Robust Binding Capability and Loss of Telomere-Binding Proteins Underlying the Evolution of Telomeric Repeat Motifs

Hobum Song[1], Seonhong Kim[2], Daisy S. Lim[3], Hee-Jung Choi[2, *], and Junho Lee[1, 2, *]

1 Institute of Molecular Biology and Genetics, Seoul National University, Seoul 08826, Korea

2 Department of Biological Sciences, Seoul National University, Seoul 08826, Korea

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

Email address

- Hobum Song: hobum1@snu.ac.kr
- Seonhong Kim: seonhongk93@snu.ac.kr
- Daisy S. Lim: daisylim@snu.ac.kr
- Hee-Jung Choi: choihj@snu.ac.kr
- Junho Lee: elegans@snu.ac.kr

*Correspondences to:

Junho Lee

Department of Biological Sciences

Seoul National University, Seoul 08826, Korea

E-mail: elegans@snu.ac.kr

Phone: +821029442663

Keywords: nematode, telomere-binding protein, telomere repeat motif, evolution, binding specificity

Abstract

Telomeres, the nucleoprotein complexes that protect the ends of linear chromosomes, are essential for maintaining the stability of eukaryotic genomes. As telomeres generally consist of repetitive DNA associated with specifically bound proteins, telomeric repeat motifs are thought to be difficult to evolve. However, a recent study identified nematodes with telomeric repeats distinct from the canonical TTAGGC motif. Here, we investigated how telomere repeats could have evolved despite the challenge posed by the specificity of telomere-binding proteins (TBPs) to the telomeric DNA in Nematoda. We performed a phylogenetic analysis and electrophoresis mobility shift assays to assess the binding affinities of two TBPs, which displayed different conservation patterns. Our results revealed that the well-conserved protein CEH-37 exhibits limited specificity, unable to distinguish telomeric repeats found in nematodes except for the TTAGGG motif, while the less conserved POT proteins displayed rigid specificity. These findings suggest that the emergence of novel telomeric repeat motifs correlated with the characteristics and evolutionary outcomes of TBPs in Nematoda. Furthermore, we confirmed that the homeodomain of CEH-37 exhibits structural homology with the Myb-related domains found in double-stranded telomeric DNA binding proteins. Our study not only revealed the dynamics of telomere evolution but also enhanced the understanding of the evolutionary relationship between proteins and DNAs.

Introduction

Genome stability in eukaryotes is inherently challenged by the continuous shortening of chromosomal ends and their misrecognition as double-strand breaks (1-3). To address these problems, eukaryotes have nucleoprotein structures at the ends of their linear chromosomes called telomeres. Telomere-binding proteins (TBPs) specifically bind to DNA sequences generally composed of G-rich repeat motifs, forming complexes that cap the chromosomal ends to prevent the activation of the DNA damage response and regulate the length of telomeres, ensuring functional telomeres. Consequently, changes in telomeric repeat motifs (TRMs) can lead to genomic instability by affecting the binding affinities of TBPs, making the evolution of telomeric sequences challenging (4).

Despite its challenging nature, telomeric sequence evolution has been observed in several lineages, including yeast, plants, and insects (5-7). Various hypotheses have been proposed to explain how this process was possible, and some of these hypotheses have been experimentally validated in certain organisms (4, 8). In yeast, the adoption of flexible TBPs to cope with rapid changes in telomeric sequences has been observed, while the coevolution of TRM and TBP has been the principle of telomere evolution in plants (5, 9, 10). In Animalia, however, no studies have been conducted to explain the relationship between TBPs and the evolution of telomeric sequences.

Recently, novel TRMs have been discovered within the Nematoda phylum, which includes one of the representative model organisms *Caenorhabditis elegans* (11). While it has been well established that the nematode TRM is TTAGGC, some species, such as *C. uteleia*, *Panagrellus redivivus* and diverse isolates in Panagrolaimidae family (e.g. LJ2284, LJ2285), and *Strongyloides ratti*, have been revealed to possess noncanonical TRMs (TTAGGT, TTAGAC, and TTAGGG, respectively) by analysis of whole genome sequence (WGS) data (Fig. 1A). Since many TBPs of *C. elegans* have been identified, *C. elegans* and related species could be good model organisms to study evolution of telomeres.

In this study, we aimed to understand how TRM evolved in Nematoda despite the challenges posed by sequence-specificities of TBPs. Based on our phylogenetic analysis, we focused on CEH-37, a well-conserved TBP, and less conserved TBPs. The binding specificities of these proteins varied according to their conservation patterns – CEH-37 exhibited flexibility, while the POT proteins displayed relatively rigid specificity. These results indicate that the flexibility or the loss and gain of TBPs may coincide with the emergence of novel TRMs. In addition, homeodomain of the evolutionary conserved CEH-37 shows structural similarity to representative double-stranded-telomeric DNA binding proteins (dsTBPs), suggesting that CEH-37 has been a major TBP in nematodes with no Myb-related proteins since ancient times.

Results

Conservation of *C. elegans* telomere-binding proteins in Nematoda

In eukaryotes, telomeric DNAs are capped by protein complexes (12, 13). In *C. elegans*, several TBPs have been identified, and TEBP-1, TEBP-2, POT-1, POT-2, and MRT-1 are thought to form a telomeric protein complex (14-21). However, there are some species in Nematoda found to possess noncanonical TRMs (11). Since TBPs need to bind effectively to telomeric DNA for telomeres to function properly, we investigated which TBPs of *C. elegans* have been conserved across other nematode species despite differences in their TRMs.

As nematodes are classified into five clades based on their small rRNA phylogeny (22), we conducted BLAST analysis of TBPs from representative species of each clade, including the ones with noncanonical TRMs (Fig. 1A and Table S1). Using a bit score threshold of 50 in BLASTP to indicate significant homology (23), we observed that the TBPs of *C. elegans* are well conserved within clade V, while it was not the same in other clades. Notably, CEH-37 and HRPA-1 are well conserved across all clades of Nematoda. In contrast, the major subunits of the *C. elegans* telomeric protein complex, TEBP and POT proteins, are not conserved in clade IV, which contains species with the TTAGAC TRM, as well as clades I and II. This suggests that the species in clades other than III and V must possess a telomeric protein complex distinct from that of *C. elegans*. Our analysis shows that TBPs experienced different selection pressure during the evolution and speciation of nematodes.

Conserved homeodomain of CEH-37 robustly binds to telomeric DNAs with limited binding specificity

We focused on CEH-37, the only dsTBP conserved regardless of TRMs to investigate how TBPs dealt with the emergence of the novel telomeric DNAs. CEH-37 is a homeodomain protein that binds to double-stranded DNA using its homeodomain, which is conserved regardless of the strain's TRM (Fig. 1B and 1C). To investigate how this protein has remained conserved despite changes in telomeric sequences, we obtained the recombinant CEH-37 homeodomains from nematodes with different TRMs (Fig. S1) and measured their binding affinities to telomeric DNAs identified in Nematoda.

First, we compared the binding affinities of CEH-37 homeodomains from *C. elegans* and *C. uteleia*, which have canonical TRM TTAGGC and noncanonical TRM TTAGGT, respectively, using electrophoresis mobility shift assay (EMSA) (Fig. 2, Table S2 and S3). Interestingly, CEH-37 homeodomains from both species bound to telomeric DNAs with TTAGGC, TTAGAC, and TTAGGT repeats, with no significant difference in their dissociation constants (K_d). However, both CEH-37

homeodomains did not effectively bind to DNA with TTAGGG repeats. These results indicate that CEH-37 does not distinguish TTAGGC, TTAGAC, and TTAGGT motifs, while it shows less preference for TTAGGG motif.

We also measured the binding affinities of CEH-37 homeodomains from strains in clade IV, where species with other noncanonical TRMs were identified (Fig. 3 and S2, Table S2 and S3). The Panagrolaimidae family in clade IV includes isolates LJ2284 and LJ2285, which possess noncanonical TRM TTAGAC, and a related species LJ2400, which has canonical TRM TTAGGC. Measurement of K_d values of CEH-37 homeodomains from LJ2284, LJ2285, and LJ2400 showed consistent results with the above experiments, indicating robust binding to telomeric DNA with limited sequence specificity, irrespective of the TRM of their origin species. These results suggest that TRMs in Nematoda evolved within the range of the well-conserved CEH-37 sequence specificity.

As our results show that CEH-37 does not effectively bind to telomeric DNA with TTAGGG repeats, we measured the binding affinities of CEH-37 homeodomains from *S. ratti*, the parasitic nematode in Clade IV expected to possess noncanonical TRM TTAGGG (11). Interestingly, the CEH-37 homeodomain from *S. ratti* also did not bind efficiently to TTAGGG-repeated DNA compared to other nematode telomeric DNAs (Fig. 3D). Contrary to a previous study, this result suggests that *S. ratti* may not have TTAGGG as its TRM, or it may rely on a different protein instead of CEH-37.

Nematodes with novel TRMs need alternative mechanisms to protect their G-rich telomeric overhangs other than capping by POT proteins.

TEBP and POT proteins form a telomeric protein complex, which is essential for telomere protection in *C. elegans* (15, 24). Interestingly, while these proteins are well conserved in *C. uteleia*, a clade V nematode species with distinct TTAGGT TRM, they are not conserved in clade IV nematodes which include other noncanonical TRMs (Fig. 1A). To investigate how these proteins contributed to the evolution of TRMs, we obtained *C. elegans* POT proteins, POT-2 and POT-3, and assessed their binding affinities to nematode telomeric DNAs (Fig. 4A, 4B, and S3A-C). The EMSA results indicate that both POT-2 and POT-3 bind efficiently to TTAGGC repeats. However, they did not bind to the DNA probe composed of TTAGGT repeats or TTAGAC repeats. These results indicate that POT proteins exhibit high sequence specificity in binding telomeric DNA, failing to bind to the novel nematode telomeric repeats.

For the emergence of novel TRMs, POT proteins must either coevolve or be lost. As the results of TBP conservation indicate, POT proteins are not conserved in clade IV nematodes, suggesting that the TTAGAC TRM in clade IV arose by the loss of POT proteins (Fig. 1A). To confirm whether this observation is a general phenomenon in clade IV nematodes, we performed a BLASTP search of every *C. elegans* POT protein (POT-1, POT-2, and POT-3) against the protein database of clade IV

nematodes, including Panagrolaimidae isolates LJ2284 and LJ2285, which possess the TTAGAC TRM (Table S4). As expected, POT proteins were not conserved in any clade IV nematodes, regardless of their TRM. This result suggests that the loss of *pot* genes preceded the emergence of the noncanonical TTAGAC TRM in clade IV. Similar to the earlier case of CEH-37, the loss of POT protein may have reduced selective pressure against the TTAGAC TRM, facilitating the emergence of the noncanonical TRM in Panagrolaimidae isolates.

While the emergence of the TTAGAC TRM can be attributed to the loss of major subunits of the telomeric protein complex, the emergence of the TTAGGT TRM in *C. uteleia* still raises the possibility of coevolution. Unlike clade IV nematodes, BLASTP analysis has shown that *C. uteleia* retains homologs of *C. elegans* POT-1 (CSP31.g24055) and POT-2 and POT-3 (CSP31.g1277). Since CSP31.g1277 (137) showed a higher bit score with *C. elegans* POT-2 and POT-3 than CSP31.g24055 (73.6) did with POT-1, we purified recombinant CSP31.g1277 to test whether it binds to its host telomeric DNA (Fig. S3D and S3E). Surprisingly, CSP31.g1277 failed to bind to any nematode telomeric DNAs, indicating that it does not function as a TBP (Fig. 4C). Since the binding affinity of CSP31.g24055 to telomeric DNAs has not yet been tested, it remains possible that *C. uteleia* possesses C-rich overhangs where CSP31.g24055 may bind, or that alternative proteins or mechanisms are involved in protecting G-rich telomeric overhangs.

CEH-37 is the only well-conserved double-stranded telomeric DNA binding protein that has the typical structure of double-stranded telomeric DNA binding proteins in nematodes.

Telomeric complexes contain dsTBP subunits, which generally have Myb-related domains as their DNA-binding domains (DBDs) (12). In *C. elegans*, TEBP proteins are identified to use Myb-containing domains to bind DNA (15, 25). These proteins are conserved within the *Caenorhabditis* genus, although there may be variations in the number of these proteins (24). While TEBP proteins are expected to form telomeric complexes, they are not conserved in nematodes outside of clade V (Fig. 1A). To maintain functional telomeres, where protein complexes cap the telomeric DNA, other TBPs might bind to the telomeric DNA in nematodes from clades I to IV.

CEH-37 homeodomain has been hypothesized to share structural homology with Myb-related domains from dsTBPs in other species, although they do not exhibit sequence homology (21, 26). To investigate whether there are additional dsTBPs in nematodes with a similar structure to the Myb-related domain, we quantitatively assessed the structural similarity of C. elegans dsTBPs with the Myb domain of the representative human TRF1 (hTRF1) using TM-align algorithm (Fig. 5 and Table S5). According to the alignment results, the CEH-37 homeodomain, along with TEBP-1 and TEBP-2, whose third Myb-containing domains (MCDs) function as DBDs, showed structural homology with the hTRF Myb domain, as they exceeded the TM-score cutoff of 0.5 (25, 27).

As mentioned above, CEH-37 is a well-conserved TBP that binds telomeric DNA through its homeodomain (Fig. 1B). To determine whether CEH-37 could function as a major dsTBP in nematode species that lack TEBP proteins, we compared its structure with representative dsTBPs from human, yeasts, and *C. elegans* (Fig. S4). CEH-37 homeodomain and DBDs of dsTBPs from other species exhibited TM-scores over 0.5, indicating the similarities in their protein folds. This result suggests that CEH-37 may have served as a key dsTBP in nematodes without TEBP proteins, given that CEH-37 is the only conserved dsTBP with structural homology to Myb-related domains.

Discussion

Chromosome ends are protected by the intricate interaction between sequence-specific TBP and telomeric DNA. However, the relationship between the evolution of TRM and TBPs in animals has been elusive. In this study, we examined how TRMs could evolve in the challenge posed by sequence-specificity of TBPs in Nematoda. Phylogenetic analysis and measurements of binding affinities of CEH-37 and POT proteins suggested that the presence of TBPs with its sequence-specificity exert selective forces on TRM evolution. Furthermore, CEH-37 is speculated to be the major dsTBP in the nematode species lacking Myb-related dsTBP TEBPs.

The evolution of telomeric sequences and their associated proteins has also been observed in yeasts. Yeasts exhibit a wide range of TRMs, from common TTAGGG repeats to long (>20 bp) and heterogeneous TRMs. This extreme diversity accompanied the adoption of flexible TBPs, such as Rap1 and Taz1, instead of the strictly sequence-specific Teb1/Tay1 (9, 13). As the new TBP emerged (i.e. Taz1) or the telomere-associated protein became a multifunctional TBP by gaining the ability to bind DNA directly (i.e. Rap1), the original TBPs either became dedicated to their secondary role as transcription factors or disappeared. The evolution of telomeres in yeasts, where TBPs have both telomeric functions and roles as transcription factors, is reminiscent of the nematode CEH-37, which is also known for its roles as a transcription factor involved in sensory neuron specification and intestinal immunity in *C. elegans* (28-30).

The emergence of the TTAGAC TRM in clade IV nematodes can be attributed not only to the flexibility of CEH-37 but also to the absence of specifically binding POT proteins, which hardly bind to TTAGAC repeats. Unlike most eukaryotes, which possess oligonucleotide/oligosaccharide-binding (OB) fold POT proteins to protect their telomeric overhangs, clade IV nematodes lack these POT proteins (Table S4). Clade IV nematodes likely evolved adaptations to compensate for the loss of POT proteins. Various strategies for protecting single-stranded telomeric DNA in the absence of POT proteins have been observed in other organisms. In budding yeast, telomeric overhangs are capped by the Cdc13-Stn1-Ten complex in place of POT proteins (13). In plants, single-stranded telomeric DNA binding proteins (ssTBPs) with or without OB folds have been identified, while some species possess blunt-ended telomeres (31, 32). These precedents suggest that investigating how nematodes adapt to the loss of POT proteins could be a compelling area for further study, with the potential to discover new mechanisms.

Given that the nematode telomeric protein complex is known to comprise TEBP and POT proteins (15, 24), it was surprising that a homolog of the 3' telomeric overhang-binding POT proteins from *C. uteleia*, a unique clade V nematode with a distinct TTAGGT TRM, hardly binds to telomeric DNA. Since the 3' telomeric overhang-binding POT proteins do not directly associate with the TEBP proteins in *C. elegans* (15), their functional significance might be reduced in the complex. Consequently, *C. uteleia* may employ other proteins to protect its 3' telomeric overhangs or lack such

overhangs. In contrast, *C. elegans* POT-1, which binds to the 5' telomeric overhangs, directly interacts with TEBP proteins (15, 24). Therefore, it is plausible that CSP31.g24055, the *C. uteleia* homolog of POT-1, binds to the novel telomeric DNA. This could have occurred either through the flexible binding specificity of *C. elegans* POT-1 or via the coevolution of CSP31.g24055 within the emergence of the novel TTAGGT TRM. Evidence supporting the former possibility comes from the case of CEH-37 in this study, as well as the flexibility observed in yeast RAP1 and Taz1 (9, 13). The latter one is supported by studies of plant evolution, where POT1 in green algae only binds to TTTAGGG repeats, while the POT1 of land plants possessing the TTAGGG TRM binds to both TTAGGG and TTTAGGG repeats (10). Identifying the mechanism by which the *C. uteleia* TRM arose would be an intriguing study.

We confirmed that CEH-37 homeodomains from every species, including *S. ratti*, do not effectively bind to TTAGGG repeats compared to other telomeric DNAs found in Nematoda, raising questions about the nature of the telomere in *S. ratti*. Previous WGS data speculated that *S. ratti* possesses a TTAGGG TRM, as this was the only TTAGGC-related motif identified in the species (11). However, TRM-containing concatemers were rare compared to TRM-containing reads from WGS data of other nematodes analyzed in the study, ranking 67th out of 67 species. Based on this information, we hypothesize that 1) *S. ratti* possesses short telomeric DNAs consisting of TTAGGG repeats, possibly protected by TBPs other than CEH-37 or by other mechanisms (e.g. adoption of rigid G-quadruplex (G4) structures), or that 2) *S. ratti* does not possess TTAGGG TRM, and the observed reads from the parasite may have originated from its mammalian host (33-35). If the first hypothesis is true, it would be interesting to investigate the role of G4 structure in telomeres by comparing telomeres of *S. ratti*, nematodes with TTAGGC TRMs which are also known to form telomeric G4 structures, and nematodes with TTAGAC TRMs which are unlikely to form telomeric G4 structures (36).

The main difference between CEH-37 and other dsTBPs is that CEH-37 uses a homeodomain instead of Myb-related domains to bind to DNA. However, its structure is similar to Myb-related domains of TBPs in other species, indicating convergent evolution (Table S5). Meanwhile, the identification of dsTBPs in many species has been predominantly focused on proteins with Myb-related domains, even though protein function is primarily determined by structure rather than sequence (37-39). The presence of CEH-37 suggests the possibility of dsTBPs with domains other than Myb-related domains existing in other species as well.

In summary, we demonstrated the dynamics of nematode telomere evolution and highlighted the role of TBPs in the evolution of telomere sequences across different nematode species. We expect our study to not only advance the field of telomere biology using nematodes as multicellular organism models but also enhance our understanding of the evolution of DNA-protein interactions beyond telomeres.

Experimental procedures

Strains

Nematodes were grown in nematode growth medium seeded with *Escherichia coli* strain OP50 at 20℃ as standard methods (40). *C. elegans* Bristol N2, *C. uteleia* JU2585, and Panagrolaimidae isolates (LJ2284, LJ2285, LJ2400, and LJ2406) were used for this study.

Conservation analysis of telomere-binding proteins

FASTA protein files of *C. elegans* (PRJNA13758) and *C. uteleia* (PRJEB12600) were obtained from WormBase ParaSite (41, 42). Protein sequences of Panagrolaimidae isolates (LJ2284, LJ2285, LJ2400, and LJ2406) were kindly produced by Lim et al (11).

To analyze conservation of telomere binding proteins of *C. elegans*, we used DIAMOND (version 2.1.8; diamond blastp -d -q -o --threads 20 --very-sensitive and diamond blastp -d -q -o --threads 20 --ultra-sensitive) (43). The conservation was evaluated by using highest bit score protein sequence with conserved domains of original proteins identified using NCBI CD-search (CDD database, version 3.21) (44-46). Multiple sequence alignment was performed by Clustal Omega using Jalview 2.11.4 (47-49). Additionally, BLAST+ provided by WormBase ParaSite on its web server or DIAMOND as mentioned above were used to analyze conservation of TEBP-POT complex subunits in clade IV nematodes (50, 51). Query sequences were used as written on Table S4 and BLAST was searched against the clade IV nematode protein database.

Construct generation

ceh-37 homeobox regions of each strain were cloned in a modified pGEX-4T-1 vector containing a HRV 3C cleavage site after the GST tag while *C. elegans pot-2* and *pot-3* were cloned in a modified pMJ806 vector containing TEV cleavage site after the MBP tag. To generate cDNA of desired genes, RNAs isolated from nematodes with TRIzol were reverse transcripted with PrimeScript RT Master Mix (Takara) except for *S. ratti*. Since we were not able to culture *S. ratti*, *ceh-37* homeobox cDNA of *S. ratti* was generated by ligating two annealed ~110bp oligos after phosphorylation. cDNAs were fused into the vector by Gibson assembly master mix (New England Biolabs).

Protein expression and purification

The plasmid containing the desired gene was transformed into *E. coli* Rosetta (DE3) cells.

Colonies were picked and inoculated into Terrific Broth, and the culture was grown at 37°C until the OD_{600} reached approximately 0.6. Protein overexpression was induced by adding IPTG to a final concentration of 0.2 mM, followed by incubation at 20°C for 18 h. The cells were harvested and resuspended in PBS, and cell lysis was performed using sonication, supplemented with PMSF and DNase I. The lysed cells were centrifuged at 14,000 rpm for 15 min, and the supernatants were collected. The supernatants from cells expressing CEH-37 or POT proteins were incubated with glutathione agarose resin or amylose agarose resin, respectively, pre-equilibrated with PBS for 1 h with gentle rolling. The resin was washed with 10 column volumes of 20 mM HEPES pH 7.5, 100 mM NaCl buffer. While the proteins were bound to the resin, HRV3C or TEV protease was added for overnight digestion at 4°C. The following day, the protein was released using the wash buffer, and subsequently loaded onto a HiTrap SP (CEH-37)/Q (POT proteins) HP 5 mL column (GE Healthcare). An appropriate salt gradient was applied using 20 mM HEPES pH 7.5 buffer, and peaks were analyzed by SDS-PAGE. The desired fractions were pooled and concentrated. Final purification was performed by loading the concentrated protein onto a Superdex 200 10/300 GL column (GE Healthcare) equilibrated with 20 mM HEPES pH 7.5, 150 mM NaCl. The peaks were analyzed by SDS-PAGE, and the desired fractions were collected and concentrated to the appropriate concentration. The concentrated protein solution was aliquoted, flash-frozen in liquid nitrogen, and stored at -80°C.

During the purification of *C. uteleia* protein CSP31.g1277, the protein precipitated after protease treatment due to the minimal difference between the isoelectric point (7.93) of tag-free CSP31.g1277 and the pH of the buffer (7.5). To address this, we initially purified His-MBP-tagged CSP31.g1277 using a HiTrap Q HP 5 mL column followed by a Superdex 200 10/300 GL column. Subsequently, the buffer was exchanged to 20 mM HEPES pH 7, 150 mM NaCl using dialysis tubing, and His-tagged TEV protease was added for cleavage at 4℃ overnight. The tag and protease were removed by Ni-NTA resin (Qiagen), and the supernatant containing purified CSP31.g1277 was collected for further experiments.

Electrophoretic mobility shift assay (EMSA)

Telomeric DNA substrates used as DNA probes were prepared by annealing oligos with forward strand having Cy5 fluorescent on its 5' end. 5 nM of DNA probes were incubated with appropriate concentration of proteins in binding buffer (20 mM Tris-HCl, 50 mM NaCl, 1 mM MgCl2, 5 mM DTT, 5% glycerol, 0.001% tween-20, and 50 ug/mL BSA) for 15 min in room temperature. Samples were loaded to 6% native gel that had been pre-run. Electrophoresis was performed for 90 min at 130 V in cold room with 0.5x TBE as a running buffer.

Structure prediction and alignment

To predict biomolecular structures, we utilized the AlphaFold Server powered by AlphaFold 3 model (52). Alignments of CEH-37 homeodomains from species with distinct TRMs were performed using the Matchmaker tool in UCSF ChimeraX (version 1.5) (53, 54).

Structural alignments were conducted using the web-based pairwise structure alignment service provided by the RCSB PDB website, employing the TM-align method (27). The structures of each protein were obtained from models stored in the AlphaFold Protein Structure Database (55, 56). The regions of proteins selected for the alignments were pre-determined DNA-binding domains specified in previous studies (44-46, 57-59).

Supporting information

This article contains supporting information.

Acknowledgements

We are grateful to Dr. J. Lim providing the sequencing data of Panagrolaimidae family isolates (LJ2284, LJ2285, LJ2400), and Dr. M-A Félix for providing *C. uteleia* JU2585. We also thank the members of Dr. J. Lee (laboratory of genes and development) and Dr. H-J Choi (laboratory of structural biology) for helpful discussion during the study.

Author contributions

H.S., H.-J.C., and J.L. conceptualization; H.S. data curation; H.S. formal analysis; J.L. funding acquisition; H.S., S.K., and D.S.L. investigation; H.S., S.K., D.S.L., and H.-J.C. methodology; H.-J.C. and J.L. supervision; H.S. visualization; H.S. writing-original draft; H.S., S.K., D.S.L., and J.L. writing-review and editing

Funding and additional information

This research was supported by the National Research Foundation of Korea (NRF) grant (2020R1A2C3003352) funded by the Korean Ministry of Science and ICT.

Conflict of interest

The authors declare that they have no conflicts of interest with the contents of this article.

References

1. Watson, J. D. (1972) Origin of concatemeric T7DNA. *Nat New Biol.* **239**, 197-201

2. 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. *J Theor Biol.* **41**, 181-190

3. De Lange, T. (2009) How telomeres solve the end-protection problem. *Science* **326**, 948-952

4. Steinberg-Neifach, O., and Lue, N. F. (2015) Telomere DNA recognition in Saccharomycotina yeast: potential lessons for the co-evolution of ssDNA and dsDNA-binding proteins and their target sites. *Front Genet.* **6**, 162

5. Červenák, F., Sepšiová, R., Nosek, J., and Tomáška, Ľ. (2021) Step-by-step evolution of telomeres: lessons from yeasts. *Genome Biol Evol.* **13**, evaa268

6. Procházková Schrumpfová, P., Schořová, Š., and Fajkus, J. (2016) Telomere-and telomerase-associated proteins and their functions in the plant cell. *Front Plant Sci.* **7**, 851

7. Kuznetsova, V., Grozeva, S., and Gokhman, V. (2020) Telomere structure in insects: A review. *J Zool Syst Evol Res.* **58**, 127-158

8. Saint-Leandre, B., and Levine, M. T. (2020) The telomere paradox: stable genome preservation with rapidly evolving proteins. *Trends Genet.* **36**, 232-242

9. Sepsiova, R., Necasova, I., Willcox, S., Prochazkova, K., Gorilak, P., Nosek, J. *et al.* (2016) Evolution of telomeres in Schizosaccharomyces pombe and its possible relationship to the diversification of telomere binding proteins. *PLoS One* **11**, e0154225

10. Shakirov, E. V., Song, X., Joseph, J. A., and Shippen, D. E. (2009) POT1 proteins in green algae and land plants: DNA-binding properties and evidence of co-evolution with telomeric DNA. *Nucleic Acids Res*. **37**, 7455-7467

11. Lim, J., Kim, W., Kim, J., and Lee, J. (2023) Telomeric repeat evolution in the phylum Nematoda revealed by high-quality genome assemblies and subtelomere structures. *Genome Res.* **33**, 1947-1957

12. De Lange, T. (2005) Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev.* **19**, 2100-2110

13. Moser, B. A., and Nakamura, T. M. (2009) Protection and replication of telomeres in fission yeast. *Biochem Cell Biol.* **87**, 747-758

14. Raices, M., Verdun, R. E., Compton, S. A., Haggblom, C. I., Griffith, J. D., Dillin, A. *et al.* (2008) C.

elegans telomeres contain G-strand and C-strand overhangs that are bound by distinct proteins. *Cell* **132**, 745-757

15. Yamamoto, I., Zhang, K., Zhang, J., Vorontsov, E., and Shibuya, H. (2021) Telomeric double-strand DNA-binding proteins DTN-1 and DTN-2 ensure germline immortality in Caenorhabditis elegans. *Elife* **10**, e64104

16. Im, S. H., and Lee, J. (2003) Identification of HMG-5 as a double-stranded telomeric DNA-binding protein in the nematode Caenorhabditis elegans. *FEBS Lett.* **554**, 455-461

17. Im, S. H., and Lee, J. (2005) PLP-1 binds nematode double-stranded telomeric DNA. *Mol Cells.* **20**, 297-302

18. Joeng, K. S., Song, E. J., Lee, K.-J., and Lee, J. (2004) Long lifespan in worms with long telomeric DNA. *Nat Genet.* **36**, 607-611

19. Meier, B., Barber, L. J., Liu, Y., Shtessel, L., Boulton, S. J., Gartner, A. *et al.* (2009) The MRT-1 nuclease is required for DNA crosslink repair and telomerase activity in vivo in Caenorhabditis elegans. *EMBO J* **28**, 3549-3563

20. Yu, X., Gray, S., and Ferreira, Helder C. (2022) POT-3 preferentially binds the terminal DNA-repeat on the telomeric G-overhang. *Nucleic Acids Res.* **51**, 610-618

21. Kim, S. H., Hwang, S. B., Chung, I. K., and Lee, J. (2003) Sequence-specific binding to telomeric DNA by CEH-37, a homeodomain protein in the nematode Caenorhabditis elegans. *J Biol Chem.* **278**, 28038-28044

22. Blaxter, M. L., De Ley, P., Garey, J. R., Liu, L. X., Scheldeman, P., Vierstraete, A. *et al.* (1998) A molecular evolutionary framework for the phylum Nematoda. *Nature* **392**, 71-75

23. Pearson, W. R. (2013) An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics.* **42**, 3.1.1-3.1.8

24. Dietz, S., Almeida, M. V., Nischwitz, E., Schreier, J., Viceconte, N., Fradera-Sola, A. *et al.* (2021) The double-stranded DNA-binding proteins TEBP-1 and TEBP-2 form a telomeric complex with POT-1. *Nat Commun.* **12**, 2668

25. Padmanaban, S., Lambacher, N. J., Tesmer, V. M., Zhang, J., Shibuya, H., and Nandakumar, J. (2024) Caenorhabditis elegans telomere-binding proteins TEBP-1 and TEBP-2 adapt the Myb module to dimerize and bind telomeric DNA. *Proc Natl Acad Sci U S A.* **121**, e2316651121

26. Moon, S., Lee, Y. W., Kim, W. T., and Lee, W. (2014) Solution structure of CEH-37 homeodomain of the nematode Caenorhabditis elegans. *Biochem Biophys Res Commun.* **443**, 370-375

27. Zhang, Y., and Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302-2309

28. Shore, D., and Nasmyth, K. (1987) Purification and cloning of a DNA binding protein from yeast that binds to both silencer and activator elements. *Cell* **51**, 721-732

29. Lanjuin, A., VanHoven, M. K., Bargmann, C. I., Thompson, J. K., and Sengupta, P. (2003) Otx/otd homeobox genes specify distinct sensory neuron identities in C. elegans. *Dev Cell.* **5**, 621-633

30. Liu, F., Zhang, H., Wang, H., Zhu, X., Li, S., Jiang, N. *et al.* (2023) The homeodomain transcription factor CEH-37 regulates PMK-1/p38 MAPK pathway to protect against intestinal infection via the phosphatase VHP-1. *Cell Mol Life Sci.* **80**, 312

31. Luo, M., Luo, H., Li, H., Liu, D., and Lu, H. (2020) Single-stranded DNA-binding proteins in plant telomeres. *Int J Biol Macromol* **165**, 1463-1467

32. Kazda, A., Zellinger, B., Rössler, M., Derboven, E., Kusenda, B., and Riha, K. (2012) Chromosome end protection by blunt-ended telomeres. *Genes Dev.* **26**, 1703-1713

33. Zahler, A. M., Williamson, J. R., Cech, T. R., and Prescott, D. M. (1991) Inhibition of telomerase by G-quartet DMA structures. *Nature* **350**, 718-720

34. Smith, J. S., Chen, Q., Yatsunyk, L. A., Nicoludis, J. M., Garcia, M. S., Kranaster, R. *et al.* (2011) Rudimentary G-quadruplex–based telomere capping in Saccharomyces cerevisiae. *Nat Struct Mol Biol.* **18**, 478-485

35. Bryan, T. M. (2020) G-quadruplexes at telomeres: friend or foe? *Molecules* **25**, 3686

36. Marquevielle, J., De Rache, A., Vialet, B., Morvan, E., Mergny, J.-L., and Amrane, S. (2022) G-quadruplex structure of the C. elegans telomeric repeat: a two tetrads basket type conformation stabilized by a non-canonical C–T base-pair. *Nucleic Acids Res.* **50**, 7134-7146

37. Li, B., Espinal, A., and Cross, G. A. (2005) Trypanosome telomeres are protected by a homologue of mammalian TRF2. *Mol Cell Biol* **25**, 5011-5021

38. Yu, E. Y., Kim, S. E., Kim, J. H., Ko, J. H., Cho, M. H., and Chung, I. K. (2000) Sequence-specific DNA recognition by the Myb-like domain of plant telomeric protein RTBP1. *J Biol Chem.* **275**, 24208-24214

39. Fulcher, N., and Riha, K. (2016) Using centromere mediated genome elimination to elucidate the functional redundancy of candidate telomere binding proteins in Arabidopsis thaliana. *Front Genet* **6**, 349

40. Brenner, S. (1974) The genetics of Caenorhabditis elegans. *Genetics* **77**, 71-94

41. Howe, K. L., Bolt, B. J., Cain, S., Chan, J., Chen, W. J., Davis, P. *et al.* (2016) WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Res.* **44**, D774-D780

42. Howe, K. L., Bolt, B. J., Shafie, M., Kersey, P., and Berriman, M. (2017) WormBase ParaSite− a comprehensive resource for helminth genomics. *Mol Biochem Parasitol.* **215**, 2-10

43. Buchfink, B., Reuter, K., and Drost, H.-G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods.* **18**, 366-368

44. Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S. *et al.* (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200-D203

45. Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R. *et al.* (2020) CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* **48**, D265-D268

46. Wang, J., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, S. *et al.* (2022) The conserved domain database in 2023. *Nucleic Acids Res.* **51**, D384-D388

47. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., andBarton, G. J. (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench Bioinformatics **25**, 1189-1191

48. Troshin, P. V., Procter, J. B., andBarton, G. J. (2011) Java bioinformatics analysis web services for multiple sequence alignment—JABAWS:MSA Bioinformatics **27**, 2001-2002

49. Troshin, P. V., Procter, J. B., Sherstnev, A., Barton, D. L., Madeira, F., andBarton, G. J. (2018) JABAWS 2.2 distributed web services for Bioinformatics: protein disorder, conservation and RNA secondary structure Bioinformatics **34**, 1939-1940

50. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-3402

51. Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I. *et al.* (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**, 2994-3005

52. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A. *et al.* (2024) Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500

53. Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H. *et al.* (2018) UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14-25

54. Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I. *et al.* (2021) UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70-82

55. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589

56. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G. *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-

sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439-D444

57. König, P., Giraldo, R., Chapman, L., and Rhodes, D. (1996) The crystal structure of the DNA-binding domain of yeast RAP1 in complex with telomeric DNA. *Cell* **85**, 125-136

58. Sigrist, C. J., De Castro, E., Langendijk-Genevaux, P. S., Le Saux, V., Bairoch, A., and Hulo, N. (2005) ProRule: a new database containing functional and structural information on PROSITE profiles. *Bioinformatics* **21**, 4060-4066

59. Uniprot Consortium (2023) UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523-D531

Abbreviations

The abbreviations used are: DBD, DNA-binding domain; dsTBP, double-stranded telomeric DNA binding protein; EMSA, electrophoresis mobility shift assay; G4, G-quadruplex; K_d, dissociation constant; MCD, Myb-containing domain; OB, oligonucleotide/oligosaccharide-binding; ssTBP, single-stranded telomeric DNA binding protein; TBP, telomere-binding proten; TRM, telomeric repeat motif; WGS, whole genome sequence.

Figures

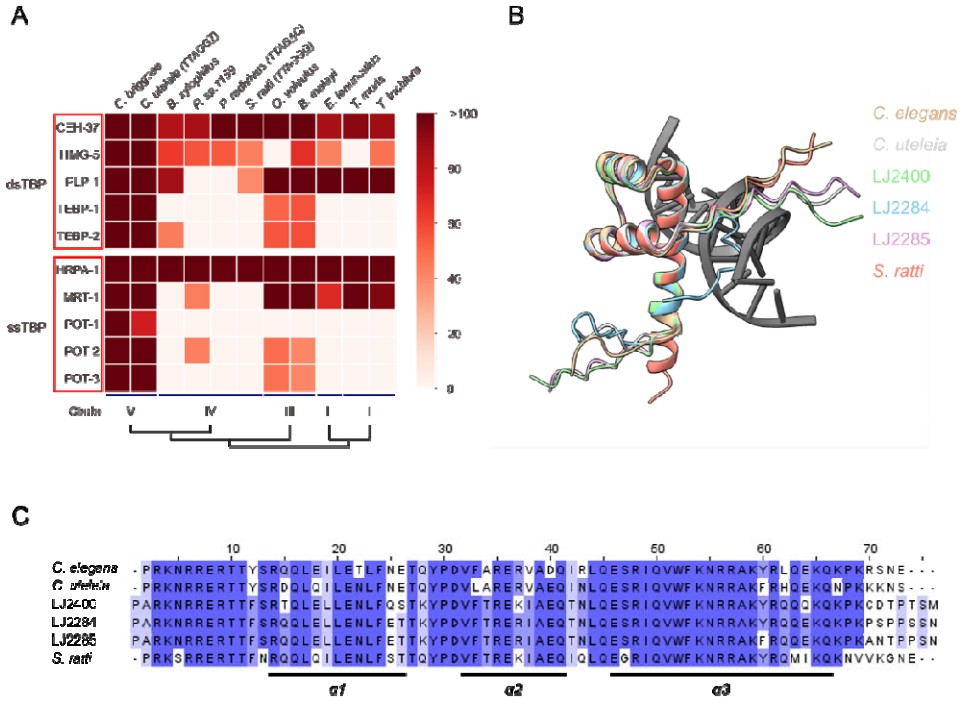

Figure 1. Conservation of telomere binding proteins in nematodes. (A) Heatmap illustrating the conservation of *C. elegans* TBPs in various nematodes according to the bit score. For nematodes with noncanonical TRMs, the TRMs are indicated in brackets after the name of the species. Nematodes with a canonical TTAGGC TRM are not explicitly labeled. Phylogenetic tree under the heatmap is based on small rRNA phylogeny, which serves as the standard for defining nematode clades. (dsTBP: double-stranded telomeric DNA binding protein, ssTBP: single-stranded telomeric DNA binding protein.) (B) Model of the DNA-CEH-37 homeodomain complex of *C. elegans* with CEH-37 homeodomains from other nematodes aligned to it. The color of the species names corresponds to the color of the CEH-37 homeodomains. (C) Multiple sequence alignment result of CEH-37 homeodomains from nematodes used in this study. The brightness of residue colors indicates the percentage identity in each column to the consensus sequence. Each α-helix region is indicated under the result of multiple sequence alignment. Sequence alignment was performed using Clustal Omega.

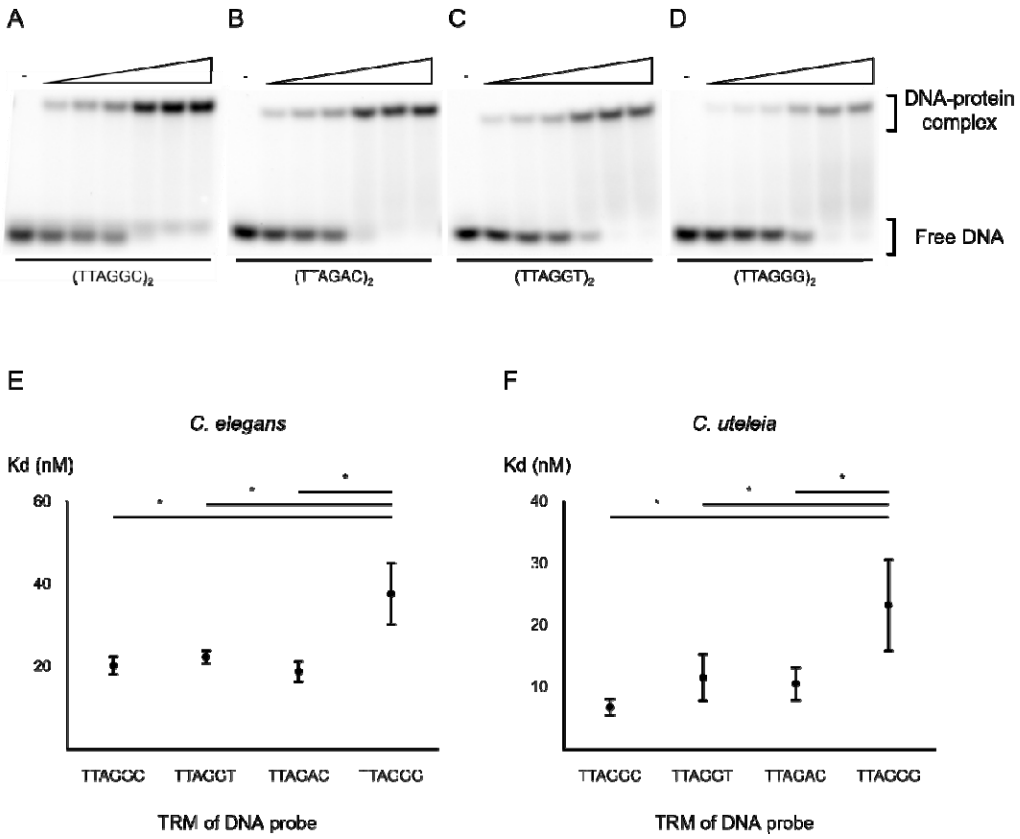

Figure 2. Binding affinity of CEH-37 homeodomains from the *Caenorhabditis* genus. (A-D) Representative images of EMSA demonstrating the binding of the *C. elegans* CEH-37 homeodomain to telomeric DNAs. 5 nM of DNA probes were incubated with increasing concentrations of proteins ranging from 5 nM to 160 nM. (E-F) K_d of *C. elegans* and *C. uteleia* CEH-37 homeodomains to each telomeric DNAs, respectively. DNA probes used for the K_d measurements consist of 2 repeats of the indicated sequences. Relationship without asterisks indicate non-significant differences (Newman-Keuls multiple comparison test, * $p<0.05$).

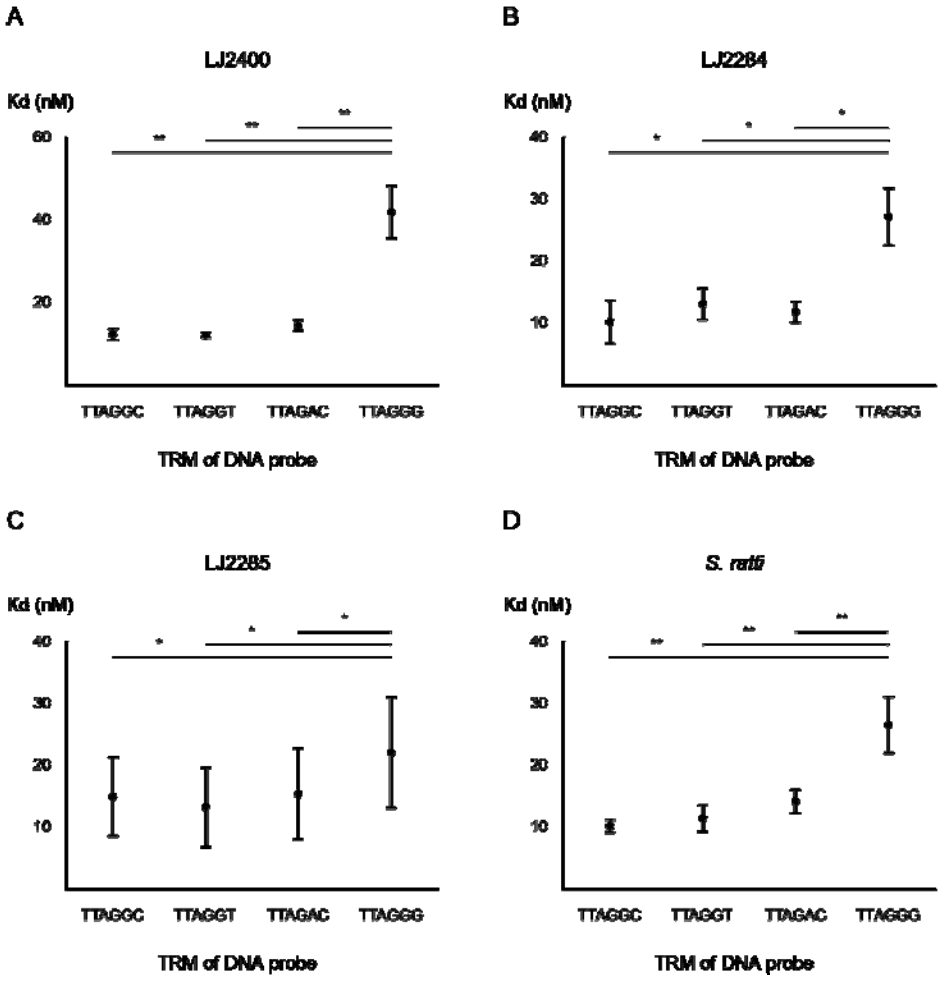

Figure 3. Binding affinity of CEH-37 homeodomains from clade IV nematodes. (A-D) K_d of CEH-37 homeodomains from LJ2400, LJ2284, LJ2285, and *S. ratti* binding to each telomeric DNA, respectively. DNA probes used for the K_d measurements consist of 2 repeats of the indicated sequences. Relationship without asterisks indicate non-significant differences (Newman-Keuls multiple comparison test, * $0.01<p<0.05$, ** $0.005<p<0.01$, *** $p<0.005$).

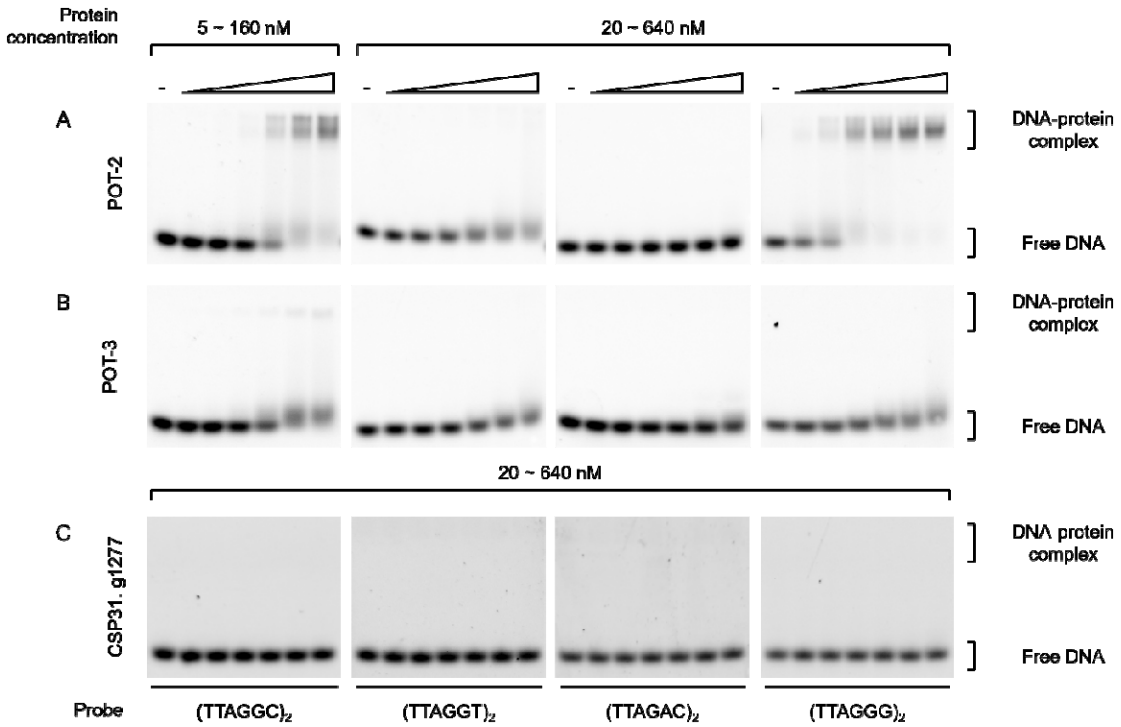

Figure 4. Analysis of binding affinity of recombinant *C. elegans* POT-2 and POT-3 to nematode telomeric repeats. (A, B, C) Representative images of EMSA demonstrating the binding of the *C. elegans* POT-2, POT-3, and *C. uteleia* CSP31.g1277 (a homolog protein of *C. elegans* POT-2 and POT-3) to telomeric DNAs, respectively. 5 nM of DNA probes were incubated with increasing concentrations of proteins as indicated.

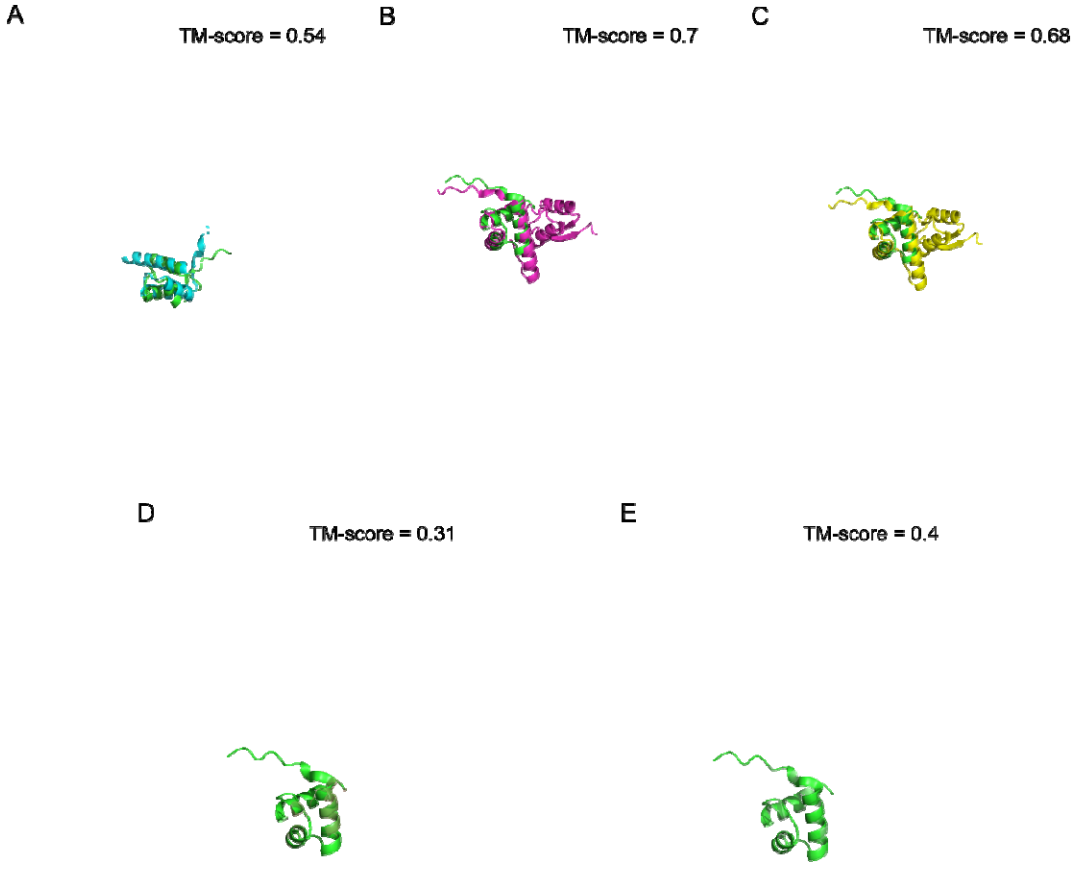

Figure 5. Models of a human TRF1 Myb domain aligned to *C. elegans* TBPs. The Myb domain of human TRF1 (green) is structurally aligned to (A) CEH-37 (cyan), (B) TEBP-1 (magenta), (C) TEBP-2 (yellow), (D) HMG-5 (salmon), and (E) PLP-1 (grey). The unaligned regions of each protein are faintly shown. TM-scores are 0.54, 0.7, 0.68, 0.31, and 0.4 respectively, where TM-score over 0.5 indicates a similar protein fold.