1 **MANUSCRIPT TITLE**

2 Telomeric repeat evolution in the phylum Nematoda revealed by high-quality genome
3 assemblies and subtelomere structures

4

5 **AUTHORS**

6 Jiseon Lim[1,2], Wonjoo Kim[1,2], Jun Kim[1,3,4,*], Junho Lee[1,2,3,*]

7 [1] Department of Biological Sciences, Seoul National University, Gwanak-ro 1, Gwanak-gu,
8 Seoul 08826, Korea

9 [2] Institute of Molecular Biology and Genetics, Seoul National University, Seoul 08826, Korea

10 [3] Research Institute of Basic Sciences, Seoul National University, Seoul 08826, Korea

11 [4] Department of Convergent Bioscience and Informatics, College of Bioscience and
12 Biotechnology, Chungnam National University, Daehak-ro 99, Daejeon 34134, Republic of
13 Korea

14 * Corresponding authors: elegans@snu.ac.kr and junkim@cnu.ac.kr

15

16 A running title: Telomeric repeat evolution in Nematoda

17

ABSTRACT

Telomeres are composed of tandem arrays of telomeric-repeat motifs (TRMs) and telomere-binding proteins (TBPs), which are responsible for ensuring end-protection and end-replication of chromosomes. TRMs are highly conserved due to the sequence specificity of TBPs, although significant alterations in TRM have been observed in several taxa, except Nematoda. We used public whole-genome sequencing datasets to analyze putative TRMs of 100 nematode species and determined that three distinct branches included specific novel TRMs, suggesting that evolutionary alterations in TRMs occurred in Nematoda. We focused on one of the three branches, the Panagrolaimidae family, and performed a *de novo* assembly of four high-quality draft genomes of the canonical (TTAGGC) and novel TRM (TTAGAC) isolates; the latter genomes revealed densely clustered arrays of the novel TRM. We then comprehensively analyzed the subtelomeric regions of the genomes to infer how the novel TRM evolved. We identified DNA damage–repair signatures in subtelomeric sequences that were representative of consequences of telomere maintenance mechanisms by alternative lengthening of telomeres. We propose a hypothetical scenario in which TTAGAC-containing units are clustered in subtelomeric regions and pre-existing TBPs capable of binding both canonical and novel TRMs aided the evolution of the novel TRM in the Panagrolaimidae family.

INTRODUCTION

Since linear chromosomes originated from a circular structure, a specific DNA–protein complex has evolved to protect the ends of the chromosomes, i.e. the telomere. Telomeres are typically composed of tandem arrays of telomeric-repeat motifs (TRMs) and telomere-binding proteins (TBPs) that can bind to the tandem arrays (Zhong et al. 1992). TBP binding interferes with the accession of other proteins. Hence, proteins that respond to DNA damage signals cannot bind to the exposed ends of linear chromosomes and the exposed ends are not recognized as DNA damage sites (Van Steensel et al. 1998; De Lange 2009; Lazzerini-Denchi and Sfeir 2016). This enables the telomere to maintain chromosome integrity.

TRM sequences are typically well conserved because TBPs attach to arrays of TRMs in a sequence-specific manner. However, dozens of variant TRMs have evolved in many taxa, including ≥150 motifs in fungi (Červenák et al. 2021) and ≥7 motifs in plants (Peska and Garcia 2020). Furthermore, except Hymenoptera, many arthropod species have unique telomere structures. For example, most Insecta species have telomere structures composed of telomere-specific retrotransposons interposed between canonical TRMs, whereas Diptera species have telomeres that are exclusively composed of retrotransposons and lack TRMs (Abad et al. 2004; Garavís et al. 2013; Zhou et al. 2022). It is yet to be elucidated how TRMs have been altered while interacting with TBPs. These diverse TRMs and telomere structures developed from the co-evolution of a TRM and several TBPs (Shakirov et al. 2009; Steinberg-Neifach and Lue 2015; Sepsiova et al. 2016; Červenák et al. 2019); however, they may have also originated from TRM-specific changes regardless of TBP changes if TBPs were capable of binding to several types of TRMs (Rotková et al. 2004; Fajkus et al. 2005; Kramara et al. 2010; Visacka et al. 2012; Červenák et al. 2019; Tomáška et al. 2019; Červenák et al. 2021).

We focused our analysis on subtelomeric regions and the DNA damage–repair signatures generated during telomere evolution, especially telomere and subtelomere reconstruction by alternative lengthening of telomeres (ALT) mechanisms (Heaphy et al. 2011; Sobinoff and Pickett 2017). ALT, which refers to alternative mechanisms that substitute telomerase function for telomere maintenance at the chromosome ends (Bryan et al. 1995), is crucial in telomerase mutant eukaryotic cells such as those in yeast, *Caenorhabditis elegans*, mouse, and human, and can thus be used to model and interpret the evolution of TRM changes (Lundblad and Blackburn 1993; Bryan et al. 1995; Seo et al. 2015; Kim et al. 2020; Kim et al. 2021a; Kim et al. 2021b). In these ALT models, homology-directed repair (HDR) mechanisms can add new telomeric components. HDRs exploit the homology between shortened telomeric sequences and tandem arrays of TRMs in other genomic loci to replicate the tandem array and adjacent sequences to the shorter end (Lydeard et al. 2007; Roumelioti et al. 2016; Kramara et al. 2018; Kim et al. 2020). Furthermore, a unique sequence inserted between TRMs, known as the template for ALT, may be used to repair shorter telomeric repeats even in wild-type *C. elegans*, resulting in the formation of a new subtelomeric structure (Kim et al. 2019a). These findings suggest that ALT mechanisms play an essential role in the evolution of telomeres and the evolutionary route can be reconstructed by studying ALT activity traces left in subtelomeric sequences.

The present study explored TRM evolution in Nematoda using public whole-genome sequencing (WGS) data obtained from 100 nematode species and our high-quality genome assemblies. We evaluated the subtelomeric regions of TTAGGC-telomere isolates of Panagrolaimidae to infer possible evolutionary paths toward TTAGAC telomeres.

RESULTS

Identification of three novel TRMs in Nematoda

We examined variations in TRMs using public short-read WGS data from Nematoda species that have matched public genome assemblies from all five major clades. A total of 100 species met the criteria, including at least 17 species from Clades I, III, IV, and V but only one from Clade II (Fig. 1A). The species and accession numbers used in this analysis are listed in Supplemental Table S1 and S2.

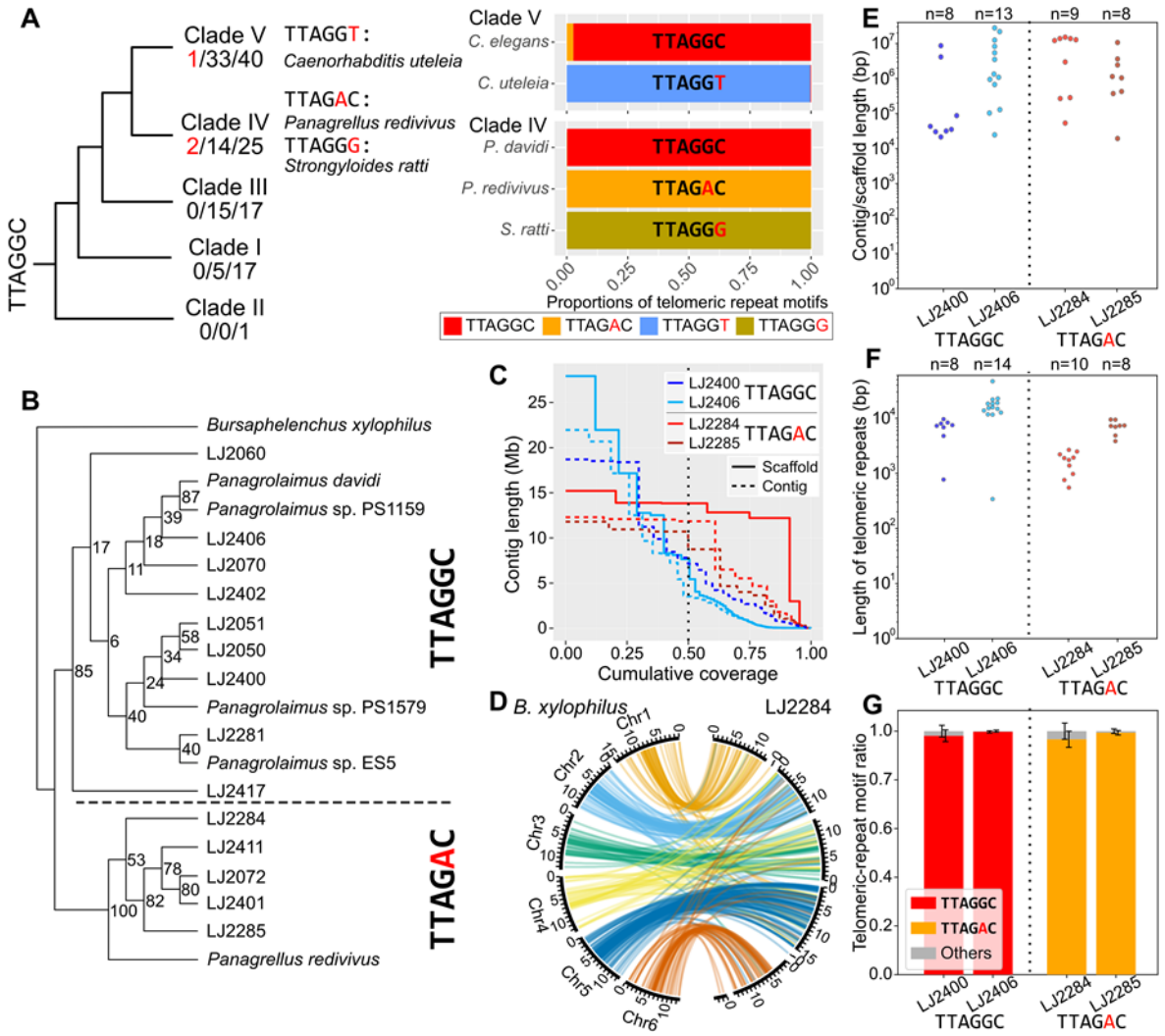


Figure 1. Putative novel TRMs in the phylum Nematoda and validation of the emergence of TTAGAC in the family Panagrolaimidae. (A) The cladogram was adapted and modified from Smythe et al. 2019 (Smythe et al. 2019) with permission from BMC Ecology and Evolution 2019, CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). The three numbers separated by '/' under each clade indicates the number of species with novel TRMs/the number of species whose TRM was identified in our analysis/total number of species we analyzed, respectively. Changes from TTAGGC, the canonical Nematoda TRM, to the novel

TRMs are highlighted in red. Below each novel TRM is a species name with the corresponding TRM. Each bar indicates the proportion of the canonical Nematoda TRM (TTAGGC) and novel TRMs in each species. The upper bar of each pair represents a control species that harbors the canonical TRM. The TRMs are represented as follows: red, TTAGGC; orange, TTAGAC; blue, TTAGGT; khaki, TTAGGG. (B) Phylogenetic relationships of 19 Panagrolaimidae species/isolates based on their 18S rDNA sequences and their putative TRMs. *Bursaphelenchus xylophilus* was used as an outgroup species. The number on each node represents the bootstrap support value. (C) Contig/scaffold length distributions of the genome assemblies. The vertical dotted line indicates N50 contig/scaffold size. (D) Synteny plot of LJ2284 compared to *B. xylophilus*. Each colored line represents a BUSCO gene shared among the genome assemblies of our four isolates and *B. xylophilus*. Orange lines indicate that corresponding BUSCO genes are in Chromosome 1 in *B. xylophilus*. Sky blue, bluish green, yellow, blue, and vermillion represent Chromosomes 2, 3, 4, 5, and 6, respectively. We used Wong's color palette designed for color-blind individuals (Wong 2011). (E–G) The vertical dotted line separates the TTAGGC-telomere isolates (left) and the TTAGAC-telomere isolates (right). (E) Length distributions of the contigs/scaffolds containing highly clustered telomeric repeats at the end. Each dot represents each contig/scaffold, and the total number of contigs/scaffolds of each isolate (n) is indicated at the top of the graph. (F) Estimated length distributions of clustered telomeric repeats at the end of the contigs/scaffolds. Each dot indicates a telomeric-repeat cluster for each contig/scaffold, and the total number of telomeric-repeat clusters for each isolate (n) is indicated at the top of the graph. (G) The

proportion of TRM types in clustered telomeric repeats. Error bars represent the standard deviation for all clustered telomeric repeats at the end of contigs/scaffolds in each isolate.

The canonical TRM of Nematoda is known to be a 6-bp sequence (TTAGGC) that covers a >1-kb length in *C. elegans* telomeres (Kim et al. 2019a; Yoshimura et al. 2019). We examined the WGS data from 100 species to identify novel TRMs with a concatemer of 5–7 bp, a high copy number, and a pattern similar to TTAGGC. We determined the TRM in 67 species using the following criterion: TRM repeats should be long concatemers for TBP binding, so TRM-like *k*-mer counts should be high enough (≥100 in 5 million short reads in our analysis) (Fig. 1A and Supplemental Table S1). However, 26 of the remaining 33 species exhibited low *k*-mer counts of TRM-like motifs, so we could not determine their TRMs (Supplemental Table S2). The remaining seven species had high *k*-mer counts, but we excluded them from the TRM-determined group because of the following reasons: we detected potential sample or host contamination in *Enoplus brevis*, two *Strongyloides*, and three *Trichinella* species, and *Diploscapter pachys* had only short concatemers (≤ 10 copies) in the middle of its reads (Supplemental Table S2; see Supplemental Note for more information).

TTAGGC, the canonical Nematoda TRM, was identified in 64 of the 67 TRM-determined species (Fig. 1A and Supplemental Table S1). We determined that the other three nematode species had few TTAGGC copies but a high copy number of one of three putative TRMs: TTAGGT, TTAGAC, and TTAGGG in *Caenorhabditis uteleia*, *Panagrellus redivivus*, and

Strongyloides ratti, respectively (Fig. 1A, Supplemental Fig. S1 and Supplemental Tables S1 and S3; see Supplemental Note and Supplemental Table S4 for further validation of the TRMs). The identified TRMs were new to Nematoda and had not been reported previously. We assumed that these three motifs evolved independently because they occurred in phylogenetically distant branches.

The emergence of a new TRM, TTAGAC, in the family Panagrolaimidae

We focused on the TTAGAC sequence of the family Panagrolaimidae to confirm the noncanonical TRM sequence identified by short-read WGS data because many different isolates of this family were widely isolated from regions across the Republic of Korea. Some were easily cultured under laboratory conditions. We used 19 short-read WGS datasets of the family Panagrolaimidae to determine whether the unique motif TTAGAC emerged once or multiple times (Supplemental Table S5). WGS datasets of five species were available from publicly accessible databases, whereas the remainder were derived from sequencing data of the isolates collected from the Republic of Korea, which were probably distinct species as they had distinguishable ribosomal DNA (rDNA) sequences (Supplemental Table S6). *B. xylophilus* was also used as the outgroup species.

We determined that *Panagrellus redivivus* and five out of fourteen collected isolates had the novel TRM, TTAGAC, whereas all four *Panagrolaimus* species and the remaining nine isolates possessed the canonical TRM, TTAGGC. Moreover, the five collected TTAGAC-TRM isolates clustered with *Panagrellus redivivus*; however, the nine TTAGGC-TRM isolates were grouped with the *Panagrolaimus* species (Fig. 1B and Supplemental Table S5). Furthermore, the

TTAGGC concatemer was not observed in any of the 6 TTAGAC-TRM species/isolates, nor was the TTAGAC concatemer found in any of the 13 TTAGGC-TRM species/isolates. These results suggest that TTAGAC sequences in Panagrolaimidae may have arisen from a single ancestor that had TTAGAC as its TRM.

High-quality genome assemblies confirmed that TTAGAC repeats were clustered at the ends of contigs or pseudo-chromosome molecules in TTAGAC-TRM isolates.

We constructed *de novo* genome assemblies of Panagrolaimidae isolates using PacBio high-fidelity (HiFi) long-read sequencing technology and Arima Hi-C technology to assess whether the newly identified TRM presented clusters at the ends of the chromosomes. Two TTAGAC-TRM isolates, LJ2284 and LJ2285, and two TTAGGC-TRM isolates, LJ2400 and LJ2406, were selected for the study. We produced 3.1 Gb (43×), 2.9 Gb (55×), 17.4 Gb (26×), and 36.1 Gb (93×) HiFi reads for LJ2284, LJ2285, LJ2400, and LJ2406, respectively. The HiFi reads were then assembled into high-quality draft genomes in which the length of the longest contig and contig N50 varied from 11.80 Mb to 21.96 Mb and 3.28 Mb to 11.84 Mb, respectively. We also generated Hi-C data for LJ2284 and LJ2406 and scaffolded their contigs into larger chunks (Supplemental Fig. S2 and Supplemental Table S7). The longest scaffold lengths increased to 15.2 Mb and 27.9 Mb, and their scaffold N50 increased to 13.8 Mb and 7.6 Mb for LJ2284 and LJ2406, respectively (Fig. 1C and Supplemental Table S8). Additionally, 91.4% of the LJ2284 contigs were connected as five scaffolds (Supplemental Fig. S2).

All of our genome assemblies had ~75% BUSCO completeness, which was comparable to that of the nearly complete genome assembly of *B. xylophilus* (71%) (Supplemental Fig. S3)

(Dayi et al. 2020). In addition, we analyzed the synteny relationships between *B. xylophilus* and our genome assemblies using their common single-copy orthologs. Pseudo-chromosome-level scaffolds of LJ2284 and contigs of LJ2285 exhibited marked collinearity for all the chromosomes, except Chromosome 4 of *B. xylophilus* (Fig. 1D, Supplemental Fig. S4 and Supplemental Note). Chromosomes 5 and 6 of *B. xylophilus* also showed collinearity with clusters of LJ2400 contigs and LJ2406 scaffolds, but the collinearity of other chromosomes was much weaker (Supplemental Fig. S4 and Supplemental Note). Based on these findings, our Panagrolaimidae genome assemblies were highly contiguous.

The constructed high-quality genome assemblies revealed that the telomeric repeats of both LJ2284 and LJ2285 changed from TTAGGC to TTAGAC. First, we confirmed whether the telomeric repeats had been adequately assembled. Our analysis revealed that each genome assembly had 8–13 contigs/scaffolds ending with telomeric repeats (Fig. 1E). Among these contigs/scaffolds, one each of LJ2284 and LJ2406 was assembled at the telomere-to-telomere level (Supplemental Tables S7 and S8). The telomeric repeats at the ends of the contigs/scaffolds in each genome assembly exhibited various ranges of length in the Panagrolaimidae isolates (Fig. 1F and Supplemental Table S9). Their lengths in LJ2284 ranged from 0.5 to 2.7 kb, but they were much longer in LJ2285 and LJ2400, ranging from 3.8 to 9.7 kb (except for one in LJ2400). LJ2406 exhibited telomeric-repeat clusters longer than 10 kb (except for one cluster). We identified that 84.9%–99.4% of raw HiFi reads were longer than 10 kb in all four isolates and that LJ2406, which has much longer telomeres than the other three isolates, had 328 reads consisting only of 10–26-kb telomeric repeats (Supplemental Fig. S5). Therefore, the TRM cluster lengths at each contig/scaffold end would be highly

206 correlated with their actual lengths (see Supplemental Note for detailed description). The
207 lengths of the clustered TRMs were comparable to those of the two *C. elegans* assemblies
208 constructed using long-read sequencing technologies (2.3–5.7 kb) (Kim et al. 2019a;
209 Yoshimura et al. 2019). Furthermore, the telomeric repeats consisted mainly of TTAGAC in
210 LJ2284 (96.67%) and LJ2285 (99.35%) and of TTAGGC in LJ2400 (98.11%) and LJ2406
211 (99.73%) (Fig. 1G, Supplemental Fig. S6 and Supplemental Table S9). These findings
212 supported our TRM study employing short-read WGS data and revealed that the TTAGAC
213 telomere had evolved in the family Panagrolaimidae.

214 **TTAGAC-containing units clustered near telomeric regions of TTAGGC-telomere**
215 **isolates**

216 Because most of the nematode species that we analyzed had a canonical Nematoda TRM
217 sequence, TTAGGC, the TTAGAC telomere probably evolved from its ancestral TTAGGC
218 telomere via a single nucleotide mutation of the telomerase RNA component gene. We
219 hypothesized that subtelomeric sequences can provide evidence regarding the evolutionary
220 mechanism because these can carry historical records of telomere dynamics. For example, it
221 has previously been documented that ALT mechanisms have been used to reconstruct
222 subtelomeric and telomeric regions in *C. elegans* and that templates for ALT in subtelomeric
223 regions can be used to repair telomere damage or maintain the integrity of the telomere
224 (Seo et al. 2015; Kim et al. 2019a; Kim et al. 2020; Kim et al. 2021b; Lee et al. 2022b). Thus, we
225 analyzed subtelomeric regions of our four Panagrolaimidae genome assemblies to trace
226 ancestral telomere damage and repair events that could explain how the TTAGAC telomere
227 evolved from the TTAGGC telomere.

Most telomeric repeats in telomeric regions (7/10 in LJ2284, 4/8 in LJ2285, 6/8 in LJ2400, and 8/14 in LJ2406) were attached to unit clusters containing TTAGGC or TTAGAC (e.g. [TTAGGCTTAATTGC]n or [TTAGACTTATTCGC]n) (Fig. 2A and Supplemental Table S10). This finding was striking, as none of the subtelomeric regions of *C. elegans* and *B. xylophilus* exhibited TRM-containing unit clusters directly attached to the telomeric regions (Fig. 2A and Supplemental Table S11). Furthermore, both TTAGGC-containing unit clusters and TTAGAC-containing unit clusters were identified in all genomes of the TTAGGC- and TTAGAC-telomere isolates, with the exception of the LJ2285 genome (TTAGAC telomere), which contained only TTAGAC-containing unit clusters. For LJ2284, LJ2400, and LJ2406, these clusters consisted of repetitive units with length distributions ranging from 14 bp to ~2 kb, and these repetitive units were tandemly repeated on an average of ≥19 copies (Supplemental Tables S10 and S12). The LJ2285 genome did not contain such same-length unit clusters; however, it contained sequence clusters consisting of 5–9 copies of units containing TTAGAC and similar sequences (10–74 bp). These TRM-containing unit clusters were not found outside the subtelomeric region in the telomere-containing contigs/scaffolds (see also Supplemental Note describing a contrasting case in *C. elegans*), raising the possibility that these clusters are associated with telomere maintenance in Panagrolaimidae. Notably, we also identified a similar subtelomeric pattern in *Caenorhabditis uteleia*, the noncanonical TTAGGT-TRM species, where TTAGGT- and/or TTAGGC-containing unit clusters were attached to or close to long TTAGGT repeat arrays (Supplemental Fig. S7 and Supplemental Note) (SRA accession: ERR8978452) (The Darwin Tree of Life Project Consortium 2022).

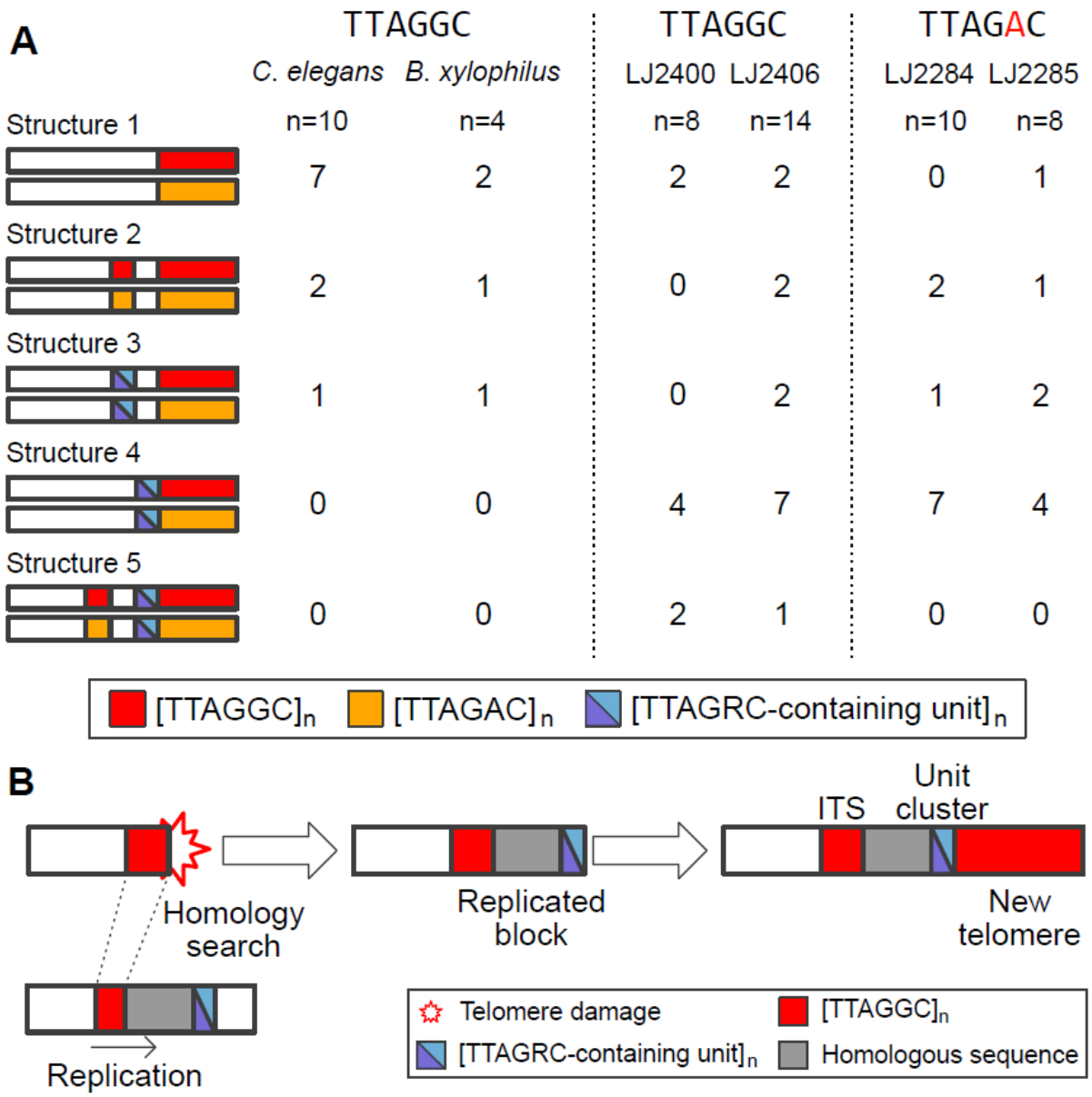


Figure 2. Subtelomere structures and a proposed model to explain clustered blocks of TRMs or TTAGRC-containing unit clusters in the subtelomeric regions. (A) Schematic representation of subtelomere structures in the contigs/scaffolds listed in Supplemental Table S10. We

256 categorized the structure of subtelomeric regions (up to 200 kb from the end of the
257 telomeric contig/scaffold) using the following criteria: (i) whether it has ITSs (shorter red
258 blocks) and/or TTAGRC-containing unit clusters (half-sky blue, half-violet blocks) and (ii)
259 whether ITSs, unit clusters, and telomeric regions were directly attached or separated by
260 other sequences (white, empty blocks). Each horizontal bar represents a subtelomere
261 structure with only a telomeric TTAGRC-containing unit cluster (shown in Structure 4 and 5)
262 and an ITS or a TTAGRC-containing unit cluster closest to the telomere (shown in Structure 2
263 or 3). Each value refers to the number of contigs/scaffolds with the corresponding structure
264 type in each species/isolate. TTAGGC-telomere species and isolates had ITSs that were
265 composed of TTAGGC, rather than TTAGAC, and TTAGAC-telomere species and isolates had
266 only TTAGAC-type ITSs, too. (B) A proposed model for generating subtelomere structures
267 responsible for Structure 5. After telomere damage, HDR mechanisms can exploit homology
268 between shortened telomeric repeats and ITSs at other loci to repair the damaged telomere.
269 A HDR mechanism, break-induced replication (BIR), may replicate sequences near the ITS,
270 creating a new homologous block of the original sequence block. If the original template
271 block exhibits a TTAGRC-containing unit cluster, the cluster would also be replicated. Finally,
272 a new telomere composed of TTAGGC repeats can be replenished by active telomerase.

273

274 In addition to the TRM-containing unit clusters, interstitial telomeric sequences (ITSs) were
275 frequently observed in the subtelomeric regions of the constructed genome assemblies,
276 which were consistent with the ITS enrichment present in *C. elegans* chromosome arms (The
277 *C. elegans* Sequencing Consortium 1998). All subtelomere structures were categorized based

on the location of the unit clusters and/or the ITSs in subtelomeric regions (up to 200 kb from the end of the telomeric contig/scaffold) (Supplemental Tables S10 and S11). We annotated a total of 12 subtypes (Supplemental Fig. S8), but for simplicity we merged them into five major subtelomere structures by considering only telomeric TTAGRC-containing unit clusters and ITSs or TTAGRC-containing unit clusters closest to the telomere (Structure 1–5 in Fig. 2A; see Supplemental Fig. S9 for examples of Structure 2–5). Notably, the motif in all ITSs was the same as their TRM. The four genome assemblies exhibited similar patterns. For most of the subtelomere structures in the four genome assemblies, the TRM-containing unit cluster was positioned directly adjacent to the telomeric region. Structure 1, which did not contain any unit cluster or ITS, accounted for fewer than a quarter of all the subtelomere structures (the lengths of telomeres in Structure 1 differ from those of ITSs in *C. elegans* and our four isolates; see Supplemental Note). In contrast, Structure 2–5 had at least one unit cluster or ITS separated from the telomeric region, which may have resulted from ancestral telomere damage and repair processes.

TTAGAC-containing units were probably used to maintain the integrity of TTAGGC telomeres

We postulated that the identified subtelomeric clusters of the TTAGRC-containing units and ITSs were evidence of ancestral telomere damage and repair via ALT mechanisms and that if true, this would help understand the role of TTAGRC-containing unit clusters in telomere maintenance. Among Structure 2–5, Structure 5 may be explained by BIR, one of the most well-known ALT mechanisms for telomere repair (Malkova et al. 1996; Bosco and Haber 1998; Kim et al. 2019a; Kim et al. 2020). Telomeres that have been damaged or shortened should

be repaired, and BIR may rely on homology between shortened telomeric repeats and ITSs at other loci (Fig. 2B, left side). The replication was then activated, and occasionally, sequences beside the ITS would also be replicated (Fig. 2B, middle). Furthermore, if the original homology block contained a TTAGRC-containing unit cluster, the cluster would also be duplicated simultaneously, resulting in a new subtelomere structure with a homology block interposed between the ITS and the TTAGRC-containing unit cluster. Finally, a new telomere composed of TTAGGC repeats could be replenished by active telomerase (Fig. 2B, right side). We then hypothesized that if Structure 5 cases in our genomes had been generated via BIR, we should be able to identify another sequence block homologous to the interposed sequence between the ITS and the TTAGRC-containing unit cluster.

To test this hypothesis, we searched for homology blocks in the sequences between an ITS and a TTAGRC-containing unit cluster from each of three contigs representing Structure 5 in whole genome sequences (ptg000074l and ptg000145l of LJ2400 and ptg000247l of LJ2406) (Fig. 2, Fig. 3, Supplemental Fig. S8, and Table S10). ptg000074l and ptg000145l of LJ2400 shared 8-kb and 9-kb homology blocks adjacent to the ITS and had a TTAGGC-containing unit cluster ~9-kb and ~200-bp away from the shared block, respectively (Fig. 3A). These 8-kb and 9-kb homology blocks pointed to the possibility of their replication via BIR using the ITS as homology.

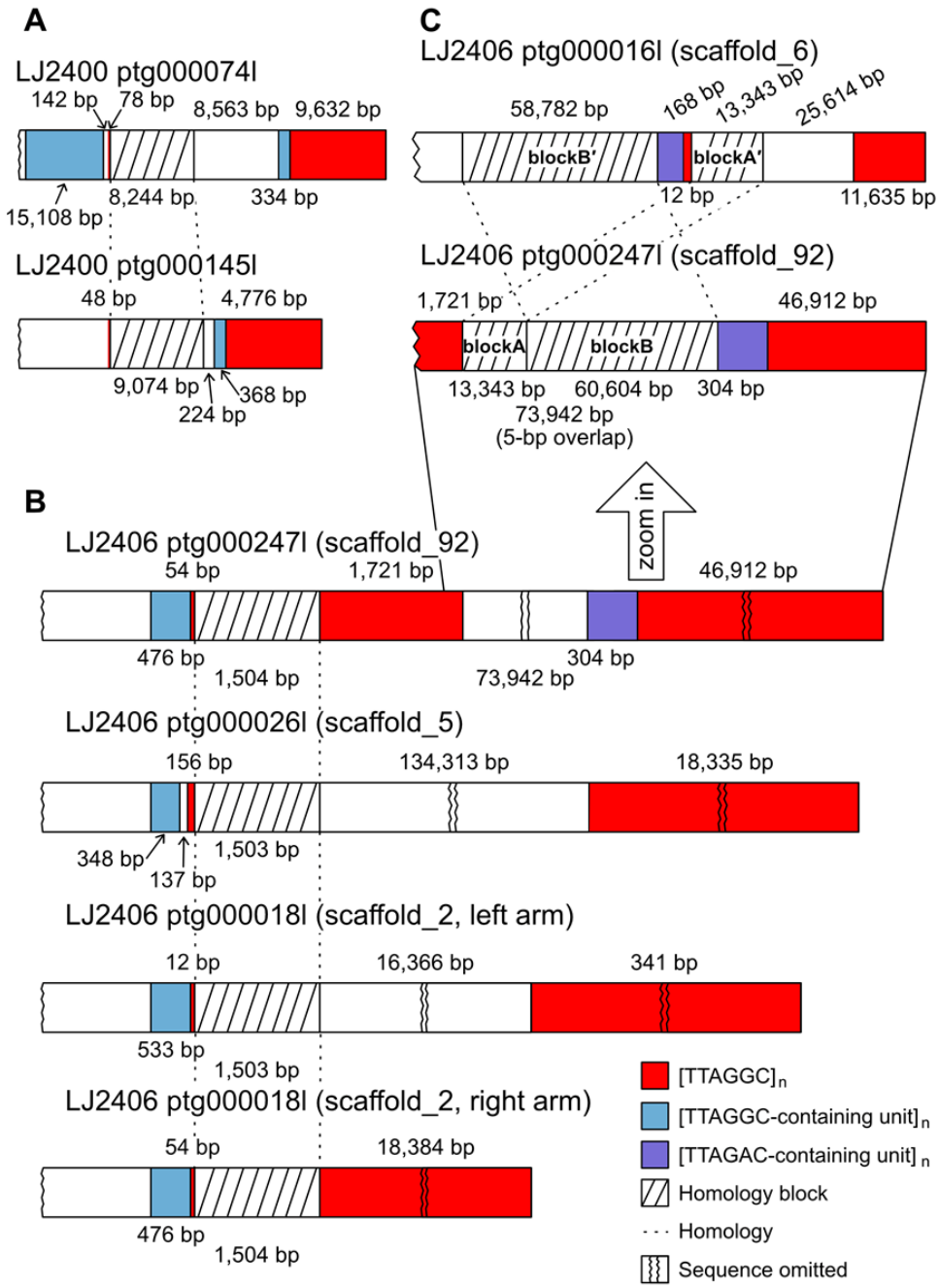


Figure 3. Traces of telomere damage and repair processes through ALT mechanisms identified in subtelomeric regions of genome assemblies of TTAGGC-telomere isolates. (A–C) Schematic representations of subtelomeres in the genome assemblies of LJ2400 and LJ2406.

Hatched boxes indicate homology blocks between two or more subtelomeres, and dashed lines indicate homologous relationships. The red boxes represent telomeric-repeat clusters. Violet and sky blue boxes represent TTAGAC-containing unit clusters and TTAGGC-containing unit clusters, respectively. Not to scale.

The ptg000247l of LJ2406 had a more evident BIR signature than the ptg000074l and ptg000145l of LJ2400, indicating that the subtelomeric TTAGRC-containing unit cluster was built via the ALT process. The subtelomeric region of ptg000247l of LJ2406 consisted of a TTAGGC-containing unit cluster, a 54-bp ITS, a 1.5-kb sequence block, a 1.7-kb ITS, a 74-kb sequence block, and a TTAGAC-containing unit cluster close to the telomeric region (Fig. 3B). By searching for this complex subtelomeric sequence in the whole genome using BLAST, we determined that the 1.5-kb and 74-kb blocks had homologous sequences in other subtelomeric regions. The 1.5-kb block between the 54-bp ITS and 1.7-kb ITS had almost identical homologous sequences in three different subtelomeric regions. These shared blocks were directly adjacent to short and variable-length ITSs near TTAGGC-containing unit clusters (Fig. 3B). The TTAGGC-containing cluster and ITS may have been utilized for homology to replicate the 1.5-kb homology blocks.

The 74-kb block between the 1.7-kb ITS and the TTAGAC-containing unit cluster possessed another trace of HDR and an additional replication of a TTAGRC-containing unit cluster after telomere damage (Fig. 3C). As the 1.7-kb length of the ITS is too long to be created by random mutation, it could represent evidence of ancestral telomere damage and repair. The

74-kb block contained at least two distinct homology blocks: the 13-kb blockA and the 61-kb blockB (Fig. 3C). The 13-kb blockA directly linked to the ITS in ptg000247l had an identical blockA′ in ptg000016l that was also directly attached to another 180-bp sequence of an ITS and a TTAGAC-containing unit cluster. The 61-kb blockB was homologous but not identical to the 59-kb blockB′ located on the opposite side of the 180-bp block of the ITS and the TTAGAC-containing unit cluster in ptg000016l. Of note, the last 5 bp of the 13-kb blockA and the first 5 bp of the 61-kb blockB were identical (i.e. TAAAT). The 61-kb blockB was further divided into four blocks, each of which also exhibited microhomologies with the adjacent blocks (Supplemental Fig. S10 and Supplemental Table S13). These data imply that telomere damage and repair may have occurred in the following order: first, the 13-kb blockA in ptg000247l was replicated by BIR using homology between the 1.7-kb ITS in ptg000247l and the 180-bp block of the ITS and the TTAGAC-containing unit cluster in ptg000016l. Next, the 61-kb blockB in ptg000247l was replicated by multiple rounds of microhomology-mediated BIR (MMBIR) and template switching events using microhomology sequences at the end of the blocks (Supplemental Fig. S10 and Supplemental Note).

This 59-kb blockB′ in ptg000016l was directly attached to the TTAGAC-containing unit cluster, which was duplicated in ptg000247l. Notably, the original TTAGAC-containing unit cluster in ptg000016l was shorter than the duplicated cluster in ptg000247l, implying that the duplicated cluster nearly doubled after MMBIR was completed. The duplicated TTAGAC-containing unit cluster in ptg000247l was directly linked to the TTAGGC telomere, indicating that the duplicated cluster was exposed at the end. These lines of evidence support the notion that even in TTAGGC-telomere species, TTAGAC-containing units could be

incorporated to constitute a telomere or at least serve as a DNA template for replenishing TRM repeats by telomerase. This use of TTAGAC-containing units in the TTAGGC-telomere isolate may have contributed to the evolution of the TTAGAC telomere from the TTAGGC telomere.

Telomere-associated protein genes exhibited similar conservation patterns among Panagrolaimidae species/isolates

We hypothesized that TTAGAC-containing unit clusters served as a component of partially stable telomere in TTAGGC-telomere species and that TBPs for TTAGGC telomere could have bound TTAGAC repeats and have been conserved in both TTAGGC-telomere and TTAGAC-telomere species. To support this hypothesis, we analyzed protein homologies in Panagrolaimidae with *B. xylophilus* as an outgroup species, in which proteins have been described to maintain or bind telomeres in *C. elegans* (Ahmed and Hodgkin 2000; Hofmann et al. 2002; Im and Lee 2003; Kim et al. 2003; Joeng et al. 2004; Im and Lee 2005; Meier et al. 2006; Boerckel et al. 2007; Raices et al. 2008; Meier et al. 2009; Ferreira et al. 2013; Shtessel et al. 2013; Dietz et al. 2021; Yamamoto et al. 2021). We identified 12 proteins that exhibited similar homology patterns between TTAGGC- and TTAGAC-telomere species/isolates (Supplemental Fig. S11). In contrast, four proteins (POT-2, TEBP-2 (also known as DTN-2), HPR-9, and MRT-1) displayed inconsistent conservation patterns. POT-2 was conserved in only two of the three TTAGGC-telomere species/isolates, and TEBP-2 remained in only one of the three TTAGAC-telomere species/isolates. HPR-9 was observed in only one TTAGAC-telomere isolate. MRT-1, required for telomerase activity in *C. elegans* (Meier et al. 2009), was not detected in *B. xylophilus* or in any of the three TTAGAC-telomere species/isolates but

was present in all three TTAGGC-telomere species/isolates. However, MRT-1 functional domains were not conserved in the TTAGGC-telomere species/isolates. Nonetheless, homologs of the four TBPs were mostly conserved in all Panagrolaimidae species and isolates (Supplemental Table S14). Despite these variations, TRT-1 was conserved in all species and isolates, suggesting that they may have active telomerase and that changes in its RNA template may have been involved in TRM evolution (Fig. 4; see Discussion).

DISCUSSION

TRMs and their protein partners are well conserved and mostly co-evolving because they must be connected to maintain the DNA–protein complex, telomeres (Shakirov et al. 2009; Steinberg-Neifach and Lue 2015; Sepsiova et al. 2016; Červenák et al. 2019). Nevertheless, TRMs have evolved across several taxa (Fulnečková et al. 2013; Garavís et al. 2013; Peska and Garcia 2020; Červenák et al. 2021). In this study, we identified three novel TRMs in Nematoda and confirmed that TTAGAC, one of the three novel TRMs, composes telomeres in a subset of Panagrolaimidae isolates. We also hypothesized that TTAGAC evolved through TTAGAC-containing unit clusters and the robustness of the TTAGGC-suited TBPs to bind tandem arrays of TTAGAC repeats in a canonical TTAGGC-telomere ancestor.

The consistent pattern of TRM changes observed in other multicellular organisms supported our hypothesis. First, sequence changes in plant TRMs also exhibit key characteristics (Peska and Garcia 2020). In particular, the basal TRM in plants, TTTAGGG, (*Arabidopsis*-type) (Richards and Ausubel 1988) evolved into _TTAGGG (vertebrate-type) (Weiss and Scherthan

2002; Sýkorová et al. 2003; Sýkorová et al. 2006), TTTTAGGG (*Chlamydomonas*-type) (Fulnečková et al. 2012), TTCAGG_ (*Genlisea*-type) (Tran et al. 2015), TTTCAGG_ (*Genlisea*-type) (Tran et al. 2015), TTTTAGG_ (*Klebsormidium*-type) (Fulnečková et al. 2013), and TTTTTTAGGG (*Cestrum*-type) (Peška et al. 2015). Briefly, these variant TRMs exhibited alterations in the lengths of the T and G arms or changes in pyrimidine (T to C) nucleotides close to the middle A. Pyrimidine nucleotide conversion, rather than changes between pyrimidine and purine nucleotides and vice versa, has also been observed in insects in which TTAGG changed to TCAGG (Mravinac et al. 2011). Similarly, in the current study, TTAGGC changed to TTAGGT (pyrimidine conversion) and TTAGAC (purine conversion) in nematodes. Considering pyrimidines (T and C) have comparable base sizes and purines (G and A) are of similar sizes, we propose that conversions between pyrimidines or between purines do not sterically hinder TBP binding, allowing pre-existing TBPs to attach to the modified TRM. Furthermore, in yeasts and plants, TBPs are sufficiently robust to bind to various types of TRM despite several base pair changes (Rotková et al. 2004; Fajkus et al. 2005; Kramara et al. 2010; Visacka et al. 2012; Červenák et al. 2019; Tomáška et al. 2019; Červenák et al. 2021). In yeast, a TBP binds at least six different types of TRMs, implying binding robustness of TBPs during TRM evolution (Červenák et al. 2019). Here we suggest that this idea could be further supported by TTAGAC-containing unit clusters in TTAGGC-telomere species.

It is still unclear whether TTAGGC-binding TBPs also bind to TTAGAC in TTAGGC-telomere Panagrolaimidae species/isolates. Traces of ALT action identified in our genome assemblies revealed that TTAGAC-containing units were used to repair telomere damage in TTAGGC-telomere isolates. Two TTAGGC-telomere isolates and one TTAGAC-telomere isolate

examined harbored both clusters of consecutive TTAGGC- and TTAGAC-containing units. Some clusters were attached directly or were adjacent to the telomeres. Specifically, the subtelomeric sequences of ptg000247l in LJ2406, a TTAGGC-telomere isolate, showed evidence of direct usage of a TTAGAC-containing unit for the repair of a damaged telomere. In the vicinity of a damaged and shortened telomere, there were homologous sequence blocks and a TTAGAC-containing unit cluster, which were probably replicated by BIR and MMBIR, and the TTAGAC-containing unit cluster was directly adjacent to its telomeric TTAGGC cluster. Thus, at least over a brief period of time, the TTAGAC-containing unit cluster was exposed at the end of the chromosome before the TTAGGC repeats were replenished.

Furthermore, the TTAGAC-containing unit cluster was not only duplicated but also elongated at the end because it was nearly twice as long as its initial cluster template in a separate subtelomeric region. This cluster might have been elongated via replication slippage based on its repetitive nature or via ALT mechanisms that can replicate TRM-containing units at the end of the chromosome. We could not determine the specific mechanisms involved in the elongation of the TTAGAC-containing unit cluster. If the elongation was achieved via ALT mechanisms, this suggests that the TTAGAC-containing unit cluster exhibits its own replication capability in the telomeric region to maintain the telomere.

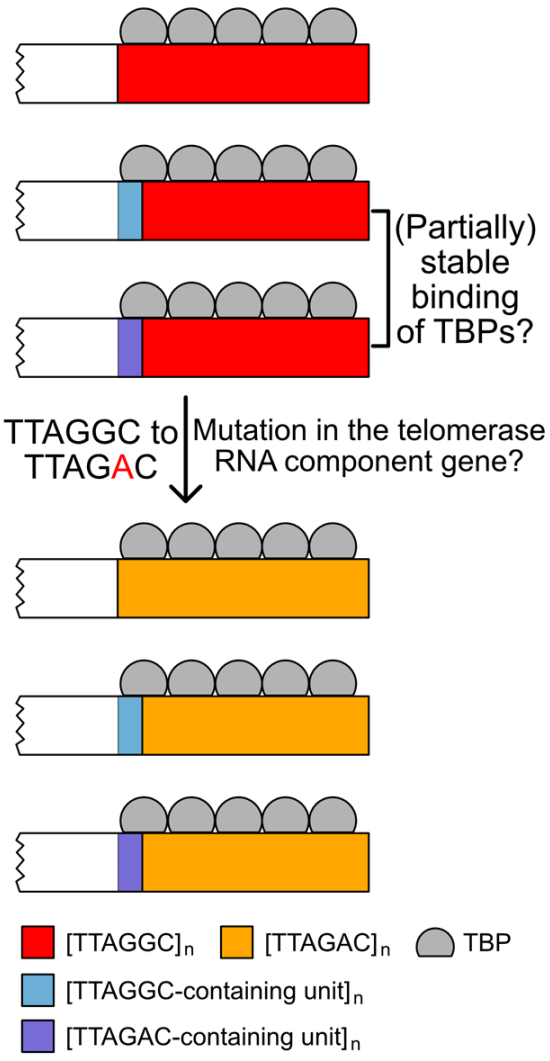
If the TTAGAC-containing unit cluster acted as a component of the telomere, the TTAGGC-suited TBPs would bind to the telomere; otherwise, the end would be recognized as a DNA damage site. This binding robustness of TBPs, if it exists as reported in fungi, may explain the subtelomere structure and TRM evolution. This robustness could be achieved via two potential ways: (1) the TTAGGC-suited TBPs also exhibited sufficient binding affinity for

454 TTAGAC, such that TTAGAC-unit clusters could be utilized for repairing telomere damage.
455 This binding affinity stabilized the TTAGAC-containing unit–TBP complex as a telomere. (2)
456 TBPs exhibited a weak binding affinity to TTAGAC and formed a partially stable telomere
457 complex, and the affinity strengthened through frequent use of TTAGAC-containing unit
458 clusters to repair telomere damage and ensure fitness. In both cases, TTAGGC-suited TBPs
459 could also stabilize TTAGAC telomeres, which thus facilitated the conversion from TTAGGC
460 telomeres to TTAGAC telomeres.

461 Our results suggest a plausible evolutionary model for the TRM conversion in the family
462 Panagrolaimidae. First, ancestral TTAGGC-telomere species may have utilized various ALT
463 mechanisms to counteract telomere damage when active telomerase cannot access the
464 damaged telomere (Fig. 4). Replication of TTAGRC-containing unit clusters could represent
465 one such ALT mechanism and would have been used to repair damaged telomeres.
466 Consequently, active telomerase may use unit clusters as a starting template to fully
467 replenish the partially repaired telomere. Because TBPs can fully or partially stabilize
468 telomere complexes consisting of TTAGAC-containing unit clusters, the evolution of the TRM
469 conversion to TTAGAC had a greater fitness than other TTAGGC variants. We could not
470 identify the telomerase RNA component gene in Panagrolaimidae. However, conservation of
471 TRT-1 and highly accurate replication of TRMs at the telomere probably support the
472 presence of an active telomerase in these species/isolates (Fig. 1G and Supplemental Fig. S6
473 and S11). TRM would have converted to TTAGAC via mutation in the telomerase RNA
474 component gene, leading to TTAGAC telomeres, although it remains unclear whether this
475 change involved a single mutation in the gene or coexistence of the original and mutated

476 genes. This transition may have occurred long ago, since TTAGAC- and TTAGGC-telomere
477 isolates had only TTAGAC- and TTAGGC-ITSs, respectively. Moreover, both telomeres and
478 ITSs consisted of the same TRM, and sequences of their ITSs were not much degenerated,
479 indicating that these ITSs resulted from recent, and probably frequent, telomere damage
480 events in these isolates.

481



482

483

Figure 4. A model for TRM evolution in Panagrolaimidae. If TBPs with affinity for TTAGGC telomeres could bind to any of the TTAGGC telomere, TTAGGC-containing unit clusters, or TTAGAC-containing unit clusters, they were able to construct a fully or partially stable telomere. This binding robustness of TBPs may have facilitated the TRM conversion from TTAGGC to TTAGAC. Even after TRM changed via mutation of the telomerase RNA component gene, if it occurred, TBPs could still bind to altered telomeric repeats, which possibly contributed to the evolution of a novel TRM, TTAGAC, in the Panagrolaimidae family.

It is unclear whether TTAGGC-suited TBPs could bind to the novel telomeric sequence; however, if this were possible, the TTAGAC cluster–TBP complex would be partially stabilized, preventing the new telomere from being recognized as a DNA damage site. This hypothesis can be evaluated directly by inducing a mutation in the telomerase RNA component gene and then measuring the robustness of TBP in terms of sequence specificity. Unfortunately, we could not identify the telomerase RNA component gene for these isolates, as the gene has not been identified even in *C. elegans*. Further studies are required to support our hypothesis. Nonetheless, we identified TTAGGC- and/or TTAGGT-containing unit clusters in putative subtelomeric regions of *Caenorhabditis uteleia*, which suggests that TRM-containing unit clusters are associated with telomere maintenance and that their usage in telomere maintenance contributes to TRM evolution (see also Supplemental Note).

Our results describe the independent evolution of three novel TRMs in Nematoda. We found that an ALT-mediated mechanism may have been used to repair ancestral telomere damage in Panagrolaimidae nematodes. Further, we hypothesized that the use of TTAGAC-containing units and the robustness of TBPs facilitated the evolution of the novel TRM. As the robustness of the TBPs may depend on the same-sized bases between canonical and novel TRMs, the evolutionary path we propose for Panagrolaimidae can also be applied to plants, insects, or other nematodes where mutations occur only between pyrimidine nucleotides or between purine nucleotides. Although our study cannot explain all TRM changes in Nematoda, it provides new insight into telomere evolution through ALT and the robustness of TBPs.

METHODS

Worm sampling and culture

Nematodes were collected from rotten fruits in the Republic of Korea (see Supplemental Table S5 for more information). They were grown in plates containing nematode growth medium seeded with *Escherichia coli* strain OP50 and contaminated by their natural microbes.

DNA/RNA extraction and sequencing

Mixed-stage worms were lysed, and genomic DNA was extracted using the Gentra Puregene Cell and Tissue Kit (QIAGEN) to obtain WGS data for Panagrolaimidae isolates, LJ2281,

LJ2284, LJ2285, LJ2400, LJ2402, and LJ2406. Macrogen (South Korea, https://www.macrogen.com/en/main) prepared DNA sequencing libraries using TruSeq Nano DNA and performed 151-bp paired-end DNA sequencing on the Illumina NovaSeq 6000 platform. One adult female worm was lysed, and its genomic DNA was amplified in the same 0.2-mL PCR tube using the REPLI-g single-cell WGA kit (QIAGEN) to obtain single-worm WGS data from LJ2050, LJ2051, LJ2060, LJ2070, LJ2072, LJ2401, LJ2411, and LJ2417. Paired-end DNA sequencing was performed using the Illumina NovaSeq 6000 platform by Theragen Bio (South Korea, https://www.theragenbio.com/en/).

Long-read DNA sequencing and short-read RNA sequencing of LJ2284, LJ2285, LJ2400, and LJ2406 were conducted as described by Kim et al. (Kim et al. 2019a) and Lee et al. (Lee et al. 2022a). In summary, genomic DNA was extracted from mixed-stage worms using phenol/chloroform/isoamyl alcohol (25:24:1) and sequenced using the HiFi mode of the PacBio Sequel II platform by Macrogen. RNA was extracted from mixed-stage worms using the TRIzol method. RNA sequencing libraries were prepared using TruSeq Nano DNA and sequenced by Macrogen using the Illumina NovaSeq 6000 platform with paired-end reads.

Preparing publicly available WGS data

We included Nematoda species in our study, whose genome assembly data were available in the GenBank database and short-read sequencing data were available in the NCBI Sequence Read Archive (SRA) (RRID:SCR_004891). The following criteria were used to filter the short-read sequencing data: Instrument, *Illumina* (if HiSeq data were available, we used HiSeq; otherwise, MiSeq, NextSeq, and Genome Analyzer II data were selected); Source, *Genomic*;

Layout, *PAIRED*; Number of Bases, *>5 Gb*; Spot number, *>5 million*; Read length, *>70 bp*. Nematode clades were classified according to the phylogenetic tree (Smythe et al. 2019). Any genus that was not included in the phylogenetic tree was excluded. WGS data for *Enoplus brevis,* which lacked genome assembly information in the GenBank database, and WGS data for *Panagrellus redivivus* (Srinivasan et al. 2013) that was supplied by Dr. P. W. Sternberg (now it can be accessible via NCBI SRA under accession number SRR25684730) were added to the filtered datasets. Detailed accession information is available in Supplemental Tables S1 and S2.

Identifying TRMs using short-read WGS data

To normalize the public WGS data, we trimmed all reads to 60 bp and used only 5 million sub-sampled reads from the R1 files using Seqtk (https://github.com/lh3/seqtk) (version 1.3-r106; *seqtk sample -s 11 - 5000000*). For each dataset, 23-mers were counted using Kounta (https://github.com/tseemann/kounta) (version 0.2.3; *kounta --kmer 23 --out*). Only tandemly repeated sequences with unit lengths of 5–7 bp and counts >99 were used to identify putative TRMs that were most frequent and most similar to the canonical TTAGGC sequence. Methods used to further validate the novel putative TRMs are described in Supplemental Methods. We averaged the counts of the 23-mers that contained the corresponding TRM concatemers to compare the number of TRMs across species. For the Panagrolaimidae isolates, we used 20 million sub-sampled reads and all 23-mers, as telomeric repeats were not evenly amplified during whole-genome amplification.

Genome assembly

HiFi reads were *de novo* assembled using Hifiasm (Cheng et al. 2021) (version 0.13-r308; *hifiasm -l0*). Contigs that were potentially contaminated with bacterial sequences were removed as described by Kim et al. (Kim et al. 2019a). Scaffolding and visualization based on the Hi-C data are described in Supplemental Methods. The completeness of the genome assembly was evaluated using BUSCO (Simão et al. 2015) (version 4.0.6; *busco -m genome -l nematoda_odb10*) with the nematoda dataset in the OrthoDB release 10. Isolate-specific repetitive sequences were identified using BuildDatabase (Flynn et al. 2020) (version 2.0.1; default options) and RepeatModeler (Flynn et al. 2020) (version 2.0.1; *RepeatModeler -database -LTRStruct*). Isolate-specific and known metazoan-repetitive sequences were masked using RepeatMasker (Chen 2004) (version 4.1.0; *RepeatMasker -lib -s* for species-specific repeats and *RepeatMasker -species metazoa -s* for metazoan repeats). We mapped the RNA-seq reads to the repeat-masked genome using HISAT2 (Kim et al. 2019b) (version 2.2.1; *hisat2-build* for genome indexing and *hisat2* with default options to map the RNA-seq reads to the corresponding genome). RNA-mapping data was used to annotate genes using BRAKER (Stanke et al. 2006; Stanke et al. 2008; Li et al. 2009; Barnett et al. 2011; Lomsadze et al. 2014; Buchfink et al. 2015; Hoff et al. 2016; Hoff et al. 2019; Brůna et al. 2021) (version 2.1.5; *braker.pl --genome -bam --softmasking*).

Preparation of 18S ribosomal DNA sequences to generate the Panagrolaimidae phylogenetic tree

For *Panagrolaimus* sp. PS1159, *Panagrolaimus davidi*, *Panagrellus redivivus*, and *Bursaphelenchus xylophilus*, we used publicly available 18S rDNA sequences (GenBank accessions: U81579.1, AJ567385.1, AF083007.1, and KJ636306.1, respectively). For the HiFi-

based *de novo* genome assemblies, we extracted 18S rDNA sequences by searching known nematode 18S rDNA PCR primers (nSSU_F_04: 5′-GCTTGTCTCAAAGATTAAGCC-3′ (Blaxter et al. 1998) and nSSU_R_82: 5′-TGATCCTTCTGCAGGTTCACCTAC-3′ (Medlin et al. 1988)) in genome assemblies using BLAST+ (Camacho et al. 2009) (version 2.7.1; *makeblastdb -input_type fasta -dbtype nucl* and *blastn -task blastn-short -outfmt 6*).

For the other TTAGAC-TRM isolates used in this study, we mapped their short-read DNA sequencing data to the LJ2285 genome assembly using BWA-MEM (version 0.7.17; *bwa mem*, default option), and the 18S rDNA sequence variation of each isolate were identified using BCFtools (Li 2011a) (version 1.13; *bcftools mpileup -Ou -f | bcftools call -Ou -mv | bcftools norm -f -Oz -o*). The indexed output VCF files were obtained using Tabix (Li 2011b) (version 1.13; default option), and the LJ2285 18S rDNA sequence was replaced with their variants of each isolate using SAMtools (Li et al. 2009) (version 1.13) and BCFtools (*samtools faidx -r | bcftools consensus -o*). For the other TTAGGC-TRM isolates, we used the LJ2400 genome assembly as a reference and repeated the procedure described above. All 18S rDNA sequences of Panagrolaimidae obtained in this study are listed in Supplemental Table S6. We generated an alignment file using all 18S rDNA sequences as input for Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/, RRID:SCR_001591, (Sievers and Higgins 2021)) (Sequence Type: DNA; Output Alignment Format: PHYLIP). We used this alignment as an input to RAxML (Stamatakis 2014) (version 8.2.12) using raxmlGUI 2.0 (Edler et al. 2021) (version 2.0.10; options: GTRGAMMA and ML+rapid bootstrap with 1000 replications) to infer the phylogenetic relationships, which were visualized using Dendroscope (Huson and Scornavacca 2012) (version 3.8.4).

Telomere and subtelomere structure in the genome assemblies

We selected candidate telomere-containing contigs using two tandemly repeated copies of TTAGRC that appeared ≥3 times in the 600-bp region at each end of the contig. We manually validated these telomeric regions using the following criteria: TRM cluster length ≥500 bp (the longest ITS length in *C. elegans*) (Yoshimura et al. 2019) or TRM cluster still located at the end after scaffolding (only for LJ2284 and LJ2406). We analyzed whether each telomere-containing contig/scaffold had clusters with ≥6 copies of TRMs or whether TTAGRC-containing units repeated tandemly in its subtelomeric region (up to 200 kb from the end of the contig/scaffold). For LJ2285, we considered sequences consisting of 5–9 copies of different units that contained TTAGRC and similar sequences as unit clusters because LJ2285 did not contain clusters composed of the same units. Subsequently, we investigated whether the unit cluster (containing ≥6 copies of units, the same TTAGRC and identity ≥85% of units) existed outside the subtelomeric region of telomere-containing contigs/scaffolds using BLAST+ (Camacho et al. 2009) (version 2.12.0; *makeblastdb -input_type fasta -dbtype nucl*; *blastn -task blastn-short -outfmt 6* for sequences <30 bp and *blastn -task megablast -outfmt* for sequences ≥30 bp). Unit sequences of LJ2284, LJ2400, and LJ2406 and unit cluster sequences in LJ2285 were denoted starting with TTAGRC (Supplemental Table S12). The BIR traces were analyzed by searching for homologous sequences between ITSs, TTAGRC-containing unit clusters, or telomeric repeats against their corresponding genome assembly using BLAST+ (Camacho et al. 2009) (version 2.7.1; *makeblastdb -input_type fasta -dbtype nucl and blastn -task megablast -outfmt 6*).

Conservation analysis of telomere-associated proteins

Protein FASTA sequences were downloaded from WormBase for *C. elegans* (Release WS281) and WormBase ParaSite (Howe et al. 2017) for *Panagrolaimus* sp. PS1159, *Panagrolaimus davidi*, *Panagrellus redivivus*, and *B. xylophilus* (Release WBPS15). Protein sequences for *Panagrolaimus* sp. PS1579 and *Panagrolaimus* sp. ES5 were not publicly available. For the four Panagrolaimidae isolates (LJ2284, LJ2285, LJ2400, and LJ2406), we used protein FASTA sequences obtained through gene annotation using BRAKER. For the non-*C. elegans* nematodes, we searched for protein sequences that were conserved in *C. elegans* using DIAMOND (Buchfink et al. 2021) (version 2.0.11; *diamond blastp -d -q -o --threads 20 --very-sensitive* and *diamond blastp -d -q -o --threads 20 --ultra-sensitive*) to characterize telomere-associated protein sequences that included TRT-1, POT-1, POT-2, POT-3, MRT-1, MRT-2, TEBP-1 (also known as DTN-1), TEBP-2 (also known as DTN-2), HPR-9, HPR-17, HUS-1, SUN-1, CEH-37, HMG-5, HRPA-1, and PLP-1. We filtered the nematode's highest bit score protein sequence for each *C. elegans* telomere-associated protein and identified the conserved domains using NCBI CD-search (Marchler-Bauer and Bryant 2004; Lu et al. 2020) (CDD database, version 3.19).

DATA ACCESS

Raw sequencing data and genome assemblies have been submitted in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject) under the accession number PRJNA845886 and in the Korean Nucleotide Archive (KoNA, https://kobic.re.kr/kona) under the BioProject accession number KAP220348 (sequencing data only).

COMPETING INTEREST STATEMENT

None declared.

ACKNOWLEDGEMENTS

This study was supported by Samsung Science and Technology Foundation [SSTF-BA1501-52]; and National Research Foundation of Korea [2019R1A6A1A10073437 to J.K.]. J. Lim was also supported by a scholarship for basic research, Seoul National University. We thank Dr. P. W. Sternberg for providing and publishing valuable sequencing data of *Panagrellus redivivus* and are grateful to the Darwin Tree of Life Project Consortium for releasing sequencing data of *Caenorhabditis uteleia*. We also thank Namhee Kim and Sungyeol Ahn for providing accession to their farms to collect rotten fruits and Dr. D. S. Lim for collaboration in collection of nematode species. We appreciate Dr. S. Sung for providing the suggestions on generating Hi-C library, and Dr. C. Kim for the help in Hi-C data processing.

AUTHOR CONTRIBUTIONS

Jiseon Lim: Conceptualization, Methodology, Formal Analysis, Investigation, Writing-Original Draft, Writing-Review & Editing. Wonjoo Kim: Investigation. Jun Kim: Conceptualization, Methodology, Investigation, Writing-Original Draft, Writing-Review & Editing. Junho Lee: Conceptualization, Writing-Review & Editing, Funding Acquisition, Supervision.

673

674 **REFERENCES**

675 Abad JP, De Pablos B, Osoegawa K, De Jong PJ, Martín-Gallardo A, Villasante A. 2004. TAHRE,
676 a novel telomeric retrotransposon from Drosophila melanogaster, reveals the origin
677 of Drosophila telomeres. *Molecular biology and evolution* **21**: 1620-1624.
678 Ahmed S, Hodgkin J. 2000. MRT-2 checkpoint protein is required for germline immortality
679 and telomere replication in C. elegans. *Nature* **403**: 159-164.
680 Barnett DW, Garrison EK, Quinlan AR, Strömberg MP, Marth GT. 2011. BamTools: a C++ API
681 and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**: 1691-1692.
682 Blaxter ML, De Ley P, Garey JR, Liu LX, Scheldeman P, Vierstraete A, Vanfleteren JR, Mackey
683 LY, Dorris M, Frisse LM. 1998. A molecular evolutionary framework for the phylum
684 Nematoda. *Nature* **392**: 71-75.
685 Boerckel J, Walker D, Ahmed S. 2007. The Caenorhabditis elegans Rad17 homolog HPR-17 is
686 required for telomere replication. *Genetics* **176**: 703-709.
687 Bosco G, Haber JE. 1998. Chromosome break-induced DNA replication leads to nonreciprocal
688 translocations and telomere capture. *Genetics* **150**: 1037-1047.
689 Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic
690 eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a
691 protein database. *NAR genomics and bioinformatics* **3**: lqaa108.
692 Bryan TM, Englezou A, Gupta J, Bacchetti S, Reddel R. 1995. Telomere elongation in immortal
693 human cells without detectable telomerase activity. *The EMBO journal* **14**: 4240-4248.
694 Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using
695 DIAMOND. *Nature methods* **18**: 366-368.
696 Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND.
697 *Nature methods* **12**: 59-60.
698 The C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode C.
699 elegans: a platform for investigating biology. *Science* **282**: 2012-2018.
700 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
701 BLAST+: architecture and applications. *BMC bioinformatics* **10**: 1-9.
702 Červenák F, Jurikova K, Devillers H, Kaffe B, Khatib A, Bonnell E, Sopkovičová M, Wellinger RJ,
703 Nosek J, Tzfati Y. 2019. Identification of telomerase RNAs in species of the Yarrowia
704 clade provides insights into the co-evolution of telomerase, telomeric repeats and
705 telomere-binding proteins. *Scientific reports* **9**: 1-15.
706 Červenák F, Sepšiová R, Nosek J, Tomáška Ľ. 2021. Step-by-step evolution of telomeres:
707 lessons from yeasts. *Genome biology and evolution* **13**: evaa268.
708 Chen N. 2004. Using Repeat Masker to identify repetitive elements in genomic sequences.
709 *Current protocols in bioinformatics* **5**: 4.10. 11-14.10. 14.
710 Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo
711 assembly using phased assembly graphs with hifiasm. *Nature methods* **18**: 170-175.

712 The Darwin Tree of Life Project Consortium. 2022. Sequence locally, think globally: The
713 Darwin tree of life project. *Proceedings of the National Academy of Sciences* **119**:
714 e2115642118.
715 Dayi M, Sun S, Maeda Y, Tanaka R, Yoshida A, Tsai IJ, Kikuchi T. 2020. Nearly complete
716 genome assembly of the pinewood nematode Bursaphelenchus xylophilus strain
717 Ka4C1. *Microbiology resource announcements* **9**: e01002-01020.
718 De Lange T. 2009. How telomeres solve the end-protection problem. *Science* **326**: 948-952.
719 Dietz S, Almeida MV, Nischwitz E, Schreier J, Viceconte N, Fradera-Sola A, Renz C, Ceron-
720 Noriega A, Ulrich HD, Kappei D. 2021. The double-stranded DNA-binding proteins
721 TEBP-1 and TEBP-2 form a telomeric complex with POT-1. *Nature communications* **12**:
722 1-20.
723 Edler D, Klein J, Antonelli A, Silvestro D. 2021. raxmlGUI 2.0: a graphical interface and toolkit
724 for phylogenetic analyses using RAxML. *Methods in Ecology and Evolution* **12**: 373-
725 377.
726 Fajkus J, Sýkorová E, Leitch AR. 2005. Telomeres in evolution and evolution of telomeres.
727 *Chromosome Research* **13**: 469-479.
728 Ferreira HC, Towbin BD, Jegou T, Gasser SM. 2013. The shelterin protein POT-1 anchors
729 Caenorhabditis elegans telomeres through SUN-1 at the nuclear periphery. *Journal of*
730 *cell biology* **203**: 727-735.
731 Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2020.
732 RepeatModeler2 for automated genomic discovery of transposable element families.
733 *Proceedings of the National Academy of Sciences* **117**: 9451-9457.
734 Fulnečková J, Hasíková T, Fajkus J, Lukešova A, Eliaš M, Sýkorova E. 2012. Dynamic evolution
735 of telomeric sequences in the green algal order Chlamydomonadales. *Genome*
736 *biology and evolution* **4**: 248-264.
737 Fulnečková J, Ševčíková T, Fajkus J, Lukešova A, Lukeš M, Vlček Č, Lang BF, Kim E, Eliaš M,
738 Sýkorova E. 2013. A broad phylogenetic survey unveils the diversity and evolution of
739 telomeres in eukaryotes. *Genome biology and evolution* **5**: 468-483.
740 Garavís M, Gonzalez C, Villasante A. 2013. On the origin of the eukaryotic chromosome: the
741 role of noncanonical DNA structures in telomere evolution. *Genome biology and*
742 *evolution* **5**: 1142-1150.
743 Heaphy CM, Subhawong AP, Hong S-M, Goggins MG, Montgomery EA, Gabrielson E, Netto
744 GJ, Epstein JI, Lotan TL, Westra WH. 2011. Prevalence of the alternative lengthening of
745 telomeres telomere maintenance mechanism in human cancer subtypes. *The*
746 *American journal of pathology* **179**: 1608-1615.
747 Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-
748 Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**:
749 767-769.
750 Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation with
751 BRAKER. In *Gene prediction*, pp. 65-95. Springer.
752 Hofmann ER, Milstein S, Boulton SJ, Ye M, Hofmann JJ, Stergiou L, Gartner A, Vidal M,
753 Hengartner MO. 2002. Caenorhabditis elegans HUS-1 is a DNA damage checkpoint
754 protein required for genome stability and EGL-1-mediated apoptosis. *Current biology*
755 **12**: 1908-1918.

Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. 2017. WormBase ParaSite− a comprehensive resource for helminth genomics. *Molecular and biochemical parasitology* **215**: 2-10.

Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic biology* **61**: 1061-1067.

Im SH, Lee J. 2003. Identification of HMG-5 as a double-stranded telomeric DNA-binding protein in the nematode Caenorhabditis elegans. *FEBS letters* **554**: 455-461.

Im SH, Lee J. 2005. PLP-1 binds nematode double-stranded telomeric DNA. *Molecules & Cells (Springer Science & Business Media BV)* **20**.

Joeng KS, Song EJ, Lee K-J, Lee J. 2004. Long lifespan in worms with long telomeric DNA. *Nature genetics* **36**: 607-611.

Kim C, Kim J, Kim S, Cook DE, Evans KS, Andersen EC, Lee J. 2019a. Long-read sequencing reveals intra-species tolerance of substantial structural variations and new subtelomere formation in C. elegans. *Genome research* **29**: 1023-1035.

Kim C, Sung S, Kim J-S, Lee H, Jung Y, Shin S, Kim E, Seo JJ, Kim J, Kim D. 2021a. Telomeres reforged with non-telomeric sequences in mouse embryonic stem cells. *Nature communications* **12**: 1-15.

Kim C, Sung S, Kim J, Lee J. 2020. Repair and reconstruction of telomeric and subtelomeric regions and genesis of new telomeres: implications for chromosome evolution. *Bioessays* **42**: 1900177.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019b. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**: 907-915.

Kim E, Kim J, Kim C, Lee J. 2021b. Long-read sequencing and de novo genome assemblies reveal complex chromosome end structures caused by telomere dysfunction at the single nucleotide level. *Nucleic acids research* **49**: 3338-3353.

Kim SH, Hwang SB, Chung IK, Lee J. 2003. Sequence-specific binding to telomeric DNA by CEH-37, a homeodomain protein in the nematode Caenorhabditis elegans. *Journal of Biological Chemistry* **278**: 28038-28044.

Kramara J, Osia B, Malkova A. 2018. Break-induced replication: the where, the why, and the how. *Trends in Genetics* **34**: 518-531.

Kramara J, Willcox S, Gunisova S, Kinsky S, Nosek J, Griffith JD, Tomaska L. 2010. Tay1 protein, a novel telomere binding factor from Yarrowia lipolytica. *Journal of Biological Chemistry* **285**: 38078-38092.

Lazzerini-Denchi E, Sfeir A. 2016. Stop pulling my strings—what telomeres taught us about the DNA damage response. *Nature reviews Molecular cell biology* **17**: 364-378.

Lee BY, Kim J, Lee J. 2022a. Intraspecific de novo gene birth revealed by presence–absence variant genes in Caenorhabditis elegans. *NAR Genomics and Bioinformatics* **4**: lqac031.

Lee BY, Kim J, Lee J. 2022b. Long-read sequencing infers a mechanism for copy number variation of template for alternative lengthening of telomeres in a wild C. elegans strain. *microPublication Biology*.

Li H. 2011a. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987-2993.

Li H. 2011b. Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics* **27**: 718-719.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Lomsadze A, Burns PD, Borodovsky M. 2014. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic acids research* **42**: e119-e119.

Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Marchler GH, Song JS. 2020. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic acids research* **48**: D265-D268.

Lundblad V, Blackburn EH. 1993. An alternative pathway for yeast telomere maintenance rescues est1− senescence. *Cell* **73**: 347-360.

Lydeard JR, Jain S, Yamaguchi M, Haber JE. 2007. Break-induced replication and telomerase-independent telomere maintenance require Pol32. *Nature* **448**: 820-823.

Malkova A, Ivanov EL, Haber JE. 1996. Double-strand break repair in the absence of RAD51 in yeast: a possible role for break-induced DNA replication. *Proceedings of the National Academy of Sciences* **93**: 7131-7136.

Marchler-Bauer A, Bryant SH. 2004. CD-Search: protein domain annotations on the fly. *Nucleic acids research* **32**: W327-W331.

Medlin L, Elwood HJ, Stickel S, Sogin ML. 1988. The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene* **71**: 491-499.

Meier B, Barber LJ, Liu Y, Shtessel L, Boulton SJ, Gartner A, Ahmed S. 2009. The MRT-1 nuclease is required for DNA crosslink repair and telomerase activity in vivo in Caenorhabditis elegans. *The EMBO journal* **28**: 3549-3563.

Meier B, Clejan I, Liu Y, Lowden M, Gartner A, Hodgkin J, Ahmed S. 2006. trt-1 is the Caenorhabditis elegans catalytic subunit of telomerase. *PLoS Genetics* **2**: e18.

Mravinac B, Meštrović N, Čavrak VV, Plohl M. 2011. TCAGG, an alternative telomeric sequence in insects. *Chromosoma* **120**: 367-376.

Peška V, Fajkus P, Fojtová M, Dvořáčková M, Hapala J, Dvořáček V, Polanská P, Leitch AR, Sýkorová E, Fajkus J. 2015. Characterisation of an unusual telomere motif (TTTTTTAGGG) n in the plant Cestrum elegans (Solanaceae), a species with a large genome. *The Plant Journal* **82**: 644-654.

Peska V, Garcia S. 2020. Origin, diversity, and evolution of telomere sequences in plants. *Frontiers in plant science* **11**: 117.

Raices M, Verdun RE, Compton SA, Haggblom CI, Griffith JD, Dillin A, Karlseder J. 2008. C. elegans telomeres contain G-strand and C-strand overhangs that are bound by distinct proteins. *Cell* **132**: 745-757.

Richards EJ, Ausubel FM. 1988. Isolation of a higher eukaryotic telomere from Arabidopsis thaliana. *Cell* **53**: 127-136.

Rotková G, Skleničková M, Dvořáčková M, Sýkorová E, Leitch AR, Fajkus J. 2004. An evolutionary change in telomere sequence motif within the plant section Asparagales had significance for telomere nucleoprotein complexes. *Cytogenetic and Genome Research* **107**: 132-138.

844 Roumelioti FM, Sotiriou SK, Katsini V, Chiourea M, Halazonetis TD, Gagos S. 2016. Alternative
845 lengthening of human telomeres is a conservative DNA replication process with
846 features of break-induced replication. *EMBO reports* **17**: 1731-1737.
847 Seo B, Kim C, Hills M, Sung S, Kim H, Kim E, Lim DS, Oh H-S, Choi RMJ, Chun J. 2015.
848 Telomere maintenance through recruitment of internal genomic regions. *Nature*
849 *communications* **6**: 1-10.
850 Sepsiova R, Necasova I, Willcox S, Prochazkova K, Gorilak P, Nosek J, Hofr C, Griffith JD,
851 Tomaska L. 2016. Evolution of telomeres in Schizosaccharomyces pombe and its
852 possible relationship to the diversification of telomere binding proteins. *PLoS One* **11**:
853 e0154225.
854 Shakirov EV, Song X, Joseph JA, Shippen DE. 2009. POT1 proteins in green algae and land
855 plants: DNA-binding properties and evidence of co-evolution with telomeric DNA.
856 *Nucleic acids research* **37**: 7455-7467.
857 Shtessel L, Lowden MR, Cheng C, Simon M, Wang K, Ahmed S. 2013. Caenorhabditis elegans
858 POT-1 and POT-2 repress telomere maintenance pathways. *G3: Genes| Genomes|*
859 *Genetics* **3**: 305-313.
860 Sievers F, Higgins DG. 2021. The clustal omega multiple alignment package. In *Multiple*
861 *sequence alignment*, pp. 3-16. Springer.
862 Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing
863 genome assembly and annotation completeness with single-copy orthologs.
864 *Bioinformatics* **31**: 3210-3212.
865 Smythe AB, Holovachov O, Kocot KM. 2019. Improved phylogenomic sampling of free-living
866 nematodes enhances resolution of higher-level nematode phylogeny. *BMC*
867 *Evolutionary Biology* **19**: 1-15.
868 Sobinoff AP, Pickett HA. 2017. Alternative lengthening of telomeres: DNA repair pathways
869 converge. *Trends in Genetics* **33**: 921-932.
870 Srinivasan J, Dillman AR, Macchietto MG, Heikkinen L, Lakso M, Fracchia KM, Antoshechkin I,
871 Mortazavi A, Wong G, Sternberg PW. 2013. The draft genome and transcriptome of
872 Panagrellus redivivus are shaped by the harsh demands of a free-living lifestyle.
873 *Genetics* **193**: 1279-1295.
874 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
875 large phylogenies. *Bioinformatics* **30**: 1312-1313.
876 Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped
877 cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**: 637-644.
878 Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with
879 a generalized hidden Markov model that uses hints from external sources. *BMC*
880 *bioinformatics* **7**: 1-11.
881 Steinberg-Neifach O, Lue NF. 2015. Telomere DNA recognition in Saccharomycotina yeast:
882 potential lessons for the co-evolution of ssDNA and dsDNA-binding proteins and
883 their target sites. *Frontiers in Genetics* **6**: 162.
884 Sýkorová E, Fajkus J, Mezníková M, Lim KY, Neplechová K, Blattner FR, Chase MW, Leitch AR.
885 2006. Minisatellite telomeres occur in the family Alliaceae but are lost in Allium.
886 *American journal of botany* **93**: 814-823.

887 Sýkorová E, Lim K, Kunická Z, Chase M, Bennett M, Fajkus J, Leitch A. 2003. Telomere
888 variability in the monocotyledonous plant order Asparagales. *Proceedings of the Royal*
889 *Society of London Series B: Biological Sciences* **270**: 1893-1904.
890 Tomáška Ľ, Nosek J, Sepšiová R, Červenák F, Juríková K, Procházková K, Neboháčová M,
891 Willcox S, Griffith JD. 2019. Commentary: single-stranded telomere-binding protein
892 employs a dual rheostat for binding affinity and specificity that drives function.
893 *Frontiers in Genetics*: 742.
894 Tran TD, Cao HX, Jovtchev G, Neumann P, Novák P, Fojtová M, Vu GT, Macas J, Fajkus J,
895 Schubert I. 2015. Centromere and telomere sequence alterations reflect the rapid
896 genome evolution within the carnivorous plant genus Genlisea. *The Plant Journal* **84**:
897 1087-1099.
898 Van Steensel B, Smogorzewska A, De Lange T. 1998. TRF2 protects human telomeres from
899 end-to-end fusions. *Cell* **92**: 401-413.
900 Visacka K, Hofr C, Willcox S, Necasova I, Pavlouskova J, Sepsiova R, Wimmerova M,
901 Simonicova L, Nosek J, Fajkus J. 2012. Synergism of the two Myb domains of Tay1
902 protein results in high affinity binding to telomeres. *Journal of Biological Chemistry*
903 **287**: 32206-32215.
904 Weiss H, Scherthan H. 2002. Aloe spp.–plants with vertebrate-like telomeric sequences.
905 *Chromosome Research* **10**: 155-164.
906 Wong B. 2011. Color blindness. *nature methods* **8**: 441.
907 Yamamoto I, Zhang K, Zhang J, Vorontsov E, Shibuya H. 2021. Telomeric double-strand DNA-
908 binding proteins DTN-1 and DTN-2 ensure germline immortality in Caenorhabditis
909 elegans. *Elife* **10**: e64104.
910 Yoshimura J, Ichikawa K, Shoura MJ, Artiles KL, Gabdank I, Wahba L, Smith CL, Edgley ML,
911 Rougvie AE, Fire AZ. 2019. Recompleting the Caenorhabditis elegans genome.
912 *Genome research* **29**: 1009-1022.
913 Zhong Z, Shiue L, Kaplan S, de Lange T. 1992. A mammalian factor that binds telomeric
914 TTAGGG repeats in vitro. *Molecular and cellular biology* **12**: 4834-4843.
915 Zhou Y, Wang Y, Xiong X, Appel AG, Zhang C, Wang X. 2022. Profiles of telomeric repeats in
916 Insecta reveal diverse forms of telomeric motifs in Hymenopterans. *Life Science*
917 *Alliance* **5**.

918

919

Telomeric repeat evolution in the phylum Nematoda revealed by high-quality genome assemblies and subtelomere structures

Jiseon Lim, Wonjoo Kim, Jun Kim, et al.

Genome Res. published online November 2, 2023
Access the most recent version at doi:10.1101/gr.278124.123