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. Yet, given their length and repetitive nature, telomeric regions are not easily reconstructed from short-read sequencing, making telomere sequencing, mapping, and variant resolution difficult problems. 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 whole genome single-molecule sequencing experiments, including *de novo* identification of telomere repeat motifs and repeat types, and also describe their sequence variation. We find that long, complex 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 haplotypes across three distinct ancestries (Ashkenazim, Chinese, and Utah) and two trios 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 thus with age (Aubert and Lansdorp 2008). Telomere length is also influenced by a variety of lifestyle factors and environmental exposures (e.g., stress, exercise, air pollution, radiation) (Shammas 2011). While human telomeres are known to consist largely of a conserved six-nucleotide repeat (TTAGGG) (Moyzis et al. 1988), several studies have identified variations of this motif in proximal telomeric regions (Allshire et al. 1989; Coleman et al. 1999; Lee et al. 2018; Bluhm et al. 2019). However, such studies were performed with oligonucleotide hybridization, PCR, immunoprecipitation, and short-read sequencing, requiring prior assumptions about specific target motifs, custom sample preparation, and targeted sequencing, and therefore preventing *de novo* identification of motif variants and their localization. Thus, long-range maps of telomeric sequence variation in the human genome are still incomplete, preliminary (Shafin et al. 2020), or have only been completed for a single genome (Jain et al. 2018; Miga et al. 2020). Therefore, completing maps of telomeres and providing new tools for such research (Nurk et al. 2020) can provide new insight into telomere biology and enable novel approaches to analyze the effects of aging, environment, and health status (Lee et al. 2018) on telomere sequence and length.

To improve our understanding of telomere sequence structure and variation, we developed *edgeCase*, a scalable framework for alignment and *de novo* telomeric motif discovery from human whole genome long-read sequencing experiments. We have validated these methods using Genome in a Bottle (Zook et al. 2019) single-molecule real-time (SMRT) sequencing datasets generated with Pacific Biosciences circular consensus sequencing (PacBio CCS) (Eid et al. 2009; Ardui et al. 2018), and short-read Illumina (Bentley et al. 2008) and 10x Genomics [Chromium] (www.10xgenomics.com) datasets, as well as with healthy donor peripheral blood mononuclear cells (PBMCs). These results provide evidence for multiple novel, non-canonical telomeric repeats, resolution of multiple chromosome-specific haplotypes with SMRT sequencing, and a new method for long-range characterization of the structure of telomeric sequences.

Results

A telomere-annotated reference genome enables recovery of telomeric reads from human long-read whole genome sequencing datasets

We first constructed an extended reference genome, *hg38ext*, that combines chromosome sequences of the *hg38* reference genome (Schneider et al. 2017; "Initial sequencing and analysis of the human genome"

2001) and human subtelomeric assemblies (Stong et al. 2014), resulting in a reference set annotated with boundaries of subtelomeric and telomeric tracts. The layout of this reference set is available in **Supplemental File S1**, and the set itself can be reproduced with a script available as **Supplemental File S2**. We then aligned to it PacBio CCS reads of seven Genome in a Bottle [GIAB] (Zook et al. 2019) human subjects (HG001 through HG007) from three different ancestries (Ashkenazim, Chinese, and Utah), which included two son/father/mother trios (**Supplemental Table S1**). In total, we observed reads uniquely mapping to the ends of chromosomes and extending into telomeric regions on 9 *p* arms and 14 *q* arms, with 43–285 such reads on the *p* arms and 34–250 on the *q* arms (**Supplemental Table S2**). Portions of reads contained in the telomeric regions were extracted for further analysis (**Figure 1**).

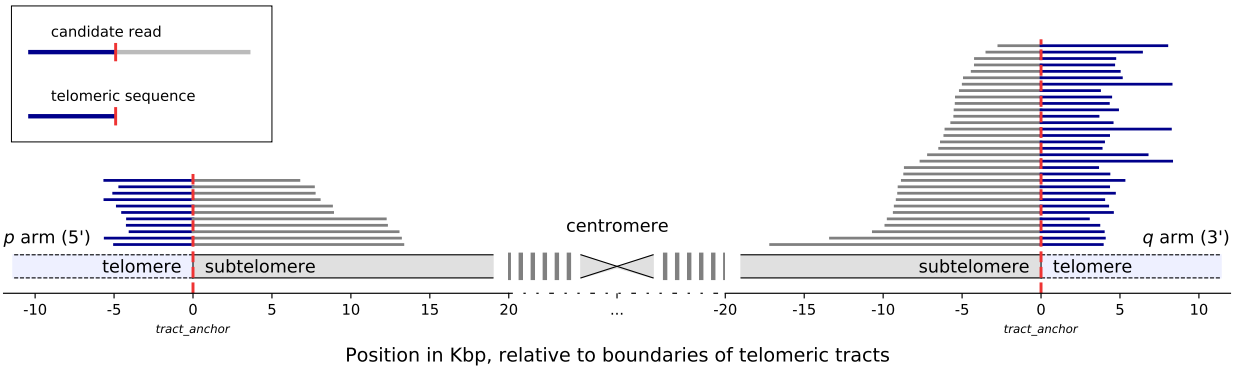

Figure 1: Mapping of candidate telomeric reads, illustrated with reads from the HG002 dataset aligning to Chromosome 12. The chromosome is displayed schematically, centered around the centromere. Vertical red dashed lines denote the position of the boundary of the annotated telomeric tract. Coordinates are given in Kbp, relative to the positions of the telomeric tract boundaries. Statistics for all chromosomes of all seven datasets are provided in **Supplemental Table S2**.

Telomeric long reads contain variations of the canonical motif

We then performed *de novo* repeat discovery in the telomeric sequences for motifs of lengths 4 through 16, and identified motifs in repeat contexts that are statistically enriched in the seven datasets. The majority of motifs were either the canonical TTAGGG / CCCTAA, its variations (e.g., TGAGGG), or a duplet of variants, such as TTAGGGTTAGGGG (**Table 1**). CG-rich motifs were also observed on the *q* arms. The top enriched motif (TTAGGG / CCCTAA) explained 27.0%–76.9% of the telomeric repeat content on the *p* arms and 49.1%–80.1% on the *q* arms, while motifs TGAGGG and TTAGGGG explained up to 8.0% and 6.6% of the repeat content overall, respectively.

We next visualized the locations of the top three enriched motifs and their reverse complements on the chromosomal ends of the HG002 dataset (**Figure 2**), as it provided the deepest coverage among the as-

sessed datasets (**Supplemental Table S2**); plots for the other six datasets are available as **Supplemental Figs. S2** and **S3**. Only the chromosomal arms cumulatively covered by at least 25 reads across datasets were plotted. These data showed that the overwhelming majority of the telomeric regions were represented by the canonical repeats, but also novel, chromosome-specific repeat motif patterns could be observed, and they were enriched for the proximal end of the telomere; these data also illustrated the positions of the repeat-rich portions of the genomes in relation to the known subtelomere-telomere boundaries, including deletions/insertions (4p, 8q) and an apparent extension of the 17p subtelomere (see Discussion).

To discern if the sequence mapping, read length, or overall coverage had any effect on the discovery or enrichment of these motifs, the motif entropies were examined as a function of their location within reads and coverage across the telomere tracks. When the locations of different motifs were examined within any 10 bp window across the length of the long reads, the entropy data showed consistency among reads and across samples (**Figure 3**). Indeed, the coverage-weighted median of normalized Shannon entropy was 0.07 for one dataset and 0.00 for the other six, while most non-zero values were contained only in the top quartile, i.e., between the 75th and the 100th percentile, indicating that locations of the variations are colinear among reads.

Arm	Motif	Percentage of sequence explainable by motif, %							Score							Combined adjusted p value
		HG001	HG002	HG003	HG004	HG005	HG006	HG007	HG001	HG002	HG003	HG004	HG005	HG006	HG007	
p	CCCTAA	76.9	67.9	56.4	60.0	64.4	41.3	27.0	0.6395	0.6149	0.4514	0.4678	0.5387	0.3257	0.1935	9.33e-113
	CCCCTAA	4.1	4.3	4.1	4.6	3.6	3.3	3.8	0.0113	0.0109	0.0096	0.0118	0.0092	0.0087	0.0116	6.41e-92
	CCTAA	2.0	1.6	2.6	3.4	1.4	1.6	1.0	0.0051	0.0043	0.0067	0.0091	0.0036	0.0041	0.0025	8.55e-74
	CCCCTAACCCTAA	3.5	2.4	3.9	4.3	3.2	3.3	2.9	0.0051	0.0030	0.0051	0.0056	0.0043	0.0044	0.0039	5.93e-83
	CCCTA	2.5	1.2	2.2	2.7	2.4	1.6	1.1	0.0053	0.0027	0.0047	0.0058	0.0055	0.0033	0.0023	1.23e-67
	CCCGAA	1.3	0.4	0.3	0.6	0.3	0.3	1.0	0.0085	0.0029	0.0019	0.0045	0.0022	0.0020	0.0068	1.36e-21
	CCCTAACCTAA	3.0	2.7	4.1	5.0	2.3	2.5	1.2	0.0043	0.0036	0.0055	0.0067	0.0031	0.0032	0.0016	2.72e-77
	CCCTACCCTAA	3.4	1.7	2.8	3.1	2.8	2.1	1.2	0.0038	0.0016	0.0028	0.0033	0.0034	0.0021	0.0014	7.27e-65
	CCCTAAA	1.2	0.9	0.8	1.1	0.7	0.7	0.9	0.0034	0.0019	0.0012	0.0024	0.0011	0.0011	0.0018	8.79e-46
	CCCAA	1.5	0.4	1.2	1.3	1.3	0.9	1.0	0.0018	0.0005	0.0018	0.0018	0.0019	0.0013	0.0017	1.37e-43
	CCCA	0.4	0.1	0.4	0.4	0.5	0.3	0.3	0.0006	0.0001	0.0008	0.0007	0.0009	0.0005	0.0006	1.36e-75
	CCCCCTAA	0.2	0.1	0.4	0.4	0.1	0.4	0.5	0.0003	0.0002	0.0005	0.0005	0.0002	0.0008	0.0011	2.08e-67
q	TTAGGG	51.7	79.9	79.6	74.8	75.3	80.1	49.1	0.4217	0.7194	0.6298	0.6008	0.6043	0.6391	0.3471	8.16e-112
	TGAGGG	1.1	2.5	0.7	1.4	4.2	2.5	8.0	0.0066	0.0150	0.0033	0.0081	0.0248	0.0148	0.0476	2.12e-56
	TTAGGGG	4.6	3.2	6.5	5.2	3.5	5.7	6.6	0.0147	0.0099	0.0172	0.0126	0.0100	0.0149	0.0213	8.85e-100
	TTAGG	1.2	1.5	3.8	3.8	1.9	4.0	1.9	0.0031	0.0038	0.0104	0.0101	0.0048	0.0106	0.0051	3.62e-94
	TTAGGGTTAGGGG	3.3	2.5	6.4	5.4	3.0	6.0	4.6	0.0049	0.0037	0.0083	0.0073	0.0039	0.0077	0.0062	1.62e-85
	TTAGGTTAGGG	1.6	2.4	5.8	6.1	2.8	5.9	2.3	0.0022	0.0032	0.0078	0.0081	0.0038	0.0081	0.0029	5.54e-74
	TAGGG	1.9	1.4	2.9	2.3	3.1	3.1	1.8	0.0036	0.0027	0.0062	0.0048	0.0065	0.0065	0.0037	8.14e-86
	TAGGGTTAGGG	2.8	2.0	3.7	3.0	3.8	4.0	2.2	0.0040	0.0019	0.0034	0.0029	0.0048	0.0042	0.0028	1.29e-64
	TTTAGGG	1.0	0.8	1.3	1.0	1.1	1.3	1.3	0.0030	0.0018	0.0026	0.0015	0.0031	0.0026	0.0029	1.12e-49
	TTGGG	1.1	0.5	1.8	1.7	2.5	1.5	1.0	0.0015	0.0008	0.0029	0.0025	0.0040	0.0021	0.0016	4.72e-69
	GCGGC	0.5	0.7	0.9	0.8	0.9	0.9	0.6	0.0012	0.0018	0.0022	0.0020	0.0022	0.0021	0.0015	6.16e-64
	TTGGGTTAGGG	1.4	0.5	2.0	2.0	2.6	1.8	0.8	0.0011	0.0003	0.0013	0.0018	0.0022	0.0013	0.0007	9.24e-30
	TTAGGGTTTAGGG	0.7	0.9	1.6	1.4	1.1	1.4	1.3	0.0006	0.0008	0.0016	0.0013	0.0009	0.0011	0.0011	2.26e-35
	TGGG	0.3	0.1	0.5	0.7	1.1	0.5	0.4	0.0004	0.0002	0.0009	0.0013	0.0024	0.0008	0.0007	9.59e-75

Table 1: Significantly enriched repeating motifs in telomeric regions of GIAB datasets HG001 through HG007. See Materials and Methods for the definition of *score*.

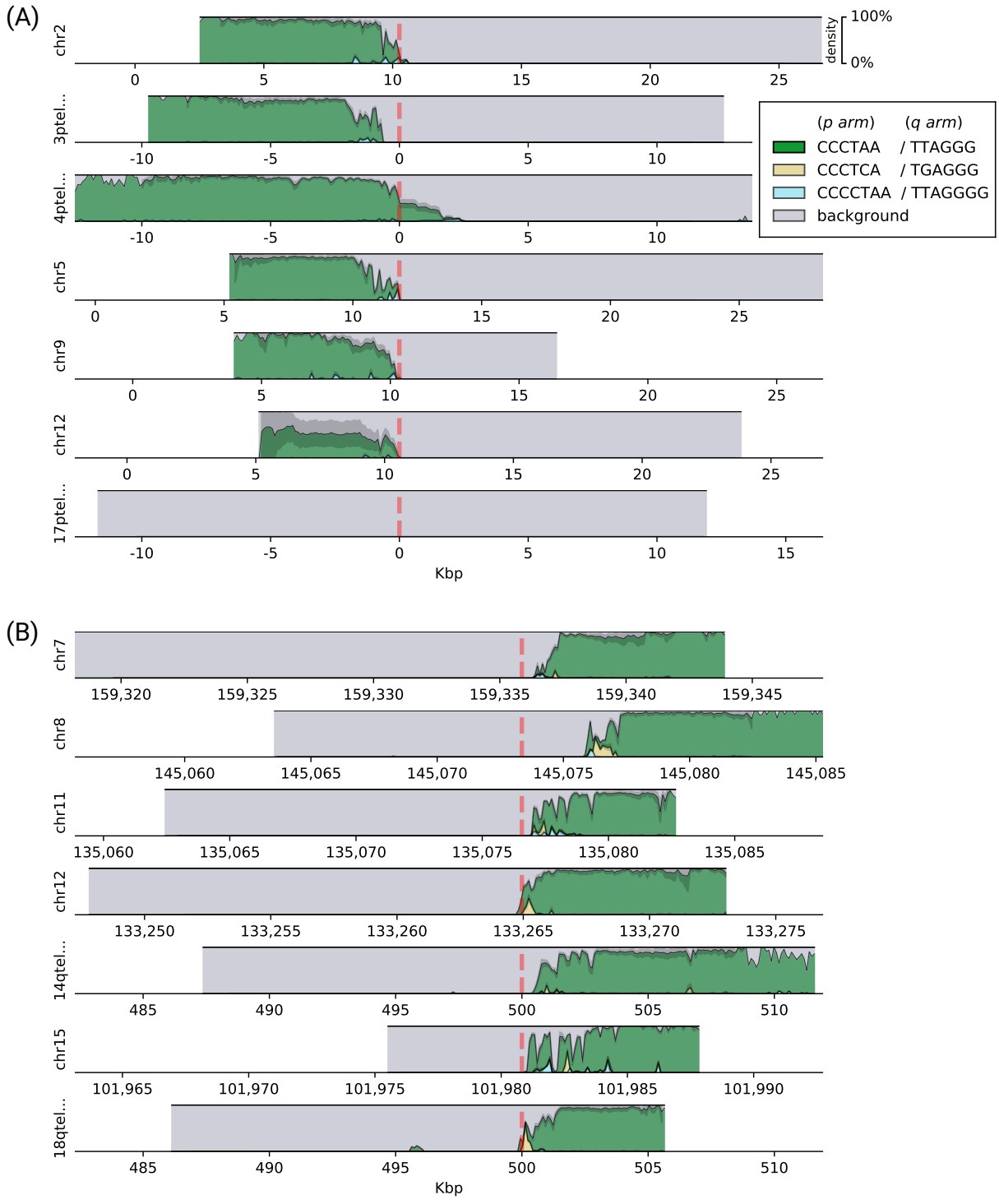

Figure 2: Densities of the top three enriched motifs at ends of chromosomal **(A)** *p* arms and **(B)** *q* arms of the HG002 dataset. *Background* represents the remaining sequence content (non-repeating sequence and not significantly enriched motifs). Reads are shown aligned to the contigs in the *hg38ext* reference set, and genomic coordinates are given in Kbp. Vertical red dashed lines denote the position of the boundary of the annotated telomeric tract.

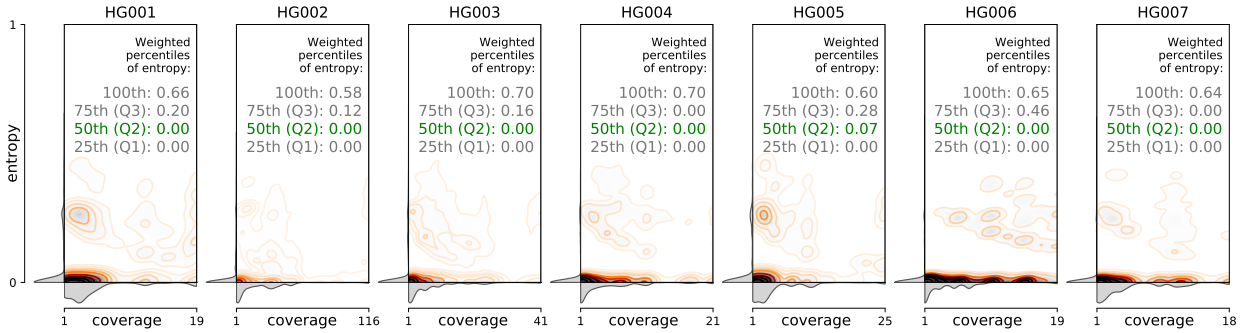

Figure 3: Distribution of motif entropies in 10 bp windows of of candidate PacBio CCS reads aligning to the same chromosomal arms in GIAB datasets HG001 through HG007, with respect to per-window coverage, and the coverage-weighted percentiles of the entropy values.

Short-read sequencing validates motif variations observed in long reads

We next validated these findings using short-read sequencing in two ways. First, we extracted telomeric candidate reads with *Telomerecat* (Farmery et al. 2018) from respective GIAB Illumina datasets, and found that they supported a definitive majority of the long-read telomeric candidates, with a median 89% of the *p* arm sequence and a median 95% of the *q* arm sequence supported (**Supplemental Fig. S1**). Second, we confirmed 13 of the enriched motifs in independently generated human short-read and linked-read genomic datasets from donated PBMCs, with the same three motifs being the most enriched (**Supplemental Table S3**).

Long-read sequencing uncovers a variety of human telomeric haplotypes

While reads agreed on colinearity of motifs, evidenced by low entropy, rarer non-zero entropy values could be attributable both to sequencing errors and to structural variations within the same subject's dataset. To investigate the latter possibility, we clustered reads on each arm of each subject by relative pairwise Levenshtein distances (Levenshtein 1966) and found that hierarchical clustering described read similarity well, resulting in high cophenetic correlation between the dendrograms and the pairwise distance matrices (**Table 2**), and in visible structure (**Figure 4**, **Figure 5**).

In this complex clustering, subject- and population-specific variation was evident and quantifiable via relative Levenshtein distances (**Table 3**; see Materials and Methods): overall, telomeric reads within a subject were more similar than within a population (adjusted Wilcoxon signed-rank test p = 4.2e-56), and telomeric reads within a population were more similar than between populations (p = 2.2e-40).

Importantly, however, this was true for most, but not all reads; 13.8% of all assessed reads (165 out of 1,192)

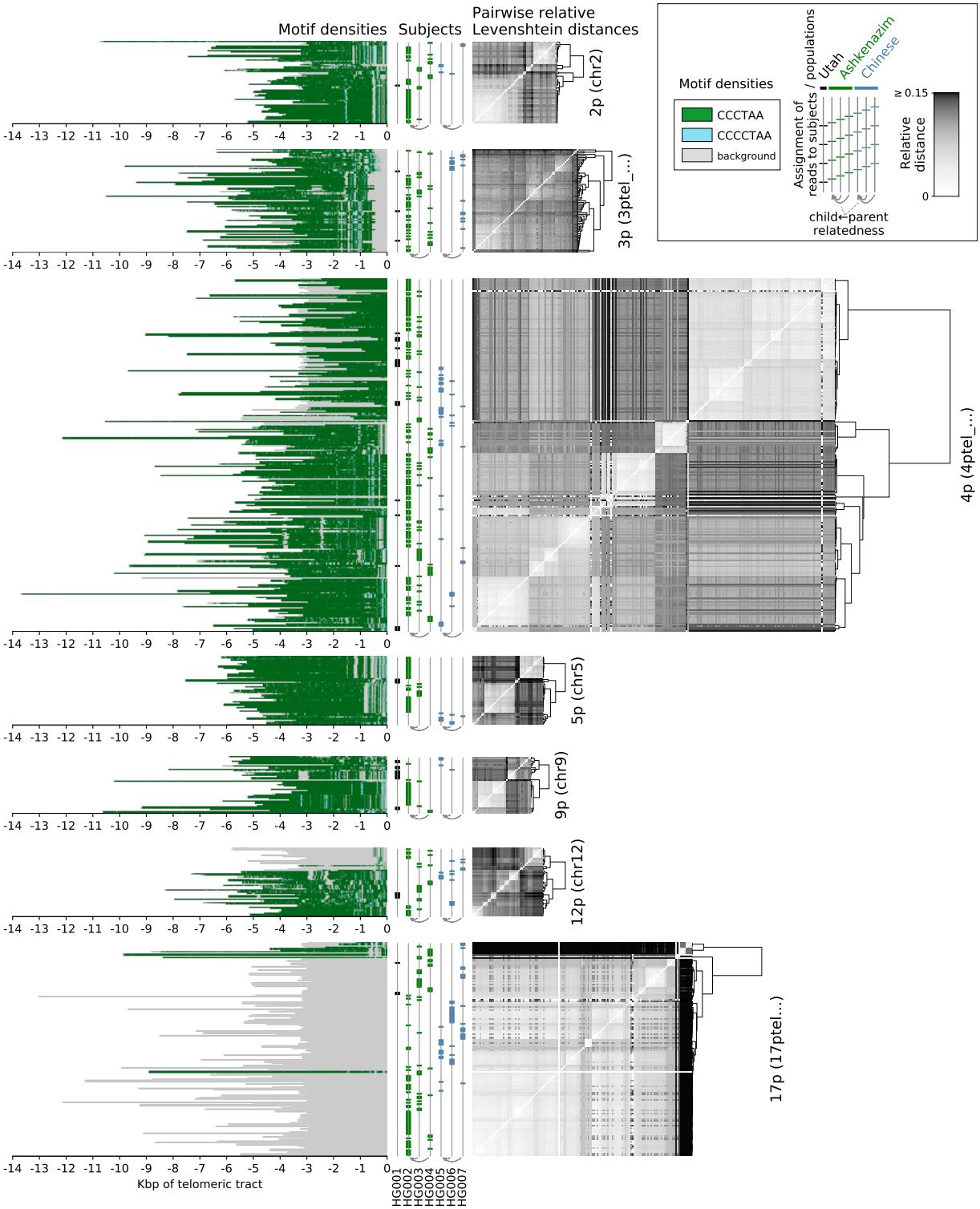

Figure 4: Clustering of reads by relative pairwise Levenshtein distances (unitless measure) on each chromosomal *p* arm of datasets HG001 through HG007, and densities of the top enriched motifs along each read. Each horizontal line represents an individual read; genomic coordinates are given in Kbp, relative to the positions of the telomeric tract boundaries. Only the chromosomal arms cumulatively covered by at least 25 reads are displayed.

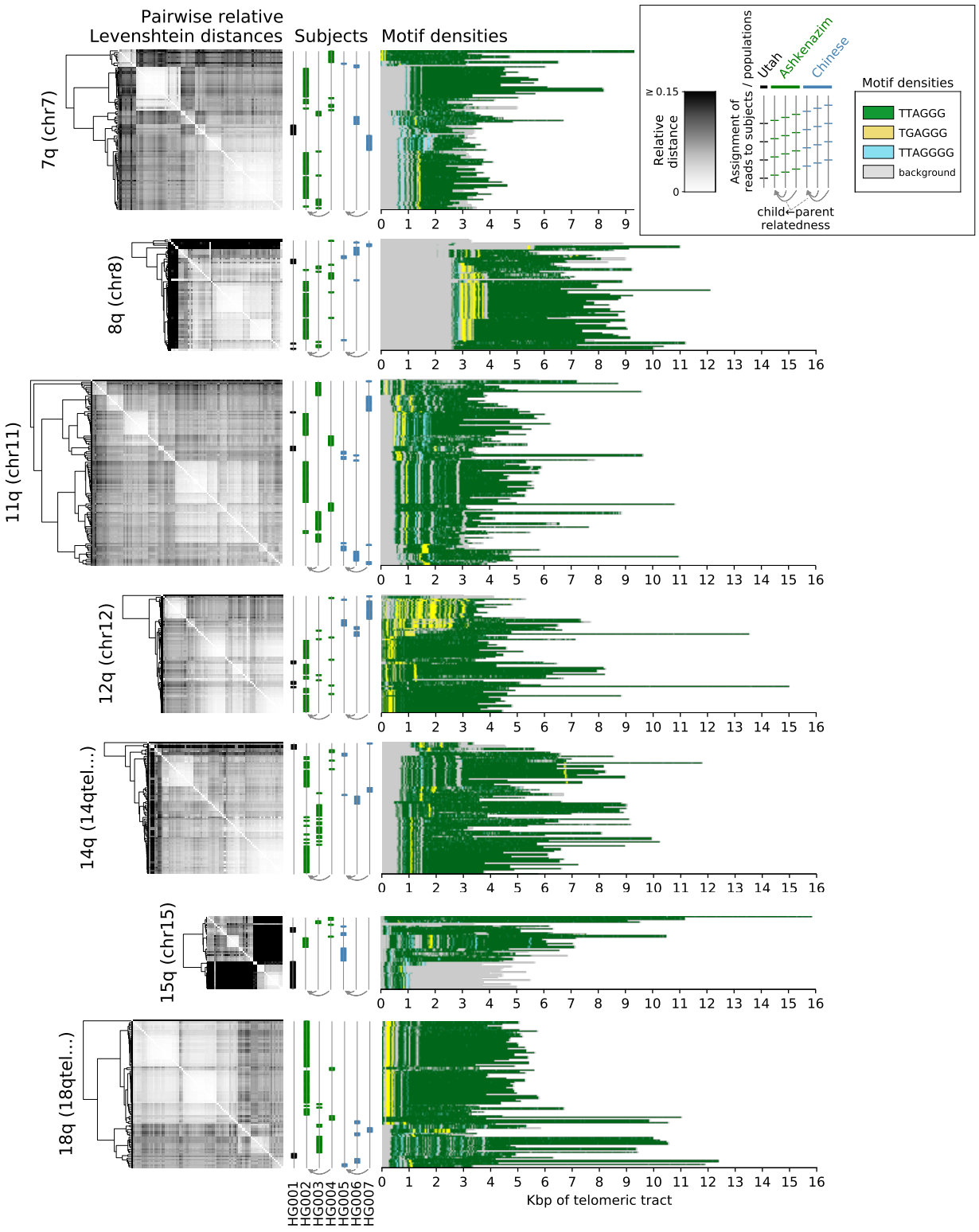

Figure 5: Clustering of reads by relative pairwise Levenshtein distances (unitless measure) on each chromosomal *q* arm of datasets HG001 through HG007, and densities of the top enriched motifs along each read. Each horizontal line represents an individual read; genomic coordinates are given in Kbp, relative to the positions of the telomeric tract boundaries. Only the chromosomal arms cumulatively covered by at least 25 reads are displayed.

Telomere	Reference contig	Cophenetic correlation	
		r	p
2p	chr2	0.631	6.8e-165
3p	3ptel_1-500K_1_12_12	0.607	1.4e-235
4p	4ptel_1-500K_1_12_12	0.490	<1.0e-300
5p	chr5	0.760	2.4e-194
9p	chr9	0.734	7.3e-119
12p	chr12	0.783	2.5e-214
17p	17ptel_1_500K_1_12_12	0.937	<1.0e-300
7q	chr7	0.838	<1.0e-300
8q	chr8	0.928	<1.0e-300
11q	chr11	0.630	<1.0e-300
12q	chr12	0.881	<1.0e-300
14q	14qtel_1-500K_1_12_12_rc	0.842	<1.0e-300
15q	chr15	0.915	<1.0e-300
18q	18qtel_1-500K_1_12_12_rc	0.682	<1.0e-300

Table 2: Measures of cophenetic correlation (Pearson's *r* and adjusted *p*-value) between the hierarchical clustering and the pairwise distance matrix on each chromosomal arm.

Comparison	Adjusted p-value
A subject's reads are closer to each other than to other subjects' reads in the trio	4.2e-56
A subject's reads are closer to each other than to subjects' reads in other populations	7.6e-107
Reads within a population are closer to each other than to reads in other populations	2.2e-40
Ashkenazim trio:	
Father's reads are closer to son's reads than to mother's reads	3.1e-11
Mother's reads are closer to son's reads than to father's reads	ns (1.00)
Chinese trio:	
Father's reads are closer to son's reads than to mother's reads	3.4e-02
Mother's reads are closer to son's reads than to father's reads	ns (0.23)

Table 3: Adjusted *p*-values of the Wilcoxon signed-rank tests on relative Levenshtein distances. For each read among all telomeric reads on each arm, closest distances to groups of reads described in the *Comparison* column are compared (see Materials and Methods).

contributed to interpopulation similarity; these reads were twice as close to reads from a different population than they were to any reads of their own subjects. This trend is observable on **Figure 4** and **Figure 5**, with subjects' and populations' reads interspersed across multiple clusters. Therefore, the captured reads reflected spectra of haplotypes, generally describing subject- and population-specific similarities, but including a sizable component that described interpopulation similarity. Paternal inheritance of variation was also observed: each father's telomeric reads were more similar to their son's than to the mother's reads in both the Ashkenazim and the Chinese trios.

Discussion

Repeat-rich, low-complexity regions of the human genome such as telomeres have been historically recalcitrant to full mapping and annotation (Miga 2015), mainly due to the alignment challenge they pose and to the read lengths required to span such areas (Treangen and Salzberg 2011). 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 (Luxton et al. 2020) of telomeres. Our results reaffirm that the canonical repeat (TTAGGG) is certainly the most dominant motif found within telomeres, but also reveal a surprising diversity of repeat variations, which are confirmed by both short and long-read sequencing technologies. This diversity of repeat sequence 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. Interestingly, repeat patterns were chromosome-specific, with different non-canonical repeats being pronounced on different chromosomes, such as TGAGGG on 12q and TTAGGGG on 15q, which may be correlated with particular biological pathways (Bluhm et al. 2019). Apart from these variations, CG-rich motifs were identified in telomeric regions of *q* arms, consistent with previously reported findings (Nergadze et al. 2009). Moreover, while short read sequencing is capable of identifying 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. Furthermore, in contrast to previously published research that utilized targeted sequencing (Allshire et al. 1989; Coleman et al. 1999; Lee et al. 2018; Bluhm et al. 2019), the method described here allows identification of multiple enriched motifs and their localization *de novo*, without any bias introduced by prior knowledge about the sequence of target motifs. These results also highlight the need of better subtelomeric and telomeric annotations in the human genome: the canonical motif was present on the *q* arm of Chromosome 8 only 2–3Kbp beyond the annotated boundary in all datasets; the candidate reads on the *p* arm of Chromosome 17 represented TTAGGG-rich and non-TTAGGG-rich haplotypes, indicating that in multiple subjects and ancestries there exists an extension of the 17p subtelomere. Strikingly, for example, the Ashkenazim son (HG002) provided only non-TTAGGG-rich 17p reads, while both the father (HG003) and the mother (HG004) had a mixture of apparently telomeric and non-telomeric 17p reads. This supports previous findings (Young et al. 2020) that the existing assemblies do not provide completely accurate subtelomeric annotation, and suggests 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 fidelity, as 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 so, long reads appear to be more suitable for this purpose as well.

The identified variations in long range contexts elucidate subject-specific, trio- and population-specific similarities of telomeric sequences, as well as a level of interpopulation similarity, and thus provide a new means of haplotype mapping and reveal the existence and motif composition of haplotype spectra on a multi-Kbp scale. Interpopulation similarity, as well as paternal inheritance of variation, provided evidence that the observed haplotypes could not be attributed to per-dataset batch effects. The lengths of PacBio CCS reads allowed resolution of uniquely mapping reads only on 23 chromosomal arms, and coverage of different arms was uneven. As such, numbers of captured telomeric reads and levels of observed similarity varied from subject to subject; in particular, maternal inheritance of haplotypes could not be determined, in contrast to statistically significant paternal inheritance. This calls for more sequencing experiments aimed to reconstruct the full picture of this variation. Clustering on a per-subject basis concealed interpopulation similarity, but underscored intra-subject variation (**Supplemental Figs. S4** and **S5**), suggesting coexistence of two or more telomeric haplotypes per chromosomal arm within each subject. Given that the reference DNA for the subjects HG001 through HG007 was extracted from growths of B lymphoblastoid cell lines, this suggests that as B cells undergo maturation, distinct clones may gain distinct variations in their telomeric sequence. This opens up avenues of investigation into the haplotypic variation among not only immune cells, but also different cell types overall, and provides a new opportunity to map, quantify, and characterize a previously unrecognized form of human genetic variation.

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 (Schneider et al. 2017; "Initial sequencing and analysis of the human genome" 2001) and the subtelomeric assemblies (Stong et al. 2014) with *minimap2* (Li 2018) 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 Chromosome 14 (chr14_KI270846v1_alt) with the quality of 52, which, in turn, mapped to the main contig of Chromosome 14 (chr14) 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. For the purposes of capturing candidate reads that uniquely align to subtelomere-telomere boundaries, subtelomeric contigs which were not previously assembled as extending completely up to the start of the telomere, and/or were not precisely localized in relation to the reference genome, such as 1p, 6p, 7p, 8p, 11p, 20p, 3q, 4q, 20q, and Xq (Stong et al. 2014; Young et al. 2020), were masked prior to downstream analyses.

Detection of telomeric sequences in long-read datasets

Seven subjects were selected for the analysis. The first individual (NA12878/HG001) came from the pilot genome of the HapMap project ("The International HapMap Project" 2003), while the other six, including the Ashkenazim Jewish Trio (son: NA24385/HG002, father: NA24149/HG003, mother: NA24143/HG004) and the Chinese Trio
(son: NA24631/HG005, father: NA24694/HG006, mother: NA24695/HG007), are members of the Personal Genome Project, whose genomes are consented for commercial redistribution and reidentification (Zook et al. 2016). These subjects are referred to throughout as HG001 through HG007, respectively.

Multiple Genome in a Bottle (Zook et al. 2019) PacBio CCS (Eid et al. 2009; Ardui et al. 2018) datasets were available and combined per each subject, with mean coverages of individual datasets ranging from ∼21x to ∼69x (**Supplemental Table S1**). We mapped these reads to *hg38ext* with *minimap2*, allowing secondary mappings, and selected reads that mapped to either end of either chromosome, having an at least 500 bp portion of their sequence mapped to the reference contig and a portion extending beyond the reference (soft- or hard-clipped in the alignment file). As each of such reads can map to multiple subtelomeres due to paralogy, we considered such multiple mappings and only retained the reads that mapped to a unique subtelomere; furthermore, out of these candidates, we only selected the ones overlapping the subtelomere and the telomere by at least 3Kbp. 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; **Figure 1**).

Evaluation of telomeric content in short- and linked-read datasets

To evaluate the concordance of telomeric reads captured by long- and short-read technologies, we extracted candidate telomeric reads from GIAB Illumina datasets for each subject (**Supplemental Table S1**) with *Telomerecat* (Farmery et al. 2018), and mapped the short reads back onto the candidate long reads from the same subject's dataset with *minimap2*, again allowing all secondary mappings. Then, we calculated the fractions of each long read that were supported by the short reads that aligned to them.

To evaluate sequence motifs in independent samples collected from human subjects (as opposed to reference cell lines), we analyzed four whole-genome Illumina datasets (mean coverage ∼104x) and three linked-read 10x datasets (mean coverage ∼28x) for one individual at different timepoints, and one additional linked-read 10x dataset (coverage ∼47x) for another individual. These data were originally obtained from astronaut subjects for an unrelated space biology experiment, and the blood samples were collected from the subjects as described in the study (Garrett-Bakelman et al. 2019). 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 (Farmery et al. 2018).

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* (Marçais and Kingsford 2011), 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 (George and Mudholkar 1983) 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. As even doubled sequences (such as ACGTACGT for motif ACGT) can partially overlap at the boundaries of repeat contexts, we quantified their presence in the telomeric reads in two distinct ways. Consider a sequence such as TTAGGG(TTAGTTAG)GGTTA: the inner (TTAG)x2 repeat can be explained by the repeats of the canonical motif extending into it from either side; the middle part of a similar sequence with a bigger number of the repeats of the 4-mer, TTAGGGTTAG(TTAGTTAG)TTAGGGTTA, can only be explained by the repeats of said 4-mer. On the one hand, the maximum fraction of the sequence that can be explained by any one motif is a useful metric, and it was calculated and reported. On the other hand, the fraction of the *k*-mers attributable to a specific motif – and not to any others – elucidates the extent of deviation from the background repeat context, and identifies motifs that most affect the sequence structure; it was calculated as well and reported as each motif's score. 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 (Minosse et al. 2006). The value of normalized entropy was a metric bounded by $[0, 1]$, with 0 describing perfect agreement and 1 describing maximum randomness. As coverage of windows drops off towards the distal end of the alignment, lower covered windows have less chance to produce entropy; we calculated percentiles of entropy as weighted by coverage minus one (thus prioritizing higher covered windows, and removing windows with the coverage of one and no entropy from the calculation). For motif visualization, we performed 1000 rounds of bootstrap of the calculated density values, this time in 100 bp rolling windows to accommodate the scale of multi-Kbp plots, and selected the lower and the upper bounds of the 95% confidence interval of bootstrap.

Identification of telomeric haplotypic variation

Within groups of reads mapping to each chromosomal arm, all relative pairwise Levenshtein distances were calculated. In short, Levenshtein distance is a string metric defined as the edit distance between two strings (sequences), equal to the minimum number of single-character insertions, deletions, and substitutions required to make these sequences identical (Levenshtein 1966). For each pair of reads, this metric was calculated and represented absolute edit distance; the relative distance was then computed as the absolute distance divided by the length of the overlap, to normalize for the variation of such lengths. Pairwise relative distances were then clustered using Ward's method via the Euclidean metric, resulting in hierarchical structure describing the extents of similarity among reads. To quantify how accurately hierarchical clustering described this similarity, cophenetic distances (Sokal and Rohlf 1962) between the hierarchies (dendrograms) and the distance matrices was calculated, and their Pearson correlation coefficients and Bonferroni-corrected *p*-values were reported.

We then traversed the distance matrices, and for each read, tracked the closest reads by category: closest reads from the same subject, from the same trio (population), and from the outgroup (other populations). For the Ashkenazim and the Chinese trios, we also tracked the closest reads between the parents and between each parent and the child. Thus, for each read, we determined whether it locally clustered within its own category (for example, with other reads of the same subject, or with other reads from the same population) or in a different one (for example, with other reads of a different population), and the value of the distances that drove either clustering. Performing the Wilcoxon signed-rank test on these values between either categories provided us with *p*-values that, after a Bonferroni correction, described whether reads tended to cluster in their own category or in a different one. Additionally, we also identified the minority of reads that did not follow the overall trend, and quantified the extent to which they did so (such as the reads that contributed to interpopulation similarity).

Data access

Healthy donor DNA came from a previous study [The NASA Twins Study] (Garrett-Bakelman et al. 2019). 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, as well as **Supplemental File S3**.

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

Allshire RC, Dempster M, Hastie ND. 1989. Human telomeres contain at least three types of G–rich repeat distributed non-randomly. *Nucl Acids Res* **17**: 4611–4627.

Ardui S, Ameur A, Vermeesch JR, Hestand MS. 2018. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research* **46**: 2159–2168.
Aubert G, Lansdorp PM. 2008. Telomeres and Aging. *Physiological Reviews* **88**: 557–579.
Bentley DR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53–59.
Bluhm A, Viceconte N, Li F, Rane G, Ritz S, Wang S, Levin M, Shi Y, Kappei D, Butter F. 2019. ZBTB10 binds the telomeric variant repeat TTGGGG and interacts with TRF2. *Nucleic Acids Research* **47**: 1896–1907.
Coleman J, Baird DM, Royle NJ. 1999. 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**: 1637–1646.
Eid J et al. 2009. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science* **323**: 133–138.
Farmery JHR, Smith ML, Lynch AG. 2018. Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Sci Rep* **8**: 1–17.
Garrett-Bakelman FE, Darshi M, Green SJ, Gur RC, Lin L, Macias BR, McKenna MJ, Meydan C, Mishra T, Nasrini J, et al. 2019. The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight. *Science* **364**: 144.
George EO, Mudholkar GS. 1983. On the convolution of logistic random variables. *Metrika* **30**: 1–13.
2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
Jain M et al. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**: 338–345.
Lee M et al. 2018. Telomere sequence content can be used to determine ALT activity in tumours. *Nucleic Acids Research* **46**: 4903–4918.
Levenshtein VI 1966. Binary codes capable of correcting deletions, insertions, and reversals. In: *Soviet physics doklady*. Vol. 10. 8, pp. 707–710.
Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.
Luxton JJ et al. 2020. Temporal Telomere and DNA Damage Responses in the Space Radiation Environment. *Cell Reports* **33**: 108435.
Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**: 764–770.
Miga KH. 2015. Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Res* **23**: 421–426.
Miga KH et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. *Nature* **585**: 79–84.
Minosse C, Calcaterra S, Abbate I, Selleri M, Zaniratti MS, Capobianchi MR. 2006. Possible Compartmentalization of Hepatitis C Viral Replication in the Genital Tract of HIV-1–Coinfected Women. *The Journal of Infectious Diseases* **194**: 1529–1536.
Moyzis RK, Buckingham JM, Cram LS, Dani M, Deaven LL, Jones MD, Meyne J, Ratliff RL, Wu JR. 1988. A highly conserved repetitive DNA sequence, (TTAGGG)n, present at the telomeres of human chromosomes. *Proceedings of the National Academy of Sciences* **85**: 6622–6626.
Nergadze SG, Farnung BO, Wischnewski H, Khoriauli L, Vitelli V, Chawla R, Giulotto E, Azzalin CM. 2009. CpG-island promoters drive transcription of human telomeres. *RNA* **15**: 2186–2194.
Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. 2020. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Research* **30**: 1291–1305.
Schneider VA et al. 2017. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**: 849–864.
Shafin K et al. 2020. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature Biotechnology* **38**: 1044–1053.
Shammas MA. 2011. Telomeres, lifestyle, cancer, and aging. *Current Opinion in Clinical Nutrition and Metabolic Care* **14**: 28–34.
Sokal RR, Rohlf FJ. 1962. The comparison of dendrograms by objective methods. *Taxon* **11**: 33–40.
Stong N et al. 2014. Subtelomeric CTCF and cohesin binding site organization using improved subtelomere assemblies and a novel annotation pipeline. *Genome Research* **24**: 1039–1050.
2003. The International HapMap Project. *Nature* **426**: 789–796.

Treangen TJ, Salzberg SL. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**: 36–46.
Young E, Abid HZ, Kwok PY, Riethman H, Xiao M. 2020. Comprehensive Analysis of Human Subtelomeres by Whole Genome Mapping. *PLOS Genetics* **16**: e1008347.
Zook JM et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* **3**: 1–26.
Zook JM et al. 2019. An open resource for accurately benchmarking small variant and reference calls. *Nat Biotechnol* **37**: 561–566.