Leveraging whole genome sequencing to estimate telomere length in plants

Michelle Zavala Paez[1], Jason Holliday[2], and Jill Hamilton[1]

[1]The Pennsylvania State University
[2]Virginia Tech

June 2, 2023

Abstract

Changes in telomere length are increasingly used to indicate species' response to environmental stress across diverse taxa. Despite this broad use, few studies have explored telomere length in plants. However, rapid advances in sequencing approaches and bioinformatic tools now allow estimation of telomere length using whole genome sequencing (WGS) data. Thus, evaluation of new approaches for measuring telomere length in plants are needed. Traditionally, telomere length has been quantified using quantitative polymerase chain reaction (qPCR). While WGS has been extensively used in humans, no study to date has compared the effectiveness of WGS in estimating telomere length in plants relative to traditional qPCR approaches. In this study, we use one hundred Populus clones re-sequenced using short-read Illumina sequencing to quantify telomere length using three different bioinformatic approaches, Computel, K-seek, and TRIP, in addition to qPCR. Overall, telomere length estimates varied across different bioinformatic approaches, but were highly correlated across methods for individual genotypes. A positive correlation was observed between WGS estimates and qPCR, however, Computel estimates exhibited the greatest correlation. Computel incorporates genome coverage into telomere length calculations, suggesting that genome coverage is likely important to telomere length quantification when using WGS data. Overall, telomere estimates from WGS provided greater precision and accuracy of telomere length estimates relative to qPCR. The findings suggest WGS is a promising approach for assessing telomere length, and as the field of telomere ecology evolves may provide added value to assaying response to biotic and abiotic environments for plants needed to accelerate plant breeding and conservation management.

Leveraging whole genome sequencing to estimate telomere length in plants

Michelle Zavala-Paez[1*]

Jason Holliday[2]

Jill A Hamilton[1]

[1] Department of Ecosystem Science and Management, Pennsylvania State University, State College, PA, USA

[2] Department of Forest Resources and Environmental Conservation, Virginia Tech, Blacksburg, VA, USA

*Correspondence:*michellezavalapaez@gmail.com*

Abstract

Changes in telomere length are increasingly used to indicate species' response to environmental stress across diverse taxa. Despite this broad use, few studies have explored telomere length in plants. However, rapid advances in sequencing approaches and bioinformatic tools now allow estimation of telomere length using whole genome sequencing (WGS) data. Thus, evaluation of new approaches for measuring telomere length in plants are needed. Traditionally, telomere length has been quantified using quantitative polymerase chain reaction (qPCR). While WGS has been extensively used in humans, no study to date has compared the effectiveness of WGS in estimating telomere length in plants relative to traditional qPCR approaches. In this study, we use one hundred*Populus* clones re-sequenced using short-read Illumina sequencing to quantify telomere length using three different bioinformatic approaches, Computel, K-seek, and TRIP, in addition to qPCR. Overall, telomere length estimates varied across different bioinformatic approaches, but were highly correlated across methods for individual genotypes. A positive correlation was observed between WGS estimates and qPCR, however, Computel estimates exhibited the greatest correlation. Computel incorporates genome coverage into telomere length calculations, suggesting that genome coverage is likely important to telomere length quantification when using WGS data. Overall, telomere estimates from WGS provided greater precision and accuracy of telomere length estimates relative to qPCR. The findings suggest WGS is a promising approach for assessing telomere length, and as the field of telomere ecology evolves may provide added value to assaying response to biotic and abiotic environments for plants needed to accelerate plant breeding and conservation management.

Keywords: telomere, plants, whole genome sequencing, qPCR, biomarker.

Introduction

Telomeres, repetitive, non-coding sequences of DNA found at the tips of chromosomes, function to maintain chromosome and genome integrity during replication . Previous research has linked telomere length loss to aging and disease in various taxa . However, new evidence suggests changes in telomere length are influenced by the environment . Indeed, telomere length loss has been associated with exposure to extreme heat in dairy cattle , increased river temperatures in fish , and developmental stress during embryogenesis in birds . While less is known about these relationships in plants, telomere length varies in response to aging and the environment in *Arabidopsis* , with impacts to fitness . Thus, telomeres may be a valuable biomarker for assessing response to environmental stress in plants. These findings highlight the need to develop new methods to quantifying telomere length for population management in natural and agronomic plant systems.

Telomere length has traditionally been estimated using the terminal restriction fragment (TRF) method . TRF measures the absolute telomere length of an individual by using the length distribution of terminal restriction fragments, which are detected by southern blotting using a probe specific for telomeric DNA . The TRF method uses genomic DNA, digested with restriction endonucleases that cut throughout the chromosome, but not within the telomeric region, assuming restriction sites are absent in the telomere region . Telomere length is then determined by quantifying the signal intensity of the entire TRF smears relative to a DNA ladder with known fragment size . Despite its utility, TRF requires large quantities of DNA (3 µg/individual), and quantification can be sensitive to DNA degradation . An alternate method, quantitative polymerase chain reaction (qPCR) has also been adopted to estimate relative telomere length for diverse taxa . The qPCR approach estimates relative telomere length using the ratio of the threshold cycle (Ct) of the telomeric region (T) to a non-variable copy number reference gene (S) relative to a reference sample . qPCR is widely used in telomere studies as it requires small quantities of DNA (20 ng/individual), provides scalability for high-throughput assessments, and produces results in a shorter amount of time than TRF . However, despite the wide utility of these approaches, recent advances in whole-genome sequencing (WGS) and new bioinformatic tools are enabling the estimation of telomere length using whole-genome sequencing. Rapid advances in high-throughput sequencing coupled with reduced costs have increased the availability

of WGS data across model and non-model systems . Thus, given the availability of new bioinformatic resources, comparison of approaches to estimating telomere length is needed, particularly in plants where there is limited evaluation of the accuracy and reliability of different approaches.

Telomere length in humans was initially estimated from WGS data by counting the number of reads containing the human telomere repeat sequence $(TTAGGG)_4$ normalized by genome coverage . To date, several bioinformatic programs, including TelSeq , Telomerecat , Telogator , TelomereHunter , and Computel have been developed to estimate telomere length in humans using WGS data. However, despite the increasing availability of WGS data for plant species, few tools assess telomere length in plants. Only three bioinformatic programs, Computel, K-seek, and Telomeric repeats identification pipeline have been developed that are able to quantify telomere length for diverse organisms more broadly and are appropriate for plants. These bioinformatics tools offer new opportunities for quantifying telomere length variation using WGS across plant systems, essential for advancing the field of telomere ecology.

Despite the biological and ecological value of quantifying telomere length variation in plants, few studies to date have either estimated telomere length or compared telomere length estimates using qPCR and WGS. In this study, we aim to fill this gap, comparing telomere length estimates using multiple bioinformatic approaches for one hundred*Populus* clones re-sequenced using Illumina short-read sequencing, comparing Computel , K-seek , and TRIP . In addition, we compare telomere length estimates for the same individuals assayed using qPCR and WGS data. This study will test new approaches to estimating telomere length and will extend the use of telomeres as a potential biomarker to assay response to environmental change needed for plant species management and breeding.

Materials and Methods

Plant Material & Greenhouse

In January 2020, dormant vegetative cuttings were collected from 100 poplar trees spanning much of the latitudinal distribution of the contact zone between *Populus trichocarpa* and *P .balsamifera* (Figure 1, Table S1). Cuttings were then transported to Blacksburg, VA where they were rooted in a greenhouse maintained at day and night temperatures of 24 and 15.5 °C, respectively, with no supplemental lighting. Following vegetative flush of rooted cuttings, leaf tissue was sampled for DNA extraction using the Qiagen plant DNeasy kit (Qiagen Inc., Valencia, CA), with minor modifications that included a phenol-chloroform extraction in place of a QIAshredder column, approximately 100 mg of leaf was used to extract DNA. Samples with low DNA concentration were subject to a secondary extraction using a modified CTAB protocol . Quality and concentration of DNA were assessed using a combination of NanoDrop and Qubit approaches, in addition to gel electrophoresis. DNA from these samples was used to quantify telomere length using both qPCR and WGS based on Illumina short-read sequencing.

Genomic libraries and sequencing

Genomic libraries were prepared using the Illumina KAPA HyperPrep kit at the Duke University Center for Genomic and Computational Biology. Briefly, the libraries were sequenced on an S4 flow cell in 2 x 150 bp format on an IlluminaNovaSeq 6000 with 64 samples per lane. Quality control and preprocessing of raw reads were done using FastQC and Trimmomatic . Sequences were trimmed to remove adapters, short reads (< 30 bp), and reads were discarded with < 30 bp of overlap between forward and reverse reads with a maximum mismatch value of 4. Following trimming, 37.33 ± 6.5 million reads were retained with an average GC length of $35.6\% \pm 0.5$ and an average read length of 141 bp (Table S1). Trimmed reads were used in the downstream analyses.

Telomere length from whole-genome sequence data

Telomere length was estimated from WGS data using three different bioinformatic approaches; Computel v1.3 , K-seek , and TRIP . All three programs estimate telomere length using reads from whole genome sequence data. However, the three programs use different parameters to estimate telomere length allowing for comparison.

Computel (v1.3) creates a telomeric reference using read length, in addition to telomeric repeat length and pattern . Telomeric reads are those reads that align to the telomere reference developed within the program. At least two replicates of the telomere sequence pattern were required to be considered a telomeric read. Once telomeric reads are identified, Computel calculates the relative genome coverage for individuals as a ratio of telomeric coverage to genome coverage. Telomeric coverage is defined as the distribution of coverage per base for the telomeric reference and genome coverage is the product of the total number of reads and read length, divided by total genome size. After estimating relative genome coverage, Computel estimates telomere length by counting telomeric reads and normalizing counts relative genome coverage. Computel then divides telomere length by the number of chromosomes present in the species to obtain mean telomere length per individual. For the purposes of our study, we modified Computel parameters to consider *Populus* genome (*P. trichocarpa* v3, NC_037285.1.) parameters, including chromosome number (n=19), genome size (434,290,000 bp), and telomere sequence pattern (TTTAGGG).

In addition to Computel, we used the bioinformatics pipeline K-seek to identify and quantify reads that contained telomeric repeats. K-seek uses a pattern-matching approach that does not consider genome coverage in telomere estimates. K-seek identifies sequence repeats ranging between 1 to 20 bp . Reads with a minimum repeat length of 50 bp per repeat are considered to avoid counting short repeat sequences scattered across the genome. After identifying and counting sequence repeats, K-seek organizes the repeat sequences in alphabetical order and size (in bp). To quantify telomeric repeat sequences, we used the canonical telomere repeat sequence specific to plants, i.e., AAACCCT , which is the reverse complement of TTTAGGG (Figure S2). To calculate mean telomere length per individual in base pairs, we applied the following equation:

Equation (1) *Mean telomere length per individual*= $\frac{\text{FTL x RL}}{2\ x\ Chr}$

Where, *FTL* is the frequency of telomeric repeats, *RL* is the telomeric repeat length (7 base pairs for TTTAGGG), two in the denominator accounts for the two ends of a chromosome and *Chr* is the number of chromosomes in *Populus* (n=19).

Finally, we used the telomeric repeats identification pipeline (TRIP) to estimate telomere length across *Populus* genotypes. Like K-seek, TRIP uses a pattern-matching approach that does not include genome coverage into telomere estimates. However, unlike K-Seek only reads with a minimum of four repeat sequences are considered . TRIP searches for repeat sequences of lengths ranging from 2 to 25 bp with no mismatches permitted. Once repeat sequences are identified, TRIP reorders the repeat sequences alphabetically and summarizes the repeat regions into count tables. The frequency of the canonical telomeric repeat sequence (AAACCCT) was used to estimate mean telomere length of an individual. Telomere length was calculated using the same equation and parameters applied with K-Seek. For subsequent analyses, all individuals were included except one which exhibited low genome coverage (12.31x) from the WGS data (Table S1).

A major difference between Computel and the two other bioinformatic approaches is the inclusion of genome coverage into telomere length estimations. Variability in genome coverage may impact telomere length estimations. Therefore, we normalized telomere length estimates by individual genome coverage using Equation 2. Telomere length values for K-seek and TRIP were re-assessed because the two programs did not previously consider genome coverage.

Equation (2) *Corrected telomere length*= $\frac{\text{FTL x RL}}{2\ x\ Chr}$ / *Genome coverage*

Genome coverage was calculated as the product of the total number of reads and read length, divided by total genome size for each individual. In the case of *Populus* , we use a genome size of 434,290,000 bp (*P. trichocarpa* v3, NC_037285.1,).

Quantitative PCR (qPCR) telomere length assay.

Quantitative polymerase chain reaction (qPCR) was used to quantify relative telomere length (rTL) using a Mx3000P QPCR System , following the methods of modified for *Populus* . qPCR quantifies the rTL for each individual as the ratio of the telomere repeat copy number to a single-copy control reference gene, relative to the reference sample . Plant-specific telomere primers were developed to assay telomere length alongside a housekeeping gene (glyceraldehyde-3-phosphate dehydrogenase; GAPDH), which is a single-copy control reference gene that exhibits limited sequence variability within the same individual.

qPCR reactions for the housekeeping GAPDH gene and telomere primers were run in triplicate on separate plates. Each qPCR reaction contained 20 ng of DNA per reaction in addition to 12.5 µl of SYBER Green Master Mix (Quantabio), 0.25 µl forward and reverse primer, 6 µl ultrapure water, and 6 µl of DNA. The qPCR cycle conditions for the GAPDH gene were 10 min at 95°C, followed by 33 cycles of 15 s at 95°C, 30 s at 59 °C and, 30 s at 72°C. qPCR conditions for the telomere region were 10 min at 95°C, followed by 20 cycles of 15 s at 95°C, 30 s at 58°C and 30 s at 72°C. Both GADPH and telomere assays were run on the same day, with*Populus* samples randomly assigned to a plate and three technical replicates per individual for each plate.

To calculate relative telomere length (rTL), we used the following formula:$2^{-\Delta\Delta Ct}$, where$\Delta\Delta Ct = \left(\mathrm{Ct}^{\mathrm{telomere}} - \mathrm{Ct}^{\mathrm{GAPDH}}\right) focal\ sample - \left(\mathrm{Ct}^{\mathrm{telomere}} - \mathrm{Ct}^{\mathrm{GAPDH}}\right)$ reference sample .*Ct* values indicate the number of PCR cycles required for the fluorescent signal to cross the threshold which is defined as the transition from linear phase to exponential phase of amplification for telomere and GAPDH reactions, respectively. *Ct* values are inversely proportional to the starting amount of the target sequence, samples with longer telomeres required fewer cycles to cross the threshold and thus had lower *Ct* -values. To standardize rTL across samples and plates, one reference *Populus* sample was run in triplicate on every plate.

For each plate, a serial dilution of DNA for a single *Populus*sample was prepared to calculate reaction efficiencies. To ensure that the DNA samples fell within the standard curve, we ran a five-point standard curve (40, 20, 10, 5, and 2.5 ng/mL of DNA) across all plates. Each point in the standard curve was run in triplicate. The reaction efficiencies for the telomere plates and GAPDH plates were within the accepted range (i.e., 100 ± 15%), with an average of 93.54 ± 3.82 % for the telomere plates and 88.78 ± 3.88 % for GAPDH plates. Additionally, we included a negative control of 6 µl of water run in triplicate on each plate, which did not produce Ct values.

Statistical analyses

Comparing bioinformatic approaches to quantifying telomere length

Telomere length estimates for each program were assessed for normality and homogeneity of variance using the Shapiro-Wilk test and Bartlett test, respectively. To achieve normality, telomere length estimates were $\log_{10}$-transformed. However, variances were unequal across computational approaches.

To compare telomere length across the three bioinformatic approaches we used non-parametric Kruskal–Wallis (H) test due to a significant difference in homogeneity of variance across telomere estimates. Following Kruskal–Wallis test, a post hoc Dunn test was performed to quantify differences between Computel, K-seek, and TRIP. To test if telomere length estimates from the same genotype are similar across the bioinformatic programs we performed Pearson pairwise correlations. All statistical analyses were performed in R v. 4.1.0 .

Comparing telomere length assayed using WGS and qPCR

To quantify the relationship between telomere length calculated using qPCR and WGS data we used a Pearson correlation. First, relative telomere length measured by qPCR was correlated with mean telomere length estimates calculated using Computel, K-seek, and TRIP. Then, to evaluate the role of genome coverage to estimates of telomere length we correlated telomere length measured by qPCR with re-assessed telomere length accounting for genome coverage (See Equation 2) from K-seek and TRIP.

Results

Telomere length assessments using whole-genome sequencing data

Telomere length estimates were significantly different across bioinformatic approaches (H = 230.06, df = 2, p< 0.001, Figure 2). A post-hoc Dunn test between telomere lengths estimated by Computel, K-Seek, and TRIP indicates significant differences for all pairwise comparisons ($p < 0.001$). On average, telomere length estimates using Computel (4,144 ± 796 bp) were substantially lower than K-seek (35,487 ± 11,326 bp) and TRIP (57,604 ± 17,393 bp). However, while the three bioinformatic approaches produced different telomere length estimates on average, there was substantial correlation between values estimated for individual genotypes (Figure S3). Correlations between genotype measures were lowest for Computel and K-seek ($r = 0.86$), but greater for Computel and TRIP ($r = 0.88$), and K-seek and TRIP ($r = 0.99$). This suggests that despite variability in the average length calculation there is a correlation between telomere length calculated across approaches. This may suggest that differences in the parameters included in the bioinformatics package, including the accounting for genome coverage and minimum number of consecutive telomeric repeats present in a read, may influence telomere length estimates on average, but telomere length calculated for the same genotype across approaches is largely correlated. Indeed, after correcting for genome coverage (Equation 2) pairwise correlations between genotype measures increased (Figure S4), but telomere length estimates were significantly different across bioinformatic approaches (H = 246.59, df = 2, p< 0.001, Figure S5). Importantly, K-seek (1,423 ± 300.65 bp) and TRIP (2,315± 451.67 bp) exhibit substantially shorter telomere length estimates after genome coverage correction.

Telomere length correlation estimated by qPCR and WGS.

Pearson correlations were used to quantify the relationship between telomere length estimated using qPCR and WGS data. Telomere length estimates from qPCR range between 2.06 to 0.67-fold change (Table S1). Telomere length calculated using Computel, K-seek and TRIP were all positively correlated with values calculated using qPCR (Fig 3). However, calculations based on Computel ($r = 0.66$, p< 0.001, Fig. 3A) were more strongly correlated than those observed for TRIP ($r = 0.52$, $p < 0.001$, Fig. 3B) or K-seek ($r = 0.50$, $p < 0.001$, Fig. 3C). Computel incorporates genome coverage into telomere length calculations, suggesting that genome coverage might be important to telomere length quantification. Indeed, after correcting for genome coverage in telomere estimates using K-seek and TRIP, re-assessed correlations were greater for both K-seek ($r = 0.62$, $p < 0.001$, Figure 4A) and TRIP ($r = 0.66$, $p < 0.001$, Figure 4B). Additionally, telomere estimates from K-seek and TRIP (Equation1) have a strong correlation with genome coverage (r = 0.81). These results point toward the importance of considering genome coverage when calculating telomere length using WGS.

Discussion

Telomere length has become an important tool for studying species' response to environmental stress, and there is a growing need to estimate telomere length rapidly and efficiently for model and non-model organisms. Recent advances in WGS and bioinformatic approaches have created new opportunities for telomere length assessment . In this study, we used whole-genome resequencing data to estimate telomere length for 100 *Populus* genotypes and compared our results to traditional qPCR measurements. We identified three bioinformatic approaches that were appropriate for telomere length assessments in plants. While average telomere length estimates varied across bioinformatic approaches, estimates for the same genotype across approaches were strongly correlated (r = 0.86 to 0.99). Furthermore, our results suggest that WGS provides a comparable approach to estimating telomere length variation relative to traditional qPCR . Indeed, correlation between telomere estimates for qPCR and WGS, particularly where genome coverage is accounted for, suggests that values are comparable. Our study demonstrates that WGS is an efficient and rapid approach for assessing telomere length in plants. This has potential applications for plant breeding and conservation

management where assessment of telomere length change acts as a biomarker to indicate individuals response to environmental stress.

Comparison of telomere length estimates from WGS data

On average, significant differences were observed across bioinformatic approaches that measured telomere length (H = 230.06, df = 2,$p < 0.001$, Figure 2). However, despite these differences, individual genotype measures were highly correlated across approaches (r = 0.86 – 0.99). This suggests that regardless of the approach, estimates are comparable, but the scale of estimates differs on average. Variation in telomere length estimates may be attributed to differences in the bioinformatic approaches, including telomere identification (i.e., alignment or matching pattern approach), minimum number of consecutive telomeric repeats required in a read, and consideration of genome coverage. Computel was initially designed to estimate mean telomere length in humans but allows species-specific modification of genome features, including genome size, number of chromosomes, and telomere sequence to allow estimation across organisms . Here, we leveraged Computel to estimate telomere length in plants. Computel uses an alignment-based method by mapping reads from WGS data to a telomere reference created within the program. Only reads that align with the telomere reference are considered telomeric reads. In contrast, K-seek was not developed to estimate telomere length but was created to identify and count simple sequence repeats from WGS in*Drosophila* . We leveraged K-seek to estimate telomere length by identifying and counting the number of predicted *Populus* telomere repeats from WGS within each genotype. K-seek considers short repeats with a minimum repeating length of 50 bp within a read so that reads containing a minimum of seven telomeric repeats were identified in the analysis. This approach decreases the probability of capturing interstitial telomeres, which are telomeric repeats localized to intrachromosomal sites . However, unlike Computel, K-seek does not include other parameters that are known to influence telomere estimates such as genome coverage . Similarly, TRIP identifies short tandem repeat sequences from WGS, but this program was specifically created for telomere identification in insects . TRIP detects reads with more than four telomeric repeats per read, and like K-seek, does not consider genome coverage in telomere length estimates. Including genome coverage can influence telomere length estimates on average by reducing potential sequencing biases . Nersisyan and Arakelyan (2015) compared human telomere length using short read sequence data with varying degrees of coverage across the same individuals (0.2, 2 and 10x). They observed that the accuracy of telomere estimates improved with higher genome coverage. In our study, we removed one individual from the analyses as an outlier as it exhibited low genome coverage (< 12.31X, Table S1). Individuals assessed in our study had a minimum genome coverage of 15X, suggesting that this may be reasonable requirement to precisely estimate telomere length using WGS. (Table S1). However, further studies comparing the impact of varying genome coverage in telomere estimations are needed to support this recommendation in plants. Therefore, while there may be benefits to the matching approach used in K-seek and TRIP, ensuring that genome coverage is considered will be essential for future comparisons.

Telomere length estimates correlated between WGS and qPCR

We observed a correlation (r = 0.50 – 0.66) between telomere length estimates using qPCR and WGS data. Previous studies using humans observed a moderate to high correlation between qPCR and WGS data (r = 0.66 – 0.95, . For *Populus* trees, lower correlations may reflect differences in bioinformatics approaches that filter out interstitial telomeric sequences and tolerate telomere repeat variants.

Telomere studies in humans avoid the inclusion of interstitial telomeric repeats by identifying telomeric reads that are aligned with the telomere region in the reference genome. Programs such as TelSeq , Telomere Hunter , and Telomerecat require alignment of reads to a reference genome prior to telomere calculations (i.e., BAM files) and telomere length estimations are based on reads that align to the telomeric region of the genome. This approach reduces the probability of capturing interstitial telomeric repeats, but the use and efficiency of the program depend heavily on the completeness of the telomere-to-telomere assembly of the reference genome . While telomere-to-telomere assemblies are available in humans most species lack high resolution across telomeric regions. For the *Populus trichocarpa* reference genome, the telomeric region contains a significant proportion of ambiguous bases (i.e., NNNN) reducing the probability of alignment to the

telomere region. In this study, we limited the estimation of telomere length to programs that used unmapped read sequence data. Thus, telomere estimates from both K-seek and TRIP may include some interstitial telomeric repeats. Nonetheless, previous studies have shown that considering consecutive telomeric repeats decreases the probability of capturing interstitial telomeric repeats and increases the correlation between qPCR and TRF telomere estimates . K-seek and TRIP considered only reads with more than four and seven consecutive repeats, respectively, reducing the probability of capturing interstitial telomeres in this study. In contrast, Computel decreases the capture of interstitial telomeres by using reads that align with the telomeric reference created by the program . For non-model species, it is possible to use sequence data unmapped to a reference genome to estimate telomere length, however consideration of potential caveats in telomere length estimations from interstitial repeats is required.

Additional bioinformatic programs such as TelomeHunter and qmotif allow telomere repeats to deviate from the typical human telomere repeat, TTAGGG . Although telomeric repeats are generally conserved within a species, deviations from the typical telomere repeat have been reported in humans and are frequently considered into telomere calculations . In plants, telomere variants that deviate from the *Arabidopsis* type (TTTAGGG) are reported between taxa, with some families such as Alliaceae exhibiting novel telomere sequences, CTCGGTTATGGG . However, to date there is limited empirical data comparing telomere repeat variation within species. In the present study, we searched only for the telomere repeat TTTAGGG previously reported as the *Populus*telomeric sequence . Despite this, potential *Populus* telomere repeat variants can be visually detected through manual inspection of the *Populus trichocarpa* reference genome. If telomeric variants within *Populus* were excluded, our correlations between WGS and qPCR may increase. Thus, the identification of intraspecific telomere repeat variation in plants, coupled with new bioinformatic approaches that include telomere repeat diversity will improve telomere estimations.

Telomere length estimates in plants are currently limited to programs that allow modification of telomere repeat pattern and species-specific genome features. Telomere repeat pattern is taxa-dependent with most vertebrates sharing the human telomere repeat pattern, TTAGGG . Multiple programs listed above, including TelomereHunter, Telseq and Telomerecat, were created to identify human telomere repeats limiting the repeat search to the vertebrate telomere type . In addition, telomere estimates for these programs are performed considering human genome features, such as number of chromosomes and genome length. Plants have different telomere repeat patterns, generally TTTAGGG, deviating from the human telomeric type . To our knowledge, the only program that allows the modification of telomere repeat patterns and genomic features is Computel . Computel allows uses species-specific genome features, including telomere pattern, number of chromosomes, and genome size. The greatest correlation between WGS and qPCR ($r = 0.66$) was observed for Computel. Previous studies indicate that Computel performs similarly to other bioinformatic approaches , but as the field of telomere ecology expands increased flexibility to modify the telomere repeat pattern and include species-specific genome features will be required to extend applications.

Accurate measurement of telomere length is needed to deploy telomeres as potential biomarkers to quantify organismal response to abiotic and biotic stress. Although qPCR has been used extensively due to its accessibility and opportunities for high throughput analysis, it provides only a relative measurement rather than an absolute measure of telomere length Furthermore, qPCR accuracy in assaying telomere length is susceptible to potential variations in the reference control gene, primer efficiency, and inter-assay variability . WGS can provide a high-resolution assessment of the telomeric regions allowing for precise quantification of absolute telomere length. In addition, WGS allows detection of mutations within the telomeric regions and permits telomere length assessment on an individual chromosome basis. Thus, while WGS can be computationally intensive and potentially cost-prohibitive for large-scale studies, WGS can enhance the accuracy of current telomere length methods, particularly for techniques involving subtelomeric primers or probes, using sequence data .

Conclusion

Telomeres are a potential biomarker for quantifying species' response to environmental stress. Therefore, it is critical to evaluate innovative approaches for estimating telomere length in both model and non-model organisms. In this study, we demonstrated that estimating telomere length from WGS is a feasible approach for *Populus* trees, offering a new opportunity to estimate telomere length for plants. Although there are potential caveats that need to be addressed in future studies, our results show that telomere length estimated using short-read whole genome sequence data yield comparable results to traditional qPCR. Importantly, we suggest that telomere length estimates derived from WGS is more accurate than those obtained from qPCR, by providing the absolute telomere length. We specifically recommend Computel to estimate telomere length using WGS as it incorporates genome coverage into telomere estimations and provides species-specific parameterization needed to correlate with qPCR telomere estimates. We suggest that telomere length estimates from WGS are sensitive to genome coverage, and this may be a major consideration to study sequence design. The results of this study open new avenues for estimating telomere length across diverse organisms and will help accelerate telomere ecological research in plants.

Acknowledgments

We thank Kyle Peer, Clay Sawyers, and Deborah Bird (Virginia Tech Reynold's Homestead Forestry Research Station) for assistance with plant propagation. We also thank to Jeffrey D Kittilson for helping with the qPCR essays. We also would like to thank to the NSF (NSF-PGR-1856450), USDA National Institute of Food and Agriculture and Hatch Appropriations (Project #PEN04809 and Accession #7003639), Schatz Center for Tree Molecular Genetics, Huck Institutes of the Life Sciences, Department of Ecosystem Science and Management, and Pennsylvania State University's Intercollege Graduate Degree Program in Ecology for the funding support.

References

Data Accessibility and Benefit-Sharing

The datasets, including qPCR results and all scripts, supporting the conclusions of this article will be available in GitHub repository upon acceptance. Demultiplexed DNA sequences will be available in the National Centre for Biotechnology Information database upon acceptance.

Benefit-Sharing Statement

Benefits from this research accrue from the sharing of our data and results on public databases as described above.

Authors' contributions

M.Z.P. contributed to designing the research, performing qPCR assays, performing computational analyses, analyzing the data, and writing the manuscript. J.H. contributed to data collection, DNA sampling and sequencing and editing the manuscript. J.A.H. contributed to data collection, designing the research, analyzing the data, and writing and editing the manuscript.

Competing interests

The authors declare that they have no conflict of interests.

Figures

Figure 1. Map of individuals sampled ($n = 100$) across the contact zone between *Populus trichocarpa* and *P. balsamifera* in the vicinity of the Rocky Mountains. Yellow dots represent the geographic location of each sampled individual. Light and dark gray areas represent *P. trichocarpa* and *P. balsamifera* distribution (Little, 1971).

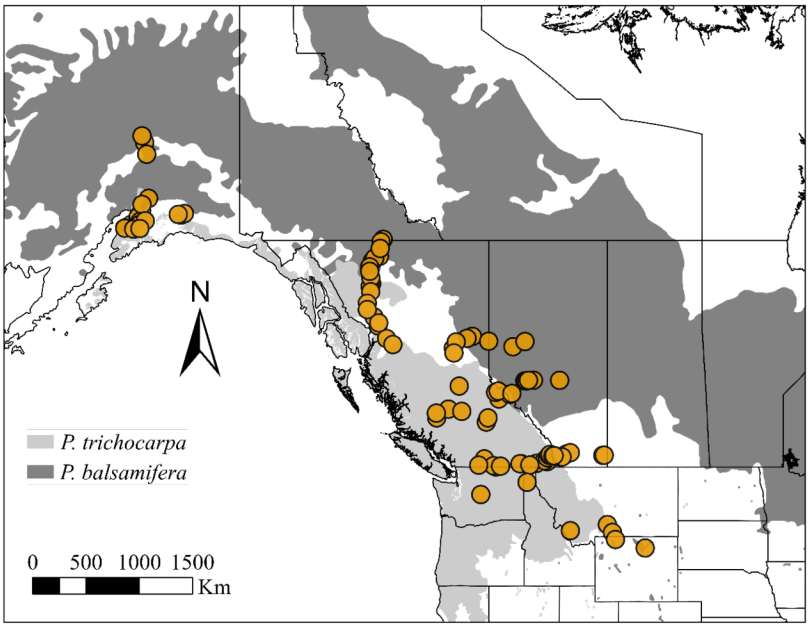

Figure 2. Box plots of telomere length estimated by three bioinformatics approaches. Horizontal line indicates the median and points are individual measurements. The three programs were significantly different based on Dunn test comparison of means ($p < 0.001$). Letters indicate significant differences between bioinformatic programs.

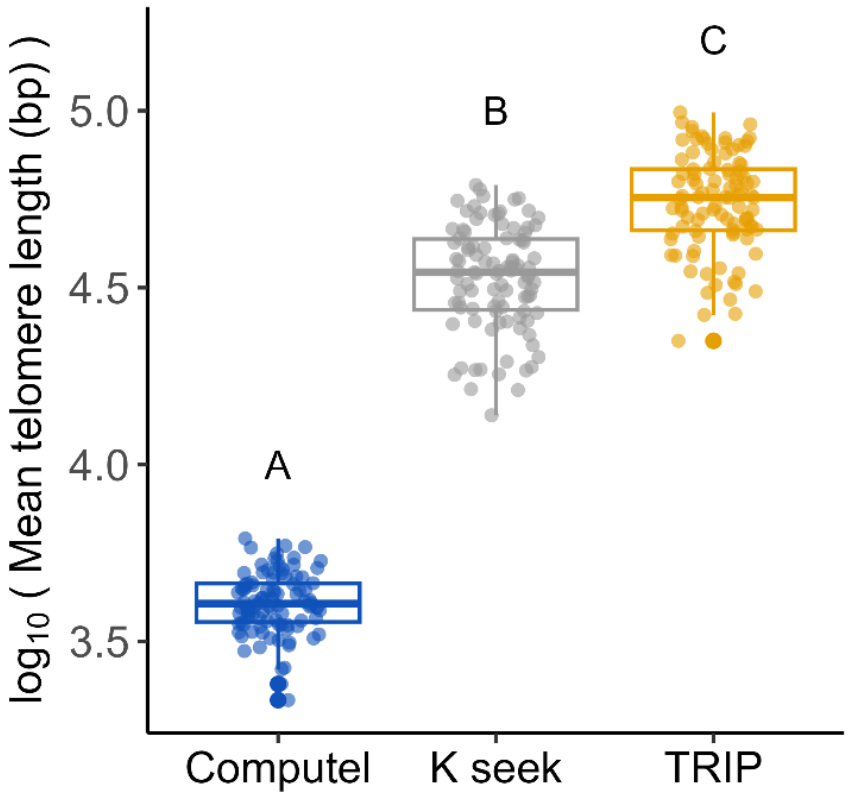

Figure 3. Telomere length estimated from sequence data using three bioinformatic tools (x-axis) compared with relative telomere length (rTL) estimated from qPCR (y - axis). A) Computel, B) K-seek, and C) TRIP. Pearson correlation coefficient (r) and p - *values*are shown in the upper left side of each plot.

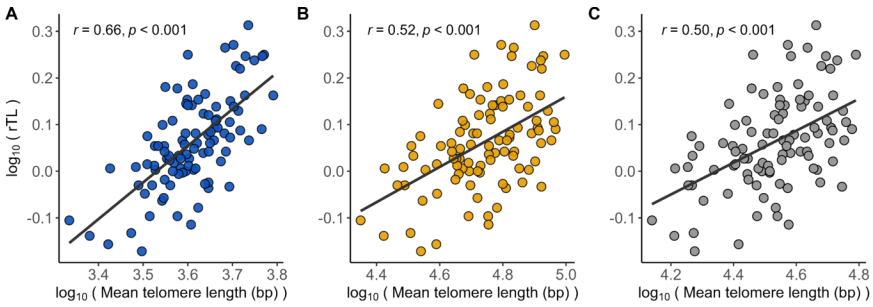

Figure 4. Correlation between telomere length corrected for genome coverage and telomere length assessed by qPCR for A) TRIP and B) K-seek. Pearson correlation coefficient (r) and p -*values* are shown in the upper left side of each plot.

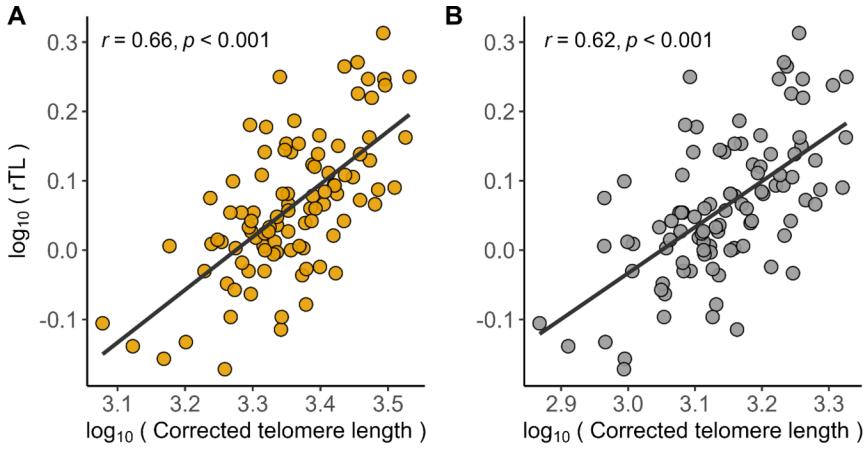

Table 1. Primers used for Quantitative polymerase chain reaction (qPCR).

Primer name	Sequence 5'-3'
GAPDH Forward	AGGGCTGCTTCCTTCAATATC
GAPDH Reverse	CGGCCGTAGGTGCATAAA
Telomere Forward	CCCCGGTTTTGGGTTTTGGGTTTTGGGTTTTGGGT
Telomere Reverse	GGGGCCCTAATCCCTAATCCCTAATCCCTAATCCCT