Haplotype Diversity and Sequence Heterogeneity of Human Telomeres

Kirill Grigorev[1,2 #], Jonathan Foox[1,2,3 #], Daniela Bezdan[1,2,3], Daniel Butler[1], Jared J. Luxton[4,5], Jake Reed[1], Miles J. McKenna[4,5], Lynn Taylor[4,5], Kerry A. George[4,5], Cem Meydan[1,2,3], Susan M. Bailey[4,5 *], Christopher E. Mason[1,2,3,6 *]

[1] Department of Physiology and Biophysics, Weill Cornell Medicine, New York, New York, USA
[2] The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medicine, New York, New York, USA
[3] The Feil Family Brain and Mind Research Institute, New York, New York, USA
[4] Department of Environmental and Radiological Health Sciences, Colorado State University, Fort Collins, CO
[5] Cell and Molecular Biology Program, Colorado State University, Fort Collins, CO
[6] The WorldQuant Initiative for Quantitative Prediction, Weill Cornell Medicine, New York, NY, USA
[#] Co-first authors
[*] Corresponding authors. Send correspondence to S.M.B. (susan.bailey@colostate.edu) and C.E.M. (chm2042@med.cornell.edu)

Abstract

Telomeres are regions of repetitive nucleotide sequences capping the ends of eukaryotic chromosomes that protect against deterioration, and whose lengths can be correlated with age and adverse health risk factors. Given their length and repetitive nature, telomeric regions are not easily reconstructed from short-read sequencing, making telomere sequence resolution a very costly and generally intractable problem. Recently, long-read sequencing, with read lengths measuring in hundreds of Kbp, has made it possible to routinely read into telomeric regions and inspect their sequence structure. Here, we describe a framework for extracting telomeric reads from single-molecule sequencing experiments, describing their sequence variation and motifs, and for haplotype inference. We find that long telomeric stretches can be accurately captured with long-read sequencing, observe extensive sequence heterogeneity of human telomeres, discover and localize non-canonical motifs (both previously reported as well as novel), confirm the presence of the non-canonical motifs in short read sequencing experiments, and report the first motif composition maps of human telomeric diplotypes on a multi-Kbp scale.

Keywords

Telomere, telomeric haplotypes, long-read sequencing, telomere sequence heterogeneity

Introduction

Telomeres are the functional ends of human chromosomes that naturally shorten with cell division and therefore with age [1]. Telomere length can also be influenced by a variety of lifestyle factors and environmental exposures (e.g., stress, exercise, air pollution, radiation) [2]. While human telomeres are known to consist largely of a conserved six-nucleotide repeat (TTAGGG) [3], several studies have identified variations of this motif in proximal telomeric regions [4–7]. However, such studies were performed with oligonucleotide hybridization, PCR, immunoprecipitation, and short-read sequencing, resulting in discovery, but not localization, of motif variants. Thus, long-range maps of telomeric sequence variation in the human genome are still lacking. Such maps can provide insight into telomere biology and enable novel approaches to analyze the effects of health status, aging, and environment on telomere structure and length.

To improve our understanding of telomere sequence structure and variation, we developed *edgeCase*, a framework for alignment, motif discovery, and haplotype inference from human telomeric reads. We have validated these methods using Genome in a Bottle [8] single-molecule real-time (SMRT) sequencing datasets generated with Pacific Biosciences circular consensus sequencing (PacBio CCS) [9, 10], and short-read Illumina [11] and 10X Genomics (Chromium) [12] datasets. These results provide evidence for multiple novel, non-canonical telomeric repeats, resolution of chromosome-specific diplotypes with SMRT sequencing, and a new method for long-range characterization of the structure of telomeric sequences.

Results

Telomeric reads are present in human long-read whole genome sequencing datasets

We aligned PacBio CCS reads of three Genome in a Bottle (GIAB) human subjects (HG001, HG002, and HG005) to a combination of the human reference genome and human subtelomeric assemblies (see Materials and Methods). In total, we observed reads mapping to the ends of chromosomes and extending into telomeric regions on 9 *p* arms and 17 *q* arms, with 256 such reads (∼10x mean coverage) in the HG001 dataset, 570 (∼22x) in HG002, and 241 (∼9x) in HG005. **??** schematically represents the alignment of such reads in the HG002 dataset; alignment plots for the other two datasets are available as a supplemental figure (**??**), and full mapping statistics are available in **??**. Illumina reads from matching GIAB datasets supported 70.8%, 63.3%, and 82.7% of the candidate PacBio CCS sequence, providing average coverages of ∼5x, ∼9x, and ∼6x, respectively.

Telomeric reads contain variations of the canonical motif

We performed *de novo* repeat discovery for motifs of lengths 4 through 16 and identified motifs in repeat contexts that are statistically enriched in the three datasets. The majority of motifs were either the canonical TTAGGG / CCCTAA, its variations (e.g., TTGGGG / CCCCAA), or a duplet of variants, such as TTAGGGTTAGGGG (**??**). CG-rich motifs were also observed on the *p* arms. The top enriched motif (TTAGGG / CCCTAA) explained 43.3%–54.4% of the telomeric repeat content on the *q* arms, and 10.0%–22.7% on the *p* arms, These top motifs, as well as 15 less enriched ones, were confirmed in independently generated human short read and linked-read genomic datasets (Supplemental methods, **??**). **??** visualizes the locations of the top four enriched motifs on the *q* arm of the HG002 dataset; only the arms covered by at least 25 reads are displayed. Plots for other datasets and arms are available as supplemental figures: **??** visualizes the top three motifs on the *p* arm of the HG002 dataset, **??** and **??** visualize datasets HG001 and HG005 respectively. Long reads on each arm agreed on the locations of different motifs within any given 10 bp window (the median of normalized Shannon entropy was 0.000 for all data, and the 3rd quartile was 0.166, 0.074, and 0.211 for the three datasets, respectively, **??**), indicating that locations of the variations are colinear among reads and are not a result of sequencing errors.

Long-read sequencing resolves human telomeric haplotypes

Sequences of telomeric reads clustered by relative pairwise Levenshtein distances [13] with varying levels of heterogeneity depending on the dataset and the chromosomal arm to which they belonged. We examined the *q* arms of the HG002 dataset to investigate this heterogeneity, as they provided the deepest coverage (**??**), and found that, on 12 out of the 15 arms, reads clustered into two prominent groups per arm when maximizing the Bayesian information criterion [14] (see Materials and Methods). Pairwise distances between the reads within these clusters were significantly lower than those for out-of-cluster pairings, implying that distinct telomeric haplotypes are present. To quantify the differences between putative haplotypes, we calculated silhouette scores [15] for these clusterings (**??**), and generated motif density plots for the four chromosome arms with the highest such scores to visualize the differences in haplotypes (**??**).

Discussion

Repeat-rich, low-complexity regions of the human genome such as telomeres have been historically recalcitrant to full mapping and annotation [16], mainly due to the alignment challenge they pose and to the read lengths required to span such areas [17]. The advent of long-read, single-molecule methods (third generation

sequencing) has provided new opportunities to map the sequence composition of a previously "dark" area of the human genome, enabling research into the sequence composition and length dynamics **[luxton2020]** of telomeres. Our results reaffirm that the canonical repeat (TTAGGG) is certainly the most dominant type of motif in telomeres, but also reveal a surprising diversity of repeat variations, which are confirmed by both short and long-read sequencing technologies. This diversity of repeats includes previously reported variants, as well as novel motifs that are characterized not only by nucleotide substitutions, but also insertions, deletions, and even motif pairing. Apart from these variations, CG-rich motifs were identified in telomeric regions of *p* arms, consistent with previously reported findings [18]. Moreover, while short read sequencing is able to identify such variants, it alone cannot reveal the relative locations of these motifs within telomeres, as repetitive short reads can neither be aligned outside of the reference genome nor provide enough overlap variability to be assembled *de novo*. Long SMRT reads, on the other hand, can be anchored to known subtelomeric sequences of the human genome and extend into the previously unmapped telomeric area. These results also highlight the need of better subtelomeric and telomeric annotations in the human genome. Four of the 40 subtelomeric assemblies [19] were homologous to regions in the reference genome far within the respective chromosomes (up to 586 Kbp into the reference sequence), and the canonical motif was present on the *q* arm of chr8 only after 2–3Kbp past the annotated boundary in all datasets, suggesting that the existing assemblies do not provide a completely accurate telomeric annotation, and that methods described herein could help to resolve these areas of reference genomes.

We observed PacBio CCS reads reaching up to 16 Kbp beyond the known regions of the genome, and resolving the underlying sequence with reasonable fidelity, measured both by the entropy of motif assignment and by pairwise Levenshtein distances between the reads belonging to the same chromosomal arms. While short reads also provided support for non-canonical motifs, the overlap between the short and the long reads was substantial, but not complete, which can be explained by the necessary bias towards the canonical motif during the selection of short reads. Therefore, telomeric regions with higher content of non-canonical repeats are less likely to be identified through the use of short reads, and instead, long reads appear to be more suitable for this purpose as well. The identified variations in long range contexts enable clustering of SMRT reads into distinct haplotypes at ends of chromosomes, and thus provide a new means of diplotype mapping and reveal the existence and motif composition of such diplotypes on a multi-Kbp scale.

Materials and Methods

The extended reference genome

We constructed the extended reference genome by performing an all-to-all alignment of all contigs in the *hg38* reference genome [20, 21] and the subtelomeric assemblies [19] with *minimap2* [22] using three settings for assembly-to-reference mapping (*asm5*, *asm10*, *asm20*). Forty subtelomeric contigs mapped to ends of *hg38* chromosomes with a mapping quality of 60, one (XpYptel) mapped with the quality of 0 and was discarded; one (14qtel) mapped to the ALT version of chr14 (chr14_KI270846v1_alt) with the quality of 52, which, in turn, mapped to the main chr14 chromosome with the quality of 60. These data and the exact match and mismatch coordinates were used to create a combined reference (*hg38ext*) in which subtelomeric contigs informed the locations of the boundaries of the telomeric tracts (*tract_anchor*). Such contigs that mapped fully within *hg38* chromosomes resulted in *tract_anchor* annotations directly on those *hg38* chromosomes; partially mapping contigs were considered as forking from the *hg38* sequence and were similarly annotated by themselves.

Detection of telomeric sequences in long-read datasets

Three subjects were selected for the analysis. The first individual (NA12878/HG001) came from the pilot genome of the HapMap project [23], while the other two, including the son from the Ashkenazi Jewish Trio (NA24385/HG002) and the son from the Chinese Trio (NA24631/HG005), are members of the Personal Genome Project, whose genomes are consented for commercial redistribution and reidentification [24]. These subjects are referred to hereafter as HG001, HG002, and HG005, respectively.

For subjects HG001 and HG005, Genome in a Bottle [8] PacBio_SequelII_CCS_11kb datasets were used (one dataset per each subject). For subject HG002, a combination of two sequencing experiments was analyzed (PacBio_CCS_10kb and PacBio_CCS_15kb). The mean coverage was $\sim$29x, $\sim$58x, and $\sim$32x for subjects HG001, HG002, and HG005, respectively. Reads were mapped to *hg38ext* with *minimap2*, and reads that mapped to either end of either chromosome and overlapped the boundary of its telomeric tract were selected for further analysis. These reads had a portion of their sequence mapped to the reference contig and a portion extending beyond the reference (soft- or hard-clipped in the alignment file). Sequences past the *tract_anchor* marker were extracted from the reads that had this marker within their mapped portion (from the 5' end to the marker on *p* arms and from the marker to the 3' end on *q* arms, accounting for forward and reverse mappings). To identify regions of the telomeres that are fully supported by both short and long reads, we extracted candidate telomeric reads from GIAB Illumina datasets (NIST_NA12878_-

HG001_HiSeq_300x, NIST_HiSeq_HG002_Homogeneity-10953946, HG005_NA24631_son_HiSeq_300x; all three ∼300x coverage) with *Telomerecat* [25], and selected those that mapped perfectly with *minimap2* (at least a 50bp-long exact match without insertions or deletions, allowing all secondary mappings) to the telomeric regions of the PacBio CCS candidates from the same subject's dataset.

Detection of telomeric sequences in short- and linked-read datasets

To evaluate sequence motifs in datasets generated by technologies other than SMRT, we generated four whole-genome Illumina datasets (mean coverage ∼104x) and three linked-read 10X datasets (mean coverage ∼28x) for one individual at different timepoints aboard the International Space Station (ISS), and one additional linked-read 10X dataset (coverage ∼47x) for another individual aboard the ISS. Blood samples were collected from astronaut subjects as described in [**twins_study**]. For each sample, 1.2ng of sorted immune cell input was aliquoted for TruSeq PCR-free WGS (short read) and standard Chromium 10X whole genome (linked-read) preparation respectively, and sequenced across one S4 flow cell on an Illumina NovaSeq 6000. From these datasets, candidate telomeric short reads were selected using Telomerecat [25].

Identification of repeat content

Overrepresentation of motifs of lengths $k \subset [4..16]$ was tested within the candidate telomeric regions of PacBio CCS reads, as well as in the candidate reads from independently generated Illumina and 10X Chromium datasets. To target motifs in repeat contexts, doubled sequences (for example, *k*-mer ACGTACGT for motif ACGT) were counted with *jellyfish* [26], and counts of *k*-mers synonymous with respect to circular shifts (for example, ACGTACGT and CGTACGTA) were summed together. For each such *k*-mer, Fisher's exact test was performed to determine whether its count is significant on the background of counts of other *k*-mers of the same length. Briefly, we considered *k*-mers with counts higher than 1.5 interquartile range above the third quartile of the distribution as potentially classifiable, and a 2×2 contingency matrix C for the test was constructed as follows: row 0 contained counts of potentially classifiable *k*-mers, row 1 contained counts of remaining (non-classifiable) *k*-mers, columns 0 and 1 contained counts of single and remaining (background) *k*-mers, respectively, i.e.: $C_{0,0} = \text{count of target } k\text{-mer}$, $C_{0,1} = \text{sum of counts of other potentially classifiable } k\text{-mers}$, $C_{1,0} = \text{median count of } k\text{-mer}$, $C_{1,1} = \text{sum of counts of other non-classifiable } k\text{-mers}$. The resultant *p*-values for each motif among the samples were combined using the Mudholkar-George method [27] within each technology (PacBio CCS, Illumina, 10X Genomics), and the Bonferroni multiple testing correction was applied Motifs in the long-read datasets for which *k*-mers yielded *p*-values below the cutoff of 0.05 were reported. Additionally, motifs that were significantly enriched in the datasets produced

by all three technologies (PacBio, Illumina, 10X), with respect to reverse-complemented equivalence, were reported.

Evaluation of sequence concordance in telomeric long reads

As telomeric reads contain long low-complexity regions and present an alignment challenge, we evaluated concordance of their sequences without realignment of their portions that extended past the reference sequence. To that end, for all reads mapping to the same chromosomal arm, we calculated densities of each identified motif in a rolling window starting from the innermost mapped position of each entire read. To evaluate whether the reads on the same arm agree on the positions of different motifs, for each read, we calculated motif densities in 10 bp windows with 10 bp smoothing to buffer insertions and deletions. For each window in each read, the motif with the highest density was selected to represent that window. Then, normalized Shannon entropy among all reads was calculated in each window as $S = \frac{-\sum_i (p_i ln p_i)}{ln N}$, where p_i is the frequency of each motif in the window and N is the number of motifs [28]. The value of normalized entropy was a metric bounded by $[0, 1]$, with 0 describing perfect agreement and 1 describing maximum randomness. For visualization, we performed 1000 rounds of bootstrap of the calculated density values in the 10 bp rolling windows, and selected the lower and the upper bounds of the 95% confidence interval of bootstrap. Of note, several chromosome arms had the *tract_anchor* position further away from the end of the contig than others ($\sim$79–586 Kbp into the chromosome sequence), and the reads mapping to these arms did not contain these motifs, suggesting that either their subtelomeric annotations were incorrect or large insertions or duplications were present in the reference genome; in light of this, reads mapping to the *p* arm of chr1, the *q* arm of chr4, and both arms of chr20 were removed from the study, and the analysis was repeated.

Extraction of telomeric haplotypes from long-read datasets

Within groups of reads mapping to each chromosome arm, all relative pairwise Levenshtein distances were calculated. In short, to calculate the absolute distance between each pair of reads, the sequences in the overlapping positions of the reads were extracted; the distance then equaled the minimum number of single-character insertions, deletions, and substitutions required to make these sequences identical. The relative distance was computed as the absolute distance divided by the length of the overlap. Relative distances were then clustered using Ward's method via the Euclidean metric.

Data access

The NASA Life Sciences Data Archive (LSDA) is the repository for all human and animal research data, including the whole genome Illumina and 10X Chromium sequencing datasets from subjects aboard the ISS that were used in this study. These datasets are protected by the terms of the Weill Cornell Medicine Internal Review Board (IRB) and can be made available to be shared upon request. LSDA has a public facing portal where data requests can be initiated (lsda.jsc.nasa.gov/Request/dataRequestFAQ); the LSDA team provides the appropriate processes, tools, and secure infrastructure for archival of experimental data and dissemination while complying with applicable rules, regulations, policies, and procedures governing the management and archival of sensitive data and information. The LSDA team enables data and information dissemination to the public or to authorized personnel either by providing public access to information or via an approved request process for information and data from the LSDA in accordance with NASA Human Research Program and JSC Institutional Review Board direction.

The software for identification of telomeric reads, *de novo* discovery of repeat motifs, haplotype inference and motif density visualization was implemented in Python and is freely available at github.com/lankycyril/edgecase.

Acknowledgements

We would like to thank the Epigenomics Core Facility at Weill Cornell Medicine, the Scientific Computing Unit (SCU), XSEDE Supercomputing Resources, as well as the STARR grants I9-A9-071, I13-0052, The Vallee Foundation, The WorldQuant Foundation, The Pershing Square Sohn Cancer Research Alliance, NASA (NNX14AH51G, NNX14AB02G, NNX17AB26G), The National Institutes of Health (R01MH117406, R01NS076465, R01CA249054, R01AI151059, P01HD067244, P01CA214274), TRISH (NNX16AO69A:0107, NNX16AO69A:0061), the LLS (9238-16, Mak, MCL-982, Chen-Kiang), and the NSF (1840275).

Author contributions

S.M.B. and C.E.M. conceived the study. K.G., J.F., and C.E.M. developed the framework and analyzed the data. D.Bu., J.J.L., M.J.M., L.T., and K.A.G. participated in collection and processing of the ISS samples. D.Be., D.Bu., J.J.L, J.R., and C.M. analyzed the data. All authors edited the manuscript.

Competing interests

The authors declare no relevant conflict of interest, although C.E.M. is a Co-Founder of Onegevity.

References

1. Aubert, G. & Lansdorp, P. M. Telomeres and Aging. *Physiological Reviews* **88** (Apr. 2008).
2. Shammas, M. A. Telomeres, lifestyle, cancer, and aging. *Current Opinion in Clinical Nutrition and Metabolic Care* **14** (Jan. 2011).
3. Moyzis, R. K. *et al.* A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proceedings of the National Academy of Sciences* **85** (Sept. 1988).
4. Allshire, R. C., Dempster, M. & Hastie, N. D. Human telomeres contain at least three types of G–rich repeat distributed non-randomly. *Nucleic Acids Research* **17** (1989).
5. Coleman, J., Baird, D. M. & Royle, N. J. The Plasticity of Human Telomeres Demonstrated by a Hypervariable Telomere Repeat Array That Is Located on Some Copies of 16p and 16q. *Human Molecular Genetics* **8** (Sept. 1999).
6. Lee, M. *et al.* Telomere sequence content can be used to determine ALT activity in tumours. *Nucleic Acids Research* **46** (Apr. 2018).
7. Bluhm, A. *et al.* ZBTB10 binds the telomeric variant repeat TTGGGG and interacts with TRF2. *Nucleic Acids Research* **47** (Jan. 2019).
8. Zook, J. M. *et al.* An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology* **37** (Apr. 2019).
9. Eid, J. *et al.* Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323** (Jan. 2009).
10. Ardui, S., Ameur, A., Vermeesch, J. R. & Hestand, M. S. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research* **46** (Feb. 2018).
11. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456** (Nov. 2008).
12. 10x Genomics. *Resolving Biology to Advance Human Health* https://www.10xgenomics.com/ (2020).
13. Levenshtein, V. I. *Binary codes capable of correcting deletions, insertions, and reversals* in *Soviet physics doklady* **10** (1966).
14. Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics* **6** (Mar. 1978).
15. *Finding Groups in Data* (eds Kaufman, L. & Rousseeuw, P. J.) (John Wiley & Sons, Inc., Mar. 1990).
16. Miga, K. H. Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Research* **23** (Sept. 2015).
17. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13** (Nov. 2011).
18. Nergadze, S. G. *et al.* CpG-island promoters drive transcription of human telomeres. *RNA* **15** (Oct. 2009).
19. Stong, N. *et al.* Subtelomeric CTCF and cohesin binding site organization using improved subtelomere assemblies and a novel annotation pipeline. *Genome Research* **24** (Mar. 2014).
20. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research* **27** (Apr. 2017).
21. Initial sequencing and analysis of the human genome. *Nature* **409** (Feb. 2001).
22. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34** (May 2018).

23. The International HapMap Project. *Nature* **426** (Dec. 2003).
24. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data* **3** (June 2016).
25. Farmery, J. H. R., Smith, M. L. & Lynch, A. G. Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Scientific Reports* **8** (Jan. 2018).
26. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27** (Jan. 2011).
27. George, E. O. & Mudholkar, G. S. On the convolution of logistic random variables. *Metrika* **30,** 1–13 (Dec. 1983).
28. Minosse, C. *et al.* Possible Compartmentalization of Hepatitis C Viral Replication in the Genital Tract of HIV-1–Coinfected Women. *The Journal of Infectious Diseases* **194** (Dec. 2006).