Published in final edited form as:

Trop Plant Biol. 2008 December ; 1(3-4): 202–215. doi:10.1007/s12042-008-9018-x.

Analysis of *Carica papaya* Telomeres and Telomere-Associated Proteins: Insights into the Evolution of Telomere Maintenance in Brassicales

E. V. Shakirov,

Department of Biochemistry and Biophysics, Texas A&M University, 2128 TAMU, College Station, TX 77843-2128, USA

S. L. Salzberg,

Center for Bioinformatics and Computational Biology, and Department of Computer Science, University of Maryland, 3125 Biomolecular Sciences Bldg, College Park, MD 20742, USA

M. Alam, and

Advanced Studies in Genomics, Proteomics and Bioinformatics, and Department of Microbiology, University of Hawaii, Honolulu, HI 96822, USA

D. E. Shippen

Department of Biochemistry and Biophysics, Texas A&M University, 2128 TAMU, College Station, TX 77843-2128, USA, dshippen@tamu.edu

Abstract

Telomeres are terminal regions of linear eukaryotic chromosomes that are critical for genome stability and continued cell proliferation. The draft assembly of the papaya genome provides an opportunity to analyze and compare the evolution of telomeric DNA sequence composition and telomere maintenance machinery in this and other organisms of the Brassicales Order, which includes Arabidopsis. Here we investigate telomere size and sequence variation at papaya chromosome ends. As with most other plant species, papaya telomeres consist of TTTAGGG repeats. However, in contrast to members of the closely related Brassicaceae family, telomeres in papaya are ~10-fold longer. Sequence analysis reveals that many centromereproximal telomere repeats in papaya harbor nucleotide substitutions and insertions of Gs and Ts. In contrast, we found very few N-to-C substitutions, and even fewer instances of nucleotide deletion, suggesting that a six-nucleotide telomere repeat is not well tolerated. The papaya genome encodes single-copy sequence homologues of several genes involved in telomere maintenance and chromosome end protection, including the Telomerase Reverse Transcriptase (TERT) and Protection Of Telomeres (POT1). Notably, unlike Arabidopsis, which encodes six Telomere Repeat binding Factor-like (TRFL) proteins that bind double-stranded telomere DNA, papaya appears to encode only two such proteins. Thus, the more streamlined genome of papaya will provide an excellent resource for comparative and functional analysis of telomeres in plants.

Keywords

TERT; Telomerase; Chromosome; POT1; TRF-like proteins

Introduction

Telomeres are nucleoprotein complexes at the ends of linear eukaryotic chromosomes that distinguish natural DNA ends from double-strand breaks and that prevent illegitimate DNA repair. In addition, telomeres solve the end replication problem by providing a means to compensate for the loss of terminal sequences due to the semi-conservative nature of DNA replication. The overall structure of telomeres is remarkably conserved across eukaryotic kingdoms (Fig. 1). Telomeric DNA is made up of short tandem repeats of TG-rich sequence: TTAGGG in vertebrates and TTTAGGG in most plants (and in some single-celled eukaryotes, such as *Plasmodium falciparum*). The length of telomeric DNA tract varies from organism to organism, ranging from 300 bp in budding yeast to over 100 kb in some plants [26]. The duplex region of the telomere terminates in a short single-stranded 3' protrusion of the G-rich strand called the G-overhang. This structure is essential for normal telomere function, and mutations that compromise the 3' overhang lead to telomere dysfunction [54].

The first telomere sequence from a higher eukaryotic organism, TTTAGGG, was cloned from the model plant *Arabidopsis thaliana* [39]. This sequence is present in most higher plants, with two notable exceptions. First, several families of the monocot Order of Asparagales harbor the six base human-type telomere repeat, TTAGGG [18]. These Asparagales species likely originated from a common ancestor that lived ~80 million years ago (mya) [55]. The second example of a non-canonical telomere repeat sequence arose ~ 15 mya in several genera of the Solanaceae family [50]. The nature and type of telomere sequence in these plants is currently unknown.

Telomere repeats are separated from the rest of the genomic DNA by a transitional sequence termed the subtelomere. The exact DNA organization of the subtelomeric region varies between organisms, but a number of features are conserved, including the presence of repetitive elements and transposons [40]. Subtelomeric regions are highly dynamic and undergo frequent inter-chromosomal exchange. As a result, large regions of subtelomeres may be shared by several different chromosome arms [22,34].

Telomeric DNA tracts are bound by two major classes of sequence-specific DNA binding proteins (Fig. 1). One group of proteins, termed Telomere Repeat binding Factors (TRF) in mammals, binds to the double-stranded region of the telomere [47]. Two TRF proteins are present in vertebrates. Upon homodimerization, both TRF1 and TRF2 bind telomeres through a Myb-type DNA binding domain [6]. TRF1 functions in telomere length control, while TRF2 participates in chromosome end protection [53]. The second group of proteins, typified by Protection Of Telomeres 1 (POT1), binds to the single-stranded G-overhang via an oligosaccharide/oligonucleotide fold (OB-fold) [5,52]. POT1, like TRF2, is essential for chromosome end protection. In mammals, both classes of telomere proteins are part of a six member telomere-specific complex termed shelterin [16].

The telomere is a substrate for the telomerase enzyme (Fig. 1). Telomerase is an unusual RNA-dependent DNA polymerase that extends the G-overhang, replenishing telomere DNA sequences lost as a consequence of the end-replication problem [15]. Telomerase is a large ribonucleoprotein complex. In mammals and apparently in *Arabidopsis*, the core telomerase enzyme is a dimer, consisting of the catalytic reverse transcriptase subunit (TERT), the RNA subunit (TER), which provides RNA template for telomere elongation, and dyskerin, an RNA binding protein necessary for telomerase RNP biogenesis *in vivo* (C. Cifuentes-Rojas and D. Shippen, unpublished; [14,28]. Additional telomerase-associated proteins may include the KU70/80 heterodimer [37,42], POT1a [49], and a variety of other proteins [15].

Relatively little is known about telomere biology in plants. Despite efforts to increase the number of plants used as models for plant telomere biology [25,56], *Arabidopsis* remains the

system of choice for studying mechanisms of telomere length homeostasis and chromosome end protection [57]. Despite its powerful genetic and molecular tools, *Arabidopsis* is not without limitations. One issue is that *Arabidopsis* belongs to a lineage that underwent several rounds of whole genome duplication [8]. As a consequence, several genes implicated in telomere biology are present in multiple copies, making their functional analysis problematic. For example, although only two *TRF* genes are present in vertebrates, *Arabidopsis* encodes at least six putative double-strand telomere binding proteins (TRF-like or TRFL), which appear to be functionally redundant [29]. Similarly, *Arabidopsis* harbors three distinct *POT1*-like genes [45,49] (A. Nelson et al., in preparation), while humans encode only a single POT1 protein [5].

Many questions remain regarding the function of these and other *Arabidopsis* telomere-related genes. Do all *Arabidopsis* TRFL and POT1 proteins have a function at telomeres? Are these functions similar to those described for their orthologues in yeast and vertebrates, or have novel roles for these proteins been invented in the plant kingdom? Thus, it is currently unclear whether *Arabidopsis* truly represents the plant kingdom with respect to its toolbox for synthesizing and maintaining chromosome ends, or whether evolution has taken different paths in various lineages of the green plants.

Papaya (Caricaceae family) and *Arabidopsis* (Brassicaceae family) belong to the Brassicales Order, and shared the last common ancestor approximately 70 mya [55]. As the most closely related organism to *Arabidopsis* with a completely sequenced genome, papaya provides a unique opportunity to investigate evolutionary aspects of telomere biology in the Brassicales Order. Strikingly, analysis of the papaya genome sequence indicates a lack of recent whole-genome duplications [34], making comparative genomic approaches in papaya and *Arabidopsis* a powerful tool for evaluating the minimal number of genes necessary for successful telomere function in flowering plants. Here we present an initial characterization of papaya telomere length and sequence. In addition, we report the identification of several telomere-related genes and evaluate in more detail the phylogenetic relationship between Brassicales *TERT* and *TRFL* genes.

Results

Telomere Length in Papaya and Other Members of the Brassicales Order

Telomere length varies tremendously across the plant kingdom, ranging from 0.3 kb in green algae [38] to 30–100 kb in barley and tobacco [17,30]. At 2.5 kb on average, telomeres in *Arabidopsis thaliana* are at the short end of this spectrum [39]. To investigate in more detail the evolution of telomere length set point in plants, we determined the size of telomere tracts in papaya using terminal restriction fragment (TRF) analysis. Initial analysis indicated that papaya telomeres are much longer than in *Arabidopsis*. To obtain a more accurate measurement of papaya telomere length, we resolved the digested DNA on SeaKem Gold Agarose, which allows efficient resolution of genomic fragments up to 50 kb and higher by conventional electrophoresis [10]. As shown in Fig. 2a, papaya telomeres appeared as a discrete pattern of bands. This profile is unusual, as telomeres in most plant species analyzed to date appear as homogeneous smears [17,41], with only a few exceptions [9,21]. Since papaya contains nine chromosome pairs and 36 telomeres, these bands may represent one or several chromosome arms with similar telomere lengths. Interestingly, the papaya TRF products are at least ten times longer than those in *Arabidopsis* and range in size from 25 kb to well over 50 kb.

To verify that the sequences detected by TRF analysis correspond to chromosome ends, DNA was preincubated with *Bal31* nuclease prior to digestion with *TruII*. *Bal31* is a non-specific exonuclease that preferentially degrades DNA ends versus more internal genomic

regions. After 30 and 60 min of *Bal31* digestion, the high molecular weight telomeric DNA migrated faster on the gel (Fig. 2a, lanes 2 and 3). With continued *Bal31* incubation, telomeric signals became smeary and the shortest telomere bands disappeared almost completely (Fig. 2a, lanes 4–6). In contrast, lower molecular weight bands, corresponding to interstitial cross-hybridizing DNA (indicated by asterisks) were insensitive to *Bal31* digestion for up to 150 min. These data confirm that the hybridizing TTTAGGG repeats represent terminal telomeric DNA.

Since telomere tracts in papaya are so much longer than those in *Arabidopsis*, we examined the range of telomere lengths in other species from the Brassicaceae family to determine which species is an outlier (Fig. 2b). Intriguingly, all the Brassicaceae species we analyzed possess short *Arabidopsis*-like telomeres (1.5–8 kb). Since *Capsella rubella* with the shortest telomeres (Fig. 2b, lane 5) and *Olimarabidopsis pumila* with the longest telomeres (Fig. 2b, lane 6) are phylogenetically distant from both *Arabidopsis thaliana* and *Brassica oleracea* [2], there is no obvious correlation between telomere size and phylogenetic relationship within Brassicaceae. Analysis of other Brassicales families will be needed to determine if the telomere length set point in this plant Order is indeed family-specific. Nevertheless, we conclude that telomere length can vary dramatically within closely related plant species.

Sequence Analysis of Subtelomeric and Telomeric DNA Regions in Papaya

Having established the total length of papaya telomeric tracts, we next examined the terminal sequence on chromosome ends in more detail. As part of the papaya genome sequencing project, 13 subtelomeric contigs were cloned and sequenced. Multiple Copia- and Gypsy-like retrotransposons and various DNA transposons were found in papaya subtelomeric regions [34]. In addition, inspection of subtelomeric regions indicated that nine of them share 0.5–1.5 kb of nearly identical DNA sequence immediately adjacent to terminal telomeric repeats [34]; Fig. 1b. A similar organization has been reported for several other eukaryotes, including yeast, *Plasmodium* and many plants [12,21,22]. Notably, the organization of subtelomeric DNA in papaya contrasts sharply with *Arabidopsis* subtelomeres, which consist of unique sequence on eight out of ten chromosome arms [23,32].

The sequence of the most terminal telomeric repeats on each chromosome arm is constantly modified through the combined action of nucleotide loss due to the end-replication problem and nucleotide gain through telomerase action. In contrast, the most centromere-proximal telomere repeats (the most internal region of the telomere repeat array that is immediately adjacent to the subtelomeric DNA tract) have likely not been acted upon by telomerase or the end-replication problem for millions of years, and thus these sequences provide an opportunity to analyze mutations that have accumulated in terminal chromosomal DNA over extended evolutionary time [36]. Mutations in centromere-proximal telomere repeats could potentially result from stochastic events, such as inter- or intra-chromosomal recombination or random nucleotide changes introduced by conventional DNA polymerase or DNA repair machineries. Alternatively, such mutations could represent misincorporations by telomerase, which occurred millions of years ago when telomerase synthesized telomeric DNA *de novo* to cap newly formed chromosome ends.

We examined the composition of papaya centromere-proximal telomere repeats for variations in telomere nucleotide content and repeat composition. We analyzed the nine available subtelomeric contigs that contain over 20 telomeric repeats (Table 1). For each contig, we calculated the total number of nucleotide deviations from the wild type sequence, TTTAGGG (Table 2). We found that 6–21% of the telomere nucleotides in each contig represent a substitution or an insertion (Table 2). A high portion of all sequenced telomere

repeats (38–76%, depending on the chromosome arm) contained at least one nucleotide substitution or insertion. One exception was the telomere tract on contig_22622, which displayed only a single nucleotide substitution out of 427 telomeric nucleotides (Table 2). The apparent lack of variant telomere repeats on this chromosome arm is intriguing and is consistent with a recent intra- or inter-chromosomal recombination with a more terminal, and less variant, telomere tract.

As an internal control, we analyzed two contigs composed exclusively of long stretches of telomeric repeats without adjacent subtelomeric sequence (Tables 1 and 2). These TELO-only contigs displayed significantly fewer mistakes. One TELO-only contig had 54 perfect telomere repeats (contig_48345), while in the other contig (contig_47027) mistakes were interspersed throughout the entire sequence (28/1020 total sequenced nucleotides, ~3%). The low number of mistakes in TELO-only contigs argues that these sequences may represent telomeric DNA derived from the extreme terminus of papaya chromosome ends.

Up to 5% of centromere-proximal telomere repeats in rice consist of the six nucleotide human repeat (TTAGGG) [36]. In contrast, among 535 centromere-proximal telomere repeats analyzed in papaya, only two instances of a single nucleotide deletion were found (Table 1, italicized repeats in contig_30940). On the other hand, there were many instances of nucleotide insertions, all of which reflected addition of either T or G (Tables 1 and 2). We classified these mistakes as T- and G-slippage [20]. There was no obvious preference towards either type of slippage overall: individual contigs differed in the ratio of the two types of mistakes (Table 2). The total frequency of T- and G-slippage was much higher in papaya than in rice telomere arrays, corresponding to 26% versus 1.2% of the total mistakes, respectively [36].

Frequency and Distribution of Individual Nucleotide Mistakes

Previous studies demonstrated that plant telomerases can extend non-telomeric DNA primers *in vitro* in a process mimicking *de novo* telomere synthesis *in vivo*. This process is error-prone and results in nucleotide misincorporation and T- and G-slippage [20]. We reasoned that if the mistakes found in the centromere-proximal portion of the papaya telomere array were introduced by telomerase at some point in papaya genome evolution, the type and frequency of nucleotide substitutions should be similar to mistakes made by plant telomerases *in vitro*. To test this idea, we analyzed individual nucleotide substitutions at each position in the papaya repeat (T1, T2, T3, A4, G5, G6 and G7). Consistent with *in vitro* generated telomerase products [20], a low number of substitutions (<2%) was observed at guanine positions 5–7 (Table 3). Similar results were also reported for rice [36]. However, in contrast to rice, where most substitutions occurred at the thymidine positions 1–3, the most commonly substituted nucleotide was adenine (42% of all substitutions in the repeat). In the majority of cases, A was changed to G (74%). Thymidines in positions 1–3 were mostly changed to G (100% for T1, 65% for T2) or to A (61% for T3). Interestingly, very few substitutions to C were observed for any nucleotide in the repeat, suggesting that this non-telomeric nucleotide is not well tolerated in papaya telomere repeats. The few substitutions to C occurred in the second (17.5%), the third (25%) and the fourth positions (21%).

Overall, repeats with a single nucleotide substitution accounted for the majority of variant repeats in papaya (>62%), followed by repeats with double point mutations (>33%). The most abundant variants were single-nucleotide substitutions tttGggg, Gtaggg and tttCggg (Table 4, repeats with an asterisk). The only exception was the variant ttAGggg, with T to A and A to G conversions. This sequence, which is present in 22 copies on six contigs, may be the product of telomerase template slippage, in which the enzyme skipped the first T and then added an extra G to produce the complete 7-nt repeat. In addition, similar to the

situation in rice [36], several contigs harbor arrays of variant repeats that are unique for a particular chromosome arm (Table 4, repeats with two asterisks). Overall, the type and frequency of nucleotide mistakes found in papaya centromere-proximal telomere repeats correlate well with mistakes generated by plant telomerases *in vitro* [20].

Identification and Sequence Analysis of Telomere-Related Genes

We queried the papaya genome for genes encoding several known regulators of telomere length homeostasis (Fig. 1). First, we identified the papaya telomerase catalytic subunit, TERT (accession EU906908). *Carica papaya* TERT harbors all the canonical functional motifs described for other TERTs (Fig. 3a). TERT has previously been characterized in *Arabidopsis* and rice, as well as several monocot species that synthesize human-type TTAGGG repeats [19,24,51]. Consistent with previous studies [51], papaya TERT protein sequence appears to diverge fast. Papaya TERT is 58% similar to *Arabidopsis* TERT and 46% similar to rice TERT overall. To analyze molecular evolution of TERT in more detail, we extended our analysis of TERT proteins to several other members of the green plant lineage with sequenced genomes, including poplar, grapevine (both dicots), *Selaginella* (a spikemoss) and two species of *Ostreococcus* (green algae) (accessions EU909207-EU909210). The availability of these novel sequences from different branches of plant evolution, as well as TERTs previously identified in two monocot Orders and in *Arabidopsis*, allowed us to reconstruct a phylogenetic tree of plant TERT proteins (Fig. 3b). In general, the molecular evolution of TERT proteins coincides with the established phylogenetic relationship of the respective plant species. Algal proteins formed a sister clade to all other plant sequences, and *Selaginella* TERT was placed as a sister node to flowering plants. Within Angiosperms, TERT proteins from dicot and monocot species formed two distinct clades.

Interestingly, in many of the functional TERT motifs, amino acid changes between papaya and *Arabidopsis* TERT proteins were observed (Fig. 3a, asterisks). In most instances, *Arabidopsis* TERT harbored an amino acid substitution that set it apart from papaya and from other plant TERTs. While the functional significance of these changes is unknown, their relative abundance suggests that TERT may experience a faster rate of molecular evolution in *Arabidopsis* than in other plants.

Several amino acid substitutions in TERT proteins from monocot species have been proposed to contribute to the synthesis of human-type repeats generated by these telomerases [51]. Specifically, the F858W substitution in motif C (numbering as in *Arabidopsis* TERT) was proposed as a potential mediator of the unusual enzymatic properties of these telomerases [51]. We found that the TERTs in the two species of green algae analyzed in this study, which synthesize the standard TTAGGG repeats, also contain W in this position (Fig. 3a). Similarly, other mutations proposed to influence TTAGGG synthesis, including A618K (motif 1) and F693A (motif A), are also found in *Selaginella* TERT, which synthesizes canonical TTAGGG telomere repeats (Fig. 3a). Thus, these residues are unlikely to be responsible for the altered telomerase specificity.

While telomerase subunits are present as single copy genes in most eukaryotic genomes, including plants, several telomere-binding proteins are encoded by multi-gene families in *Arabidopsis*. To determine if this is also true for papaya, we used the *Arabidopsis* *POT1a* and *POT1b* sequences to query the papaya genome for a *POT1* orthologue. We found evidence for only a single full-length *POT1* gene in papaya (accession EU887728). The predicted protein shares ~58% similarity to both AtPOT1a and AtPOT1b throughout its sequence, suggesting a remarkably fast divergence rate for POT1 proteins. Although a second locus with sequence similarity to the 5' end of *CpPOT1* gene was detected, this

segment appears to be a pseudogene, based on the presence of multiple in-frame stop codons and the lack of a sequence corresponding to the 3' end of the gene (data not shown).

We next searched for *TRFL* genes, which encode double-strand telomere DNA binding proteins [29]. We performed BLAST searches using the characteristic DNA binding domain of Arabidopsis TRFL proteins. In contrast to Arabidopsis, which harbors six *TRFL* genes, we detected only two papaya *TRFL* genes, designated *CpTRFL1* and *CpTRFL2* (accessions EU909205 and EU909206). Both papaya genes encode proteins with a highly conserved C-terminal Myb and Myb-extension domain, which in plants is essential for specific telomeric DNA binding *in vitro* [27,29,48]. In addition, both TRFL proteins in papaya contain a Telobox motif found in all double-strand telomere binding proteins [7] (Fig. 4a). Furthermore, papaya TRFL proteins also contain a central domain (CD), which is present in all plant TRFL proteins, but not in TRF proteins from mammals [29]. The function of this 80 amino acid region is currently unknown. Phylogenetic analysis indicated that *CpTRFL1* is most closely related to *AtTBP1* and *AtTRFL9*, while *CpTRFL2* is most similar to *AtTRFL2* (Fig. 4b). Notably, *AtTBP1* has been shown to negatively regulate telomere length in Arabidopsis [27]. Thus, *TRFL1* and *TRFL2* in papaya may be more closely related to the ancestral *TRFL* genes in Brassicales.

Discussion

Evolution of Telomere Length Homeostasis in Brassicales

Telomere length is tightly regulated and a species-specific length set point is achieved for each organism. Plants exhibit a wide variety of telomere lengths, but thus far there has been no concerted, systematic analysis of telomere length in phylogenetically related plant species. Previous studies indicated that telomeres in representatives of two dicot families, Brassicaceae (*Arabidopsis thaliana*) and Caryophyllaceae (*Silene latifolia*) possess short telomeres in the range of 2–5 kb [39,41], while members of the Solanaceae family (tomato, tobacco) harbor telomeres in the size range of 30–100 kb [17,21]. To further investigate the origin of this remarkable variation in telomere length, we examined telomere lengths in two closely related families of the Brassicales Order, Caricaceae (papaya) and Brassicaceae. Our data indicate that Brassicaceae species bear relatively short telomeres in the same range as Arabidopsis telomeres. In contrast, papaya telomeres are an order of magnitude longer, more like telomeres in members of the distantly related Solanaceae family. Thus, it now appears that not only species-specific, but also family-specific telomere length set points are established in plants. Analysis of additional members of the Caricaceae family as well as representatives from other closely related families of plants will be necessary to more precisely establish the timing of the telomere set point change within the Brassicales Order.

The TRF-like proteins are key players in modulating telomerase access. In yeast and human cells, the number of these proteins bound to the telomere is precisely measured, and this information is used to help determine whether a particular telomere tract should be extended by telomerase [26,33]. This “protein-counting” mechanism allows for preferential extension of the shortest telomeres in the population. Thus, telomerase recruitment may be more efficient or less regulated in plant species with very long telomeres. Arabidopsis telomerase has the intrinsic ability to synthesize exceptionally long telomere tracts *in vivo*; KU mutants generate telomeres twice the size of wild type in a single plant generation [42]. Thus, negative regulation of telomerase is a major mechanism controlling telomere length.

The absolute number of *TRF-like* genes or species-specific differences in their expression levels may influence telomere length set point. It is notable that papaya encodes only two TRFL proteins, as opposed to six in Arabidopsis [29]. If telomerase regulation is controlled through the combined action of all available TRFL proteins, this could account for greater

inhibition of telomerase in Arabidopsis than in papaya, resulting in very different telomere lengths. An argument against this model is that rice possesses short telomeres (5–10 kb) [25], and yet encodes only two TRFL proteins). Thus, it is unlikely that differences in the number of *TRFL* genes will be sufficient to account for dramatic variation in telomere length set point among related plant species. In support of this conclusion, different natural accessions of Arabidopsis, rice and maize display two-fold or higher differences in the length of their telomeres [11,35,44]. Other features of TRFL proteins, such as their relative affinity for telomeric DNA or differences in their interactions with other shelterin factors, are likely to contribute to species-specific telomere length set point.

Variant Telomere Repeats in the Centromere-Proximal Region of Papaya Telomere Arrays

Just as telomere length varies widely among different organisms, so does telomerase fidelity. In budding yeast, telomerase naturally produces variant repeats TG₁₋₃, not all of which provide proper binding sites for the yeast double-strand telomere-binding protein Rap1p [33]. In plants, telomerase fidelity has been analyzed in an *in vitro* chromosome healing assay [20]. This study examined a wide range of plant telomerases from monocot and dicot species. Most plant telomerases were found to be error-prone, with error rates between 1.0×10^{-2} and 1.8×10^{-3} . These enzymes generated a higher number of nucleotide substitutions, insertion or deletions in the first few synthesized repeats than in the repeats added more distal to the primer 3' terminus. Importantly, the number of G- and T-slippages was remarkably high, accounting for up to 52% of all mistakes [20].

To test if similar properties could be attributed to telomerase activity *in vivo*, we examined centromere-proximal telomere repeats immediately adjacent to the subtelomere in papaya. These repeats have likely been generated by telomerase millions of years ago when current subtelomeric chromosomal DNA became terminal, and thus these sequences may provide clues concerning *de novo* telomere formation by the papaya telomerase. Multiple nucleotide substitutions and insertions were found in centromere-proximal telomere repeats on most chromosome arms. In addition, we detected very few nucleotide deletions and a much higher number of nucleotide insertions. Overall, the nucleotide variations in papaya telomere repeats are remarkably similar to mistakes generated by plant telomerases during *de novo* telomere synthesis *in vitro* [20].

Several lines of evidence suggest that telomere sequence variations detected in subtelomere-adjacent repeats are not simply sequencing mistakes. First, if mistakes were introduced during cloning or sequencing steps, they would be expected to be distributed equally throughout telomeric DNA on all contigs. Instead, the number and type of telomere mistakes differs between contigs. Second, although telomeric repeats are notoriously difficult to clone and sequence, two contigs that consist exclusively of telomeric repeats were obtained (Table 2). While in one of them (contig_47027), only 3% of all nucleotides differ from the perfect canonical sequence, the sequence of the second contig_48345 (54 repeats in total), contained no mistakes at all. Similarly, the 12639-nt long contig_22622 contains just one mistake in the telomeric region.

The number and types of nucleotide changes in the centromere-proximal telomere repeats of papaya correlate well with results reported for humans. Human variant telomere repeats are abundant and occur mostly in centromere-proximal telomere repeats [1,4]. Moreover, mistakes in the 6 bp periodicity in human telomeres are rare, and are attributed to replication slippage or occasional intra-allelic recombination, rather than mistakes introduced by telomerase [3]. These findings argue that genetic processes underlying molecular evolution of telomere repeat regions are conserved between plants and animals.

Similar to the situation in rice [36], our data suggest that despite the increased number of nucleotide mistakes, most papaya centromere-proximal variant telomere repeats may still be functional with regard to DNA binding by TRFL proteins. Biochemical studies and crystal structures of plant TRFL proteins indicate that their preferred binding site is GGGTTT, although TRFL proteins will bind to variants of this sequence with less affinity [48,57]. This core sequence is preserved in the vast majority of variant papaya telomere repeats analyzed in our study.

Evolution of Telomere-Related Genes in Plants

Several amino acid changes in monocot TERT proteins have been proposed to influence telomerase activity and to promote the synthesis of human-type telomere repeats in these plants [51]. To test this hypothesis in a phylogenetically broader context, we obtained and examined novel TERT sequences from six additional plant species. No correlation between the amino acid changes in motifs C, 1 and A of monocot TERTs and the synthesis of human-type telomere repeats were observed, indicating that the switch in the sequence of telomere repeats synthesized by some monocot telomerases does not result from changes in TERT. Rather, we believe the cause is more likely to stem from a single nucleotide deletion in the template region of the RNA subunit of telomerase.

Our comparative analysis of plant TERT sequences also revealed that Arabidopsis TERT is evolving faster than other plant TERTs, as it harbors several amino acid substitutions in conserved functional motifs that are not found in other TERTs. This finding correlates well with recent evidence of positive selection in a telomerase-associated protein POT1a (E. Sharikov, et. al, in preparation), and suggests that evolutionary pressure may be acting on several components of the Arabidopsis telomerase holoenzyme.

An important finding from our analysis of the papaya genome is the striking amplification of genes encoding telomere-binding proteins in Arabidopsis versus papaya. As expected from two rounds of genome duplication in the specific lineage leading to Arabidopsis, members of the Brassicaceae family encode at least two highly divergent POT1 proteins [46], while *POT1* is a single-copy gene in papaya. Thus, we can now estimate the timing of *POT1* gene duplication in Brassicaceae to ~ 34 mya, when the last common ancestor of all Brassicaceae underwent a whole-genome duplication [43].

A more extreme example of increased gene copy number is apparent for the *TRFL* family. Current data indicate that TRFL proteins negatively regulate telomere length in plants [25,27,56]. However, T-DNA disruptions of five TRFL genes in Arabidopsis did not result in obvious defects in telomere biology [29]. Mutation in the sixth gene, *TBPI*, leads to a mild degree of telomere elongation, which only becomes apparent in the third generation of mutants [27]. Thus, it is likely that members of *TRFL* gene family in Arabidopsis are functionally redundant.

The identification of only two *TRFL* genes in papaya suggests that plants, like humans, may need only two TRFL proteins for successful telomere length regulation and chromosome end protection. In support of this model, only two *TRFL* genes can be detected in the genomes of rice and the moss *Physcomitrella patens*. Nevertheless, amplification of *TRFL* gene family members is a common theme in other representatives of the plant kingdom, including poplar and *Selaginella*, which encode three or more TRFL proteins. While the functional importance of multiple *TRFL* genes remains unclear, additional members of this protein family could provide an extra level of regulation for telomere length homeostasis and end protection.

In conclusion, the availability of the *Carica papaya* and *Arabidopsis thaliana* genomes provides a unique opportunity to evaluate evolution of plant telomeres in two highly related organisms. Specifically, the discovery of a small number of *TRFL* genes in papaya and the elucidation of their phylogenetic relationship to the members of the larger *TRFL* gene family in *Arabidopsis* will significantly focus future investigations of this gene family. Detailed analysis of the streamlined papaya genome [34] will provide a powerful tool for facilitating investigations of other gene families that underwent extensive amplification in the Brassicaceae family, but not in its close relative Caricaceae.

Methods

Plant Materials and Growth Conditions

Seeds for Brassicaceae family members *Arabidopsis thaliana*, *A. suecica*, *A. arenosa*, *A. lyrata*, *Capsella rubella*, *Olimarabidopsis pumila* and *Brassica oleracea* were obtained from ABRC (Ohio State University, Columbus), cat. ## CS 60000, CS3895, CS3901, CS22696, CS22697, CS22562 and CS29002, respectively. Papaya samples were obtained from commercial cultivars. Seeds were sown in soil, cold-treated overnight at 4°C, then placed in an environmental growth chamber and grown under a 16/8-hr light/dark photoperiod at 23°C.

Telomere Length Analysis and Bal31 Digestion

DNA from individual plants was extracted as described [13]. TRF analysis was performed with *TruII* (Fermentas) restriction enzyme and ³²P 5' end-labeled (T₃AG₃)₄ oligonucleotide as a probe [19]. Gel electrophoresis was performed using high gel strength, low EEO (≤0.05) SeaKem Gold Agarose (Lonza, Cat. # 50152), specifically developed for efficient resolution of genomic DNA fragments up to 50 kb and higher by conventional electrophoresis. Electrophoresis was performed in 0.3% agarose in 1X TAE buffer for 48 h at 25 V. Radioactive signals were scanned by a STORM PhosphorImager (Molecular Dynamics), and the data were analyzed by IMAGEQUANT software (Molecular Dynamics). For the *Bal31* exonuclease assay, 100 µg of papaya genomic DNA was incubated with 50 units of *Bal31* (New England Biolabs) or with H₂O (0 min time point) in 1X *Bal31* reaction buffer at 30°C. Equal amounts of sample were removed at 30 min intervals for 150 min. Reactions were stopped by the addition of 20 mM EGTA and heating to 65°C for 15 min. DNA in each sample was precipitated with isopropanol and ammonium acetate, followed by *TruII* digestion and blotting as described above.

BLAST Searches, Gene Predictions and Phylogenetic Analysis

BLAST searches of the papaya genome were performed using the tblastn option available at the papaya genome portal (<http://asgpb.mhpc.hawaii.edu/tools/tools.php>) with *Arabidopsis* telomere-related proteins as a query. Individual cDNA sequences were deduced using a combination of EST alignments and various gene prediction programs available at the papaya genome portal. TERT sequences from poplar (*Populus trichocarpa*), grapevine (*Vitis vinifera*), *Selaginella moellendorffii*, *Ostreococcus taurus* and *Ostreococcus lucimarinus* were obtained from the corresponding genome portals (http://genome.jgi-psf.org/euk_cur1.html) using similar approaches. Predicted plant cDNAs were deposited into the GenBank under the following accession numbers: EU887728 (CpPOT1), EU906908 (CpTERT), EU909205 (CpTRFL1), EU909206 (CpTRFL2), EU909207 (PtTERT), EU909208 (SmTERT), EU909209 (OtTERT), EU909210 (OtTERT). While this work was in preparation, the sequence of VvTERT was deposited into the GenBank under the accession CAO40853. Alignment and phylogenetic analysis of TERT and TRFL sequences was performed with MEGA3 software [31] using neighbor-joining method with 10,000 bootstrap replicates. The sequences of the papaya subtelomeric contigs

were deposited into GenBank (EU910963-EU910972). GenBank accession for subtelomeric contig_30940 is ABIM01030891.

Acknowledgments

We thank Tom McKnight and members of the Shippen laboratory for helpful discussions, and Ray Ming for coordinating collaborative work on the papaya genome. This work was supported by NIH GM 065383 and NSF MCB-0349993 to D.E. S., by NIH GM 083873 to S.L.S., and by the University of Hawaii and the US Department of Defense grant number W81XWH0520013 and the Maui High Performance Computing Center to M.A.

References

1. Allshire RC, Dempster M, Hastie ND. Human telomeres contain at least three types of G-rich repeat distributed non-randomly. *Nucleic Acids Res* 1989;17:4611–4627.10.1093/nar/17.12.4611 [PubMed: 2664709]
2. Bailey CD, Koch MA, Mayer M, Mummenhoff K, O Kane SL Jr, Warwick SI, Windham MD, Al-Shehbaz IA. Toward a global phylogeny of the Brassicaceae. *Mol Biol Evol* 2006;23:2142–2160.10.1093/molbev/msl087 [PubMed: 16916944]
3. Baird DM, Jeffreys AJ, Royle NJ. Mechanisms underlying telomere repeat turnover, revealed by hypervariable variant repeat distribution patterns in the human Xp/Yp telomere. *EMBO J* 1995;14:5433–5443. [PubMed: 7489732]
4. Baird DM, Coleman J, Rosser ZH, Royle NJ. High levels of sequence polymorphism and linkage disequilibrium at the telomere of 12q: implications for telomere biology and human evolution. *Am J Hum Genet* 2000;66:235–250.10.1086/302721 [PubMed: 10631154]
5. Baumann P, Cech TR. Pot1, the putative telomere endbinding protein in fission yeast and humans. *Science* 2001;292:1171–1175.10.1126/science.1060036 [PubMed: 11349150]
6. Bianchi A, Smith S, Chong L, Elias P, de Lange T. TRF1 is a dimer and bends telomeric DNA. *EMBO J* 1997;16:1785–1794.10.1093/emboj/16.7.1785 [PubMed: 9130722]
7. Bilaud T, Koering CE, Binet-Brasselet E, Ancelin K, Pollice A, Gasser SM, Gilson E. The telobox, a Myb-related telomeric DNA binding motif found in proteins from yeast, plants and human. *Nucleic Acids Res* 1996;24:1294–1303.10.1093/nar/24.7.1294 [PubMed: 8614633]
8. Bowers JE, Chapman BA, Rong J, Paterson AH. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* 2003;422:433–438.10.1038/nature01521 [PubMed: 12660784]
9. Broun P, Ganai MW, Tanksley SD. Telomeric arrays display high levels of heritable polymorphism among closely related plant varieties. *Proc Natl Acad Sci USA* 1992;89:1354–1357.10.1073/pnas.89.4.1354 [PubMed: 1741387]
10. Burland V, Curtis FP, Kusakawa N. Agarose gel analysis of 15–40-kb PCR amplimers. *Biotechniques* 1996;21:142–144. [PubMed: 8816249]
11. Burr B, Burr FA, Matz EC, Romero-Severson J. Pinning down loose ends: mapping telomeres and factors affecting their length. *Plant Cell* 1992;4:953–960. [PubMed: 1356536]
12. Chan CS, Tye BK. Organization of DNA sequences and replication origins at yeast telomeres. *Cell* 1983;33:563–573.10.1016/0092-8674(83)90437-3 [PubMed: 6345000]
13. Cocciolone SM, Cone KC. Pl-Bh, an anthocyanin regulatory gene of maize that leads to variegated pigmentation. *Genetics* 1993;135:575–588. [PubMed: 7694886]
14. Cohen SB, Graham ME, Lovrecz GO, Bache N, Robinson PJ, Reddel RR. Protein composition of catalytically active human telomerase from immortal cells. *Science* 2007;315:1850–1853.10.1126/science.1138596 [PubMed: 17395830]
15. Collins K. The biogenesis and regulation of telomerase holoenzymes. *Nat Rev Mol Cell Biol* 2006;7:484–494.10.1038/nrm1961 [PubMed: 16829980]
16. de Lange T. Shelterin: the protein complex that shapes and safeguards human telomeres. *Genes Dev* 2005;19:2100–2110.10.1101/gad.1346005 [PubMed: 16166375]

17. Fajkus J, Kovarik A, Kralovics R, Bezdek M. Organization of telomeric and subtelomeric chromatin in the higher plant *Nicotiana tabacum*. *Mol Gen Genet* 1995;247:633–638.10.1007/BF00290355 [PubMed: 7603443]
18. Fajkus J, Sykorova E, Leitch AR. Telomeres in evolution and evolution of telomeres. *Chromosome Res* 2005;13:469–479.10.1007/s10577-005-0997-2 [PubMed: 16132812]
19. Fitzgerald MS, Riha K, Gao F, Ren S, McKnight TD, Shippen DE. Disruption of the telomerase catalytic subunit gene from *Arabidopsis* inactivates telomerase and leads to a slow loss of telomeric DNA. *Proc Natl Acad Sci USA* 1999;96:14813–14818.10.1073/pnas.96.26.14813 [PubMed: 10611295]
20. Fitzgerald MS, Shakirov EV, Hood EE, McKnight TD, Shippen DE. Different modes of de novo telomere formation by plant telomerases. *Plant J* 2001;26:77–87.10.1046/j.1365-313x.2001.01010.x [PubMed: 11359612]
21. Ganai MW, Lapitan NL, Tanksley SD. Macrostructure of the tomato telomeres. *Plant Cell* 1991;3:87–94. [PubMed: 1668615]
22. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, Carlton JM, Pain A, Nelson KE, Bowman S, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 2002;419:498–511.10.1038/nature01097 [PubMed: 12368864]
23. Heacock M, Spangler E, Riha K, Puizina J, Shippen DE. Molecular analysis of telomere fusions in *Arabidopsis*: multiple pathways for chromosome end-joining. *EMBO J* 2004;23:2304–2313.10.1038/sj.emboj.7600236 [PubMed: 15141167]
24. Heller-Uszynska K, Schnippenkoetter W, Kilian A. Cloning and characterization of rice (*Oryza sativa* L) telomerase reverse transcriptase, which reveals complex splicing patterns. *Plant J* 2002;31:75–86.10.1046/j.1365-313X.2001.01337.x [PubMed: 12100484]
25. Hong JP, Byun MY, Koo DH, An K, Bang JW, Chung IK, An G, Kim WT. Suppression of RICE TELOMERE BINDING PROTEIN 1 results in severe and gradual developmental defects accompanied by genome instability in rice. *Plant Cell* 2007;19:1770–1781.10.1105/tpc.107.051953 [PubMed: 17586654]
26. Hug N, Lingner J. Telomere length homeostasis. *Chromosoma* 2006;115:413–425.10.1007/s00412-006-0067-3 [PubMed: 16741708]
27. Hwang MG, Cho MH. *Arabidopsis thaliana* telomeric DNA-binding protein 1 is required for telomere length homeostasis and its Myb-extension domain stabilizes plant telomeric DNA binding. *Nucleic Acids Res* 2007;35:1333–1342.10.1093/nar/gkm043 [PubMed: 17272298]
28. Kannan K, Nelson AD, Shippen DE. Dyskerin is a component of the *Arabidopsis* telomerase RNP required for telomere maintenance. *Mol Cell Biol* 2008;28:2332–2341.10.1128/MCB.01490-07 [PubMed: 18212040]
29. Karamysheva ZN, Surovtseva YV, Vespa L, Shakirov EV, Shippen DE. A C-terminal Myb extension domain defines a novel family of double-strand telomeric DNA-binding proteins in *Arabidopsis*. *J Biol Chem* 2004;279:47799–47807.10.1074/jbc. M407938200 [PubMed: 15364931]
30. Kilian A, Stiff C, Kleinhofs A. Barley telomeres shorten during differentiation but grow in callus culture. *Proc Natl Acad Sci USA* 1995;92:9555–9559.10.1073/pnas.92.21.9555 [PubMed: 11607583]
31. Kumar S, Tamura K, Nei M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform* 2004;5:150–163.10.1093/bib/5.2.150 [PubMed: 15260895]
32. Kuo HF, Olsen KM, Richards EJ. Natural variation in a subtelomeric region of *Arabidopsis*: implications for the genomic dynamics of a chromosome end. *Genetics* 2006;173:401–417.10.1534/genetics.105.055202 [PubMed: 16547105]
33. Marcand S, Gilson E, Shore D. A protein-counting mechanism for telomere length regulation in yeast. *Science* 1997;275:986–990.10.1126/science.275.5302.986 [PubMed: 9020083]
34. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, et al. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). *Nature* 2008;452:991–996.10.1038/nature06856 [PubMed: 18432245]

35. Mizuno H, Wu J, Kanamori H, Fujisawa M, Namiki N, Saji S, Katagiri S, Katayose Y, Sasaki T, Matsumoto T. Sequencing and characterization of telomere and subtelomere regions on rice chromosomes 1S, 2S, 2L, 6L, 7S, 7L and 8S. *Plant J* 2006;46:206–217.10.1111/j.1365-313X.2006.02684.x [PubMed: 16623884]
36. Mizuno H, Wu J, Katayose Y, Kanamori H, Sasaki T, Matsumoto T. Chromosome-specific distribution of nucleotide substitutions in telomeric repeats of rice (*Oryza sativa* L.). *Mol Biol Evol* 2008;25:62–68.10.1093/molbev/msm227 [PubMed: 17940210]
37. Peterson SE, Stellwagen AE, Diede SJ, Singer MS, Haimberger ZW, Johnson CO, Tzoneva M, Gottschling DE. The function of a stem-loop in telomerase RNA is linked to the DNA repair protein Ku. *Nat Genet* 2001;27:64–67. [PubMed: 11138000]
38. Petracek ME, Lefebvre PA, Silflow CD, Berman J. Chlamydomonas telomere sequences are A+T-rich but contain three consecutive G-C base pairs. *Proc Natl Acad Sci USA* 1990;87:8222–8226.10.1073/pnas.87.21.8222 [PubMed: 2236035]
39. Richards EJ, Ausubel FM. Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. *Cell* 1988;53:127–136.10.1016/0092-8674(88)90494-1 [PubMed: 3349525]
40. Riethman H, Ambrosini A, Paul S. Human subtelomere structure and variation. *Chromosome Res* 2005;13:505–515.10.1007/s10577-005-0998-1 [PubMed: 16132815]
41. Riha K, Fajkus J, Siroky J, Vyskot B. Developmental control of telomere lengths and telomerase activity in plants. *Plant Cell* 1998;10:1691–1698. [PubMed: 9761795]
42. Riha K, Watson JM, Parkey J, Shippen DE. Telomere length deregulation and enhanced sensitivity to genotoxic stress in *Arabidopsis* mutants deficient in Ku70. *EMBO J* 2002;21:2819–2826.10.1093/emboj/21.11.2819 [PubMed: 12032094]
43. Schranz ME, Mitchell-Olds T. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. *Plant Cell* 2006;18:1152–1165.10.1105/tpc.106.041111 [PubMed: 16617098]
44. Shakirov EV, Shippen DE. Length regulation and dynamics of individual telomere tracts in wild-type *Arabidopsis*. *Plant Cell* 2004;16:1959–1967.10.1105/tpc.104.023093 [PubMed: 15258263]
45. Shakirov EV, Surovtseva YV, Osbun N, Shippen DE. The *Arabidopsis* Pot1 and Pot2 proteins function in telomere length homeostasis and chromosome end protection. *Mol Cell Biol* 2005;25:7725–7733.10.1128/MCB.25.17.7725-7733.2005 [PubMed: 16107718]
46. Smogorzewska A, de Lange T. Regulation of telomerase by telomeric proteins. *Annu Rev Biochem* 2004;73:177–208.10.1146/annurev.biochem.73.071403.160049 [PubMed: 15189140]
47. Sue SC, Hsiao HH, Chung BC, Cheng YH, Hsueh KL, Chen CM, Ho CH, Huang TH. Solution structure of the *Arabidopsis thaliana* telomeric repeat-binding protein DNA binding domain: a new fold with an additional C-terminal helix. *J Mol Biol* 2006;356:72–85.10.1016/j.jmb.2005.11.009 [PubMed: 16337232]
48. Surovtseva YV, Shakirov EV, Vespa L, Osbun N, Song X, Shippen DE. *Arabidopsis* POT1 associates with the telomerase RNP and is required for telomere maintenance. *EMBO J* 2007;26:3653–3661.10.1038/sj.emboj.7601792 [PubMed: 17627276]
49. Sykorova E, Lim KY, Chase MW, Knapp S, Leitch IJ, Leitch AR, Fajkus J. The absence of *Arabidopsis*-type telomeres in *Cestrum* and closely related genera *Vestia* and *Sessea* (Solanaceae): first evidence from eudicots. *Plant J* 2003;34:283–291.10.1046/j.1365-313X.2003.01731.x [PubMed: 12713535]
50. Sykorova E, Leitch AR, Fajkus J. Asparagales telomerases which synthesize the human type of telomeres. *Plant Mol Biol* 2006;60:633–646.10.1007/s11103-005-5091-9 [PubMed: 16649103]
51. Theobald DL, Mitton-Fry RM, Wuttke DS. Nucleic acid recognition by OB-fold proteins. *Annu Rev Biophys Biomol Struct* 2003;32:115–133.10.1146/annurev.biophys.32.110601.142506 [PubMed: 12598368]
52. van Steensel B, Smogorzewska A, de Lange T. TRF2 protects human telomeres from end-to-end fusions. *Cell* 1998;92:401–413.10.1016/S0092-8674(00)80932-0 [PubMed: 9476899]
53. Wei C, Price M. Protecting the terminus: t-loops and telomere end-binding proteins. *Cell Mol Life Sci* 2003;60:2283–2294.10.1007/s00018-003-3244-z [PubMed: 14625675]
54. Wikstrom N, Savolainen V, Chase MW. Evolution of the angiosperms: calibrating the family tree. *Proc Biol Sci* 2001;268:2211–2220.10.1098/rspb.2001.1782 [PubMed: 11674868]

55. Yang SW, Kim SK, Kim WT. Perturbation of NgTRF1 expression induces apoptosis-like cell death in tobacco BY-2 cells and implicates NgTRF1 in the control of telomere length and stability. *Plant Cell* 2004;16:3370–3385.10.1105/tpc.104.026278 [PubMed: 15528297]
56. Yu EY, Kim SE, Kim JH, Ko JH, Cho MH, Chung IK. Sequence-specific DNA recognition by the Myb-like domain of plant telomeric protein RTBP1. *J Biol Chem* 2000;275:24208–24214.10.1074/jbc.M003250200 [PubMed: 10811811]
57. Zellinger B, Riha K. Composition of plant telomeres. *Biochim Biophys Acta* 2007;1769:399–409. [PubMed: 17383025]

Fig. 1.

A proposed model of papaya telomere structure. **a** Telomeric DNA repeats are shown in *blue* and subtelomeric DNA in *red*. Based on studies from Arabidopsis and other organisms, as well as analysis of the papaya genome (this study), it is likely that papaya telomeres are bound by both double-strand (*TRFL*) and single-strand (*POT1*) telomere proteins. The core telomerase enzyme is composed of dyskerin, TERT and TER subunits and associates with the G-overhang in S-phase. The KU70/80 heterodimer as well as yet unidentified shelterin components associate with telomeres through protein-protein interactions. **b** Diagram of papaya telomere organization. The terminal telomere array is over 50 kb in length (*blue*), including variant telomere repeats (*pale blue*) at the junction between telomere repeats and the subtelomeric sequence (*red*). 0.5–1.5 kb of subtelomeric DNA is followed by chromosome-specific regions (*orange*) with multiple insertions of Copia- and Gypsy-type retrotransposons and DNA transposons

Fig. 2.

Telomere length analysis in members of the Brassicales Order. **a** TRF analysis of papaya telomeres. Lane 1, *TruII* digestion of genomic DNA without prior *Bal31* treatment (0 min), lanes 2–6, *TruII* digestion of genomic DNA after various incubation periods with *Bal31* exonuclease. Molecular weight markers are shown on the left. Asterisks indicate cross-hybridizing interstitial telomeric DNA bands, which are not sensitive to *Bal31* digestion for up to 150 min. **b** TRF analysis of telomeric DNA in members of the Brassicaceae family. Results are shown for genomic DNA from *Arabidopsis thaliana* (lane 1), *Arabidopsis suecica* (lane 2), *Arabidopsis arenosa* (lane 3), *Arabidopsis lyrata* (lane 4), *Capsella rubella* (lane 5), *Olimarabidopsis pumila* (lane 6) and *Brassica oleracea* (lane 7)

Fig. 3.

Phylogenetic analysis and motif conservation of plant TERT proteins. **a** Alignment of conserved TERT motifs from plants and animals. *Top panel*: Diagram of TERT proteins. TERTs are large polypeptides that share nine functionally important motifs. Reverse transcriptase motifs are shown in *light grey*, and N-terminal telomerase specific motifs are shown in *dark grey*. *Bottom panel*: Alignment of conserved TERT motifs from plants and animals indicates a high degree of inter-kingdom conservation. *Asterisks* denote amino acid positions that differ in AtTERT relative to other TERTs. *Arrows* denote amino acids predicted to play a role in the synthesis of human-type telomeric repeat in some Asparagales species [51]. Abbreviations: Cp, *Carica papaya*; At, *Arabidopsis thaliana*; Pt, *Populus trichocarpa* (poplar); Vv, *Vitis vinifera* (grapevine); Sc, *Scilla peruviana*; It, *Iris tectorum*; De, *Doryanthes excelsa*; Zm, *Zea mays* (maize); Os, *Oryza sativa* (rice); Sm, *Selaginella moellendorffii*; Ol, *Ostreococcus lucimarinus*; Ot, *Ostreococcus tauri*; Cf, *Canis familiaris* (dog); Mm, *Mus musculus* (mouse), Xl, *Xenopus laevis*. **(B)** Phylogenetic tree of plant TERTs. Vertebrate sequences from humans (hTERT), dog (CfTERT), mouse (MmTERT) and frog (XITERT) were used as outgroups. Brackets on the right indicate species with common taxonomic positions. The tree was constructed by the neighbor-joining method after sequence alignment of TERT proteins from plants and animals. Numbers at each node indicate the percentage of 10,000 bootstrap replicates for statistical support

Fig. 4. Sequence alignment and phylogenetic tree of TRFL proteins from papaya and Arabidopsis. **a** A diagram showing relative positions of CD and MYB with Myb-extension domains (*top panel*). Sequence alignment of the conserved MYB with Myb-extension domains of papaya and Arabidopsis TRFL proteins is shown (*bottom panel*). A characteristic feature of MYB domain containing telomere-binding proteins is the highly conserved Telobox motif. **b** Phylogenetic tree of papaya and Arabidopsis TRFL proteins. Alignment and tree reconstruction performed as described in Fig. 3b

Papaya telomere sequences

Table 1

Contig #	Telomere sequence
contig_30940	(tttaggg)G(ttG ^{agg})(ttG ^{ggg})(ttG ^{ggg})(GttC ^{ggg})G(GttC ^{ggg}) <i>ttagg</i> (tttaggg) ₅ G(tttaggg) ₂ G(tttaggg)G (ttC ^{ggg}) ₂ (ttG ^{ggg}) ₂ (ttAG ^{ggg})(ttC ^{ggg})(ttC ^{ggg}) ₃ (tttaggg) ₁₂ (ttG ^{ggg}) ₂ (ttAG ^{ggg}) <i>T</i> (tttaggg) ₂ (ttAG ^{ggg}) ₂ (<i>ttagg</i>)(ttAG ^{ggg})(ttAG ^{ggg})(ttAG ^{ggg})G7(tttaggg)G7(tttaggg) ₂ (Gtttaggg) ₄ G <i>T</i> (tttaggg) ₂ (Gtttaggg)G(tttaggg)(Gtttaggg)G7(tttaggg)(ttAG ^{ggg}) <i>T</i> (tttaggg) ₂ (tttaggg) (ttAG ^{ggg}) <i>T</i> (tttaggg) ₁₂
contig_45657	(tttaggg) ₃ (ttG ^{ggg})(tttaggg)(ttG ^{ggg}) ₃ (tttaggg) ₃ (Gtttaggg) ₈ (Gtttaggg) ₂ (ttAG ^{ggg})(Gtttaggg)(ttG ^{ggg}) (tttaggg) ₂ (Gtttaggg)(tttaggg)(Gtttaggg) ₅ (tttaggg) ₆
contig_33510	(tttaggg) ₇ G(tttaggg) ₁₃ (GttAG ^{ggg})(tttaggg) ₂ (GttAG ^{ggg})(tttaggg) ₁₉ (tttaggT)(tttaggg)(tttaggT)(tttaggg) ₂ (tttaggT) (tttaggg) ₇ (tttaggg) ₇ (tttaggg) ₇ (TTAGtaggg)(tttaggg) ₂ (tttaggg) <i>T</i> (tttaggTT)(tttaggg)G(<i>T</i> tttaggg) ₂ (tttaggg)G(<i>T</i> tttaggg) ₃ G (Ttttaggg) <i>T</i> (ttT ^{ggg})(Ttttaggg) ₅ (tttaggg) ₂ (ttT ^{ggg})(tttaggg)
contig_28422	(ttAGGggg)(ttAG ^{ggg})(ttAG ^{ggg}) ₆ (ttAG ^{ggg})G(ttCG ^{ggg}) ₅ (ttCG ^{ggg}) ₄ (GttG ^{gTg})G(ttCG ^{ggg}) ₃ (ttCG ^{ggg})(ttAG ^{ggg})G(ttAG ^{ggg}) (ttCG ^{ggg})(tttaggg) ₅ (tttaggg) ₄ (tttaggg) ₂ (ttGG ^{agg})(tttaggg)(tttaggg)(GGtaggg) ₂ (ttAGG ^{ggg}) (tttaggg) ₂ (ttAGG ^{ggg})(tttaggg)(GGtaggg) ₄ G(tttaggg) <i>T</i> (tttaggg)G7(tttaggg) ₃ (GGtaggg) ₄ G(tttaggg) <i>T</i> (tttagggGGG) ₂ (tttaggg) ₃ <i>T</i> (tttagggGGG)(tttaggg) ₇
contig_38705	(tttaggg) ₉ (Gtttaggg)(tttaggg) ₂ G(ttAG ^{ggg})(ttG ^{ggg})(tttaggg)(Gtttaggg)(tttaggg)(tttaggg)
contig_27826	(GttT ^{ggg}) ₂ (ttT ^{ggg})(GttC ^{ggg})(GttAG ^{ggg}) ₂ GttAG ^{ggg} (Gtttaggg)(tttaggg) ₁₅ Gtttaggg)(tttaggg)(Gtttaggg) (tttaggg)(Gtttaggg)(tttaggg)(Gtttaggg)(tttaggg)
contig_26718	(Atttaggg)(tttG ^{ggg})(ttAG ^{ggg})G(ttGG ^{ggg})(ttAG ^{ggg})G(ttG ^{ggg}) ₃ (ttC ^{ggg}) ₂ (ttG ^{ggg})(ttC ^{ggg})(ttC ^{ggg}) (tttG ^{ggg})(tttaggg) ₁₇ (ttAaggg)(tttaggg) ₁₀
contig_32257	(Atttaggg)(tttG ^{ggg})(ttAG ^{ggg})G(ttGG ^{ggg})(ttAG ^{ggg})G(ttG ^{ggg}) ₃ (ttC ^{ggg}) ₂ (ttG ^{ggg})(ttC ^{ggg})(ttC ^{ggg}) (tttG ^{ggg})(tttaggg) ₁₇ (ttAaggg)(tttaggg) ₁₀ G(ttAaggg)(tttaggg) ₃ G(tttaggg) ₂ (tttaggg) ₂ (ttAG ^{ggg})G (ttAaggg)(tttaggg) ₂ G(tttaggg) ₁₂
contig_22622	(tttaggg) ₉ (tttaA ^{ggg})(tttaggg) ₅₁
contig_47027 (TELO only)	(tttaggg) ₂₈ (Gtttaggg) ₃ (tttaggg)(Gtttaggg) ₂ (tttaggg) ₇ (Gtttaggg) ₃ (tttaggg)(Gtttaggg)(tttaggg)G(tttaggg) ₂ <i>T</i> (ttC ^{ggg})(tttaggg)G(ttT ^{ggg})(Ttttaggg) ₂ (ttT ^{ggg}) <i>T</i> (tttaggg) ₂ (tttaggg) ₇ (Ctaggg) <i>T</i> (tttaggg) ₂ (tttaggg) ₂ <i>T</i> (tttaggg) ₂ <i>T</i> (tttaggg) ₈ (tttaggC)(tttaggg) ₅₉
contig_48345 (TELO only)	(tttaggg) ₅₄

Contig sequences were trimmed of non-telomeric DNA. Telomeric DNA starting with the first full TTTAGGG repeat is shown in the 5'-3' orientation in lower case. Mistakes in telomeric sequence are shown in upper case, and mistakes classified as T- or G-slippage are italicized. The two repeats with a single nucleotide deletion in contig_30940 are italicized. Contigs 47027 and 48345 consist only of telomeric repeats

Table 2

Distribution and type of errors in the centromere-proximal telomeric repeats

Configs	Mutant repeats	Mutant nucleotides	G-slippage (%)	T-slippage (%)	Misincorporations (%)
config_30940	59/97 (~60%)	86/696 (~13%)	10	12	78
config_22622	1/61 (~1.6%)	1/427 (~0.2%)	0	0	100
config_27826	13/32 (~40%)	21/226 (~9%)	10	0	90
config_28422	62/82 (~76%)	126/603 (~21%)	18	5	77
config_26718	17/44 (~39%)	22/310 (~7%)	9	0	91
config_32257	27/71 (~38%)	32/504 (~6%)	22	0	78
config_45657	23/45 (~51%)	26/315 (~8%)	0	0	100
config_33510	34/82 (~41%)	45/605 (~7.4%)	11	53	36
config_38705	8/21 (~38%)	9/150 (~6%)	33	0	67
config_47027 ^a (TELO only)	27/144 (~19%)	28/1,020 (~3%)	7	36	57
config_48345 ^a (TELO only)	0/54	0/378	0	0	0

^aConfigs containing only telomere repeats

Table 3

Nucleotide substitutions in papaya telomere repeats

Nucleotide in TTTAGGG repeat	T1	T2	T3	A4	G5	G6	G7
Frequency of substitutions	22.34%	12.1%	20.5%	42.12%	0.37%	0.73%	1.83%
Substituted to	G (100%)	G (65%)	G (14%)	G (74%)	nd	nd	nd
	A (0%)	A (17.5%)	A (61%)	T (5%)			
	C (0%)	C (17.5%)	C (25%)	C (21%)			

Table 4

Frequency and distribution of variant repeats at papaya telomeres

Mutant repeat	Contigs with mutant repeat	Total number of mutant repeats	contig-specific?	Mutant repeat	Contigs with mutant repeat	Total number of mutant repeats	contig-specific?
tAAG _{ggg} ^a	6	22	no	tAGC _{ggg} ^b	1	3	yes
tttG _{ggg} ^a	5	35	no	tttaggT ^b	1	3	yes
Gtttaggg ^a	5	35	no	GttT _{ggg} ^b	1	2	yes
tttC _{ggg} ^a	4	17	no	tAtC _{ggg}	1	2	no
tttT _{ggg}	3	4	no	ttG _{ggg}	1	1	no
ttA _{ggg} ^b	2	7	yes	ttAG _{gg} T	1	1	no
tCtaggg ^b	2	6	yes	tttaAgg	1	1	no
GtAG _{ggg}	2	3	no	GtAG _{gg} T	1	1	no
GttG _{ggg}	2	2	no	GttG _g Tg	1	1	no
ttGG _{ggg}	2	2	no	tAtG _{ggg}	1	1	no
ttCG _{ggg} ^b	1	14	yes	ttGG _{ggg}	1	1	no
GGtaggg ^b	1	11	yes	GtG _{agg}	1	1	no
tGtaggg ^b	1	10	yes	tttagTT	1	1	no
GttC _{ggg} ^b	1	5	yes	tttaggC	1	1	no

^aMost common variant repeats present on multiple contigs

^bUnique variant repeats present only on a particular chromosome arm