Dear editors of Genome Research, dear reviewers,

Thank you very much for reviewing our manuscript and providing thorough and constructive comments, criticism, and for the careful assessment of the submission. We have carefully evaluated and addressed all suggestions and comments of the reviewers, which led to a significant revision of the paper, inclusion of new data, and additional novel results, and hope that our responses and additions to the paper are satisfactory.

One of the important concerns of the reviewers was the fact that the original submission only used data from three Genome in a Bottle (GIAB) datasets; at the moment of completing the first version of the paper, those were the only ones with PacBio CCS data available. Since then, such data for all seven GIAB subjects became available, and we rectified this by analyzing all these data and including them in this version. We would like to mention that this allowed us to come to additional, novel findings which, in our opinion, make the paper stronger. Of note, we discover spectra of telomeric haplotypes that exhibit both intra- and inter-subject and intra- and inter-population similarities, which supersedes the previously limited investigation of intra-subject diplotypes and presents a new result of a larger scale. We include these findings in the Results, Methods, and Discussion sections, and add a mention of them to the Abstract.

Please find below our point-by-point replies to the reviewers’ comments.

**Reviewer #1**:

A: We would like to thank the referee for the careful assessment of our manuscript and the constructive criticism.

Q: *My main criticism of this paper is that, while the specific methodologies and approaches presented in this paper haven’t yet been applied to the problem of variant telomere repeat identification and haplotype determination, several others have and have provided biologically relevant information equivalent to or better than the current study. For example, Baird et al. (1995 & 2000, plus several STELA papers) provided detailed information on variant telomere repeat organization and allele-specific telomere haplotypes for multiple telomeres using alternative methods, and multiple groups have used alternative methods including molecular combing and DNA sequence read analyses to establish the existence of and novel roles for variant telomere repeats (e.g., in the ALT mechanism of telomere maintenance).*

A: Thank you for raising these important concerns. However, we would like to stress that the methods that we present in this manuscript have significant advantages over the methods they are being compared to. Baird et al., 1995 & 2000, publish telomere diplotype sequences for just two chromosomes (12q and Xp/Yp) using expensive and highly targeted Sanger sequencing (Thermo Fisher BigDye), in contrast to the method presented here, which is scalable and applicable to conventional third-generation WGS datasets. Importantly, Baird et al. were also limited to investigating previously known motifs – TTAGGG, TTGGGG, TCAGGG, – while *edgeCase*, the method we present, allows for *de novo* identification of motifs, and, without such priors, not only confirms these previously known variations, but uncovers a set of other variations; moreover, it does so at a scale an order of magnitude larger than Baird et al., 1995 & 2000.  
Lee et al., 2018, describe a similarly scarce set of non-canonical motifs, and are limited by a fundamental bias common to all short-read methods, where candidate reads are selected by the content of the canonical motif.

Finally, STELA, while being an important method for measuring telomere lengths, does not investigate motif composition or repeat organization; on the flip side, the research that we present is focused specifically on sequence composition and not on telomere lengths.

Q: *So while this specific method holds promise for a more general approach to analyzing haplotypes in 10 kb-sized subtelomere-adjacent regions, in my view it is incumbent on the authors to generate or analyze enough high-quality data to show that it does a better job than these other methods. With a maximum read length of 10-15 kb, this method still suffers from longstanding limitations in getting definitive mappings to anchor telomeric haplotypes for a substantial number of chromosome ends (including most subtelomeres with more extensive subtelomeric segmental duplications and structural variations), a limitation likely to be overcome only by much longer accurate reads capable of spanning these segmental duplication regions.*

A: Thank you for drawing attention to this valid point. We agree that the length limitation of third-generation sequencing may not currently allow complete resolution of all telomeric sequences; however, *edgeCase* already resolves haplotypes at multi-Kbp lengths, capturing arguably the most pertinent regions of the telomere (as the canonical motif has been shown to become overwhelmingly prevalent towards the distal regions of the telomere, Kreiter et al., 1995), and can be readily applied to any conventional HiFi (CCS) sequencing experiments, eliminating the need for custom sample preparation and targeted sequencing, so if/when longer and more accurate data is generated in the future, any researcher will be able to immediately apply to it the framework that we make available here.

* Kreiter, M., Irion, V., Ward, J. and Morin, G., 1995. The fidelity of human telomerase. In *Nucleic acids symposium series* (No. 33, pp. 137-139).

Q: *In sum, while this is a potentially useful tool, I am not convinced it will generate new and biologically useful data beyond what is already known given the current limitations of the datasets used and available.*

A: We would like to underscore that our findings provide a method to assess enrichment of motifs in the data without prior knowledge of motifs, and is not only able to confirm previous findings, but already uncovers a variety of non-canonical motifs that have not been previously reported. In addition, in this revision of the paper, we include more subjects (for a total of seven, i.e., all Genome in a Bottle reference individuals), and observe a spectrum of haplotypes that exhibits both intra- and inter-subject and interpopulation similarities, which is a novel finding.

**Reviewer #2**:

A: We would like to thank the referee for the careful assessment of our manuscript and drawing attention to multiple pertinent questions.

Q: *In the manuscript, “Haplotype Diversity and Sequence Heterogeneity of Human Telomeres”, the introduction does not talk at all about why identifying telomere sequence, haplotype and structural variants would be useful.*

A: Thank you for highlighting this issue. Improved knowledge of the telomeric sequence and variation can aid in undestanding of its implications in aging and disease; telomere sequence content has been correlated with biological pathways, such as, for instance, ALT activity in tumors (Lee et al., 2018). We addressed it in the Introduction section – *“[...] long-range maps of telomeric sequence […] can provide insight into telomere biology and enable novel approaches to analyze the effects of aging, environment, and health status on telomere sequence [...]”.* We would also like to note that previously published efforts did not present a framework for localizing the array of non-canonical motif variants, and the approach that we describe, *edgeCase*, can provide further insights by doing so; moreover, *edgeCase* serves as a tool that can be readily applied to a wealth of clinical samples and can aid in investigating the effects of telomere sequence on disease phenotypes.

* Lee, M., Teber, E.T., Holmes, O., Nones, K., Patch, A.M., Dagg, R.A., Lau, L.M.S., Lee, J.H., Napier, C.E., Arthur, J.W. and Grimmond, S.M., 2018. Telomere sequence content can be used to determine ALT activity in tumours. *Nucleic acids research*, 46(10), pp.4903-4918.

Q: *Also seems like a comparison to Taub et al., 2019, is warranted.*

A: While Taub et al., 2019 present undoubtedly interesting and novel results, their research is aimed at established genetic determinants of telomere length, rather than telomere sequence composition. Thus, we thank the reviewer for noting this research to us, but also believe this to be the only comment not relevant to the content of this paper.

* Taub, M.A., Weinstock, J.S., Iyer, K.R., Yanek, L.R., Conomos, M.P., Brody, J.A., Keramati, A., Laurie, C.A., Arvanitis, M., Smith, A.V. and Lane, J., 2019. Novel genetic determinants of telomere length from a multi-ethnic analysis of 75,000 whole genome sequences in TOPMed. *bioRxiv*, p.749010.

Q: *I am confused about the relationship of the sequences used and why these were chosen from the rest of the data in GIAB; it would be more biologically interesting if the relationships were described in the introduction, not just the methods, and any identified variants compared to unrelated sequences as well.*

*a) The subjects are talked about in methods, but please comment on them in the main text; what would you expect the comparison of their haplotypes to be like? It would be helpful to describe those subjects earlier in the text, as they are referred to consistently.*

*b) Why didn’t you choose some related individuals as well to assess the biological implications of telomeric haplotypes? There are two son/father/mother trios of Ashkenazi Jewish and Han Chinese ancestry genomes available on GIAB.*

A: Thank you for raising this extremely fitting concern. During the development of our framework, PacBio CCS data for only three GIAB subjects had been published, but in this revision, we were able to include all seven subjects, as the relevant data was already available, thus including both son/father/mother trios. This made it possible to investigate not only the parental relationships in the data, but inter- and intra- population variation as well. We analyzed all these data in an integrated fashion, included all the results and the additional methods we used, and presented evidence for haplotypic variation that describes such relationships. We also made sure to describe these seven subjects earlier in the text, and include a table outlining their relationships and the data that was used (Supplemental Table S1).

Q: *c) Further, why were the ISS samples used for the illumina short-read data validation? It would seem that other samples - even the illumina short read trio samples - would be more relevant here - as no specific conclusions were drawn from the ISS samples in this manuscript.*

A: We agree that the narrative in the original submission made it a little confusing that the validation was actually performed with short reads both from the trio samples and, separately, from unrelated subjects (ISS). We restructured the narrative to make this point clearer. GIAB Illumina data was used first, as it was the most relevant for the seven subjects: we both compared the sequence data as-is between the two technologies within each subject (section “Evaluation of telomeric content in short- and linked-read datasets” and Supplemental Fig. S7). The data from unrelated subjects was used, specifically, to compare the findings between cell lines (GIAB) and human samples (Supplemental Table S3), and the samples that we used were available to us from a previous unrelated study; we made sure to reword this in the Methods section to make it more transparent that the provenance of the samples was unrelated to the specific focus of this paper.

Q: *The authors generated the “extended reference genome” using hg38 all to all alignments with minimap. First - I argue this should be in the main results section - as initially how the telomeric tract boundary locations were generated was unclear. This “extended reference” should be made publicly available for assessment and reproducibility. Finally, it is not at all clear that using hg38 to determine the subtelomeric regions of HG002 is the correct method - I would have instead suggested using rich data available on this sample to reassemble the subtelomeric region specific to this individual.*

A: Thank you for this suggestion. We added the extended reference genome to the Results section, and included two supplemental files (Supplemental File S1 and S2), which provide both the extended reference and its annotation. We agree that alternative references, arising from other research such as *de novo* reassembly,can be potentially used for this purpose. In this paper, we present a framework that can be used with any such reference; however, the subtelomeric assemblies mapped to the *hg38* reference genome exceptionally well (see Materials and Methods), which enabled us to use this combination to select telomeric candidates.

Q: *Figure 1: The schematics of the pacbio reads into the telomere are very helpful. Why don’t they also show the illumina reads and where they overlap? How did they determine the overlap into the telomere with the short reads?*

A: We addressed this helpful criticism by including Supplemental Fig. S7, which shows the mapping of short reads back onto the long reads. We would like to mention here, however, that due to the multi-Kbp scale, the figure has limited usefulness, as small interruptions in coverage may not be visible.

Q: *Table 1 - enriched compared to . . . . the rest of hg38? Or compared to random? More specifics are needed. And what is going on with the p.value=0? What test was done here?   
a) The authors state “The top enriched motif (TTAGGG / CCCTAA) explained 43.3%-54.4% of the telomeric repeat content...” how are they deining telomeric repeat content? Is it just the section of reads past the demarcated end?*

A: We expanded the Materials and Methods section to provide more detail. The p-values result from a Fisher’s exact test, where the potential enriched motifs are compared against the counts of background motifs in the data. The p-values that equaled to zero were due to the behavior of Python, where extremely low p-values are rounded down. We made sure to identify such situations and verified that these p-values are indeed close to zero and in these cases are smaller than 1e-300; instead of reporting zeros, we now report this directly (<1e-300).

We also included the description of how the telomeric repeat content was calculated to address the concern about the definition of it.

Q: *Figure 2:*

*a) Is the figure implying that ~100% of the stretch is TTAGGG repeats? That doesn’t make sense combined with the Table 1 result of ~0.5 abundance of k-mers? Perhaps the window used to represent the densities here is too wide, given that it shows an almost ubiquitous presence of the TTAGGG?*

*b) Why don’t the canonical sequences make up a greater percentage at the beginning of the reads - this suggests the “boundary of the annotated telomere tract” may be incorrect?*

A: Thank you for identifying this important discrepancy. We rectified this problem. In the original submission, we were in fact underestimating the fraction explained by each motif, as reported in Table 1. To be precise, given that motif contexts may slightly overlap, there are at least two ways to quantify this extent. We now include both measures (maximum fraction explainable by the motif, as well as the fraction of *k*-mers specifically attributable to this motif), describe in detail how they differ and how these values were obtained, and report them (Table 1). Motif densities in Figure 2 visually represent the sequence explainable by the motif, and now are consistent with the values reported in Table 1.

It is also a completely accurate assessment that the canonical sequences, on some chromosomes, do not make up a big percentage immediately after the tract boundary (such as on chromosome 8). We note that this could be a consequence of the innacuracy of the *hg38* reference genome. This calls back to one of the previous questions that concerned the selection of *hg38* for the construction of the extended reference. We argue that this choice serves a purpose, as we identify potential problems in this widely used reference, and draw attention to the need of research specifically addressing reassembly of the genome.

We would like to note that we are currently working together with the T2T consortium to attain better assemblies and understanding of distal regions of human chromosomes, which is a distinct *assembly* project that presents a fundamentally different set of issues and goals, while the current paper simultaneously demonstrates the application of our framework (*edgeCase*), uncovers motif variation in GIAB reference samples, and highlights the need for improvements in the reference genome.

Q: *c) Overall this is an interesting way to show the enriched motifs, but biologically what does this mean, why is it interesting?*

A: Knowledge about the distribution of enriched motifs can pave way to better understanding of implications of telomere variation in disease phenotypes (for example, see Lee et al., 2018, which we now mention in the Introduction section). In addition, the motif density plots indicate, both within subjects (Figure 2, Supplemental Figs. S1–S6) and across subjects and populations (Figures 4 and 5), that telomeres of different chromosomes have different characteristic secondary motifs, which we now also note in the Discussion section. This can serve as a jumping off point for investigating implications in biological pathways (for instance, the significance of one of such motifs has already been demonstrated by Bluhm et al., 2019).

Moreover, our results suggest the possibility that different cell types, and in particular, even different B cell clones can have different telomeric haplotypes (i.e., distributions of motifs), with potential implications in immunology.

* Bluhm, A., Viceconte, N., Li, F., Rane, G., Ritz, S., Wang, S., Levin, M., Shi, Y., Kappei, D. and Butter, F., 2019. ZBTB10 binds the telomeric variant repeat TTGGGG and interacts with TRF2. *Nucleic acids research*, 47(4), pp.1896-1907.

Q: *Supp Fig 5 compares the motif entropies in all 3 datasets. This is interesting and I wish they would comment more on this. I think it backs up well that these sequencing methods can actually be used to evaluate motif distribution/potential haplotypes.*

A: Thank you for this positive assessment. We expanded on the entropy calculation in the Materials and Methods section, as well as provided more comprehensive plots and metrics in the Results section; as a result, these data are now presented as Figure 3 in the main body of the manuscript.

Q: *Figure 3:*

*a. A little bit of space defining levenshtein distance and silhouette score wouldn’t be amiss.*

*b. Did they look to see if the clustering showed snps in the subtelomeric(aka mapped) genome regions that they could use to confirm their haplotyping?*

*c. Again though, why no comment on the biological implications of haplotype identification? And I’d suggest as before that looking at trio data strengthens this result substantially. If the haplotype differences are the same/difference depending on relation this would very interesting.*

A: Following the inclusion of the remaining GIAB subjects, we significantly reworked this section of the paper. We included a more thorough explanation of the Levenshtein distance, and performed clustering in the global, multi-subject context; this context enabled us to investigate the “big picture” of the haplotypic variation with evidence for intra- and inter-subject similarities, which also suggests the existence of multiple haplotypes in B lymphoblastoid cell lines. As in this new global context it did not make sense to constrain the clustering to a specific number of subclusters (where the silhouette score was an applicable metric), we opted now to describe these haplotype spectra with other methods, such as the cophenetic distance and relative distance comparisons with the Wilcoxon signed-rank test, the description of both of which we include in the revised Materials and Methods section. This also allowed us to inspect and quantify the extent of haplotype similarity within trios, both overall and on the per-chromosome basis. As tempting and as apposite the suggestion to correlate our data with subtelomeric SNVs was, we had to abstain from such inferences: for each read, the bigger the portion of it lay in the telomere, the smaller the portion was in the subtelomere, and vice versa, and therefore, per-read correlation of such data was statistically informative for only a very small subset of reads. Given that we now have evidence of haplotypes supported from multiple subjects, we argue that these findings provide a more reliable approach to validation, compared to correlations with subtelomeric variants.

**Reviewer #3**:

Q: *Did the authors examine the CHM13 assembly from T2T consortium? This would be interesting to analyze since it is the among the best GRCh38-level assemblies to date.*

A: Thank you for this constructive and fitting suggestion. In fact, we are currently working together with the T2T consortium to attain better assemblies and understanding of distal regions of human chromosomes. However, this research presents a fundamentally different set of issues and goals, as the current CHM13 assembly is a continuously evolving moving target. In particular, to do the T2T effort justice, CHM13 subtelomeres need to be assembled and annotated *de novo* as well, and the telomeric annotations to be built on top of this, which makes it a project that we are indeed currently pursuing, but a project distinct from the one presented in this manuscript. The project that we are presenting here capitalizes, in particular, on the fact that we are able to investigate variation of the telomeric sequences in the same cell type (B lymphoblasts, rather than the hydatidiform mole in the case of T2T) of multiple subjects, and in fact ties into the answer we provide to your next question, as in this revision of the paper we include all seven Genome in a Bottle (GIAB) datasets and investigate intra- and inter-subject variation.

Q: *Is it possible to perform an orthogonal comparison for telomere-based haplotypes by examining parental data for HG002 and HG005?*

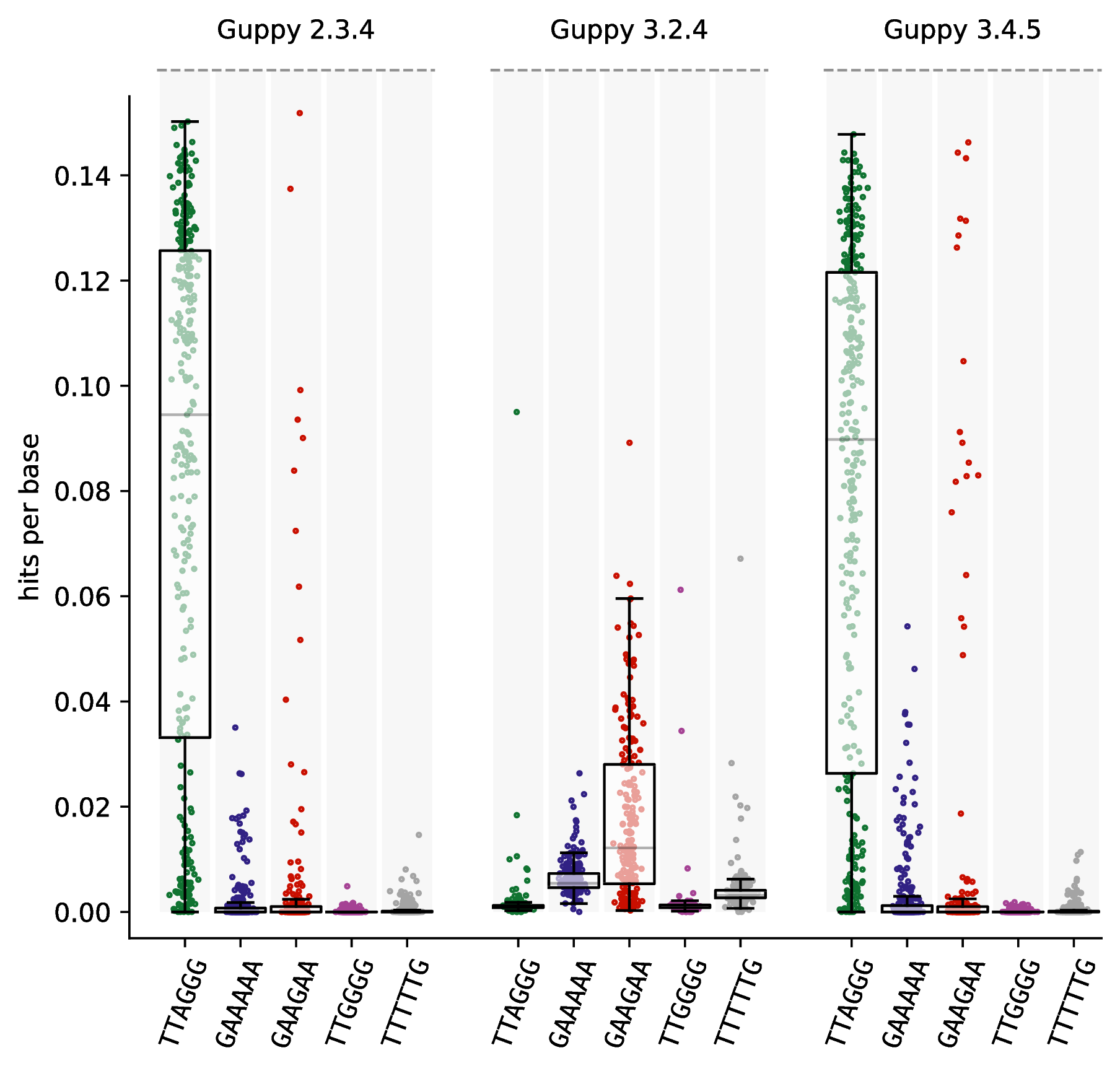
A: We would like to thank you for this extremely fitting and applicable recommendation. During the development of our framework, PacBio CCS data for only three GIAB subjects had been published, but in this revision, we were able to include all seven subjects, as the relevant data was already available. This made it possible to investigate not only the parental relationships in the data, but inter- and intra- population variation as well. We analyzed all these data in an integrated fashion, included all the results and the additional methods we used, and presented evidence for haplotypic variation that describes such relationships.

Q: *Would long nanopore reads help in this regard? If so, how much more resolution could be achieved for a deeply sequenced genome like HG002?*

A: We investigated the applicability of ONT long-read data for this purpose and found that currently it is not trustworthy enough to be used for this purpose. Specifically, we found that base calling models are not sufficiently trained to reliably recognize low-complexity genomic regions, likely because the training data for these models was lacking in telomeric sequences. This was evident from the fact that different versions of the *Guppy* basecaller from Oxford Nanopore (https://community.nanoporetech.com) did not agree on the sequence content of the same telomeric reads. We compared the results of the basecalls of HG002 nanopore reads from three different versions of *Guppy* – 2.3.4, 3.2.4, and 3.4.5 – and found that among the five top enriched motifs (see included Figure), in addition to highly suspect six-mers GAAAAA and GAAGAA, the versions did not agree even on the prevalence of the canonical TTAGGG motif, where it was significantly less frequent in version 3.2.4 than in versions 2.3.4 (Mann-Whitney U test, adjusted *p* = 2.3e-65) and 3.4.5 (*p =* 9.7e-64). While it is tempting to discard version 3.2.4 as the “odd one out,” there is not enough evidence to suggest that either other version is currently reliable for the analysis of telomeric sequences.

Q: *Did the authors look at nanopore-based assemblies from the Shafin et al. 2020 paper for HG002 and HG005?*

A: Following the assession of Nanopore telomeric data, we, unfortunately, had to decide against using data generated from ONT experiments.



**Figure**: Distributions of per-read fractions explainable by the top five enriched motifs on the *q* arms of the HG002 GIAB ONT dataset, as basecalled by three different versions of the *Guppy* software. The vertical axis (hits per base) represents the number of *k-*mers contributing to the motif in each read, divided by the length of the read.

Q: *Is there a possibility of any biases in the interpretation due to these data arising from cell lines vs. human samples?*

A: Thank you for raising such a valid concern. There was indeed such a possibility; in order to assess whether the non-canonical motifs that we found can be observed in human samples, we validated our findings using several sequencing experiments from human subjects that were previously generated for an unrelated project in the lab, and found that at least 15 non-canonical motifs are validated, with the same top four motifs as in the assessed cell lines (see Supplemental Table S3 submitted along with the manuscript). Following this validation, we focused again on the cell lines, as there is an advantage in continuously elucidating novel information from an agreed-upon set of reference samples, with an additional benefit of assessing within-trio and interpopulation relationships.

Q: *I felt that the methods section includes information that really should be in the results. My suggestion would be to rework the section.*

A: This, again, is a great suggestion and we thank you for bringing our attention to it. As we revised the paper and included more data and, therefore, results, we tried to address it by not only adding the relevant results to the appropriate section, but also highlighted the extended genome in the Results section.