Profiles of telomeric repeats in Insecta reveal diverse forms of telomeric motifs in Hymenopterans

Yihang Zhou1,2,3,* Yi Wang4,5,*, Xiao Xiong1,2,3, Arthur G Appel3,6, Chao Zhang1, Xu Wang2,3,6,7

Telomeres consist of highly conserved simple tandem telomeric repeat motif (TRM): (TTAGG)n in arthropods, (TTAGGG)n in vertebrates, and (TTAGGG)n in most plants. TRM can be detected from chromosome-level assembly, which typically requires long-read sequencing data. To take advantage of short-read data, we developed an ultra-fast Telomeric Repeats Identiﬁcation Pipeline and evaluated its performance on 91 species. With proven accuracy, we applied Telomeric Repeats Identiﬁcation Pipeline in 129 insect species, using 7 Tbp of short-read sequences. We conﬁrmed (TTAGG)n as the TRM in 19 orders, suggesting it is the ancestral form in insects. Systematic proﬁling in Hymenopterans revealed a diverse range of TRMs, including the canonical 5-bp TTAGG (bees, ants, and basal sawﬂies), three independent losses of tandem repeat form TRM (Ichneumonoids, hunting wasps, and gall-forming wasps (TTATTGGGG). Our results identiﬁed extraordinary evolutionary ﬂuidity of Hymenopteran TRMs, and rapid evolution of TRM and repeat abundance at all evolutionary scales, providing novel insights into telomere evolution.

DOI 10.26508/lsa.202101163 | Received 19 July 2021 | Revised 4 March 2022 | Accepted 4 March 2022 | Published online 1 April 2022

Introduction

Telomeres are vital and highly conserved DNA–protein complexes located at the ends of eukaryotic chromosomes that protect chromosome ends from deterioration or fusion with neighboring chromosomes. During DNA replication, an RNA primer is required by the polymerase, and the lagging strand at the chromosomal ends cannot be replicated, shortening the chromosome by 50–100 bp per division in normal diploid human cells (Wright et al, 1997; Chow et al, 2012). To solve this issue, an RNA-dependent DNA polymerase called telomerase can restore the telomere length by adding telomeric sequences using an RNA template. Telomere dynamics and regulation play critical roles in aging, cancer, cell proliferation, and gene repression (Blackburn, 1991, 2000; Zakian, 1995; Verdu & Karlseder, 2007). Misregulation of telomeres may lead to senescence and genomic instability (Greider, 1996; Sahin et al, 2011).

In most eukaryotic organisms, telomeres are comprised of G-rich short tandem telomeric repeat motif (TRM). For example, human telomeres are composed of ~230 Kb region of 6-bp (TTAGGG)n TRM (de Lange, 2004). TRM is highly conserved in animals (Moyzis et al, 1988; Meyne et al, 1989). All vertebrates examined share the same 6-bp TRM (TTAGGG)n and are believed to evolve from a common ancestor over 400 million years ago (Mya) (Meyne et al, 1989; Traut et al, 2007). This high-level TRM sequence conservation is presumably driven by the conserved TRM binding machinery and the telomerase RNA component (TERC). TERC is an RNA component of the telomerase ranging from 312 to 559 nucleotides in size in vertebrates (Logeswaran et al, 2021) and serves as the template for telomere elongation (Counter et al, 1998; Chen et al, 2000; Lundblad, 2002) (Fig 1A). In mammals, six telomere-associated proteins (TRF1, TRF2, POT1, TIN2, TPP1, and RAP1) form a complex named Shelterin, which is an essential protein complex mediating telomere protection (de Lange, 2018).

TRM variants were found in single-cell eukaryotes and basal invertebrates, such as (TGGG)n in Tetrahymena and (TTAGGG)n in Caenorhabditis elegans (Zakian, 1995). Plant TRMs are more variable and, in some cases, longer than 6-bp (Peska & Garcia, 2020). A 7-bp TRM (TTAGGG)n was discovered in Arabidopsis (Richards & Ausubel, 1988) and Nicotiana tabacum (Fajkus et al, 1995), which is the common type found in 44 species belonging to 14 families of angiosperms, gymnosperms, and bryophytes (Fuchs et al, 1995). Many Asparagus species have the vertebrate-type 6-bp (TTAGGG)n, and (TTGGG)n were also found in this order (Sykorova et al, 2003). Two intermingled TRM variants (TTCAGG and TTTCAGG) were reported in the carnivorous genus Genlisea (Tran et al, 2015). An unusually long telomeric sequence (TTTTTGGGG)n was discovered in Cestrum

© 2022 Zhou et al. https://doi.org/10.26508/lsa.202101163
et al., 2004), scarabs (Coleoptera) (Sahara et al., 1999), alder (Lorite et al., 2002), northern caddis (Megaloptera) (Frydrychová et al., 2004), and others. A few motif sequence variations, such as (TCAGG)n, were discovered in Coleoptera (Mravinac et al., 2011). The canonical (TTAGG)n was not observed in the parasitoid wasps (Werren et al., 2010; Gokhman et al., 2013) and Dipteran (Lopez et al., 1999; Gorab, 2003), are known to lack TRM, which could be due to limited resolution to detect short tandem repeats using whole-genome sequencing data. A bioinformatic pipeline named Telomeric Repeats Identification Pipeline (TRIP). The input is short-read whole-genome sequencing (WGS) data with a minimum depth of 10, which is readily available from public databases and can be obtained from genome resequencing. TRIP provides de novo TRM identification, which is not limited by prior knowledge of candidate motifs. Negative results from TRIP are also meaningful, indicating potential loss of simple tandem repeat as TRM and suggesting a complex form of telomere composition, such as collections of retrotransposons in Drosophila (Levis et al., 1993). We applied this pipeline to explore the predicted TRMs of 129 species, primarily focusing on TRM evolution in Insecta. Our methodology will facilitate TRM discovery and will shed light on the telomere function and evolution.

Results

TRIP, an ultra-fast and precise bioinformatic pipeline for TRM identification

TRMs have a high abundance among the short tandem repeats in the genome. According to this feature, we developed a TRIP for TRM identification from Illumina short-read genome sequencing data by profiling short tandem repeats in sequence reads (Fig 1B).
Specifically, RepeatDetector is used to extract 2–25-bp short tandem repeat sequences from Illumina reads. Based on the results, RepeatSummary will reorder the repeat monomers alphabetically, count the total number of repeat-containing reads, and compute the number and the total length for each repeat unit (Fig 1B). TRIP outputs 19 summary statistics and makes inferences of candidate TRMs, based primarily on the number of repeats containing reads, the total tandem repeat length, and the percent of repeat region in the sequencing reads (Table S1). A final TRM is called as the single best candidate with threefold or higher abundance than other repeat units (Table S2). If this condition is not met, the single best candidate will be evaluated by additional evidence from experimental results in the literature or high-quality genome assembly to make the final TRM call (Table S2). We refer the final TRM calls from the TRIP pipeline as predicted TRMs.

To evaluate the performance, we applied the TRIP pipeline on 91 species with TRM information reported in the literature across four phyla and nine classes, including plants, vertebrate and invertebrate species (Supplemental Data 1). For the 41 species with simple tandem type TRMs, TRIP correctly identified all of them from short-read data (Supplemental Data 2). The remaining 50 species are known to have retrotransposon/complex telomeres (e.g., Dipterans) or lack TTAGG/TTAGGG TRMs (Supplemental Data 1). The TRIP pipeline determined the absence of any simple tandem repeat type TRM repeat motifs for 50 species (Supplemental Data 3). For the 91 species, TRIP has 100% accuracy and zero false positives. The average computational processing speed is 10 Kbp of sequencing read data per second per process when run on a Linux machine with Intel Xeon Processor E5-2660 CPU at 2.20 GHz, with negligible memory cost.

Independent identification and confirmation of a previously reported (TTATGGG)n TRM in a jewel wasp species N. vitripennis (Nv)

In 2020, Dalla Benetta et al (2020) assembled the Nv PSR strain using PacBio and Oxford Nanopore long-read technologies, with significantly improved genome completeness (Dalla Benetta et al, 2020). The authors observed TTATGGG repeats at chromosome ends and confirmed the chromosomal termini location using FISH in the PSR strain. We independently identified (TTATGGG)n TRM in a different Nv strain, AsymCx, using the TRIP pipeline with 125.8 Gbp of 10× Genomics genome reshuffling data (Fig 2A). With 426.2× genome coverage, we performed comprehensive profiling of short tandem repeats in the Nv genome. In fact, (TTATGGG)n is the most abundant short tandem repeat in Nv (Fig 2A), accounting for 1.4% of the genome with over 4,111 Kbp in length per haploid genome. We examined its genomic locations by assembling the potential sub-telomeric contigs using pair-ended read data (Supplemental Data 2). The unmapped reads were assembled into sub-telomeric contigs. (C) Assembled sub-telomeric regions were aligned to Nv reference genome assembly v2.1, and they were mapped to eight chromosomal end scaffolds. Sequence identities between sub-telomeric regions and Nv scaffolds were labeled for each hit.

The TTATGGG type of motif was found at 7 of 10 chromosome ends, totaling 21 Kbp in length (Table S4). This number is much less than the TRIP estimate (over 4 Mbp) because the long simple repeat sequences are difficult to assemble, especially at the chromosome ends.

TTAGG is the ancestral TRM across 19 Insecta orders

After confirming the unprecedented long-form TRM (TTATGGG)n, we collected 7 Tbp whole-genome short-read resequencing data of 129 Insecta species across 21 orders (117 from publicly available databases and 12 generated in this research, see Supplemental Data 4). The remaining 10 insect orders were not included in this analysis due to a lack of data (Table S5). If the chromosomal terminal location of the predicted TRM was supported by experimental
evidence or chromosome-level genome assembly, the reference was cited and discussed (Supplemental Data 4). If the terminal position was not demonstrated by experimental support for the identified TRMs, the TRIP call was referred to as predicted TRM (Supplemental Data 4).

(TTAGG)n motifs were reported to be the TRM in non-insect arthropod lineages, such as Chelicerata (spiders, tick, etc.) (Vitkova et al., 2005), Myriapoda (millipedes, centipedes, etc.) (Vitkova et al., 2005), and Crustaceans (crabs, lobsters, etc.) (Vitkova et al., 2005). Both Insecta and Entognatha belong to ecollass Hexapoda, and we examined the three sister groups in Entognatha. (TTAGG)n was previously found in Collombola (springtails) by a staining method (Vitkova et al., 2005). We independently confirmed this finding in Folsomia candida. No telomere research was reported in Diplura, and we discovered (TTAGG)n as the predicted TRM in Campodea ausgen (Supplemental Data 4-6), suggesting this 5-bp unit is the ancestral form of TRM in arthropods (Fig 3).

For the 129 insect species we investigated, 110 candidate TRMs were discovered in 63 species by TRIP according to the criteria summarized in Table S1. Among these 63 species, 78% (49/63) have a single candidate TRM (Supplemental Data 6), and Mantophasmatodea (lack of data). No previous telomere research was reported in them (Fig 3 and Supplemental Data 4) except for Neuroptera (no (TTAGG)n telomere, we identified (TTAGG)n as the TRM in at least one species of 18 orders. Among these 18 orders with probes (results summarized in Fig 3). (TTAGG)n was found as the TRM in at least one species of 18 orders. Among these 18 orders with known (TTAGG)n telomere, we identified the (TTAGG)n TRM in 16 of them (Fig 3 and Supplemental Data 4) except for Neuroptera (no simple tandem repeat type TRM found) and Mantophasmatodea (lack of data). No previous telomere research was reported in Strepsiptera, and we identified (TTAGG)n as the predicted TRM for the first time in Xenos vesparum (Fig 3 and Supplemental Data 4). We also performed whole-genome sequencing of Lepisma saccharina, which is a living fossil in the basal order Zygentoma (see the Materials and Methods section). (TTAGG)n motif was found as the predicted TRM, although a lack of telomerase activity was reported in this silverfish species (Korandova et al., 2014). Collectively, (TTAGG)n was found in 19 insect orders, indicating it is the ancestral form of telomeric tandem repeat in insects.

Dipterans, including fruit flies and mosquitoes, have lost the TRM. Instead, they use retrotransposon elements as telomerases to protect the chromosomal ends from deterioration. We examined 16 Dipteran species, and all of them lack a tandem repeat type predicted TRM (Supplemental Data 6). Among them, Drosophila melanogaster has 13, Aedes albopictus has 20 candidate motifs (Table S6), and none of them were qualified using the TRIP TRM calling requirements (Supplemental Data 6). The TRM in orders Odonata and Ephemeroptera is still unknown, and previous research confirmed the lack of (TTAGG)n (Frydrychova et al., 2004; Kuznetsova et al., 2017). Public sequencing data are available for dragonfly (Libellula Angelina and Ladona fulva) (Consortium isK, 2013), dainty damselfly (Coenagrion scitulum) (Consortium isK, 2013), and mayfly (Cloeon dipterus) (Almudi et al., 2020). However, no predicted TRMs could be identified in any of these species using TRIP (Supplemental Data 1). We concluded that the ancestral TRM type (TTAGG)n is missing in Odonata and Ephemeroptera, which was an independent loss event compared with Dipterans (Fig 3). The mechanisms of chromosome end protection warrant further study in these orders.

**Diverse Hymenoptera TRMs—(TTAGG)n is the ancestral TRM form in Hymenopterans**

(TTAGG)n was reported as the TRM in hymenopterans, and a longer 8-bp form was reported in a single wasp species N. vitripennis (Dalla Benetta et al., 2020), suggesting diverse forms of TRM in this order. To perform a comprehensive investigation of TRMs in Hymenoptera, we analyzed a total of 71 Hymenopteran species/strains in 14 superfamilies, and the results were shown on the Hymenopteran phylogeny (Fig 4). Sawflies are basal Hymenoptera, and the canonical insect TRM (TTAGG)n was found in two species in the superfamily Tenthredinidae using FISH (Gokhman & Kuznetsova, 2018). Genomic short-read data were available in three sawfly superfamilies, Tenthredinidae, Orussoidae, and Cephoidea, and TRIP identified (TTAGG)n as the predicted TRM in all of them (Fig 4 and Supplemental Data 4), providing strong evidence that TTAGG is the ancestral TRM in basal Hymenopterans. (TTAGG)n was also reported as the TRM form in advanced social insect clades, such as Apoidea (bees) and Formicoidea (ants). We confirmed the presence of (TTAGG)n as the TRM using TRIP (Fig 4 and Supplemental Data 4).

**Diverse Hymenoptera TRMs—elongated 8-bp TRMs in multiple parasitoid chalcid wasps with even longer motif variants in fig wasps and miniature wasps**

Parasitoid wasps are the most abundant Hymenopteran and account for 20% of all insect species (Smith et al., 2008). Chalcidoidea is the largest Hymenopteran superfamily with more than half-million estimated species (Heraty, 2017). Among the 22 chalcid families, we generated/collected genome sequencing data for 10 families in 25 species/strains and found 8 or 9-bp TRMs in eight families (Fig 4 and Supplemental Data 4). The 8-bp (TTATTGG)n was discovered to be the predicted TRM in Pteromalidae (Nasonia, Trichomalopsis, and Muscidifurax), Torymidae (Torymus auratus), Ormyridae (Ormyrus ponticus and Ormyrus nitidulus), Megaschmidae (Megaschmid stigmatizans and Bothanomyia dorsalis), Eurytomidae (Eurytoma adlerioides), and Euepilidae (Eupelminus annulatus), suggesting the 8-bp unit is the dominant form in chalcid wasps (Fig 4). Interestingly, the basal family Trichogrammatidae, which are minute egg parasitoids with an adult body length of less than 1 mm (Lindsey et al., 2018), were discovered to have a 9-bp predicted TRM (TTACTTGGG)n by TRIP, with independent data support from two species Trichomalopsis pretiosum and Trichogramma evanescens (Fig 4 and Supplemental Data 4). We infer that this novel TRM originated from the 8-bp ancestral Chalcidoidea form by a single cytosine insertion in the repeat unit because the 9-bp repeat type is Trichogrammatidae specific. Pollinating-fig wasps belong to the family Agraonidae. Their larvae feed inside the fig syconia, and pollinate the fig. This remarkable mutualism originated 66~75 Mya (Heraty et al., 2013; Xiao et al., 2013, 2021; Peters et al., 2018). Ceratosolen solmsi is an important pollinator for its host tropical fig tree Ficus hispida, and we identified a 9-bp candidate TRM (TTATTGGGG)n, which differs from the 8-bp Chalcidoidea motif by a single Guanine addition (Fig 4 and
Supplemental Data 4). Although this candidate TRM did not survive TRIP filtering, we observed at least four TRM regions at the very end of the assembled scaffolds (Xiao et al., 2013) (Table S7). This 9-bp TRM was also found at scaffold end in Ceratosolen fusciceps, another fig wasp species within this genus, providing independent evidence of this 9-bp novel TRM form (Table S7). To check whether this 9-bp TRM is common in all fig-pollinating wasps, we examined PacBio long-read assemblies of species from three additional fig wasp genera (Kradibia, Dolichoris, and Eupristina): Kradibia gibbosae, Dolichoris vasculosae, Eupristina koningsbergeri (Xiao et al., 2021), and Eupristina verticillate (Zhang et al., 2020). The 8-bp (TTATTGGG)n TRM was discovered at scaffold ends for all four species (Table S7), suggesting it is the ancestral form in Agaonidae.

We conclude that the 8 to 9-bp TRM transition occurred in the

Figure 3. (TTAGG)n is the common and ancestral telomeric repeat motif (TRM) in Insecta.

TRM profiles for insects (shaded) and arthropod outgroup species (nonshaded) summarized on the Insecta phylogeny adapted from Misof et al. (2014). Previous reports of presence (+) or absence (−) of (TTAGG)n were listed under the literature column. TRM calls from Telomeric Repeats Identification Pipeline (TRIP) in this study are represented by “+” sign or the “absent” label (no tandem repeat TRM form observed). The “variable” label stands for novel TRM variants found. The branches are shaded in green if the TRIP results agree with previous literature reports in all species within this taxon. The branches are shared in blue if novel TRM information is provided by TRIP. Branches without genomic data are shaded in gray. For TTAGG “+” families identified by TRIP, the total repeat lengths are shown in horizontal bar plots for one representative species per branch.
Figure 4. Profile of telomeric repeat motif (TRM) in Hymenopterans revealed elongated repeat monomers in parasitoid wasps. TRM identified from previous reports using (TTAGG)n probes are summarized in the middle panel next to the Hymenoptera phylogeny at the family level, adapted from Peters et al (2017, 2018). The “+” sign means confirmed (TTAGG)n TRM, and the “−” sign stands for lack of such signal. For Telomeric Repeats Identification Pipeline (TRIP) “+” families, the abundance (Kb per haploid genome) of candidate TRM are shown in horizontal bar plots for representative species. The branches are shaded in green if the TRIP results agree with previous literature report in any species within this taxon. The branches are shared in blue if novel TRM information is provided by TRIP.
common ancestor of *C. solmsi* and *C. fuscicep* after diverging from other Agaonidae genera. Based on a calibrated fig wasp phylogeny of 12 species (Xiao et al., 2021), we estimated this 9-bp TRM occurred between 55 and 43 Mya. Non-pollinating fig wasp species *Apocrypta bakeri* and *Syncebia* sp. (Xiao et al., 2021) in the family Pteromalidae have the 8-bp (TTATTGGG)n TRM (Table S8), which is the same form as other Pteromalidae species we examined. The Ceratosolen-specific TRM sequence turnover occurred within 55 million years after its divergence from *Kladibia*, showing a tendency of increasing TRM length in Chalcid wasps.

**Diverse Hymenoptera TRMs—lack of TRM in Ichneumonoids, hunting wasps, and gall-forming wasps suggests multiple loss events during hymenopteran evolution**

Despite the extensive sampling across Hymenoptera, we did not detect any predicted TRM in Ichneumonoids, hunting wasps, or gall-forming wasps. Ichneumonoidae is the second-largest Hymenoptera superfamily with ~150,000 described species. In this study, nine species in the families Ichneumonidae and Braconidae were explored, and we did not identify any short tandem repeat form TRMs (Fig 4), which is consistent with the negative results from previous research using (TTAGG)n probes (Gokhman et al., 2014), suggesting the loss of ancestral (TTAGG)n telomeric repeats. Solitary hunting wasp families Crabronidae (digger wasps) and Sphecidae (sand wasp) lack TRM in our TRIP results, which are consistent with the negative signals in (TTAGG)n FISH experiments (Menezes et al., 2017). In stingless wasps, TRIP cannot identify any predicted TRMs in three superfamilies (Pompiloidea, Vespoidea, and Chrysidioidea). Together with the negative results of (TTAGG)n in Scolioidae (Menezes et al., 2017) in the literature, we hypothesized that a large-scale loss of simple tandem repeat type of TRM occurred in stingless wasp families (Fig 5 and Supplemental Data 4). Ten species in the gall-forming wasp family Cynipidae were also investigated, and no predicted TRMs were identified, suggesting a different mechanism of chromosome end protection in these herbivores.

**Diverse Hymenoptera TRMs—inferring TRMs by ancestral state reconstruction**

A total of 22 families/superfamilies had no previous research about TRM in the literature, and they also lack genomic data for TRM identification by TRIP (Fig 4). To extend the knowledge of potential TRMs in these taxa, ancestral state reconstruction was performed based on known TRM traits using six maximum parsimony and maximum likelihood methods implemented in PastML (Ishikawa et al., 2019). Among these 22 families/superfamilies, 19 had the same inferred TRM state from all six methods (Fig 5). Based on the ancestral TRM inference, basal sawflies (Pamphilioidea, Xyeloidea, Xiphydioidae, and Siricoidea) have (TTAGG)n form TRM (Fig 5), and the five sting wasp taxa (Thynninoidea, Tiphinoidea, Evanioidea, Trigonaloidea, and Stephanoidea) lack tandem repeat type TRMs (Fig 5). DELTRAN (delayed transformation) and DOWNPASS methods suggest that elongated TRM first appeared ~100 Mya in Trichogrammatidae, whereas the MAP (maximum a posteriori) method supports the occurrence of ~130 Mya in Mymaridae (Fig 5). Estimates from the ACCTRAN (accelerated transformation) method indicate an even more ancient origin ~210 Mya in the ancestor of Platagastroidea (Fig 5). Additional experimental and genomic data are needed to resolve the exact evolutionary origin of the elongated TRMs. The current evidence suggests its origin between 100 and 210 Mya in parasitoid wasps.

**Characterization of TRM using TRIP and chromosome assembly revealed rapid evolution of TRM abundance in the parasitoid wasp family Pteromalidae**

To confirm that the (TTATTGGG)n TRM is the common form rather than a specific case in *Nv*, we performed Illumina short-read sequencing of the genomes of 11 wasp strains in six species of the family Pteromalidae, with an average coverage of 295× (see the Materials and Methods section). (TTATTGGG)n was identified as the predicted TRM in all of them (Fig 6A–E). In addition, we performed PacBio long-read sequencing of four parasitoid wasp species (Xiong et al., 2021) and achieved chromosome-level assemblies (Table S9). These wasps have a haploid karyotype of *n* = 5 (Goodpasture, 1974; Gokhman & Westendorff, 2000; Silva-Junior et al., 2000). (TTATTGGG)n was found at all 10 chromosome termini in all four species (Table S9), serving as a definitive validation of the TRM calls. We also quantified the total telomeric repeat length per haploid genome and discovered that the three *Nv* strains (AsymCx, Labl, and V12.1) have an average length of 3,856 Kbp, a 6.5-fold increase in abundance compared with five other related wasp species (*P = 0.019*, Mann–Whitney U test), despite having the same chromosome number (Fig 6F). *Trichomalopsis*, *Muscidifurax*, and *Nasonia* only diverged within 4.9 Mya, suggesting that rapid TRM expansion occurred in *Nv* within 1 million years since it diverged from its close relative *Nasonia giraulti* (*Ng*).

**Discussion**

**A fast and accurate pipeline for TRM discovery from short-read genome sequencing data**

We developed a TRIP to identify predicted TRM by using massive publicly available short-read sequencing data with the following advantages. It makes use of existing WGS short-read data, whereas the hybridization and FISH approaches require prior knowledge or speculation about the TRM for probe design, which is not capable of identifying novel TRMs or confirming the absence of TRM. The TRIP pipeline can perform de novo TRM detection, and negative results suggest a lack of simple tandem repeat-type telomere. High-quality chromosome-level assembly from a combination of long-read sequencing, Hi-C, and optical mapping can be used for direct identification of TRM at chromosome ends. Still, it is currently cost-prohibitive to use for all species. Not all chromosomes were assembled with intact termini because chromosome ends are more likely to be damaged during DNA extraction and fragmentation. As short-read data
become available for more species, the TRIP pipeline will serve as a standard genome analysis tool for predicted TRM identification and provide strong candidate motifs for other experimental methods. Although long-read sequencing and genome assembly were performed in selected species to support the terminal location of these identified TRMs, further experimental analysis of the chromosome terminus location at the cytogenetic level has yet to be performed.
The effect of interstitial telomere-like sequences on TRIP TRM identification

TRIP is a short tandem repeat–based approach, so it is subject to the issue of spurious calls of interstitial telomeric repeat–like sequences. Identification of highly abundant telomere–like tandem repeats does not mean they serve as telomeric repeats. In this study, we tested TRIP in 91 species spanning nine classes and did not observe any false-positive cases. Although we cannot completely exclude this possibility, interstitial telomeric repeat–like sequences are extremely unlikely to affect TRIP results. This is because we only allow 100% match in tandem motif identification in the sequencing reads. Interstitial telomeric repeats were not known to be under selective pressure so they could mutate freely. Most sequencing reads containing these interstitial telomeric repeat–like were excluded from the repeat counts because of the non–perfect match in the motif sequences. If noncanonical TRMs were identified, experimental confirmation using the Bal31 digestion or FISH approaches, as well as direct identification from chromosome–level assemblies would be needed to validate the findings.

Ancestral (TTAGG)n TRM is conserved in most insect orders

Our extensive profiling of TRMs in major insect orders provided additional evidence to confirm that (TTAGG)n is the ancestral TRM in Insecta, as reported and summarized in the literature (Sahara et al, 1999; Sasaki & Fujiwara, 2000; Lorite et al, 2002; Frydrychova et al, 2004; Vitkova et al, 2005; Korandova et al, 2014; Menezes et al, 2017;
Kuznetsova et al., 2019; de Castro et al., 2020). In addition to the insect orders, we discovered the (TTAGG)n TRM in Diplura, which is the immediate outgroup of Insecta (Fig 3). Nineteen of 25 informative Insecta orders have the (TTAGG)n motif, which confirms (TTAGG)n as the ancestral Insecta TRM, through this largest-scale survey across Insecta to date.

Rapid evolving elongated TRM in parasitoid wasps indicates intriguing evolutionary fluidity of the Hymenopteran TRM

TRM sequences are highly conserved. In addition to the multiple insect orders we discussed, ancestral TRM (TTAGG)n was also identified in Chelicerata, Myriapoda, Protura, and Diplura, suggesting that it remained unchanged for more than 550 million years, before the Cambrian explosion. In deuterostomes, (TTAGGG)n was found to be the TRM in all species investigated so far (Jain & Cooper, 2010; Kondo & Akasaka, 2012; Bliznina et al., 2021), which diverged ~600 Mya (Delsuc et al., 2018). In contrast, the telomere repeat motif in parasitoid wasps evolved from (TTAGG)n to a long repeat pattern (TTATTTGGG)n, which occurred 110–220 Mya. An even longer TRM, (TTACTTGGG)n, was identified in Trichogrammatidae, by adding a C in the AT-rich region, which was presumably occurred ~100 Mya (Fig 4). As the common form in parasitoid wasps, (TTATTTGGG)n was also found in E. verticillata, a pollinating-fig wasp species in the family Agraonidae. Another 9-bp variant (TTATTGGGG)n was identified in Ceratosolen, which is an example of novel TRM in the same family within 55 million years. Compared to the unaltered insect and deuterostome ancestral TRMs (TTAGG and TTAGGG), the TRM sequence turnovers we identified in parasitoid Chalcid wasps are extremely recent, suggesting rapid evolution of telomere sequences.

Concerted sequence changes in the TERC template domain and telomere-binding proteins in parasitoid wasps?

Telomerase is responsible for restoring the telomere length after each cell division. As a reverse transcriptase, telomerase is a ribonucleoprotein using its RNA component (TERC) as a template to elongate the telomere. TERC RNA lacks sequence conservation even among closely related species, and the functional conservation is presumably achieved through secondary structures (Logeswaran et al., 2021). The repeat template regions in the TERC RNA have 100% sequence identity to the TRM. However, changes in TRM sequence during evolution may not require alterations in the conserved TERC template region. For example, the major forms of insect (TTAGG) and deuterostome (TTAGGG) TRMs are compatible with the TTA(G)n pattern, which is TTA followed by adjacent Gs. They could be synthesized from the same TERC core template TTAGGG (Fig 1). Surprisingly, the newly discovered elongated TRMs in parasitoid wasps (TTATTTGGG and others) are not compatible with the ancestral TTA(G)n pattern. Therefore, sequence evolution in the TERC template must have occurred, unless a unique and complicated template shift mechanism is responsible for the synthesis of the longer TRM variants. The fig wasp TRM (TTATTTGGG) may share the same TERC template as TTATTTGGG containing species, but the 9-bp Trichogrammatidae TRM contains an extra C in the middle of the unit (TTACTTGGG), and an insertion of a C is expected to occur in the 8-bp ancestral Chalcidoidea TERC template region. In addition to TERC evolution, there are more than 500 telomere-interacting proteins in human and mouse, many of which directly bind the (TTAGGG)n sequences (Shore, 1997; Luo et al., 2015; Braun et al., 2018). The rapid evolution of TRM in chalcid wasps may cause concerted evolution in the binding sites of telomere-binding proteins, to maintain the critical function of telomere elongation and regulation. Characterizing the TERC RNA telomerase, and telomere-binding proteins in Hymenopterans will allow researchers to test these hypotheses formally and elucidate telomere-associated RNA and protein evolution.

New insights on the canonical TRM pattern

The canonical TRM discovered in previous studies could be summarized as the formula T_{104}A_{20}G_{n} (TTAGG, TTAGGG, TTTAGG, etc.) (Peska & Garcia, 2020). There are a few exceptions, but the outliers were not commonly found in a monophyletic group. Our discoveries shed new light on the TRM formula, and the common pattern is refined as W_{104}G_{2}W_{104}G_{n} (W: A/T, b = 0 or 1, d = 2–4), which consists of an A/T region followed by a G-overhang with 2–4 consecutive Gs. At most, a single C insertion can be tolerated in the A/T region to accommodate the (TTAGG)n TRM found in Coleoptera, and the (TTACTTGGG)n motif in Trichogramma discovered in this research. This formula is also consistent with the TRM discovered in yeast (Wellinger & Zakian, 2012) and green algae (Fulneckova et al., 2012). However, in Saccharomycetaceae yeast, a large number of diverse long forms of G-quadruplexes related TRM were found through tandem repeat predictions (Cervenak et al., 2019; Peska et al., 2021). The TTATTTGGGCTCGG telomeric repeat discovered in Allium plants is another long-form TRM (Fajkus et al., 2016). These outliers were not consistent with the formula we proposed.

Origin of elongated and variable TRMs in parasitoid wasps, and diverse forms of telomeric repeats in Hymenopterans

In Hymenopterans, the canonical 5-bp (TTAGG)n is likely to be the ancestral form TRM because it is identified in all basal sawfly taxa. At least three independent TRM loss events occurred in Ichneumonoids, hunting wasps, and gall-forming wasps at ~210 Mya. Surprisingly, we discovered an 8-bp (TTATTTGGG)n TRM common in parasitoid species, which is 60% longer than known Metazoan motifs. The 8-bp or longer TRM forms were only present in Chalcidoidea, which is a superfamily of parasitoid wasps that experienced massive radiation ~200 Mya (Peters et al., 2017). It is a mystery why rapid sequence evolution occurred on the highly conserved telomeric repeats with essential functions to protect the chromosome termini. In addition, total TRM repeat length expansion was observed within the genus and species level (Fig 6), indicating active evolution within 1 million years. The evolutionary origin and sequence turnover of these long-form TRMs warrant further study.

Materials and Methods

The TRIP workflow

TRIP was designed for the identification of candidate TRMs using automated tandem repeats abundance profiling from short-read
sequencing data 75–250 bp in length. With provided species information, TRIP can automatically create a local repository, fetch data from European Nucleotide Archive and NCBI, process the short-read data, generate summary tables, visualize the repeats distribution, and return the final TRM identification results. The read processing step is achieved by an ultra-fast tandem repeat detector RepeatMaster. RepeatMaster is composed of RepeatDetector and RepeatSummary. The WGS read data in FASTQ or compressed format will be processed by RepeatDetector to output an intermediate table of all short tandem repeats with unit size ranging from 2 to 25 bp. Subsequently, RepeatSummary will summarize and quantify the repeat abundance in a tab-delimited table for downstream TRIP analysis. RepeatMaster is offered as independent software, which enables TRIP to handle local data in addition to publicly available data. TRIP will compute 19 parameters (Table S1) from RepeatMaster output, five of which are used to characterize and identify the TRM candidates (Table S1). A set of two filtering criteria is applied to make the final TRM call (Table S2). The sequencing processing speed is ~10 Kbp per second per process with negligible memory cost. The tested working environment is Linux 2.6.32 using Intel Xeon Processor E5-2660 CPU at 2.20 GHz. TRIP is open-source under GPL-3.0 License.

Data acquisition from publicly available databases

Short-read WGS data used in this research were downloaded from NCBI (NCBI Resource Coordinators, 2018), Sequence Read Archive (Leinonen et al., 2011), and European Nucleotide Archive (Harrison et al., 2021). We selected sequencing datasets with at least 10× haploid genome coverage (can be specified using `avg_genome_cov` parameter in TRIP), preferably with a reference genome assembly. The accession number for all datasets used were listed in Supplemental Table S10 from RepeatMaster output, Table for downstream TRIP analysis. RepeatMaster is offered as open-source under GPL-3.0 License.

The quality and size of DNA was determined using the Agilent Genomic DNA ScreenTape Assay. The DNA integrity number (DIN) is 9.3, suggesting perfect DNA integrity. A 10× Genomic library was constructed using the Chromium Genome Reagent Kits v2 on the Chromium Controller (10x Genomics Inc.) with 1.2 ng input DNA. After quality control, the library was sequenced on an Illumina HiSeq X machine at the Genomic Services Lab at the HudsonAlpha Institute for Biotechnology.

A total of 895 million 150-bp reads were obtained (134.2 Gbp of sequences). An initial de novo assembly of *L. saccharina* genome was performed by the Supernova assembler version v2.1.1 (Weisenfeld et al., 2017) using linked reads information. To improve the assembly, we carried out an additional de novo assembly using MEGAHIT version 1.2.9 (Li et al., 2015), using high-quality reads trimmed by Trimomatic v0.39 (Bolger et al., 2014). These two assemblies were combined using quickmerge (Chakraborty et al., 2016). Scaffolds with a size greater than 2 Kb were kept in the final assembly.

**Genome sequencing and assembly of parasitoid wasp species**

The genomes of 11 strains of parasitoid wasps were sequenced in our previous research (Wang et al., 2020; Lin et al., 2021; Xiong et al., 2021). These strains are highly inbred and are maintained in our laboratory under constant lighting conditions at 25°C. *N. vitripennis* reference strain (AsymCx), a laboratory strain derived from isolates in the Netherlands (LabI), and a laboratory strain originated from Rochester, New York (V12.1), were sequenced using 10× Genomics approach. Six strains from three closely related species in the *Nasonia* genus were also included: *N. giraulti* RVX strain, *N. giraulti* G10 strain with *Ng* nuclear genome and *Nv* mitochondrial genome background, *N. giraulti* V19_008U strain collected in Virginia, *Nasonia oneida* NONY 11/36 strain, *N. oneida* NONYAUD108 strain, and *Nasonia longicornis* MTMC13A strain. Two species outside the *Nasonia* genus, Trichomalopsis sarcophaga, and Muscidifurax unicurator, were selected for DNA extraction and genome sequencing.

**Reconstruction of ancestral TRM states in Hymenopterans**

Construction of ancestral TRM states was performed using PastML (Ishikawa et al., 2019) with six methods, including three maximum parsimony ACCTRAN (accelerated transformation), DELTRAN (delayed transformation), and DOWNPASS, as well as three maximum likelihood methods JOINT, MAP (maximum a posteriori), and MPPA (marginal posterior probabilities approximation). The Newick tree file and TRM traits (TTAGG, absent, and novel) were provided as input.

**Data Availability**

The TRIP pipeline can be downloaded at https://github.com/XuWangLab/TRIP; http://github.com/XuWangLab/2020_insect_Telmere sppData. The genome assembly of *L. saccharina* can be accessed at NCBI using accession number https://www.ncbi.nlm.nih.gov/nuccore/JAGDQP000000000, and the raw sequencing data have been deposited to NCBI database and

---

Elongated and rapid evolving telomeric motifs in parasitoids Zhou et al. https://doi.org/10.26508/lsa.202101163 vol 5 | no 7 | e202101163 11 of 15

Supplementary Information

Supplementary information is available at https://doi.org/10.26508/lsa.202101163.

Acknowledgements

This project is supported by an Auburn University IGP Award 180271 and the USDA National Institute of Food and Agriculture Hatch project 1018100. X Wang is supported by National Science Foundation EPSCoR RII Track-4 Research Fellowship (1928770), an Alabama Agriculture Experiment Station (AAES) ARES Agriculture Research Enhancement, Exploration and Development (AgR-SEED) award, and a laboratory start-up fund from Auburn University College Veterinary Medicine. C Zhang is supported by grants from the National Key Research and Development Program of China (Grant No. 2017YFA0103902 & 2019YFA0111400), the National Natural Science Foundation of China (Grant No. 31712183), and Innovative Research Team of High-level Local Universities in Shanghai (NO. SSMU-ZDCX2018700). Y Zhou and X Xiong are supported by the Auburn University Presidential Graduate Research Fellowship and College of Veterinary Medicine Dean’s Fellowship. Part of Fig 1 was created with http://BioRender.com. We thank Dr. John H. Werren from the University of Rochester for his valuable discussions and suggestions. We acknowledge the Auburn University Easley Cluster for support of this work.

Author Contributions

Y Zhou: data curation, software, formal analysis, validation, investigation, visualization, methodology, and writing—original draft, review, and editing.
Y Wang: software, methodology, and writing—review and editing.
X Xiong: investigation and writing—review and editing.
A Appel: resources and writing—review and editing.
C Zhang: conceptualization, resources, supervision, funding acquisition, and writing—review and editing.
X Wang: conceptualization, resources, data curation, formal analysis, supervision, funding acquisition, investigation, visualization, methodology, project administration, and writing—original draft, review, and editing.

Conflict of Interest Statement

The authors declare that they have no conflict of interest.

References

Elongated and rapid evolving telomeric motifs in parasitoids  Zhou et al.

https://doi.org/10.26508/lsa.202101163  vol 5 | no 7 | e20210163


License: This article is available under a Creative Commons License (Attribution 4.0 International, as described at https://creativecommons.org/licenses/by/4.0/).