

## 1 **Neotelomeres and Telomere-Spanning Chromosomal Arm Fusions in Cancer 2 Genomes Revealed by Long-Read Sequencing**

3  
4 Kar-Tong Tan<sup>1,2,3</sup>, Michael K. Slevin<sup>1,4</sup>, Mitchell L. Leibowitz<sup>1,2,3</sup>, Max Garrity-Janger<sup>1,2,3</sup>,  
5 Heng Li<sup>5,6,\*</sup>, Matthew Meyerson<sup>1,2,3,4,7,\*</sup>

6 <sup>1</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA

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

8 <sup>3</sup>Department of Genetics, Harvard Medical School, Boston, MA 02215, USA

9 <sup>4</sup>Center for Cancer Genomics, Dana-Farber Cancer Institute, Boston, MA 02215, USA

10 <sup>5</sup>Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215, USA

11 <sup>6</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02215, USA

12 <sup>7</sup>Lead contact

13  
14 \*Correspondence: [hli@jimmy.harvard.edu](mailto:hli@jimmy.harvard.edu) (H.L.), [matthew\\_meyerson@dfci.harvard.edu](mailto:matthew_meyerson@dfci.harvard.edu)  
15 (M.M.)

### 16 **Abstract**

17  
18 Alterations in the structure and location of telomeres are key events in cancer genome  
19 evolution. However, previous genomic approaches, unable to span long telomeric  
20 repeat arrays, could not characterize the nature of these alterations. Here, we applied  
21 both long-read and short-read genome sequencing to assess telomere repeat-  
22 containing structures in cancers and cancer cell lines. Using long-read genome  
23 sequences that span telomeric repeat arrays, we defined four types of telomere repeat  
24 variations in cancer cells: neotelomeres where telomere addition heals chromosome  
25 breaks, chromosomal arm fusions spanning telomere repeats, fusions of neotelomeres,  
26 and peri-centromeric fusions with adjoined telomere and centromere repeats. Analysis  
27 of lung adenocarcinoma genome sequences identified somatic neotelomere and  
28 telomere-spanning fusion alterations. These results provide a framework for systematic  
29 study of telomeric repeat arrays in cancer genomes, that could serve as a model for  
30 understanding the somatic evolution of other repetitive genomic elements.

### 31 **Keywords**

32 Telomere, long-read sequencing, neotelomeres, arm fusions, repetitive elements

42 **Introduction**

43

44 Cancer is driven by alterations to the genome. The continued invention and  
45 application of new methods has enabled the characterization of genomic alterations in  
46 cancer with much greater scale and resolution. The development of massively parallel  
47 short-read sequencing over the past fifteen years has greatly accelerated our efforts to  
48 characterize the cancer genome by enabling the detailed and rapid characterization of  
49 somatic and germline variants in tens of thousands of samples<sup>1-7</sup>, and led directly to the  
50 discovery of many cancer driving genetic alterations that are now being targeted by  
51 emerging therapeutics. The recent development and application of linked-read genome  
52 sequencing of long molecules with barcoded short-reads then facilitated the  
53 characterization of more complex structural variations, and of genomic alterations at the  
54 haplotype-level in cancer<sup>8-12</sup>.

55 Despite these advances in genome technology, the identification and  
56 characterization of somatic alterations at repetitive elements, which constitute  
57 approximately half the human genome<sup>13-15</sup>, still remain significant challenges. Repetitive  
58 elements and duplicated sequences in the human genome are typically 100 to 8000 bp  
59 in size<sup>15</sup>, although centromeres are much longer arrays of repetitive elements, and can  
60 be broadly classified into three main classes. First, repetitive elements include tandem  
61 repeats of specific DNA sequences<sup>15,16</sup>, including short tandem repeats (1-6 bp repeat  
62 unit) in the form of microsatellites, and longer repeat units forming minisatellites<sup>16</sup>.  
63 Telomeres and centromeres, which are key structures in a chromosome, are largely  
64 comprised of long tandem repeats<sup>15</sup>. Second, repetitive elements include interspersed  
65 repeats, identical or nearly identical sequences spread out across the human genome<sup>15</sup>,  
66 including short interspersed nuclear elements (SINEs; typically 100-300 bp in length)  
67 such as Alu repeats, and long interspersed nuclear elements (LINEs, typically >300 bp  
68 in length) such as L1 repeats<sup>15</sup>. Third, “low copy repeats”, or segmental duplicates, also  
69 occur in the genome. These large repetitive sequences are blocks of DNA that are 1-  
70 400 kilobases in size, occur as at least two copies, share high sequence similarity  
71 (>90%)<sup>17,18</sup>, and are potential hotspots of chromosomal rearrangements and  
72 instability<sup>19,20</sup>. Although sophisticated computational methods have been developed to  
73 infer somatic alterations in repetitive regions using short-reads, comprehensive  
74 characterization of somatic alterations in these regions still cannot be completely  
75 achieved.

76 Telomeres are a salient example of highly repetitive structures of particular  
77 importance in cancer that cannot yet be readily resolved by current sequencing  
78 methodologies. Human telomeres, which act as protective caps on the ends of  
79 chromosomes are composed of ~2-10 kb (TTAGGG)<sub>n</sub> tandem repeats<sup>21,22</sup>. Somatic  
80 integration of telomeric sequences into non-telomeric DNA in tumor samples has also  
81 been observed<sup>23</sup>, though the origin and structures of these sequences remain unclear.  
82 As the short-read sequencing that is typically performed, such as 2 x 150 bp paired  
83 reads, is unable to fully span the 2 kb – 10 kb long highly repetitive telomeres, much  
84 remains unknown about telomere structures in cancer.

85 The study of telomere structure is important in cancer genomics because  
86 telomere maintenance is crucial in cancer pathogenesis. Cancer cell immortality  
87 requires a mechanism to activate telomerase or otherwise maintain telomeres, and is a

88 key “hallmark of cancer”<sup>24,25</sup>. Telomerase, the enzyme which adds telomeric repeats to  
89 the ends of chromosomes, has been estimated to undergo reactivation in as many as  
90 90% of human cancers and was shown experimentally to be critical for malignant  
91 transformation<sup>26–31</sup>. The reactivation of telomerase activity in cancer is driven in part by  
92 promoter mutations, amplifications and translocations in the telomerase catalytic subunit  
93 gene, *TERT*<sup>32–35</sup>, and also by amplification of the RNA component of telomerase,  
94 *TERC*, in cancer<sup>35,36</sup>. In some cancer types, genetic inactivation of the *ATRX* and *DAXX*  
95 genes are also associated with telomere elongation, independent of telomerase, by the  
96 alternative lengthening of telomeres (ALT) pathway<sup>37,38</sup>.

97 The emergence of long-read genome sequencing now makes it possible to  
98 analyze somatic alterations in highly repetitive regions, such as telomeric repeats, with  
99 greater precision and detail. Recently, the first telomere-to-telomere human genome  
100 was assembled using long-reads that can span large, complex, or repetitive genomic  
101 sequences, including telomeric repeats. This assembly relied upon PacBio high-fidelity  
102 (HiFi) sequencing, which can generate long reads with an accuracy of 99.8% and an  
103 average length of 13.5 kb<sup>39</sup>, as well as ultra-long-read nanopore sequencing, which can  
104 generate reads of over 100 kb<sup>40</sup>. Using long reads, the repetitive telomeres can be  
105 spanned and mapped uniquely to the human genome. However, long-read sequencing  
106 is still significantly more expensive than short-read sequencing. Given that high-  
107 coverage short-read genome sequences are now widely available, a cost-effective  
108 strategy at this time is to leverage short-read sequencing datasets to identify samples  
109 with potentially interesting telomeric alterations *in silico*, and to subject these samples to  
110 more detailed analysis by long-read sequencing.

111 Here, we explored the structure of previously unresolved telomeric events in the  
112 cancer genome. We used large databanks of short-read genome sequencing datasets  
113 to identify candidate telomeric alterations in the genome of 326 cancer cell line and 95  
114 primary lung adenocarcinoma samples using TelFuse, a computational method to  
115 profile ectopic intra-chromosomal telomeric repeat sites. Then, using PacBio HiFi and  
116 Nanopore long-read genome sequencing in three cell lines with high numbers of  
117 putative telomeric variants, we resolved the structure of these alterations in combination  
118 with spectral karyotyping, copy number and allelic ratio analysis. Long-read genome  
119 sequencing of these samples led directly to the discoveries of neotelomeres, telomere-  
120 spanning chromosomal arm fusion events, and complex telomeric alterations that were  
121 not previously resolvable using short-read genome sequencing. These findings also  
122 validate recent experimental observations on neotelomere formation<sup>41</sup>. Our study  
123 creates a framework that can be applied to the examination of other highly repetitive  
124 sequences that are likely to be of biological significance in disease, including  
125 centromere arrays, transposable element insertions, and microsatellite repeats.

126  
127

128 **Results**

129

130 Identification of ectopic telomeric repeat sequences

131

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

139 We developed the analytic method, TelFuse, to identify ectopic telomeric repeats  
140 within the cancer genome, and to estimate telomere length of each chromosomal arm  
141 with long-read sequencing respectively (**Figure S1A**). TelFuse identifies ectopic  
142 telomeric repeat sequences (TTAGGG)<sub>n</sub> or (CCCTAA)<sub>n</sub> that are absent from the  
143 germline and mapped to intrachromosomal regions (i.e. at least 500 kb from  
144 chromosomal ends) (**Methods, Figure S1A-B**). TelFuse begins by identifying read pairs  
145 that contain at least 2 perfect consecutive telomeric repeats (at least 12 base pairs of  
146 telomere sequence) with adjacent sequences that map to intra-chromosomal sites.  
147 Paired read sequences that are fully aligned to the reference genome are removed,  
148 eliminating telomeric repeats in the reference, which include ancient chromosome  
149 fusion events<sup>42,43</sup>. To ensure the specificity of our calls, we also developed a series of  
150 filters (**Figure S1A-B, Methods**) to remove spurious sites caused by artefacts induced  
151 during the mapping process (**Figure S1A-B**), assessed by a variety of quality control  
152 metrics (**Methods**). Those sites that pass all filters and are at least 500 kb from the  
153 GRCh38 reference genome chromosome terminus, a sufficient distance to avoid sub-  
154 telomere sequences<sup>44-46</sup>, are considered candidate sites of ectopic telomere sequence.

155

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

158

159 To assess the landscape of ectopic telomere repeats in cancer, we began by  
160 analysis of cancer cell line data, which allows assessment of multiple cancer types and  
161 which provides high sequencing depth due to 100% cancer cell purity. We applied  
162 TelFuse to whole genome sequencing datasets from 326 cancer cell line DNA  
163 specimens from the Cancer Cell Line Encyclopedia (CCLE)<sup>7</sup>, and detected 240  
164 candidate ectopic intra-chromosomal telomeric repeat sequence sites in 34% of cell  
165 lines (112/326) (**Figure 1C-D and Table S1 and S2**). Analysis of the orientation of the  
166 telomere repeats further defined these candidates as corresponding to 149 candidate  
167 sites with telomeric repeat sequences in the standard orientation, and 91 candidate  
168 sites with telomeric repeat sequences in the reverse orientation (**Figure 1C-D**). An  
169 additional 42 candidate sites with telomeric repeat sequences within softclipped  
170 sequences, but not on the first 12 base-pairs, were also detected (**Table S3**); these  
171 were not analyzed in depth. These data indicate that genomic events involving telomeric  
172 repeat sequences can be readily detected in cancer cell lines from short-read genome  
173 sequencing using TelFuse.

174

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

176

177        Although short-read sequencing (2 x 101bp for the CCLE dataset<sup>7</sup>) can detect  
178 ectopic telomeric sequences, the length and repetitive nature of these sequences,  
179 which can span 10 kb in length<sup>21,22</sup>, renders their structures indecipherable based on  
180 short-read data alone, and cannot distinguish between possible modes of generation of  
181 these sites. Therefore, we decided to perform high-depth long-read sequencing of  
182 selected cell line genomes.

183        We selected the U2-OS osteosarcoma cancer cell line, with 55 candidate  
184 telomeric repeat sites from short read sequencing (46 in the standard orientation and 9  
185 in the inverse orientation), the Hs-746T gastric carcinoma cell line with 6 candidate  
186 events (1 standard and 5 inverse orientation), and the NCI-H1184 small cell lung cancer  
187 cell line with 6 candidate events (5 standard and 1 inverted orientation) (**Figure S1C-E**),  
188 together with its matched normal sample (NCI-BL1184). These samples were selected  
189 due to the high frequencies of ectopic telomeric events (**Figure S1C-E**). Notably, the  
190 U2-OS cell line was found to be highly rearranged, with ectopic telomeric sites found  
191 near regions with changes in sequencing coverage and allelic ratios (**Figure S2**). We  
192 then performed PacBio HiFi and Oxford Nanopore long-read genome sequencing  
193 (**Figure 1E**). We achieved a median genomic coverage of 49x, 62x, 65x and 73x for the  
194 U2OS, Hs-746T, NCI-BL1184 and NCI-H1184 cell lines respectively with Nanopore  
195 long-read genome sequencing (**Figure S3, Table S4 and S5**). With PacBio HiFi  
196 sequencing, we achieved a median genomic coverage of 19x, 20x, 19x, and 23x for the  
197 same four cell lines using high quality PacBio HiFi reads, and a median coverage of  
198 29x, 31x, 33x and 36x when all PacBio reads were considered (**Figure S3, Table S4**  
199 **and S5**). Nanopore sequencing data had a median read length of 6-13 kb (N50: 18-21  
200 kb), while the PacBio HiFi data had a median read length of 15-17 kb (N50: 16-19 kb)  
201 (**Figure S3, Table S4 and S5**). In parallel, to assess chromosomal scale structures of  
202 these events, we also performed spectral karyotyping.

203        Long-read sequencing of cancer cell line genomes revealed two major types of  
204 structural alterations containing telomere repeat sequences that comprised either  
205 telomeric repeat sequences of > 1 kb flanked on one end by chromosomal sequence  
206 with no other flanking DNA (seen in 46 of 51 examples sequenced) or telomeric repeat  
207 sequences of at least few hundred base-pairs flanked on both sides by chromosomal  
208 sequence (seen in 12 of 15 candidate events sequenced). Telomeres flanked on only  
209 one-end with chromosomal sequence are consistent with neotelomere structures, which  
210 might be generated through telomerase activity<sup>41</sup>. Telomeres flanked on both sides with  
211 chromosomal sequence are likely to be sites of chromosome fusion or other  
212 translocation events.

213

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

215

216        Long-read sequencing analyses demonstrated that the ectopic telomere repeat  
217 sequences in the standard orientation were long and unbounded and therefore  
218 consistent with neotelomere addition. For example, a candidate ectopic telomere repeat  
219 sequence site adjacent to sequence from chrX:103,320,553 in the U2-OS

220 osteosarcoma cell line was observed to contain at least seven tandem (TTAGGG)<sub>n</sub>  
221 repeats in the short-read sequencing data (**Figure 2A**), and a reduction in sequencing  
222 coverage, corresponding to the position of the telomeric repeats, at this chromosomal  
223 position (**Figure 2B**). Upon analysis of this region in long-read sequencing data sets,  
224 both PacBio HiFi and Oxford Nanopore, long telomeric repeats of ~3-10kb in the  
225 standard orientation could be readily observed (**Figure 2C**), where the variation in  
226 telomere length between reads might be explained by active telomere sequence loss or  
227 telomere maintenance after DNA replication, in different cells across the population.  
228 These data support a model where breakage of the chrXq arm was capped by  
229 generation of novel telomeric sequence representing a neotelomere (**Figure 2D**).

230 Another example of a neotelomere is seen in the Hs-746T cell line, within  
231 chromosome arm 21p at chr21:10,547,397 where an ectopic telomeric repeat site was  
232 observed. Short-read sequencing showed at least six tandem (CCCTAA)<sub>n</sub> repeats  
233 (**Figure 2E**). At this location, fluctuation in both sequencing coverage and allelic ratios  
234 could be observed (**Figure 2F**). Analysis of both PacBio HiFi and Nanopore long-read  
235 genome sequencing data again revealed long telomeric repeats (~5-10 kb) in the  
236 standard orientation with reference to the break point at this site (**Figure 2G**), lending  
237 support to the existence of a neotelomere which had likely formed following breakage of  
238 the chr21p arm (**Figure 2H**). Similar observations were made at other ectopic  
239 neotelomeric sites, such as chr7:24,302,169 in the U2-OS cell line (**Figure S4A-D**), and  
240 chr1:214,460,753 in the NCI-H1184 small cell lung carcinoma cell line (**Figure S4E-H**),  
241 further supporting the idea that these ectopic telomeric sites in the standard orientation  
242 detected by short-reads represent neotelomere addition events.

243 In all, among 51 sites predicted by TelFuse as containing standard orientation  
244 telomere repeat sequences in these three cancer cell lines using short-read genome  
245 sequencing data, 46 of these sites could be readily demonstrated to represent long  
246 telomere repeats suggestive of neotelomeres, using the long-read genome sequencing  
247 data (**Figure 2I, Table S6**). No telomeric long reads could be found at the other 5 sites.  
248 Together, our results indicate that short telomeric repeats in the standard orientation,  
249 observed with short-read sequencing data, represent neotelomeres with long telomeric  
250 repeats as confirmed by long-read genome sequencing.

251 To assess the relationship between neotelomeres and chromosomal alterations,  
252 and to support our neotelomere calls, we performed spectral karyotyping of the U2-OS  
253 cancer cell line, with detailed karyotyping for ten randomly selected cells (representative  
254 cell shown in **Figure S5A**). Integrative analysis of sequencing coverage, allelic ratios  
255 and long-read data inferred two copies of chromosome X in U2-OS cells, one complete  
256 copy and one truncated chromosome X. Concordant with a neotelomere detected by  
257 long-read genome sequencing data (**Figure 2A-D**), a shorter chromosome X with q-arm  
258 deletion was observed by spectral karyotyping in 7/10 cells assessed (**Figure 2J**),  
259 together with a full-length chromosome X in 10/10 cells karyotyped. Thus, the spectral  
260 karyotyping analysis confirms that neotelomeres identified by long-read sequencing can  
261 be correlated with chromosomal truncations observed by cytogenetics.

262 We also observed a significant level of chromosomal heterogeneity (**Figure**  
263 **S5B-C, Table S7**). Heterogeneities we observed included slight variations in  
264 chromosome number between each cell (N=76-80) (**Table S7**) and heterogeneity in  
265 translocation events between cells that were concordant with a prior study<sup>27</sup>.

266 Specifically, while a t(4;22) translocation could be observed in 10/10 cells assessed  
267 (**Figure S5C**), a t(15;19) translocation was only observed in 6/10 cells assessed. This  
268 cellular heterogeneity might explain why long-read sequencing was unable to validate 5  
269 of the 51 candidate sites that were detected in the population of cells sequenced by  
270 CCLE. Therefore, heterogeneity in tumor cell populations remains a complication in  
271 identifying ectopic telomeric events.

272  
273 Telomere repeat-spanning chromosomal arm fusions in cancer resolved by long-read  
274 genome sequencing

275  
276 We next explored sites with ectopic telomeric repeat sequences found in the  
277 inverted orientation with respect to the breakpoint; long-read sequencing revealed that  
278 these sites largely represent chromosomal arm fusion events. At one candidate site at  
279 position chr4:30,909,846, we observed eight inverted telomeric repeats (CCCTAA)<sub>n</sub>  
280 (~48 bp) using short-read sequencing data (**Figure 3A**). At this position, a significant  
281 change in sequencing coverage and change in allelic ratio were also observed in  
282 support of the fusion event (**Figure 3B**). Analyzing this region with both PacBio HiFi and  
283 Nanopore long-read genome sequencing, we observed ~650bp of inverted (CCCTAA)<sub>n</sub>  
284 repeats after the breakpoint (**Figure 3C**), followed by 5-8 kb of sequences on chr22q  
285 sub-telomeres (**Figure 3C**). Individual long-reads that cover the whole event suggest  
286 that the inverted (CCCTAA)<sub>n</sub> repeat sequences formed via the fusion of the chr22q arm  
287 with its short telomere to an intra-chromosomal site (**Figure 3D**).

288 We also observed more complex fusion events, including evidence for the  
289 formation of a neotelomere followed by a subsequent chromosomal fusion. At  
290 chr11:84,769,636, five inverted ectopic telomeric repeats (CCCTAA)<sub>n</sub> (~30 bp) were  
291 detected at the breakpoint with short-read sequencing (**Figure 3E**). At this site, a drastic  
292 change in allelic ratios was observed despite minimal changes in copy number  
293 estimated from sequencing coverage (**Figure 3F**), suggesting changes to one of the  
294 parental chromosomes despite no overall changes in chromosomal number. Using both  
295 PacBio HiFi and Nanopore long-read sequencing data, we observed ~1750 bp of  
296 inverted (CCCTAA)<sub>n</sub> telomeric repeats at this site (**Figure 3G**). Surprisingly, we could  
297 further observe >5kb of sequences corresponding to an intra-chromosomal site on the  
298 chr11p arm, suggesting that the neotelomere was the consequence of multiple steps. It  
299 may have first formed on the centromeric side of the chr11p breakpoint  
300 (chr11p:43,002,345), which then subsequently fused to the breakpoint on chr11q at  
301 position 84,769,636 (**Figure 3H**).

302 To assess if telomere-spanning chromosomal fusions could be detected in other  
303 samples, we again examined long-read genome sequencing data of the Hs-746T  
304 gastric adenocarcinoma and NCI-H1184 lung adenocarcinoma cell lines. Inverted  
305 ectopic telomeric repeats that were identified using TelFuse were confirmed as sites of  
306 chromosomal arm fusion events with long-read data in the Hs-746T sample (**Figure S6**)  
307 at the sites chr11:79,325,679 and chr1:244,201,717, but not for the single candidate site  
308 in the NCI-H1184 sample (**Table S6**). Again, the discrepancy between long- and short-  
309 read data in our study could be caused by heterogeneity in the cancer cell lines.  
310 Overall, across 15 inverted telomeric repeat sites predicted by TelFuse in these cell

311 lines, 12 of these events (80%) could be validated as chromosomal arm fusion events  
312 using long-read genome sequencing (**Figure 3I, Table S6**).

313 We further investigated chromosomal arm fusion events for their concordance  
314 with spectral karyotyping results of the U2-OS cells. Consistent with the t(4;22) fusion  
315 seen in long-read sequencing (**Figure 3A-D**), a fusion between chromosome 22 and  
316 chromosome 4 was observed by spectral karyotyping in 5/10 cells assessed (**Figure**  
317 **3J**). As such, these results suggest that telomere-spanning chromosomal arm fusion  
318 events detected by long-read sequencing are concordant with the chromosomal scale  
319 observations.

320  
321 Length distribution of neotelomeres matches that of normal telomeres  
322

323 Because short telomeres lead to chromosomal fusion events, we hypothesized  
324 that neotelomeres would have similar lengths to unaltered telomeres at chromosome  
325 ends, whereas fusion events, which might have resulted from telomere attrition, would  
326 be shorter. To assess telomere length, we developed an approach (TelSize) to estimate  
327 the length of telomeric repeats in long read sequences (**Methods**) that accounts for  
328 noise in telomeric long-reads which are interspersed with errors and/or *bona fide*  
329 deviations from the standard “TTAGGG” repeat motif (**Figure S7A**).

330 Using TelSize, we can estimate the length of telomere repeat regions on single  
331 chromosomes. We applied the TelSize approach to establish the length of telomeres  
332 found at each of the chromosomal arms, and at intra-chromosomal telomeric sites. As  
333 the sub-telomeric region of the GRCh38 reference genome has not been fully  
334 assembled, we first assessed the reliability of assigning telomeric long reads to their  
335 respective arms for the CHM13 cell line for which the genome has been fully assembled  
336 (**Figure S7B**). TelSize was used to generate telomere length estimates for all of the cell  
337 lines with long read sequencing data (**Figure S8**).

338 We then assessed the length of telomeres at each neotelomere, at each natural  
339 telomere found on each chromosomal arm, and each chromosomal arm fusion event.  
340 For example, in a site of neotelomere addition at position chrX:103,320,553 in DNA  
341 from U2-OS cells that was described in an earlier (**Figure 2A-D**), TelSize predicts a  
342 telomere length of at least 4988 bp from a single nanopore read (**Figure 4A**). In a site  
343 of chromosome arm fusion between positions chr4:30,909,846 and the chr22 telomeric  
344 end (**Figure 3A-D**) in DNA from U2-OS cells, TelSize predicts a telomere length of 632  
345 bp from a single nanopore read (**Figure 4B**), with intra-chromosomal and sub-telomeric  
346 sequences flanking these sites. Most neotelomeres identified were multi-kilobasepair  
347 long with an average telomere length of ~5kb in both the U2-OS and the Hs-746T  
348 cancer cell lines (**Figure 4C-D, Figure S9A-B**). In contrast to neotelomeres and normal  
349 chromosomal arms, and consistent with our hypothesis, we see that telomeres at  
350 chromosomal arm fusion events tend to be relatively short and were largely only a few  
351 hundred base pairs long in U2-OS but longer in the small number of examples in Hs-  
352 746T (**Figure 4E-F, Figure S9C-D**), suggesting that chromosomal arms with short  
353 telomeres are more likely to undergo fusion events.

354 By composite analysis of data corresponding to each class of events, we see that  
355 structurally unaltered normal chromosomal ends (p- and q-arms) have similar median  
356 telomere length (~5kb) and similar length distribution to neotelomeres (**Figure 4G-H**,

357 **Figure S9E-F)** in both the U2-OS and Hs-746T cancer cell lines. Conversely, telomeric  
358 repeats at chromosomal arm fusions are significantly shorter as compared to the other  
359 classes of events (**Figure 4G-H, Figure S9E-F**). Together, these results indicate that  
360 neotelomeres have similar telomere length as natural telomeres and are thus possibly  
361 functional. Our results also suggest that chromosomal arms with short telomeres are  
362 more likely to undergo telomere-spanning chromosomal arm fusion events  
363

364 Somatically altered ectopic telomere repeat sequences in lung adenocarcinoma  
365 genomes  
366

367 Given the results of long read analysis that demonstrated both neotelomere  
368 events, corresponding largely to the standard telomeric repeat orientation, and  
369 telomere-spanning chromosome fusion events, corresponding largely to the inverted  
370 telomeric repeat orientation, in cancer cell lines, we sought to determine whether similar  
371 events could be observed as somatic genome alterations in primary human cancers.  
372 We applied TelFuse to 95 pairs of lung adenocarcinoma tumor/normal genome  
373 sequences from The Cancer Genome Atlas, or TCGA (TCGA-LUAD) (**Table S8**). This  
374 analysis identified 34 sites with ectopic telomere sequences in the standard orientation,  
375 and 46 sites with ectopic telomere sequences in the inverted orientation (**Tables S9**  
376 and **S10**). Putative sites of ectopic telomeric repeat sequences could be seen across  
377 the genome on almost all chromosome arms, without a particular distribution in the  
378 genome at this resolution of sample number and events (**Figure 5A**). These ectopic  
379 telomere sequences, in both the standard and inverted orientations, could be in either  
380 the centromeric or counter-centromeric direction (**Figure 5A**).  
381

382 Among the standard orientation ectopic telomere repeats in the TCGA-LUAD  
383 sequence data, 32/34 sites were confirmed as somatic alterations and therefore as  
384 putative somatically generated neotelomeres by comparing the lung adenocarcinoma  
385 DNA sequence with the matched normal sequence. In addition, 44/46 of the inverted  
386 orientation repeats were confirmed as somatic alterations that are likely to represent  
387 telomere-spanning chromosomal arm fusions (**Figure 5B**). Together, among the set of  
388 80 potential neotelomeres and chromosomal arm fusion events detected in the TCGA-  
389 LUAD tumor samples, we found that 72/80 (90%) events were only detected in the  
390 tumor sample (**Figure 5B, Table S9**), suggesting that a large majority of calls made in  
391 tumor samples by TelFuse are somatic, even though no matched normal samples were  
392 assessed in our initial analysis.

393 We then performed a deeper inspection of these somatic ectopic telomere repeat  
394 sites that were detected in primary tumors. At the ectopic telomeric repeat site at  
395 chr1:214,760,486 in the patient TCGA-44-4112, at least 10 TTAGGG repeats could be  
396 observed in the primary tumor by short reads, coupled with a drop in sequencing  
397 coverage (**Figure 5C**), which is consistent with the presence of a neotelomeric site. At  
398 another site chr17:31,537,163 in the patient TCGA-49-4507, at least 6 inverted  
399 telomeric repeats of TTAGGG could be seen in the primary tumor sample by short-  
400 reads (**Figure 5D**), which may indicate the presence of a chromosomal arm fusion  
401 event given our observations with long-read sequencing of cancer cell lines. Notably,  
402 similar observations were also made at other sites with somatic ectopic telomeric repeat  
sequences that are consistent with potential neotelomeric or chromosomal arm fusion

403 events (**Figure S10**) in primary lung adenocarcinoma samples. Together, this suggests  
404 that ectopic telomeric repeats in both the standard and inverted orientation can be  
405 readily observed in primary lung adenocarcinoma samples, and suggest that  
406 neotelomeres and chromosomal arm fusion events are similarly present in primary  
407 tumor samples.

408 All together, we observed ectopic telomeric repeats in the standard orientation  
409 and inverted orientation in 26% and 31% of the TCGA-LUAD cohort respectively  
410 (**Figure 5E**), which may point to the potential existence of neotelomeric events and  
411 chromosomal arm fusions in these samples respectively. Of note, as many as 49% of  
412 samples displayed either a neo-telomeric or chromosomal arm fusion signal, suggesting  
413 that these events are relatively common in primary tumor samples. Although this  
414 suggests an active mechanism for generation of telomeric events in cancers, we were  
415 unable to ascertain strong sequence signatures suggestive of specific telomere  
416 insertion mechanisms (**Figure S11**).

417

#### 418 Germline variations leading to ectopic telomeric repeat insertions

419 Interestingly, we also observed 8 likely germline examples of ectopic telomere  
420 repeat sequence alterations across 4 different individuals in the TCGA-LUAD cohort  
421 (**Figure 5B, Table S9**). A deeper exploration of these events was performed to assess  
422 the structure and features associated with these sites (**Figure S12**). Two ectopic  
423 telomeric sites were found on the chr12q arm in both blood and tumor samples of  
424 TCGA-44-6778 at the sites chr12:54,480,142 and chr12:54,494,011, and were noted to  
425 contain a 14 kb deletion, coupled to an insertion of 6x CCCTAA repeat sequences  
426 (**Figure S12A**). In both blood and tumor samples of the same individual at the sites  
427 chr12:25,085,740 and chr12:25,085,754 on chr12p, an insertion of 7x CCCTAA repeats  
428 was observed in tandem with duplication of a neighboring 14 bp region (**Figure S12B**).  
429 A similar germline deletion event of 13 bp, coupled with the insertion of telomeric repeat  
430 sequences, was found in TCGA-62-A470 at chr4:184,711,090 (**Figure S12C**), while a  
431 duplication of 19 bp was coupled to a telomeric repeat insertion at chr6:170,186,789 in  
432 TCGA-44-5643 (**Figure S12D**). Ectopic telomeric repeats could also be observed in  
433 TCGA-55-6987 at low allelic frequencies in both tumor and the adjacent normal sample  
434 (**Figure S12E**), which may point to contamination of the normal sample or to somatic  
435 mosaicism. Together, these results indicate that ectopic telomeric repeats might be  
436 frequent germline variants, perhaps as a result of DNA repair in the presence of active  
437 telomerase<sup>41</sup>.

438

#### 439 Neotelomeres and chromosomal arm fusion events disrupt protein coding genes and 440 are highly prevalent in cancer cell lines

441 In addition to allowing chromosome fusions to occur and capping truncated  
442 chromosomes, insertion of telomeric DNA might also disrupt genes, including tumor  
443 suppressors, leading to associated functional impact. To assess this possibility, we  
444 evaluated ectopic telomere sites in this study for overlap with protein coding genes.  
445 Among sites that we detected, 47% (112/240) and 47% (34/72) of sites were found to  
446 colocalize to a protein coding gene in cancer cell lines and primary lung  
447 adenocarcinomas respectively (**Table S2 and S9**).

448 Notable genes with insertion events include *PTPN2*, a gene related to  
449 immunotherapy response<sup>48</sup>, where a neotelomere was found within the first intron,  
450 leading to a corresponding loss of the first exon and the promoter region (**Figure 6A**).  
451 Chromosomal fusion events were also found to disrupt genes, including events that led  
452 to the loss of more than half of the 5' region of the *KLF15* and *FOXN3* genes (**Figure**  
453 **6B-C**). We also observed one complex event involving telomeric DNA wherein a short  
454 neotelomere on chr1p within the *RUNX3* gene then fused to the centromere of  
455 chr22/21/14. This event caused the loss of most of the gene (**Figure 6D**). Gene  
456 disruption events were also observed by long-read genome sequencing in the *NRDC*  
457 and *TENM4* genes in cancer cell lines (**Figure S13A-B**). Interestingly, the *PTPN2*,  
458 *NRDC*, *FOXN3*, and *RUNX3* genes identified in our study have putative functional roles  
459 in cancer, suggesting that the disruption of protein coding genes by neotelomeres and  
460 chromosomal arm fusions may contribute to tumorigenesis. Thus, our results indicate  
461 that neotelomeres and chromosomal arm fusion may represent an important but poorly  
462 appreciated mechanism for gene disruption.

463 We next assessed if these gene disruption events from telomeric insertion can  
464 also be observed in primary tumor samples. In the lung adenocarcinoma sample TCGA-  
465 62-A46O, a putative neotelomere could be observed using short-read data within the  
466 gene encoding the ETS family transcription factor, *ETV6* which is known to be  
467 associated with leukemia and congenital fibrosarcoma<sup>47,49,50</sup> (**Figure 6E**). Another  
468 putative neotelomere event was observed within the gene encoding centromere protein  
469 F, *CENPF* which is thought to play a role in chromosome segregation during mitosis<sup>51-53</sup>  
470 (**Figure 6F**). Putative neotelomeres and chromosomal arm fusion events were also  
471 found within the protein arginine methyltransferase gene, *PRMT7*, and the forkhead box  
472 transcription factor, *FOXP4*, genes respectively (**Figure S13C-D**). Of note, due to the  
473 size and scale at which these neotelomeres and chromosomal arm fusion events occur,  
474 they are likely to fully disrupt these genes. Therefore, our results indicate that the  
475 formation of neotelomeres and telomere-spanning chromosomal arm fusions may  
476 represent a mechanism for gene disruption, in addition to their roles in defining gross  
477 chromosomal structure.

478  
479  
480  
481  
482

483 **Discussion**

484

485 While alterations in telomere sequences are key events in cancer genome  
486 evolution, the precise nucleotide-level structure of these alterations has been hitherto  
487 inaccessible because of the inability of short-read sequence data to resolve longer  
488 repetitive sequences. Here, using long-read sequencing technologies, we delineated  
489 four types of alterations in telomere repeat sequences. First, we provide evidence that  
490 cancer cell line and primary cancer genomes contain long (several kilobase) additions  
491 of telomere repeat sequences to intra-chromosomal sites, in the standard telomere  
492 orientation (**Figure 7A**). Second, we identify telomeric repeat sequences of varying  
493 length that bridge the end of one chromosome to an intra-chromosomal site on a  
494 different chromosome (**Figure 7B**). These telomeric repeats are consistent with  
495 karyotyping analyses that have observed the attachment of chromosomal fragments to  
496 the ends of existing chromosomes<sup>54-59</sup>, which are key events in cancer genome  
497 evolution. Third, we observe more complex alterations where the formation of a  
498 neotelomere is followed by the fusion of the neotelomere to a second intra-  
499 chromosomal location (**Figure 7C**). Fourth, we observe fusions that link centromeric to  
500 telomeric sequence repeats (**Figure 7D**). The implications of several of these alterations  
501 are described below.

502 A previous study, analyzing short-read genome sequencing of patients' cancer  
503 samples from the Pan-Cancer Analysis of Whole Genomes (PCAWG) project, was able  
504 to identify a number of intra-chromosomal telomeric repeat insertion sites<sup>23</sup>. In  
505 comparison to this previous study, our work shows that telomere length at these repeat  
506 insertion sites can be estimated using long-read sequencing, and the underlying  
507 sequence structure can be analyzed in the context of adjacent sequences compared to  
508 free telomeric ends. This technical advance allowed us to differentiate intra-  
509 chromosomal telomeric repeat sites based on the orientation of the telomeric repeat  
510 sequences. By integrative analysis of long-read genome sequencing, spectral  
511 karyotyping, coverage analysis, and short-read genome sequencing, we demonstrated  
512 the existence of multi-kilo base-pair long neotelomeres at sites of putative chromosomal  
513 arm breakages, corresponding to telomeric repeats in the standard orientation. We  
514 further provided evidence for the presence of these standard orientation telomere  
515 repeats representing neotelomeres in primary tumor sequence data of lung  
516 adenocarcinoma (LUAD) from TCGA. Further, powered by long-read genome  
517 sequencing, we were able to reliably show that sites with inverted telomeric repeat  
518 sequences represent fusion of chromosomal arms spanning short telomere sequence  
519 repeats, also found in TCGA LUAD data. Together, our study provides support for the  
520 existence of neotelomeres and chromosomal arm fusion events in cancer genomes,  
521 and also provides insights into the cause of their occurrence.

522 A recent experimental study generated double-strand breaks in cells over-  
523 expressing telomerase, leading to the addition of neotelomeres at a subset of these  
524 breaks<sup>41</sup>. Our study provides genomic evidence for a signature of neotelomere addition  
525 in cancer cell lines and cancer genomes, complementary to this experimental evidence.  
526 The location and the unbounded structure of these repeats suggest that they are likely  
527 to be functional neotelomeres. Taken together, the cellular experiments and genomic

528 observations support a model where neotelomere addition by telomerase, nucleating at  
529 sites of double strand breaks, can be a common step in tumorigenesis.

530 The generation of new chromosomes via chromosomal rearrangements is a key  
531 element of cancer genome evolution and also occurs during the course of evolution and  
532 speciation<sup>60-62</sup>. Some of our findings using long-read sequencing of cancer genomes  
533 mirror long-standing observations in genomes of many organisms. Interstitial telomeric  
534 repeats have been identified in the genomes of many vertebrates, including primates  
535 and the pygmy tree shrew<sup>63-65</sup>, akin to those found at sites of chromosomal arm fusions  
536 in cancer cell lines (**Figure 7B**). Furthermore, interstitial telomeric sequences have been  
537 observed close to centromeres in the genomes of diverse organisms including Chinese  
538 hamster, Arabidopsis, and the European grayling<sup>65-67</sup>. These structures, termed  
539 pericentromeric telomeric repeats, were similarly observed by long-read genome  
540 sequencing in the U2-OS cancer cell line in our study (**Figure 7D**). Overall, the study of  
541 telomere repeat alterations also provide an understanding into how new chromosomes  
542 originate during the course of evolution and speciation, as well as during cancer  
543 genome evolution.

544 Looking at the genome beyond telomeric repeats, repetitive elements constitute  
545 approximately half the human genome<sup>13-15</sup>. However, we have not yet been able to  
546 understand genome structure and alterations at a detailed level because of the  
547 technical limitations of short-read sequencing, which is unable to span or completely  
548 delineate the precise structure of these repeat elements. Here, using telomeres as a  
549 salient example, we show how long-read genome sequencing can be used to drive  
550 discoveries of functional importance in highly repetitive regions of the cancer genome,  
551 and also inform the analysis of existing short-read data. As a bridge to a future where  
552 universal long-read sequencing is technically and economically feasible, our study  
553 provides a framework to assess short-read genome sequencing data for genome  
554 alterations within highly repetitive regions, that can be followed by long-read sequencing  
555 and complete analysis of selected samples. Significantly, given that >95% of repetitive  
556 sequences in the genome are estimated to be <8 kb in length<sup>15</sup>, long-read sequencing  
557 data that is typically generated at >10 kb in length (**Figure S3**) would enable the  
558 majority of previously neglected alterations in the cancer genome to be completely  
559 resolved. Thus, our study highlights the utility of long-read genome sequencing in the  
560 study of chromosomal scale structures in cancer and beyond. This analysis may have  
561 functional implications as we observed the disruption of protein coding genes by  
562 neotelomeres and chromosomal arm fusions. More broadly, the identification of these  
563 gene disruptions points to the potential role that other repetitive elements may play in  
564 gene disruption as well as activation events and to the discovery opportunity provided  
565 by long-read cancer genome sequencing.

566 There are a few limitations associated with our study. First, in contrast to a recent  
567 yeast genomic study in which the end of each telomere was tagged<sup>68</sup>, it is difficult to  
568 assess if telomeric repeats containing long-reads analyzed in our study captured the  
569 telomeres end-to-end. As such, telomere length estimates made in our study may  
570 underestimate the true length of telomeres. Further, it is also known that the sub-  
571 telomeres at normal chromosomal arms contain telomere-like sequences and short  
572 internal telomeric repeats close to long stretches of perfect (TTAGGG)<sub>n</sub> repeats<sup>44,69</sup>.

573 However, it is unclear if these sequences should be included in the computation of  
574 telomere length estimates performed in our study.

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

584  
585

586 **Acknowledgements**

587 We thank all members of the Matthew Meyerson and Heng Li labs for helpful comments  
588 and inputs on the work. We would also thank Jidong Shan (Albert Einstein College of  
589 Medicine) for generating spectral karyotyping results, and for inputs on the analysis of  
590 cytogenetics data. We further thank Jodi Hirschman for assistance with edits to our  
591 manuscript.

592

593 **Funding**

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

600

601 **Author contributions**

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

610

611

612 **Declaration of interests**

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

620

621 **Figure Legends**

622  
623 **Figure 1 Classes of ectopic telomeric repeats found in cancer cell genomes.** (A)  
624 Schematic of sequence and positions of normal telomeres at chromosomal termini. (B)  
625 Schematic of ectopic telomeric repeats found at abnormal locations away from  
626 chromosomal termini. Standard orientation: (TTAGGG)<sub>n</sub> on the right side of a breakpoint  
627 and (CCCTAA)<sub>n</sub> on the left side of the breakpoint in the 5' to 3' direction (same as  
628 normal telomere in Fig. 1A). Inverted orientation: (CCCTAA)<sub>n</sub> on the right side of a  
629 breakpoint and (TTAGGG)<sub>n</sub> on the left side of the breakpoint in the 5' to 3' direction.  
630 Note that faded chromosomal segment is not part of derivative chromosome. (C)  
631 Genome-wide localization of ectopic telomeric repeats in cancer cell line genomes  
632 (n=326) identified using short-read genome sequencing. Red: ectopic telomeric  
633 sequences in the standard orientation. Blue: ectopic telomeric sequences in the inverted  
634 orientation. Position of telomeric repeats relative to the breakpoint is indicated by arrows  
635 oriented in different directions. (D) Percentage of cancer cell lines in the CCLE with  
636 ectopic telomeric sequences in either orientation. Total sample number as indicated. (E)  
637 Flow-chart of long-read genome sequencing and cytogenetic analyses in cancer cell  
638 lines, with the indicated validation criteria.  
639

640 **Figure 2 Neotelomeres in cancer genomes revealed by long-read genome**  
641 **sequencing.** (A-H) Genomic analysis of telomere repeat alterations in the standard  
642 orientation that were detected (A-D) in the U2-OS osteosarcoma cell line at  
643 chrX:103,320,553, and (E-H) in the Hs-746T cell line at chr21:10,547,397. (A) IGV  
644 screenshots of short-read genome sequencing data. Ectopic telomeric repeats  
645 (TTAGGG)<sub>n</sub> are shown in color. (B) Sequencing coverage and allelic ratios of  
646 chromosome X. Orange semi-oval: site of the neotelomeric event. (C) IGV screenshots  
647 depicting long telomeric repeat sequences (TTAGGG)<sub>n</sub> with PacBio HiFi and Nanopore  
648 long-read sequencing at the site shown in (A). (D) Schematic of neotelomere location  
649 on chromosome Xq. (E) IGV screenshots of short-read genome sequencing data.  
650 Ectopic telomeric repeats (CCCTAA)<sub>n</sub> are shown in color. (F) Sequencing coverage and  
651 allelic ratios of chromosome 21. Orange semi-oval: site of the neotelomeric event. (G)  
652 IGV screenshots depicting long telomeric repeat sequences (CCCTAA)<sub>n</sub> with PacBio  
653 HiFi and Nanopore long-read sequencing at the site shown in (E). (H) Schematic of  
654 neotelomere location on chromosome 21p. (I) Percentage of ectopic telomeric repeat  
655 sites in the standard orientation, found by short-read genome sequencing using  
656 TelFuse, that were validated by long-read genome sequencing. (J) Spectral karyogram  
657 of chrX in ten U2-OS single cells assessed by spectral karyotyping with corresponding  
658 karyotype labels. First label: total # of X chromosomes and their derivatives observed in  
659 given cell. Second label: karyotypes of the aberrant X chromosomes or derivatives.  
660 Asterisk (\*): truncated X chromosome. See also Figure S4.

661  
662 **Figure 3 Chromosomal arm fusions in cancer genomes revealed by long-read**  
663 **genome sequencing.** (A-H) Genomic analysis of telomere repeat alterations in the  
664 inverted orientation that were detected in the U2-OS osteosarcoma cell line (A-D) at the  
665 site chr4:30,909,846, and (E-H) at the site chr11:84,769,636. (A) IGV screenshots of  
666 short-read genome sequencing data. Ectopic telomeric repeats (CCCTAA)<sub>n</sub> are shown

667 in color. **(B)** Sequencing coverage and allelic ratios of chromosome 4. Orange semi-  
668 oval: site of the ectopic telomere repeat sequence. **(C)** IGV screenshots of PacBio HiFi  
669 and Nanopore long-read sequencing data at the site shown in **(A)**. Ectopic telomeric  
670 repeats in the inverted orientation contained ~650 bp of (CCCTAA)<sub>n</sub> telomeric repeat  
671 sequences followed by chr22q sub-telomeric sequences, indicative of a chromosomal  
672 arm fusion event of chr22q to the site at chr4:30,909,846. **(D)** Schematic of telomere-  
673 spanning fusion event between chromosomes 22q-ter and 4p. **(E)** IGV screenshots of  
674 short-read genome sequencing data. Ectopic telomeric repeats (CCCTAA)<sub>n</sub> are shown  
675 in color. **(F)** Sequencing coverage and allelic ratios of chromosome 11. Orange semi-  
676 oval: site of the ectopic telomere repeat sequence. **(G)** IGV screenshots of PacBio HiFi  
677 and Nanopore long-read sequencing at the site shown in **(E)**. ~1750 bp of (CCCTAA)<sub>n</sub>  
678 telomeric repeat sequences are found sequences corresponding to chr11p  
679 (chr11:43,002,345), suggestive of a complex event consistent with the formation of a  
680 neotelomere on chr11p, followed by a chromosomal arm fusion event of this  
681 neotelomere to the site on chr11q (chr11:84,769,636). **(H)** Schematic telomere-  
682 spanning fusion event between chromosome arms 11q (with a predicted neotelomere)  
683 and 11p. **(I)** Percentage of new telomeric sites in the inverted orientation that were  
684 predicted by TelFuse from short-read genome sequencing, and then validated by long-  
685 read genome sequencing as telomere-spanning chromosome arm fusion events. **(J)**  
686 Spectral karyogram of chromosome 22 for which a chromosomal arm fusion was  
687 detected with chromosome 4. Ten U2-OS single cells assessed are as indicated. The  
688 fusion event between chromosome 22 (yellow) and chromosome 4 (blue) is indicated by  
689 a red arrow. See also Figure S6.  
690

691 **Figure 4 Neotelomeres have similar telomere length distribution as normal**  
692 **telomeres, while telomeric repeats at sites with chromosomal arm fusions are**  
693 **short. (A-B)** Telomeric repeat signal observed at a representative Nanopore read with  
694 **(A)** a neotelomere in U2-OS DNA at chrX:103,320,553, and **(B)** a chromosomal arm  
695 fusion event in U2-OS DNA at chr4:30,909,846. The length of telomeric repeats on each  
696 long-read was estimated from these telomeric repeat signal profiles. Boxplots depicting  
697 the distribution of telomere length found at each neotelomere assessed by Nanopore  
698 sequencing for the **(C)** U2-OS and **(D)** Hs-746T cell lines. Boxplot depicting length of  
699 telomeric repeats assessed using Nanopore sequencing for each chromosomal arm  
700 fusion event in the **(E)** U2-OS and **(F)** Hs-746T cell lines. Note: telomere length for  
701 neotelomeres and normal chromosomal arms were only estimated using long-reads  
702 reads that start or end in telomeric repeats, while length of telomeric repeats at  
703 chromosomal arm fusions were estimated using long-reads with telomeric repeats in the  
704 middle of the read. Aggregated telomeric length of all long-reads at the normal  
705 chromosomal arms (p- and q-arms), neotelomeres, and chromosomal arm fusion events  
706 in the **(G)** U2-OS and **(H)** Hs-746T cell lines. P-values indicated in the plots were  
707 calculated using the two-sided Wilcoxon Rank Sum test. See also Figure S9.  
708

709  
710 **Figure 5 Putative neotelomeres and chromosomal arm fusion events are detected**  
711 **as somatic alterations in primary lung adenocarcinoma genomes. (A)** Genome-  
712 wide distribution of putative neotelomeres and chromosomal arm fusion events in lung

713 adenocarcinoma patient samples from The Cancer Genome Atlas (TCGA) (n=95).  
714 Neotelomeres were inferred from ectopic telomeric sequences in the standard  
715 orientation, while chromosomal arm fusion events were inferred from ectopic telomeric  
716 sequences in the inverted orientation, as described in Figure 1B, using short-read  
717 genome sequencing data. **(B)** Proportion of telomeric alterations (neotelomeres/arm  
718 fusions) that were found to be germline or somatic. **(C-D)** Examples of neotelomeres  
719 and chromosomal arm fusion events detected in tumor samples from patients with lung  
720 adenocarcinoma. **(C)** Neotelomere in tumor DNA from case TCGA-44-4112 at the site  
721 chr1:214,760,486. **(D)** Chromosomal arm fusion in tumor DNA from case TCGA-49-  
722 4507 at the site chr17:31,537,163. Top panels: sequencing coverage at the sites of  
723 interest. Bottom panels: IGV screenshots corresponding to the neotelomere or  
724 chromosomal arm fusion events in the normal and tumor samples. **(E)** Frequency of  
725 neotelomeres and chromosomal arm fusion events in lung adenocarcinoma patient  
726 tumor samples in TCGA.  
727

728 **Figure 6 Neotelomeres and chromosomal arm fusion events disrupt protein**  
729 **coding genes in cancer cell lines and patient samples. (A)** Disruption of the *PTPN2*  
730 gene in the U2-OS osteosarcoma cell line at chr18:12,875,538 with addition of a  
731 neotelomere. **(B)** Disruption of the *KLF15* gene in the Hs-746T gastric adenocarcinoma  
732 cell line associated with a chromosomal arm fusion event at chr3:126,349,603. **(C)** A  
733 chromosomal arm fusion event in the U2-OS cell line between a broken chromosome  
734 14 and the telomere arm of chromosome 21q/22q/19q associated with disruption of the  
735 *FOXN3* gene at chr14:89,300,563. **(D)** A neotelomere in the U2-OS cell line coupled to  
736 fusion to a centromere leads to disruption of the *RUNX3* gene at chr1:24,906,321. **(E)** A  
737 putative neotelomere associated with disruption of the *ETV6* gene in a lung  
738 adenocarcinoma tumor sample derived from the patient TCGA-62-A46O at the site  
739 chr12:11,696,012. **(F)** A putative neotelomere associated with disruption of the *CEPF*  
740 gene in a lung adenocarcinoma tumor sample derived from the patient TCGA-53-7624  
741 at the site chr1:214,609,478. See also Figure S13.  
742

743 **Figure 7 Possible models that can account for the different types of telomeric**  
744 **repeat sequences observed in this study. (A)** A neotelomere can form after a  
745 chromosomal arm breakage event. This leads to the generation of a smaller  
746 chromosome with a neotelomere, similar in repeat length to telomeres found on a  
747 normal chromosomal arm. **(B)** Chromosome arm fusion where a broken chromosomal  
748 arm can fuse to another chromosome with very short telomeres. This generates a larger  
749 chromosome with interstitial telomeric repeat sequences in the middle of the  
750 chromosome. **(C)** Complex alteration where neotelomere formation is followed by the  
751 fusion of this neotelomere to another chromosomal fragment. This leads to the  
752 observation of long-reads in our study which contains telomeric repeat sequences,  
753 flanked on both sides by intra-chromosomal sequences. **(D)** A complex telomeric  
754 alteration involving a chromosomal arm break at or very near to the centromere, which  
755 is fused to another chromosomal arm with very short telomeres. The resultant new  
756 chromosome has pericentromeric telomeric repeat sequences. Purple line: parts of the  
757 model supported by long-read genome sequencing data.  
758

759

760 **STAR Methods**

761

762 **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	Cat# 101-908-100

	(PacBio)	
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	To be uploaded to SRA database (pending accession number)
Nanopore MinION long-read sequencing dataset	This paper	To be uploaded to SRA database (pending accession number)
PacBio HiFi long-read sequencing datasets	This paper	To be uploaded to SRA database (pending accession number)
Illumina short-read sequencing datasets	This paper	To be uploaded to SRA database (pending accession number)
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>70,71</sup>	<a href="https://gdc.cancer.gov/about-data/publications/pancanatlas">https://gdc.cancer.gov/about-data/publications/pancanatlas</a>
dbSNP (build 151)	Sherry et al <sup>72</sup>	<a href="ftp://ftp.ncbi.nlm.nih.gov/snps/organisms/human_9606_b151_GRCh38p7/VCF/common_all_20180418.vcf.gz">ftp://ftp.ncbi.nlm.nih.gov/snps/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>73</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™
NCI-BL1184 cells	American Type Culture	Cat# CRL-5949™

NCI-H1184 cells	Collection (ATCC)	
	American Type Culture Collection (ATCC)	Cat# CRL-5858™
Hs-746T cells	American Type Culture Collection (ATCC)	Cat# HTB-135
<b>Software and algorithms</b>		
TelFuse	This paper	<a href="https://github.com/ktan8/tel_tools/">https://github.com/ktan8/tel_tools/</a>
TelSize	This paper	<a href="https://github.com/ktan8/tel_tools/">https://github.com/ktan8/tel_tools/</a>
Minimap2 v2.17-r941	Li <sup>74</sup>	<a href="https://github.com/lh3/minimap2">https://github.com/lh3/minimap2</a>
BWA-MEM v0.7.17-r1188	Li <sup>75</sup>	<a href="https://github.com/lh3/bwa">https://github.com/lh3/bwa</a>
SAMtools v1.10	Li et al <sup>76</sup>	<a href="https://github.com/samtools/samtools">https://github.com/samtools/samtools</a>
R v4.2.0	R Foundation for Statistical Computing <sup>77</sup>	<a href="https://www.r-project.org/">https://www.r-project.org/</a>
Python v3.7.4	Van Rossum et al <sup>78</sup>	<a href="https://www.python.org/">https://www.python.org/</a>
Perl v5.26.2	Wall et al <sup>79</sup>	<a href="http://www.perl.org/">http://www.perl.org/</a>
Integrative Genomics Viewer (IGV)	Thorvaldsdóttir et al <sup>80</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>81</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)	-

763

764

## 765 Resource availability

766

### 767 Lead contact

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

770

### 771 Materials availability

772 This study did not generate new unique reagents.

773

774

775

776 **Methods details**

777

778 CCLE whole genome sequencing dataset

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

783

784 Lung adenocarcinoma whole genome sequencing dataset

785 Whole genome short-read sequencing dataset of lung adenocarcinoma patients<sup>70,71</sup>  
786 from The Cancer Genome Atlas were downloaded from the GDC Data Portal  
787 (<https://portal.gdc.cancer.gov/>). The list of accession numbers corresponding to  
788 samples analyzed for this study is indicated in Table S8.

789

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

792 Candidate short read pairs with at least two consecutive telomeric repeat sequences  
793 (TTAGGG)<sub>2</sub> in either reads in the pair were first extracted to narrow down the number of  
794 read pairs for subsequent analysis. Specifically, this was done by applying a custom  
795 Python script in the TelFuse package to each whole genome sequencing dataset.

796

797 Candidate read pairs were then remapped to the reference genome (GRCh38) with  
798 BWA-MEM (version 0.7.17-r1188)<sup>75</sup> with default parameters. A custom Python script in  
799 the TelFuse package was then used to extract all sites with soft-clipped regions on the  
800 mapped reads. Soft-clipped sequences from all reads at each unique genomic site was  
801 then used to generate a consensus sequence. A corresponding average sequence  
802 identity of the soft-clipped sequences to the consensus was also calculated.

803

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

812

813 The candidate sites were then further subdivided into sites with telomeric repeats in the  
814 standard or inverted orientation, depending on the orientation of telomeric repeat  
815 sequences with respect to the genomic loci of interest.

816

817 Cell culture

818 U2-OS cells (ATCC® HTB-96™) were cultured in McCoy’s 5A Medium Modified (ATCC  
819 cat no. 30-2007) with 10% FBS. Cell lines NCI-BL1184 (ATCC cat no. CRL-5949™) and  
820 NCI-H1184 (ATCC cat no. CRL-5858™) were cultured in ATCC-formulated RPMI-1640  
821 Medium (ATCC cat no. 30-2001) supplemented with FBS at 10%. Hs-746T cells (ATCC

822 cat no. HTB-135) were cultured in ATCC-formulated Dulbecco's Modified Eagle's  
823 Medium (ATCC cat no. 30-2002) supplemented with 10% FBS

824

#### 825 High molecular weight DNA extraction

826 High molecular weight (HMW) DNA was isolated using a Monarch® Genomic DNA  
827 Purification Kit (NEB, cat no. T3010S). DNA was quantified with a Qubit™ HS dsDNA  
828 assay (ThermoFisher, cat no. Q32851) followed by verification of HMW DNA integrity by  
829 electrophoresis on an Agilent 4200 TapeStation (Genomic DNA ScreenTape, cat no.  
830 5067-5366).

831

#### 832 MinION Library Preparation

833 Sequencing libraries were prepared for the Oxford Nanopore Technologies (ONT)  
834 platform using the ONT Genomic DNA Ligation kit (ONT, cat no. SQK-LSK109).  
835 Briefly, HMW U2OS DNA was fragmented to ~20 Kb using a Covaris® g-TUBE (cat no.  
836 520079) followed by SPRI-cleanup (Agencourt® AMPure XP, Beckman Coulter, cat no.  
837 A63881). Fragmented material was quantified with a Qubit™ dsDNA HS Assay Kit  
838 (Invitrogen™, Catalog number: Q32851). One microgram of HMW U2OS DNA was end-  
839 repaired and A-tailed (NEBNext® Companion Module for  
840 Oxford Nanopore Technologies® Ligation Sequencing, cat no. E7180S) followed by  
841 adapter ligation. For sequencing 100 fmols of library material was loaded on an R9 flow  
842 cell (cat no. FLO-MIN106D).

843

#### 844 PromethION Library Preparation

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

857

#### 858 PacBio HiFi Library Preparation

859 For CCS library preparation, ≥3 ug of high molecular weight genomic DNA (more than  
860 50% of fragments ≥40 kb) was sheared to ~15 kb using the Megaruptor 3 (Diagenode  
861 B06010003), followed by DNA repair and ligation of PacBio adapters using the PacBio  
862 SMRTbell Express Template Prep Kit 2.0 (100-938-900) and removal of incomplete  
863 ligation products with the SMRTbell Enzyme Clean Up Kit 2.0 (PacBio 101-938-500).  
864 Libraries were then size-selected for 15 kb +/- 20% using the PippinHT with 0.75%  
865 agarose cassettes (Sage Science). Following quantification with the Qubit dsDNA High  
866 Sensitivity assay (Thermo Q32854), libraries were diluted to 60 pM per SMRT cell,  
867 hybridized with PacBio V5 sequencing primer, and bound with SMRT seq polymerase

868 using Sequel II Binding Kit 2.2 (PacBio 101-908-100). CCS sequencing was performed  
869 on the Sequel IIe instrument using 8M SMRT Cells (101-389-001) and Sequel II  
870 Sequencing 2.0 Kit (101-820-200), utilizing PacBio's adaptive loading feature with a 2  
871 hour pre-extension time and 30 hour movie time per SMRT cell. Initial quality filtering,  
872 basecalling, adapter marking, and CCS error correction was done automatically on  
873 board the Sequel IIe.

874

875 Base calling of Nanopore sequencing data

876 Base calling of Nanopore sequencing data in this study was performed using Bonito  
877 (Version 0.3.5) with the default dna\_r9.4.1 basecalling model. However, the default  
878 Nanopore basecalling model leads to frequent strand-specific base calling errors at  
879 telomeric repeats in our dataset, with  $(TTAGGG)_n$  being miscalled as  $(TTAAAA)_n$ , and  
880  $(CCCTAA)_n$  being miscalled as  $(CTTCTT)_n$  and  $(CCCTGG)_n$ , akin to what we had  
881 previously reported<sup>81</sup>. As such, telomeric reads were extracted using a pipeline that we  
882 had previously developed, followed by re-basecalling using a basecalling model that  
883 was previously tuned to correct these errors<sup>81</sup>.

884

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

886 Long reads containing telomeric repeats were extracted by first enumerating the  
887 number of  $(TTAGGG)_2$  and  $(CCCTAA)_2$  motifs on each read using custom Perl scripts.  
888 Long reads containing at least four of these motifs were then defined as candidate  
889 telomeric repeats. Of note, a low cutoff was deliberately set here to more sensitively  
890 identify long-reads with telomeric repeats for detailed analysis by TelSize.

891

892 Estimation of telomere length from noisy long reads

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

904

905 Specifically, long-reads were classified into five different classes: full telomeric – long-  
906 reads that contains telomeric repeat sequences end-to-end, left telomeric – long-reads  
907 that contains telomeric repeat sequences on the left edge of the long-read, right  
908 telomeric – long-reads that contains telomeric repeat sequences on the right edge of the  
909 long-read, intra-telomeric – long-reads that contains telomeric repeat sequences in the  
910 middle of the single long-read, and non-telomeric – long-reads that do not contain  
911 significant telomeric repeat signal throughout the long-read. These telomeric repeat  
912 signal can also occur as either  $(TTAGGG)_n$  or  $(CCCTAA)_n$  repeats, and these  
913 information are further reported.

914

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

917

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

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

928

929 For these telomeric repeat containing reads, sequences corresponding to the telomeric  
930 repeat region were trimmed off. The remaining non-telomeric sequences of each read  
931 were then mapped to the GRCh38 reference genome with minimap2 (Version 2.17-  
932 r941). Primary read mappings in the PAF format were then extracted and analyzed  
933 using custom R scripts in order to assess mapping coordinates of these sequences. For  
934 each site of interest that was identified using short-read data, telomeric repeat  
935 containing long-reads that mapped to a  $\pm 100$  bp region of each site were extracted.  
936 Telomere length estimates for long-reads at each neotelomeric and chromosomal arm  
937 fusion sites were then reported as per Figure 4.

938

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

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

948

#### 949 Copy number profiles

950 To generate copy number profiles of the cancer cell lines from the CCLE, the total  
951 sequencing coverage of each 10 kb bin was calculated using the bedcov function  
952 SAMtools (v1.10)<sup>76</sup> with default parameters. The coverage was then normalized to a  
953 per-basepair level and is as depicted.

954

955 For lung adenocarcinoma samples which has a matched normal samples, the  
956 normalized sample coverage across each chromosome was calculated as follows. The  
957 sequencing coverage for each 10kb bin was calculated for both the tumor and matched  
958 normal sample using the bedcov function in SAMtools (v1.10)<sup>76</sup>. These values were

959 then normalized by the total read count of each dataset, and the ratio between the  
960 tumor and normal sample calculated to obtain the normalized sample coverage.  
961

## 962 Analysis of BAF

963 As no matched normal samples were sequenced for each of the cancer cell lines,  
964 heterozygous germline variants cannot be directly assessed and used in the generation  
965 of allelic ratio plots. Allelic ratios was thus assessed using a set of common germline  
966 SNPs from the dbSNP database (GRCh38.p7 build 151)<sup>72</sup>  
967 ([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  
968 database as SNPs that are found with a minor allele frequency of at least 0.01 in the  
969 1000 genomes project.  
970

971 Custom Python scripts and SAMtools mpileup (v1.10) were then used to enumerate all  
972 four possible bases at each SNP site (base quality  $\geq 20$ ). The allelic ratio was then  
973 calculated as the ratio of the variant base (as defined by dbSNP) count versus the sum  
974 of the reference and variant base count. Only sites with a coverage of at least 15x were  
975 plotted.  
976

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

978 The sequence signature at each new telomeric and chromosomal arm fusion site was  
979 analyzed using the consensus soft-clipped sequences identified by TelTools, and the  
980 sequence extracted from the reference genome at each site. The sequence signature at  
981 each new telomere and chromosomal arm fusion was then analyzed by (i) identifying  
982 the frequency of each telomeric 6-mer in each soft-clipped sequence, and by (ii)  
983 assessing the sequence motif of the telomeric region and genomic region.  
984

## 985 Spectral Karyotyping

986 DNA Spectral Karyotyping Hybridization was performed according to the protocol of  
987 commercial spectral karyotyping paint probes from Applied Spectral Imaging (5315  
988 Avenida Encinas, Suite 150, Carlsbad, CA92008). Briefly, the slides were dropped in  
989 Thermotron and aged for 3-5 days in a 37°C oven. The slides were then checked under  
990 the microscope before hybridization. A series of four steps were then performed on  
991 these slides to generate the spectro karyotype of the cell lines: (1) Trypsin Treatment:  
992 The slides were washed briefly in Earl's medium, and then treated with Trypsin/EDTA  
993 solution. Washing was then performed in water and then dehydrated in ethanol series of  
994 70%, 80% and 100% for 2 minutes each followed by air-dying of the slides. (2) Chromosome Denaturation: The slides were treated in 2XSSC buffer for 2 minutes and  
995 then dehydrated in Ethanol series for 2 minutes each. Denaturation of the slides was  
996 then performed at 72°C in denaturation solution for 1.5 minutes. This is followed  
997 immediately by placing the slides in cold ethanol series to dehydrate the slides, and  
998 then air drying. (3) Probe Denaturation and hybridization: The probe was denatured by  
999 incubating the probe at 80°C in a water bath for 7 minutes. The denatured Spectral  
1000 Karyotyping reagent was then applied to the denaturized chromosome preparation and  
1001 incubated at 37°C for 5-6 days. (4) Detection, imaging and karyotyping: The slides were  
1002 washed in 0.4XSSC at 72°C for 2 minutes and then dipped in 4XSSC/Tween-20 for 1  
1003 minutes. Cy5 staining reagent was then applied and incubated at 37°C for 40 minutes.  
1004

1005 The slides were then washed 3 times in washing solution, and then mounted with anti-  
1006 fade DAPI. After which, the slides are ready for spectral imaging. Rearrangements were  
1007 defined with nomenclature rules from international Committee in Standard Genetic  
1008 Nomenclature for Human.

1009

1010

1011 **Data and code availability**

1012 TelFuse and TelSize developed for this study are available at  
1013 <https://github.com/ktan8/teltools/>. Long-read genome sequencing data generated for  
1014 this study would be deposited in the SRA database prior to the publication of the  
1015 manuscript.

1016

1017

1018 **References**

1019 1. Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D.,  
1020 Weerasinghe, A., Colaprico, A., Wendl, M.C., Kim, J., Reardon, B., et al. (2018).  
1021 Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*.  
1022 10.1016/j.cell.2018.02.060.

1023 2. Huang, K. lin, Mashl, R.J., Wu, Y., Ritter, D.I., Wang, J., Oh, C., Paczkowska, M.,  
1024 Reynolds, S., Wyczalkowski, M.A., Oak, N., et al. (2018). Pathogenic Germline  
1025 Variants in 10,389 Adult Cancers. *Cell* 173, 355-370.e14.  
1026 10.1016/J.CELL.2018.03.039.

1027 3. Campbell, P.J., Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L., Stein, L.D.,  
1028 Perry, M.D., Nahal-Bose, H.K., Ouellette, B.F.F., Li, C.H., et al. (2020). Pan-  
1029 cancer analysis of whole genomes. *Nat.* 2020 5787793 578, 82–93.  
1030 10.1038/s41586-020-1969-6.

1031 4. Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Brujin, E., Shale, C.,  
1032 Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-  
1033 cancer whole-genome analyses of metastatic solid tumours. *Nature* 575.  
1034 10.1038/s41586-019-1689-y.

1035 5. Degasperi, A., Zou, X., Dias Amarante, T., Martinez-Martinez, A., Koh, G.C.C.,  
1036 Dias, J.M.L., Heskin, L., Chmelova, L., Rinaldi, G., and Wang, V.Y.W. (2022).  
1037 Substitution mutational signatures in whole-genome–sequenced cancers in the  
1038 UK population. *Science* (80- ). 376, abl9283.

1039 6. Imielinski, M., Berger, A.H., Hammerman, P.S., Hernandez, B., Pugh, T.J., Hodis,  
1040 E., Cho, J., Suh, J., Capelletti, M., Sivachenko, A., et al. (2012). Mapping the  
1041 hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* 150,  
1042 1107–1120. 10.1016/j.cell.2012.08.029.

1043 7. Ghandi, M., Huang, F.W., Jané-Valbuena, J., Kryukov, G. V., Lo, C.C., McDonald,  
1044 E.R., Barretina, J., Gelfand, E.T., Bielski, C.M., Li, H., et al. (2019). Next-  
1045 generation characterization of the Cancer Cell Line Encyclopedia. *Nature*.  
1046 10.1038/s41586-019-1186-3.

1047 8. Zheng, G.X.Y., Lau, B.T., Schnall-Levin, M., Jarosz, M., Bell, J.M., Hindson, C.M.,  
1048 Kyriazopoulou-Panagiotopoulou, S., Masquelier, D.A., Merrill, L., Terry, J.M., et  
1049 al. (2016). Haplotyping germline and cancer genomes with high-throughput  
1050 linked-read sequencing. *Nat. Biotechnol.* 10.1038/nbt.3432.

1051 9. Xia, L.C., Bell, J.M., Wood-Bouwens, C., Chen, J.J., Zhang, N.R., and Ji, H.P.  
1052 (2018). Identification of large rearrangements in cancer genomes with barcode  
1053 linked reads. *Nucleic Acids Res.* 46. 10.1093/NAR/GKX1193.

1054 10. Greer, S.U., Nadauld, L.D., Lau, B.T., Chen, J., Wood-Bouwens, C., Ford, J.M.,  
1055 Kuo, C.J., and Ji, H.P. (2017). Linked read sequencing resolves complex genomic  
1056 rearrangements in gastric cancer metastases. *Genome Med.* 9, 1–17.  
1057 10.1186/S13073-017-0447-8/FIGURES/5.

1058 11. Viswanathan, S.R., Ha, G., Hoff, A.M., Wala, J.A., Carrot-Zhang, J., Whelan,  
1059 C.W., Haradhvala, N.J., Freeman, S.S., Reed, S.C., Rhoades, J., et al. (2018).  
1060 Structural Alterations Driving Castration-Resistant Prostate Cancer Revealed by  
1061 Linked-Read Genome Sequencing. *Cell*. 10.1016/j.cell.2018.05.036.

1062 12. Tan, K.-T., Kim, H., Carrot-Zhang, J., Zhang, Y., Kim, W.J., Kugener, G., Wala,  
1063 J.A., Howard, T.P., Chi, Y.-Y., Beroukhim, R., et al. (2021). Haplotype-resolved

1064 germline and somatic alterations in renal medullary carcinomas. *Genome Med.*  
1065 Vol. 13.

1066 13. Schmid, C.W., and Deininger, P.L. (1975). Sequence organization of the human  
1067 genome. *Cell* 6, 345–358. 10.1016/0092-8674(75)90184-1.

1068 14. Batzer, M.A., and Deininger, P.L. (2002). Alu repeats and human genomic  
1069 diversity. *Nat. Rev. Genet.* 2002 35 3, 370–379. 10.1038/nrg798.

1070 15. Treangen, T.J., and Salzberg, S.L. (2011). Repetitive DNA and next-generation  
1071 sequencing: computational challenges and solutions. *Nat. Rev. Genet.* 2011 131  
1072 13, 36–46. 10.1038/nrg3117.

1073 16. Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution.  
1074 *Nat. Rev. Genet.* 2004 56 5, 435–445. 10.1038/nrg1348.

1075 17. Eichler, E.E. (2001). Recent duplication, domain accretion and the dynamic  
1076 mutation of the human genome. *Trends Genet.* 17, 661–669. 10.1016/S0168-  
1077 9525(01)02492-1.

1078 18. Bailey, J.A., Yavor, A.M., Massa, H.F., Trask, B.J., and Eichler, E.E. (2001).  
1079 Segmental Duplications: Organization and Impact Within the Current Human  
1080 Genome Project Assembly. *Genome Res.* 11, 1005. 10.1101/GR.187101.

1081 19. Sharp, A.J., Locke, D.P., McGrath, S.D., Cheng, Z., Bailey, J.A., Vallente, R.U.,  
1082 Pertz, L.M., Clark, R.A., Schwartz, S., Segraves, R., et al. (2005). Segmental  
1083 Duplications and Copy-Number Variation in the Human Genome. *Am. J. Hum.  
1084 Genet.* 77, 78. 10.1086/431652.

1085 20. Vollger, M.R., Guitart, X., Dishuck, P.C., Mercuri, L., Harvey, W.T., Gershman, A.,  
1086 Diekhans, M., Sulovari, A., Munson, K.M., Lewis, A.P., et al. (2022). Segmental  
1087 duplications and their variation in a complete human genome. *Science* (80-. ).  
1088 376.  
1089 10.1126/SCIENCE.ABJ6965/SUPPL\_FILE/SCIENCE.ABJ6965\_MDAR\_REPROD  
1090 UCIBILITY\_CHECKLIST.PDF.

1091 21. Factor-Litvak, P., Susser, E., Kezios, K., McKeague, I., Kark, J.D., Hoffman, M.,  
1092 Kimura, M., Wapner, R., and Aviv, A. (2016). Leukocyte telomere length in  
1093 newborns: Implications for the role of telomeres in human disease. *Pediatrics* 137.  
1094 10.1542/PEDS.2015-3927/-DCSUPPLEMENTAL.

1095 22. Canela, A., Vera, E., Klatt, P., and Blasco, M.A. (2007). High-throughput telomere  
1096 length quantification by FISH and its application to human population studies.  
1097 *Proc. Natl. Acad. Sci. U. S. A.* 104, 5300–5305.  
1098 10.1073/PNAS.0609367104/SUPPL\_FILE/09367FIG5.JPG.

1099 23. Sieverling, L., Hong, C., Koser, S.D., Ginsbach, P., Kleinheinz, K., Hutter, B.,  
1100 Braun, D.M., Cortés-Ciriano, I., Xi, R., Kabbe, R., et al. (2020). Genomic footprints  
1101 of activated telomere maintenance mechanisms in cancer. *Nat. Commun.*  
1102 10.1038/s41467-019-13824-9.

1103 24. Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: The next  
1104 generation. *Cell* 144, 646–674.  
1105 10.1016/J.CELL.2011.02.013/ATTACHMENT/3F528E16-8B3C-4D8D-8DE5-  
1106 43E0C98D8475/MMC1.PDF.

1107 25. Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. *Cancer Discov.* 12,  
1108 31–46. 10.1158/2159-8290.CD-21-1059.

1109 26. Li, Y., and Tergaonkar, V. (2014). Noncanonical functions of telomerase:

1110        implications in telomerase-targeted cancer therapies. *Cancer Res.* **74**, 1639–  
1111        1644. 10.1158/0008-5472.CAN-13-3568.

1112        27. Kim, N.W., Piatyszek, M.A., Prowse, K.R., Harley, C.B., West, M.D., Ho, P.L.C.,  
1113        Coville, G.M., Wright, W.E., Weinrich, S.L., and Shay, J.W. (1994). Specific  
1114        Association of Human Telomerase Activity with Immortal Cells and Cancer.  
1115        *Science* (80- ). **266**, 2011–2015. 10.1126/science.7605428.

1116        28. Meyerson, M., Counter, C.M., Eaton, E.N., Ellisen, L.W., Steiner, P., Caddle,  
1117        S.D., Ziaugra, L., Beijersbergen, R.L., Davidoff, M.J., Liu, Q., et al. (1997).  
1118        hEST2, the Putative Human Telomerase Catalytic Subunit Gene, Is Up-Regulated  
1119        in Tumor Cells and during Immortalization. *Cell* **90**, 785–795. 10.1016/S0092-  
1120        8674(00)80538-3.

1121        29. Kolquist, K.A., Ellisen, L.W., Counter, C.M., Meyerson, M., Tan, L.K., Weinberg,  
1122        R.A., Haber, D.A., and Gerald, W.L. (1998). Expression of TERT in early  
1123        premalignant lesions and a subset of cells in normal tissues. *Nat. Genet.* **19**, 182–  
1124        186. 10.1038/554.

1125        30. Li, Y., and Tergaonkar, V. (2016). Telomerase reactivation in cancers:  
1126        Mechanisms that govern transcriptional activation of the wild-type vs. mutant  
1127        *TERT* promoters. *Transcription* **7**, 44–49. 10.1080/21541264.2016.1160173.

1128        31. Yuan, X., Larsson, C., and Xu, D. (2019). Mechanisms underlying the activation of  
1129        TERT transcription and telomerase activity in human cancer: old actors and new  
1130        players. *Oncogene* **38**, 6172–6183. 10.1038/s41388-019-0872-9.

1131        32. Huang, F.W., Hodis, E., Xu, M.J., Kryukov, G. V., Chin, L., and Garraway, L.A.  
1132        (2013). Highly recurrent TERT promoter mutations in human melanoma. *Science*  
1133        **339**, 957–959. 10.1126/SCIENCE.1229259.

1134        33. Horn, S., Figl, A., Rachakonda, P.S., Fischer, C., Sucker, A., Gast, A., Kadel, S.,  
1135        Moll, I., Nagore, E., Hemminki, K., et al. (2013). TERT promoter mutations in  
1136        familial and sporadic melanoma. *Science* (80- ). **339**, 959–961.  
1137        10.1126/SCIENCE.1230062/SUPPL\_FILE/HORN.SM.PDF.

1138        34. Killela, P.J., Reitman, Z.J., Jiao, Y., Bettegowda, C., Agrawal, N., and Diaz, L.A.  
1139        (2013). TERT promoter mutations occur frequently in gliomas and a subset of  
1140        tumors derived from cells with low rates of self-renewal. **110**, 6021–6026.  
1141        10.1073/pnas.1303607110.

1142        35. Barthel, F.P., Wei, W., Tang, M., Martinez-Ledesma, E., Hu, X., Amin, S.B.,  
1143        Akdemir, K.C., Seth, S., Song, X., Wang, Q., et al. (2017). Systematic analysis of  
1144        telomere length and somatic alterations in 31 cancer types. *Nat. Genet.*  
1145        10.1038/ng.3781.

1146        36. Cao, Y., Bryan, T.M., and Reddel, R.R. (2008). Increased copy number of the  
1147        TERT and TERC telomerase subunit genes in cancer cells. *Cancer Sci.* **99**,  
1148        1092–1099. 10.1111/J.1349-7006.2008.00815.X.

1149        37. Jiao, Y., Shi, C., Edil, B.H., De Wilde, R.F., Klimstra, D.S., Maitra, A., Schulick,  
1150        R.D., Tang, L.H., Wolfgang, C.L., Choti, M.A., et al. (2011). DAXX/ATRX, MEN1,  
1151        and mTOR pathway genes are frequently altered in pancreatic neuroendocrine  
1152        tumors. *Science* (80- ). **331**, 1199–1203.  
1153        10.1126/SCIENCE.1200609/SUPPL\_FILE/JIAO-SOM.PDF.

1154        38. Heaphy, C.M., De Wilde, R.F., Jiao, Y., Klein, A.P., Edil, B.H., Shi, C.,  
1155        Bettegowda, C., Rodriguez, F.J., Eberhart, C.G., Hebbar, S., et al. (2011). Altered

1156 telomeres in tumors with ATRX and DAXX mutations. *Science* (80- ). 333, 425.  
 1157 10.1126/SCIENCE.1207313/SUPPL\_FILE/HEAPHY.SOM.REV1.PDF.

1158 39. Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion,  
 1159 G.T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N.D., et al. (2019).  
 1160 Highly-accurate long-read sequencing improves variant detection and assembly  
 1161 of a human genome. *bioRxiv*. 10.1101/519025.

1162 40. Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R.,  
 1163 Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and  
 1164 assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* 36, 338–  
 1165 345. 10.1038/nbt.4060.

1166 41. Kinzig, C.G., Zakusilo, G., Takai, K.K., and de Lange, T. (2022). Neotelomere  
 1167 formation by human telomerase. *bioRxiv*, 2022.10.31.514589.  
 1168 10.1101/2022.10.31.514589.

1169 42. Ijdo, J.W., Baldini, A., Ward, D.C., Reeders, S.T., and Wells, R.A. (1991). Origin  
 1170 of human chromosome 2: an ancestral telomere-telomere fusion. *Proc. Natl.  
 1171 Acad. Sci. U. S. A.* 88, 9051. 10.1073/PNAS.88.20.9051.

1172 43. Fan, Y., Linardopoulou, E., Friedman, C., Williams, E., and Trask, B.J. (2002).  
 1173 Genomic Structure and Evolution of the Ancestral Chromosome Fusion Site in  
 1174 2q13–2q14.1 and Paralogous Regions on Other Human Chromosomes. *Genome  
 1175 Res.* 12, 1651. 10.1101/GR.337602.

1176 44. Stong, N., Deng, Z., Gupta, R., Hu, S., Paul, S., Weiner, A.K., Eichler, E.E.,  
 1177 Graves, T., Fronick, C.C., Courtney, L., et al. (2014). Subtelomeric CTCF and  
 1178 cohesin binding site organization using improved subtelomere assemblies and a  
 1179 novel annotation pipeline. *Genome Res.* 24, 1039–1050. 10.1101/gr.166983.113.

1180 45. Riethman, H., Ambrosini, A., Castaneda, C., Finklestein, J., Hu, X.L., Mudunuri,  
 1181 U., Paul, S., and Wei, J. (2004). Mapping and initial analysis of human  
 1182 subtelomeric sequence assemblies. *Genome Res.* 14, 18–28.  
 1183 10.1101/GR.1245004.

1184 46. Ambrosini, A., Paul, S., Hu, S., and Riethman, H. (2007). Human subtelomeric  
 1185 dupilon structure and organization. *Genome Biol.* 8, 1–13. 10.1186/GB-2007-8-  
 1186 7-R151/TABLES/2.

1187 47. Hock, H., and Shimamura, A. (2017). ETV6 in Hematopoiesis and Leukemia  
 1188 Predisposition. *Semin. Hematol.* 54, 98.  
 1189 10.1053/J.SEMINHEMATOL.2017.04.005.

1190 48. Manguso, R.T., Pope, H.W., Zimmer, M.D., Brown, F.D., Yates, K.B., Miller, B.C.,  
 1191 Collins, N.B., Bi, K., La Fleur, M.W., Juneja, V.R., et al. (2017). In vivo CRISPR  
 1192 screening identifies Ptpn2 as a cancer immunotherapy target. *Nature* 547.  
 1193 10.1038/nature23270.

1194 49. TR, G., GF, B., M, L., and DG, G. (1994). Fusion of PDGF receptor beta to a  
 1195 novel ets-like gene, tel, in chronic myelomonocytic leukemia with t(5;12)  
 1196 chromosomal translocation. *Cell* 77. 10.1016/0092-8674(94)90322-0.

1197 50. Knezevich, S.R., McFadden, D.E., Tao, W., Lim, J.F., and Sorensen, P.H.B.  
 1198 (1998). A novel ETV6-NTRK3 gene fusion in congenital fibrosarcoma. *Nat. Genet.*  
 1199 18. 10.1038/ng0298-184.

1200 51. Liao, H., Winkfein, R.J., Mack, G., Rattner, J.B., and Yen, T.J. (1995). CENP-F is  
 1201 a protein of the nuclear matrix that assembles onto kinetochores at late G2 and is

1202 rapidly degraded after mitosis. *J. Cell Biol.* 130, 507. 10.1083/JCB.130.3.507.

1203 52. Zhu, X., Mancini, M.A., Chang, K.H., Liu, C.Y., Chen, C.F., Shan, B., Jones, D.,  
1204 Yang-Feng, T.L., and Lee, W.H. (1995). Characterization of a novel 350-kilodalton  
1205 nuclear phosphoprotein that is specifically involved in mitotic-phase progression.  
1206 *Mol. Cell. Biol.* 15, 5017–5029. 10.1128/MCB.15.9.5017.

1207 53. Zhu, X., Chang, K.H., He, D., Mancini, M.A., Brinkley, W.R., and Lee, W.H.  
1208 (1995). The C Terminus of Mitosin Is Essential for Its Nuclear Localization,  
1209 Centromere/Kinetochore Targeting, and Dimerization. *J. Biol. Chem.* 270, 19545–  
1210 19550. 10.1074/JBC.270.33.19545.

1211 54. Grigorova, M., Staines, J.M., Ozdag, H., Caldas, C., and Edwards, P.A.W. (2004).  
1212 Possible causes of chromosome instability: Comparison of chromosomal  
1213 abnormalities in cancer cell lines with mutations in BRCA1, BRCA2, CHK2 and  
1214 BUB1. In *Cytogenetic and Genome Research* 10.1159/000077512.

1215 55. Davidson, J.M., Gorringe, K.L., Chin, S.F., Orsetti, B., Besret, C., Courtay-Cahen,  
1216 C., Roberts, I., Theillet, C., Caldas, C., and Edwards, P.A.W. (2000). Molecular  
1217 cytogenetic analysis of breast cancer cell lines. *Br. J. Cancer.* 10.1054/bjoc.2000.1458.

1218 56. Abdel-Rahman, W.M., Katsura, K., Rens, W., Gorman, P.A., Sheer, D., Bicknell,  
1219 D., Bodmer, W.F., Arends, M.J., Wyllie, A.H., and Edwards, P.A.W. (2001).  
1220 Spectral karyotyping suggests additional subsets of colorectal cancers  
1221 characterized by pattern of chromosome rearrangement. *Proc. Natl. Acad. Sci. U.*  
1222 *S. A.* 10.1073/pnas.041603298.

1223 57. Grigorova, M., Lyman, R.C., Caldas, C., and Edwards, P.A.W. (2005).  
1224 Chromosome abnormalities in 10 lung cancer cell lines of the NCI-H series  
1225 analyzed with spectral karyotyping. *Cancer Genet. Cytogenet.* 10.1016/j.cancergenryo.2005.03.007.

1226 58. Sirivatanauskorn, V., Sirivatanauskorn, Y., Gorman, P.A., Davidson, J.M., Sheer,  
1227 D., Moore, P.S., Scarpa, A., Edwards, P.A.W., and Lemoine, N.R. (2001). Non-  
1228 random chromosomal rearrangements in pancreatic cancer cell lines identified by  
1229 spectral karyotyping. *Int. J. Cancer.* 10.1002/1097-  
1230 0215(200002)9999:9999<::AID-IJC1049>3.3.CO;2-3.

1231 59. Edwards, P. SKY Karyotypes and FISH analysis of Epithelial Cancer Cell Lines.

1232 60. Livingstone, K., and Rieseberg, L. (2004). Chromosomal evolution and speciation:  
1233 A recombination-based approach. *New Phytol.* 10.1046/j.1469-  
1234 8137.2003.00942.x.

1235 61. Fischer, G., James, S.A., Roberts, I.N., Oliver, S.G., and Louis, E.J. (2000).  
1236 Chromosomal evolution in *Saccharomyces*. *Nature.* 10.1038/35013058.

1237 62. Dutrillaux, B. (1979). Chromosomal evolution in Primates: Tentative phylogeny  
1238 from *Microcebus murinus* (Prosimian) to man. *Hum. Genet.*  
1239 10.1007/BF00272830.

1240 63. Mazzoleni, S., Schillaci, O., Sineo, L., and Dumas, F. (2017). Distribution of  
1241 Interstitial Telomeric Sequences in Primates and the Pygmy Tree Shrew  
1242 (Scandentia). *Cytogenet. Genome Res.* 10.1159/000467634.

1243 64. Lin, K.W., and Yan, J. (2008). Endings in the middle: Current knowledge of  
1244 interstitial telomeric sequences. *Mutat. Res. - Rev. Mutat. Res.*  
1245 10.1016/j.mrrev.2007.08.006.

1248 65. Meyne, J., Baker, R.J., Hobart, H.H., Hsu, T.C., Ryder, O.A., Ward, O.G., Wiley,  
1249 J.E., Wurster-Hill, D.H., Yates, T.L., and Moyzis, R.K. (1990). Distribution of non-  
1250 telomeric sites of the (TTAGGG)<sub>n</sub> telomeric sequence in vertebrate  
1251 chromosomes. *Chromosoma*. 10.1007/BF01737283.

1252 66. Ocalewicz, K., Furgala-Selezniow, G., Szmyt, M., Lisboa, R., Kucinski, M., Lejk,  
1253 A.M., and Jankun, M. (2013). Pericentromeric location of the telomeric DNA  
1254 sequences on the European grayling chromosomes. *Genetica*. 10.1007/s10709-  
1255 013-9740-7.

1256 67. Faravelli, M., Moralli, D., Bertoni, L., Attolini, C., Chernova, O., Raimondi, E., and  
1257 Giulotto, E. (1998). Two extended arrays of a satellite DNA sequence at the  
1258 centromere and at the short-arm telomere of Chinese hamster chromosome 5.  
1259 *Cytogenet. Cell Genet.* 10.1159/000015171.

1260 68. Sholes, S.L., Karimian, K., Gershman, A., Kelly, T.J., Timp, W., and Greider, C.W.  
1261 (2022). Chromosome-specific telomere lengths and the minimal functional  
1262 telomere revealed by nanopore sequencing. *Genome Res.* 32, 616–628.  
1263 10.1101/GR.275868.121/-/DC1.

1264 69. Grigorev, K., Foox, J., Bezdan, D., Butler, D., Luxton, J.J., Reed, J., McKenna,  
1265 M.J., Taylor, L., George, K.A., Meydan, C., et al. (2021). Haplotype diversity and  
1266 sequence heterogeneity of human telomeres. *Genome Res.* 31, 1269–1279.  
1267 10.1101/gr.274639.120.

1268 70. Carrot-Zhang, J., Yao, X., Devarakonda, S., Deshpande, A., Damrauer, J.S.,  
1269 Silva, T.C., Wong, C.K., Choi, H.Y., Felau, I., Robertson, A.G., et al. (2021).  
1270 Whole-genome characterization of lung adenocarcinomas lacking the  
1271 RTK/RAS/RAF pathway. *Cell Rep.* 10.1016/j.celrep.2021.108707.

1272 71. Campbell, J.D., Alexandrov, A., Kim, J., Wala, J., Berger, A.H., Pedamallu, C.S.,  
1273 Shukla, S.A., Guo, G., Brooks, A.N., Murray, B.A., et al. (2016). Distinct patterns  
1274 of somatic genome alterations in lung adenocarcinomas and squamous cell  
1275 carcinomas. *Nat. Genet.* 10.1038/ng.3564.

1276 72. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigelski, E.M., and  
1277 Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. *Nucleic Acids  
1278 Res.* 29, 308–311.

1279 73. Nurk, S., Koren, S., Rhee, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A.,  
1280 Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The  
1281 complete sequence of a human genome. *Science* (80- ). 376.  
1282 10.1126/science.abj6987.

1283 74. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences.  
1284 *Bioinformatics* 34, 3094–3100. 10.1093/bioinformatics/bty191.

1285 75. Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs  
1286 with BWA-MEM. *arXiv Prepr. arXiv*, 3997.

1287 76. Li, H. (2011). A statistical framework for SNP calling, mutation discovery,  
1288 association mapping and population genetical parameter estimation from  
1289 sequencing data. *Bioinformatics* 27, 2987–2993. 10.1093/bioinformatics/btr509.

1290 77. Team, R.C. (2021). R: A Language and Environment for Statistical Computing. R  
1291 Found. Stat. Comput.

1292 78. Van Rossum, G., and Drake, F.L. (2009). Python 3 Reference Manual;  
1293 CreateSpace. Scotts Val. CA.

1294 79. Larry Wall (1994). The PERL Programming Language. Dr. Dobb's J. Softw. Tools  
1295 19.

1296 80. Thorvaldsdóttir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative  
1297 Genomics Viewer (IGV): High-performance genomics data visualization and  
1298 exploration. *Brief. Bioinform.* 14. 10.1093/bib/bbs017.

1299 81. Tan, K.T., Slevin, M.K., Meyerson, M., and Li, H. (2022). Identifying and  
1300 correcting repeat-calling errors in nanopore sequencing of telomeres. *Genome*  
1301 *Biol.* 23, 1–16. 10.1186/S13059-022-02751-6/FIGURES/2.

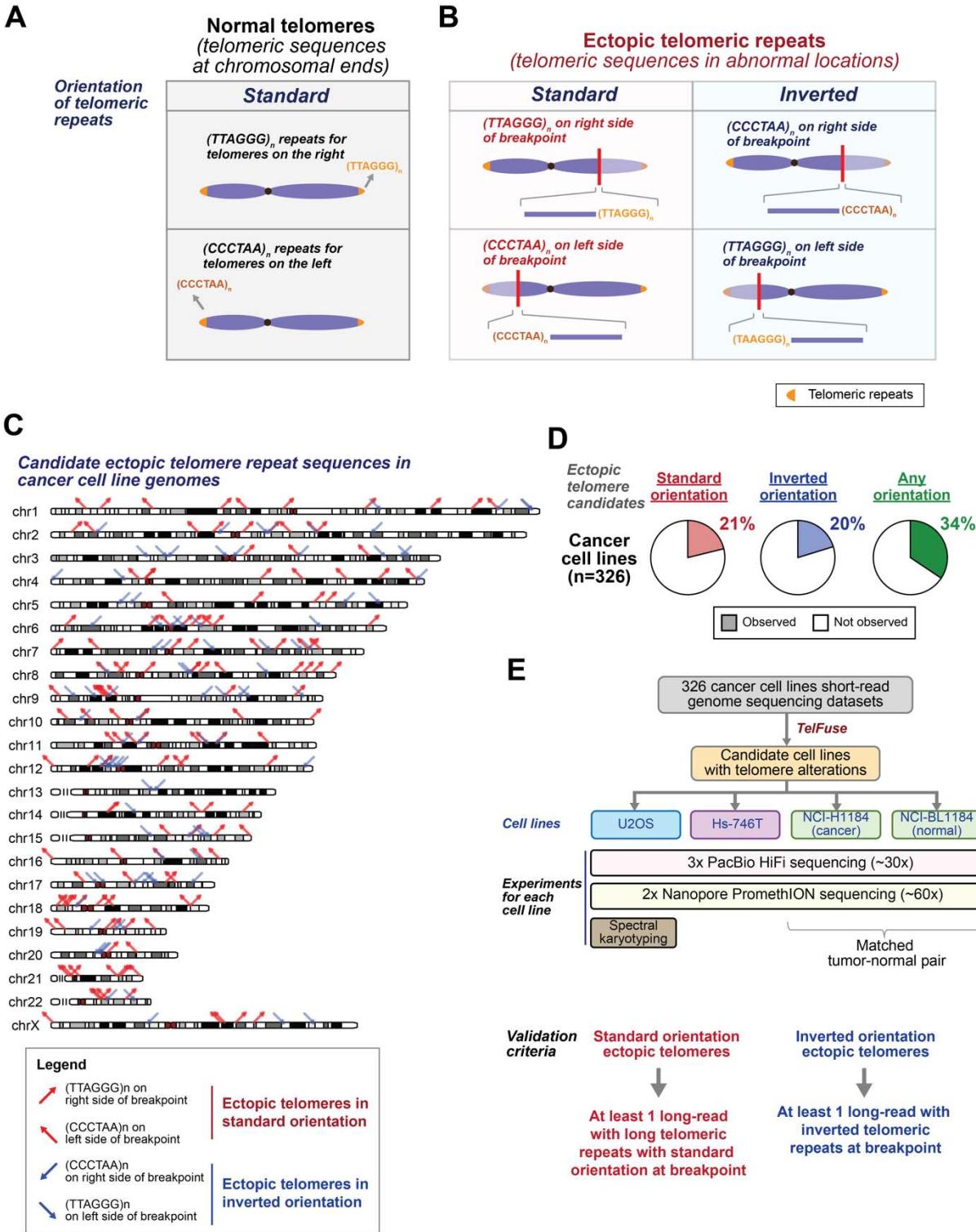
1302

1303

1304

1305

# Figure 1

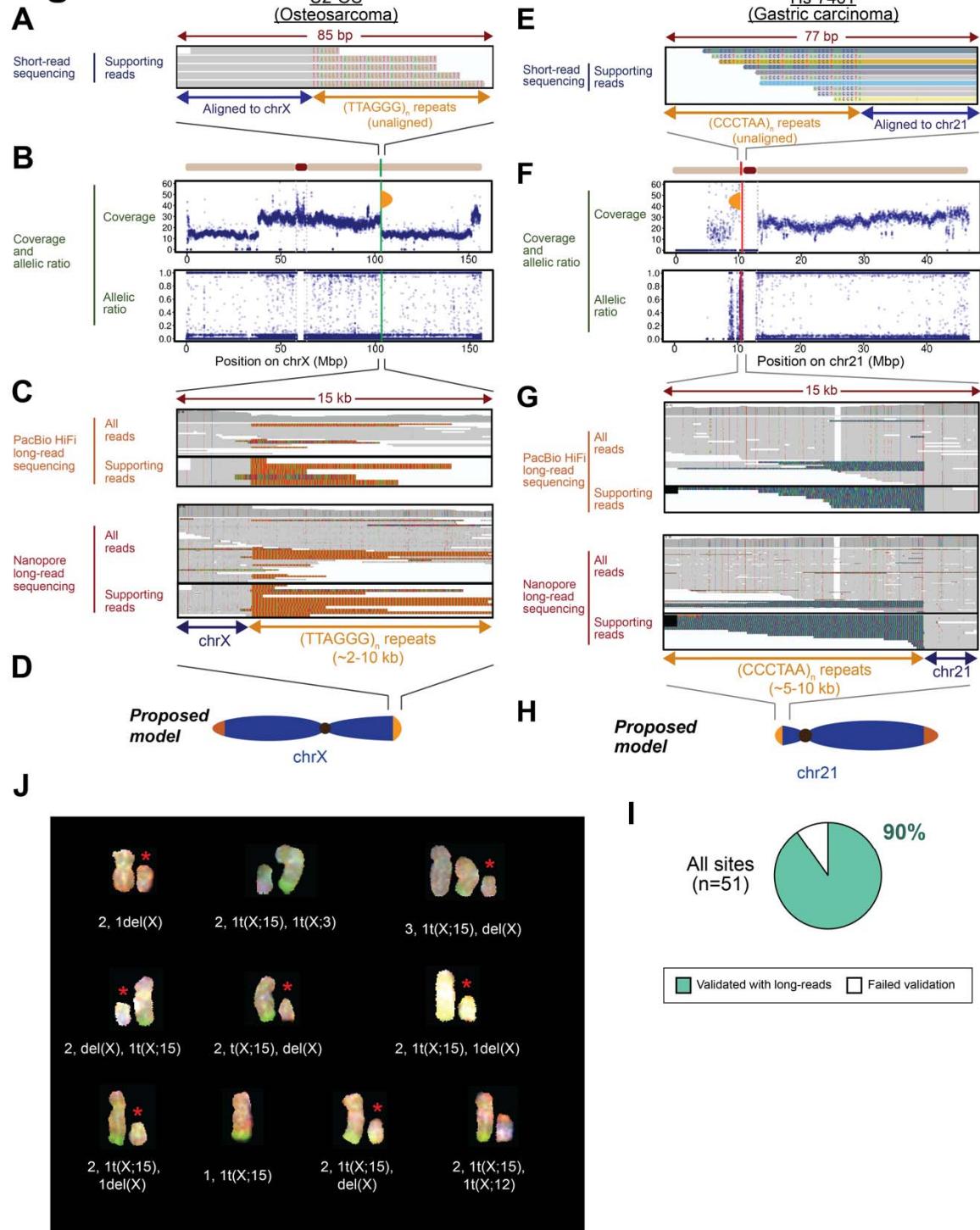


1308 **Figure 1 Classes of ectopic telomeric repeats found in cancer cell genomes. (A)**  
1309 Schematic of sequence and positions of normal telomeres at chromosomal termini. **(B)**  
1310 Schematic of ectopic telomeric repeats found at abnormal locations away from  
1311 chromosomal termini. Standard orientation: (TTAGGG)<sub>n</sub> on the right side of a breakpoint  
1312 and (CCCTAA)<sub>n</sub> on the left side of the breakpoint in the 5' to 3' direction (same as  
1313 normal telomere in Fig. 1A). Inverted orientation: (CCCTAA)<sub>n</sub> on the right side of a  
1314 breakpoint and (TTAGGG)<sub>n</sub> on the left side of the breakpoint in the 5' to 3' direction.  
1315 Note that faded chromosomal segment is not part of derivative chromosome. **(C)**  
1316 Genome-wide localization of ectopic telomeric repeats in cancer cell line genomes  
1317 (n=326) identified using short-read genome sequencing. Red: ectopic telomeric  
1318 sequences in the standard orientation. Blue: ectopic telomeric sequences in the inverted  
1319 orientation. Position of telomeric repeats relative to the breakpoint is indicated by arrows  
1320 oriented in different directions. **(D)** Percentage of cancer cell lines in the CCLE with  
1321 ectopic telomeric sequences in either orientation. Total sample number as indicated. **(E)**  
1322 Flow-chart of long-read genome sequencing and cytogenetic analyses in cancer cell  
1323 lines, with the indicated validation criteria.

1324

1325

## Figure 2

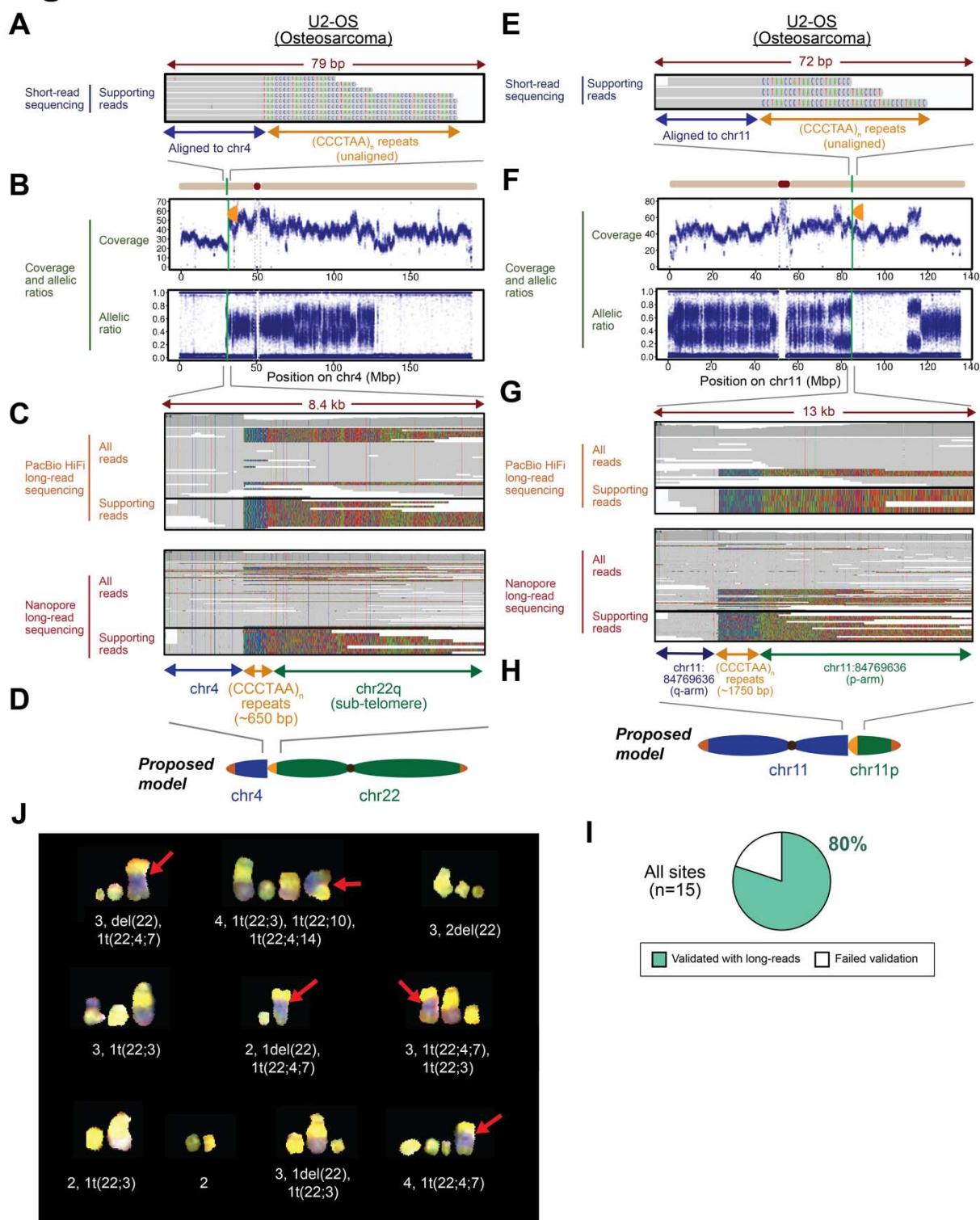


1328 **Figure 2 Neotelomeres in cancer genomes revealed by long-read genome**  
1329 **sequencing. (A-H)** Genomic analysis of telomere repeat alterations in the standard  
1330 orientation that were detected **(A-D)** in the U2-OS osteosarcoma cell line at  
1331 chrX:103,320,553, and **(E-H)** in the Hs-746T cell line at chr21:10,547,397. **(A)** IGV  
1332 screenshots of short-read genome sequencing data. Ectopic telomeric repeats  
1333 (TTAGGG)<sub>n</sub> are shown in color. **(B)** Sequencing coverage and allelic ratios of  
1334 chromosome X. Orange semi-oval: site of the neotelomeric event. **(C)** IGV screenshots  
1335 depicting long telomeric repeat sequences (TTAGGG)<sub>n</sub> with PacBio HiFi and Nanopore  
1336 long-read sequencing at the site shown in **(A)**. **(D)** Schematic of neotelomere location  
1337 on chromosome Xq. **(E)** IGV screenshots of short-read genome sequencing data.  
1338 Ectopic telomeric repeats (CCCTAA)<sub>n</sub> are shown in color. **(F)** Sequencing coverage and  
1339 allelic ratios of chromosome 21. Orange semi-oval: site of the neotelomeric event. **(G)**  
1340 IGV screenshots depicting long telomeric repeat sequences (CCCTAA)<sub>n</sub> with PacBio  
1341 HiFi and Nanopore long-read sequencing at the site shown in **(E)**. **(H)** Schematic of  
1342 neotelomere location on chromosome 21p. **(I)** Percentage of ectopic telomeric repeat  
1343 sites in the standard orientation, found by short-read genome sequencing using  
1344 TelFuse, that were validated by long-read genome sequencing. **(J)** Spectral karyogram  
1345 of chrX in ten U2-OS single cells assessed by spectral karyotyping with corresponding  
1346 karyotype labels. First label: total # of X chromosomes and their derivatives observed in  
1347 given cell. Second label: karyotypes of the aberrant X chromosomes or derivatives.  
1348 Asterisk (\*): truncated X chromosome. See also Figure S4.

1349

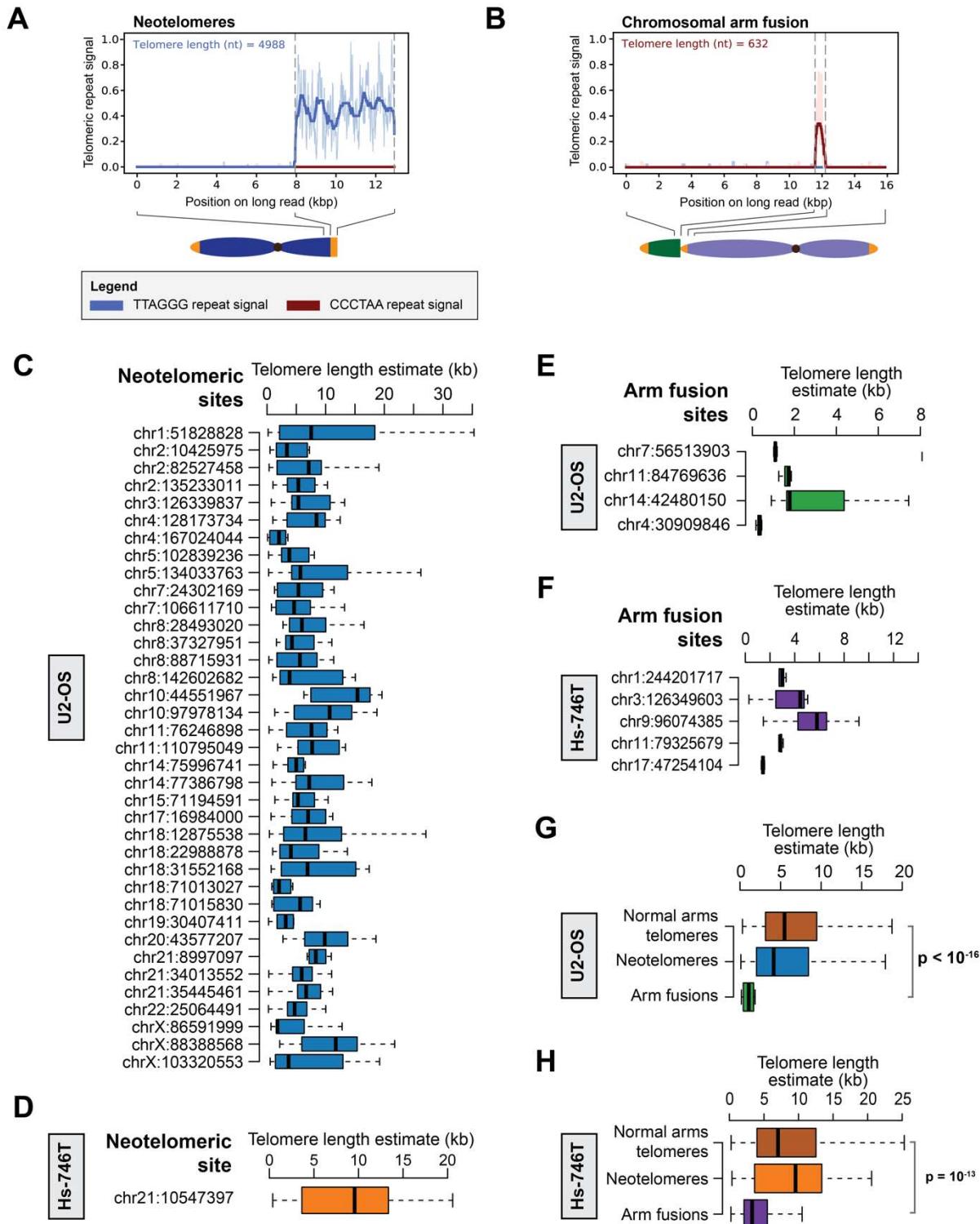
1350

## Figure 3



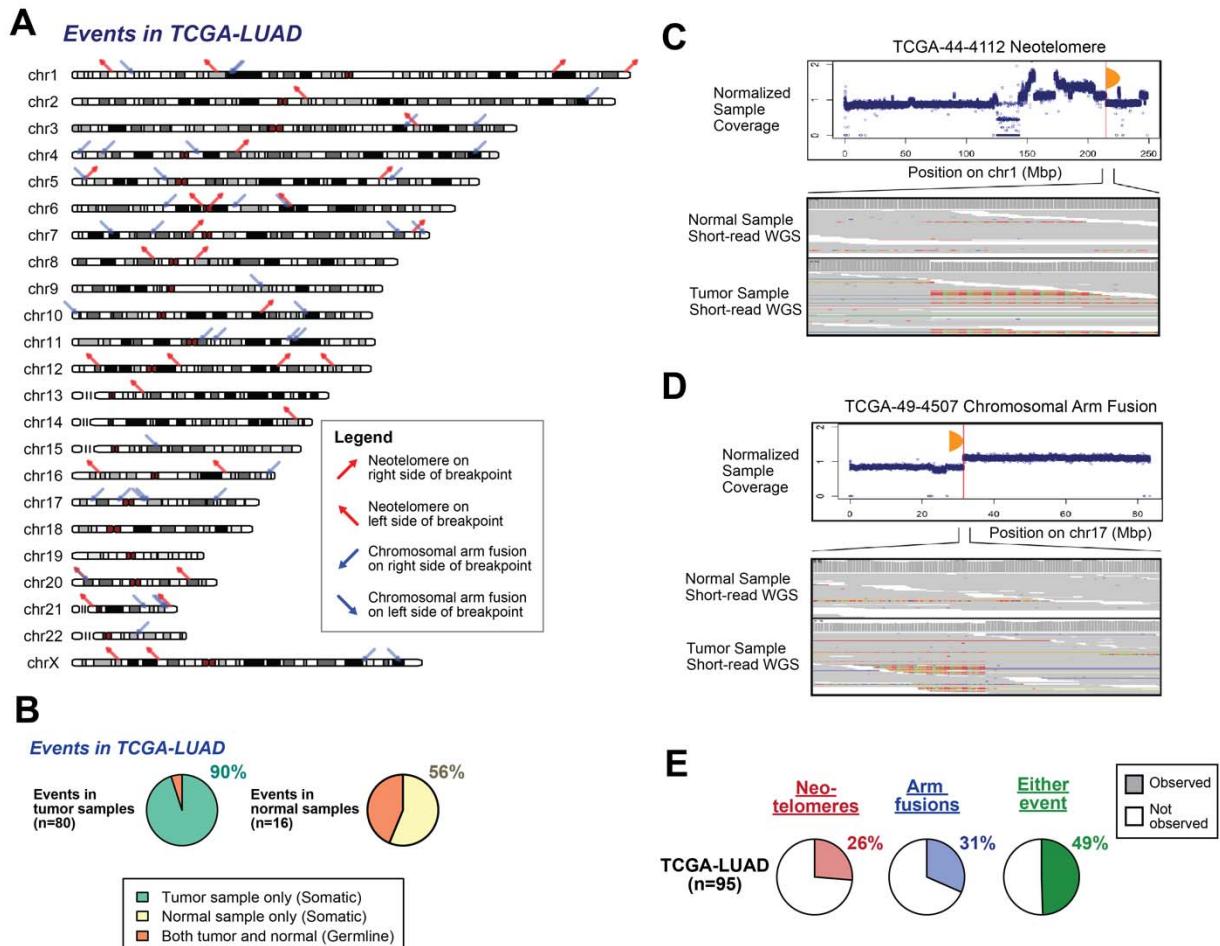
1353 **Figure 3 Chromosomal arm fusions in cancer genomes revealed by long-read**  
1354 **genome sequencing. (A-H)** Genomic analysis of telomere repeat alterations in the  
1355 inverted orientation that were detected in the U2-OS osteosarcoma cell line **(A-D)** at the  
1356 site chr4:30,909,846, and **(E-H)** at the site chr11:84,769,636. **(A)** IGV screenshots of  
1357 short-read genome sequencing data. Ectopic telomeric repeats (CCCTAA)<sub>n</sub> are shown  
1358 in color. **(B)** Sequencing coverage and allelic ratios of chromosome 4. Orange semi-  
1359 oval: site of the ectopic telomere repeat sequence. **(C)** IGV screenshots of PacBio HiFi  
1360 and Nanopore long-read sequencing data at the site shown in **(A)**. Ectopic telomeric  
1361 repeats in the inverted orientation contained ~650 bp of (CCCTAA)<sub>n</sub> telomeric repeat  
1362 sequences followed by chr22q sub-telomeric sequences, indicative of a chromosomal  
1363 arm fusion event of chr22q to the site at chr4:30,909,846. **(D)** Schematic of telomere-  
1364 spanning fusion event between chromosomes 22q-ter and 4p. **(E)** IGV screenshots of  
1365 short-read genome sequencing data. Ectopic telomeric repeats (CCCTAA)<sub>n</sub> are shown  
1366 in color. **(F)** Sequencing coverage and allelic ratios of chromosome 11. Orange semi-  
1367 oval: site of the ectopic telomere repeat sequence. **(G)** IGV screenshots of PacBio HiFi  
1368 and Nanopore long-read sequencing at the site shown in **(E)**. ~1750 bp of (CCCTAA)<sub>n</sub>  
1369 telomeric repeat sequences are found sequences corresponding to chr11p  
1370 (chr11:43,002,345), suggestive of a complex event consistent with the formation of a  
1371 neotelomere on chr11p, followed by a chromosomal arm fusion event of this  
1372 neotelomere to the site on chr11q (chr11:84,769,636). **(H)** Schematic telomere-  
1373 spanning fusion event between chromosome arms 11q (with a predicted neotelomere)  
1374 and 11p. **(I)** Percentage of new telomeric sites in the inverted orientation that were  
1375 predicted by TelFuse from short-read genome sequencing, and then validated by long-  
1376 read genome sequencing as telomere-spanning chromosome arm fusion events. **(J)**  
1377 Spectral karyogram of chromosome 22 for which a chromosomal arm fusion was  
1378 detected with chromosome 4. Ten U2-OS single cells assessed are as indicated. The  
1379 fusion event between chromosome 22 (yellow) and chromosome 4 (blue) is indicated by  
1380 a red arrow. See also Figure S6.

## Figure 4



1385 **Figure 4 Neotelomeres have similar telomere length distribution as normal**  
1386 **telomeres, while telomeric repeats at sites with chromosomal arm fusions are**  
1387 **short. (A-B)** Telomeric repeat signal observed at a representative Nanopore read with  
1388 **(A)** a neotelomere in U2-OS DNA at chrX:103,320,553, and **(B)** a chromosomal arm  
1389 fusion event in U2-OS DNA at chr4:30,909,846. The length of telomeric repeats on each  
1390 long-read was estimated from these telomeric repeat signal profiles. Boxplots depicting  
1391 the distribution of telomere length found at each neotelomere assessed by Nanopore  
1392 sequencing for the **(C)** U2-OS and **(D)** Hs-746T cell lines. Boxplot depicting length of  
1393 telomeric repeats assessed using Nanopore sequencing for each chromosomal arm  
1394 fusion event in the **(E)** U2-OS and **(F)** Hs-746T cell lines. Note: telomere length for  
1395 neotelomeres and normal chromosomal arms were only estimated using long-reads  
1396 reads that start or end in telomeric repeats, while length of telomeric repeats at  
1397 chromosomal arm fusions were estimated using long-reads with telomeric repeats in the  
1398 middle of the read. Aggregated telomeric length of all long-reads at the normal  
1399 chromosomal arms (p- and q-arms), neotelomeres, and chromosomal arm fusion events  
1400 in the **(G)** U2-OS and **(H)** Hs-746T cell lines. P-values indicated in the plots were  
1401 calculated using the two-sided Wilcoxon Rank Sum test. See also Figure S9.  
1402  
1403  
1404

## Figure 5



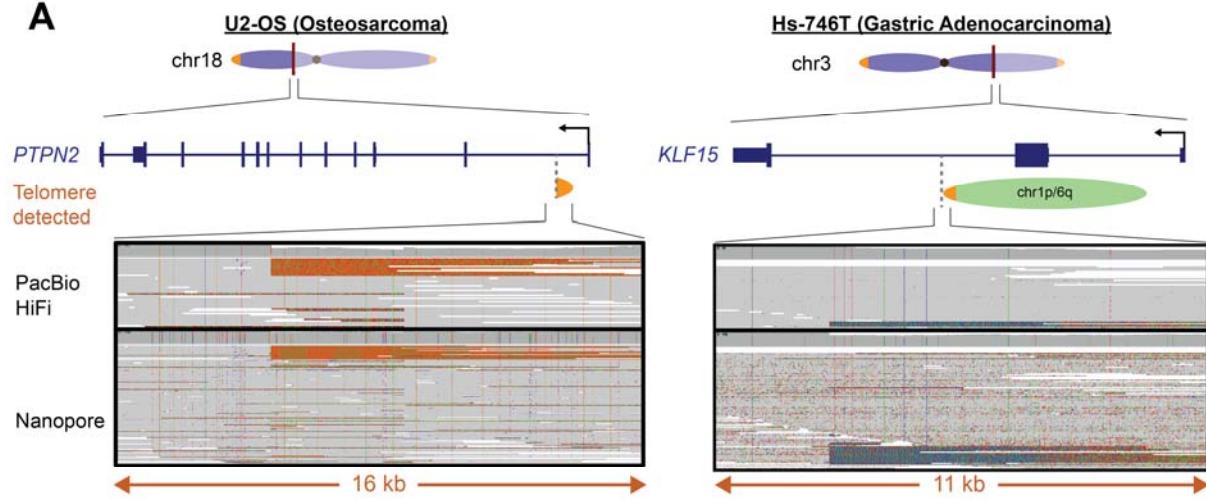
1407 **Figure 5 Putative neotelomeres and chromosomal arm fusion events are detected**  
1408 **as somatic alterations in primary lung adenocarcinoma genomes. (A)** Genome-  
1409 wide distribution of putative neotelomeres and chromosomal arm fusion events in lung  
1410 adenocarcinoma patient samples from The Cancer Genome Atlas (TCGA) (n=95).  
1411 Neotelomeres were inferred from ectopic telomeric sequences in the standard  
1412 orientation, while chromosomal arm fusion events were inferred from ectopic telomeric  
1413 sequences in the inverted orientation, as described in Figure 1B, using short-read  
1414 genome sequencing data. **(B)** Proportion of telomeric alterations (neotelomeres/arm  
1415 fusions) that were found to be germline or somatic. **(C-D)** Examples of neotelomeres  
1416 and chromosomal arm fusion events detected in tumor samples from patients with lung  
1417 adenocarcinoma. **(C)** Neotelomere in tumor DNA from case TCGA-44-4112 at the site  
1418 chr1:214,760,486. **(D)** Chromosomal arm fusion in tumor DNA from case TCGA-49-  
1419 4507 at the site chr17:31,537,163. Top panels: sequencing coverage at the sites of  
1420 interest. Bottom panels: IGV screenshots corresponding to the neotelomere or  
1421 chromosomal arm fusion events in the normal and tumor samples. **(E)** Frequency of  
1422 neotelomeres and chromosomal arm fusion events in lung adenocarcinoma patient  
1423 tumor samples in TCGA.

1424

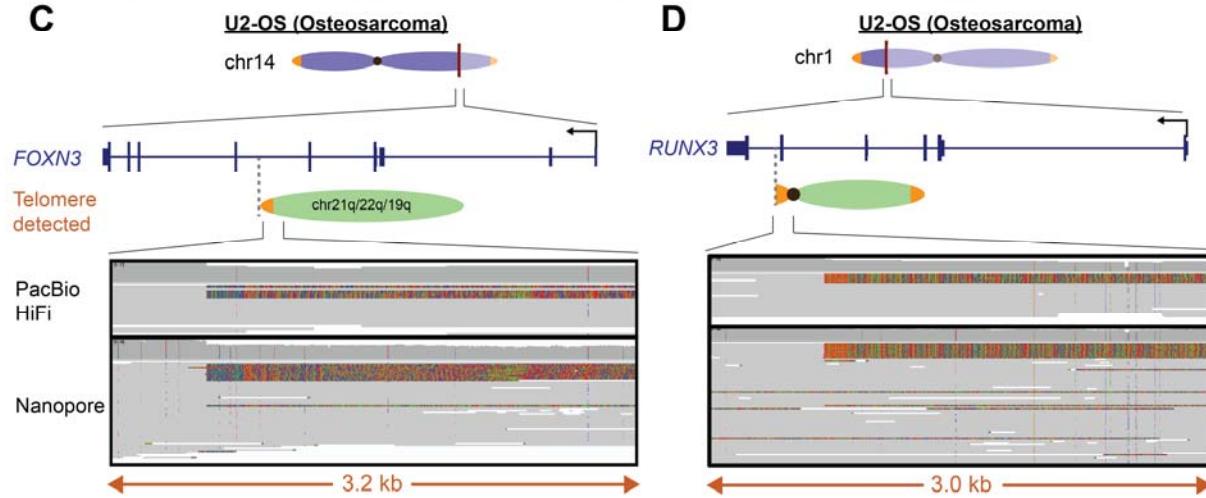
1425

## Figure 6

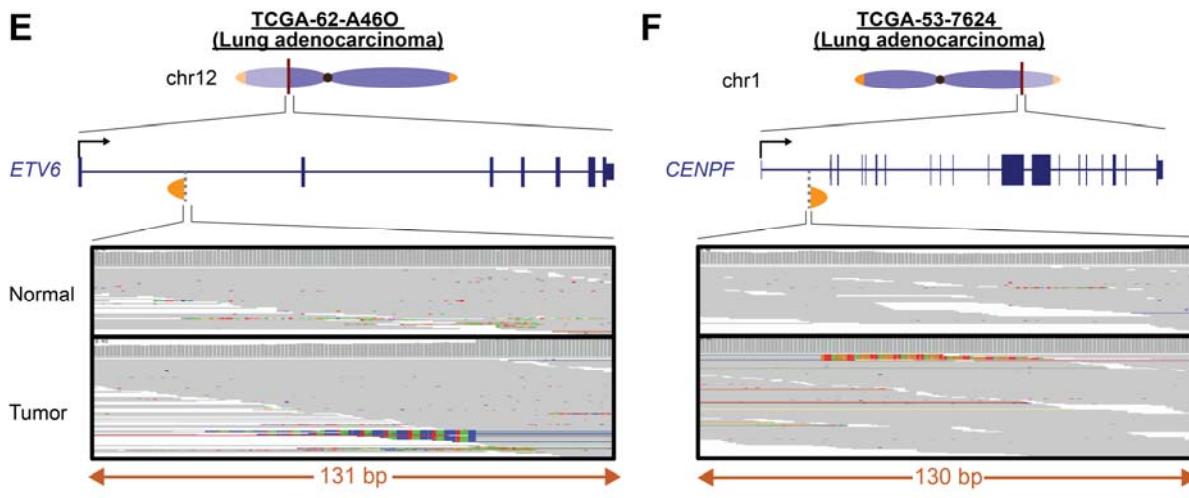
**A**



**C**



**E**

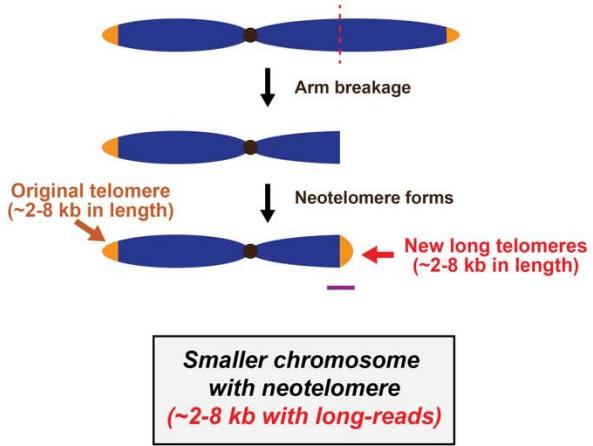


1428 **Figure 6 Neotelomeres and chromosomal arm fusion events disrupt protein**  
1429 **coding genes in cancer cell lines and patient samples. (A)** Disruption of the *PTPN2*  
1430 gene in the U2-OS osteosarcoma cell line at chr18:12,875,538 with addition of a  
1431 neotelomere. **(B)** Disruption of the *KLF15* gene in the Hs-746T gastric adenocarcinoma  
1432 cell line associated with a chromosomal arm fusion event at chr3:126,349,603. **(C)** A  
1433 chromosomal arm fusion event in the U2-OS cell line between a broken chromosome  
1434 14 and the telomere arm of chromosome 21q/22q/19q associated with disruption of the  
1435 *FOXN3* gene at chr14:89,300,563. **(D)** A neotelomere in the U2-OS cell line coupled to  
1436 fusion to a centromere leads to disruption of the *RUNX3* gene at chr1:24,906,321. **(E)** A  
1437 putative neotelomere associated with disruption of the *ETV6* gene in a lung  
1438 adenocarcinoma tumor sample derived from the patient TCGA-62-A46O at the site  
1439 chr12:11,696,012. **(F)** A putative neotelomere associated with disruption of the *CEPF*  
1440 gene in a lung adenocarcinoma tumor sample derived from the patient TCGA-53-7624  
1441 at the site chr1:214,609,478. See also Figure S13.  
1442  
1443

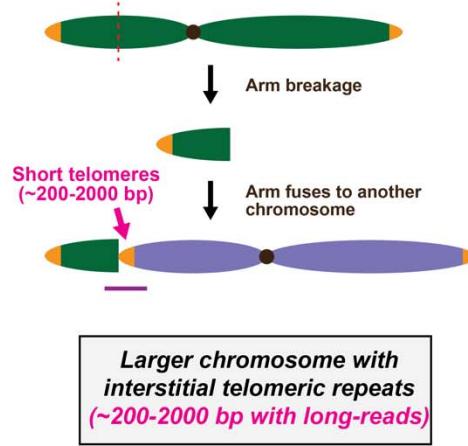
## Figure 7

### Simple events

**A**

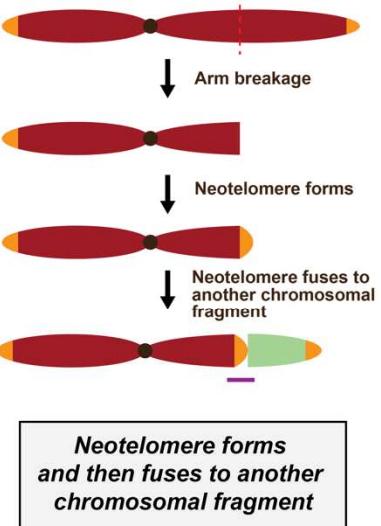


**B**

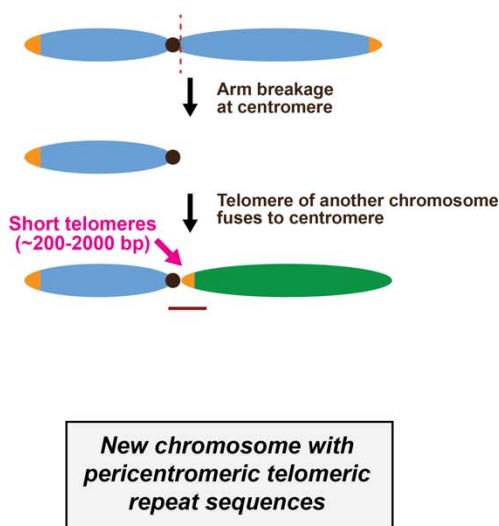


### Complex events

**C**



**D**



#### Legend

► Telomere   ● Centromere   | Breakage site   — Long reads

1446 **Figure 7 Possible models that can account for the different types of telomeric  
1447 repeat sequences observed in this study. (A)** A neotelomere can form after a  
1448 chromosomal arm breakage event. This leads to the generation of a smaller  
1449 chromosome with a neotelomere, similar in repeat length to telomeres found on a  
1450 normal chromosomal arm. **(B)** Chromosome arm fusion where a broken chromosomal  
1451 arm can fuse to another chromosome with very short telomeres. This generates a larger  
1452 chromosome with interstitial telomeric repeat sequences in the middle of the  
1453 chromosome. **(C)** Complex alteration where neotelomere formation is followed by the  
1454 fusion of this neotelomere to another chromosomal fragment. This leads to the  
1455 observation of long-reads in our study which contains telomeric repeat sequences,  
1456 flanked on both sides by intra-chromosomal sequences. **(D)** A complex telomeric  
1457 alteration involving a chromosomal arm break at or very near to the centromere, which  
1458 is fused to another chromosomal arm with very short telomeres. The resultant new  
1459 chromosome has pericentromeric telomeric repeat sequences. Purple line: parts of the  
1460 model supported by long-read genome sequencing data.

1461

1462

1463 **Supplemental Information**

1464

1465 **Supplementary Figure Legends**

1466

1467 **Figure S1 Overview of the TelFuse methodology and the distribution of ectopic**  
1468 **telomeric events in cancer cell line genomes. (A)** Overview of the TelFuse  
1469 computational method to identify ectopic telomere repeat sequences with short-reads  
1470 genome sequencing. Raw sites were first identified from partially mapped read-pairs  
1471 which also contain telomeric repeats on unmapped portions of these reads. A stringent  
1472 set of filters was then applied to ensure the specificity of these calls. These candidate  
1473 sites were then classified by the orientation of telomeric repeats after the breakpoint.  
1474 Specifically, sites with (TTAGGG)<sub>n</sub> repeats after the breakpoint have telomeric repeats  
1475 in the standard orientation, akin to a standard telomere. Conversely, sites containing  
1476 (CCCTAA)<sub>n</sub> telomeric repeats after the breakpoint have telomeric repeat sequences in  
1477 the inverted orientation. **(B)** Detailed overview of method used for the identification of  
1478 ectopic telomeric repeat sites. **(C-E)** Number of ectopic telomeric repeat sites found in  
1479 the genome of each cancer cell line (n=326 cell lines) in **(C)** either orientation, in **(D)** the  
1480 standard orientation, and in **(E)** the inverted orientation. The three cell lines (U2-OS, Hs-  
1481 746T, and NCI-H1184) which have a high frequency of ectopic telomeric events, and  
1482 which were selected for long-read genome sequencing are as indicated.

1483

1484 **Figure S2 Circos plot depicting locations of ectopic telomeric repeat sequences**  
1485 **identified in the U2-OS cell line by TelFuse.** Locations of ectopic telomeric repeats in  
1486 the standard orientation are labelled with red triangles, while those in the inverted  
1487 orientation are labelled with green triangles. Triangles pointing in the anti-clockwise  
1488 direction indicates that the ectopic telomeric repeat sequences are found on the anti-  
1489 clockwise edge, while triangles pointing in the clockwise directions indicates that ectopic  
1490 telomeric repeat sequences are on the clockwise edge. The allelic ratios are depicted in  
1491 red, while the sequencing coverage is labelled in blue. The inner most circle depicts  
1492 translocation events detected in the cell line.

1493

1494 **Figure S3 Sequencing quality of long-read genome sequencing datasets**  
1495 **generated as part of this study.** Sequencing quality metrics for long-read genome  
1496 sequencing data generated using the Nanopore PromethION platform (2x flow cells per  
1497 cell line), and the PacBio platforms (3x SMRT flow cells per cell line) are as indicated.  
1498 The PacBio dataset was further divided into the full dataset consisting of all reads, and  
1499 the PacBio HiFi dataset consisting of reads with a read quality  $\geq 0.9$ . **(A)** Read length  
1500 distribution of the Nanopore and PacBio long-read genome sequencing datasets  
1501 generated in this study. Each sequencing run is indicated separately. The median read  
1502 length for each run are also indicated in the legend for each plot. **(B)** The cumulative  
1503 fraction of sequenced bases above a particular read length for each run is as indicated  
1504 in the plots. The N50 (i.e. minimum read length at which half the bases were  
1505 sequenced) for each run is also indicated in the legend of each plot. **(C)** Fraction of the  
1506 genome sequenced above the stated sequencing depth for each sample is as indicated  
1507 in the plots. The median sequencing coverage across the human genome is also  
1508 indicated in the legend of each plot. **(D)** Distribution of sequence divergence of long-

1509 reads generated by each platform and for each platform is as indicated. Sequence  
1510 divergence (i.e. how much each long-read differs from the reference genome)  
1511 information for each long-read was extracted from long-reads aligned to the GRCh38  
1512 reference genome using minimap2.

1513  
1514 **Figure S4 Additional examples of neotelomeres discovered using long-read**  
1515 **genome sequencing, related to Figure 2. (A-H)** Genomic analysis of telomere repeat  
1516 alterations in the standard orientation that were detected **(A-D)** in the U2-OS  
1517 osteosarcoma cell line at chr7:24,302,169, and **(E-H)** in the NCI-H1184 small cell lung  
1518 carcinoma cell line at chr1:214,460,753, but not in the matched normal cell line (NCI-  
1519 BL1184). **(A)** IGV screenshots of short-read genome sequencing data. Ectopic  
1520 telomeric repeats (CCCTAA)<sub>n</sub> are shown in gold. **(B)** Sequencing coverage and allelic  
1521 ratios of chromosome 7. Orange semi-oval: site of the neotelomeric event. **(C)** IGV  
1522 screenshots depicting long telomeric repeat sequences (TTAGGG)<sub>n</sub> with PacBio HiFi  
1523 (read quality  $\geq 0.9$ ) and Nanopore long-read sequencing at the site shown in **(A)**. **(D)**  
1524 Schematic of neotelomere location on chromosome 7p. **(E)** IGV screenshots of short-  
1525 read genome sequencing data. Ectopic telomeric repeats (TTAGGG)<sub>n</sub> are shown in  
1526 gold. **(F)** Sequencing coverage and allelic ratios of chromosome 1. Orange semi-oval:  
1527 site of the neotelomeric event. **(G)** IGV screenshots depicting long telomeric repeat  
1528 sequences (CCCTAA)<sub>n</sub> with PacBio HiFi (read quality  $\geq 0.9$ ) and Nanopore long-read  
1529 sequencing at the site shown in **(E)**. **(H)** Schematic of neotelomere location on  
1530 chromosome 1q.

1531  
1532 **Figure S5 Degree of chromosomal heterogeneity between cells is chromosome**  
1533 **specific. (A)** Spectral karyogram of a representative U2-OS cell (Cell 01-01) analyzed  
1534 in this study. Chromosomes observed were assigned to each of the 24 possible  
1535 autosomes and sex chromosomes. Chromosomes that could not be assigned were  
1536 labelled as marker chromosomes 'M'. Spectral karyogram of **(B)** chromosome 4 with  
1537 low levels of chromosomal heterogeneity and **(C)** chromosome 15 with high levels of  
1538 chromosomal heterogeneity in ten cells assessed. Red arrows in **(B)** highlights the  
1539 chromosome with translocation between chromosome 4 and 22. Red arrows in **(C)**  
1540 highlights the chromosome with translocation between chromosome 15 and 19.

1541  
1542 **Figure S6 Additional examples of chromosomal arm fusion events revealed by**  
1543 **long-read genome sequencing, related to Figure 3. (A-H)** Genomic analysis of  
1544 telomere repeat alterations in the inverted orientation that were detected in the Hs-746T  
1545 gastric carcinoma cell line **(A-D)** at the site chr11:79,325,679, and **(E-H)** at the site  
1546 chr1:244,201,717. **(A)** IGV screenshots of short-read genome sequencing data. Ectopic  
1547 telomeric repeats (TTAGGG)<sub>n</sub> are shown in color. **(B)** Sequencing coverage and allelic  
1548 ratios of chromosome 11. Orange semi-oval: site of the ectopic telomere repeat  
1549 sequence. **(C)** IGV screenshots of PacBio HiFi (read quality  $\geq 0.9$ ) and Nanopore long-  
1550 read sequencing data at the site shown in **(A)**. Ectopic telomeric repeats in the inverted  
1551 orientation contained ~4.2 kb of (TTAGGG)<sub>n</sub> telomeric repeat sequences followed by  
1552 chr3q/19p sub-telomeric sequences, indicative of a chromosomal arm fusion event of  
1553 chr3q/19p to the site at chr11:79,325,679. **(D)** Schematic of telomere-spanning fusion  
1554 event between chromosomes 3q/19p-ter and 11q. **(E)** IGV screenshots of short-read

1555 genome sequencing data. Ectopic telomeric repeats (TTAGGG)<sub>n</sub> are shown in color. **(F)**  
1556 Sequencing coverage and allelic ratios of chromosome 1. Orange semi-oval: site of the  
1557 ectopic telomere repeat sequence. **(G)** IGV screenshots of PacBio HiFi (read quality  $\geq$   
1558 0.9) and Nanopore long-read sequencing at the site shown in **(E)**. Ectopic telomeric  
1559 repeats in the inverted orientation contained  $\sim$ 3.5 kb of (TTAGGG)<sub>n</sub> telomeric repeat  
1560 sequences followed by chr6q sub-telomeric sequences, indicative of a chromosomal  
1561 arm fusion event of chr6q to the site at chr1:244,201,717. **(H)** Schematic of telomere-  
1562 spanning fusion event between chromosomes 6q-ter and 1q.  
1563

1564 **Figure S7 Estimation of telomere length from telomeric long-reads.** **(A)** Schematic  
1565 depicting how the telomeric region from a single telomeric long-read is defined. The  
1566 TTAGGG motif on a single telomeric long-read is highlighted in yellow on the left, and a  
1567 concentration of telomeric repeats can be observed towards the end of the telomeric  
1568 long-read. The telomeric region from the single long-read can then be defined to  
1569 estimate telomere length on the single long-read. A zoomed-in view of the boundary  
1570 between the sub-telomeric and telomeric region is provided on the right. **(B)** Telomere  
1571 length estimate for each chromosomal arm in the CHM13 cell line determined using  
1572 PacBio HiFi long-read genome sequencing.  
1573

1574 **Figure S8 Telomere length estimates for the cell lines sequenced in this study**  
1575 **using different sequencing platforms.** **(A)** Number of telomeric reads of each class in  
1576 the long-read datasets generated in this study. Long-reads containing telomeric repeats  
1577 were split into four different classes depending on where telomeric repeats were  
1578 observed in the long-read. These four classes are: Full – Long-reads that contains  
1579 telomeric repeat sequences end-to-end, Left – Long-reads that contains telomeric  
1580 repeat sequences on the left edge of the long-read, Right – Long-reads that contains  
1581 telomeric repeat sequences on the right edge of the long-read, and Intra – Long-reads  
1582 that contains telomeric repeat sequences in the middle of the single long-read. The type  
1583 of telomeric repeat sequences observed is also further indicated (i.e. if the reads  
1584 contain (TTAGGG)<sub>n</sub> or (CCCTAA)<sub>n</sub> repeats). Results for the four cell lines sequenced in  
1585 this study by Nanopore, PacBio (All reads), or PacBio HiFi (read quality  $\geq$  0.9)  
1586 sequencing are as indicated. **(D-F)** Telomere length estimates for the four classes of  
1587 telomeric reads in the four cell lines sequenced. Results for each of the sequencing  
1588 platforms: **(D)** Nanopore, **(E)** PacBio (All reads) and **(F)** PacBio HiFi (read quality  $\geq$  0.9)  
1589 are as indicated. **(G-I)** Telomere length estimates for telomeric reads derived from either  
1590 the “forward” strand (i.e. containing (TTAGGG)<sub>n</sub> repeats) or “reverse” strand (i.e.  
1591 containing (CCCTAA)<sub>n</sub> repeats) are as indicated. Results for each of the sequencing  
1592 platforms: **(G)** Nanopore, **(H)** PacBio (All reads) and **(I)** PacBio HiFi (read quality  $\geq$  0.9)  
1593 are as indicated.  
1594

1595 **Figure S9 Length of telomeric repeats at neotelomeres and chromosomal arm**  
1596 **fusion events as estimated using PacBio HiFi sequencing, related to Figure 4.** The  
1597 length of telomeric repeats on each long-read was estimated from these telomeric  
1598 repeat signal profiles. Boxplots depicting the distribution of telomere length found at  
1599 each neotelomere assessed by PacBio HiFi for the **(A)** U2-OS and **(B)** Hs-746T cell  
1600 lines. Boxplot depicting length of telomeric repeats assessed using PacBio HiFi for each

1601 chromosomal arm fusion event in the **(C)** U2-OS and **(D)** Hs-746T cell lines. Note:  
1602 telomere length for neotelomeres and normal chromosomal arms were only estimated  
1603 using long-reads reads that start or end in telomeric repeats, while length of telomeric  
1604 repeats at chromosomal arm fusions were estimated using long-reads with telomeric  
1605 repeats in the middle of the read. Aggregated telomeric length of all long-reads at the  
1606 normal chromosomal arms (p- and q-arms), neotelomeres, and chromosomal arm  
1607 fusion events in the **(E)** U2-OS and **(F)** Hs-746T cell lines. P-values indicated in the  
1608 plots were calculated using the two-sided Wilcoxon Rank Sum test.  
1609

1610 **Figure S10 Representative examples of neotelomeres and chromosomal arm**  
1611 **fusion events detected in patients with lung adenocarcinoma. (A-D)** Normalized  
1612 tumor sequencing coverage of chromosomes with neotelomeres and chromosomal arm  
1613 fusion events predicted by TelFuse analysis of short-read genome sequencing are as  
1614 depicted. Sequencing coverage of the tumor was normalized to the matched normal  
1615 sample (Methods). IGV screenshots of the tumor and matched normal samples with  
1616 neotelomeres and chromosomal arm fusion events are also as indicated. The sites and  
1617 samples represented in the plot are **(A)** the putative chromosomal arm fusion site  
1618 chr22:49,418,106 in TCGA-75-7031, **(B)** the putative neotelomere site chr1:17,644,075  
1619 in TCGA-44-5643, **(C)** the putative chromosomal arm fusion site chr11:96,570,712 in  
1620 TCGA-86-8673, and **(D)** the putative neotelomere site chr12:11,696,012 in TCGA-62-  
1621 A46O.  
1622

1623 **Figure S11 Little to no sequence preference associated with neotelomeres and**  
1624 **chromosomal arm fusion events. (A-B)** The first 6 base-pairs of each stretch of  
1625 telomeric repeat sequence at each **(A)** neotelomere or **(B)** chromosomal arm fusion  
1626 event was assessed and classified into one of six possible circular permutations  
1627 representing the telomeric repeat sequence. **(A)** (top) Schematic illustrating the  
1628 telomeric repeat sequences that are found directly after a breakpoint, and at a  
1629 neotelomere. The first 6 base-pairs of the neotelomere after the breakpoint can occur in  
1630 anyone of six possible circular permutations of the TTAGGG sequence. (bottom) Bar  
1631 plots depicting the frequency of each six possible circular permutations observed on the  
1632 first 6 base-pairs of the neotelomere in the CCLE and TCGA-LUAD cohorts. **(B)** (top)  
1633 Schematic illustrating the telomeric repeat sequences that are found directly after a  
1634 breakpoint, and at a chromosomal arm fusion site. The first 6 base-pairs of the  
1635 chromosomal arm fusion after the breakpoint can occur in anyone of six possible  
1636 circular permutations of the TTAGGG sequence. (bottom) Bar plots depict the frequency  
1637 of each six possible circular permutations observed on the first 6 base-pairs of the  
1638 chromosomal arm fusions observed in the CCLE and TCGA-LUAD cohorts. p-values in  
1639 **(A)** and **(B)** were calculated using the chi-squared test under the assumption that all six  
1640 circular permutations are expected to be observed at the same frequency. The  
1641 expected frequencies are indicated by a grey dotted line, and the number of events  
1642 assessed for each cohort is indicated in the header of each plot. **(C-D)** Sequence logo  
1643 plot representing the frequencies of nucleotides observed near the breakpoints of  
1644 neotelomere and chromosomal arm fusion events. **(C)** (top) Schematic of the  
1645 neotelomere, and the three main regions (genomic region flanking the neotelomere,  
1646 telomeric repeats corresponding to the neotelomeres, and genomic region of the broken

1647 chromosomal fragment that was detached from the neotelomere) associated with these  
1648 events. (bottom) Logo plots representing frequencies of nucleotides in the three main  
1649 regions around a neotelomeric event. **(D)** (top) Schematic of a chromosomal arm fusion  
1650 event, and the four main regions around the breakpoint of the chromosomal arm fusion  
1651 event (genomic region flanking the arm fusion event, telomeric repeats corresponding to  
1652 the chromosomal arm that fused to this site, sub-telomeric region of the arm that  
1653 underwent fusion, and the genomic region of the remaining chromosomal fragment that  
1654 was detached following the chromosomal arm fusion event). (bottom) Frequency of  
1655 nucleotides in the three regions around the breakpoint of a chromosomal arm fusion  
1656 event. **(E-F)** Coverage profiles in the  $\pm 200\text{kb}$  region surrounding a **(E)** neotelomere or  
1657 **(F)** telomere fusion event in the CCLE cohort. The line depicts the median coverage  
1658 observed across all sites, while the shaded area represents the interquartile range.  
1659

1660 **Figure S12 Putative germline ectopic telomeric events observed in lung**  
1661 **adenocarcinoma tumor samples from patients. (A)** Ectopic telomeric repeat  
1662 sequences in the inverted orientation at the site chr12:54,480,142, and in the standard  
1663 orientation at the site chr12:54,494,011 in both the normal (blood) and tumor (lung  
1664 adenocarcinoma) sample for the patient TCGA-44-6778. IGV screenshots depicting  
1665 these observations are as indicated. These observations point to a model where  $\sim 6x$   
1666 (CCCTAA)<sub>n</sub> repeats have integrated into this locus at chr12q, coupled with a deletion of  
1667 regions B and C indicated in the figure. **(B)** Ectopic telomeric repeat sequences in the  
1668 standard orientation were found at the site chr12:25,085,740, and in the inverted  
1669 orientation at the site chr12:25,085,754 in both the normal (blood) and tumor (lung  
1670 adenocarcinoma) sample for the patient TCGA-44-6778. IGV screenshots depicting  
1671 these observations are as indicated. These observations point to a model where  $\sim 7x$   
1672 (CCCTAA)<sub>n</sub> repeats have integrated into this locus at chr12p, coupled with a duplication  
1673 of region B for the event on the left. The event on the right represents the insertion of  
1674 the telomeric repeats without duplication of region B. **(C)** Ectopic telomeric repeat  
1675 sequences in the inverted orientation at the site chr4:184,711,090, and in the standard  
1676 orientation at the site chr4:184,711,103 in both the normal (blood) and tumor (lung  
1677 adenocarcinoma) sample for the patient TCGA-62-A470. IGV screenshots depicting  
1678 these observations are as indicated. These observations point to a model where  $\sim 3x$   
1679 (CCCTAA)<sub>n</sub> repeats have integrated into this locus at chr4q, coupled with a deletion of  
1680 region B found on the reference genome. **(D)** Ectopic telomeric repeat sequences in the  
1681 inverted orientation was found at the site chr6:170,186,789, and in the standard  
1682 orientation at the site chr6:170,186,808 in both the normal (blood) and tumor (lung  
1683 adenocarcinoma) sample for the patient TCGA-44-5643. IGV screenshots depicting  
1684 these observations are as indicated. These observations point to a model where  $>9x$   
1685 (TTAGGG)<sub>n</sub> repeats have integrated into this locus at chr6q, coupled with a duplication  
1686 of region B in the reference genome. **(E)** Ectopic telomeric repeat sequences in the  
1687 inverted orientation was found at the site chr2:192,206,320 in both the normal (adjacent  
1688 lung tissue) and tumor (lung adenocarcinoma) sample for the patient TCGA-55-6987.  
1689 IGV screenshots depicting these observations are as indicated. These observations  
1690 point to a model where 4x (TTAGGG)<sub>n</sub> repeats have integrated into this locus at chr2q,  
1691 together with sub-telomeric sequences corresponding to either chr7q/9q/5q, suggesting  
1692 that a chromosomal arm fusion event has potentially occurred here.

1693

1694 **Figure S13 Additional examples of neotelomeric and chromosomal arm fusion**  
1695 **events which led to gene disruptions, related to Figure 6. (A-B)** Schematic and IGV  
1696 screenshots depicting gene disrupting events caused by neotelomeres or chromosomal  
1697 arm fusion events in cancer cell lines. These were observed in **(A)** the U2-OS  
1698 osteosarcoma cell line where a neotelomere could be observed in the middle of the  
1699 *NRDC* gene at the site chr1:51,828,828, and **(B)** the Hs-746T gastric adenocarcinoma  
1700 cell line where a chromosomal arm fusion event could be observed within the *TENM4*  
1701 gene at the site chr11:79,325,679. **(C-D)** Somatic neotelomere and chromosomal arm  
1702 fusion events observed in primary lung adenocarcinoma tumor samples. These were  
1703 observed in the tumor sample of **(C)** patient TCGA-55-A48Y in the middle of the *FOXP4*  
1704 gene at the position chr6:41,573,027 where a putative chromosomal arm fusion is  
1705 observed, and in **(D)** patient TCGA-55-A493 in the *PRMT7* gene at the position  
1706 chr16:68,349,160, where a putative neotelomere could be observed.  
1707

1708 **Supplementary Table Legends**

1709

1710 **Table S1 SRA accession numbers of short-read genome sequencing data for**  
1711 **cancer cell lines analyzed in this study.**

1712

1713 **Table S2 Detailed information of ectopic telomeric sites identified in cancer cell**  
1714 **lines analyzed in this study.** Sites indicated in this table has perfect telomeric repeat  
1715 sequences on the first 12 base-pairs of the event.

1716

1717 **Table S3 Detailed information of ectopic telomeric sites identified in cancer cell**  
1718 **lines analyzed in this study without perfect telomeric repeat sequences on the**  
1719 **first 12 base-pairs.** Sites indicated in this table do not have perfect telomeric repeat  
1720 sequences on the first 12 base-pairs of the event but contains at least 12 base-pairs of  
1721 telomeric repeat sequences within the soft-clipped sequences.

1722

1723 **Table S4 Sequencing statistics of each long-read genome sequencing run**  
1724 **generated for this study.**

1725

1726 **Table S5 Sequencing statistics for each sample analyzed by long-read genome**  
1727 **sequencing for this study.** Multiple runs for the same sample were aggregated into a  
1728 single dataset, and their corresponding sequencing metrics are as indicated.

1729

1730 **Table S6 Sites assessed, and validation status as determined by long-read**  
1731 **genome sequencing.**

1732

1733 **Table S7 Spectral karyotyping results of ten U2-OS cells.**

1734

1735 **Table S8 Genomic Data Commons accession numbers for TCGA Lung**  
1736 **adenocarcinoma patient samples analyzed in this study.**

1737

1738 **Table S9 Detailed information of ectopic telomeric sites identified in tumor**  
1739 **samples in the cohort of lung adenocarcinoma samples analyzed.**

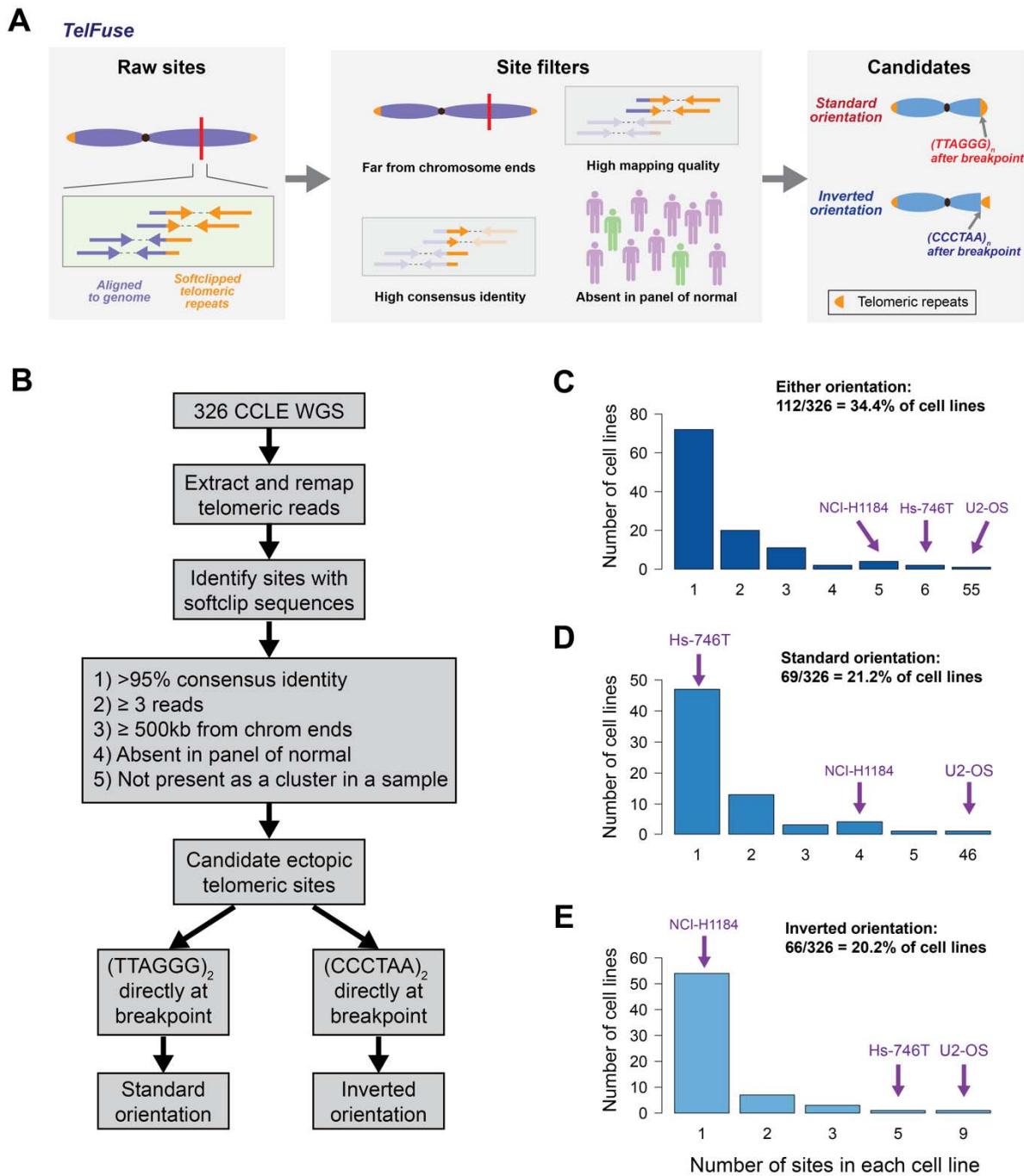
1740

1741 **Table S10 Detailed information of ectopic telomeric sites identified in normal**  
1742 **samples in the cohort of lung adenocarcinoma samples analyzed.**

1743

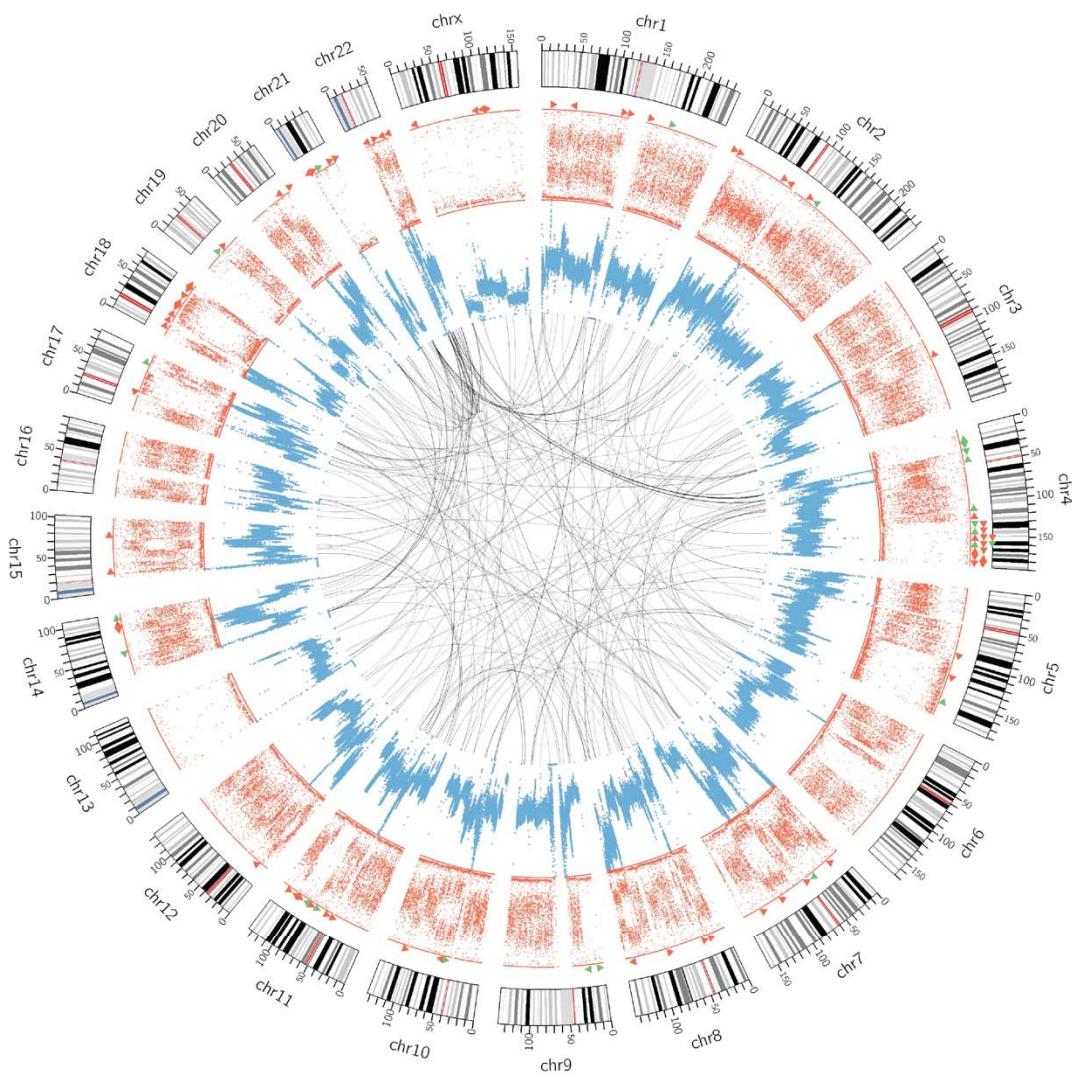
1744

## Supplementary Figure S1



1746 **Figure S1 Overview of the TelFuse methodology and the distribution of ectopic**  
1747 **telomeric events in cancer cell line genomes. (A)** Overview of the TelFuse  
1748 computational method to identify ectopic telomere repeat sequences with short-reads  
1749 genome sequencing. Raw sites were first identified from partially mapped read-pairs  
1750 which also contain telomeric repeats on unmapped portions of these reads. A stringent  
1751 set of filters was then applied to ensure the specificity of these calls. These candidate  
1752 sites were then classified by the orientation of telomeric repeats after the breakpoint.  
1753 Specifically, sites with (TTAGGG)<sub>n</sub> repeats after the breakpoint have telomeric repeats  
1754 in the standard orientation, akin to a standard telomere. Conversely, sites containing  
1755 (CCCTAA)<sub>n</sub> telomeric repeats after the breakpoint have telomeric repeat sequences in  
1756 the inverted orientation. **(B)** Detailed overview of method used for the identification of  
1757 ectopic telomeric repeat sites. **(C-E)** Number of ectopic telomeric repeat sites found in  
1758 the genome of each cancer cell line (n=326 cell lines) in **(C)** either orientation, in **(D)** the  
1759 standard orientation, and in **(E)** the inverted orientation. The three cell lines (U2-OS, Hs-  
1760 746T, and NCI-H1184) which have a high frequency of ectopic telomeric events, and  
1761 which were selected for long-read genome sequencing are as indicated.  
1762  
1763

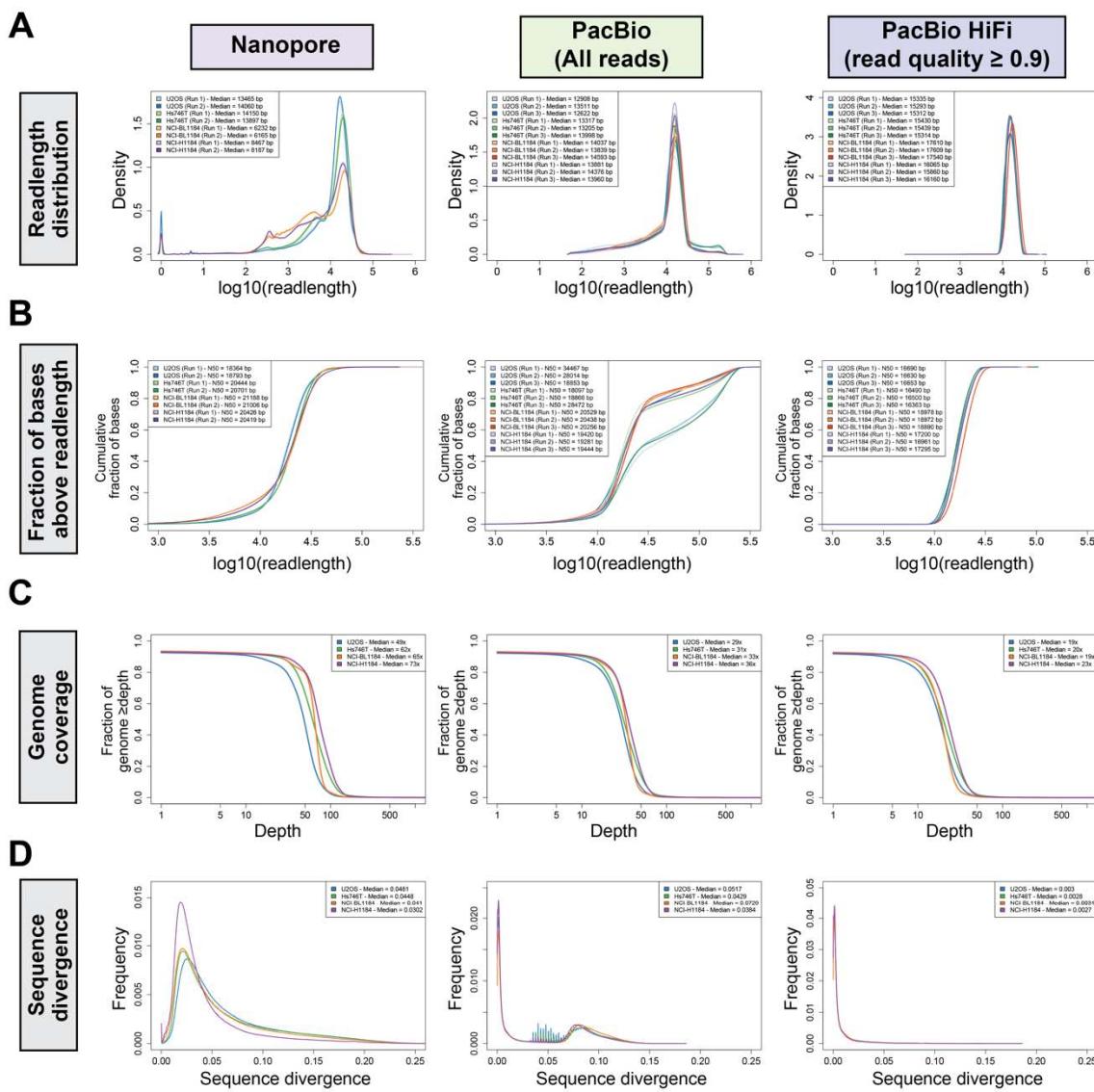
## Supplementary Figure S2



1764  
1765

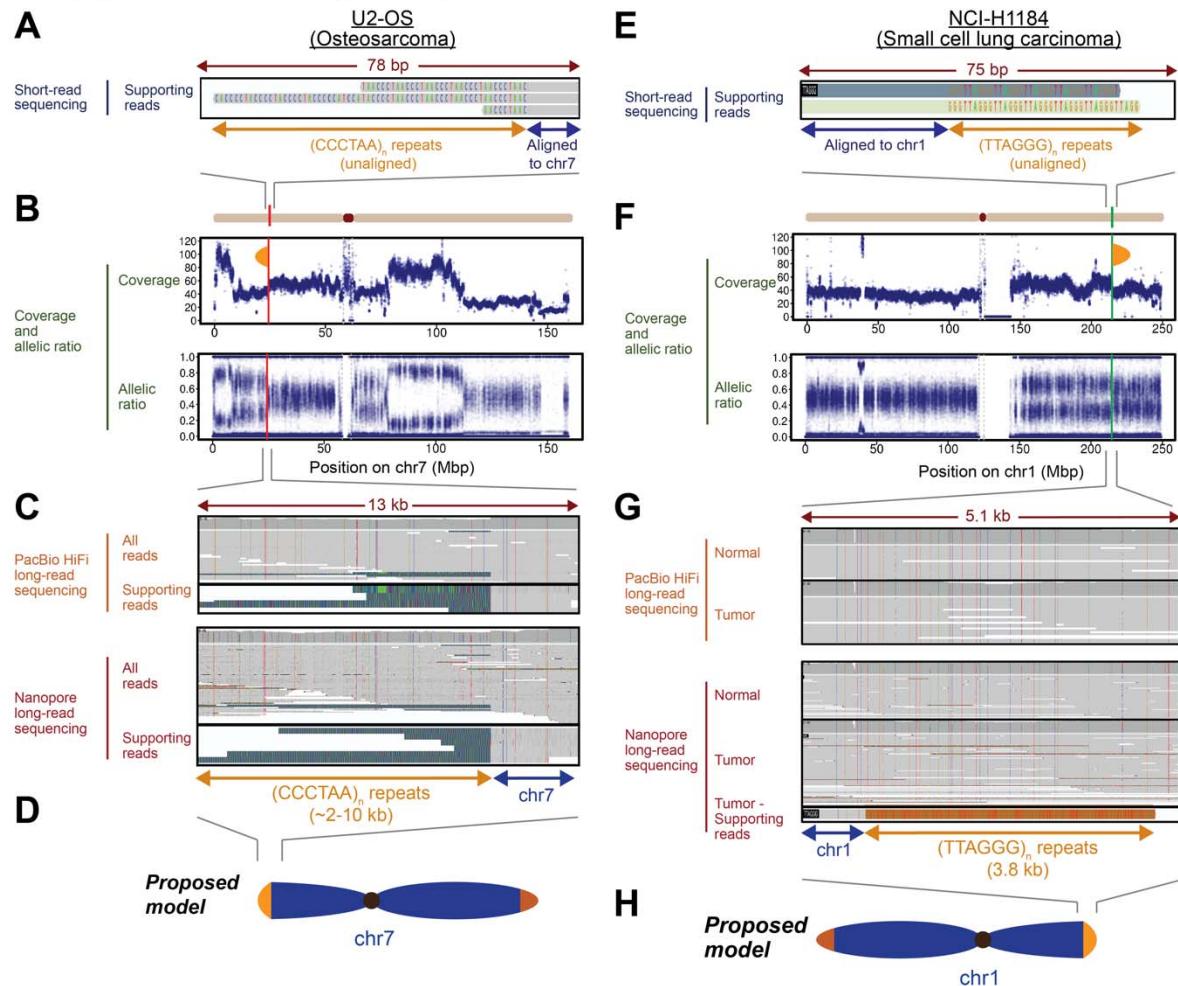
1766 **Figure S2 Circos plot depicting locations of ectopic telomeric repeat sequences**  
1767 **identified in the U2-OS cell line by TelFuse.** Locations of ectopic telomeric repeats in  
1768 the standard orientation are labelled with red triangles, while those in the inverted  
1769 orientation are labelled with green triangles. Triangles pointing in the anti-clockwise  
1770 direction indicates that the ectopic telomeric repeat sequences are found on the anti-  
1771 clockwise edge, while triangles pointing in the clockwise directions indicates that ectopic  
1772 telomeric repeat sequences are on the clockwise edge. The allelic ratios are depicted in  
1773 red, while the sequencing coverage is labelled in blue. The inner most circle depicts  
1774 translocation events detected in the cell line.  
1775  
1776

## Supplementary Figure S3



1779 **Figure S3 Sequencing quality of long-read genome sequencing datasets**  
1780 **generated as part of this study.** Sequencing quality metrics for long-read genome  
1781 sequencing data generated using the Nanopore PromethION platform (2x flow cells per  
1782 cell line), and the PacBio platforms (3x SMRT flow cells per cell line) are as indicated.  
1783 The PacBio dataset was further divided into the full dataset consisting of all reads, and  
1784 the PacBio HiFi dataset consisting of reads with a read quality  $\geq 0.9$ . **(A)** Read length  
1785 distribution of the Nanopore and PacBio long-read genome sequencing datasets  
1786 generated in this study. Each sequencing run is indicated separately. The median read  
1787 length for each run are also indicated in the legend for each plot. **(B)** The cumulative  
1788 fraction of sequenced bases above a particular read length for each run is as indicated  
1789 in the plots. The N50 (i.e. minimum read length at which half the bases were  
1790 sequenced) for each run is also indicated in the legend of each plot. **(C)** Fraction of the  
1791 genome sequenced above the stated sequencing depth for each sample is as indicated  
1792 in the plots. The median sequencing coverage across the human genome is also  
1793 indicated in the legend of each plot. **(D)** Distribution of sequence divergence of long-  
1794 reads generated by each platform and for each platform is as indicated. Sequence  
1795 divergence (i.e. how much each long-read differs from the reference genome)  
1796 information for each long-read was extracted from long-reads aligned to the GRCh38  
1797 reference genome using minimap2.  
1798  
1799

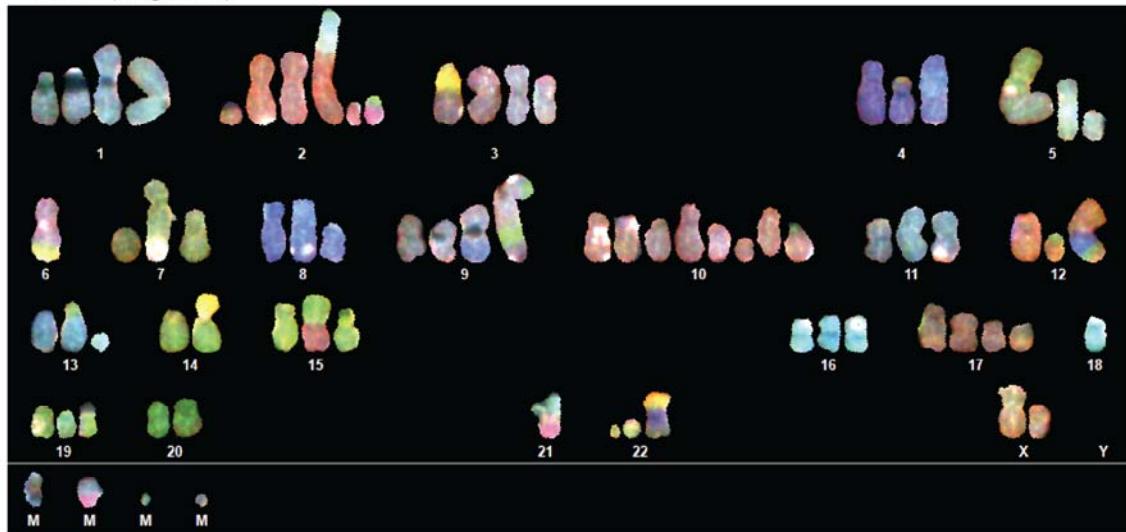
## Supplementary Figure S4



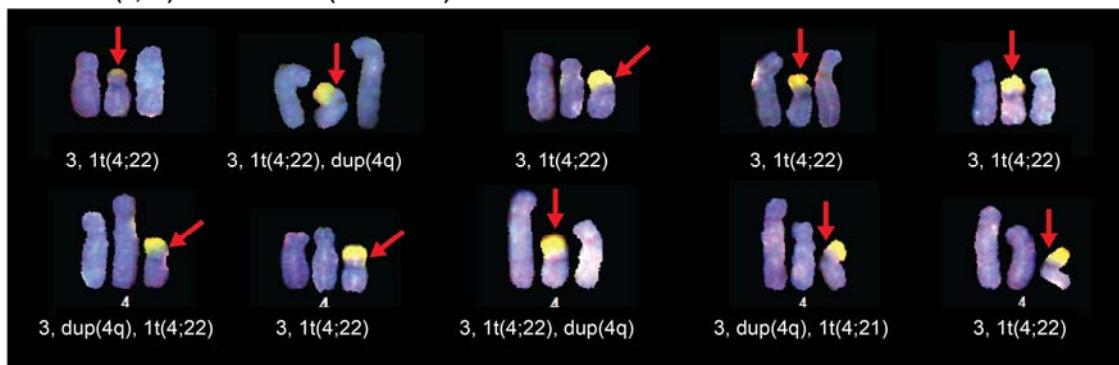
1802 **Figure S4 Additional examples of neotelomeres discovered using long-read**  
1803 **genome sequencing, related to Figure 2. (A-H)** Genomic analysis of telomere repeat  
1804 alterations in the standard orientation that were detected **(A-D)** in the U2-OS  
1805 osteosarcoma cell line at chr7:24,302,169, and **(E-H)** in the NCI-H1184 small cell lung  
1806 carcinoma cell line at chr1:214,460,753, but not in the matched normal cell line (NCI-  
1807 BL1184). **(A)** IGV screenshots of short-read genome sequencing data. Ectopic  
1808 telomeric repeats (CCCTAA)<sub>n</sub> are shown in gold. **(B)** Sequencing coverage and allelic  
1809 ratios of chromosome 7. Orange semi-oval: site of the neotelomeric event. **(C)** IGV  
1810 screenshots depicting long telomeric repeat sequences (TTAGGG)<sub>n</sub> with PacBio HiFi  
1811 (read quality  $\geq 0.9$ ) and Nanopore long-read sequencing at the site shown in **(A)**. **(D)**  
1812 Schematic of neotelomere location on chromosome 7p. **(E)** IGV screenshots of short-  
1813 read genome sequencing data. Ectopic telomeric repeats (TTAGGG)<sub>n</sub> are shown in  
1814 gold. **(F)** Sequencing coverage and allelic ratios of chromosome 1. Orange semi-oval:  
1815 site of the neotelomeric event. **(G)** IGV screenshots depicting long telomeric repeat  
1816 sequences (CCCTAA)<sub>n</sub> with PacBio HiFi (read quality  $\geq 0.9$ ) and Nanopore long-read  
1817 sequencing at the site shown in **(E)**. **(H)** Schematic of neotelomere location on  
1818 chromosome 1q.  
1819  
1820

## Supplementary Figure S5

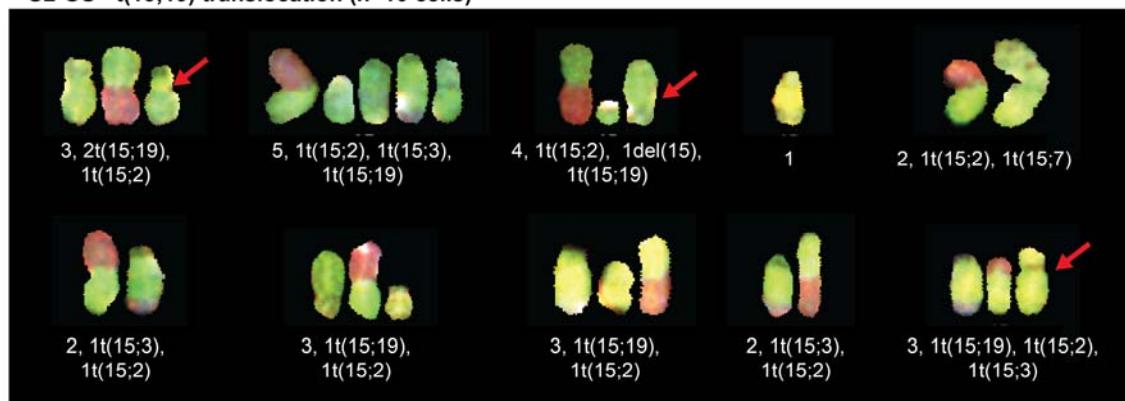
**A** U2-OS (single cell)



**B** U2-OS - t(4;22) translocation (n=10 cells)



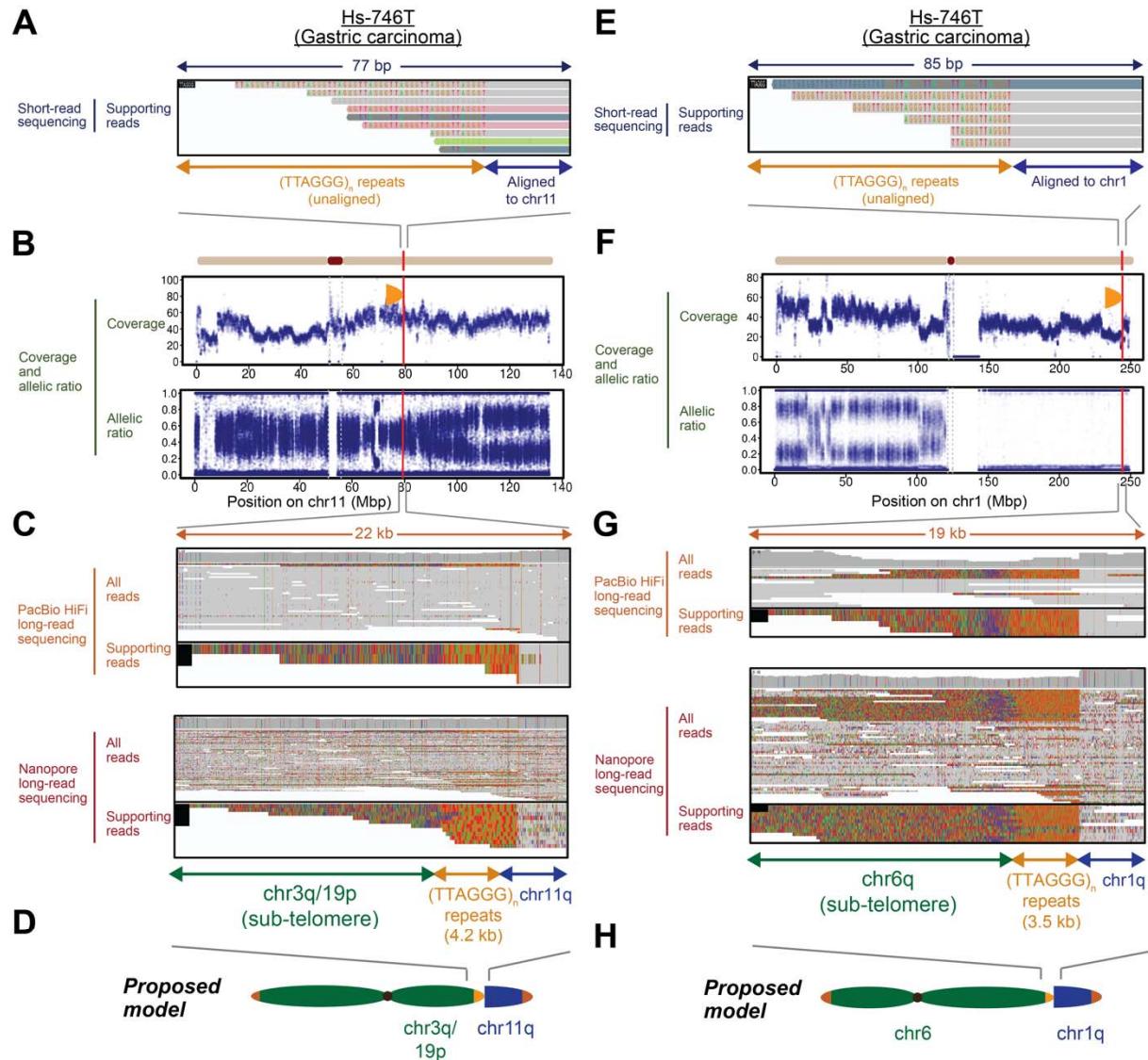
**C** U2-OS - t(15;19) translocation (n=10 cells)



1821  
1822

1823 **Figure S5 Degree of chromosomal heterogeneity between cells is chromosome**  
1824 **specific. (A)** Spectral karyogram of a representative U2-OS cell (Cell 01-01) analyzed  
1825 in this study. Chromosomes observed were assigned to each of the 24 possible  
1826 autosomes and sex chromosomes. Chromosomes that could not be assigned were  
1827 labelled as marker chromosomes 'M'. Spectral karyogram of **(B)** chromosome 4 with  
1828 low levels of chromosomal heterogeneity and **(C)** chromosome 15 with high levels of  
1829 chromosomal heterogeneity in ten cells assessed. Red arrows in **(B)** highlights the  
1830 chromosome with translocation between chromosome 4 and 22. Red arrows in **(C)**  
1831 highlights the chromosome with translocation between chromosome 15 and 19.  
1832  
1833

## Supplementary Figure S6



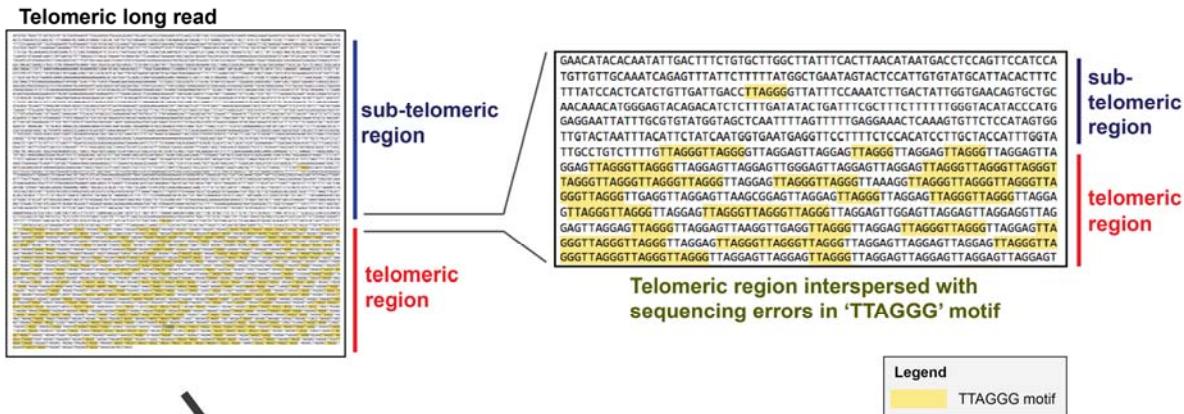
1836 **Figure S6 Additional examples of chromosomal arm fusion events revealed by**  
1837 **long-read genome sequencing, related to Figure 3. (A-H)** Genomic analysis of  
1838 telomere repeat alterations in the inverted orientation that were detected in the Hs-746T  
1839 gastric carcinoma cell line **(A-D)** at the site chr11:79,325,679, and **(E-H)** at the site  
1840 chr1:244,201,717. **(A)** IGV screenshots of short-read genome sequencing data. Ectopic  
1841 telomeric repeats (TTAGGG)<sub>n</sub> are shown in color. **(B)** Sequencing coverage and allelic  
1842 ratios of chromosome 11. Orange semi-oval: site of the ectopic telomere repeat  
1843 sequence. **(C)** IGV screenshots of PacBio HiFi (read quality  $\geq 0.9$ ) and Nanopore long-  
1844 read sequencing data at the site shown in **(A)**. Ectopic telomeric repeats in the inverted  
1845 orientation contained  $\sim 4.2$  kb of (TTAGGG)<sub>n</sub> telomeric repeat sequences followed by  
1846 chr3q/19p sub-telomeric sequences, indicative of a chromosomal arm fusion event of  
1847 chr3q/19p to the site at chr11:79,325,679. **(D)** Schematic of telomere-spanning fusion  
1848 event between chromosomes 3q/19p-ter and 11q. **(E)** IGV screenshots of short-read  
1849 genome sequencing data. Ectopic telomeric repeats (TTAGGG)<sub>n</sub> are shown in color. **(F)**  
1850 Sequencing coverage and allelic ratios of chromosome 1. Orange semi-oval: site of the  
1851 ectopic telomere repeat sequence. **(G)** IGV screenshots of PacBio HiFi (read quality  $\geq$   
1852 0.9) and Nanopore long-read sequencing at the site shown in **(E)**. Ectopic telomeric  
1853 repeats in the inverted orientation contained  $\sim 3.5$  kb of (TTAGGG)<sub>n</sub> telomeric repeat  
1854 sequences followed by chr6q sub-telomeric sequences, indicative of a chromosomal  
1855 arm fusion event of chr6q to the site at chr1:244,201,717. **(H)** Schematic of telomere-  
1856 spanning fusion event between chromosomes 6q-ter and 1q.

1857

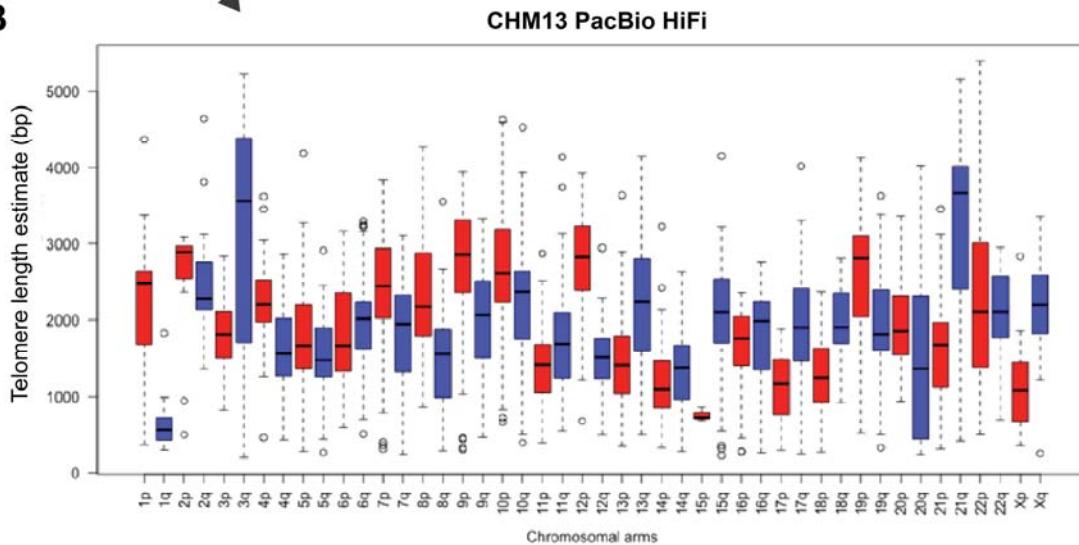
1858

## Supplementary Figure S7

A

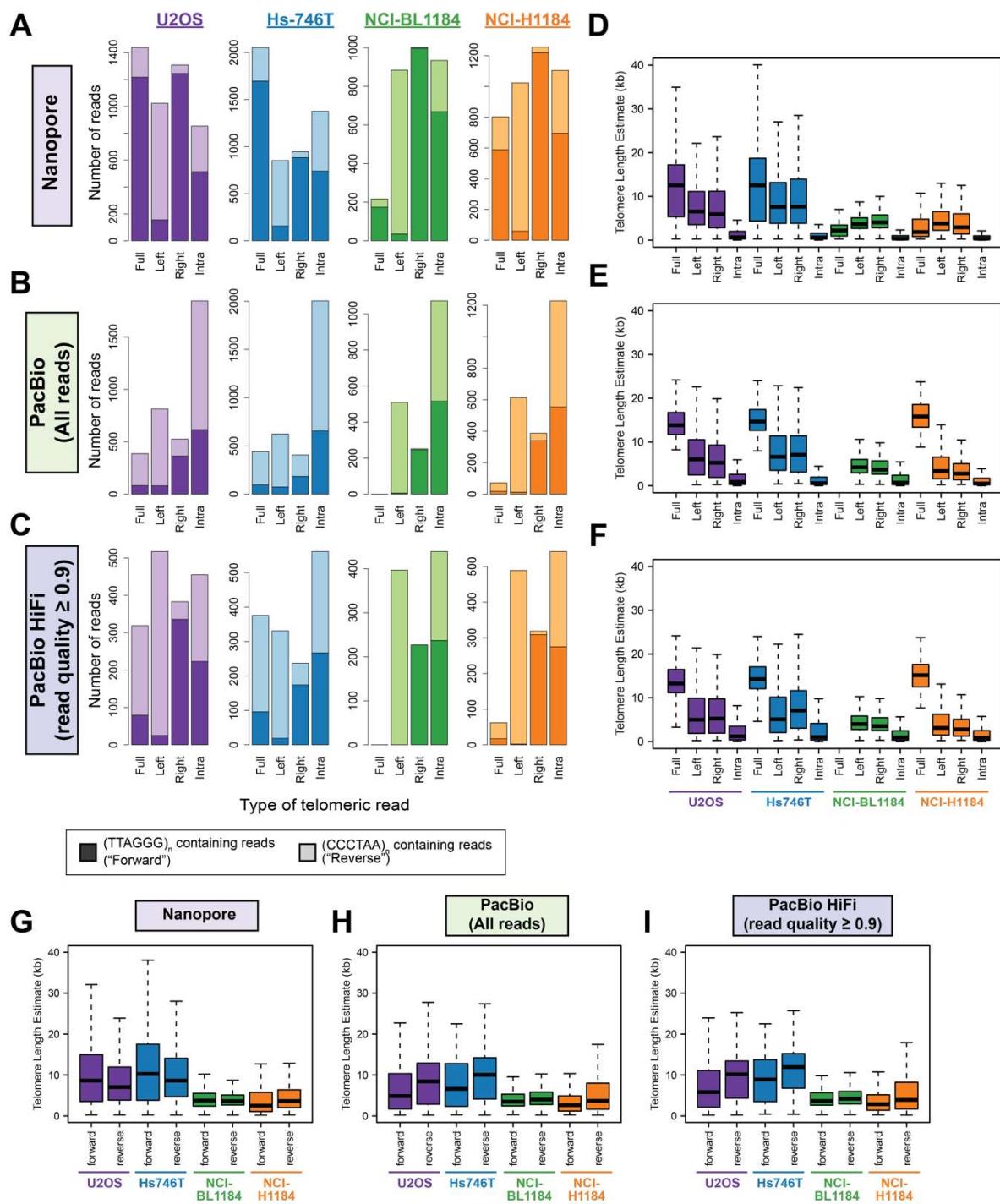


B



1861 **Figure S7 Estimation of telomere length from telomeric long-reads. (A)** Schematic  
1862 depicting how the telomeric region from a single telomeric long-read is defined. The  
1863 TTAGGG motif on a single telomeric long-read is highlighted in yellow on the left, and a  
1864 concentration of telomeric repeats can be observed towards the end of the telomeric  
1865 long-read. The telomeric region from the single long-read can then be defined to  
1866 estimate telomere length on the single long-read. A zoomed-in view of the boundary  
1867 between the sub-telomeric and telomeric region is provided on the right. **(B)** Telomere  
1868 length estimate for each chromosomal arm in the CHM13 cell line determined using  
1869 PacBio HiFi long-read genome sequencing.  
1870  
1871

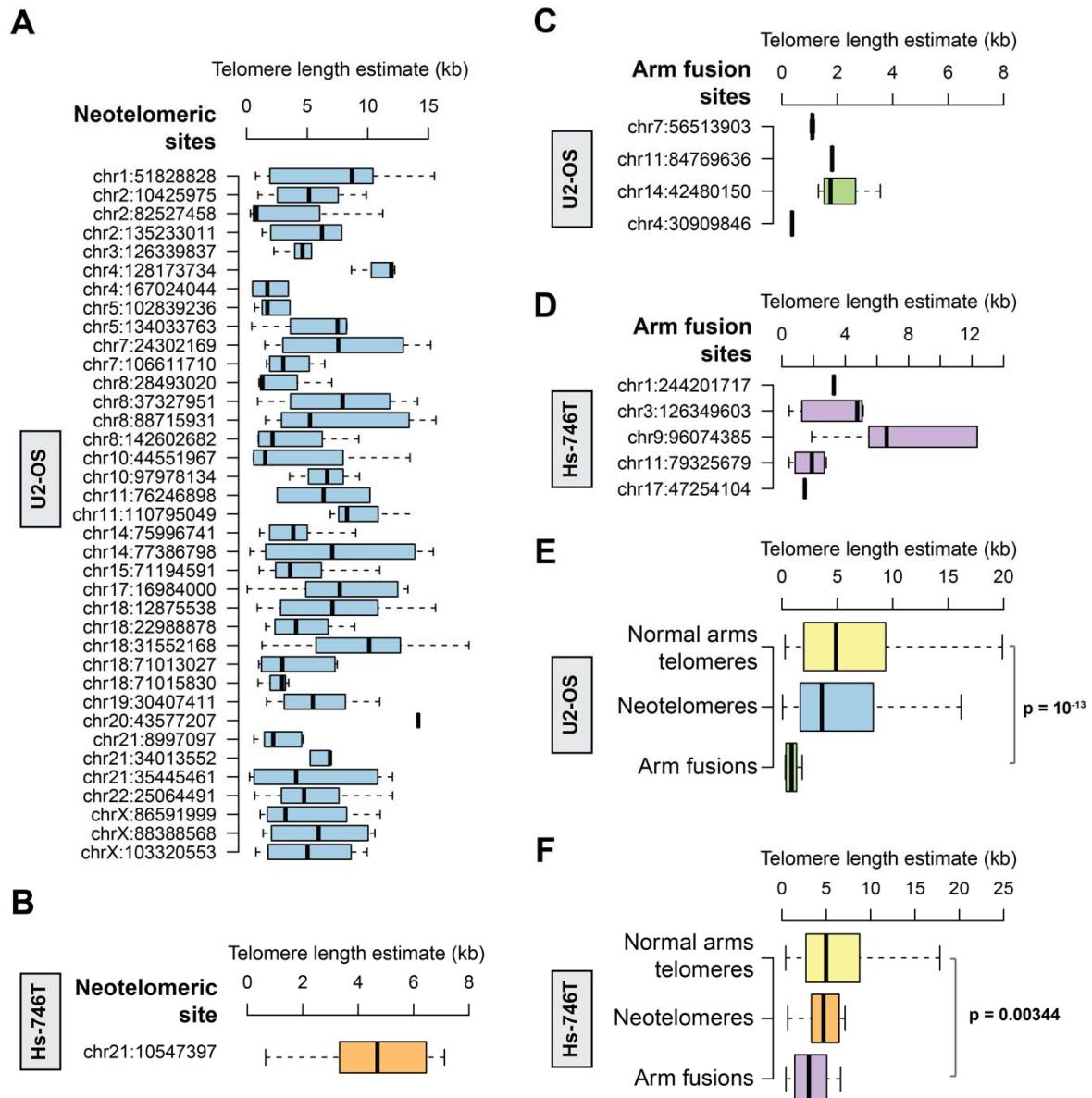
## Supplementary Figure S8



1874 **Figure S8 Telomere length estimates for the cell lines sequenced in this study**  
1875 **using different sequencing platforms. (A)** Number of telomeric reads of each class in  
1876 the long-read datasets generated in this study. Long-reads containing telomeric repeats  
1877 were split into four different classes depending on where telomeric repeats were  
1878 observed in the long-read. These four classes are: Full – Long-reads that contains  
1879 telomeric repeat sequences end-to-end, Left – Long-reads that contains telomeric  
1880 repeat sequences on the left edge of the long-read, Right – Long-reads that contains  
1881 telomeric repeat sequences on the right edge of the long-read, and Intra – Long-reads  
1882 that contains telomeric repeat sequences in the middle of the single long-read. The type  
1883 of telomeric repeat sequences observed is also further indicated (i.e. if the reads  
1884 contain (TTAGGG)<sub>n</sub> or (CCCTAA)<sub>n</sub> repeats). Results for the four cell lines sequenced in  
1885 this study by Nanopore, PacBio (All reads), or PacBio HiFi (read quality  $\geq 0.9$ )  
1886 sequencing are as indicated. **(D-F)** Telomere length estimates for the four classes of  
1887 telomeric reads in the four cell lines sequenced. Results for each of the sequencing  
1888 platforms: **(D)** Nanopore, **(E)** PacBio (All reads) and **(F)** PacBio HiFi (read quality  $\geq 0.9$ )  
1889 are as indicated. **(G-I)** Telomere length estimates for telomeric reads derived from either  
1890 the “forward” strand (i.e. containing (TTAGGG)<sub>n</sub> repeats) or “reverse” strand (i.e.  
1891 containing (CCCTAA)<sub>n</sub> repeats) are as indicated. Results for each of the sequencing  
1892 platforms: **(G)** Nanopore, **(H)** PacBio (All reads) and **(I)** PacBio HiFi (read quality  $\geq 0.9$ )  
1893 are as indicated.

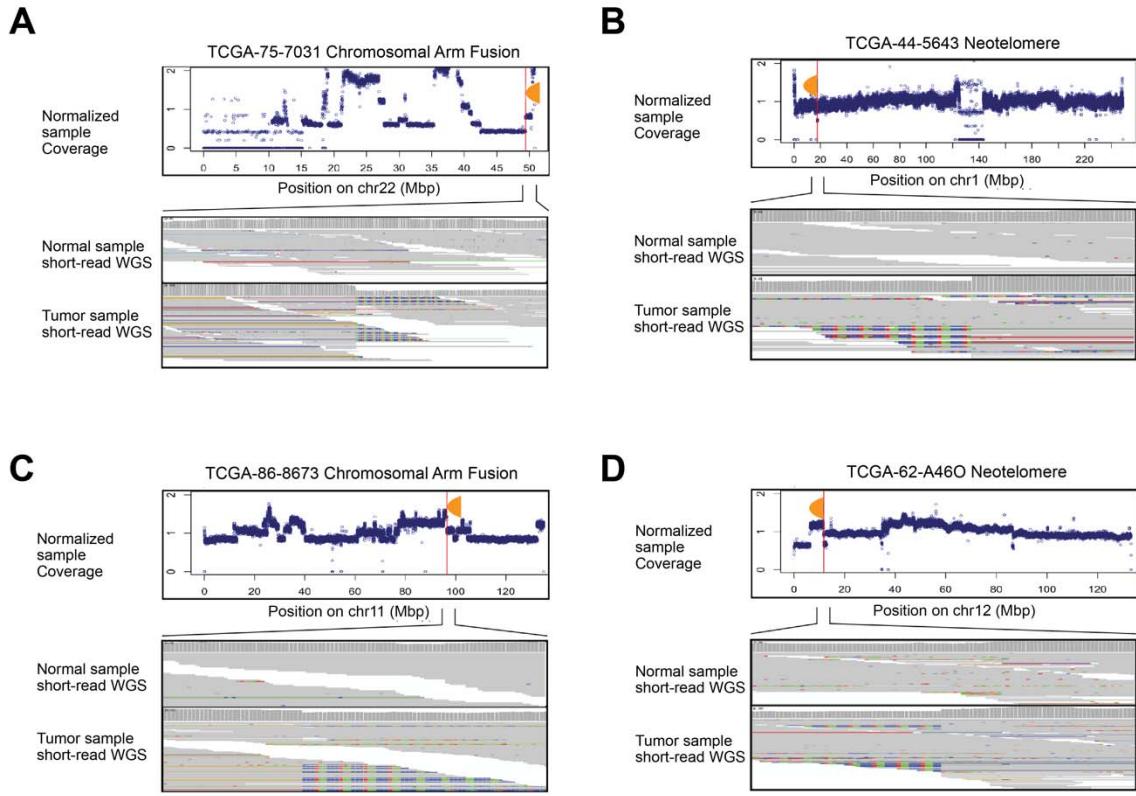
1894  
1895

## Supplementary Figure S9



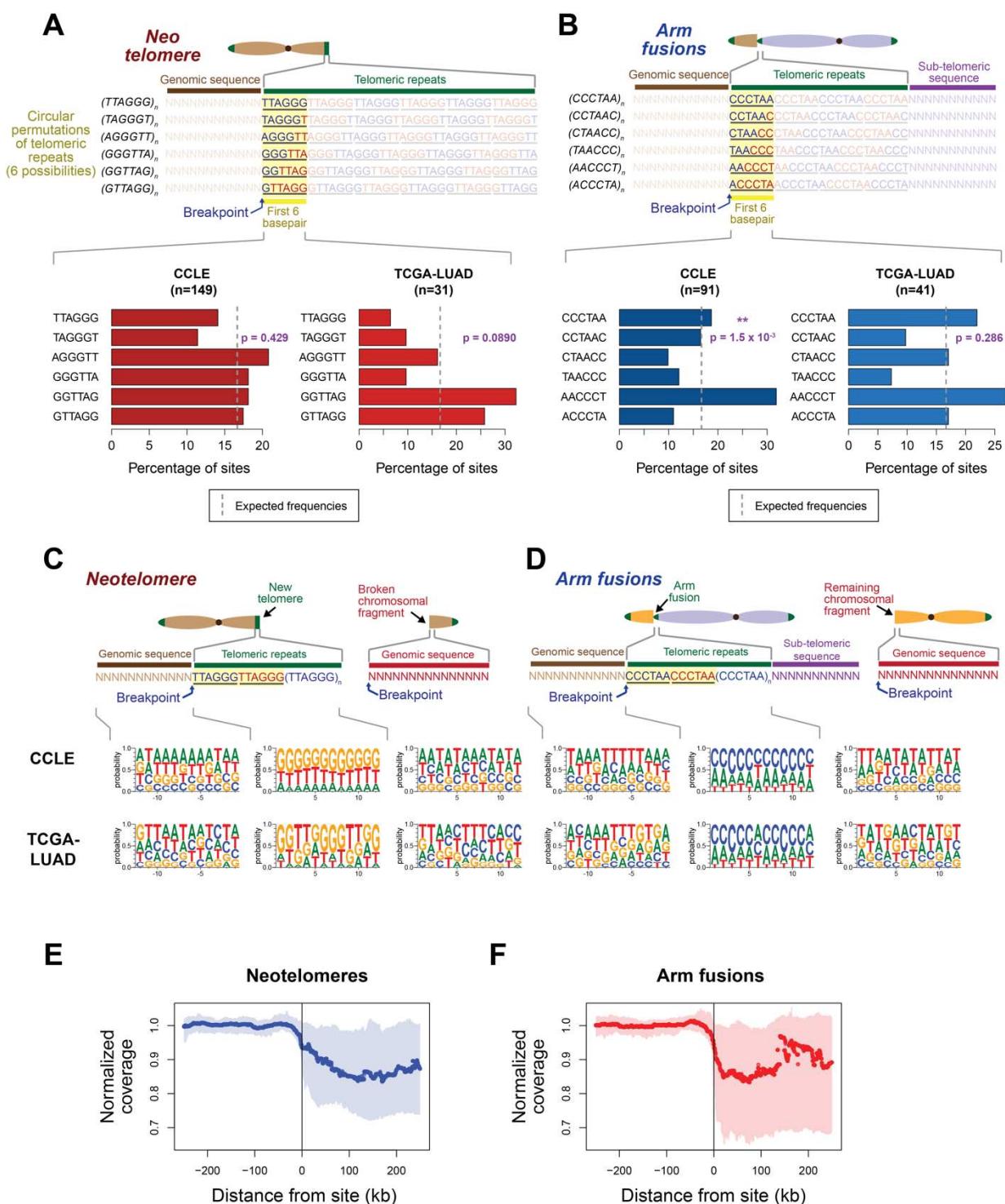
1898 **Figure S9 Length of telomeric repeats at neotelomeres and chromosomal arm**  
1899 **fusion events as estimated using PacBio HiFi sequencing, related to Figure 4.** The  
1900 length of telomeric repeats on each long-read was estimated from these telomeric  
1901 repeat signal profiles. Boxplots depicting the distribution of telomere length found at  
1902 each neotelomere assessed by PacBio HiFi for the **(A)** U2-OS and **(B)** Hs-746T cell  
1903 lines. Boxplot depicting length of telomeric repeats assessed using PacBio HiFi for each  
1904 chromosomal arm fusion event in the **(C)** U2-OS and **(D)** Hs-746T cell lines. Note:  
1905 telomere length for neotelomeres and normal chromosomal arms were only estimated  
1906 using long-reads reads that start or end in telomeric repeats, while length of telomeric  
1907 repeats at chromosomal arm fusions were estimated using long-reads with telomeric  
1908 repeats in the middle of the read. Aggregated telomeric length of all long-reads at the  
1909 normal chromosomal arms (p- and q-arms), neotelomeres, and chromosomal arm  
1910 fusion events in the **(E)** U2-OS and **(F)** Hs-746T cell lines. P-values indicated in the  
1911 plots were calculated using the two-sided Wilcoxon Rank Sum test.  
1912  
1913

## Supplementary Figure S10



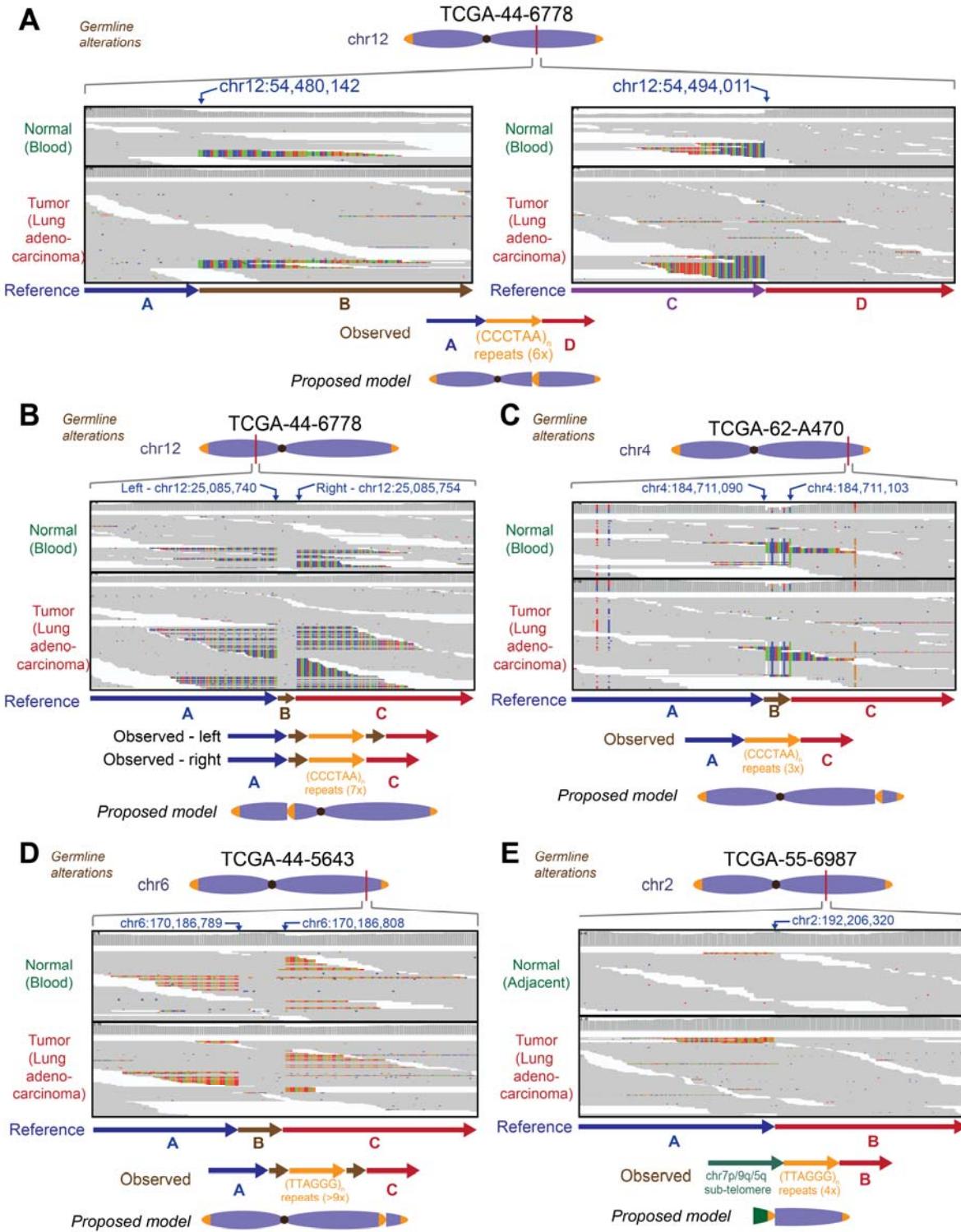
1916 **Figure S10 Representative examples of neotelomeres and chromosomal arm**  
1917 **fusion events detected in patients with lung adenocarcinoma. (A-D)** Normalized  
1918 tumor sequencing coverage of chromosomes with neotelomeres and chromosomal arm  
1919 fusion events predicted by TelFuse analysis of short-read genome sequencing are as  
1920 depicted. Sequencing coverage of the tumor was normalized to the matched normal  
1921 sample (Methods). IGV screenshots of the tumor and matched normal samples with  
1922 neotelomeres and chromosomal arm fusion events are also as indicated. The sites and  
1923 samples represented in the plot are **(A)** the putative chromosomal arm fusion site  
1924 chr22:49,418,106 in TCGA-75-7031, **(B)** the putative neotelomere site chr1:17,644,075  
1925 in TCGA-44-5643, **(C)** the putative chromosomal arm fusion site chr11:96,570,712 in  
1926 TCGA-86-8673, and **(D)** the putative neotelomere site chr12:11,696,012 in TCGA-62-  
1927 A46O.  
1928  
1929

## Supplementary Figure S11



1932 **Figure S11 Little to no sequence preference associated with neotelomeres and**  
1933 **chromosomal arm fusion events. (A-B)** The first 6 base-pairs of each stretch of  
1934 telomeric repeat sequence at each **(A)** neotelomere or **(B)** chromosomal arm fusion  
1935 event was assessed and classified into one of six possible circular permutations  
1936 representing the telomeric repeat sequence. **(A)** (top) Schematic illustrating the  
1937 telomeric repeat sequences that are found directly after a breakpoint, and at a  
1938 neotelomere. The first 6 base-pairs of the neotelomere after the breakpoint can occur in  
1939 anyone of six possible circular permutations of the TTAGGG sequence. (bottom) Bar  
1940 plots depicting the frequency of each six possible circular permutations observed on the  
1941 first 6 base-pairs of the neotelomere in the CCLE and TCGA-LUAD cohorts. **(B)** (top)  
1942 Schematic illustrating the telomeric repeat sequences that are found directly after a  
1943 breakpoint, and at a chromosomal arm fusion site. The first 6 base-pairs of the  
1944 chromosomal arm fusion after the breakpoint can occur in anyone of six possible  
1945 circular permutations of the TTAGGG sequence. (bottom) Bar plots depict the frequency  
1946 of each six possible circular permutations observed on the first 6 base-pairs of the  
1947 chromosomal arm fusions observed in the CCLE and TCGA-LUAD cohorts. p-values in  
1948 **(A)** and **(B)** were calculated using the chi-squared test under the assumption that all six  
1949 circular permutations are expected to be observed at the same frequency. The  
1950 expected frequencies are indicated by a grey dotted line, and the number of events  
1951 assessed for each cohort is indicated in the header of each plot. **(C-D)** Sequence logo  
1952 plot representing the frequencies of nucleotides observed near the breakpoints of  
1953 neotelomere and chromosomal arm fusion events. **(C)** (top) Schematic of the  
1954 neotelomere, and the three main regions (genomic region flanking the neotelomere,  
1955 telomeric repeats corresponding to the neotelomeres, and genomic region of the broken  
1956 chromosomal fragment that was detached from the neotelomere) associated with these  
1957 events. (bottom) Logo plots representing frequencies of nucleotides in the three main  
1958 regions around a neotelomeric event. **(D)** (top) Schematic of a chromosomal arm fusion  
1959 event, and the four main regions around the breakpoint of the chromosomal arm fusion  
1960 event (genomic region flanking the arm fusion event, telomeric repeats corresponding to  
1961 the chromosomal arm that fused to this site, sub-telomeric region of the arm that  
1962 underwent fusion, and the genomic region of the remaining chromosomal fragment that  
1963 was detached following the chromosomal arm fusion event). (bottom) Frequency of  
1964 nucleotides in the three regions around the breakpoint of a chromosomal arm fusion  
1965 event. **(E-F)** Coverage profiles in the  $\pm 200$ kb region surrounding a **(E)** neotelomere or  
1966 **(F)** telomere fusion event in the CCLE cohort. The line depicts the median coverage  
1967 observed across all sites, while the shaded area represents the interquartile range.  
1968  
1969

## Supplementary Figure S12



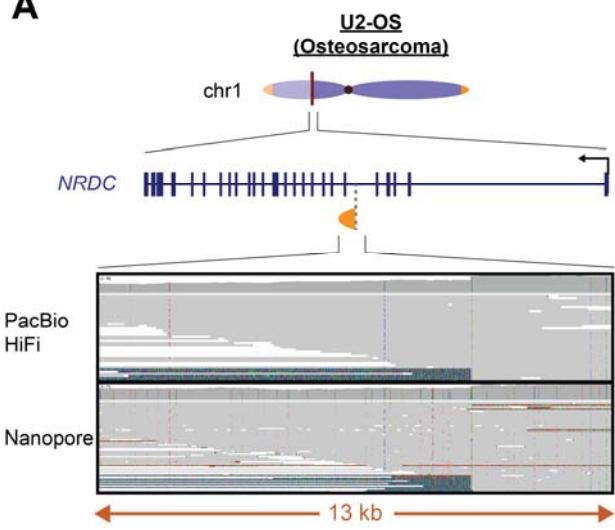
1970  
1971

1972 **Figure S12 Putative germline ectopic telomeric events observed in lung**  
1973 **adenocarcinoma tumor samples from patients. (A)** Ectopic telomeric repeat  
1974 sequences in the inverted orientation at the site chr12:54,480,142, and in the standard  
1975 orientation at the site chr12:54,494,011 in both the normal (blood) and tumor (lung  
1976 adenocarcinoma) sample for the patient TCGA-44-6778. IGV screenshots depicting  
1977 these observations are as indicated. These observations point to a model where ~6x  
1978 (CCCTAA)<sub>n</sub> repeats have integrated into this locus at chr12q, coupled with a deletion of  
1979 regions B and C indicated in the figure. **(B)** Ectopic telomeric repeat sequences in the  
1980 standard orientation were found at the site chr12:25,085,740, and in the inverted  
1981 orientation at the site chr12:25,085,754 in both the normal (blood) and tumor (lung  
1982 adenocarcinoma) sample for the patient TCGA-44-6778. IGV screenshots depicting  
1983 these observations are as indicated. These observations point to a model where ~7x  
1984 (CCCTAA)<sub>n</sub> repeats have integrated into this locus at chr12p, coupled with a duplication  
1985 of region B for the event on the left. The event on the right represents the insertion of  
1986 the telomeric repeats without duplication of region B. **(C)** Ectopic telomeric repeat  
1987 sequences in the inverted orientation at the site chr4:184,711,090, and in the standard  
1988 orientation at the site chr4:184,711,103 in both the normal (blood) and tumor (lung  
1989 adenocarcinoma) sample for the patient TCGA-62-A470. IGV screenshots depicting  
1990 these observations are as indicated. These observations point to a model where ~3x  
1991 (CCCTAA)<sub>n</sub> repeats have integrated into this locus at chr4q, coupled with a deletion of  
1992 region B found on the reference genome. **(D)** Ectopic telomeric repeat sequences in the  
1993 inverted orientation was found at the site chr6:170,186,789, and in the standard  
1994 orientation at the site chr6:170,186,808 in both the normal (blood) and tumor (lung  
1995 adenocarcinoma) sample for the patient TCGA-44-5643. IGV screenshots depicting  
1996 these observations are as indicated. These observations point to a model where >9x  
1997 (TTAGGG)<sub>n</sub> repeats have integrated into this locus at chr6q, coupled with a duplication  
1998 of region B in the reference genome. **(E)** Ectopic telomeric repeat sequences in the  
1999 inverted orientation was found at the site chr2:192,206,320 in both the normal (adjacent  
2000 lung tissue) and tumor (lung adenocarcinoma) sample for the patient TCGA-55-6987.  
2001 IGV screenshots depicting these observations are as indicated. These observations  
2002 point to a model where 4x (TTAGGG)<sub>n</sub> repeats have integrated into this locus at chr2q,  
2003 together with sub-telomeric sequences corresponding to either chr7q/9q/5q, suggesting  
2004 that a chromosomal arm fusion event has potentially occurred here.

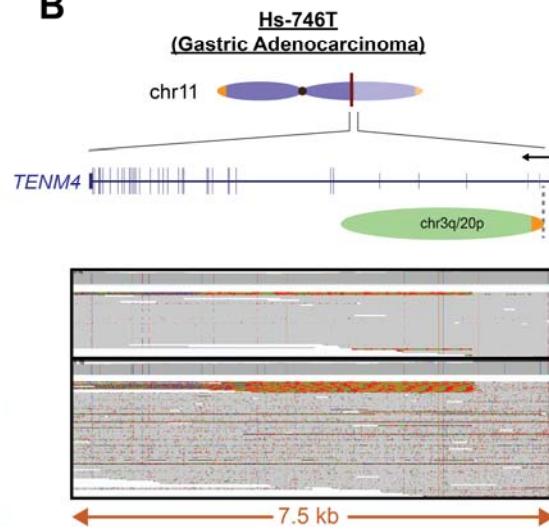
2005  
2006

## Supplementary Figure S13

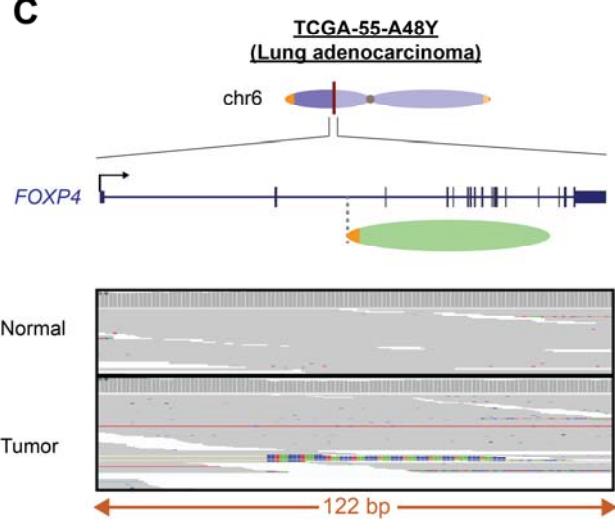
**A**



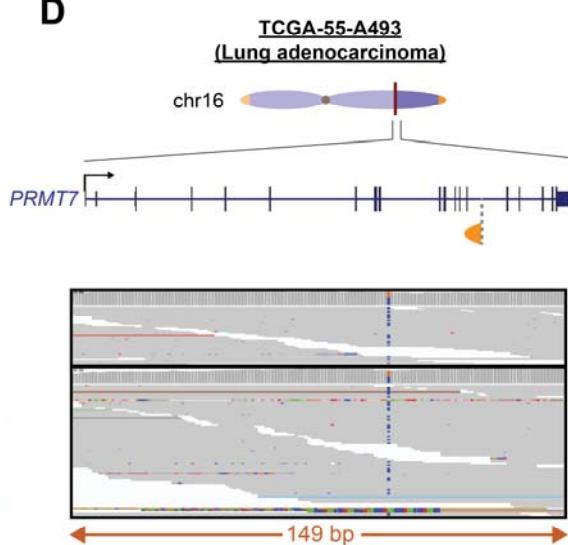
**B**



**C**



**D**



2009 **Figure S13 Additional examples of neotelomeric and chromosomal arm fusion**  
2010 **events which led to gene disruptions, related to Figure 6. (A-B)** Schematic and IGV  
2011 screenshots depicting gene disrupting events caused by neotelomeres or chromosomal  
2012 arm fusion events in cancer cell lines. These were observed in **(A)** the U2-OS  
2013 osteosarcoma cell line where a neotelomere could be observed in the middle of the  
2014 *NRDC* gene at the site chr1:51,828,828, and **(B)** the Hs-746T gastric adenocarcinoma  
2015 cell line where a chromosomal arm fusion event could be observed within the *TENM4*  
2016 gene at the site chr11:79,325,679. **(C-D)** Somatic neotelomere and chromosomal arm  
2017 fusion events observed in primary lung adenocarcinoma tumor samples. These were  
2018 observed in the tumor sample of **(C)** patient TCGA-55-A48Y in the middle of the *FOXP4*  
2019 gene at the position chr6:41,573,027 where a putative chromosomal arm fusion is  
2020 observed, and in **(D)** patient TCGA-55-A493 in the *PRMT7* gene at the position  
2021 chr16:68,349,160, where a putative neotelomere could be observed.  
2022  
2023