uncertain. Based on prior studies in other organisms, at least two processes could generate these new telomeres in human cells. First, new telomeres could be generated through a telomerase-mediated process termed telomere healing.<sup>55–57</sup> For instance, telomerase can add new telomeric repeats to double-stranded breaks in



**Figure 6. Possible models for the different types of telomeric repeat sequences observed in this study**

(A) A neotelomere can form after a chromosomal arm breakage event. This leads to the generation of a smaller chromosome with a neotelomere, similar in repeat length to telomeres found on a normal chromosomal arm.

(B) Chromosome arm fusion where a broken chromosomal arm can fuse to another chromosome with very short telomeres. This generates a larger chromosome with interstitial telomeric repeat sequences in the middle of the chromosome.

(C) Complex alteration where neotelomere formation is followed by the fusion of this neotelomere to another chromosomal fragment. This leads to the observation of long-reads in our study which contains telomeric repeat sequences, flanked on both sides by intra-chromosomal sequences.

(D) A complex telomeric alteration involving a chromosomal arm break at or very near to the centromere, which is fused to another chromosomal arm with very short telomeres. The resultant new chromosome has pericentromeric telomeric repeat sequences. Purple line: parts of the model supported by long-read genome sequencing data.

mology sequences with telomeric DNA (0–3 nt). A subset of sites we identified (22%) are also characterized by short (approximately 1–10 nt) non-telomeric sequences, which may be indicative of non-templated insertions. The observation of short non-telomeric sequences and/or short microhomology sequences at these new telomeric sites are characteristic of a process mediated by non-homologous

end-joining, suggesting that our results may be more concordant with a DNA recombination-driven telomere capture model, or, alternatively, a more error prone process for neotelomere formation by telomerase.

Our study reveals a potential connection between ALT and the occurrence of neotelomeres and chromosomal arm fusion events. Samples with *ATRX/DAXX* mutations, indicative of ALT positivity, showed increased neotelomeric and chromosomal arm fusion events, consistent with a prior study on telomere insertions.<sup>18</sup> ALT can contribute to these events through various pathways, such as the abundance of extra-chromosomal telomeric circles,<sup>66,67</sup> which can recombine to form a neotelomere. Also, frequent telomeric recombination between sister chromatids<sup>68</sup> in ALT samples could facilitate the formation of these events.

Our results also point to the existence of an additional non-ALT-related pathway that is involved in the formation of ectopic telomeric events, because ALT alone cannot fully account for all ectopic telomeric events identified in our study. Specifically, it is important to note that samples with *ATRX/DAXX* mutations constitute less than 2% of the samples analyzed, significantly fewer than samples with neotelomeric and chromosomal arm fusion events (approximately 19.5%). Moreover, ALT is rarely

*Saccharomyces cerevisiae*,<sup>58–62</sup> at or near sequences with homology to telomerase template RNA.<sup>58</sup> In humans, a recent experimental study observed the addition of neotelomeres at a subset of induced double-stranded breaks.<sup>63</sup> Second, new telomeres could be generated through DNA recombination, in a telomerase-independent manner termed telomere capture, for example, by translocation of a non-homologous chromosome to a chromosome with terminal deletion leading to the introduction of a new telomeric cap,<sup>64</sup> or by the stabilization of a broken chromosome through the stabilization of a newly captured telomere.<sup>65</sup>

Previous biochemical assays reported that two to four nucleotides of complementarity with the telomerase RNA template at the 3' termini is sufficient for telomerase to recognize these sites to add (TTAGGG)<sub>n</sub> repeats to generate a neotelomere.<sup>46</sup> However, the majority of neotelomeric sites identified in our study do not display this two- to four-nucleotide complementarity signal, suggesting that the majority of neotelomeres identified in our study do not proceed via the telomerase-driven and perfect complementarity model proposed by Morin.<sup>46</sup> In contrast, at sites of both new telomeres and arm fusions in primary cancer samples, we observed relatively short microho-

observed in epithelial malignancies like those involving lung carcinoma.<sup>69,70</sup> Despite the lack of reported ALT in lung carcinoma, we see that these ectopic telomeric alterations were observed in more than 30% of lung squamous cell carcinoma and lung adenocarcinoma (Figure 5A). Thus, we believe that neotelomeres and chromosomal arm fusion events are unlikely to be exclusive to ALT-positive samples.

The generation of new chromosomes via chromosomal rearrangements is a key element of cancer genome evolution and also occurs during the course of evolution and speciation.<sup>71–73</sup> Some of our findings using long-read sequencing of cancer genomes mirror long-standing observations in genomes of many organisms. Interstitial telomeric repeats have been identified in the genomes of many vertebrates, including primates and the pygmy tree shrew,<sup>74–76</sup> akin to those found at sites of chromosomal arm fusions in cancer cell lines (Figure 6B). Furthermore, interstitial telomeric sequences have been observed close to centromeres in the genomes of diverse organisms including Chinese hamster, *Arabidopsis*, and the European grayling.<sup>76–78</sup> These structures, termed pericentromeric telomeric repeats, were similarly observed by long-read genome sequencing in U2-OS cells in our study (Figure 6D). Interestingly, Turkalo et al.<sup>79</sup> had performed an analysis of U2-OS cells by Micro-C analysis (a method akin to Hi-C) and identified contacts involving terminal chromosomal regions. Their results are, therefore, consistent with the chromosomal arm fusion events we observed in U2-OS cells. Additionally, a competitive balance between *de novo* telomere addition (telomere healing) and homologous recombination has been observed to stabilize chromosome ends following a double-stranded break in *Plasmodium falciparum*.<sup>80</sup> Notably, the length of telomeric repeats was also found to increase more in healed telomeres as compared with non-healed telomeres after irradiation.<sup>81</sup> Overall, the study of telomere repeat alterations also provide an understanding into how new chromosomes originate during the course of evolution and speciation, as well as during cancer genome evolution.

Looking at the genome beyond telomeric repeats, repetitive elements constitute approximately one-half of the human genome.<sup>13–15</sup> Here, using telomeres as a salient example, we show how long-read genome sequencing can be used to drive discoveries of functional importance in highly repetitive regions of the cancer genome. Our study provides a framework to assess short-read genome sequencing data for genome alterations within highly repetitive regions, that can be followed by long-read sequencing and complete analysis of selected samples. Significantly, given that more than 95% of repetitive sequences in the genome are estimated to be less than 8 kb in length,<sup>15</sup> long-read sequencing data that is typically generated at more than 10 kb in length (Figure S4) would enable the majority of previously neglected alterations in the cancer genome to be completely resolved. More broadly, analysis of repetitive elements in cancer genomes is a great discovery opportunity provided by long-read cancer genome sequencing.

### Limitations of the study

There are a few limitations associated with our study. First, in contrast with a recent yeast genomic study in which the end of each telomere was tagged,<sup>82</sup> it is difficult to assess whether te-

lomeric repeats containing long-reads analyzed in our study captured the telomeres end-to-end. As such, telomere length estimates in our study may underestimate the true length of both normal telomeres and neotelomeres. Further, it is also known that sub-telomeres at normal chromosomal arms contain telomere-like sequences and short internal telomeric repeats close to long stretches of perfect (TTAGGG)<sub>n</sub> repeats.<sup>38,83</sup> However, it is unclear if these sequences should be included in the computation of telomere length estimates; as such, we have not included them in our length estimates. Additionally, long-read sequencing was only performed on cancer cell lines in our study due to difficulties in acquiring DNA of cancer samples analyzed by short-read data. As such, it is possible that the observations we have made on neotelomeres and chromosomal arm fusion events from cancer cell lines may not fully extend to primary cancer samples. The U2-OS cell line, which was selected for long-read sequencing, constitutes 43 of 46 (93%) of all the neotelomeres observed and 7 of 12 (58%) of all the chromosomal arm fusion events, is known to maintain its telomeres by the ALT pathway<sup>84,85</sup>; it is, therefore, possible that a significant number of events validated by long-read sequencing may have characteristics that are limited to ALT samples. These results may, therefore, not extend fully to samples that maintain their telomeres by non-ALT pathways.

### Conclusion

We have used long-read sequencing to demonstrate the generation of neotelomeres, and of chromosome arm fusions that span telomere repeats, in human cancer cell lines. We then provided evidence for these alterations in primary human cancer genomes. This study provides detailed insight into the process of telomere maintenance in human cancer. Further long-read sequencing studies of cancer genomes could help to elucidate the potential role of somatic alterations in highly repetitive regions of the human genome in cancer pathogenesis. More broadly, long-read sequencing analyses may also provide insights into chromosomal rearrangements that drive genetic diseases and evolution.

### STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Cell lines
  - Human participants
- METHOD DETAILS
  - CCLE whole genome sequencing dataset
  - TCGA whole genome sequencing dataset
  - PCAWG and ICGC-NON-PCAWG whole genome sequencing datasets
  - Annotations for PCAWG datasets
  - Identification of candidate new telomeres and chromosomal arm fusion events from short reads
  - High molecular weight DNA extraction

- MinION library preparation
- PromethION library preparation
- PacBio HiFi library preparation
- Base calling of nanopore sequencing data
- Extraction of candidate telomeric long reads for detailed analysis by TelSize
- Estimation of telomere length from noisy long reads
- Benchmarking of TelFuse by simulation experiments
- Analysis of telomeric repeat length at neotelomeres and chromosomal arm fusion sites
- Analysis of telomeric repeat length at normal chromosomal arms
- Copy number profiles
- Analysis of BAF
- SV variant allele frequency
- Sequence signatures at sites with new telomeres and chromosomal arm fusion events
- Spectral Karyotyping
- **QUANTIFICATION AND STATISTICAL ANALYSIS**

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.xgen.2024.100588>.

## ACKNOWLEDGMENTS

We thank all members of the Meyerson and Li labs for helpful comments and input on the work. We further thank Jodi Hirschman for edits to the manuscript and Leslie Gaffney for assistance in figure preparation.

K.T.T. was supported by a PhRMA Foundation Informatics Fellowship and a NUS Development Grant from the National University of Singapore. M.L.L. is supported by the National Cancer Institute (Grant No. F32CA281185). M.M. is supported by an American Cancer Society Research Professorship. This work was supported by grants from the National Cancer Institute (Grant No. R35 CA197568 to M.M.) and the National Human Genome Research Institute (NHGRI) (Grant Nos. R01 HG010040, U01 HG010961, and U41 HG010972 to H.L.).

## AUTHOR CONTRIBUTIONS

K.T.T. and M.M. initiated the study of telomeres in cancer with long-read genome sequencing. K.T.T. developed computational methods and designed computational analyses with input from H.L. and M.M. K.T.T. performed most computational analyses in this study. M.G.J. assisted with computational analysis of TCGA-LUAD dataset and long-read sequencing data. M.S. generated DNA samples of cancer cell lines used for long-read genome sequencing and performed an initial long-read sequencing run. J.S. generated spectral karyotyping results and provided input on the analysis of cytogenetics data. K.T.T. wrote the initial draft of the manuscript with input from M.M., M.L.L., and H.L. M.M. and H.L. jointly supervised the work. All authors read, revised, and approved the submission of the manuscript.

## DECLARATION OF INTERESTS

M.M. is a consultant for DelveBio, Isabl, and Bayer; receives research support from Bayer and Janssen; has patents for EGFR mutations for lung cancer diagnosis issued, licensed, and with royalties paid from LabCorp and has issued patents and patents pending licensed to Bayer; and was a founding advisor of, consultant to, and equity holder in Foundation Medicine, shares of which were sold to Roche. H.L. is a consultant of Integrated DNA Technologies and is on the Scientific Advisory Boards of Sentieon and Innozeen.

Received: December 14, 2022

Revised: November 9, 2023

Accepted: May 30, 2024

Published: June 24, 2024

## REFERENCES

1. Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018). Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* 173, 371–385.e18. <https://doi.org/10.1016/j.cell.2018.02.060>.
2. Huang, K.-L., Mashl, R.J., Wu, Y., Ritter, D.I., Wang, J., Oh, C., Paczkowska, M., Reynolds, S., Wyczalkowski, M.A., Oak, N., et al. (2018). Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* 173, 355–370.e14. <https://doi.org/10.1016/j.cell.2018.03.039>.
3. Campbell, P.J., Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L., Stein, L.D., Perry, M.D., Nahal-Boise, H.K., Ouellette, B.F.F., Li, C.H., et al. (2020). Pan-cancer analysis of whole genomes. *Nature* 578, 82–93. <https://doi.org/10.1038/s41586-020-1969-6>.
4. Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. *Nature* 575, 210–216. <https://doi.org/10.1038/s41586-019-1689-y>.
5. Degasperi, A., Zou, X., Amarante, T.D., Martinez-Martinez, A., Koh, G.C.C., Dias, J.M.L., Heskin, L., Chmelova, L., Rinaldi, G., Wang, V.Y.W., et al. (2022). Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. *Science* 376, abt9283. <https://doi.org/10.1126/science.abt9283>.
6. Imielinski, M., Berger, A.H., Hammerman, P.S., Hernandez, B., Pugh, T.J., Hodis, E., Cho, J., Suh, J., Capelletti, M., Sivachenko, A., et al. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 150, 1107–1120. <https://doi.org/10.1016/j.cell.2012.08.029>.
7. Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G.V., Lo, C.C., McDonald, E.R., Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* 569, 503–508. <https://doi.org/10.1038/s41586-019-1186-3>.
8. Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M., Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., et al. (2016). Haplotype germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* 34, 303–311. <https://doi.org/10.1038/nbt.3432>.
9. Xia, L.C., Bell, J.M., Wood-Bouwens, C., Chen, J.J., Zhang, N.R., and Ji, H.P. (2018). Identification of large rearrangements in cancer genomes with barcode linked reads. *Nucleic Acids Res.* 46, e19. <https://doi.org/10.1093/nar/gkx1193>.
10. Greer, S.U., Nadauld, L.D., Lau, B.T., Chen, J., Wood-Bouwens, C., Ford, J.M., Kuo, C.J., and Ji, H.P. (2017). Linked read sequencing resolves complex genomic rearrangements in gastric cancer metastases. *Genome Med.* 9, 57. <https://doi.org/10.1186/s13073-017-0447-8>.
11. Viswanathan, S.R., Ha, G., Hoff, A.M., Wala, J.A., Carrot-Zhang, J., Whelan, C.W., Haradhvala, N.J., Freeman, S.S., Reed, S.C., Rhoades, J., et al. (2018). Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by Linked-Read Genome Sequencing. *Cell* 174, 433–447.e19. <https://doi.org/10.1016/j.cell.2018.05.036>.
12. Tan, K.-T., Kim, H., Carrot-Zhang, J., Zhang, Y., Kim, W.J., Kugener, G., Wala, J.A., Howard, T.P., Chi, Y.-Y., Beroukhi, R., et al. (2021). Haplotype-resolved germline and somatic alterations in renal medullary carcinomas. *Genome Med.* 13, 114. <https://doi.org/10.1186/s13073-021-00929-4>.
13. Schmid, C.W., and Deininger, P.L. (1975). Sequence organization of the human genome. *Cell* 6, 345–358. [https://doi.org/10.1016/0092-8674\(75\)90184-1](https://doi.org/10.1016/0092-8674(75)90184-1).
14. Batzer, M.A., and Deininger, P.L. (2002). Alu repeats and human genomic diversity. *Nat. Rev. Genet.* 3, 370–379. <https://doi.org/10.1038/nrg798>.

15. Treangen, T.J., and Salzberg, S.L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 13, 36–46. <https://doi.org/10.1038/nrg3117>.
16. Factor-Litvak, P., Susser, E., Kezios, K., McKeague, I., Kark, J.D., Hoffman, M., Kimura, M., Wapner, R., and Aviv, A. (2016). Leukocyte Telomere Length in Newborns: Implications for the Role of Telomeres in Human Disease. *Pediatrics* 137, e20153927. <https://doi.org/10.1542/peds.2015-3927>.
17. Canela, A., Vera, E., Klatt, P., and Blasco, M.A. (2007). High-throughput telomere length quantification by FISH and its application to human population studies. *Proc. Natl. Acad. Sci. USA* 104, 5300–5305. <https://doi.org/10.1073/pnas.0609367104>.
18. Sieverling, L., Hong, C., Koser, S.D., Ginsbach, P., Kleinheinz, K., Hutter, B., Braun, D.M., Cortés-Ciriano, I., Xi, R., Kabbe, R., et al. (2020). Genomic footprints of activated telomere maintenance mechanisms in cancer. *Nat. Commun.* 11, 733. <https://doi.org/10.1038/s41467-019-13824-9>.
19. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: The next generation. *Cell* 144, 646–674. <https://doi.org/10.1016/j.cell.2011.02.013>.
20. Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discov.* 12, 31–46. <https://doi.org/10.1158/2159-8290.CD-21-1059>.
21. Li, Y., and Tergaonkar, V. (2014). Noncanonical functions of telomerase: implications in telomerase-targeted cancer therapies. *Cancer Res.* 74, 1639–1644. <https://doi.org/10.1158/0008-5472.CAN-13-3568>.
22. Kim, N.W., Piatyszek, M.A., Prowse, K.R., Harley, C.B., West, M.D., Ho, P.L., Coviello, G.M., Wright, W.E., Weinrich, S.L., and Shay, J.W. (1994). Specific Association of Human Telomerase Activity with Immortal Cells and Cancer. *Science* 266, 2011–2015. <https://doi.org/10.1126/science.7605428>.
23. Meyerson, M., Counter, C.M., Eaton, E.N., Ellisen, L.W., Steiner, P., Caddle, S.D., Ziaugra, L., Beijersbergen, R.L., Davidoff, M.J., Liu, Q., et al. (1997). hEST2, the Putative Human Telomerase Catalytic Subunit Gene, Is Up-Regulated in Tumor Cells and during Immortalization. *Cell* 90, 785–795. [https://doi.org/10.1016/S0092-8674\(00\)80538-3](https://doi.org/10.1016/S0092-8674(00)80538-3).
24. Kolquist, K.A., Ellisen, L.W., Counter, C.M., Meyerson, M., Tan, L.K., Weinberg, R.A., Haber, D.A., and Gerald, W.L. (1998). Expression of TERT in early premalignant lesions and a subset of cells in normal tissues. *Nat. Genet.* 19, 182–186. <https://doi.org/10.1038/554>.
25. Li, Y., and Tergaonkar, V. (2016). Telomerase reactivation in cancers: Mechanisms that govern transcriptional activation of the wild-type vs. mutant *TERT* promoters. *Transcription* 7, 44–49. <https://doi.org/10.1080/21541264.2016.1160173>.
26. Yuan, X., Larsson, C., and Xu, D. (2019). Mechanisms underlying the activation of TERT transcription and telomerase activity in human cancer: old actors and new players. *Oncogene* 38, 6172–6183. <https://doi.org/10.1038/s41388-019-0872-9>.
27. Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G.V., Chin, L., and Garraway, L.A. (2013). Highly recurrent TERT promoter mutations in human melanoma. *Science* 339, 957–959. <https://doi.org/10.1126/science.1229259>.
28. Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S., Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in familial and sporadic melanoma. *Science* 339, 959–961. <https://doi.org/10.1126/science.1230062>.
29. Killela, P.J., Reitman, Z.J., Jiao, Y., Bettgowda, C., Agrawal, N., Diaz, L.A., Friedman, A.H., Friedman, H., Gallia, G.L., Giovannella, B.C., et al. (2013). TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc. Natl. Acad. Sci. USA* 110, 6021–6026. <https://doi.org/10.1073/pnas.1303607110>.
30. Barthel, F.P., Wei, W., Tang, M., Martinez-Ledesma, E., Hu, X., Amin, S.B., Akdemir, K.C., Seth, S., Song, X., Wang, Q., et al. (2017). Systematic analysis of telomere length and somatic alterations in 31 cancer types. *Nat. Genet.* 49, 349–357. <https://doi.org/10.1038/ng.3781>.
31. Cao, Y., Bryan, T.M., and Reddel, R.R. (2008). Increased copy number of the TERT and TERC telomerase subunit genes in cancer cells. *Cancer Sci.* 99, 1092–1099. <https://doi.org/10.1111/J.1349-7006.2008.00815.X>.
32. Jiao, Y., Shi, C., Edil, B.H., De Wilde, R.F., Klimstra, D.S., Maitra, A., Schultek, R.D., Tang, L.H., Wolfgang, C.L., Choti, M.A., et al. (2011). DAXX/ATRAX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science* 331, 1199–1203. <https://doi.org/10.1126/science.1200609>.
33. Heaphy, C.M., De Wilde, R.F., Jiao, Y., Klein, A.P., Edil, B.H., Shi, C., Bettgowda, C., Rodriguez, F.J., Eberhart, C.G., Hebbar, S., et al. (2011). Altered telomeres in tumors with ATRX and DAXX mutations. *Science* 333, 425.
34. Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* 37, 1155–1162. <https://doi.org/10.1038/s41587-019-0217-9>.
35. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–345. <https://doi.org/10.1038/nbt.4060>.
36. Ijdo, J.W., Baldini, A., Ward, D.C., Reeders, S.T., and Wells, R.A. (1991). Origin of human chromosome 2: An ancestral telomere-telomere fusion. *Proc. Natl. Acad. Sci. USA* 88, 9051–9055. <https://doi.org/10.1073/pnas.88.20.9051>.
37. Fan, Y., Linardopoulou, E., Friedman, C., Williams, E., and Trask, B.J. (2002). Genomic structure and evolution of the ancestral chromosome fusion site in 2q13-2q14.1 and paralogous regions on other human chromosomes. *Genome Res.* 12, 1651–1662. <https://doi.org/10.1101/gr.337602>.
38. Stong, N., Deng, Z., Gupta, R., Hu, S., Paul, S., Weiner, A.K., Eichler, E.E., Graves, T., Fronick, C.C., Courtney, L., et al. (2014). Subtelomeric CTCF and cohesin binding site organization using improved subtelomere assemblies and a novel annotation pipeline. *Genome Res.* 24, 1039–1050. <https://doi.org/10.1101/gr.166983.113>.
39. Riethman, H., Ambrosini, A., Castaneda, C., Finklestein, J., Hu, X.L., Mudunuri, U., Paul, S., and Wei, J. (2004). Mapping and initial analysis of human subtelomeric sequence assemblies. *Genome Res.* 14, 18–28. <https://doi.org/10.1101/gr.1245004>.
40. Ambrosini, A., Paul, S., Hu, S., and Riethman, H. (2007). Human subtelomeric duplication structure and organization. *Genome Biol.* 8, R151. <https://doi.org/10.1186/gb-2007-8-7-r151>.
41. 1000 Genomes Project Consortium; Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. *Nature* 526, 68–74.
42. Ewing, A.D., Houlahan, K.E., Hu, Y., Ellrott, K., Caloian, C., Yamaguchi, T.N., Bare, J.C., P'ng, C., Waggott, D., Sabelnykova, V.Y., et al. (2015). Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat. Methods* 12, 623–630. <https://doi.org/10.1038/nmeth.3407>.
43. Lee, A.Y., Ewing, A.D., Ellrott, K., Hu, Y., Houlahan, K.E., Bare, J.C., Espiritu, S.M.G., Huang, V., Dang, K., Chong, Z., et al. (2018). Combining accurate tumor genome simulation with crowdsourcing to benchmark somatic structural variant detection. *Genome Biol.* 19, 188. <https://doi.org/10.1186/s13059-018-1539-5>.
44. International Cancer Genome Consortium; Hudson, T.J., Anderson, W., Artez, A., Barker, A.D., Bernabé, R.R., Bernabé, R.R., Bhan, M.K., Calvo, F., Eerola, I., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. <https://doi.org/10.1038/nature08987>.

45. Ding, L., Bailey, M.H., Porta-Pardo, E., Thorsson, V., Colaprico, A., Bertrand, D., Gibbs, D.L., Weerasinghe, A., Huang, K.-L., Tokheim, C., et al. (2018). Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* 173, 305–320. <https://doi.org/10.1016/j.cell.2018.03.033>.
46. Morin, G.B. (1991). Recognition of a chromosome truncation site associated with alpha-thalassaemia by human telomerase. *Nature* 353, 454–456. <https://doi.org/10.1038/353454a0>.
47. Manguso, R.T., Pope, H.W., Zimmer, M.D., Brown, F.D., Yates, K.B., Miller, B.C., Collins, N.B., Bi, K., LaFleur, M.W., Juneja, V.R., et al. (2017). In vivo CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. *Nature* 547, 413–418. <https://doi.org/10.1038/nature23270>.
48. Grigorova, M., Staines, J.M., Ozdag, H., Caldas, C., and Edwards, P.A.W. (2004). Possible causes of chromosome instability: comparison of chromosomal abnormalities in cancer cell lines with mutations in BRCA1, BRCA2, CHK2 and BUB1. *Genome Res.* 104, 333–340. <https://doi.org/10.1159/000077512>.
49. Davidson, J.M., Gorringer, K.L., Chin, S.F., Orsetti, B., Besret, C., Courta-Cahen, C., Roberts, I., Theillet, C., Caldas, C., and Edwards, P.A. (2000). Molecular cytogenetic analysis of breast cancer cell lines. *Br. J. Cancer* 83, 1309–1317. <https://doi.org/10.1054/bjoc.2000.1458>.
50. Abdel-Rahman, W.M., Katsura, K., Rens, W., Gorman, P.A., Sheer, D., Bicknell, D., Bodmer, W.F., Arends, M.J., Wyllie, A.H., and Edwards, P.A. (2001). Spectral karyotyping suggests additional subsets of colorectal cancers characterized by pattern of chromosome rearrangement. *Proc. Natl. Acad. Sci. USA* 98, 2538–2543. <https://doi.org/10.1073/pnas.041603298>.
51. Grigorova, M., Lyman, R.C., Caldas, C., and Edwards, P.A.W. (2005). Chromosome abnormalities in 10 lung cancer cell lines of the NCI-H series analyzed with spectral karyotyping. *Cancer Genet. Cytogenet.* 162, 1–9. <https://doi.org/10.1016/j.cancergencyto.2005.03.007>.
52. Sirivatanauksorn, V., Sirivatanauksorn, Y., Gorman, P.A., Davidson, J.M., Sheer, D., Moore, P.S., Scarpa, A., Edwards, P.A., and Lemoine, N.R. (2001). Non-random chromosomal rearrangements in pancreatic cancer cell lines identified by spectral karyotyping. *Int. J. Cancer* 91, 350–358.
53. Edwards, P. SKY Karyotypes and FISH Analysis of Epithelial Cancer Cell Lines.
54. Muyas, F., Gómez Rodríguez, M.J., Cortes-Ciriano, I., and Flores, I. (2022). The ALT pathway generates telomere fusions that can be detected in the blood of cancer patients. Preprint at bioRxiv. <https://doi.org/10.1101/2022.01.25.477771>.
55. BRYANT, P.S.P.E. (1998). Review Chromosome healing, telomere capture and mechanisms of radiation-induced chromosome breakage. *Int. J. Radiat. Biol.* 73, 1–13. <https://doi.org/10.1080/095530098142653>.
56. Putnam, C.D., Pennaneach, V., and Kolodner, R.D. (2004). Chromosome healing through terminal deletions generated by de novo telomere additions in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* 101, 13262–13267. <https://doi.org/10.1073/pnas.0405443101>.
57. Harrington, L.A., and Greider, C.W. (1991). Telomerase primer specificity and chromosome healing. *Nature* 353, 451–454. <https://doi.org/10.1038/353451a0>.
58. Ribeyre, C., and Shore, D. (2013). Regulation of telomere addition at DNA double-strand breaks. *Chromosoma* 122, 159–173. <https://doi.org/10.1007/s00412-013-0404-2>.
59. Diede, S.J., and Gottschling, D.E. (1999). Telomerase-mediated telomere addition in vivo requires DNA primase and DNA polymerases alpha and delta. *Cell* 99, 723–733. [https://doi.org/10.1016/S0092-8674\(00\)81670-0](https://doi.org/10.1016/S0092-8674(00)81670-0).
60. Kramer, K.M., and Haber, J.E. (1993). New telomeres in yeast are initiated with a highly selected subset of TG1-3 repeats. *Genes Dev.* 7, 2345–2356. <https://doi.org/10.1101/gad.7.12a.2345>.
61. Mangahas, J.L., Alexander, M.K., Sandell, L.L., and Zakian, V.A. (2001). Repair of chromosome ends after telomere loss in *Saccharomyces*. *Mol. Biol. Cell* 12, 4078–4089. <https://doi.org/10.1091/mbc.12.12.4078>.
62. Schulz, V.P., and Zakian, V.A. (1994). The *saccharomyces* PIF1 DNA helicase inhibits telomere elongation and de novo telomere formation. *Cell* 76, 145–155. [https://doi.org/10.1016/0092-8674\(94\)90179-1](https://doi.org/10.1016/0092-8674(94)90179-1).
63. Kinzig, C.G., Zakusilo, G., Takai, K.K., Myler, L.R., and de Lange, T. (2022). ATR blocks telomerase from converting DNA breaks into telomeres. *Science* 383, 763–770. <https://doi.org/10.1126/science.adg3224>.
64. Meltzer, P.S., Guan, X.Y., and Trent, J.M. (1993). Telomere capture stabilizes chromosome breakage. *Nat. Genet.* 4, 252–255. <https://doi.org/10.1038/ng0793-252>.
65. Flint, J., Rochette, J., Craddock, C.F., Dodé, C., Vignes, B., Horsley, S.W., Kearney, L., Buckle, V.J., Ayyub, H., and Higgs, D.R. (1996). Chromosomal stabilisation by a subtelomeric rearrangement involving two closely related Alu elements. *Hum. Mol. Genet.* 5, 1163–1169. <https://doi.org/10.1093/hmg/5.8.1163>.
66. Tokutake, Y., Matsumoto, T., Watanabe, T., Maeda, S., Tahara, H., Sakamoto, S., Niida, H., Sugimoto, M., Ide, T., and Furuchi, Y. (1998). Extra-chromosomal telomere repeat DNA in telomerase-negative immortalized cell lines. *Biochem. Biophys. Res. Commun.* 247, 765–772. <https://doi.org/10.1006/bbrc.1998.8876>.
67. Cesare, A.J., and Griffith, J.D. (2004). Telomeric DNA in ALT Cells Is Characterized by Free Telomeric Circles and Heterogeneous t-Loops. *Mol. Cell Biol.* 24, 9948–9957. <https://doi.org/10.1128/mcb.24.22.9948-9957.2004>.
68. Londoño-Vallejo, J.A., Der-Sarkissian, H., Cazes, L., Bacchetti, S., and Reddel, R.R. (2004). Alternative Lengthening of Telomeres Is Characterized by High Rates of Telomeric Exchange. *Cancer Res.* 64, 2324–2327. <https://doi.org/10.1158/0008-5472.CAN-03-4035>.
69. Heaphy, C.M., Subhawong, A.P., Hong, S.M., Goggins, M.G., Montgomery, E.A., Gabrielson, E., Netto, G.J., Epstein, J.I., Lotan, T.L., Westra, W.H., et al. (2011). Prevalence of the alternative lengthening of telomeres telomere maintenance mechanism in human cancer subtypes. *Am. J. Pathol.* 179, 1608–1615. <https://doi.org/10.1016/j.ajpath.2011.06.018>.
70. Zhang, J.-M., and Zou, L. (2020). Alternative lengthening of telomeres: from molecular mechanisms to therapeutic outlooks. *Cell Biosci.* 10, 30. <https://doi.org/10.1186/s13578-020-00391-6>.
71. Livingstone, K., and Rieseberg, L. (2004). Chromosomal evolution and speciation: a recombination-based approach. *New Phytol.* 161, 107–112. <https://doi.org/10.1046/j.1469-8137.2003.00942.x>.
72. Fischer, G., James, S.A., Roberts, I.N., Oliver, S.G., and Louis, E.J. (2000). Chromosomal evolution in *Saccharomyces*. *Nature* 405, 451–454. <https://doi.org/10.1038/35013058>.
73. Dutrillaux, B. (1979). Chromosomal evolution in Primates: Tentative phylogeny from *Microcebus murinus* (Prosimian) to man. *Hum. Genet.* 48, 251–314. <https://doi.org/10.1007/BF00272830>.
74. Mazzoleni, S., Schillaci, O., Sineo, L., and Dumas, F. (2017). Distribution of Interstitial Telomeric Sequences in Primates and the Pygmy Tree Shrew (Scandentia). *Genome Res.* 151, 141–150. <https://doi.org/10.1159/000467634>.
75. Lin, K.W., and Yan, J. (2008). Endings in the middle: current knowledge of interstitial telomeric sequences. *Mutat. Res.* 658, 95–110. <https://doi.org/10.1016/j.mrrev.2007.08.006>.
76. Meyne, J., Baker, R.J., Hobart, H.H., Hsu, T.C., Ryder, O.A., Ward, O.G., Wiley, J.E., Wurster-Hill, D.H., Yates, T.L., and Moyzis, R.K. (1990). Distribution of non-telomeric sites of the (TTAGGG)<sub>n</sub> telomeric sequence in vertebrate chromosomes. *Chromosoma* 99, 3–10. <https://doi.org/10.1007/BF01737283>.
77. Ocalewicz, K., Furgala-Selezniow, G., Szmyt, M., Lisboa, R., Kucinski, M., Lejk, A.M., and Jankun, M. (2013). Pericentromeric location of the

- telomeric DNA sequences on the European grayling chromosomes. *Genetica* 141, 409–416. <https://doi.org/10.1007/s10709-013-9740-7>.
78. Faravelli, M., Moralli, D., Bertoni, L., Attolini, C., Chernova, O., Raimondi, E., and Giulotto, E. (1998). Two extended arrays of a satellite DNA sequence at the centromere and at the short-arm telomere of Chinese hamster chromosome 5. *Cell Genet.* 83, 281–286. <https://doi.org/10.1159/000015171>.
79. Turkalo, T.K., Maffia, A., Schabort, J.J., Regalado, S.G., Bhakta, M., Blanchette, M., Spierings, D.C.J., Lansdorp, P.M., and Hockemeyer, D. (2023). A non-genetic switch triggers alternative telomere lengthening and cellular immortalization in ATRX deficient cells. *Nat. Commun.* 14, 939. <https://doi.org/10.1038/s41467-023-36294-6>.
80. Calhoun, S.F., Reed, J., Alexander, N., Mason, C.E., Deitsch, K.W., and Kirkman, L.A. (2017). Chromosome End Repair and Genome Stability in *Plasmodium falciparum*. *mBio* 8, e00547-17. <https://doi.org/10.1128/mBio.00547-17>.
81. Reed, J., Kirkman, L.A., Kafsack, B.F., Mason, C.E., and Deitsch, K.W. (2021). Telomere length dynamics in response to DNA damage in malaria parasites. *iScience* 24, 102082. <https://doi.org/10.1016/j.isci.2021.102082>.
82. Sholes, S.L., Karimian, K., Gershman, A., Kelly, T.J., Timp, W., and Greider, C.W. (2022). Chromosome-specific telomere lengths and the minimal functional telomere revealed by nanopore sequencing. *Genome Res.* 32, 616–628. <https://doi.org/10.1101/gr.275868.121>.
83. Grigorev, K., Foox, J., Bezdan, D., Butler, D., Luxton, J.J., Reed, J., McKenna, M.J., Taylor, L., George, K.A., Meydan, C., et al. (2021). Haplotype diversity and sequence heterogeneity of human telomeres. *Genome Res.* 31, 1269–1279. <https://doi.org/10.1101/gr.274639.120>.
84. Min, J., Wright, W.E., and Shay, J.W. (2017). Alternative lengthening of telomeres can be maintained by preferential elongation of lagging strands. *Nucleic Acids Res.* 45, 2615–2628. <https://doi.org/10.1093/nar/gkw1295>.
85. Loe, T.K., Li, J.S.Z., Zhang, Y., Azeroglu, B., Boddy, M.N., and Denchi, E.L. (2020). Telomere length heterogeneity in ALT cells is maintained by PML-dependent localization of the BTR complex to telomeres. *Genes Dev.* 34, 650–662. <https://doi.org/10.1101/gad.333963.119>.
86. Carrot-Zhang, J., Yao, X., Devarakonda, S., Deshpande, A., Damrauer, J.S., Silva, T.C., Wong, C.K., Choi, H.Y., Felau, I., Robertson, A.G., et al. (2021). Whole-genome characterization of lung adenocarcinomas lacking the RTK/RAS/RAF pathway. *Cell Rep.* 34, 108707. <https://doi.org/10.1016/j.celrep.2021.108707>.
87. Campbell, J.D., Alexandrov, A., Kim, J., Wala, J., Berger, A.H., Pedamallu, C.S., Shukla, S.A., Guo, G., Brooks, A.N., Murray, B.A., et al. (2016). Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* 48, 607–616. <https://doi.org/10.1038/ng.3564>.
88. 1000 Genomes Project Consortium; Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. *Nature* 526, 68–74. <https://doi.org/10.1038/nature15393>.
89. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29, 308–311.
90. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bizikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, C., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. *Science* 376, 44–53. <https://doi.org/10.1126/science.abj6987>.
91. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34, 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>.
92. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at arXiv, 3997. <https://doi.org/10.48550/arXiv.1303.3997>.
93. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetic parameter estimation from sequencing data. *Bioinformatics* 27, 2987–2993. <https://doi.org/10.1093/bioinformatics/btr509>.
94. R Core Team (2021). A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).
95. Van Rossum, G., and Drake, F.L. (2009). Python 3 Reference Manual (CreateSpace).
96. Wall, L. (1994). The PERL Programming Language. Dr. Dobbs's J. Softw. Tools 19.
97. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinform.* 14, 178–192. <https://doi.org/10.1093/bib/bbs017>.
98. Tan, K.T., Slevin, M.K., Meyerson, M., and Li, H. (2022). Identifying and correcting repeat-calling errors in nanopore sequencing of telomeres. *Genome Biol.* 23, 180. <https://doi.org/10.1186/s13059-022-02751-6>.
99. Ding, Z., Mangino, M., Aviv, A., Spector, T., and Durbin, R.; UK10K Consortium (2014). Estimating telomere length from whole genome sequence data. *Nucleic Acids Res.* 42, e75. <https://doi.org/10.1093/nar/gku181>.
100. Lee, M., Hills, M., Conomos, D., Stutz, M.D., Dagg, R.A., Lau, L.M.S., Reddel, R.R., and Pickett, H.A. (2014). Telomere extension by telomerase and ALT generates variant repeats by mechanistically distinct processes. *Nucleic Acids Res.* 42, 1733–1746. <https://doi.org/10.1093/nar/gkt1117>.
101. Conomos, D., Stutz, M.D., Hills, M., Neumann, A.A., Bryan, T.M., Reddel, R.R., and Pickett, H.A. (2012). Variant repeats are interspersed throughout the telomeres and recruit nuclear receptors in ALT cells. *J. Cell Biol.* 199, 893–906. <https://doi.org/10.1083/jcb.201207189>.
102. Coleman, J., Baird, D.M., and Royle, N.J. (1999). The Plasticity of Human Telomeres Demonstrated by a Hypervariable Telomere Repeat Array That Is Located on Some Copies of 16p and 16q. *Hum. Mol. Genet.* 8, 1637–1646. <https://doi.org/10.1093/hmg/8.9.1637>.
103. Baird, D.M., Jeffreys, A.J., and Royle, N.J. (1995). Mechanisms underlying telomere repeat turnover, revealed by hypervariable variant repeat distribution patterns in the human Xp/Yp telomere. *EMBO J.* 14, 5433–5443. <https://doi.org/10.1002/j.1460-2075.1995.tb00227.x>.
104. Allshire, R.C., Dempster, M., and Hastie, N.D. (1989). Human telomeres contain at least three types of G-rich repeat distributed non-randomly. *Nucleic Acids Res.* 17, 4611–4627. <https://doi.org/10.1093/nar/17.12.4611>.
105. Zook, J.M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C.E., Alexander, N., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* 3, 160025. <https://doi.org/10.1038/sdata.2016.25>.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE                                                                                                      | SOURCE                                                                        | IDENTIFIER                                                                                                                          |
|--------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|
| <b>Chemicals, peptides, and recombinant proteins</b>                                                                     |                                                                               |                                                                                                                                     |
| McCoy's 5A Modified Medium                                                                                               | American Type Culture Collection (ATCC)                                       | Cat# 30-2007                                                                                                                        |
| ATCC-formulated RPMI-1640 Medium                                                                                         | American Type Culture Collection (ATCC)                                       | Cat# 30-2001                                                                                                                        |
| ATCC-formulated Dulbecco's Modified Eagle's Medium                                                                       | American Type Culture Collection (ATCC)                                       | Cat# 30-2002                                                                                                                        |
| <b>Critical commercial assays</b>                                                                                        |                                                                               |                                                                                                                                     |
| Monarch® Genomic DNA Purification Kit                                                                                    | New England Biolabs (NEB)                                                     | Cat# T3010S                                                                                                                         |
| Qubit™ HS dsDNA assay                                                                                                    | ThermoFisher - Invitrogen                                                     | Cat# Q32851 and Q32854                                                                                                              |
| ONT Genomic DNA Ligation kit                                                                                             | Oxford Nanopore Technologies (ONT)                                            | Cat# SQK-LSK109                                                                                                                     |
| NEBNext® Companion Module for Oxford Nanopore Technologies® Ligation Sequencing                                          | New England Biolabs (NEB)                                                     | Cat# E7180S                                                                                                                         |
| Agilent 4200 TapeStation (Genomic DNA ScreenTape)                                                                        | Agilent                                                                       | Cat# 5067-5366                                                                                                                      |
| Nanopore R9 MinION flow cell                                                                                             | Oxford Nanopore Technologies (ONT)                                            | Cat# FLO-MIN106D                                                                                                                    |
| NEBnext FFPE DNA Repair Mix                                                                                              | New England Biolabs (NEB)                                                     | Cat# M6630S                                                                                                                         |
| NEBNext Ultra II End Repair/dA tailing Module                                                                            | New England Biolabs (NEB)                                                     | Cat# E7442L                                                                                                                         |
| Nanopore PromethION R9.4.1 flow cell                                                                                     | Oxford Nanopore Technologies (ONT)                                            | Cat# FLO-PRO002                                                                                                                     |
| PacBio SMRTbell Express Template Prep Kit 2.0                                                                            | Pacific Biosciences (PacBio)                                                  | Cat# 100-938-900                                                                                                                    |
| SMRTbell Enzyme Clean Up Kit 2.0                                                                                         | Pacific Biosciences (PacBio)                                                  | Cat# 101-938-500                                                                                                                    |
| BluePippin™ Dye Free 0.75% Agarose Gel Cassettes                                                                         | Sage Science                                                                  | Cat# BHZ7510                                                                                                                        |
| Sequel II Binding Kit 2.2                                                                                                | Pacific Biosciences (PacBio)                                                  | Cat# 101-908-100                                                                                                                    |
| Sequel IIe 8M SMRT Cells                                                                                                 | Pacific Biosciences (PacBio)                                                  | Cat# 101-389-001                                                                                                                    |
| Sequel II Sequencing 2.0 Kit                                                                                             | Pacific Biosciences (PacBio)                                                  | Cat# 101-820-200                                                                                                                    |
| Agencourt® AMPure XP                                                                                                     | Beckman Coulter                                                               | Cat# A63881                                                                                                                         |
| Commercial spectral karyotyping paint probes from Applied Spectral Imaging                                               | Applied Spectral Imaging (5315 Avenida Encinas, Suite 150, Carlsbad, CA92008) | –                                                                                                                                   |
| <b>Deposited data</b>                                                                                                    |                                                                               |                                                                                                                                     |
| Nanopore PromethION long-read sequencing datasets                                                                        | This paper                                                                    | PRJNA1107807                                                                                                                        |
| Nanopore MinION long-read sequencing dataset                                                                             | This paper                                                                    | PRJNA1107807                                                                                                                        |
| PacBio HiFi long-read sequencing datasets                                                                                | This paper                                                                    | PRJNA1107807                                                                                                                        |
| Illumina short-read sequencing datasets                                                                                  | This paper                                                                    | PRJNA1107807                                                                                                                        |
| Whole genome short-read sequencing dataset from the Cancer Cell Line Encyclopedia                                        | Ghandi et al. <sup>7</sup>                                                    | PRJNA523380                                                                                                                         |
| Whole genome short-read sequencing dataset of lung adenocarcinoma patients from The Cancer Genome Atlas                  | Carrot-Zhang et al. and Campbell et al. <sup>86,87</sup>                      | <a href="https://gdc.cancer.gov/about-data/publications/pancanatlas">https://gdc.cancer.gov/about-data/publications/pancanatlas</a> |
| Whole genome short-read sequencing dataset of acute myeloid leukemia and colorectal cancers from The Cancer Genome Atlas | Ding et al. <sup>45</sup>                                                     | <a href="https://gdc.cancer.gov/about-data/publications/pancanatlas">https://gdc.cancer.gov/about-data/publications/pancanatlas</a> |

(Continued on next page)

### Continued

| REAGENT or RESOURCE                          | SOURCE                                               | IDENTIFIER                                                                                                                                                                                                      |
|----------------------------------------------|------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| NA12878 (SRR622457)                          | Auton et al. <sup>88</sup>                           | <a href="ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR622/SRR622457">ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR622/SRR622457</a>                                                                                           |
| PCAWG short-read sequencing datasets         | Campbell et al. <sup>3</sup>                         | <a href="https://dcc.icgc.org/pcawg">https://dcc.icgc.org/pcawg</a>                                                                                                                                             |
| Anotations for PCAWG datasets                | Campbell et al. <sup>3</sup>                         | <a href="https://dcc.icgc.org/releases/PCAWG">https://dcc.icgc.org/releases/PCAWG</a>                                                                                                                           |
| ICGC sequencing datasets                     | Hudson et al. <sup>44</sup>                          | <a href="https://dcc.icgc.org/">https://dcc.icgc.org/</a>                                                                                                                                                       |
| dbSNP (build 151)                            | Sherry et al. <sup>89</sup>                          | <a href="ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b151_GRCh38p7/VCF/common_all_20180418.vcf.gz">ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b151_GRCh38p7/VCF/common_all_20180418.vcf.gz</a> |
| GRCh38 reference genome                      | UCSC Genome Browser                                  | <a href="https://hgdownload.soe.ucsc.edu/downloads.html">https://hgdownload.soe.ucsc.edu/downloads.html</a>                                                                                                     |
| CHM13 reference genome                       | Nurk et al. <sup>90</sup>                            | <a href="https://github.com/marbl/CHM13">https://github.com/marbl/CHM13</a>                                                                                                                                     |
| <b>Experimental models: Cell lines</b>       |                                                      |                                                                                                                                                                                                                 |
| U2OS cells                                   | American Type Culture Collection (ATCC)              | Cat# HTB-96 <sup>TM</sup> ; RRID:CVCL_0042                                                                                                                                                                      |
| NCI-BL1184 cells                             | American Type Culture Collection (ATCC)              | Cat# CRL-5949 <sup>TM</sup> ; RRID:CVCL_2635                                                                                                                                                                    |
| NCI-H1184 cells                              | American Type Culture Collection (ATCC)              | Cat# CRL-5858 <sup>TM</sup> ; RRID:CVCL_1458                                                                                                                                                                    |
| Hs-746T cells                                | American Type Culture Collection (ATCC)              | Cat# HTB-135; RRID:CVCL_0333                                                                                                                                                                                    |
| <b>Software and algorithms</b>               |                                                      |                                                                                                                                                                                                                 |
| TelFuse                                      | This paper                                           | <a href="https://github.com/ktan8/teltools/">https://github.com/ktan8/teltools/</a>                                                                                                                             |
| TelSize                                      | This paper                                           | <a href="https://github.com/ktan8/teltools/">https://github.com/ktan8/teltools/</a>                                                                                                                             |
| Minimap2 v2.17-r941                          | Li <sup>91</sup>                                     | <a href="https://github.com/lh3/minimap2">https://github.com/lh3/minimap2</a>                                                                                                                                   |
| BWA-MEM v0.7.17-r1188                        | Li <sup>92</sup>                                     | <a href="https://github.com/lh3/bwa">https://github.com/lh3/bwa</a>                                                                                                                                             |
| SAMtools v1.10                               | Li et al. <sup>93</sup>                              | <a href="https://github.com/samtools/samtools">https://github.com/samtools/samtools</a>                                                                                                                         |
| R v4.2.0                                     | R Foundation for Statistical Computing <sup>94</sup> | <a href="https://www.r-project.org/">https://www.r-project.org/</a>                                                                                                                                             |
| Python v3.7.4                                | Van Rossum et al. <sup>95</sup>                      | <a href="https://www.python.org/">https://www.python.org/</a>                                                                                                                                                   |
| Perl v5.26.2                                 | Wall et al. <sup>96</sup>                            | <a href="http://www.perl.org/">http://www.perl.org/</a>                                                                                                                                                         |
| Integrative Genomics Viewer (IGV)            | Thorvaldsdóttir et al. <sup>97</sup>                 | <a href="https://software.broadinstitute.org/software/igv/">https://software.broadinstitute.org/software/igv/</a>                                                                                               |
| Bonito v0.3.5                                | Oxford Nanopore Technologies (ONT)                   | <a href="https://github.com/nanoporetech/bonito">https://github.com/nanoporetech/bonito</a>                                                                                                                     |
| Bonito basecalling model for telomeric reads | Tan et al. <sup>98</sup>                             | <a href="https://github.com/ktan8/nanopore_telomere_basecall">https://github.com/ktan8/nanopore_telomere_basecall</a>                                                                                           |
| <b>Other</b>                                 |                                                      |                                                                                                                                                                                                                 |
| Covaris® g-TUBE                              | Covaris®                                             | Cat# 520079                                                                                                                                                                                                     |
| Megaruptor 3 system                          | Diagenode                                            | B06010003                                                                                                                                                                                                       |
| PippinHT                                     | Sage Science                                         | Cat# HTP0001                                                                                                                                                                                                    |
| Sequel IIe instrument                        | Pacific Biosciences (PacBio)                         | –                                                                                                                                                                                                               |

## RESOURCE AVAILABILITY

### Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Matthew Meyerson ([matthew\\_meyerson@dfci.harvard.edu](mailto:matthew_meyerson@dfci.harvard.edu)).

### Materials availability

This study did not generate new unique reagents.

### Data and code availability

- Long-read genome sequencing data generated for this study is accessible at NCBI under the BioProject ID PRJNA1107807 and are publicly available as of the date of publication. The accession number is also listed in the [key resources table](#).
- TelFuse and TelSize software tools developed for this study are available at <https://github.com/ktan8/teltools/> or on Zenodo at <https://doi.org/10.5281/zenodo.11118081>. The DOI is listed in the [key resources table](#).
- Any additional information required to reanalyze the data reported in this paper is available from the [lead contact](#) upon request.

## EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

### Cell lines

U2-OS cells (ATCC HTB-96), NCI-BL1184 (ATCC cat no. CRL-5949) and NCI-H1184 (ATCC cat no. CRL-5858) and Hs-746T cells (ATCC cat no. HTB-135) cell lines were purchased directly from ATCC and used in this study. The identity of these cell lines was verified by comparing genomic sequencing data of these cell lines from those published previously by the Cancer Cell Line Encyclopedia. U2-OS cells were derived from a female donor, while NCI-BL1184, NCI-H1184 and Hs-746T cells were derived from male donors.

Cell lines were grown using culture and temperature conditions as stipulated by ATCC. Detailed information can be found on the ATCC website under the section “Handling information” for each cell line. Briefly, U2-OS cells were cultured in McCoy’s 5A Medium Modified (ATCC cat no. 30–2007) with 10% FBS. Cell lines NCI-BL1184 and NCI-H1184 were cultured in ATCC-formulated RPMI-1640 Medium (ATCC cat no. 30–2001) supplemented with FBS at 10%. Hs-746T cells were cultured in ATCC-formulated Dulbecco’s Modified Eagle’s Medium (ATCC cat no. 30–2002) supplemented with 10% FBS. All cell lines were cultured at 37°C.

### Human participants

Human participants were not recruited as part of this study. However, genomics datasets derived from human donors were analyzed as part of this study. Information about these donors can be found from the primary publication from which these datasets were derived from Campbell et al., Hudson et al., Carrot-Zhang et al., and Campbell et al.<sup>3,44,86,87</sup>

## METHOD DETAILS

### CCLE whole genome sequencing dataset

CCLE dataset<sup>7</sup> was downloaded from the European Nucleotide Archive under the study accession number (PRJNA523380). Specifically, only whole genome sequencing (WGS) datasets from the study was obtained. A full list of accession numbers corresponding to the CCLE WGS dataset used in this study is indicated in Table S1.

### TCGA whole genome sequencing dataset

Whole genome short-read sequencing dataset of lung adenocarcinoma,<sup>86,87</sup> acute myeloid leukemia and colorectal cancer patients from The Cancer Genome Atlas were downloaded from the GDC Data Portal (<https://portal.gdc.cancer.gov/>). The list of accession numbers corresponding to samples analyzed for this study are indicated in Tables S8 and S17.

### PCAWG and ICGC-NON-PCAWG whole genome sequencing datasets

Whole genome short-read sequencing dataset of PCAWG and ICGC-NON-PCAWG<sup>3,44</sup> datasets were obtained from the ICGC data portal (<https://dcc.icgc.org/>). The list of accession numbers corresponding to samples analyzed for this study is indicated in Tables S15, S16, S18, and S19.

### Annotations for PCAWG datasets

Annotations for PCAWG datasets were obtained from the PCAWG annotation repository (<https://dcc.icgc.org/releases/PCAWG>). Specifically, mutation calls of *ATRX/DAXX* genes (“Frame\_Shift\_Del”, “Frame\_Shift\_Ins”, “Nonsense\_Mutation”, “Splice\_Site”) and *TERT* promoter mutations (chr5:1295228 C>T, chr5:1295250 C>T) were obtained from the “final\_consensus\_passonly.snv\_mnv\_indel.icgc.public.txt” file from the “consensus\_snv\_indel” sub-directory. In addition, copy number and ploidy annotations (“consensus.20170217.purity.ploidy.txt”) were obtained from the “consensus\_cnv” sub-directory and gene expression annotations (“tophat\_star\_fpkms.v2 aliquot\_gl.tsv”) were obtained from the “transcriptome/gene\_expression” sub-directory. Structural variant calls were obtained from Table S1 of the PCAWG marker paper.<sup>3</sup>

### Identification of candidate new telomeres and chromosomal arm fusion events from short reads

Candidate short read pairs with at least two consecutive telomeric repeat sequences (TTAGGG)<sub>2</sub> in either reads in the pair were first extracted to narrow down the number of read pairs for subsequent analysis. 2x TTAGGG repeats was used as a cutoff as maximal correlation between telomere length from Southern blot measurements with telomere length estimates from short-reads was achieved with this cutoff,<sup>99</sup> and it was also applied in a different tool for telomere length estimation with short-read data.<sup>99</sup> Further, this would enable us to maximize sensitivity of our approach. Detection of repeats was performed using a custom Python script in the TelFuse package to each whole genome sequencing dataset.

Given the existence of telomere variant repeats (e.g., TCAGGG, TGAGGG, TTGGGG, GTAGGG, etc.) in a wide range of samples,<sup>100,101</sup> we also provided the option to analyze other types of variant repeats in TelFuse. However, because these telomeric variant repeats are also commonly found in the proximal region in the subtelomere,<sup>102–104</sup> and therefore difficult to interpret in the context of neotelomeres and chromosomal arm fusion events in our study, they were not analyzed here.

Candidate read pairs were then remapped to the reference genome (GRCh38) with BWA-MEM (version 0.7.17-r1188)<sup>92</sup> with default parameters. A custom Python script in the TelFuse package was then used to extract all sites with soft-clipped regions on

the mapped reads. Soft-clipped sequences from all reads at each unique genomic site was then used to generate a consensus sequence. A corresponding average sequence identity of the soft-clipped sequences to the consensus was also calculated.

To then filter this list of candidate sites for potential new telomeres and chromosomal arm fusion events, a series of filters were applied. Specifically, we ensured that (i) each site is supported by at least 3 reads, (ii) has an average sequence identity to the consensus of  $\geq 95\%$ , (iii) average mapping quality  $\geq 30$ , (iv) found more than 500kb from each end of the chromosome as defined by the reference genome, (v) is not found in more than one sample in the “panel of normal” constructed from these samples, and (vi) contains the circular permutations of (TTAGGG)<sub>2</sub> or (CCCTAA)<sub>2</sub> sequence in the soft-clipped sequences immediately after the breakpoint.

The candidate sites were then further subdivided into sites with telomeric repeats in the standard or inverted orientation, depending on the orientation of telomeric repeat sequences found on the first 12 bp of the softclipped sequences with respect to the genomic loci of interest. Sites which do not have perfect telomeric repeats on the first 12 bp, but in other parts of the softclipped sequences were defined as “neotelomere-like” or “arm-fusion-like” sites.

### High molecular weight DNA extraction

High molecular weight (HMW) DNA was isolated using a Monarch Genomic DNA Purification Kit (NEB, cat no. T3010S). DNA was quantified with a Qubit HS dsDNA assay (ThermoFisher, cat no. Q32851) followed by verification of HMW DNA integrity by electrophoresis on an Agilent 4200 TapeStation (Genomic DNA ScreenTape, cat no. 5067–5366).

### MinION library preparation

Sequencing libraries were prepared for the Oxford Nanopore Technologies (ONT) platform using the ONT Genomic DNA Ligation kit (ONT, cat no. SQK-LSK109). Briefly, HMW U2OS DNA was fragmented to  $\sim 20$  kb using a Covaris g-TUBE (cat no. 520079) followed by SPRI-cleanup (Agencourt AMPure XP, Beckman Coulter, cat no. A63881). Fragmented material was quantified with a Qubit dsDNA HS Assay Kit (Invitrogen, Catalog number: Q32851). One microgram of HMW U2OS DNA was end-repaired and A-tailed (NEBNext Companion Module for Oxford Nanopore Technologies Ligation Sequencing, cat no. E7180S) followed by adapter ligation. For sequencing 100 fmols of library material was loaded on an R9 flow cell (cat no. FLO-MIN106D).

### PromethION library preparation

Sequencing libraries for PromethION sequencing was prepared using the Genomic DNA by Ligation kit (SQK-LSK109) provided by Oxford Nanopore Technologies according to the recommended protocol (Version GDE\_9063\_v109\_revT\_14Aug2019) with slight modifications to the amount of input DNA used and the equipment used for shearing of the DNA. Briefly, 2.5  $\mu$ g of high molecular weight genomic DNA was sheared to 20kb using a Megaruptor 3 system (Diagenode, cat no. B06010003). DNA repair and end-prep was then performed using the NEBnext FFPE DNA Repair Mix and NEBNext Ultra II End Repair/dA tailing Module reagents in accordance with the manufacturer’s instructions followed by cleanup with AMPure XP beads. Ligation of adapters was then performed using the Ligation Sequencing kit (SQK-LSK109) according to manufacturer’s instructions, followed by loading onto a PromethION R9.4.1 flowcell (Oxford Nanopore, cat no. FLO-PRO002).

### PacBio HiFi library preparation

For CCS library preparation,  $\geq 3$   $\mu$ g of high molecular weight genomic DNA (more than 50% of fragments  $\geq 40$  kb) was sheared to  $\sim 15$  kb using the Megaruptor 3 (Diagenode B06010003), followed by DNA repair and ligation of PacBio adapters using the PacBio SMRTbell Express Template Prep Kit 2.0 (100-938-900) and removal of incomplete ligation products with the SMRTbell Enzyme Clean Up Kit 2.0 (PacBio 101-938-500). Libraries were then size-selected for 15 kb  $\pm$  20% using the PippinHT with 0.75% agarose cassettes (Sage Science). Following quantification with the Qubit dsDNA High Sensitivity assay (Thermo Q32854), libraries were diluted to 60 p.m. per SMRT cell, hybridized with PacBio V5 sequencing primer, and bound with SMRT seq polymerase using Sequel II Binding Kit 2.2 (PacBio 101-908-100). CCS sequencing was performed on the Sequel iIe instrument using 8M SMRT Cells (101-389-001) and Sequel II Sequencing 2.0 Kit (101-820-200), utilizing PacBio’s adaptive loading feature with a 2 h pre-extension time and 30 h movie time per SMRT cell. Initial quality filtering, basecalling, adapter marking, and CCS error correction was done automatically on board the Sequel iIe.

### Base calling of nanopore sequencing data

Base calling of Nanopore sequencing data in this study was performed using Bonito (Version 0.3.5) with the default dna\_r9.4.1 base-calling model. However, the default Nanopore basecalling model leads to frequent strand-specific base calling errors at telomeric repeats in our dataset, with (TTAGGG)<sub>n</sub> being miscalled as (TTAAAA)<sub>n</sub>, and (CCCTAA)<sub>n</sub> being miscalled as (CTTCTT)<sub>n</sub> and (CCCTGG)<sub>n</sub>, akin to what we had previously reported.<sup>98</sup> As such, telomeric reads were extracted using a pipeline that we had previously developed, followed by re-basecalling using a basecalling model that was previously tuned to correct these errors.<sup>98</sup>

### Extraction of candidate telomeric long reads for detailed analysis by TelSize

Long reads containing telomeric repeats were extracted by first enumerating the number of (TTAGGG)<sub>2</sub> and (CCCTAA)<sub>2</sub> motifs on each read using custom Perl scripts. Long reads containing at least four of these motifs were then defined as candidate telomeric

repeats. Of note, a low cutoff was deliberately set here to more sensitively identify long-reads with telomeric repeats for detailed analysis by TelSize.

### Estimation of telomere length from noisy long reads

As the telomeric long reads generated by Nanopore sequencing were relatively noisy, the length of telomeric repeats could not be readily inferred from the reads. To address this, we scanned each telomeric long read for instances of the telomeric repeat sequence (TTAGGG), or its reverse complement (CCCTAA). A vector representing positions where each of these motifs were observed was then generated. We then applied a moving average filter with window size 50 on this profile, followed by a moving median filter with window size 501. A minimum telomeric repeat signal of  $\geq 0.35$  was then applied to define a region as telomeric. The size of the telomeric repeat region was then established to determine the length of telomeric repeats on the long read, the localization of these sequences on the long-reads, and if (CCCTAA)<sub>n</sub> or (TTAGGG)<sub>n</sub> repeats were observed.

Specifically, long-reads were classified into five different classes: full telomeric – long-reads that contains telomeric repeat sequences end-to-end, left telomeric – long-reads that contains telomeric repeat sequences on the left edge of the long-read, right telomeric – long-reads that contains telomeric repeat sequences on the right edge of the long-read, intra-telomeric – long-reads that contains telomeric repeat sequences in the middle of the single long-read, and non-telomeric – long-reads that do not contain significant telomeric repeat signal throughout the long-read. These telomeric repeat signal can also occur as either (TTAGGG)<sub>n</sub> or (CCCTAA)<sub>n</sub> repeats, and these information are further reported.

This package for telomeric long read extraction and estimation (telSize) is available at the following github repository (<https://github.com/ktan8/teltools/>).

### Benchmarking of TelFuse by simulation experiments

The BAMSurgeon software (commit:75201d4) (<https://github.com/adamewing/bamsurgeon>)<sup>42,43</sup> was used to introduced neotelomeric and chromosomal arm fusion variants into the NA12878<sup>105</sup> and BL1184 datasets. This was done by introducing a long telomeric repeat into a locus, such that one end of the break point would look akin to a neotelomere with (TTAGGG)<sub>n</sub> directly after the break, while the other end of the break would appear similar to a arm fusion event with (CCCTAA)<sub>n</sub> repeats directly after the break. Specifically, 2000 random locations across chromosomes 1–22 and chromosome X were first selected using a customized R script. The length of telomere repeats at each of these sites was also determined randomly generated from a normal distribution ( $\mu = 6000$ ,  $\sigma = 1000$ ). The BAMSurgeon software (“addsv.py” script) was then used to introduce long telomeric insertions into each of these sites using the recommended parameters (--aligner mem --keepsecondary --seed 1234) and a variant allele frequency of 0.3 at all sites, leading to the introduction of a neotelomeric event on one end of the insertion event, and a chromosomal arm fusion event on the other end of the insertion event.

A total of 1513 and 1558 telomere insertions were successfully introduced into the NA12878 and BL1184 datasets respectively. Telomeric insertions could not be successfully introduced into the remaining 400+ sites as the sequencing reads or reference genome sequences at these sites failed predefined filtering criteria in the BAMSurgeon software (e.g., discordant read fraction >0.1, N nucleotides in the reference genome at these sites, best contig too short to make mutation, etc.).

The modified bam file was then processed using the TelFuse pipeline using the same parameters as that applied for the analysis of the real whole genome sequencing datasets in this study. Specifically, the “panel of normal” used for evaluation was generated from the 326 whole genome sequencing datasets from the CCLE. As there is a minor difference between the coordinate of the ground truth site introduced by BAMSurgeon, and that identified by TelFuse due to instances of microhomology between the introduced telomeric repeats and the reference genome, we considered sites identified by TelFuse to be true positives if they are within  $\pm 5$  bp of the ground truth coordinate. Significantly, with a more stringent window size of  $\pm 1$  bp, ~87% of sites identified using a  $\pm 5$  bp window size were still captured. Sites in the list of simulated neotelomeric/arm fusion sites that were not captured by TelFuse were deemed as false positives. Sites that are present in the ground truth list, but not detected by TelFuse were classified as false negatives.

The following metrics used for the evaluation of the performance of TelFuse in the simulated datasets are defined as follows:

Precision: Positive predictive value (PPV) = true positives/(true positives + false positives)

Sensitivity: True positive rate (TPR) = true positives/(true positives + false negatives)

Balanced F score ( $F_1$ ) =  $2 \times (\text{PPV} \times \text{TPR}) / (\text{PPV} + \text{TPR})$

### Analysis of telomeric repeat length at neotelomeres and chromosomal arm fusion sites

To assess length of telomeric repeats at neotelomeres and chromosomal arm fusion sites, only left telomeric, right telomeric, and intra-telomeric reads were considered. Specifically, for neotelomeric events, only reads with telomeric repeat regions found at the 5' or 3' end of the read (i.e., left telomeric and right telomeric reads) were considered to ensure that these reads correspond to a terminal region of a genomic locus. In the context of chromosomal arm fusion events, we require that the telomeric region be situated within the long-read (i.e., intra-telomeric reads that are flanked by non-telomeric repeats on both sides) to ensure that reads analyzed at these loci represent chromosomal arm fusion events.

For these telomeric repeat containing reads, sequences corresponding to the telomeric repeat region were trimmed off. The remaining non-telomeric sequences of each read were then mapped to the GRCh38 reference genome with minimap2 (Version 2.17-r941). Primary read mappings in the PAF format were then extracted and analyzed using custom R scripts in order to assess

mapping coordinates of these sequences. For each site of interest that was identified using short-read data, telomeric repeat containing long-reads that mapped to a  $\pm 100$  bp region of each site were extracted. Telomere length estimates for long-reads at each neotelomeric and chromosomal arm fusion sites were then reported as per [Figure 4](#).

### Analysis of telomeric repeat length at normal chromosomal arms

To assess length of telomeric repeats at normal chromosomal arms, only left telomeric and right telomeric reads were considered, akin to the neotelomeric sites. Sequences corresponding to the telomeric region were similar trimmed off. The remaining non-telomeric repeat sequences were mapped to the CHM13 v2.0 reference genome using minimap2 (Version 2.17-r941) as the sub-telomeric region of this reference genome is complete in contrast to the GRCh38 reference genome. Reads that mapped to the terminal 500kb region of each chromosomal arm were classified as telomeric reads originating from normal chromosomal arms.

### Copy number profiles

To generate copy number profiles of the cancer cell lines from the CCLE, the total sequencing coverage of each 10 kb bin was calculated using the bedcov function SAMtools (v1.10)<sup>93</sup> with default parameters. The coverage was then normalized to a per-basepair level and is as depicted.

For lung adenocarcinoma samples which have a matched normal samples, the normalized sample coverage across each chromosome was calculated as follows: The sequencing coverage for each 10kb bin was calculated for both the tumor and matched normal sample using the bedcov function in SAMtools (v1.10).<sup>93</sup> These values were then normalized by the total read count of each dataset, and the ratio between the tumor and normal sample calculated to obtain the normalized sample coverage.

### Analysis of BAF

As no matched normal samples were sequenced for each of the cancer cell lines, heterozygous germline variants cannot be directly assessed and used in the generation of allelic ratio plots. Allelic ratios were thus assessed using a set of common germline SNPs from the dbSNP database (GRCh38.p7 build 151)<sup>89</sup> ([ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human\\_9606\\_b151\\_GRCh38p7/VCF/common\\_all\\_20180418.vcf.gz](ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b151_GRCh38p7/VCF/common_all_20180418.vcf.gz)). Specifically, the list of common SNPs are defined by the dbSNP database as SNPs that are found with a minor allele frequency of at least 0.01 in the 1000 genomes project.

Custom Python scripts and SAMtools mpileup (v1.10) were then used to enumerate all four possible bases at each SNP site (base quality  $\geq 20$ ). The allelic ratio was then calculated as the ratio of the variant base (as defined by dbSNP) count versus the sum of the reference and variant base count. Only sites with a coverage of at least 15x were plotted.

### SV variant allele frequency

The clonality of the ectopic telomeric events was assessed by assessing the allelic frequency associated with the breakpoints of these structural variants. Specifically to do so, read pairs that support the ectopic telomeric events (minimum of 5 bp overlap) were enumerated (Count A). Read pairs that support the reference sequence were similar counted (Count B). The variant allele frequency associated with the ectopic telomeric event was calculated as  $A/(A + B)$ . To correct for tumor purity and ploidy, the raw variant allele frequency values for these events were divided by the tumor purity and the ploidy values obtained from the PCAWG consortium (consensus\_cnv/consensus.20170217.purity.ploidy.txt),<sup>3</sup> and then normalized to a ploidy value of two. Notably, as the sites presented in our study were generated from the hg38 reference genome, these sites had to be lifted over to hg19 coordinates before clonality information could be inferred from the hg19 mapped BAM files generated by the PCAWG consortium. The source code to perform this procedure can be found in the teltools Github repository.

### Sequence signatures at sites with new telomeres and chromosomal arm fusion events

The sequence signature at each new telomeric and chromosomal arm fusion site was analyzed using the consensus soft-clipped sequences identified by TelTools, and the sequence extracted from the reference genome at each site. The sequence signature at each new telomere and chromosomal arm fusion was then analyzed by (i) identifying the frequency of each telomeric 6-mer in each soft-clipped sequence, and by (ii) assessing the sequence motif of the telomeric region and genomic region.

### Spectral Karyotyping

DNA Spectral Karyotyping Hybridization was performed according to the protocol of commercial spectral karyotyping paint probes from Applied Spectral Imaging (5315 Avenida Encinas, Suite 150, Carlsbad, CA92008). Briefly, the slides were dropped in Thermo-tron and aged for 3–5 days in a 37°C oven. The slides were then checked under the microscope before hybridization. A series of four steps were then performed on these slides to generate the spectro karyotype of the cell lines: (1) Trypsin Treatment: The slides were washed briefly in Earl's medium, and then treated with Trypsin/EDTA solution. Washing was then performed in water and then dehydrated in ethanol series of 70%, 80% and 100% for 2 min each followed by air-drying of the slides. (2) Chromosome Denaturation: The slides were treated in 2XSSC buffer for 2 min and then dehydrated in Ethanol series for 2 min each. Denaturation of the slides was then performed at 72°C in denaturation solution for 1.5 min. This is followed immediately by placing the slides in cold ethanol series to dehydrate the slides, and then air drying. (3) Probe Denaturation and hybridization: The probe was denatured by incubating the probe at 80°C in a water bath for 7 min. The denatured Spectral Karyotyping reagent was then applied to the denatured chromosome

preparation and incubated at 37°C for 5–6 days (4) Detection, imaging and karyotyping: The slides were washed in 0.4XSSC at 72°C for 2 min and then dipped in 4XSSC/Tween 20 for 1 min. Cy5 staining reagent was then applied and incubated at 37°C for 40 min. The slides were then washed 3 times in washing solution, and then mounted with anti-fade DAPI. After which, the slides are ready for spectral imaging. Rearrangements were defined with nomenclature rules from international Committee in Standard Genetic Nomenclature for Human.

### QUANTIFICATION AND STATISTICAL ANALYSIS

The Wilcoxon Rank Sum tests, Chi-squared test, linear regression and the Kolmogorov-Smirnov tests were performed using R (v4.2.0) as part of analyses in this study. Statistical details of these tests are provided in the figure legends, or in the text. Precision measures (mean or median) are specified in the text when reported. Exact values of *n* for statistical tests and what it represents, are provided directly in the figure, figure legends, or in the main text. A *p*-value <0.05 was deemed statistically significant in this study. Non-parametric tests were applied as indicated above, and assumptions about the population distribution was therefore not made.