TARPON - a Telomere Analysis and Research Pipeline Optimized for Nanopore

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

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

* Correspondence: peter@baumannlab.org and

nathanieldeimler.research@gmail.com

Keywords: telomere length, Nanopore, pipeline

Abstract

Long-read sequencing has transformed many areas of biology and holds significant promise for telomere research by enabling nucleotide-level resolution of chromosome arm–specific telomere length analyses in both model organisms and humans. However, the adoption of new technologies, particularly in clinical or diagnostic contexts, requires careful validation to identify potential technical and computational limitations. 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, 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 adjacent non-telomeric sequence on the opposite end are excluded. A sliding-window approach defines the subtelomere-telomere boundary, followed by quality filtering to remove low-quality or subtelomeric reads that passed earlier steps.

The pipeline outputs customizable statistics, text-based summaries, and publication-ready visualizations (HTML, PNG, PDF). While default settings are optimized for diagnostic workflows, all parameters are easily adjustable via the GUI or command line to support diverse applications. These include telomere analyses in variant-rich samples (e.g., ALT-positive tumors) and organisms with non-canonical telomeric repeats such as insects (GTTAG) and plants (GGTTTAG). TARPON is the first complete and experimentally validated pipeline for Nanopore-based telomere analysis requiring no pre-processing or prior bioinformatics expertise, while offering flexibility for advanced users.

Introduction

Telomeres, the structures that protect the ends of linear eukaryotic chromosomes, are comprised of G-rich DNA repeat sequences, proteins, and telomeric RNA (Grill and Nandakumar 2021). In humans, telomeres span from 3–15 kb and terminate in a 100 to 200 nucleotide single-stranded G-rich 3' overhang (Makarov, Hirose, and Langmore 1997; McElligott and Wellinger 1997). Each time a cell duplicates, approximately 50–150 base pairs of terminal sequence are lost from each chromosome end due to the end replication problem (Olovnikov 1973; Watson 1972), resulting in progressive telomere shortening (Daniali et al. 2013; Demanelis et al. 2020). Telomeric repeat arrays protect the integrity of the coding regions of the genome by temporarily buffering this sequence attrition. However, in the absence of mechanisms that replenish telomeric DNA, progressive telomere shortening eventually triggers cellular senescence as telomeres reach a critical length (Allsopp and Harley 1995). Telomere length is maintained in highly proliferative cells by telomerase, a ribonucleoprotein complex comprised at its core of the catalytic protein Telomerase Reverse Transcriptase (TERT) and a non-coding Telomerase RNA subunit, TR (Greider and Blackburn 1985; Lingner et al. 1997). Telomere Biology Disorders (TBDs) are a symptomatically heterogeneous group of syndromes associated with abnormally short telomeres or telomere instability (Savage 2022; Revy, Kannengiesser, and Bertuch 2023). The broad spectrum of symptoms resembles the aging process, ranging from changes in skin pigmentation and nail dystrophy to severe effects such as pulmonary fibrosis and total bone marrow failure (Rossiello et al. 2022).

Accurate measurement of telomere length is therefore essential for both research and clinical applications. Multiple techniques have been developed including, but not limited to, terminal restriction fragment (TRF) analysis by Southern blotting, quantitative PCR (qPCR), and fluorescent in situ hybridization (FISH) (Yu, Byun, and Park 2024). Terminal restriction fragment (TRF) length analysis has long

been considered the gold standard in research settings (Moyzis et al. 1988; Southern 1979). In this method, genomic DNA is digested by restriction enzymes that frequently cut within the genome but not within telomeric repeats. Digested DNA is then resolved on an agarose gel and hybridized with a labeled telomere-specific probe. Although this method shows low inter-laboratory variation (Martin-Ruiz et al. 2015), reproducibility is limited by poorly described protocols for digestion and gel quantification (Aubert, Hills, and Lansdorp 2012). Moreover, variations in subtelomeric sequences or DNA modifications may lead to apparent inter-individual differences in telomere length (Steinert, Shay, and Wright 2004).

Quantitative PCR (qPCR) estimates telomere length indirectly by comparing the quantity of telomeric repeat amplification products to a single-copy gene product (Cawthon 2002). Its simplicity, low cost, and minimal DNA input requirements have enabled widespread use in clinical samples and large-scale comparisons (Njajou et al. 2009; Sun et al. 2021). However, qPCR has been shown to yield variable results depending on laboratory protocols (Martin-Ruiz et al. 2015), DNA extraction methods (Cunningham et al. 2013), and storage conditions. For instance, samples stored in 4% formaldehyde showed increased telomere length measurements over time, a bias not observed with samples preserved in RNAlater (Koppelstaetter et al. 2005).

Flow-FISH, currently the clinical gold standard, uses fluorescently labeled peptide nucleic acid (PNA) probes to detect telomeric DNA in permeabilized cells, with fluorescence intensity measured by flow cytometry (Rufer et al. 1999). It avoids certain biases associated with TRF and qPCR and enables direct comparison to reference standards (Alter et al. 2007). However, Flow-FISH requires fresh blood samples, limiting its use on archived or bio-banked material.

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

Third generation long-read sequencing has emerged as a powerful tool across biological disciplines, including aging and telomere research. Oxford Nanopore Technologies (ONT) sequencing has enabled nucleotide-resolution analysis of human telomeres (Sanchez et al. 2024; Schmidt et al. 2024; Karimian et al. 2024). In parallel, PacBio HiFi sequencing has also been used to generate high-throughput, single-molecule telomere length measurements at nucleotide resolution across diverse human cell lines and patient samples (Tham et al. 2023). However, telomeres represent only ~0.01% of the human genome, and whole-genome sequencing yields relatively few telomeric reads. To address this, two enrichment strategies have been developed: (1) a physical enrichment using biotinylated duplexes and streptavidin-coated beads ("duplex-enriched") (Karimian et al. 2024) and (2) a library preparation method that captures telomeric ends using a telomere-specific splint and ONT's adapter overhang ("splint-enriched") (Sanchez et al. 2024; Schmidt et al. 2024). Both strategies increase telomeric read recovery and append a known capture probe to the distal telomere end, confirming full-length capture.

Unlike traditional approaches that return a single mean or median telomere length per sample, long-read sequencing enables single-molecule resolution of telomere length distributions. With sufficient subtelomeric sequence, reads can be aligned to the genome, allowing chromosome arm–specific telomere assignment, as shown in yeast (Sholes et al. 2021) and the human cell line HG002 (Karimian et al.

2024; Sanchez et al. 2024; Schmidt et al. 2024). However, this approach is currently limited to samples with high-quality subtelomeric reference assemblies (Rautiainen et al. 2023).

Despite these promising advances, no standardized or validated pipeline exists for the analysis of Nanopore-based telomere data. Here, we present **TARPON** (Telomere Analysis and Research Pipeline Optimized for Nanopore), an experimentally validated, end-to-end pipeline for analyzing telomere reads obtained via ONT sequencing. TARPON is implemented in Nextflow and can be executed via the command line or through the user-friendly EPI2ME graphical interface, which requires no programming experience. The pipeline supports both duplex- and splint-enriched libraries and includes automated quality control, capture probe detection, telomere length quantification, and comprehensive reporting. Parameters can be modified through either interface, enabling both novice and expert users to tailor TARPON to a wide range of experimental applications.

Methods

Sample Selection and DNA Isolation

Genomic DNA for the below described telomere enrichment protocols had multiple sources. HG002 genomic DNA was purchased directly from the Coriell Institute for Medical Research (NA24385); this sample originates from an immortalized lymphoblastoid cell line and was chosen for analysis because of its publicly available, high quality subtelomeric reference genome. Clinical sample genomic DNA (WB60) originates from whole blood of a 60-year-old individual and was extracted using Chemagen Chemagic DNA Blood Kit Special for 5 mL on the Chemagen Robot MSM1 following manufacturer's instruction. HEK293T genomic

DNA was isolated using the Qiagen Blood and Cell Culture DNA Midi Kit (Cat. no. 13343) following manufacturer's recommendations.

Splint Capture Based Telomere Enrichment

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

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

A 0.5x SPRI bead purification was performed as above, and DNA was eluted into 32 µL nuclease-free water. Sequencing adapter ligation was performed using 10,000 U Quick T4 DNA Ligase (NEB E6056) and 5 µL native adapter in 1x NEBNext Quick Ligation buffer at 21°C for 20 minutes. Another 0.5x SPRI bead cleanup was

carried out, followed by two washes with long fragment buffer (LFB). DNA was eluted into 25 µL elution buffer and loaded onto a R10.4.1 MinION flow cell for a 72-hour sequencing run.

Duplex Capture Based Telomere Enrichment

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

This duplex, hereafter referred to as the "capture probe," was ligated to 3 µg genomic DNA in 50 µL reactions (16 replicates) containing 75 pmol capture probe, 1x HiFi Taq Ligase Buffer (B0647SVIAL), and 1 µL HiFi Taq DNA Ligase (M0647SVIAL) using a MasterCycler X50s. Ligations ran for 5 minutes at 65°C, followed by 25 cycles of 1 minute at 65°C and 3 minutes at 45°C with 1°C/s ramp rate. EcoRV-HF (20 U) was added and incubated at 37°C for 2 hours.

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

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

Basecalling and TARPON Pipeline Execution

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

Strand-specific filtering is applied via the `--c_strand_only` flag for splint-enriched libraries. Filtering thresholds, barcode demultiplexing, capture probe sequence, and GUI/CLI 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 both bioinformaticians 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 outputs multiple files summarizing each processing step, including telomere read counts, read-level length and quality statistics, and 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 additional software dependencies is required, as TARPON utilizes Docker containers through Nextflow, inherently avoiding version incompatibility issues.

For users less familiar with the command line or bioinformatics, TARPON is integrated into Oxford Nanopore Technologies (ONT) EPI2ME agent. This platform

assists with the installation of Java, Docker, and Nextflow on the local machine, enabling GUI-based operation of the pipeline. Integration into EPI2ME is achieved by entering the GitHub repository URL into the EPI2ME "Add Workflows" utility (Fig. 1b). Once loaded, the user can specify pipeline inputs and adjust parameters either through the GUI (Fig. 1c) or the command line. Configurable options include the capture probe sequence, barcode files for sample demultiplexing, and parameter settings for telomeric read isolation, capture probe and barcode detection, subtelomere-to-telomere boundary identification, and the stringency of each filtering step. Additional guidance on all configurable parameters is available in the README file provided with the pipeline.

First Pass Telomeric Read Isolation

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

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

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

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

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

To further confirm that reads identified using the ten non-consecutive telomeric repeat 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 approximately 1–5% of reads, independent of enrichment method (Fig. 2e). Only reads passing this strand-filtering step are used in subsequent analysis.

Identification of the Telomeric Capture Probe

The strategy of capturing telomeres via ligation of oligonucleotides or partial duplexes to the 5′ end of the C-strand or 3′ end of the G-strand has its roots in earlier ligation-based assays such as the amplification and sequencing of single telomeres (Baird et al. 2003; Trujillo, Bunch, and Baumann 2005; Sfeir et al. 2005). Building on this foundation, more recent sequencing-based protocols have refined the technique by integrating capture probes into high-throughput library preparation workflows (Sanchez et al. 2024; Schmidt et al. 2024; Karimian et al. 2024).

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

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

Greater than 75% of splint-enriched telomeric reads contained a capture probe when two mismatches were allowed within the 12-nucleotide probe sequence. In contrast, only ~50% of duplex-enriched reads contained a detectable probe under the same criteria. This reduction is likely due to two factors. First, during duplex-enrichment, only one strand of each DNA fragment needs to carry a capture probe for successful streptavidin pulldown, potentially resulting in half of the resulting telomeric reads lacking a probe. Second, only the reverse complement of the GGTTAG permutation was used as a capture probe, which anneals in register with the 5′ end of the C-strand at the double- to single-strand junction (Sfeir et al. 2005). In contrast, the G-strand overhang is more heterogeneous in terminal sequence, and the use of all six possible telomeric repeat permutations may increase probe ligation to G-strand ends. This adjustment may be particularly important when capturing G-strands from telomerase-negative cells where bias for a specific 3’ end permutation is minimal (Sfeir et al. 2005).

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

Increasing the number of allowed mismatches when identifying the capture probe also increases the number of off-target sequences detected. While the majority of probe matches are located within the final 200 bp of the read (or the first 200 bp in C-strand reads), a small number of matches appear between 4 kb and 5 kb from the read end (Fig. 3d). These likely reflect subtelomeric sequences with similarity to the capture probe rather than true probe ligation sites, and represent computational artifacts introduced by relaxed stringency. To mitigate this, an additional filter was introduced to ensure that the capture probe is only accepted if it is the first match found after the identification of twenty telomeric repeats. This 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 (Sanchez et al. 2024; Karimian et al. 2024; Schmidt et al. 2024).

To develop a consistent and biologically informed algorithm for telomere start detection, we first visualized telomeric sequences across thousands of reads. We observed that the frequency of canonical telomeric repeats (GGTTAG) alone did not consistently define a clear boundary when measured in 100 bp sliding windows (Fig. 4a, blue lines). In contrast, a strong and persistent increase in one-nucleotide variants of GGTTAG emerged in many reads (Fig. 4a, orange lines). Once these variant-containing windows surpassed a certain threshold, they typically maintained >50% signal density across the remainder of the telomere. We refer to

this combined pattern of canonical repeats and single-nucleotide variants as "telomere+1N". The region enriched for the variants is known as the variant repeat–rich (VRR) region and varied in length from a few hundred base pairs (Fig. 4a, left) to several kilobases (Fig. 4a, middle and right) before transitioning into invariant GGTTAG repeats. Considering a potential functional relevance of this telomere repeat containing region, we included it in the telomere length output. This decision is supported by observations of reads that are essentially devoid of a terminal stretch of homogeneously wildtype repeats yet terminate in a capture probe (Supplemental Fig. 2). Accordingly, TARPON calculates the telomere length from the start of the VRR-region to the distal end of the telomere, as defined by the position of the capture probe.

The telomere length metric requires the ability to accurately detect the start of the VRR-region and the associated frequency spike of telomere+1N repeats. However, while plotting the percentage of telomere+1N repeats within the first 300 bp of the read, a notable bimodal distribution emerged: most reads contained less than 15% telomere+1N repeats, while a smaller subset exceeded 80% (Fig. 4b). This latter group represents telomere-containing sequences that begin within the VRR-region or even within homogenous GGTTAG repeat sequence. These reads were excluded from further analysis as proximally truncated, since the full length of the VRR-region or telomere sequence cannot be determined with confidence (for C-strand reads, this uncertainty applies to the read end since all reads have been reverse complemented into the G-strand orientation). To avoid skewing length estimates by including these truncated telomeric reads, it was necessary to exclude 15 to 50% of sequences from further analysis (Fig. 4c). This percentage was at the higher end of the range in duplex-enriched datasets, likely due to an abundance of

short reads corresponding to partial C-strand sequences (Supplemental Fig. 3), which may have arisen during library preparation and fill-in synthesis of the G-strand overhang.

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

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

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

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

Filtering for High Confidence Telomeric Sequences

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

A second validation step was performed to confirm the accurate identification of the VRR-region start by examining the 2 kb of sequence immediately upstream (in the subtelomeric direction) of the predicted boundary. This region was analyzed to ensure it was not composed primarily of telomere+1N repeats (Fig. 5h). Elevated telomere+1N signal in this region can arise either from genuine subtelomeric structure, consistent across reads (Fig. 5i), or from basecalling artifacts (Fig. 5j). Reads resembling Fig. 5j, which exhibit a sharp transition from fully telomeric sequence to 0% telomere+1N content, are attributed to sequencing artifacts introduced by dorado v0.7.0. These segments typically show reduced quality scores relative to the rest of the telomeric sequence (Supplemental Fig. 5) and lack consistent VRR-region repeat patterns.

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

Read Assignment to Specific Chromosome Arms

Chromosome-specific telomere length differences have been reported in *S. cerevisiae* (Button and Astell 1986; O'Donnell et al. 2023), and underlying regulatory mechanisms have recently been identified (Sholes et al. 2021; Teplitz et al. 2024).

These findings have sparked considerable interest in applying chromosome arm–level telomere analysis to other model organisms as well as humans. Accurate assignment of each telomeric read to a unique chromosome arm is therefore an important goal.

Aligning telomeric reads from HG002 cells to a high-quality, HG002-specific telomere-to-telomere (T2T) reference assembly enables assignment of over 90% of reads to a single chromosome arm (Fig. 6a, HG002-SE, solid red line). However, aligning the same dataset to other subtelomeric references or general-purpose T2T assemblies (e.g., STONG or CHM13) yields substantially lower performance: only ~60% of reads map uniquely, many are multimapped with low quality scores, and the proportion of unmapped reads increases as the minimum mapping quality (MAPQ) threshold is raised. This issue becomes more pronounced when aligning telomeric reads from non-HG002 samples—such as the clinical sample WB60 (SE and DE libraries) —to the HG002, CHM13, or STONG references. In these cases, only ~50% of telomeric reads map uniquely with a MAPQ ≥ 20, reflecting poor alignment specificity (Fig. 5a, WB60-SE). Data derived from the HEK293T cell line was not included for analysis due to its hypotriploid status.

Furthermore, because neither enrichment technique should bias against any particular chromosome arm, an even distribution of telomeric reads across all 92 chromosome arms is expected. This was indeed observed when aligning HG002-derived telomeric reads to the HG002-specific reference: approximately 1.1% of telomeric sequences aligned to each arm, with no arm exceeding 2% of the total. The most enriched arm contained only 2.1 times more reads than the 25th percentile arm, reflecting a near-uniform distribution. Normalization does not rely on the

chromosome arm with the fewest mapped reads, in order to avoid introducing bias from low-level mapping artifacts.

In contrast, alignments of the same HG002 dataset to other reference genomes (e.g., CHM13 or CHM13+STONG) introduced significant bias. In these cases, the most enriched chromosome arm contained 11.7-fold and 64.6-fold more telomeric reads, respectively, than arms in the lowest quartile (Fig. 6b). The effect was also seen in clinical samples aligned to these references, where maximum fold differences ranged from 17.7x to 70x (Fig. 6b). These disparities were accompanied by higher mismatch and gap counts when aligning clinical datasets to non-matching references, as compared to HG002 aligned to its own reference (Fig. 6c). These results clearly demonstrate that aligning telomeric sequences to a reference genome derived from a different cell line or individual does not provide reliable chromosome arm–specific telomere measurements. Improved methodologies will be required to achieve accurate arm-specific telomere length estimates and to investigate chromosome-specific telomere dynamics in human populations.

Extended Features of TARPON

Beyond its default configuration, TARPON includes a range of optional parameters for more specialized analysis of telomeric sequences and enrichment protocols. For duplex-enriched datasets, the `--strand_comparison` flag provides strand-specific enrichment metrics, including relative abundance, filtering outcomes, and telomere length distributions for C- and G-strand reads. The `--detailed_stats` option outputs additional read-level statistics and visualizations, such as repeat composition and telomere length-to-quality score comparisons, in text, graphical, and HTML formats.

For users exploring telomeric enrichment strategies, the `--restriction_digest` flag accepts a comma-separated list of restriction enzyme recognition sites and returns the number of telomeric sequences affected by each cut site. In samples containing mutant telomeric repeats such as in the case of a telomerase RNA template mutation (Hinchie et al. 2024), the `--mutant` flag enables these sequences to be included in all filtering and boundary-identification steps, while also returning statistics related to mutant versus wildtype telomerase processivity.

To reduce computational overhead, TARPON also supports a hybrid basecalling strategy. When `--pod5_directory` is specified, telomeric reads are first isolated from fast basecalled data and then re-basecalled using super-high accuracy (SUP) models. Input file formats do not need to be modified in advance: FASTQ and BAM files are both accepted, and TARPON will convert files to UBAM format internally as needed. Integration with Nextflow and the use of Docker containers or conda environments helps prevent versioning issues and dependency conflicts during execution.

Discussion

TARPON is a flexible and modular pipeline designed to analyze telomeric sequences from Oxford Nanopore Technologies (ONT) long-read sequencing data. It performs a full analysis workflow from fast or SUP basecalled data, including telomeric read isolation, capture probe and barcode identification, and subtelomere-to-telomere boundary detection. The output includes relevant quality metrics and telomere length statistics in graphical and tabular form.

While default parameters are optimized for telomerase-positive human samples, all settings can be customized to accommodate different organisms,

enrichment strategies, or specific experimental goals. This flexibility allows TARPON to support a wide range of research contexts, including species with noncanonical telomere repeats, mutant telomerase variants, and different enrichment chemistries. TARPON continues to evolve with support for new features such as strand bias analysis and telomeric variant detection.

Comparison to Other Telomere Analysis Software

While several software tools are available for analyzing telomere content from Illumina whole-genome sequencing (WGS) data, only TeloBP (Karimian et al. 2024) and Telometer (Sanchez et al. 2024) have been designed specifically for nanopore-based telomere sequencing. Both tools require experience with command-line interfaces, manual installation of dependencies, and troubleshooting version conflicts. Neither TeloBP nor Telometer provides a complete analysis pipeline: users must first isolate and demultiplex telomeric reads, remove low-quality or chimeric sequences, and ensure the presence of a capture probe prior to estimating telomere length or identifying the subtelomere-to-telomere boundary.

In Telometer, reads must be aligned to a subtelomeric reference genome, and any read that does not map within the first or last 30 kb of the reference is discarded. TeloBP requires input in FASTQ or compressed FASTQ format and filters out chimeric reads before proceeding. Both tools perform limited quality filtering—Telometer excludes reads with an average quality score below 9, while TeloBP relies on external preprocessing.

In contrast, TARPON is accessible both via command-line and through a graphical user interface (GUI) (ONT's EPI2ME) for users less familiar with scripting. Nextflow's integration with Docker containers eliminates dependency issues entirely.

TARPON accepts fast or SUP basecalled output directly from the sequencer, requiring no data preprocessing. Unlike existing tools, TARPON also evaluates read integrity and completeness, automatically filtering out chimeric and truncated telomeres prior to analysis.

While TeloBP, Telometer, and TARPON all use a jumping window approach to identify the subtelomere-to-telomere boundary, only TARPON allows users to adjust relevant parameters without modifying source code. In TeloBP, changes require directly 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 the 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 exceed 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 calculated as: (observed – expected) / expected × 100. This value is then plotted, and the area under the curve (AUC) is computed using a series of 83 sliding windows. If the AUC in a given window is less than –50 (suggesting the region is not telomeric) and the next window contains fewer or equal telomeric repeats, this defines a new threshold from which further analysis begins. From this threshold, the differences in AUC values between adjacent sliding windows are calculated. If the absolute difference between two AUC values is less than 0.2, or if the current difference is less than 0.2 and the next difference exceeds 0.2, the boundary between the telomere and subtelomere is defined. Telomere length is then estimated as the distance from this boundary to the end of the read.

TARPON takes a different approach, using a threshold-based method to identify the transition from subtelomeric to telomeric sequence. It incorporates both telomere+1N repeat density and local sequence context, minimizing false positives arising from repetitive subtelomeric regions and reducing overestimation due to terminal artifacts. After boundary detection, TARPON applies additional filtering steps not found in TeloBP or Telometer: the telomeric region must contain more than 60% telomere+1N repeats, and the 2 kb region upstream of the boundary must contain less than 10% telomere+1N. These filters eliminate reads with internal

sequence artifacts, such as the example in Fig. 5j. In this case, TeloBP and Telometer assign telomere lengths of 18 bp and 342 bp, respectively. In reality, the read contains over 3 kb of telomeric sequence, but since it is unclear if the erroneous stretch of basecalls is similar in length to the bona fide repeat sequence, the read is excluded from analysis in TARPON.

At the distal end, TARPON calculates telomere length up to the start of the capture probe. In contrast, if reads are not preprocessed before using TeloBP, telomere length estimates will be inflated due to the inclusion of non-telomeric sequence contributed by the capture probe or ONT sequencing adapters.

Telometer, TeloBP, and TARPON all return tabular output files (CSV or TSV) listing read IDs and corresponding telomere lengths. TARPON includes additional metadata such as the coordinates of the telomeric region within each read, strand specificity, and read-level quality metrics. It also generates a suite of visual outputs summarizing pipeline execution, filtering steps, and telomere length distributions. A precompiled HTML report—automatically launched in the EPI2ME GUI upon completion—provides an accessible, sample-by-sample overview. For multiplexed datasets, side-by-side comparisons are included by default.

Despite their workflow differences, all previously published analyses rely on alignment to a reference genome for chromosome arm assignment. This method works well for HG002 data aligned to an HG002-specific reference but yields poor accuracy for other cell lines or clinical datasets, where alignment to non-matching references leads to increased mapping bias and reduced reliability in chromosome arm–specific telomere length estimation.

Multiplexing and Cost Reduction

At the time of manuscript preparation, the cost to sequence a telomere enriched sample was 95€ for nanopore specific library preparation reagents (SQK-LSK114), 34€ for third party reagents (NEBNext Companion Module v2, E7672S), and 570€ for a R10.4.1 MinION FlowCell assuming purchase of a pack of 12 FlowCells, plus the cost of sample generation/collection/enrichment. This results in a minimum total cost per sample of 700€. Therefore, it is crucial that reliable multiplexing methods are developed for telomere sequencing to reduce the cost per sample without decreasing the number of telomeric reads per sample. The multiplexing protocols described previously (Sanchez et al. 2024; Schmidt et al. 2024; Karimian et al. 2024) allow for sample pooling after duplex or splint ligation prior to library preparation. Multiplexing introduces additional computational challenges, which TARPON already addresses. If a sample file is provided that contains unique barcode sequences (often found within the duplex sequence as in Karimian et al., 2024) together with a capture probe, the capture probe sequence is first used to identify the terminal end of the telomere. The following 100bp are then used to demultiplex the samples based on the barcodes found within the sample file. If no capture probe sequence is provided, but a sample file is, the barcodes found within that sample file will be used to both demultiplex the data and determine the end of the telomere. If only a capture probe is provided without a sample file, TARPON assumes the dataset was not multiplexed.

Applications of Nanopore Telomere Sequencing

Assessing telomere length distributions at nucleotide resolution opens new avenues for studying telomere dynamics in aging and senescent cell populations. Nanopore sequencing enables a clearer definition of the subtelomere-to-telomere

boundary and provides the opportunity to investigate the functional relevance of the variant repeat-rich (VRR) region. It also allows for high-resolution studies of cancer cells exhibiting recombination-based telomere maintenance or elevated telomere+1N repeat content, as well as the detailed characterization of telomerase template mutations.

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

Importantly, this approach is not restricted to human samples. TARPON can be used to study telomere length dynamics in non-human organisms with non-heterogeneous telomeric repeats, including invertebrates and plants. This is particularly valuable in species where few control samples exist, making FlowFISH inapplicable. However, we want to caution that parameters for boundary detection may require optimization on a species-specific basis. For organisms with highly heterogeneous telomeric repeats, such as *Schizosaccharomyces pombe* or *Saccharomyces cerevisiae*, TARPON is not currently recommended. In such cases, specialized tools like the soon to be released pombeTARPON, designed to account for sequence heterogeneity, are more appropriate.

Conclusion

Nanopore sequencing is a rapidly evolving technology that offers a unique opportunity to explore telomere-related questions previously inaccessible with Sanger, short-read, or PacBio sequencing platforms. While several tools exist for analyzing telomeric sequences from short-read whole-genome sequencing data, few address the distinct challenges posed by nanopore reads. Among those that do, most require bioinformatics 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.

Acknowledgement

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 all other members of the Baumann Laboratory for their insightful discussions. This work was funded in part by an

Alexander von Humboldt Professorship awarded to P.B. at JGU. We thank the Institute for Quantitative and Computational Biosciences, the Nucleic Acid Core Facility, and the Computational Systems Genetics Group for computing resources.

References

Allsopp, Richard C., and Calvin B. Harley. 1995. "Evidence for a Critical Telomere Length in Senescent Human Fibroblasts." *Experimental Cell Research* 219 (1): 130–36. https://doi.org/10.1006/excr.1995.1213.

Alter, Blanche P., Gabriela M. Baerlocher, Sharon A. Savage, Stephen J. Chanock, Babette B. Weksler, Judith P. Willner, June A. Peters, Neelam Giri, and Peter M. Lansdorp. 2007. "Very Short Telomere Length by Flow Fluorescence in Situ Hybridization Identifies Patients with Dyskeratosis Congenita." *Blood* 110 (5): 1439–47. https://doi.org/10.1182/blood-2007-02-075598.

Aubert, Geraldine, Mark Hills, and Peter M. Lansdorp. 2012. "Telomere Length Measurement—Caveats and a Critical Assessment of the Available Technologies and Tools." *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, Telomeres and disease, 730 (1): 59–67. https://doi.org/10.1016/j.mrfmmm.2011.04.003.

Baird, Duncan M., Jan Rowson, David Wynford-Thomas, and David Kipling. 2003. "Extensive Allelic Variation and Ultrashort Telomeres in Senescent Human Cells." *Nature Genetics* 33 (2): 203–7. https://doi.org/10.1038/ng1084.

Button, L. L., and C. R. Astell. 1986. "The Saccharomyces Cerevisiae Chromosome III Left Telomere Has a Type X, but Not a Type Y', ARS Region." *Molecular and Cellular Biology* 6 (4): 1352–56. https://doi.org/10.1128/mcb.6.4.1352-1356.1986.

Cawthon, Richard M. 2002. "Telomere Measurement by Quantitative PCR." *Nucleic Acids Research* 30 (10): e47. https://doi.org/10.1093/nar/30.10.e47.

Cunningham, Julie M., Ruth A. Johnson, Kristin Litzelman, Halcyon G. Skinner, Songwon Seo, Corinne D. Engelman, Russell J. Vanderboom, et al. 2013. "Telomere Length Varies by DNA Extraction Method: Implications for Epidemiologic Research." *Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 22 (11): 2047–54. https://doi.org/10.1158/1055-9965.EPI-13-0409.

Daniali, Lily, Athanase Benetos, Ezra Susser, Jeremy D. Kark, Carlos Labat, Masayuki Kimura, Kunj K. Desai, Mark Granick, and Abraham Aviv. 2013. "Telomeres Shorten at Equivalent Rates in Somatic Tissues of Adults." *Nature Communications* 4 (1): 1597. https://doi.org/10.1038/ncomms2602.

Demanelis, Kathryn, Farzana Jasmine, Lin S. Chen, Meytal Chernoff, Lin Tong, Dayana Delgado, Chenan Zhang, et al. 2020. "Determinants of Telomere Length across Human Tissues." *Science (New York, N.Y.)* 369 (6509): eaaz6876. https://doi.org/10.1126/science.aaz6876.

Greider, C. W., and E. H. Blackburn. 1985. "Identification of a Specific Telomere Terminal Transferase Activity in Tetrahymena Extracts." *Cell* 43 (2 Pt 1): 405–13. https://doi.org/10.1016/0092-8674(85)90170-9.

Grill, Sherilyn, and Jayakrishnan Nandakumar. 2021. "Molecular Mechanisms of Telomere Biology Disorders." *Journal of Biological Chemistry* 296 (January). https://doi.org/10.1074/jbc.REV120.014017.

Hinchie, Angela M., Samantha L. Sanford, Kelly E. Loughridge, Rachel M. Sutton, Anishka H. Parikh, Agustin A. Gil Silva, Daniel I. Sullivan, et al. 2024. "A Persistent Variant Telomere Sequence in a Human Pedigree." *Nature Communications* 15 (1): 4681. https://doi.org/10.1038/s41467-024-49072-9.

Karimian, Kayarash, Aljona Groot, Vienna Huso, Ramin Kahidi, Kar-Tong Tan, Samantha Sholes, Rebecca Keener, et al. 2024. “Human Telomere Length Is Chromosome End–Specific and Conserved across Individuals.” *Science* 384 (6695): 533–39. https://doi.org/10.1126/science.ado0431.

Koppelstaetter, Christian, Paul Jennings, Kathrin Hochegger, Paul Perco, Rudolf Ischia, Henryk Karkoszka, and Gert Mayer. 2005. “Effect of Tissue Fixatives on Telomere Length Determination by Quantitative PCR.” *Mechanisms of Ageing and Development* 126 (12): 1331–33. https://doi.org/10.1016/j.mad.2005.08.003.

Lingner, Joachim, Timothy R. Hughes, Andrej Shevchenko, Matthias Mann, Victoria Lundblad, and Thomas R. Cech. 1997. “Reverse Transcriptase Motifs in the Catalytic Subunit of Telomerase.” *Science* 276 (5312): 561–67. https://doi.org/10.1126/science.276.5312.561.

Makarov, Vladimir L, Yoko Hirose, and John P Langmore. 1997. “Long G Tails at Both Ends of Human Chromosomes Suggest a C Strand Degradation Mechanism for Telomere Shortening.” *Cell* 88 (5): 657–66. https://doi.org/10.1016/S0092-8674(00)81908-X.

Martin-Ruiz, Carmen M, Duncan Baird, Laureline Roger, Petra Boukamp, Damir Krunic, Richard Cawthon, Martin M Dokter, et al. 2015. “Reproducibility of Telomere Length Assessment: An International Collaborative Study.” *International Journal of Epidemiology* 44 (5): 1673–83. https://doi.org/10.1093/ije/dyu191.

McElligott, Richard, and Raymund J. Wellinger. 1997. “The Terminal DNA Structure of Mammalian Chromosomes.” *The EMBO Journal* 16 (12): 3705–14. https://doi.org/10.1093/emboj/16.12.3705.

Moyzis, R. K., J. M. Buckingham, L. S. Cram, M. Dani, L. L. Deaven, M. D. Jones, J. Meyne, R. L. Ratliff, and J. R. Wu. 1988. “A Highly Conserved Repetitive DNA Sequence, (TTAGGG)n, Present at the Telomeres of Human Chromosomes.” *Proceedings of the National Academy of Sciences of the United States of America* 85 (18): 6622–26. https://doi.org/10.1073/pnas.85.18.6622.

Njajou, Omer T., Wen-Chi Hsueh, Elizabeth H. Blackburn, Anne B. Newman, Shih-Hsuan Wu, Rongling Li, Eleanor M. Simonsick, Tamara M. Harris, Steve R. Cummings, and Richard M. Cawthon. 2009. “Association Between Telomere Length, Specific Causes of Death, and Years of Healthy Life in Health, Aging, and Body Composition, a Population-Based Cohort Study.” *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 64A (8): 860–64. https://doi.org/10.1093/gerona/glp061.

O’Donnell, Samuel, Jia-Xing Yue, Omar Abou Saada, Nicolas Agier, Claudia Caradec, Thomas Cokelaer, Matteo De Chiara, et al. 2023. “Telomere-to-Telomere Assemblies of 142 Strains Characterize the Genome Structural Landscape in Saccharomyces Cerevisiae.” *Nature Genetics* 55 (8): 1390–99. https://doi.org/10.1038/s41588-023-01459-y.

Olovnikov, A. M. 1973. “A Theory of Marginotomy. The Incomplete Copying of Template Margin in Enzymic Synthesis of Polynucleotides and Biological Significance of the Phenomenon.” *Journal of Theoretical Biology* 41 (1): 181–90. https://doi.org/10.1016/0022-5193(73)90198-7.

Rautiainen, Mikko, Sergey Nurk, Brian P. Walenz, Glennis A. Logsdon, David Porubsky, Arang Rhie, Evan E. Eichler, Adam M. Phillippy, and Sergey Koren. 2023. “Telomere-to-Telomere Assembly of Diploid Chromosomes with Verkko.” *Nature Biotechnology* 41 (10): 1474–82. https://doi.org/10.1038/s41587-023-01662-6.

Revy, Patrick, Caroline Kannengiesser, and Alison A. Bertuch. 2023. “Genetics of Human Telomere Biology Disorders.” *Nature Reviews. Genetics* 24 (2): 86–108. https://doi.org/10.1038/s41576-022-00527-z.

Rossiello, Francesca, Diana Jurk, João F. Passos, and Fabrizio d’Adda di Fagagna. 2022. “Telomere Dysfunction in Ageing and Age-Related Diseases.” *Nature Cell Biology* 24 (2): 135–47. https://doi.org/10.1038/s41556-022-00842-x.

Rufer, Nathalie, Tim H. Brümmendorf, Steen Kolvraa, Claus Bischoff, Kaare Christensen, Louis Wadsworth, Michael Schulzer, and Peter M. Lansdorp. 1999. “Telomere Fluorescence Measurements in Granulocytes and T Lymphocyte Subsets Point to a High Turnover of Hematopoietic Stem Cells and Memory T Cells in Early Childhood.” *Journal of Experimental Medicine* 190 (2): 157–68. https://doi.org/10.1084/jem.190.2.157.

Sanchez, Santiago E., Yuchao Gu, Yan Wang, Anudeep Golla, Annika Martin, William Shomali, Dirk Hockemeyer, Sharon A. Savage, and Steven E. Artandi. 2024. “Digital Telomere Measurement by Long-Read Sequencing Distinguishes Healthy Aging from Disease.” *Nature Communications* 15 (1): 5148. https://doi.org/10.1038/s41467-024-49007-4.

Savage, Sharon A. 2022. “Dyskeratosis Congenita and Telomere Biology Disorders.” *Hematology. American Society of Hematology. Education Program* 2022 (1): 637–48. https://doi.org/10.1182/hematology.2022000394.

Schmidt, Tobias T., Carly Tyer, Preeyesh Rughani, Candy Haggblom, Jeffrey R. Jones, Xiaoguang Dai, Kelly A. Frazer, et al. 2024. “High Resolution Long-Read Telomere Sequencing Reveals Dynamic Mechanisms in Aging and Cancer.” *Nature Communications* 15 (1): 5149. https://doi.org/10.1038/s41467-024-48917-7.

Sfeir, Agnel J., Weihang Chai, Jerry W. Shay, and Woodring E. Wright. 2005. “Telomere-End Processing.” *Molecular Cell* 18 (1): 131–38. https://doi.org/10.1016/j.molcel.2005.02.035.

Sholes, Samantha L., Kayarash Karimian, Ariel Gershman, Thomas J. Kelly, Winston Timp, and Carol W. Greider. 2021. “Chromosome-Specific Telomere Lengths and the Minimal Functional Telomere Revealed by Nanopore Sequencing.” *Genome Research*, October, gr.275868.121. https://doi.org/10.1101/gr.275868.121.

Southern, E. M. 1979. “Measurement of DNA Length by Gel Electrophoresis.” *Analytical Biochemistry* 100 (2): 319–23. https://doi.org/10.1016/0003-2697(79)90235-5.

Steinert, Susanne, Jerry W. Shay, and Woodring E. Wright. 2004. “Modification of Subtelomeric DNA.” *Molecular and Cellular Biology* 24 (10): 4571–80. https://doi.org/10.1128/MCB.24.10.4571-4580.2004.

Sun, Guozhu, Hui Cao, Yang Bai, Jueheng Wang, Yuxun Zhou, Kai Li, and Jun-hua Xiao. 2021. “A Novel Multiplex qPCR Method for Assessing the Comparative Lengths of Telomeres.” *Journal of Clinical Laboratory Analysis* 35 (9): e23929. https://doi.org/10.1002/jcla.23929.

Teplitz, Gabriela M., Emeline Pasquier, Erin Bonnell, Evelina De Laurentiis, Louise Bartle, Jean-François Lucier, Samantha Sholes, Carol W. Greider, and Raymund J. Wellinger. 2024. “A Mechanism for Telomere-Specific Telomere Length Regulation.” bioRxiv. https://doi.org/10.1101/2024.06.12.598646.

Tham, Cheng-Yong, LaiFong Poon, TingDong Yan, Javier Yu Peng Koh, Muhammad Khairul Ramlee, Vania Swee Imm Teoh, Suihan Zhang, et al. 2023. “High-Throughput Telomere Length Measurement at Nucleotide Resolution Using the PacBio High Fidelity Sequencing Platform.” *Nature Communications* 14 (1): 281. https://doi.org/10.1038/s41467-023-35823-7.

Trujillo, Kelly M., Jeremy T. Bunch, and Peter Baumann. 2005. “Extended DNA Binding Site in Pot1 Broadens Sequence Specificity to Allow Recognition of Heterogeneous Fission Yeast Telomeres.” *The Journal of Biological Chemistry* 280 (10): 9119–28. https://doi.org/10.1074/jbc.M414511200.

Watson, J. D. 1972. “Origin of Concatemeric T7 DNA.” *Nature: New Biology* 239 (94): 197–201. https://doi.org/10.1038/newbio239197a0.

Yu, Hyeon Jong, Yoon Hwan Byun, and Chul-Kee Park. 2024. “Techniques for Assessing Telomere Length: A Methodological Review.” *Computational and Structural Biotechnology Journal* 23 (April):1489–98. https://doi.org/10.1016/j.csbj.2024.04.011.

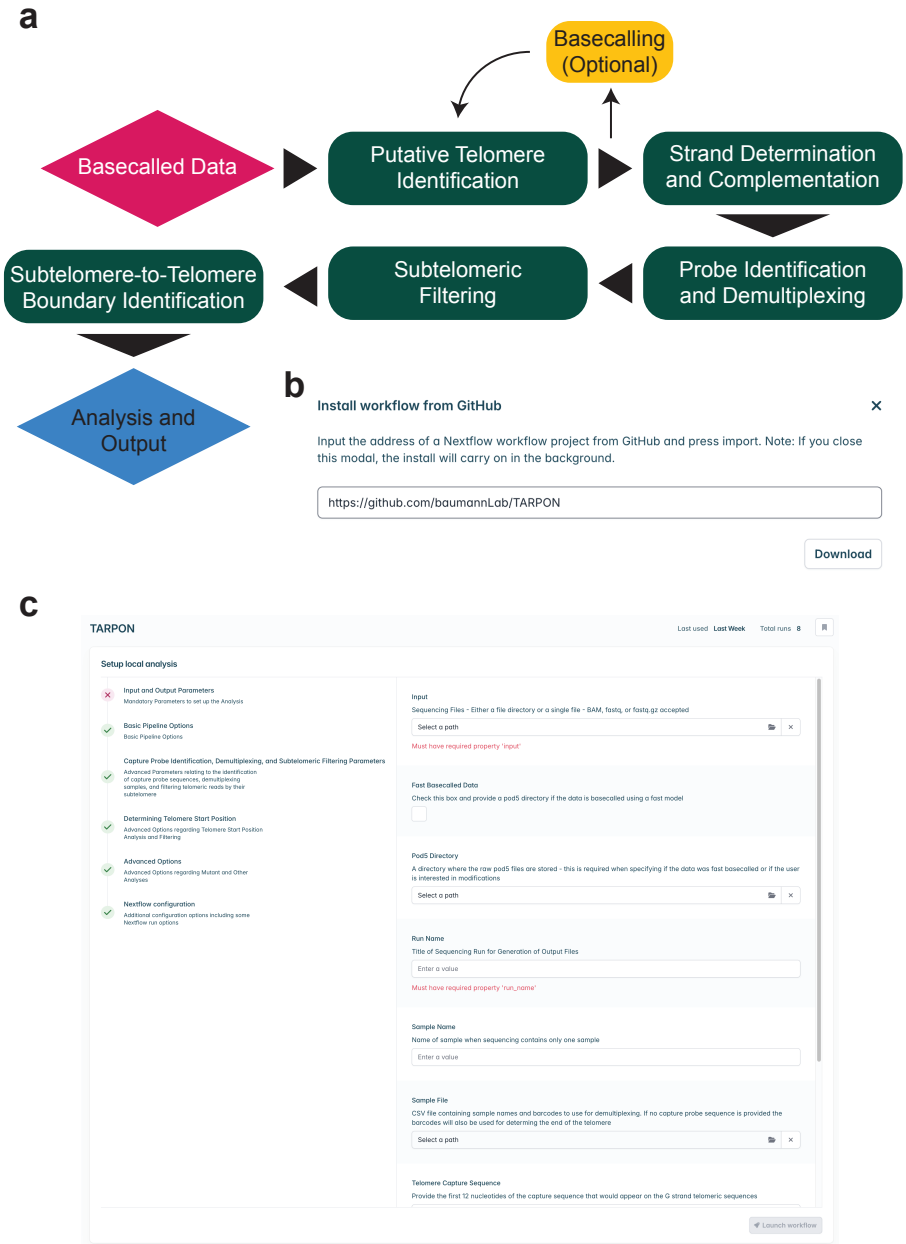

Figure 1. Pipeline Processes and Implementation into EPI2ME. (a) Graphical display of the pipeline workflow and strategy. (b) To install TARPON into EPI2ME the user can input the URL of the TARPON GitHub repository into a pop-up window that appears after clicking “Import Workflow”. (c) Once added to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

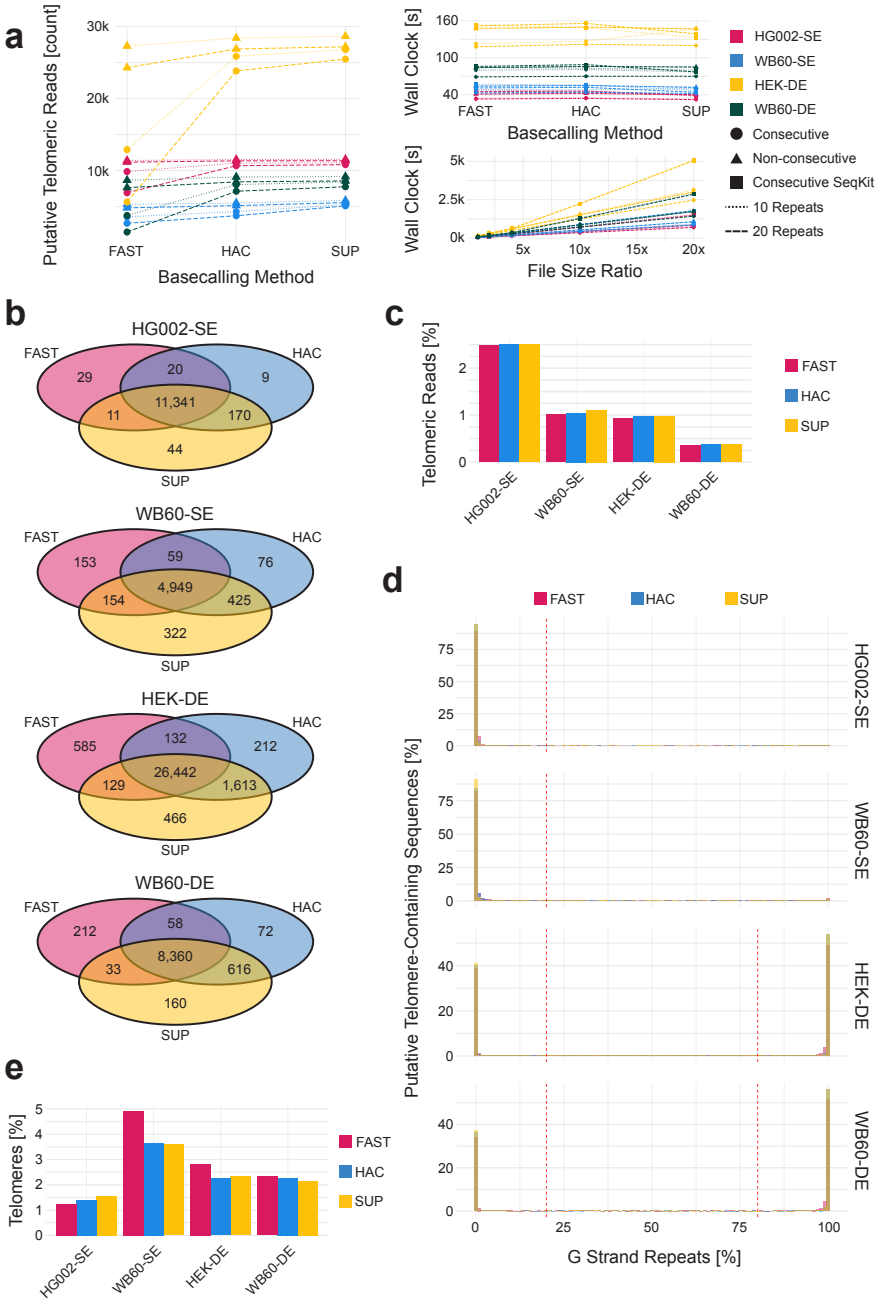

Figure 2. Putative Read Isolation and Chimera Filtering (a) Number of putative telomeric reads isolated using fast, high accuracy (HAC), or super high accuracy (SUP) basecalling models across the different isolation methods (left), the speed at which these methods perform (top right), and the speed at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

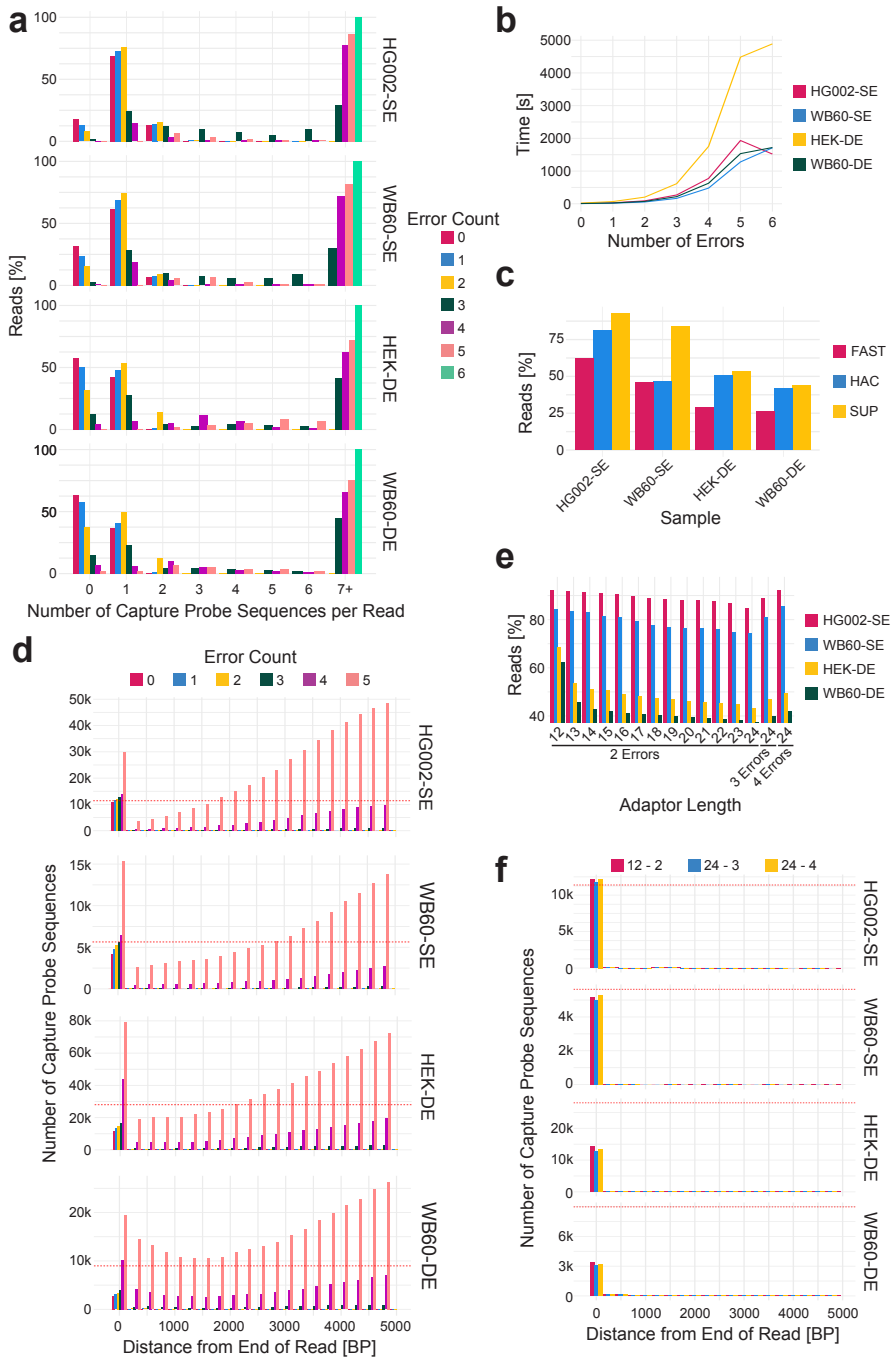

Figure 3. Identifying the end of a telomeric sequence using an enrichment technique specific capture probe (a) Number of capture probe sequences found in a read increasing the number of allowed errors within the capture probe sequence. (b) Change in wall-clock time associated with increasing the number of allowed errors within the capture probe sequence. (c) Percentage of reads in which a capture probe sequence is successfully found using a 12 nucleotide capture probe sequence with two errors. (d) Distance between the identified capture probe sequence and the end of the sequence in 200bp bins. (e) Percentage of reads where the capture probe sequence was successfully identified as the capture probe sequence increases in length. (f) Distance between the capture probe sequence and the end of the sequence in 200bp bins while increasing the length of the capture probe sequence and number of allowed errors.

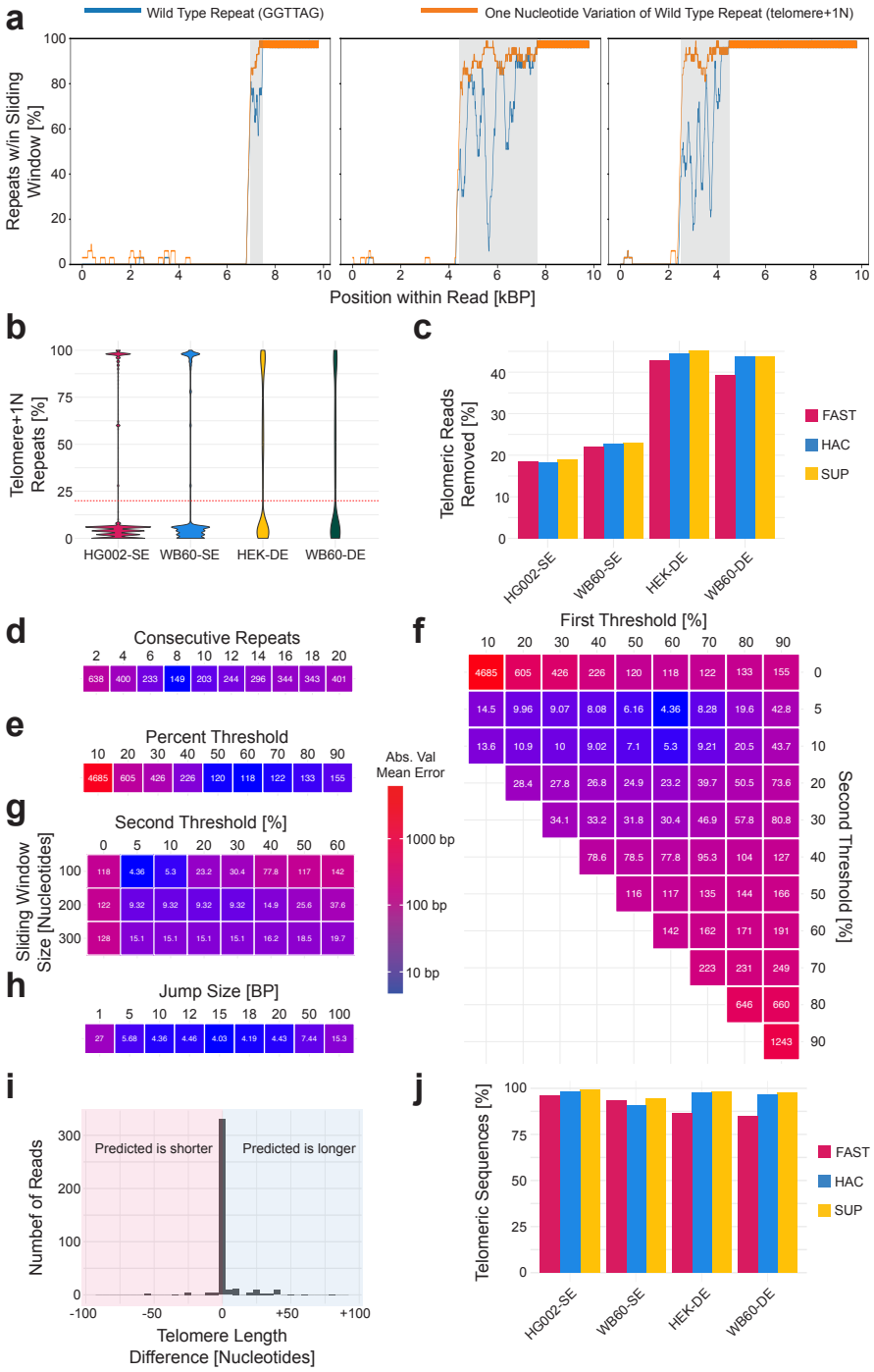

Figure 4. Identifying the Subtelomere-to-Telomere Boundary (a) Three telomeric reads with the percentage of telomere repeats in a 100bp sliding window (blue lines) and the percentage of telomere+1N repeats in the same window (orange lines). The variant repeat rich (VRR) region is highlighted in gray. (b) Percentage of telomere+1N repeats in the first 300bp of a read with a 20% threshold indicated by a dotted red line. (c) Percentage of telomeric reads removed for containing greater than 20% telomere+1N repeats in the first 300bp of the sequence. (d) Mean absolute value difference in predicted versus manually annotated start of the VRR-region when identifying the start of the VRR-region by the first occurrence of consecutive telomere+1N repeats, (e) when identifying the start of the VRR-region by a 100bp jumping window with a 10bp jump and a minimum threshold of telomere+1N repeats that must be maintained for the remainder of the read, and (f) when identifying the start of the VRR-region by a 100bp jumping window with a 10bp jump and a minimum threshold of telomere+1N repeats for that must be exceeded and not decrease past the second

threshold. (g) Mean absolute value difference in predicted versus manually annotated start of the VRR-region when increasing the sliding window size and (h) increasing the interval size to decrease computational time when a 60%/5% threshold system are used with a sliding window size of 100bp. (i) Differences between manually annotated start sites and computationally identified start site. (j) Percentage of telomeric sequences where the VRR-region start site was determined.

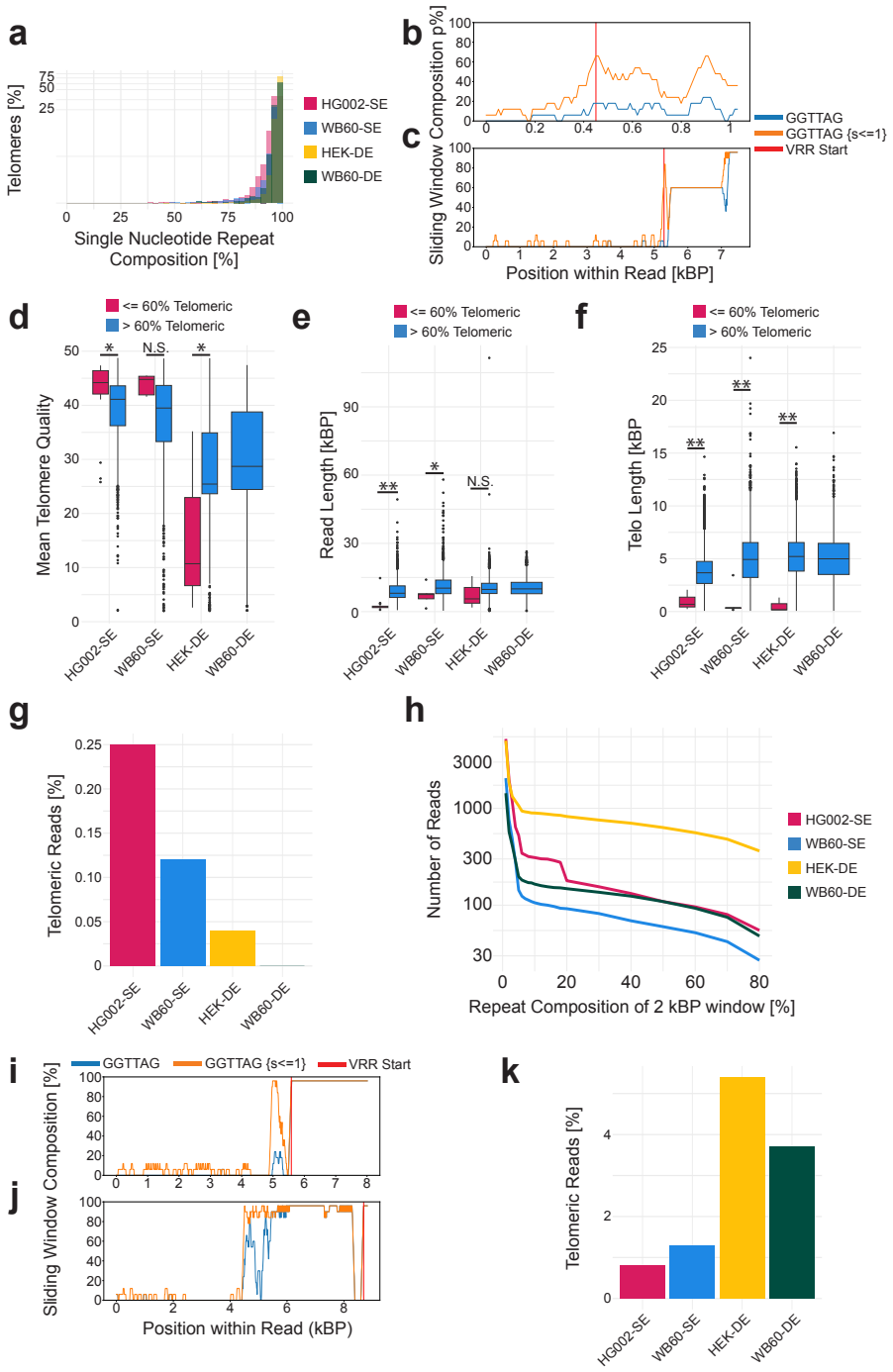

Figure 5. Filtering of Telomeric Sequences. (a) Percentage of telomere composed of telomeric+1N repeats. (b) An example non-telomeric sequence passing all filtering criteria that is erroneous and (c) an example telomeric sequence passing all filtering criteria that is not erroneous and should be retained. (d) Mean telomere quality of sequences in which the telomeres are composed of greater than or less than 60% telomere+1N repeats. (e) Read length of filtered telomeric sequences and (f) telomere length of filtered sequences. (g) Percentage of filtered reads

for a telomere+1N composition below 60%. (h) Percentage of telomere+1N repeats in the 2kb immediately prior to the VRR-region start site in the subtelomeric direction. (i) A high telomere+1N percentage in prior 2kb read that should not be removed from the analysis and (j) a high percentage read that should be removed from the analysis. (k) Percentage of telomeric sequences removed for containing a high proportion of telomere+1N repeats before the VRR-region start site. (* = p.val < 0.05, ** = p.val < 0.005)

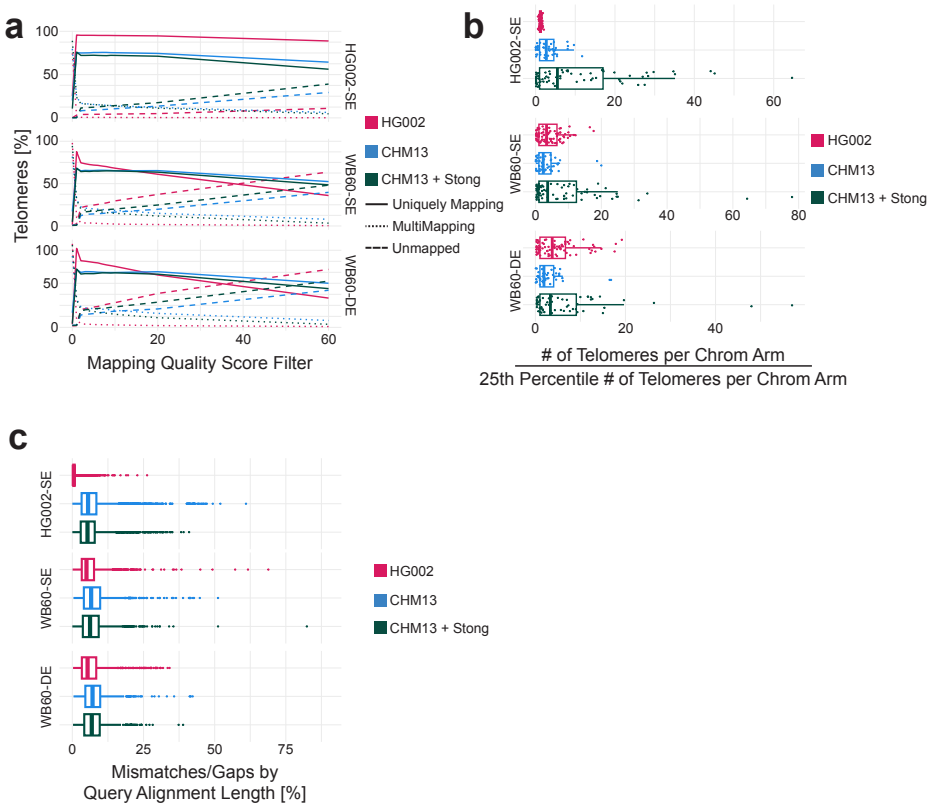

Figure 6. Chromosome Arm Specific Telomere Length. (a) Mapping statistics of 3 different datasets to 3 different reference genomes. (b) Distribution of telomeric sequences across chromosome arms when aligning to different reference genomes. (c) Alignment mismatches and gaps within alignments to different references.