TARPON - a Telomere Analysis and Research Pipeline Optimized for Nanopore

Nathaniel Deimler[1,2,*], David Ho[1,2], Paul Norbert[3], Peter Baumann[1,2,4,*]

[1]Department of Biology, Hanns-Dieter-Hüsch-Weg 15, Johannes Gutenberg University, 55128 Mainz, Germany; [2]Institute for Quantitative and Computational Biosciences, Johannes von Müller Weg 6, 55128 Mainz, Germany; [3]Institute for the History, Philosophy, and Ethics of Medicine, Johannes Gutenberg University Medical Center, Mainz Germany
[4]Institute of Molecular Biology, Ackermannweg 4, 55128 Mainz, Germany

* Correspondence: peter@baumannlab.org and
nathanieldeimler.research@gmail.com

Keywords: telomere length, Nextflow, long-read

Abstract

Background

Long-read sequencing has transformed many areas of biology and holds significant promise for telomere research by enabling analysis of nucleotide-level resolution chromosome arm–specific telomere length in both model organisms and humans. However, the adoption of new technologies, particularly in clinical or diagnostic contexts, requires careful validation to recognize potential technical and computational limitations.

Results

We present **TARPON** (Telomere Analysis and Research Pipeline Optimized for Nanopore), a best-practices Nextflow pipeline designed for the analysis of telomeres sequenced on the Oxford Nanopore Technologies (ONT) platform. TARPON can be executed via the command line or integrated into ONT's EPI2ME agent, providing a user-friendly graphical interface for those without computational training. Nextflow's container-based architecture eliminates dependency conflicts, thereby streamlining deployment across platforms.

TARPON isolates telomeric repeat–containing reads, assigns strand specificity, and identifies enrichment probes that can be used both for demultiplexing and for confirming capture-based library preparation. To ensure that the analysis is restricted to full-length telomeres, reads lacking a capture probe or non-telomeric sequence on the opposite end are excluded. A sliding-window approach defines the subtelomere-to-telomere boundary, followed by quality filtering to remove low-quality or subtelomeric reads that passed earlier

46 steps. The pipeline generates customizable statistics, text-based summaries, and publication-
47 ready visualizations (HTML, PNG, PDF). While default settings are optimized for diagnostic
48 workflows, all parameters are easily adjustable via the GUI or command line to support
49 diverse applications. These include telomere analyses in variant-rich samples (e.g., ALT-
50 positive tumors) and organisms with non-canonical telomeric repeats such as some insects
51 (GTTAG) and certain plants (GGTTTAG).

52 *Conclusions*

53 TARPON is the first complete and experimentally validated pipeline for Nanopore-
54 based telomere analysis requiring no data pre-processing or prior bioinformatics expertise,
55 while offering flexibility for advanced users.

56

57 **Introduction**

58 Telomeres, the structures that protect the ends of linear eukaryotic chromosomes, are
59 comprised of G-rich DNA repeat sequences, proteins, and telomeric RNA (1). In humans,
60 telomeres span from 3–15 kb and terminate in a 100 to 200 nucleotide single-stranded G-rich
61 3' overhang (2,3). Each time a cell duplicates, approximately 50–150 base pairs of terminal
62 sequence are lost from each chromosome end due to the end replication problem (4,5),
63 resulting in progressive telomere shortening (6,7). Telomeric repeat arrays protect the
64 integrity of the coding regions of the genome by temporarily buffering this sequence attrition.
65 However, in the absence of mechanisms that replenish telomeric DNA, progressive telomere
66 shortening eventually triggers cellular senescence as telomeres reach a critical length (8).
67 Telomere length is maintained in highly proliferative cells by telomerase, a ribonucleoprotein
68 complex comprised at its core of the catalytic protein Telomerase Reverse Transcriptase
69 (TERT) and a non-coding Telomerase RNA subunit, TR (9,10). Telomere Biology Disorders
70 (TBDs) are a symptomatically heterogeneous group of syndromes associated with abnormally
71 short telomeres or telomere instability (11,12). The broad spectrum of symptoms resembles
72 the aging process, ranging from changes in skin pigmentation and nail dystrophy to severe
73 effects such as pulmonary fibrosis and total bone marrow failure (13).

74 Accurate measurement of telomere length is therefore essential for both research and
75 clinical applications. Multiple techniques have been developed including, but not limited to,
76 terminal restriction fragment (TRF) analysis by Southern blotting, quantitative PCR (qPCR),
77 and fluorescent in situ hybridization (FISH) (14). Terminal restriction fragment (TRF) length
78 analysis has long been considered the gold standard in research settings (15,16). In this
79 method, genomic DNA is digested by restriction enzymes that frequently cut within the
80 genome but not within telomeric repeats. Digested DNA is then resolved on an agarose gel

81 and hybridized with a labeled telomere-specific probe. Although this method shows low inter-
82 laboratory variation (17), reproducibility is limited by poorly described protocols for
83 digestion and gel quantification (18). Moreover, variations in subtelomeric sequences or
84 DNA modifications may lead to apparent inter-individual differences in telomere length (19).

85 Quantitative PCR (qPCR) estimates telomere length indirectly by comparing the
86 quantity of telomeric repeat amplification products to a single-copy gene product (20). Its
87 simplicity, low cost, and minimal DNA input requirements have enabled widespread use in
88 clinical samples and large-scale comparisons (21,22). However, qPCR has been shown to
89 yield variable results depending on laboratory protocols (17), DNA extraction methods (23),
90 and storage conditions. For instance, samples stored in 4% formaldehyde showed increased
91 telomere length measurements over time, a bias not observed with samples preserved in
92 RNAlater (24).

93 Flow-FISH, currently the clinical gold standard, uses fluorescently labeled peptide
94 nucleic acid (PNA) probes to detect telomeric DNA in permeabilized cells, with fluorescence
95 intensity measured by flow cytometry (25). It avoids certain biases associated with TRF and
96 qPCR and enables direct comparison to reference standards (26). However, Flow-FISH
97 requires fresh blood samples, limiting its use on archived or bio-banked material.

98 In summary, each of the established methods for telomere length measurement has
99 specific strengths and limitations. A reliable, reproducible, and user-friendly method that
100 works across a wide range of sample types—including fresh and archived specimens—
101 remains a critical but unmet need.

102 Third generation long-read sequencing has emerged as a powerful tool across
103 biological disciplines, including aging and telomere research. Oxford Nanopore Technologies

104 (ONT) sequencing has enabled nucleotide-resolution analysis of human telomeres (27–29).
105 In parallel, PacBio HiFi sequencing has also been used to generate high-throughput, single-
106 molecule telomere length measurements at nucleotide resolution across diverse human cell
107 lines and patient samples (30). However, telomeres represent only ~0.01% of the human
108 genome, and whole-genome sequencing yields relatively few telomeric reads. To address this,
109 two enrichment strategies have been developed: a physical enrichment using biotinylated
110 duplexes and streptavidin-coated beads ("duplex-enriched") (29) and a library preparation
111 method that captures telomeric ends using a telomere-specific splint and ONT's adapter
112 overhang ("splint-enriched") (27,28). Both strategies increase telomeric read recovery and
113 append a known capture probe to the distal telomere end, confirming full-length capture.

114 Unlike traditional approaches that return a single statistic (usually mean or median)
115 that represents the telomere length per sample, long-read sequencing enables single-molecule
116 resolution of telomere length distributions. With sufficient subtelomeric sequence, reads can
117 be aligned to the genome, allowing chromosome arm–specific telomere assignment, as shown
118 in yeast (31) and the human cell line HG002 (27–29). However, this approach is currently
119 limited to samples with high-quality subtelomeric reference assemblies (32).

120 Despite these promising advances, no standardized or validated pipeline exists for the
121 analysis of Nanopore-based telomere data. Here, we present **TARPON** (Telomere Analysis
122 and Research Pipeline Optimized for Nanopore), an experimentally validated, end-to-end
123 pipeline for analyzing telomere reads obtained via ONT sequencing. TARPON is
124 implemented in Nextflow and can be executed via the command line or through the user-
125 friendly EPI2ME graphical interface, which requires no programming experience. The
126 pipeline supports both duplex- and splint-enriched libraries and includes automated quality
127 control, capture probe detection, telomere length quantification, and comprehensive

128 reporting. Parameters can be modified through either interface, enabling both novice and
129 expert users to tailor TARPON to a wide range of experimental applications.

130 **Methods**

131 *Sample Selection and DNA Isolation*

132 Genomic DNA for the below described telomere enrichment protocols had multiple
133 sources. HG002 genomic DNA was purchased directly from the Coriell Institute for Medical
134 Research (NA24385); this sample originates from an immortalized lymphoblastoid cell line
135 and was chosen for analysis because of its publicly available, high quality subtelomeric
136 reference genome. Clinical sample genomic DNA (WB60) originates from whole blood of a
137 60-year-old individual and was extracted using Chemagen Chemagic DNA Blood Kit Special
138 for 5 mL on the Chemagen Robot MSM1 following manufacturer's instruction. HEK293T
139 genomic DNA was isolated using the Qiagen Blood and Cell Culture DNA Midi Kit (Cat. no.
140 13343) following manufacturer's recommendations.

141 *Splint Capture Based Telomere Enrichment*

142 Telomeric sequencing via the ONT Telo-Seq protocol was conducted following
143 manufacturer's recommendations. Briefly, 15 µg of genomic DNA was incubated with Telo
144 Adapter mix (TLO, 100 nM) in 200 µL reactions containing 1x rCutSmart Buffer, 1 mM
145 rATP, and 50 U/µL Quick T4 DNA Ligase (NEB36056) at 35°C for 16 hours, followed by
146 heat inactivation at 65°C for 10 minutes. The ligation reaction was split into four 50 µL
147 aliquots. EcoRV-HF (20 U, NEB R3195S) was added to each aliquot and incubated at 37°C
148 for 30 minutes to digest bulk genomic DNA, followed by heat inactivation at 65°C for 20
149 minutes. The aliquots were pooled and subjected to 3' dA tailing using NEBNext dA-tailing
150 reaction buffer and 45 U Klenow Exo- (NEB E6053) in a 250 µL reaction at 37°C for 30
151 minutes.

152 SPRI bead purification (1x) was performed using AMPure XP beads (Beckman
153 Coulter A63882). Beads were pelleted on a magnet and washed twice with 80% ethanol. DNA
154 was eluted at 37°C for 15 minutes with 198 µL nuclease-free water. DNA (196 µL) was
155 incubated in a 200 µL reaction containing 100 nM Telo-Splint (TLS; 2 µL of 10 µM stock),
156 50 mM NaCl (2 µL of 5 M), and held at 50°C for 1 hour. The Telo-Splint contains a 5-nt
157 overhang complementary to the ONT native adapter. The Telo-Splint/TLO combination is
158 hereafter referred to as the "capture probe" to differentiate it from the ONT sequencing
159 adapter.

160 A 0.5x SPRI bead purification was performed as above, and DNA was eluted into 32
161 µL nuclease-free water. Sequencing adapter ligation was performed using 10,000 U Quick
162 T4 DNA Ligase (NEB E6056) and 5 µL native adapter in 1x NEBNext Quick Ligation buffer
163 at 21°C for 20 minutes. Another 0.5x SPRI bead cleanup was carried out, followed by two
164 washes with long fragment buffer (LFB). DNA was eluted into 25 µL elution buffer and
165 loaded onto a R10.4.1 MinION flow cell for a 72-hour sequencing run.

166 *Duplex Capture Based Telomere Enrichment*

167 Telomeric sequencing via a duplex capture approach was conducted using a modified
168 version of Karimian et al. (2024). A biotinylated oligo (Oligo BLoli9407, Supplemental Table
169 1) was annealed to a complementary oligo with a three-telomeric-repeat overhang (Oligo
170 BLoli9406, Supplemental Table 1) in 1x Annealing Buffer [20 mM Tris/Ac (pH 7.5), 50 mM
171 NaCl, 2 mM $MgCl_2$] at 20 nM final concentration. The mix was heated to 99°C for 1 minute
172 and cooled to room temperature at 1°C/min.

173 This duplex, hereafter referred to as the "capture probe," was ligated to 3 µg genomic
174 DNA in 50 µL reactions (16 replicates) containing 75 fmol capture probe, 1x HiFi Taq Ligase
175 Buffer (B0647SVIAL), and 1 µL HiFi Taq DNA Ligase (M0647SVIAL) using a
176 MasterCycler X50s. Ligations ran for 5 minutes at 65°C, followed by 25 cycles of 1 minute

177 at 65°C and 3 minutes at 45°C with 1°C/s ramp rate. EcoRV-HF (20 U) was added and
178 incubated at 37°C for 2 hours.

179 MyOne C1 streptavidin beads (16 µL) were washed and resuspended in 816 µL 2x
180 B&W buffer [10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 2 M NaCl]. Digestion reactions were
181 pooled, combined with beads, and rotated at room temperature for 45 minutes. Beads were
182 collected on a magnet, the supernatant discarded, and the beads washed twice with 300 µL
183 2x B&W, once with 300 µL Elution Buffer [10 mM Tris-HCl, pH 8.5], followed by
184 resuspension in 72 µL 1x rCutSmart Buffer.

185 PvuI-HF (3 µL, 60 U, NEB R3150L) was added and incubated at 37°C for 90 minutes
186 with rotation. The reaction was heated to 65°C for 20 minutes to release digested DNA, which
187 was retained after pelleting the beads. Library preparation was completed using the ONT
188 LSK114 kit (GDE_9161_v114_revX_13Dec2024) following the manufacturer's protocol.

189 *Basecalling and TARPON Pipeline Execution*

190 Pod5 files were basecalled using Dorado v0.7.0 with Fast, High Accuracy (HAC), and
191 Super Accuracy (SUP) v5.0.0 models on NVIDIA L4 GPUs. To reduce compute time,
192 telomeric reads were first identified in fast basecalled data using TARPON and re-basecalled
193 with SUP if `--pod5_directory` and `--fast_basecalled` flags were specified. TARPON is
194 implemented in Nextflow v24.04.3 with Docker containerization. Input data may be pre-
195 basecalled with fast or SUP basecalling models; no preprocessing is required. TARPON
196 detects telomeric reads based on the presence of a user-defined telomeric repeat (default:
197 GGTTAG).

198 Strand-specific filtering is applied via the `--c_strand_only` flag for splint-enriched
199 libraries. Filtering thresholds, barcode demultiplexing, capture probe sequence, and GUI/CLI
200 options are detailed in the GitHub README. The instructions on how to run this pipeline

from the command line or the EPI2ME GUI and the pipeline's source code are available at https://github.com/baumannlab/TARPON.

Results

Overview and Use of TARPON

TARPON addresses the computational challenges associated with telomere sequencing using Oxford Nanopore long-read technologies by providing an integrated analysis pipeline suitable for bioinformaticians, researchers, and clinicians. The pipeline requires no preprocessing of the data and accepts a range of input formats, including FASTQ, compressed FASTQ, and BAM files. It supports data that has been basecalled using any of ONT's fast basecalling models as well as the super-high accuracy (SUP) model, simplifying usage and eliminating the need for manual data manipulation prior to analysis.

Once provided with input files, TARPON identifies putative telomeric reads, assigns strand specificity, detects the subtelomere-to-telomere boundary, and applies several filtering steps. These include the removal of reads lacking a terminal capture probe, reads dominated by telomeric repeats at the 5' end, reads with extended regions of erroneous repeat calls, and reads in which the telomere start site is misidentified. The pipeline generates multiple files summarizing each processing step, including telomere read counts, read-level length and quality statistics, as well as bulk telomere length distributions, returned in both tabular and graphical formats. If the sequencing run is multiplexed, TARPON automatically separates all statistics and plots in a sample-specific manner. In addition to the generation of raw files (PDF, PNG, and TXT), TARPON compiles all relevant information and figures into an HTML report (Fig. 1a).

For the more experienced user, TARPON can be cloned directly from GitHub and run from the command line after installation of Java, Nextflow, and Docker. No installation of

225 additional software dependencies is required, as TARPON utilizes Docker containers through
226 Nextflow, inherently avoiding version incompatibility issues.

227 For users less familiar with the command line, TARPON is integrated into Oxford
228 Nanopore Technologies (ONT) EPI2ME agent. This platform assists with the installation of
229 Java, Docker, and Nextflow on a local machine, enabling GUI-based operation of the
230 pipeline. Integration into EPI2ME is achieved by entering the GitHub repository URL into
231 the EPI2ME "Add Workflows" utility (Fig. 1b). Once loaded, the user can specify pipeline
232 inputs and adjust parameters either through the GUI (Fig. 1c) or the command line.
233 Configurable options include the capture probe sequence, barcode files for sample
234 demultiplexing, and parameter settings for telomeric read isolation, capture probe and
235 barcode detection, subtelomere-to-telomere boundary identification, and the stringency of
236 each filtering step. Additional guidance on all configurable parameters is available in the
237 README file provided with the pipeline.

238 *First Pass Telomeric Read Isolation*

239 Regardless of which telomere enrichment technique is used, the low abundance of
240 telomeric DNA relative to bulk genomic DNA results in a high proportion of non-telomeric
241 reads in the sequencing output. Performing telomere-specific functions, such as identifying
242 the subtelomere-to-telomere boundary, on the full dataset would unnecessarily increase
243 computational demands and analysis time. TARPON addresses this by first isolating putative
244 telomere-containing reads based on the presence of a user-defined telomeric repeat motif,
245 which defaults to the canonical vertebrate repeat GGTTAG.

246 To establish an efficient and consistent strategy for first-pass telomeric read isolation,
247 we tested multiple parameter combinations across two splint-enriched sequencing runs
248 (HG002-SE and WB60-SE) and two duplex-enriched runs (HEK-DE and WB60-DE). Raw
249 pod5 files, which store the original electrical signal data generated by the nanopore sensor

250 and serve as the source for all downstream read information, were basecalled using dorado
251 v0.7.0 with fast, high accuracy (HAC), and super high accuracy (SUP) models. Our goal was
252 to identify parameters that would yield the same subset of telomeric reads regardless of
253 basecalling model, while minimizing computational overhead even when input file size
254 increased up to twenty-fold.

255 Among the tested strategies, requiring a read to contain at least ten non-consecutive
256 instances of the telomeric repeat (default: GGTTAG) consistently identified the greatest
257 number of reads across all runs. Importantly, this criterion was robust as it produced near-
258 identical read sets across all three basecalling models (Fig. 2a, left). In contrast, increasing
259 the threshold to 20 repeats or requiring repeats to be consecutive led to decreased sensitivity
260 in fast and HAC data relative to SUP, whether detected using custom Python scripts or seqkit
261 grep. These results demonstrate that the 10 non-consecutive repeat threshold provides a
262 basecalling model–agnostic balance between sensitivity and consistency.

263 Runtime performance for read isolation was also evaluated. Although basecalling
264 model had minimal impact on isolation time, the choice of detection method did: seqkit
265 grep was slightly slower than custom Python scripts. Additionally, identifying non-
266 consecutive repeats took longer than consecutive ones. However, input file size had the
267 greatest influence on runtime. For instance, HEK-DE (2.9 million reads) required
268 substantially more processing time than WB60-DE (1.8 million reads), while both HG002-
269 SE and WB60-SE had under one million reads. Across all strategies, processing time
270 increased linearly with input file size (Fig. 2a, lower right). At 20X input size, seqkit
271 grep became significantly slower than custom scripts, likely due to the higher proportion of
272 small, non-telomeric reads in WB60-SE, HEK-DE, and WB60-DE.

273 To further confirm that reads identified using the ten non-consecutive telomeric repeat
274 criterion originated from the same raw signal data, we compared read IDs across basecalling

models for each of the sequencing runs. This analysis demonstrated that the same subset of reads was isolated, regardless of whether fast, HAC, or SUP basecalling was used (Fig. 2b). Because the telomeric reads identified in fast and SUP basecalled datasets are identical, TARPON allows users to isolate candidate telomeric reads from fast basecalled data and then selectively re-basecall only these reads using the super high accuracy model. This approach can result in substantial savings in computational resources and user time, since SUP basecalling is performed on only ~1–2% of the dataset (Fig. 2c). To enable this functionality, the user must specify the location of the pod5 files using the --pod5_directory flag and indicate that the data were initially basecalled using a fast model by including the --fast_basecalled parameter.

Strand Orientation of Isolated Telomeric Sequences

Depending on the pre-sequencing enrichment technique, users may expect to sequence either telomeric C strands alone (e.g., HG002-SE and WB60-SE, Fig. 2d) or a combination of both C and G strands (e.g., HEK-DE and WB60-DE, Fig. 2d). For C-strand specific splint-enriched sequencing, any read with more than 20% G-strand telomeric repeats is removed from the analysis. In these cases, the --c_strand_only parameter should be set to ensure proper filtering. For duplex-enriched sequencing, where both strand orientations are expected, reads with mixed C and G strand identity, defined as containing between 20% and 80% G-strand repeats, are excluded. These are likely chimeric reads or sequences containing only subtelomeric regions that met the ten-repeat threshold.

Telomeric repeats within the subtelomeric portion of a read do not affect this filtering step due to their relatively low abundance compared to the telomeric region. Reads classified as C-strand telomeric reads (<20% G-strand content) are reverse complemented into G-strand orientation for downstream analysis. The original strand identity is retained as a tag in the resulting BAM file to permit future discrimination. This initial filtering step removes

300 approximately 1–5% of reads, independent of enrichment method (Fig. 2e). Only reads
301 passing this strand-filtering step are used in subsequent analysis.

302 *Identification of the Telomeric Capture Probe*

303 The strategy of capturing telomeres via ligation of oligonucleotides or partial duplexes
304 to the 5′ end of the C-strand or 3′ end of the G-strand has its roots in earlier ligation-based
305 assays such as the amplification and sequencing of single telomeres (33–35). Building on this
306 foundation, more recent sequencing-based protocols have refined the technique by integrating
307 capture probes into high-throughput library preparation workflows (27–29).

308 While the primary purpose of the capture probe is to facilitate enrichment of telomeric
309 DNA fragments, it also serves as a terminal tag that can be identified computationally. This
310 enables verification that a full-length telomere is present within a read and helps exclude
311 fragments that were truncated during library preparation or sequencing. Furthermore, capture
312 probe identification prevents inclusion of extraneous sequences—such as ONT adapter
313 elements or ligation artifacts—in telomere length measurements.

314 Due to the relatively high error rate of single-pass nanopore sequencing, exact
315 matching of the 12-nucleotide capture probe fails to detect the intended sequence in 20–50%
316 of reads (Fig. 3a, pink bars). To improve sensitivity, we tested the effect of permitting a
317 limited number of mismatches. Allowing six errors led to detection of seven or more putative
318 probes per read, an unrealistic outcome that reflects extensive off-target matching (Fig. 3a,
319 light green; Fig. 3b). Allowing two mismatches resulted in the highest proportion of reads
320 with a single identifiable capture probe across all sequencing runs (Fig. 3a, yellow bars),
321 consistent with biological expectations. Increasing the mismatch threshold to three resulted
322 in additional off-target detections (Fig. 3a) and a notable increase in runtime (Fig. 3b).

323 Greater than 75% of splint-enriched telomeric reads contained a capture probe when
324 two mismatches were allowed within the 12-nucleotide probe sequence. In contrast, only

325 ~50% of duplex-enriched reads contained a detectable probe under the same criteria. This
326 reduction is likely due to two factors. First, during duplex-enrichment, only one strand of
327 each DNA fragment needs to carry a capture probe for successful streptavidin pulldown,
328 potentially resulting in half of the resulting telomeric reads lacking a probe. Second, only the
329 reverse complement of the GGTTAG permutation was used as a capture probe, which anneals
330 in register with the 5′ end of the C-strand at the double- to single-strand junction (35). In
331 contrast, the G-strand overhang is more heterogeneous in terminal sequence, and the use of
332 all six possible telomeric repeat permutations may increase probe ligation to G-strand ends.
333 This adjustment may be particularly important when capturing G-strands from telomerase-
334 negative cells where bias for a specific 3' end permutation is minimal (35).

335 In duplex-enriched datasets such as HEK-DE and WB60-DE, approximately 30% of
336 G-strand telomeres contain a capture probe, compared to 75% of C-strand telomeres
337 (Supplemental Fig. 1). Additionally, capture probe detection is improved by SUP basecalling
338 basecalling (Fig. 3c). For this reason, we strongly recommend either running TARPON on
339 fast basecalled data with the pod5 directory specified for selective re-basecalling or using pre-
340 basecalled SUP data as input.

341 Increasing the number of allowed mismatches when identifying the capture probe also
342 increases the number of off-target sequences detected. While the majority of probe matches
343 are located within the final 200 bp of the read (or the first 200 bp in C-strand reads), a small
344 number of matches appear between 4 kb and 5 kb from the end of the read (Fig. 3d). These
345 likely reflect subtelomeric sequences with similarity to the capture probe rather than true
346 probe ligation sites, and represent computational artifacts introduced by relaxed stringency.
347 To mitigate this, an additional filter was introduced to ensure that the capture probe is only
348 accepted if it is the first match found after the identification of twenty telomeric repeats. This
349 positional constraint improves specificity by eliminating internal subtelomeric matches. In

splint-enriched datasets, the number of capture probes identified slightly exceeds the number of reads due to abnormal ligation products in which multiple capture probes are present in tandem. This phenomenon is not observed in duplex-enriched libraries, where a substantial fraction of telomeric reads lack a capture probe entirely. Nonetheless, a small number of duplex reads do contain two adapter sequences, as seen in Fig. 3a.

When using capture probe sequences longer than 12 nucleotides, a higher mismatch allowance is recommended to maintain sensitivity. As probe length increases from 12 to 24 nucleotides, retaining only two allowed mismatches leads to a drop in detection efficiency, with fewer reads identified as containing a capture probe (Fig. 3e). However, this effect is mitigated by increasing the mismatch threshold. For example, allowing three or four mismatches in 24-nucleotide probes restores capture probe detection to expected levels (Fig. 3e). Importantly, increasing both the length of the capture probe and the number of tolerated errors does not result in additional off-target matches within telomeric or subtelomeric regions (Fig. 3f), indicating that probe specificity is preserved under these conditions.

Subtelomere-to-Telomere Boundary Identification

The subtelomere-to-telomere boundary in humans has remained poorly characterized, largely due to the repetitive nature of this region and the historical lack of sequencing technologies capable of resolving it at high resolution. Nanopore sequencing now offers the opportunity to define this boundary on a single-read basis and across chromosome arms. Despite this potential, there is currently no consensus in the field on how to delineate the subtelomere-to-telomere transition. Three recent studies have used distinct methodologies to estimate telomere length from long-read data, complicating cross-study comparisons (27–29).

To develop a consistent and biologically informed algorithm for telomere start detection, we first visualized telomeric sequences across thousands of reads. We observed

375 that the frequency of canonical telomeric repeats (GGTTAG) alone did not consistently
376 define a clear boundary when measured in 100 bp sliding windows (Fig. 4a, blue lines). In
377 contrast, a strong and persistent increase in one-nucleotide variants of GGTTAG emerged in
378 many reads (Fig. 4a, orange lines). Once these variant-containing windows surpassed a
379 certain threshold, they typically maintained >50% signal density across the remainder of the
380 telomere. We refer to this combined pattern of canonical repeats and single-nucleotide
381 variants as "telomere+1N". The region enriched for the variants is referred to as the variant
382 repeat–rich (VRR) region and varied in length from a few hundred base pairs (Fig. 4a, left)
383 to several kilobases (Fig. 4a, middle and right) before transitioning into invariant GGTTAG
384 repeats. Considering a potential functional relevance of this telomere repeat containing
385 region, we included it in the telomere length output. This decision is supported by
386 observations of reads that are essentially devoid of a terminal stretch of homogeneously
387 wildtype repeats yet terminate in a capture probe (Supplemental Fig. 2). Accordingly,
388 TARPON calculates the telomere length from the start of the VRR-region to the distal end of
389 the telomere, as defined by the position of the capture probe.

390 The telomere length metric requires the ability to accurately detect the start of the
391 VRR-region and the associated frequency spike of telomere+1N repeats. However, while
392 plotting the percentage of telomere+1N repeats within the first 300 bp of the read, a notable
393 bimodal distribution emerged: most reads contained less than 15% telomere+1N repeats,
394 while a smaller subset exceeded 80% (Fig. 4b). This latter group represents telomere-
395 containing sequences that begin within the VRR-region or even within homogenous
396 GGTTAG repeat sequence. These reads were excluded from further analysis as proximally
397 truncated, since the full length of the VRR-region or telomere sequence cannot be determined
398 with confidence (for C-strand reads, this uncertainty applies to the read end since all reads
399 have been reverse complemented into the G-strand orientation). To avoid underestimating

400 telomere length distributions by including these truncated telomeric reads, it was necessary
401 to exclude 15 to 50% of sequences from further analysis (Fig. 4c). This percentage was at the
402 higher end of the range in duplex-enriched datasets, likely due to an abundance of short reads
403 corresponding to partial C-strand sequences (Supplemental Fig. 3), which may have arisen
404 during library preparation and fill-in synthesis of the G-strand overhang.

405 To validate candidate algorithms for identifying the start of the VRR-region, 100 reads
406 from each sample were selected to represent a variety of subtelomere-to-telomere boundary
407 patterns. These reads were manually annotated to create a truth set. The accuracy of each
408 algorithm was evaluated by calculating the mean absolute error between the predicted and
409 manually annotated VRR-region start positions. Lower values indicate higher accuracy.

410 The first method tested involved identifying the first occurrence of eight consecutive
411 telomere+1N repeats. This approach yielded a mean absolute error of 149 nucleotides and
412 was ultimately unsuccessful (Fig. 4d). A second method used a 100 bp jumping window,
413 applied in 10 bp increments along the read. When the frequency of telomere+1N repeats in a
414 window surpassed a predefined threshold, the first telomere+1N repeat within that window
415 was marked as the VRR-region start. To prevent misidentification due to telomere-like
416 islands in the subtelomeric region (Supplemental Fig. 4), this threshold was required to be
417 maintained throughout the remainder of the read. Using a 60% threshold, this method reduced
418 the mean absolute error to 118 nucleotides. Further refinement involved a dual-threshold
419 strategy: once the primary 60% frequency threshold was reached, the telomere+1N repeat
420 content could not fall below a secondary threshold of 5% for the remainder of the telomere.
421 This approach dramatically improved performance, achieving a mean absolute error of 4.36
422 nucleotides (Fig. 4f).

423 We also tested the effect of window size and step size. Larger window sizes decreased
424 accuracy at low secondary thresholds (Fig. 4g), while increasing the jump interval from 10

425 bp to 15 bp slightly improved precision by ~0.3 nucleotides (Fig. 4h). Most VRR-region start
426 sites were predicted with near single-nucleotide accuracy. For those that were inaccurate, the
427 error typically resulted in slight overestimation of telomere length by 10 to 50 nucleotides
428 (Fig. 4i). Using a 100 bp window, a 15 bp increment, and the dual-threshold criteria, the
429 algorithm successfully identified the VRR-region start in over 80% of telomeric reads (Fig.
430 4j).

431 *Filtering for High Confidence Telomeric Sequences*

432 To ensure that telomeric reads were accurately identified, we examined the proportion
433 of telomere+1N repeats between the VRR-region start and the capture probe (Fig. 5a). On
434 average, this region contained greater than 80% telomere+1N content. Reads with less than
435 80% fell into two main categories: (1) sequences in which a telomere start was identified, but
436 the read originated from subtelomeric or interstitial telomeric regions (Fig. 5b); or (2) true
437 telomeric reads that included atypical subtelomeric structures, unusually long VRR-regions,
438 and relatively short stretches of invariant telomeric repeats (Fig. 5c). To exclude ambiguous
439 or misidentified reads while retaining biologically valid telomeric sequences, we tightened
440 the threshold, removing all reads with less than 60% telomere+1N content. This removed
441 cases like Fig. 5b while retaining complex, valid reads like those shown in Fig. 5c. Reads
442 discarded at this step tended to have slightly higher basecalling quality scores (Fig. 5d) but
443 were much shorter in both total read length (Fig. 5e) and telomere length (Fig. 5f) compared
444 to retained reads. These patterns are consistent with a subtelomeric origin, likely representing
445 sequences that failed to extend fully into the VRR- or telomeric region during sequencing.
446 This filtering step excluded fewer than 1% of telomeric reads overall (Fig. 5g).

447 A second validation step was performed to confirm the accurate identification of the
448 VRR-region start by examining the 2 kb of sequence immediately upstream (in the
449 subtelomeric direction) of the predicted boundary. This region was analyzed to ensure it was

450 not composed primarily of telomere+1N repeats (Fig. 5h). Elevated telomere+1N signal in
451 this region can arise either from genuine subtelomeric structure, consistent across reads (Fig.
452 5i), or from basecalling artifacts (Fig. 5j). Reads resembling Fig. 5j, which exhibit a sharp
453 transition from fully telomeric sequence to 0% telomere+1N content, are attributed to
454 sequencing artifacts introduced by dorado v0.7.0. These segments typically show reduced
455 quality scores relative to the rest of the telomeric sequence (Supplemental Fig. 5) and lack
456 consistent VRR-region repeat patterns.

457 To address these cases, the VRR-region detection algorithm was refined while
458 retaining the core thresholds: 60% telomere+1N content in a 100 bp jumping window, with
459 no drop below 5% for the remainder of the telomere. An additional requirement was
460 introduced: the repeat content must remain below the 5% threshold for at least 15 consecutive
461 windows before concluding that no VRR-region start site is present. This refinement ensures
462 accurate detection in cases like Fig. 5i while filtering out artifacts such as Fig. 5j. Reads
463 with >10% telomere+1N content in the 2 kb upstream of the predicted start site were
464 excluded, resulting in the removal of approximately 1–5% of telomeric sequences (Fig. 5k).

465

466 *Read Assignment to Specific Chromosome Arms*

467 Chromosome arm-specific telomere length differences have been reported in *S.*
468 *cerevisiae* (36,37), and underlying regulatory mechanisms have recently been identified .
469 These findings have sparked considerable interest in applying chromosome arm–level
470 telomere analysis in other model organisms including humans. Accurate assignment of each
471 telomeric read to a unique chromosome arm is therefore an important goal.

472 Aligning telomeric reads from HG002 cells to a high-quality, HG002-specific
473 telomere-to-telomere (T2T) reference assembly enables assignment of over 90% of reads to
474 individual chromosome arms (Fig. 6a, HG002-SE, solid red line). However, aligning the

475 same dataset to other subtelomeric references or general-purpose T2T assemblies (e.g.,
476 STONG or CHM13) yields substantially lower performance: only ~60% of reads map
477 uniquely, many are multimapping with low quality scores, and the proportion of unmapped
478 reads increases as the minimum mapping quality (MAPQ) threshold is raised. This issue is
479 illustrated here when aligning telomeric reads from non-HG002 samples—such as the clinical
480 sample WB60 (SE and DE libraries) —to the HG002, CHM13, or STONG references. In
481 these cases, only ~50% of telomeric reads map uniquely with a MAPQ ≥ 20, indicative of
482 poor alignment specificity (Fig. 5a, WB60-SE). Data derived from the HEK293T cell line
483 was not included for analysis due to its hypotriploid status.

484 Furthermore, since neither enrichment technique has an intrinsic bias against specific
485 chromosome arms, an even distribution of telomeric reads across all 92 chromosome arms
486 should be expected. This was indeed observed when aligning HG002-derived telomeric reads
487 to the HG002-specific reference: approximately 1.1% of telomeric sequences aligned to each
488 arm, with no arm exceeding 2% of the total. The most enriched arm contained only 2.1 times
489 more reads than the 25th percentile arm, reflecting a near-uniform distribution. To avoid
490 introducing bias from low-level mapping artifacts, normalization does not rely on the
491 chromosome arm with the fewest mapped reads.

492 In contrast, alignments of the same HG002 dataset to other reference genomes (e.g.,
493 CHM13 or CHM13+STONG) introduced significant bias. In these cases, the most enriched
494 chromosome arm contained 11.7-fold and 64.6-fold more telomeric reads, respectively, than
495 arms in the lowest quartile (Fig. 6b). The effect was also seen in clinical samples aligned to
496 the same references, where maximum fold differences ranged from 17.7x to 70x (Fig. 6b).
497 These disparities were accompanied by higher mismatch and gap counts when aligning
498 clinical datasets to non-matching references, as compared to HG002 aligned to its own
499 reference (Fig. 6c). These results clearly demonstrate that aligning telomeric sequences to a

500 reference genome derived from a different individual or cell line does not provide reliable
501 chromosome arm–specific telomere measurements.

502 To test whether de novo clustering techniques can provide chromosome arm-specific
503 telomere length information, telomeric sequences derived from HG002 sequencing data were
504 clustered de novo and then compared to the alignment of the same sequences to the HG002
505 subtelomeric reference. Parameters optimized with HG002 were then applied to other datasets
506 for validation. Telogator2 (31,38,39), a pipeline designed to derive chromosome arm-specific
507 telomeric length from long reads, can be used to cluster telomeric reads without alignment to
508 a reference using pairwise alignment and hierarchical clustering. Perfect clustering would
509 result in the formation of 92 clusters with 100% of the input telomeric sequences divided
510 evenly across the clusters (~1.08% of the sequences per cluster). When full length telomeric
511 sequences were used as input for Telogator2, only 55 clusters were generated (Fig. 6d) with
512 ~40% of the telomeres clustering together (Supplemental Fig. 6a). A cluster is defined when
513 containing greater than 0.02% of the input telomeric sequences.

514 To focus clustering primarily on the VRR-region, telomeric sequences were trimmed
515 to the sequence found immediately before or after the subtelomere-to-telomere boundary.
516 When clustered under the recommended parameters similar trends to full length reads are
517 observed for reads with 3 kb or 4 kb of subtelomeric sequence prior to the subtelomere-to-
518 telomere boundary regardless of the distance of telomeric sequence used after the VRR-
519 region start (Fig. 6d). However, removing all subtelomeric sequence or trimming the
520 subtelomere to the 1 or 2 kb immediately before the VRR-region start site (subtelomere)
521 notably increased the number of identified clusters (Fig. 6d). With 0 - 2 kb of subtelomeric
522 sequence, increasing the distance after the subtelomere-to-telomere boundary to 3 or 4 kb
523 increases the percentage of telomeric sequences (>95%) found within clusters (Fig. 6e). More
524 importance is therefore placed on the amount of subtelomeric sequence used in the analysis

525 than the amount of telomeric sequence; however, trimming the telomere sequences is
526 beneficial as it decreases computational time.

527 When telomeres are trimmed 1 kb before and 3 kb after the subtelomere-to-telomere
528 boundary and clustered, the telomeres are distributed equally across the chromosome arms,
529 directly comparable to the distribution seen in alignments (Fig. 6f). Furthermore, the read IDs
530 belonging to each chromosome arm by alignment are often uniquely found in a single cluster
531 (Fig. 6g) demonstrating the accuracy of the clustering methodology, which can additionally
532 be visualized by plotting the individual reads in each cluster and comparing the VRR-region
533 structures (Supplemental Fig. 7).

534 When the parameters established for HG002 were tested on clinical samples the
535 results varied with the number of telomeric sequences used. For B2_duplex and B2_simplex
536 only 2,034 and 3,170 telomeric were available, respectively. Here we see a slightly higher
537 cluster count than expected (Supplemental Fig. 6b) and a slightly decreased percentage of
538 telomeric sequences included in clusters (~90%) (Supplemental Fig. 6c) indicating that the
539 clustering parameters may need to be fine-tuned in a sample-specific context. Nevertheless,
540 the results from the clustering approach were superior compared to the alignment to a non-
541 sample specific reference. Importantly, when the number of telomeric reads was increased to
542 6,266, approximately 92 clusters were obtained, and more sequences were assigned to
543 clusters. We therefore recommend a minimum of 5,000 telomeric reads per sample when
544 performing de novo-based chromosome arm-specific telomere length analysis. It is further
545 recommended that at least 1 kb of subtelomeric sequence is included in the analysis to avoid
546 formation of large clusters due to high sequence similarity within the VRR-region
547 (Supplemental Fig. 6d).

548 The computational time required for Telogator2 is substantially larger than
549 performing a simple alignment and performance of the software varies with parallelization;

550 best results were identified when using a minimum of 10 threads. It is also important to note
551 that while the phrase "Chromosome Arm-Specific Telomere Lengths" is used in this method,
552 a more appropriate terminology would be "telomere allele specific telomere length" as the de
553 novo methodology would not allow for the determination of which chromosome arm the
554 cluster is derived from without a high quality subtelomere reference genome.
555
556 *Extended Features of TARPON*
557 Beyond its default configuration, TARPON includes a range of optional parameters
558 for more specialized analysis of telomeric sequences and enrichment protocols. For duplex-
559 enriched datasets, the --strand_comparison flag provides strand-specific enrichment metrics,
560 including relative abundance, filtering outcomes, and telomere length distributions for C- and
561 G-strand reads. The --detailed_stats option outputs additional read-level statistics and
562 visualizations, such as repeat composition and telomere length-to-quality score comparisons,
563 in text, graphical, and HTML formats.
564 For users exploring telomeric enrichment strategies, the --restriction_digest flag accepts
565 a comma-separated list of restriction enzyme recognition sites and returns the number of
566 telomeric sequences affected by each restriction site. In samples containing mutant telomeric
567 repeats such as in the case of a telomerase RNA template mutation (40), the --mutant flag
568 enables these sequences to be included in all filtering and boundary-identification steps, while
569 also returning statistics related to mutant versus wildtype telomerase processivity.
570 To reduce computational overhead, TARPON also supports a hybrid basecalling
571 strategy. When --pod5_directory is specified, telomeric reads are first isolated from fast
572 basecalled data and then re-basecalled using SUP models. Input file formats do not need to
573 be modified in advance: FASTQ and BAM files are both accepted, and TARPON will convert
574 files to UBAM format internally as needed. Integration with Nextflow and the use of Docker

575 containers or conda environments helps prevent versioning issues and dependency conflicts
576 during execution.
577
578 **Discussion**
579 TARPON is a flexible and modular pipeline designed to analyze telomeric sequences
580 from Oxford Nanopore Technologies (ONT) long-read sequencing data. It performs a full
581 analysis workflow from fast or SUP basecalled data, including telomeric read isolation,
582 capture probe and barcode identification, and subtelomere-to-telomere boundary detection.
583 The output includes quality metrics and telomere length statistics in graphical and tabular
584 form.
585 While default parameters are optimized for telomerase-positive human samples, all
586 settings can be customized to accommodate different organisms, enrichment strategies, or
587 specific experimental goals. This flexibility allows TARPON to support a wide range of
588 research contexts, including species with noncanonical telomere repeats, mutant telomerase
589 variants, and different enrichment chemistries. Future releases of TARPON will support
590 additional features such as strand bias analysis and telomeric variant detection.
591
592 *Comparison With Other Telomere Analysis Software*
593 While several software tools are available for analyzing telomere content from
594 Illumina whole-genome sequencing (WGS) data, only TeloBP (29), Telometer (27), and wf-
595 teloseq have been designed specifically for nanopore-based telomere sequencing with the
596 latter being released by ONT while this manuscript was in preparation. TeloBP and Telometer
597 both require experience with command-line interfaces, manual installation of dependencies,
598 and troubleshooting version conflicts. Neither TeloBP nor Telometer provides a complete
599 analysis pipeline: users must first isolate and demultiplex telomeric reads, remove low-quality

600 or chimeric sequences, and ensure the presence of a capture probe prior to estimating telomere
601 length or identifying the subtelomere-to-telomere boundary.

602 In Telometer, reads must be aligned to a subtelomeric reference genome, and any read
603 that fails to map within the first or last 30 kb of the reference is discarded. TeloBP requires
604 input in FASTQ or compressed FASTQ format and filters out chimeric reads before
605 proceeding. Both tools perform limited quality filtering: Telometer excludes reads with an
606 average quality score below 9 and TeloBP relies on external preprocessing.

607 In contrast, TARPON and wf-teloseq are accessible both via command-line and
608 through a graphical user interface (GUI) (ONT's EPI2ME) for users less familiar with
609 scripting. Nextflow's integration with Docker containers eliminates dependency issues
610 entirely. TARPON accepts fast or SUP basecalled output directly from the sequencer,
611 requiring no data preprocessing. Reads must be demultiplexed prior to executing wf-teloseq
612 while TARPON will internally handle all sample demultiplexing. Filtering steps prior to
613 subtelomere-to-telomere boundary identification also differ between TARPON and wf-
614 teloseq. While both identify a capture probe at the end of the telomere, wf-teloseq does so
615 indirectly by requiring all input reads to be demultiplexed. TARPON ensures the entire
616 telomere has been sequenced and the read is not chimeric prior to telomere boundary
617 identification, while wf-teloseq filters incompletely sequenced telomeres after boundary
618 identification. Additionally, wf-teloseq requires a minimum read length of 120 bp and at least
619 100 telomeric repeats: this parameter is not customizable and effectively requires a telomere
620 to be composed of at least 600 canonical telomeric nucleotides, potentially eliminating a
621 subgroup of telomeres with high biological relevance.

622 While TeloBP, Telometer, and TARPON all use a jumping window approach to
623 identify the subtelomere-to-telomere boundary, only TARPON allows users to adjust relevant
624 parameters without modifying source code. wf-teloseq uses a sliding window approach

through the linear convolution of two arrays of different sizes but does not allow for any user flexibility. Customization in TeloBP and wf-teloseq requires editing the Python files within the cloned GitHub repository. For Telometer, users must first locate the pip package installation and modify Python scripts within those directories. By contrast, TARPON supports parameter customization either through standard Nextflow command-line syntax or via the EPI2ME GUI prior to workflow execution, providing more accessible and flexible control over the boundary detection process.

The details of how each tool defines and filters telomeric regions also vary. Telometer uses a 120 bp sliding window that advances in 12 bp increments to identify telomeric regions. If GGTTAG repeats make up more than 10% of a window, the start of a telomeric region is defined. If a subsequent window falls below the 10% threshold, a gap is introduced. If this gap exceeds 100 nucleotides and the frequency of telomere+1N repeats within the gap is also less than 10%, the telomeric region is terminated. A new telomeric region may then be defined further along the read. After read processing, Telometer applies several filtering steps. If no telomeric regions were identified, or if no window within a telomeric region contains more than 75% GGTTAG, the read is excluded from further analysis. If the average quality score of the gap between two telomeric regions is ≤9, any telomeric region following that gap is discarded. Telomeric regions are also excluded if they do not begin or end within 100 bp of a read boundary. Telomere length is then calculated from the first telomeric region that passes all filtering criteria. Finally, if the estimated telomere length plus 50 nucleotides exceeds the total read length, the read is omitted from further analysis.

TeloBP uses a similar but more complex approach to identify the subtelomere-to-telomere boundary. When executed with default parameters, a 100 bp jumping window with 6 bp intervals scans the read in the telomere-to-subtelomere direction for the presence of "GGG" motifs. For each window, the deviation from the expected composition (50%) is

650 calculated as: (observed – expected) / expected × 100. This value is then plotted, and the area
651 under the curve (AUC) is computed using a series of 83 sliding windows. If the AUC in a
652 given window is less than –50 (suggesting the region is not telomeric) and the next window
653 contains fewer or equal telomeric repeats, this defines a new threshold from which further
654 analysis begins. From this threshold, the differences in AUC values between adjacent sliding
655 windows are calculated. If the absolute difference between two AUC values is less than 0.2,
656 or if the current difference is less than 0.2 and the next difference exceeds 0.2, the boundary
657 between the telomere and subtelomere is defined. Telomere length is then estimated as the
658 distance from this boundary to the end of the read.

659 TARPON takes a different approach, using a threshold-based method to identify the
660 transition from subtelomeric to telomeric sequence. It incorporates both telomere+1N repeat
661 density and local sequence context, minimizing false positives arising from repetitive
662 subtelomeric regions and reducing overestimation due to terminal artifacts. Notably, wf-
663 teloseq first converts the telomeric sequence into a binary array: 1 for telomeric sequence or
664 telomeric variants and 0 for non-telomeric sequence. wf-teloseq does not refer to variants as
665 telomere+1N repeats, but as basecalling variants i.e. CACCCT, ACCCCT, CCCAAA,
666 CCCCGA, etc. and greatly reduces the length of the VRR-region. After binary sequence
667 conversion, the resultant array is smoothed by scipy.ndimage.median_filter. If there is a high
668 enough density of wild type telomeric repeats within the variant repeat rich region, the
669 smoothing may allow for the capture of the VRR-region. However, this smoothing occurs in
670 a sequence specific manner and would affect each chromosome arm differently creating a
671 systematic bias for chromosome arm-specific telomere length analysis. The smoothed binary
672 array is then compared to a mock telomere boundary (an array of 61) composed of 30
673 nucleotides of telomeric sequences (1 after binary conversion), a 0 value, and then 30
674 nucleotides of -1 value. These two arrays are compared using np.convolve which returns the

675 arrays discrete, linear convolution. The telomeric boundary is identified as the last occurrence
676 of the minimum value found within the newly calculated convolution. As wf-teloseq is
677 designed to operate on C-strand telomeric sequences only, the final occurrence within a read
678 would be centromere proximal. Since wf-teloseq utilizes a different approach than TARPON
679 to identify the subtelomere-to-telomere boundary, the 400 manually curated telomeric
680 sequences were used to calculate the mean absolute value error of wf-teloseq as 157
681 nucleotides, compared to the 4.36 nucleotide mean absolute value error of TARPON. This
682 large error was primarily driven by the identification of telomere islands within the
683 subtelomeric sequences (Supplemental figure 7b). While these wild type islands may support
684 telomere stability and exhibit shelterin binding, the intervening subtelomeric sequence most
685 likely does not and would result in a non-biological chromosome-arm specific length bias.
686 After removal of these erroneous called boundaries, the mean absolute value error decreased
687 to 42 nucleotides.

688 After boundary detection, TARPON and wf-teloseq apply additional filtering steps
689 not found in TeloBP or Telometer. In TARPON, the telomeric region must contain more than
690 60% telomere+1N repeats, and the 2 kb region upstream of the boundary must contain less
691 than 10% telomere+1N. These filters eliminate reads with internal sequencing artifacts that
692 arise through erroneous basecalling and exhibit reduced quality scores, such as the example
693 in Fig. 5j. In this case, TeloBP and Telometer assign telomere lengths of 18 bp and 342 bp,
694 respectively. In reality, the read contains over 3 kb of telomeric sequence, but since it is
695 unclear if the erroneous stretch of basecalls is similar in length to the bona fide repeat
696 sequence, the read is excluded from analysis in TARPON. wf-teloseq first ensures the
697 telomere boundary identified is greater than 61 nucleotides away from the start of the read
698 and 30 nucleotides away from the end of the read. While the README of wf-teloseq states
699 the distance as 60 nucleotides from the end of the read, within the source code this parameter

700 is divided by 2 during the condition operation. The first 30% of the identified telomere (C
701 strand sequences start at the distal end of the telomere) must be composed of greater than
702 80% telomeric repeats. The subtelomeric portion of the telomeric read is then confirmed to
703 be composed of less than 25% CCC and a median sequence quality greater than 9. Lastly wf-
704 teloseq filters for known telomere basecalling artifacts and if it finds 5 of such artifacts within
705 500bp of each other the read is discarded from the analysis.

706 TARPON calculates telomere length up to the start of the capture probe and wf-
707 teloseq calculates telomere length as the distance between the capture probe, trimmed off
708 during demultiplexing, and the identified telomere boundary. In contrast, if reads are not
709 preprocessed before using TeloBP, telomere length estimates will be inflated due to the
710 inclusion of non-telomeric sequence contributed by the capture probe and / or ONT
711 sequencing adapters.

712 Telometer, TeloBP, wf-teloseq, and TARPON all return tabular output files (CSV or
713 TSV) listing read IDs and corresponding telomere lengths. TARPON includes additional
714 metadata such as the coordinates of the telomeric region within each read, strand specificity,
715 and read-level quality metrics. It also generates a suite of visual outputs summarizing pipeline
716 execution, filtering steps, and telomere length distributions. A precompiled HTML report—
717 automatically launched in the EPI2ME GUI upon completion—provides an accessible,
718 sample-by-sample overview. For multiplexed datasets, side-by-side comparisons are
719 included by default. While TARPON and wf-teloseq are both complete analysis pipelines
720 (excluding the lack of demultiplexing in wf-teloseq), they differ in key functionalities: wf-
721 teloseq is only applicable to C-strand telomeric sequences, offers no user flexibility making
722 it difficult to adapt to specific use cases such as ALT positive cell lines which may contain a
723 higher proportion of variant telomeric repeats. Additionally, the discrepancies that exist
724 between the source code of wf-teloseq and the README make it difficult for non-

725 computational users to identify how their telomeric sequences are being processed to ensure
726 no bias is being introduced.

727 *Chromosome Arm-Specific Analysis*

728 Despite their workflow differences, all previously published analysis tools rely on
729 alignment to a reference genome for chromosome arm assignment. This method works well
730 for HG002 data aligned to an HG002-specific reference but yields poor accuracy for other
731 cell lines or clinical datasets, where alignment to non-matching references leads to
732 increased mapping bias and reduced reliability in chromosome arm–specific telomere
733 length estimation. TARPON is the first pipeline to provide a solution to this dilemma.

734 While pre-existing tools to cluster telomeric sequences such as Telogator2 exist for
735 measuring allele-specific telomere length, performance was underwhelming when using full
736 length telomeric sequences. However, when focusing on the variant regions that differ
737 between chromosome arms (the variant repeat rich region) between 90-95 clusters of equal
738 size were obtained even with low input read counts. It is important to note that Telogator2
739 inherently maps clusters back to a subtelomeric specific reference; however, when executed
740 within TARPON this is not allowed and Telogator2 is terminated after cluster formation.
741 Telogator2 utilizes pairwise alignment and hierarchical clustering resulting in long run
742 times. In the future, other programs that can accurately assign de novo telomeres reads in a
743 cluster specific manner should be evaluated. Additionally, while Telogator2 serves as a
744 solid foundation for de novo clustering other software that may decrease computation
745 requirements or increase accuracy should be explored.

746

747 *Multiplexing and Cost Reduction*

748 At the time of manuscript preparation, the cost to sequence a telomere enriched sample
749 was 95€ for nanopore specific library preparation reagents (SQK-LSK114), 34€ for third

750 party reagents (NEBNext Companion Module v2, E7672S), and 570€ for a R10.4.1 MinION
751 FlowCell assuming purchase of a pack of 12 FlowCells, plus the cost of sample
752 generation/collection/enrichment. This results in a minimum total cost per sample of 700€.
753 To reduce this considerable cost per sample, it is crucial that reliable multiplexing methods
754 are developed without decreasing the number of telomeric reads per sample. The multiplexing
755 protocols described previously (27–29) allow for sample pooling after duplex or splint
756 ligation prior to library preparation. Multiplexing introduces additional computational
757 challenges, which TARPON already addresses. If a sample file is provided that contains
758 unique barcode sequences (often found within the duplex sequence as in Karimian et al.,
759 2024) together with a capture probe, the capture probe sequence is first used to identify the
760 terminal end of the telomere. The following 100bp are then used to demultiplex the samples
761 based on the barcodes found within the sample file. If no capture probe sequence is provided,
762 but a sample file is, the barcodes found within that sample file will be used to both
763 demultiplex the data and determine the end of the telomere. If only a capture probe is provided
764 without a sample file, TARPON assumes the dataset was not multiplexed.
765
766 *Applications of Nanopore Telomere Sequencing*
767 Assessing telomere length distributions at nucleotide resolution opens new avenues
768 for studying telomere dynamics in aging and senescent cell populations. Nanopore
769 sequencing enables a clearer definition of the subtelomere-to-telomere boundary and provides
770 the opportunity to investigate the functional relevance of the variant repeat-rich (VRR)
771 region. It also allows for high-resolution studies of cancer cells exhibiting recombination-
772 based telomere maintenance or elevated telomere+1N repeat content, as well as the detailed
773 characterization of the effects of telomerase template mutations on telomere dynamics.

774 The utility of nanopore-based telomere sequencing extends well beyond traditional
775 research. FlowFISH, the current clinical gold standard for telomere length diagnostics,
776 provides only a median telomere length per sample and is unavailable in many clinical
777 settings. Nanopore sequencing, by contrast, offers detailed telomere length distributions and
778 can be performed in any laboratory equipped with a MinION device, significantly reducing
779 turnaround time for clinical assessments.

780 Importantly, telomere analysis with TARPON is not limited to human samples.
781 TARPON can be used to study telomere length dynamics in any organism with non-
782 heterogeneous telomeric repeats, including many invertebrates and plants. However, we want
783 to caution that parameters for boundary detection may require optimization on a species-
784 specific basis. For non-vertebrate taxa, where telomere repeats can be highly heterogeneous,
785 such as *Schizosaccharomyces pombe* (41), TARPON is not currently recommended. In such
786 cases, the soon to be released pombeTARPON, designed to account for sequence
787 heterogeneity, will be more appropriate. Additionally, while TARPON is designed for easy
788 clinical implementation Nanopore Sequencing nor TARPON have been clinically validated
789 and should at this time not be the sole clinical diagnostic methodology.

790

791 **Conclusion**

792 Nanopore sequencing is a rapidly evolving technology that offers a unique
793 opportunity to explore telomere-related questions previously inaccessible with Sanger, short-
794 read Illumina, or PacBio sequencing platforms. While several tools exist for analyzing
795 telomeric sequences from short-read whole-genome sequencing data, few address the distinct
796 challenges posed by nanopore reads. Among those that do, most require bioinformatics
797 expertise and substantial preprocessing, and are limited to command-line interfaces.

TARPON is the first fully automated and GUI-accessible telomere analysis pipeline tailored to nanopore sequencing. It supports both splint- and duplex-enriched telomeric libraries and is designed for ease of use with experimentally validated defaults and seamless integration into the EPI2ME platform. No command-line experience or manual data manipulation is required for standard operation. At the same time, TARPON offers advanced users full flexibility to adjust parameters for specialized research questions, including non-human samples and atypical telomeric features.

By generating accessible tabular and graphical outputs, including a complete HTML report, TARPON empowers researchers and clinicians alike to analyze telomere length with precision and transparency. The pipeline is publicly available at https://github.com/baumannlab/TARPON.

Ethics Approval and Consent to Participate

The Clinical Ethics Committee of the Johannes Gutenberg University Medical Center, acting in its capacity as an Institutional Review Board (IRB), has reviewed the research project and determined that all human genomic material used in the study is fully anonymized with no means of re-identification, informed consent for research use has been obtained for all human-derived samples, including secondary use, and the public datasets are used in accordance with their respective data use agreements. Given that these conditions are implemented, no formal ethical approval is required.

Consent for Publication

Consent for publication was obtained for the data relating to all individuals in this study.

Availability of Data and Materials

The datasets supporting the conclusions in this article are included as additional files. TARPON is publicly available at https://github.com/baumannlab/TARPON.

Competing Interests

The authors declare that they have no competing interests.

Funding

This work was funded in part by an Alexander von Humboldt Professorship awarded to P.B. at JGU.

Authors' Contributions

D.V.H performed all relevant laboratory work relating to the sequencing of simplex enriched telomeric samples HG002-SE and B2-SE. N.D. performed all relevant laboratory work relating to samples HEK293-DE and B2-DE as well as all relevant computational work presented here. P.B. secured funding, aided in project development, and assisted in preparation of the manuscript. All authors read and approved the final manuscript. P.N., in his capacity as the Head of Clinical Ethics at Johannes Gutenberg University Medical Center, ensured that all ethical regulations were met during the duration of the study and aided in obtaining ethics committee approval.

Acknowledgements

The authors would like to thank Dr. Lars Erichsen for donating HEK293T cells, the lab of Dr. Susann Schweiger for access to genomic DNA of a 60-year-old individual, systems administrator Robert Vettel, as well as Dr. Zoe Gill and other members of the Baumann Laboratory for insightful discussions. We thank the Institute for Quantitative and Computational Biosciences, the Nucleic Acid Core Facility at JGU, and the Computational Systems Genetics Group at UMC for computing resources.

847 **Figure Legends**

848 Figure 1. Pipeline Processes and Implementation into EPI2ME. (a) Graphical display of the
849 pipeline workflow and strategy. (b) To install TARPON into EPI2ME the user can input the
850 URL of the TARPON GitHub repository into a pop-up window that appears after clicking
851 "Import Workflow". (c) Once added to EPI2ME the user can launch the workflow through
852 the GUI and modify all relevant parameters without accessing the command line or editing
853 the source code.

854 Figure 2. Putative Read Isolation and Chimera Filtering (a) Number of putative telomeric
855 reads isolated using fast, high accuracy (HAC), or super high accuracy (SUP) basecalling
856 models across the different isolation methods (left), the speed at which these methods
857 perform (top right), and the speed at which these methods perform on different file sizes
858 (bottom right). (b) Overlapping read ids of first-pass isolated telomeric sequences for four
859 sequencing runs using ten non-consecutive telomeric repeats. (c) Percentage of reads in a
860 sequencing run that are identified using ten non-consecutive telomeric repeats. (d)
861 Percentage of telomeric repeats identified in a read that are G rich repeats for four
862 sequencing runs across different basecalling models. (e) Percentage of telomeric reads
863 removed for containing G strand repeats when a C strand enrichment technique is used or
864 chimeric reads when a duplex capture is used.

865

866 Figure 3. Identifying the end of a telomeric sequence using an enrichment technique
867 specific capture probe (a) Number of capture probe sequences found in a read increasing the
868 number of allowed errors within the capture probe sequence. (b) Change in wall-clock time
869 associated with increasing the number of allowed errors within the capture probe sequence.
870 (c) Percentage of reads in which a capture probe sequence is successfully found using a 12-

871 nucleotide capture probe sequence allowing for two errors. (d) Distance between the
872 identified capture probe sequence and the end of the sequence in 200bp bins. (e) Percentage
873 of reads where the capture probe sequence was successfully identified as the capture probe
874 sequence increases in length. (f) Distance between the capture probe sequence and the end
875 of the sequence in 200bp bins while increasing the length of the capture probe sequence and
876 number of allowed errors.

877

878 Figure 4. Identifying the Subtelomere-to-Telomere Boundary (a) Three telomeric reads
879 with the percentage of telomere repeats in a 100bp sliding window (blue lines) and the
880 percentage of telomere+1N repeats in the same window (orange lines). The variant repeat
881 rich (VRR) region is highlighted in gray. (b) Percentage of telomere+1N repeats in the first
882 300bp of a read with a 20% threshold indicated by a dotted red line. (c) Percentage of
883 telomeric reads removed for containing greater than 20% telomere+1N repeats in the first
884 300bp of the sequence. (d) Mean absolute value difference in predicted versus manually
885 annotated start of the VRR-region when identifying the start of the VRR-region by the first
886 occurrence of consecutive telomere+1N repeats, (e) when identifying the start of the VRR-
887 region by a 100bp jumping window with a 10bp jump and a minimum threshold of
888 telomere+1N repeats that must be maintained for the remainder of the read, and (f) when
889 identifying the start of the VRR-region by a 100bp jumping window with a 10bp jump and
890 a minimum threshold of telomere+1N repeats for that must be exceeded and not decrease
891 past the second threshold. (g) Mean absolute value difference in predicted versus manually
892 annotated start of the VRR-region when increasing the sliding window size and (h)
893 increasing the interval size to decrease computational time when a 60%/5% threshold
894 system are used with a sliding window size of 100bp. (i) Differences between manually

895 annotated start sites and computationally identified start site. (j) Percentage of telomeric
896 sequences where the VRR-region start site was determined.

897 Figure 5. Filtering of Telomeric Sequences. (a) Percentage of telomere composed of
898 telomeric+1N repeats. (b) An example non-telomeric sequence passing all filtering criteria
899 that is erroneous and (c) an example telomeric sequence passing all filtering criteria that is
900 not erroneous and should be retained. (d) Mean telomere quality of sequences in which the
901 telomeres are composed of greater than or less than 60% telomere+1N repeats. (e) Read
902 length of filtered telomeric sequences and (f) telomere length of filtered sequences. (g)
903 Percentage of filtered reads for a telomere+1N composition below 60%. (h) Percentage of
904 telomere+1N repeats in the 2kb immediately prior to the VRR-region start site in the
905 subtelomeric direction. (i) A high telomere+1N percentage in prior 2kb read that should not
906 be removed from the analysis and (j) a high percentage read that should be removed from
907 the analysis. (k) Percentage of telomeric sequences removed for containing a high
908 proportion of telomere+1N repeats before the VRR-region start site. (* = p.val < 0.05, ** =
909 p.val < 0.005)

910

911 Figure 6. Chromosome Arm-Specific Telomere Length**.** (a) Mapping statistics of 3 different
912 datasets to 3 different reference genomes. (b) Distribution of telomeric sequences across
913 chromosome arms when aligning to different reference genomes. (c) Alignment mismatches
914 and gaps within alignments to different references. (d) Number of clusters composed of
915 greater than 0.02% of the input telomeric sequences across different clustering parameters
916 and (e) the percentage of input telomeric sequences in the clusters. (f) Telomeric sequence
917 cluster distribution compared to the distribution of telomeric sequences during alignment

918 (red). (g) Number of clusters that each chromosome arm is related to based on the aligned
919 read ids.

920 **Supplemental Figure Legends**

921 Supplemental Figure 1. Percentage of telomeric sequences from four sequencing runs that
922 contain a capture probe separated by strand.

923

924 Supplemental Figure 2. Three examples of telomeric sequences that contain a capture probe
925 but terminate within the variant repeat rich region. Blue lines represent the frequency of
926 GGTTAG repeats within a 100bp sliding window, orange lines represent the frequency of
927 all telomere+1N repeats within a 100bp sliding window, and red lines represent the VRR-
928 region start site.

929

930 Supplemental Figure 3. Percentage of telomeric sequences that contain less than 20%
931 telomere+1N repeats in the first 300bp of the read opposite of the capture probe separated
932 by strand.

933

934 Supplemental Figure 4. Example telomeric sequence containing a subtelomeric island that
935 is rich in telomere+1N repeats.

936

937 Supplemental Figure 5. Average quality score in 100bp segments of telomeric reads that
938 contain large stretches of erroneous sequence that is an artifact of basecalling.

939

940 Supplemental Figure 6. (a) Percentage of telomeres aligning back to each chromosome arm
941 or present in clusters when full length telomeric sequences are used. (b) Number of clusters
942 when non-HG002 samples are clustered using Telogator2 and (c) the percentage of
943 telomeric reads composing the clusters. (d) Percentage of telomeres found within each
944 cluster for non-HG002 samples.
945
946 Supplemental Figure 7. (a) Five example reads from three randomly chosen clusters
947 showing the variant repeat rich region patterns. Blue lines represent the frequency of
948 GGTTAG repeats and orange lines represent the frequency of telomere+1N repeats in a
949 100bp sliding window. (b) Behavior of wf-teloseq in the presence of a subtelomeric island
950 results in the edge of the island being identified as the telomere start site.
951

952 **Supplemental Table Legends**

953 Supplemental Table 1. Oligos used to form wild type capture duplexes.

954 **Supplemental Data**

955 Putative telomeric sequences containing at least ten non-consecutive telomeric repeats are
956 provided for the four samples described in this study as supplemental data files in UBAM format.
957 Additional file 1 contains HG002-SE data, Additional file 2 contains WB60-SE data, Additional file
958 3 contains HEK293-DE data, and Additional file 4 contains WB60-DE data.

959 **References**

960 1. Grill S, Nandakumar J. Molecular mechanisms of telomere biology disorders. J Biol Chem.
961 2021 Jan 1;296:100064.

962 2. Makarov VL, Hirose Y, Langmore JP. Long G Tails at Both Ends of Human Chromosomes
963 Suggest a C Strand Degradation Mechanism for Telomere Shortening. Cell. 1997 Mar
964 7;88(5):657–66.

965 3. McElligott R, Wellinger RJ. The terminal DNA structure of mammalian chromosomes. EMBO
966 J. 1997 Jun 15;16(12):3705–14.

967 4. Olovnikov AM. A theory of marginotomy. The incomplete copying of template margin in
968 enzymic synthesis of polynucleotides and biological significance of the phenomenon. J Theor
969 Biol. 1973 Sep 14;41(1):181–90.

970 5. Watson JD. Origin of concatemeric T7 DNA. Nature New Biol. 1972 Oct 18;239(94):197–201.

971 6. Daniali L, Benetos A, Susser E, Kark JD, Labat C, Kimura M, et al. Telomeres shorten at
972 equivalent rates in somatic tissues of adults. Nat Commun. 2013 Mar 19;4(1):1597.

973 7. Demanelis K, Jasmine F, Chen LS, Chernoff M, Tong L, Delgado D, et al. Determinants of
974 telomere length across human tissues. Science. 2020 Sep 11;369(6509):eaaz6876.

975 8. Allsopp RC, Harley CB. Evidence for a Critical Telomere Length in Senescent Human
976 Fibroblasts. Exp Cell Res. 1995 Jul 1;219(1):130–6.

977 9. Greider CW, Blackburn EH. Identification of a specific telomere terminal transferase activity in
978 Tetrahymena extracts. Cell. 1985 Dec;43(2 Pt 1):405–13.

979 10. Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR. Reverse Transcriptase
980 Motifs in the Catalytic Subunit of Telomerase. Science. 1997 Apr 25;276(5312):561–7.

981 11. Savage SA. Dyskeratosis congenita and telomere biology disorders. Hematol Am Soc Hematol
982 Educ Program. 2022 Dec 9;2022(1):637–48.

983 12. Revy P, Kannengiesser C, Bertuch AA. Genetics of human telomere biology disorders. Nat Rev
984 Genet. 2023 Feb;24(2):86–108.

985 13. Rossiello F, Jurk D, Passos JF, d'Adda di Fagagna F. Telomere dysfunction in ageing and age-
986 related diseases. Nat Cell Biol. 2022 Feb;24(2):135–47.

987 14. Yu HJ, Byun YH, Park CK. Techniques for assessing telomere length: A methodological
988 review. Comput Struct Biotechnol J. 2024 Apr 10;23:1489–98.

989 15. Moyzis RK, Buckingham JM, Cram LS, Dani M, Deaven LL, Jones MD, et al. A highly
990 conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human
991 chromosomes. Proc Natl Acad Sci U S A. 1988 Sep;85(18):6622–6.

992 16. Southern EM. Measurement of DNA length by gel electrophoresis. Anal Biochem. 1979 Dec
993 1;100(2):319–23.

994 17. Martin-Ruiz CM, Baird D, Roger L, Boukamp P, Krunic D, Cawthon R, et al. Reproducibility
995 of telomere length assessment: an international collaborative study. Int J Epidemiol. 2015
996 Oct;44(5):1673–83.

997 18. Aubert G, Hills M, Lansdorp PM. Telomere length measurement—Caveats and a critical
998 assessment of the available technologies and tools. Mutat Res Mol Mech Mutagen. 2012 Feb
999 1;730(1):59–67.

1000 19. Steinert S, Shay JW, Wright WE. Modification of Subtelomeric DNA. Mol Cell Biol. 2004 May
1001 15;24(10):4571–80.

1002 20. Cawthon RM. Telomere measurement by quantitative PCR. Nucleic Acids Res. 2002 May
1003 15;30(10):e47.

1004 21. Njajou OT, Hsueh WC, Blackburn EH, Newman AB, Wu SH, Li R, et al. Association Between
1005 Telomere Length, Specific Causes of Death, and Years of Healthy Life in Health, Aging, and

1006 Body Composition, a Population-Based Cohort Study. J Gerontol A Biol Sci Med Sci. 2009
1007 Aug;64A(8):860–4.

1008 22. Sun G, Cao H, Bai Y, Wang J, Zhou Y, Li K, et al. A novel multiplex qPCR method for
1009 assessing the comparative lengths of telomeres. J Clin Lab Anal. 2021;35(9):e23929.

1010 23. Cunningham JM, Johnson RA, Litzelman K, Skinner HG, Seo S, Engelman CD, et al. Telomere
1011 length varies by DNA extraction method: implications for epidemiologic research. Cancer
1012 Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2013
1013 Nov;22(11):2047–54.

1014 24. Koppelstaetter C, Jennings P, Hochegger K, Perco P, Ischia R, Karkoszka H, et al. Effect of
1015 tissue fixatives on telomere length determination by quantitative PCR. Mech Ageing Dev. 2005
1016 Dec 1;126(12):1331–3.

1017 25. Rufer N, Brümmendorf TH, Kolvraa S, Bischoff C, Christensen K, Wadsworth L, et al.
1018 Telomere Fluorescence Measurements in Granulocytes and T Lymphocyte Subsets Point to a
1019 High Turnover of Hematopoietic Stem Cells and Memory T Cells in Early Childhood. J Exp
1020 Med. 1999 Jul 19;190(2):157–68.

1021 26. Alter BP, Baerlocher GM, Savage SA, Chanock SJ, Weksler BB, Willner JP, et al. Very short
1022 telomere length by flow fluorescence in situ hybridization identifies patients with dyskeratosis
1023 congenita. Blood. 2007 Sep 1;110(5):1439–47.

1024 27. Sanchez SE, Gu Y, Wang Y, Golla A, Martin A, Shomali W, et al. Digital telomere
1025 measurement by long-read sequencing distinguishes healthy aging from disease. Nat Commun.
1026 2024 Jun 18;15(1):5148.

1027 28. Schmidt TT, Tyer C, Rughani P, Haggblom C, Jones JR, Dai X, et al. High resolution long-read
1028 telomere sequencing reveals dynamic mechanisms in aging and cancer. Nat Commun. 2024 Jun
1029 18;15(1):5149.

1030 29. Karimian K, Groot A, Huso V, Kahidi R, Tan KT, Sholes S, et al. Human telomere length is
1031 chromosome end–specific and conserved across individuals. Science. 2024 May
1032 3;384(6695):533–9.

1033 30. Tham CY, Poon L, Yan T, Koh JYP, Ramlee MK, Teoh VSI, et al. High-throughput telomere
1034 length measurement at nucleotide resolution using the PacBio high fidelity sequencing
1035 platform. Nat Commun. 2023 Jan 17;14(1):281.

1036 31. Sholes SL, Karimian K, Gershman A, Kelly TJ, Timp W, Greider CW. Chromosome-specific
1037 telomere lengths and the minimal functional telomere revealed by nanopore sequencing.
1038 Genome Res. 2021 Oct 26;gr.275868.121.

1039 32. Rautiainen M, Nurk S, Walenz BP, Logsdon GA, Porubsky D, Rhie A, et al. Telomere-to-
1040 telomere assembly of diploid chromosomes with Verkko. Nat Biotechnol. 2023
1041 Oct;41(10):1474–82.

1042 33. Baird DM, Rowson J, Wynford-Thomas D, Kipling D. Extensive allelic variation and ultrashort
1043 telomeres in senescent human cells. Nat Genet. 2003 Feb;33(2):203–7.

1044 34. Trujillo KM, Bunch JT, Baumann P. Extended DNA binding site in Pot1 broadens sequence
1045 specificity to allow recognition of heterogeneous fission yeast telomeres. J Biol Chem. 2005
1046 Mar 11;280(10):9119–28.

1047 35. Sfeir AJ, Chai W, Shay JW, Wright WE. Telomere-End Processing. Mol Cell. 2005
1048 Apr;18(1):131–8.

1049 36. Button LL, Astell CR. The Saccharomyces cerevisiae chromosome III left telomere has a type
1050 X, but not a type Y', ARS region. Mol Cell Biol. 1986 Apr;6(4):1352–6.

1051 37. O'Donnell S, Yue JX, Saada OA, Agier N, Caradec C, Cokelaer T, et al. Telomere-to-telomere
1052 assemblies of 142 strains characterize the genome structural landscape in Saccharomyces
1053 cerevisiae. Nat Genet. 2023 Aug;55(8):1390–9.

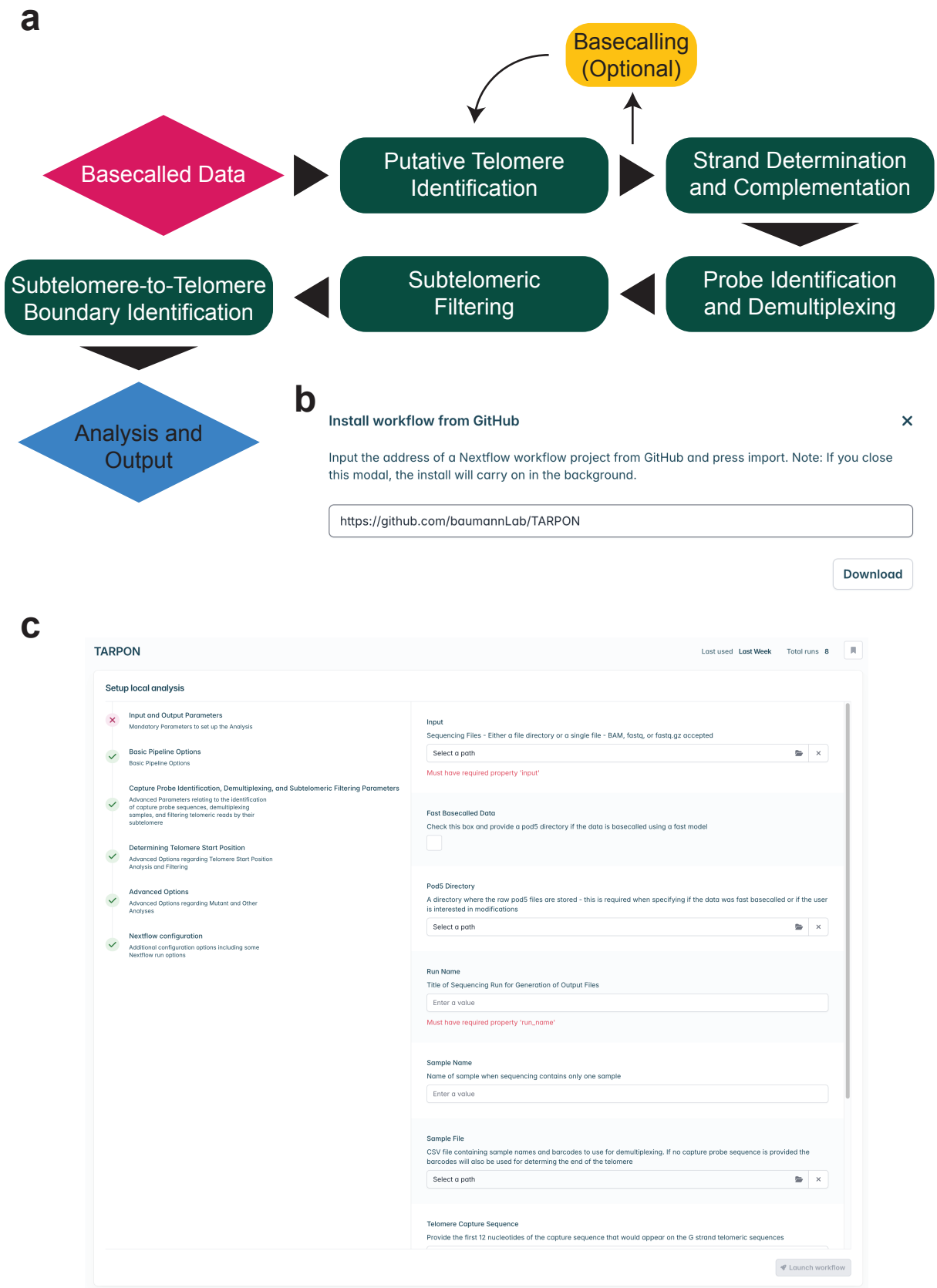
1054 38. Stephens Z, Kocher JP. Characterization of telomere variant repeats using long reads enables
1055 allele-specific telomere length estimation. BMC Bioinformatics. 2024 May 17;25(1):194.

1056 39. Teplitz GM, Pasquier E, Bonnell E, De Laurentiis E, Bartle L, Lucier JF, et al. A mechanism for
1057 telomere-specific telomere length regulation. bioRxiv. 2024 Jun 12;2024.06.12.598646.

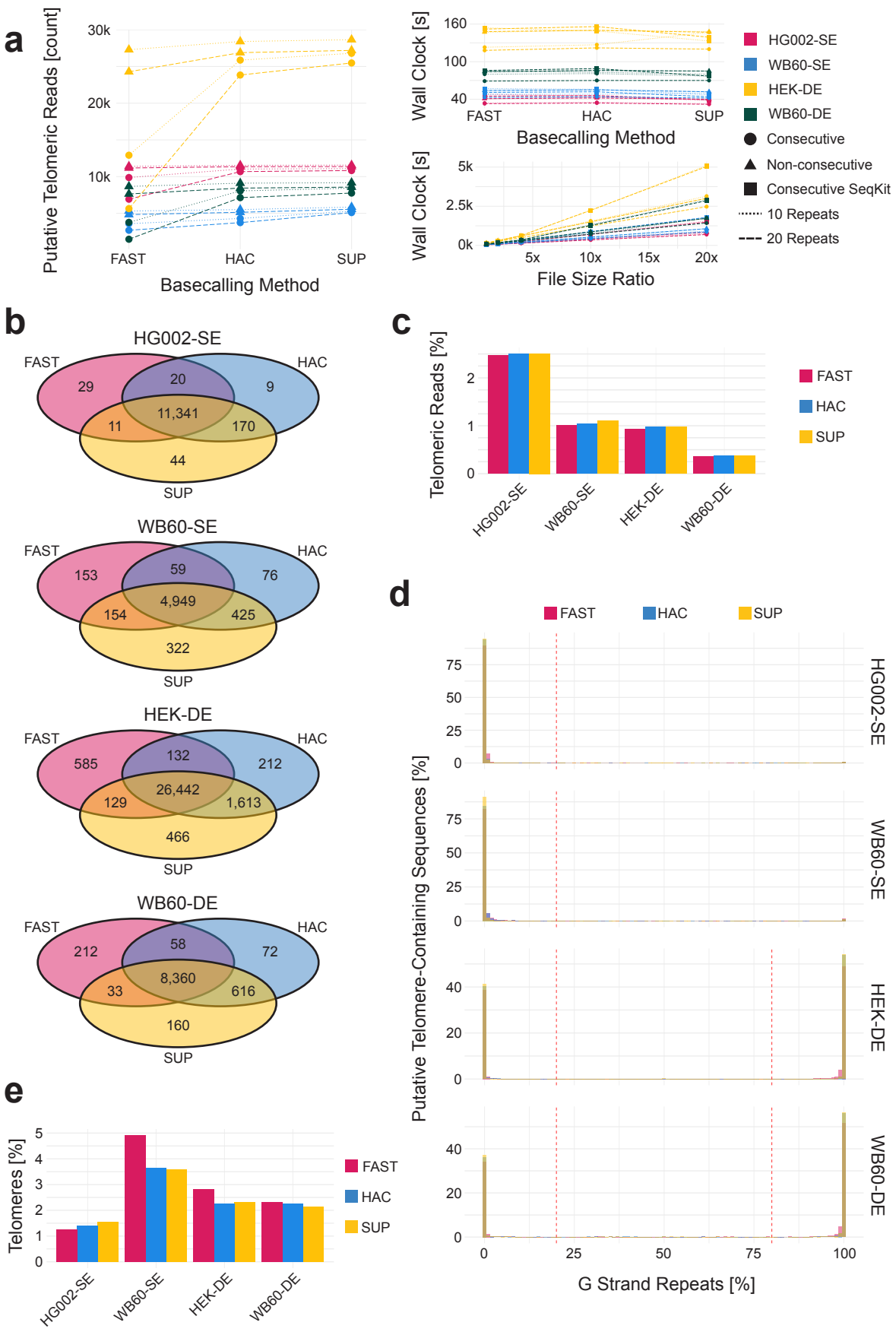
1058 40. Hinchie AM, Sanford SL, Loughridge KE, Sutton RM, Parikh AH, Gil Silva AA, et al. A
1059 persistent variant telomere sequence in a human pedigree. Nat Commun. 2024 Jun
1060 1;15(1):4681.

1061 41. Kanoh J. Roles of Specialized Chromatin and DNA Structures at Subtelomeres in
1062 Schizosaccharomyces pombe. Biomolecules. 2023 May;13(5):810.

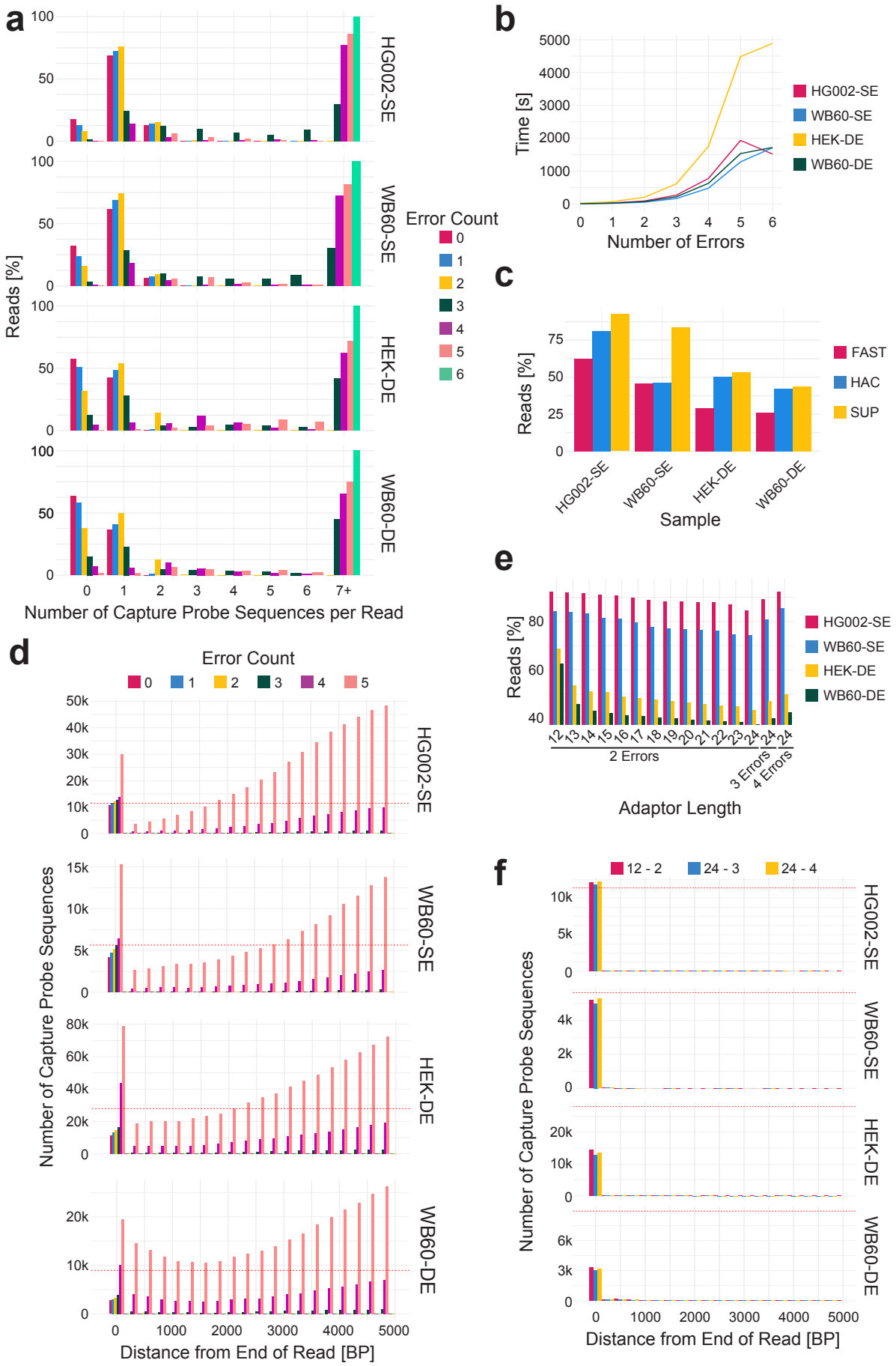
1063



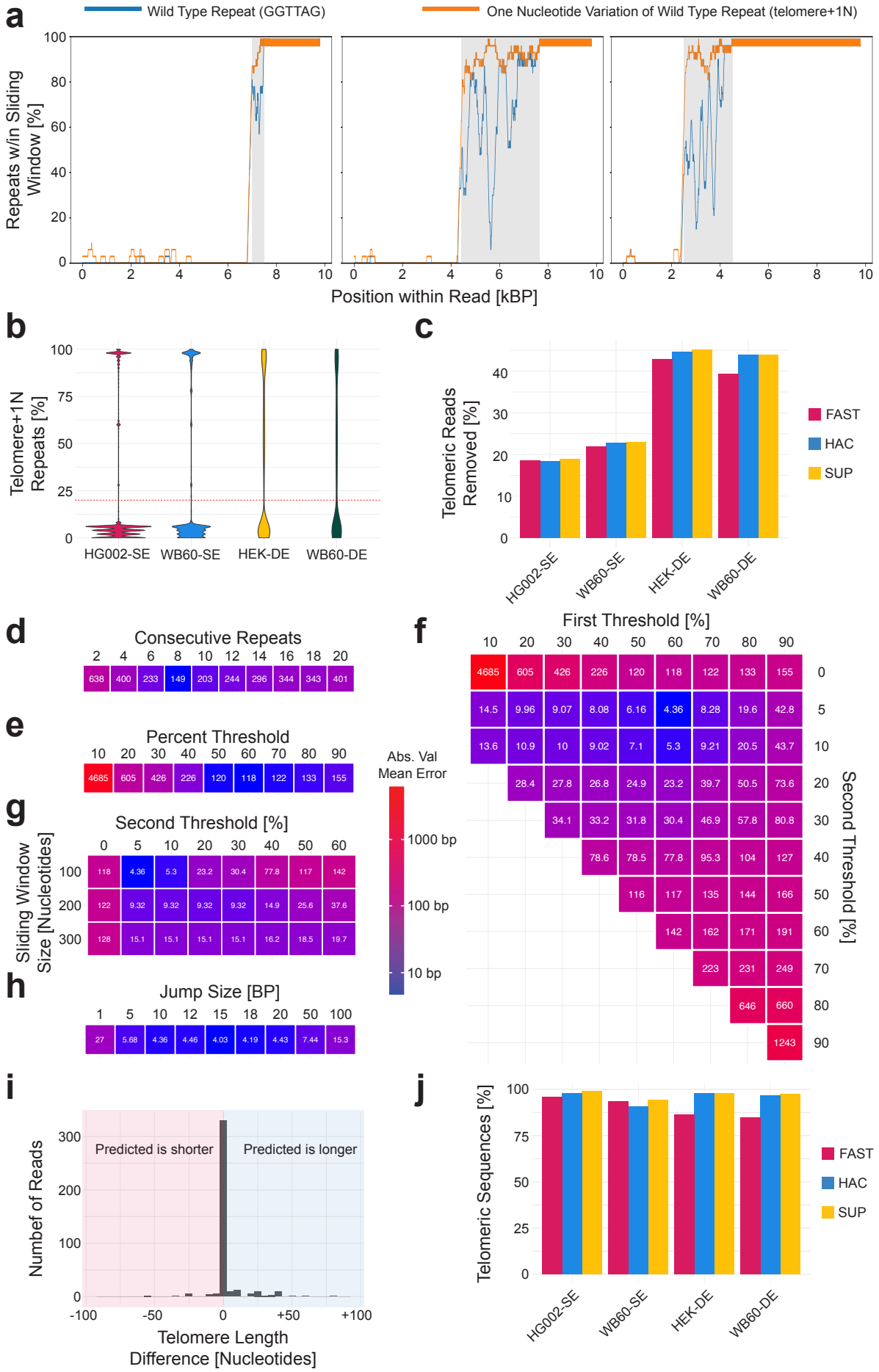
Deimler, 2025; Figure 1



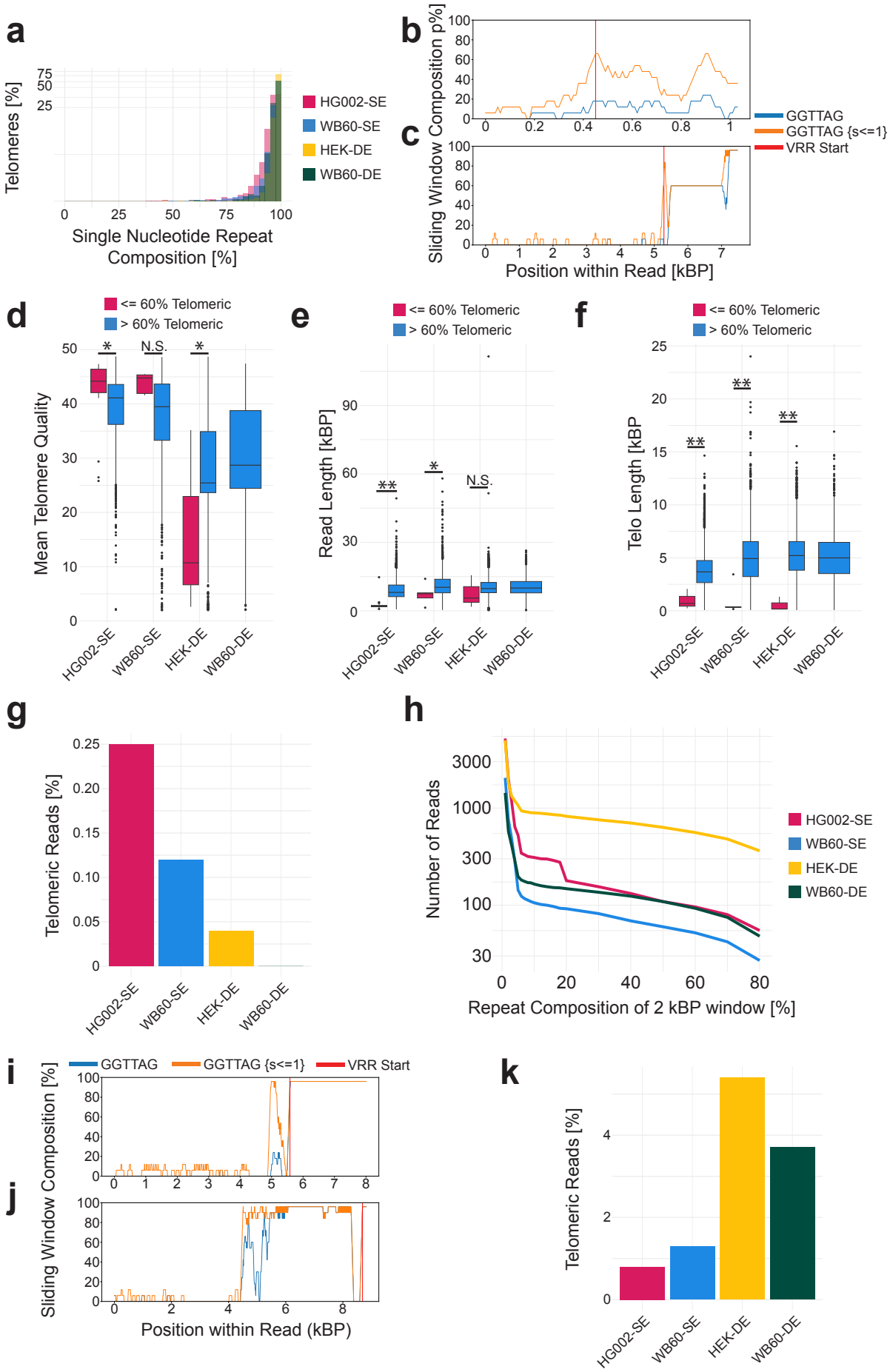
Deimler, 2025; Figure 2



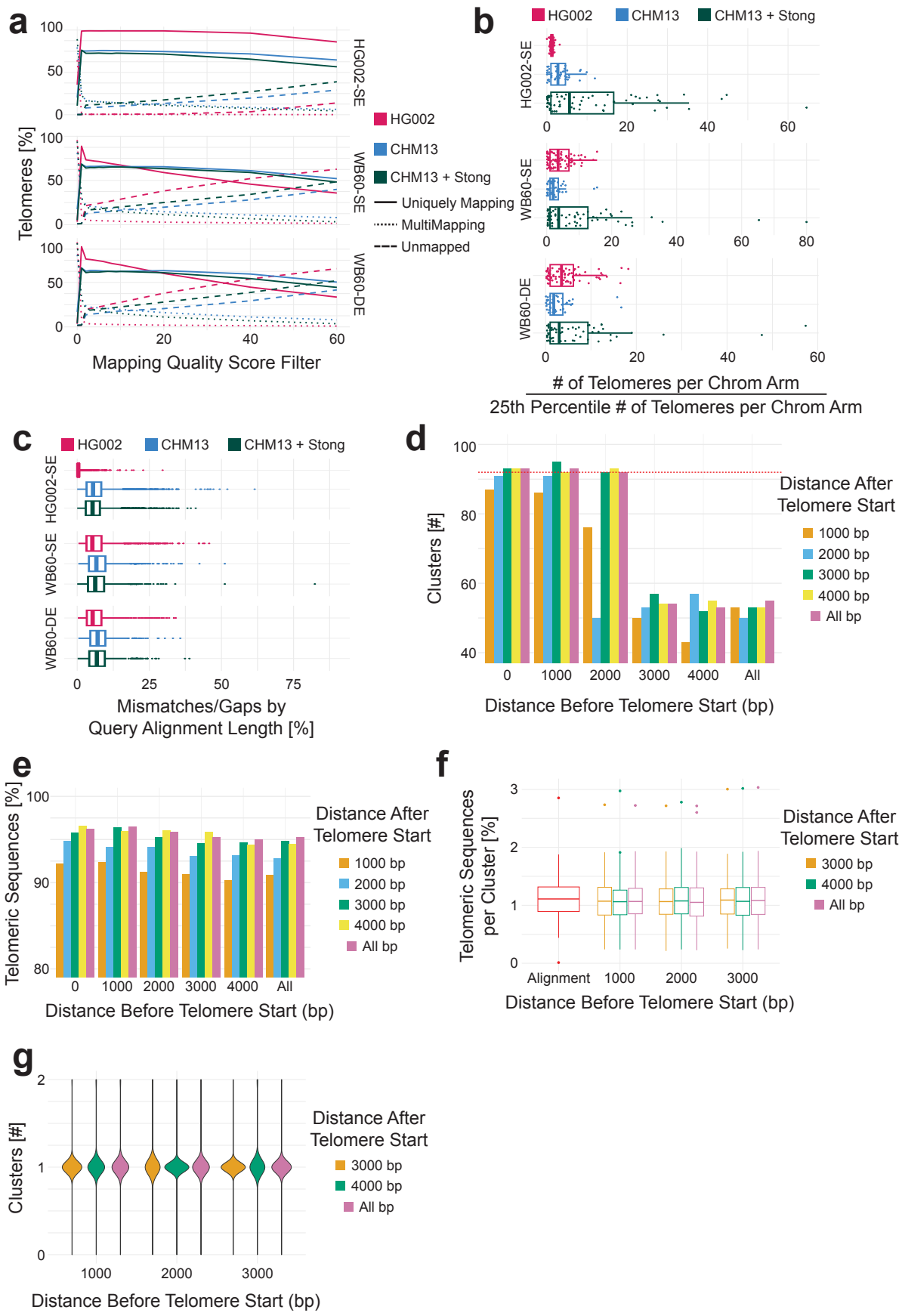
Deimler, 2025; Figure 3



Deimler, 2025; Figure 4



Deimler, 2025; Figure 5



Deimler, 2025; Figure 6