

A comprehensive tandem repeat catalog of the human genome

Corresponding Author: Dr Inanc Birol

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Comments on: "A comprehensive tandem repeat catalog of the human genome",
a manuscript submitted for publication to Nature Communications

This manuscript presents an overview of the occurrence of tandem repeats (TRs) in human haplotypes assembled from long read sequencing data of 272 individuals. Most TR genotyping approaches involve inferring TR alleles from an alignment of sequencing reads against a reference genome. This can introduce biases, as loci that are highly different in an individual's genome compared to the reference may be improperly represented in the alignment.

The authors present an alternative approach where TRs are annotated directly in contigs assembled from long sequencing reads, circumventing the need for a reference genome. Since this approach was performed on 272 individuals, the authors were able to generate a catalogue of TR polymorphisms across a population of healthy samples.

The findings presented in this manuscript will be of interest to researchers studying TRs and their variability in human populations. Furthermore, the TR variation catalogue may be relevant to researchers interested in tracing the genetic basis of rare and complex disorders.

For the most part, methodologies used for the collection and generation of assemblies are sound and are clearly described. The results section discusses a range of characteristics and implications derived from the TR catalogue. The section describing the variability of coding STRs (lines 330-357) is particularly strong, as it highlights how TR variability can give rise to a range of viable protein products across a healthy population.

However, I do have concerns about some of the analyses presented in the manuscript, which I outline below. My main concern is that the inclusion of many short repeat loci (especially STRs with impure units) inflates the TR catalogue, which makes it unwieldy to work with and may obscure interesting effects when performing statistics on data aggregated across the catalogue. I feel the arguments put forward to justify the inclusion of such short STRs are not sufficiently convincing.

The strategy further risks that the resulting TR catalogue including many low complexity regions that are not in reality TRs.

Major points

lines 120-123: I have a major concern that the inclusion of STRs with impure units and copy numbers between two and three inflates the TR catalogue and affects statistics performed on the datasets presented in this manuscript. Loci with such low copy numbers are not considered STRs in any recent studies, as short patterns resembling repeats may appear by chance, especially in low complexity regions.

When I filter the CHM13 TR catalogue using STR copy number thresholds described in Mousavi et al., (2019; reference 28 in the manuscript), it removes over 10 million loci with unit length 2 – 6 bp. This constitutes roughly three-quarters of STRs and over half the total number of TRs in the CHM13 catalogue. Importantly, the four loci presented as justification for including short and impure STR loci (lines 454-460) are all longer than the Mousavi thresholds and would therefore not be removed by implementing this filter.

Could the authors clarify why they chose to include such low copy number loci as STRs? Do they observe length polymorphisms at loci consisting of only two repeat units? I feel a re-evaluation of the minimal locus length for STRs is warranted. Further, it would be useful to discuss the purity threshold for STRs.

lines 298-365: I have two questions about the analyses presented in this section. First, it is unclear to me why the STR allele lengths are clustered. What is the basis for saying small differences in allele length are per definition genotyping errors?

Second, the clustering ('binning' may be a more appropriate term for one-dimensional data) approach seems quite ad-hoc. What is the 5% difference criterion based on? Could the authors clarify why they chose not to use some established algorithm that works on one-dimensional data? (DBSCAN, KDE-based clustering, etc.).

line 311 & 316: This goes back to my previous comment about lines 120-123: it seems likely that the very low percentage of polymorphic STRs is due to the fact that the set of STRs described in the paper is inflated by very short, impure loci.

from line 367: Section "Motif heterogeneity" would benefit from expanding on potential molecular mechanisms generating different motif variants observed for the same STR. Presumably these motifs are homologous, so by aligning unit copies their evolutionary molecular histories can be reconstructed. I expect that this will explain differences observed in Fig. 3. Without further analyses, this section is not very informative.

line 401: Jam et al., (2023; reference 54 in the manuscript) also presented an STR-based PCA of 1000G samples in their Supplementary Figure 7. A comparison to this analysis seems appropriate here.

Minor points

line 27: The term 'polymorphism' may be more appropriate than 'divergence', here.

line 48: Range of read lengths is unclearly specified. Consider changing to '15 kb to 100 kb'.

lines 63-69: To me, these sentences discussing CHM13 seem a bit disconnected from the rest of the narrative and break the flow of reading. Maybe the authors could consider restructuring this section to fit better or remove it.

lines 135-136: It is unclear to me what is the basis for the expectation that TRs with modest deviations from the reference should be reliably captured by assemblies. Could the authors please specify which part of the 'analysis below' confirms this expectation? Additionally, this statement could be strengthened by performing alignment-based STR variant calling on (a subset) of samples using another method besides Straglr that also genotypes shorter loci. These results could then be compared to the results of the assembly-based approach.

lines 175-177 & 189-190: I am curious how many genotypes are missing due to the fact that the genomic region in a locus was filtered out during repeat detection, and how many genotypes are missing because a repeat locus is not present in a given haplotype (i.e., region was assembled and analysed with TRF, but no repeat was returned). Would it be possible to include these statistics in the manuscript?

lines 178 & 191: I feel the qualifier 'small' may not be appropriate for a set of 1.4 million or 1.6 million loci. Maybe this could be rephrased.

line 210: The description of criterion (d) is unclear to me: could the authors clarify how a SD of 100 or more indicates limited variability?

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)

The work by Chiu, Birol et al. describes constructing a tandem repeat catalog of the human genome, obtained by analyzing long-read sequencing data from hundreds of human donors. This could become a useful resource and, at a technical level, the execution of the analyses seems sound.

However, the catalog as provided (in vast BED files that, even in their compressed form, are >3GB, i.e. larger than the human genome sequence itself) is too large to be useful to most readers, including this reviewer. I wasn't even able to perform basic due diligence by checking on specific TRs whose properties I know about.

The authors' criterion for inclusion in this catalog are extremely permissive – far too permissive, I would say. Even trivially short TRs like "AGAG" (that would have been apparent from genome sequences long before the current work) meet the authors' criteria, and in fact it sounds like they comprise the vast majority of the 18 million elements. This renders the resulting catalog enormous and unwieldy – I couldn't even uncompress it on my computer. And it is therefore hard for many readers to use effectively, and was hard for me to perform basic due diligence as a reviewer. This took a long time, delayed my review, and I was still unsuccessful in the end.

If this is to become a useful resource, I strongly suggest making available additional files that contain subsets of the catalog that are much more likely to be of interest in common use cases. The one of clearest interest would be one containing just the polymorphic tandem repeats (as the authors note, "the overwhelming majority of STRs genotyped in our samples are non-polymorphic"). And perhaps another limited to the TRs that affect gene coding sequences, or that have at least a threshold number of repeats (such as 10).

In its current state, the catalog is dominated by trivially short repeats that were apparent long before the current work and that make the useful parts of the catalog effectively inaccessible to most readers.

Minor comments:

The manuscript would ideally include some summary descriptive figures or tables reporting how many of the TRs contain 10, 20, 30, 40 or more repeats, and how many contain 2, 3, 4, 5 or more alleles. This is much more interesting than the "18 million" number that is emphasized in the abstract – a number that, again, arises mainly from trivial short repeats that were apparent long before the current work.

Some very straightforward analyses would greatly expand interest to readers. For example,

- what features of a TR (number of copies, length of motif, genic/nongenic, sequence content) affect the likelihood that it is polymorphic in human populations?
- which specific TRs show the greatest differentiation in length across human populations?

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have addressed all the comments.

(Remarks on code availability)

The authors have addressed all the comments.

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

It is helpful that the giant catalog files can be now navigated with tabix (they were impossible to evaluate on the first submission) – though like most readers, I don't have tabix on my laptop, and it took quite some time and help to figure out how to install it. It is worth thinking about how to make the catalog easy for most readers to access.

Once I had figured out tabix, this did allow me (finally!) to do some due diligence on the study's results for well-known repeats. And overall, the results look sensible. So, that was reassuring w.r.t. quality of the catalog.

There really does need to be some kind of README file explaining the various fields of the catalog (did I miss this?). For example, it is not clear how impure repeats were handled – I am guessing that perhaps the reported "copy number" includes inexact instances of the motif that interrupt the repeat (unless they are close to the end), but I couldn't find where this is stated or explained. For the polymorphic repeats (a small fraction of the total) with impurities/interruptions, it would be so much more helpful to have the catalog include the full sequences of the observed alleles (including where the interruptions are, and thus how long the "pure" section of the repeat is) and the frequencies/counts of these alleles, instead of just the light summary numbers (copy number, etc.). I continue to think that the catalog is dominated by short and somewhat trivial repeats (making it unnecessarily large) and would be more useful if it contained more information (such as precise allele sequences) about the subset of (generally longer) repeats that are polymorphic.

For the commonly polymorphic repeats (a small fraction of the total), I would strongly suggest including a VCF with the sequences of the individual alleles that were ascertained.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewer #1 (Remarks to the Author):

Comments on: "A comprehensive tandem repeat catalog of the human genome", a manuscript submitted for publication to Nature Communications

This manuscript presents an overview of the occurrence of tandem repeats (TRs) in human haplotypes assembled from long read sequencing data of 272 individuals. Most TR genotyping approaches involve inferring TR alleles from an alignment of sequencing reads against a reference genome. This can introduce biases, as loci that are highly different in an individual's genome compared to the reference may be improperly represented in the alignment.

The authors present an alternative approach where TRs are annotated directly in contigs assembled from long sequencing reads, circumventing the need for a reference genome. Since this approach was performed on 272 individuals, the authors were able to generate a catalogue of TR polymorphisms across a population of healthy samples.

The findings presented in this manuscript will be of interest to researchers studying TRs and their variability in human populations. Furthermore, the TR variation catalogue may be relevant to researchers interested in tracing the genetic basis of rare and complex disorders.

For the most part, methodologies used for the collection and generation of assemblies are sound and are clearly described. The results section discusses a range of characteristics and implications derived from the TR catalogue. The section describing the variability of coding STRs (lines 330-357) is particularly strong, as it highlights how TR variability can give rise to a range of viable protein products across a healthy population.

However, I do have concerns about some of the analyses presented in the manuscript, which I outline below. My main concern is that the inclusion of many short repeat loci (especially STRs with impure units) inflates the TR catalogue, which makes it unwieldy to work with and may obscure interesting effects when performing statistics on data aggregated across the catalogue. I feel the arguments put forward to justify the inclusion of such short STRs are not sufficiently convincing.

The strategy further risks that the resulting TR catalogue including many low complexity regions that are not in reality TRs.

Major points

lines 120-123: I have a major concern that the inclusion of STRs with impure units and copy numbers between two and three inflates the TR catalogue and affects statistics performed on the datasets presented in this manuscript. Loci with such low copy numbers are not considered STRs in any recent studies, as short patterns resembling repeats may appear by chance, especially in low complexity regions.

When I filter the CHM13 TR catalogue using STR copy number thresholds described in Mousavi et al., (2019; reference 28 in the manuscript), it removes over 10 million loci with unit length 2 – 6 bp. This constitutes roughly three-quarters of STRs and over half the total number of TRs in the CHM13 catalogue. Importantly, the four loci presented as justification for including short and impure STR loci (lines 454-460) are all longer than the Mousavi thresholds and would therefore not be removed by implementing this filter.

Could the authors clarify why they chose to include such low copy number loci as STRs? Do they observe length polymorphisms at loci consisting of only two repeat units? I feel a re-evaluation of the minimal locus length for STRs is warranted. Further, it would be useful to discuss the purity threshold for STRs.

Author's response: We thank the reviewer for raising the question about the high number of loci in our catalog. We agree that a lot of the TRs with two repeat copies are not polymorphic and therefore we filtered out all TRs with average repeat copies less than three. We also did not originally apply any filtering based on the number of successful genotypes, but upon closer analysis we conclude that many loci with only intermediate number of calls (Supplementary Fig. 2) are potentially artefacts. To bolster confidence in the validity of the genotyped loci, we applied further filtering by only keeping loci with at least 400 alleles genotyped. Together, these changes reduce the number of loci by more than 2/3 to ~5 million in the final catalog (lines 172-176). While we still think that a comprehensive catalog would be valuable for the community, we also acknowledge that some users may be particularly interested in polymorphic TRs. We therefore are also making available a subset of only polymorphic loci (those with non-identical genotypes) (line 826).

We did not impose any purity threshold in our genotyping pipeline as it was designed to process genotypes uniformly regardless of motif lengths. Applying a purity threshold for VNTRs may result in a loss of important loci as they are prone to exhibit sequence variations in their long motifs. Taking non-integer copy numbers as a proxy for sequence impurity, approximately 55% of the genotyped STRs showed some degree of impurity (lines 164-167). However, this number may also be contributed to by repeat boundary imprecision as a result of the lenient TRF parameters employed.

lines 298-365: I have two questions about the analyses presented in this section. First, it is unclear to me why the STR allele lengths are clustered. What is the basis for saying small differences in allele length are per definition genotyping errors? Second, the clustering ('binning' may be a more appropriate term for one-dimensional data) approach seems quite ad-hoc. What is the 5% difference criterion based on? Could the authors clarify why they chose not to use some established algorithm that works on one-dimensional data? (DBSCAN, KDE-based clustering, etc.).

Author's response: It is hard to establish that small variations in allelic tandem repeat lengths were biological or technological artefacts, originating from sequencing or bioinformatics. We therefore attempted to cluster our results to identify major alleles and avoid over-estimating size polymorphisms. However, a universal distance threshold for size clustering is difficult to determine. We attempted to use various algorithms (including those mentioned by the reviewer) with different parameters without much satisfaction. The use of a custom clustering algorithm with an arbitrary relative distance of 5% seemed to exhibit the best performance on some loci as seen in the original supplementary figures 6 and 7. However, we realized that the extent of size variability of some highly polymorphic loci made the identification of major alleles via clustering unattainable. This is also evident in the original supplementary figures 6 and 7. We therefore replaced the use of clustering with violin plots for visualizing the size variability of polymorphic loci (lines 311-320). Violin plots not only effectively convey the extent and nature of variability but also avoid the need to arbitrarily choose clustering parameters. The new Supplementary figure 6 clearly illustrates cases of extremely polymorphic loci with a large size range (top) and cases where two or more major alleles exist within a smaller size range. However, we kept the clustering result in the BEGAIN locus analysis (lines 320-323) as it helped effectively identify the 12 closely-sized major alleles.

line 311 & 316: This goes back to my previous comment about lines 120-123: it seems likely that the very low percentage of polymorphic STRs is due to the fact that the set of STRs described in the paper is inflated by very short, impure loci.

Author's response: After applying the new filtering, the percentage of STRs with more than one allele length is ~33% (line 187 and Fig. 1f)

from line 367: Section "Motif heterogeneity" would benefit from expanding on potential molecular mechanisms generating different motif variants observed for the same STR. Presumably these motifs are homologous, so by aligning unit copies their evolutionary molecular histories can be reconstructed. I expect that this will explain differences observed in Fig. 3. Without further analyses, this section is not very informative.

Author's response: The loci in the seven cases with association between repeat size and motif sequences are all heterozygous. Per the reviewer's suggestion, we attempted to infer the evolutionary molecular history of these alleles by constructing the phylogenetic tree based on Clustal alignments of these sequences (lines 404-414, Supplementary Fig. 11). We can observe clear separation of the two different alleles on the phylogenetic tree, with placement of intermediate sized alleles at the branching points. In terms of potential molecular mechanisms, in four of the seven loci short repeat tracts corresponding to the short allele coexist in juxtaposition to the expanded alleles, suggesting a possible expansion event with simultaneous motif conversion. We do not have enough data to perform a more thorough evolutionary analysis based on motif heterogeneity.

The purpose of this vignette was to investigate whether the association between motif sequence and size observed in some pathogenic TR alleles is also observed in healthy populations. From the catalog data, the seven supporting cases we observed appear to be exceptions rather than the norm.

line 401: Jam et al., (2023; reference 54 in the manuscript) also presented an STR-based PCA of 1000G samples in their Supplementary Figure 7. A comparison to this analysis seems appropriate here.

Author's response: Thanks to the reviewer for pointing us to this analysis. The clustering patterns in the PCA results of Jam et al. is very similar to ours, with EAS and AFR samples forming their own distinct clusters while SAS and AMR samples overlapped in close proximity to the EUR samples. We added this comparison to the manuscript (lines 441-443), as the reviewer suggests. We also point out the small difference in clustering pattern we observed in comparison to the analysis performed by Frontanilla et al (lines 443-445).

Minor points

line 27: The term 'polymorphism' may be more appropriate than 'divergence', here.

Author's response: Since the term "polymorphic" was already used in the first part of the sentence, we think it would be better to replace the word 'divergence' with 'differences' instead of 'polymorphism' to avoid repeating the same word.

line 48: Range of read lengths is unclearly specified. Consider changing to '15 kb to 100 kb'.

Author's response: We thank the reviewer for pointing this out. The use of "≥" is unnecessary and was removed from the sentence.

lines 63-69: To me, these sentences discussing CHM13 seem a bit disconnected from the rest of the narrative and break the flow of reading. Maybe the authors could consider restructuring this section to fit better or remove it.

Author's response: We agree. We have removed the description of CHM13 in the mentioned paragraph and moved on to continue describing the emergence of high-quality haplotype-resolved assemblies to maintain the focus in the preceding paragraph about using LRS for TR genotyping (lines 62-68).

lines 135-136: It is unclear to me what is the basis for the expectation that TRs with modest deviations from the reference should be reliably captured by assemblies. Could the authors please specify which part of the 'analysis below' confirms this expectation? Additionally, this statement could be strengthened by performing alignment-based STR variant calling on (a subset) of samples using another method besides Straglr that also genotypes shorter loci. These results could then be compared to the results of the assembly-based approach.

Author's response: Small TRs generally present a less challenging problem for sequence assemblers than expanded, long alleles. For example, from the size distribution (top) panel of the upset plot in Supplementary Figure 4, over 60% of matching alleles between assembly-based and read-based (both ONT and HiFi) results are less than 1kb in size, thereby corroborating our assertion. This analysis was, however, limited to loci where at least one allele was at least 100bp larger than the reference (heterozygous loci may contribute smaller alleles for the analysis). To increase the sample size of alleles with modest size deviations from the reference, we identified over 15,000 loci from the catalog with the largest allele only between 20 to 100 bp larger than the reference. We used TRGT, an alignment-based STR variant caller, to genotype these loci from the minimap2 alignments of 28 PacBio HiFi samples. The Pearson coefficient of correlation was 0.99 between the assembly-based and TRGT genotypes of these loci (Supplementary Fig. 5). This analysis, which serves to complement the original evaluation we performed on large TRs, has been added to the "Assembly-based vs read-based genotypes" section between lines 270-279.

lines 175-177 & 189-190: I am curious how many genotypes are missing due to the fact that the genomic region in a locus was filtered out during repeat detection, and how many genotypes are missing because a repeat locus is not present in a given haplotype (i.e., region was assembled and analysed with TRF, but no repeat was returned). Would it be possible to include these statistics in the manuscript?

Author's response: The analysis has been done and the results are presented in the sections where the catalog summarization is first presented (lines 178-185). The results vary by sample but on average 120,000 genotypes (2.3%) were missing per haplotype. 56% of the missing loci were not assembled or partially assembled. Only 7% were missing because of failed TRF detection. About 20% were lost in the filtering steps within the assembly-based genotyping pipeline, including loci which cannot be confidently mapped back to the reference genome. The remaining 18% of the missing loci have overlaps with other loci in the catalog. They are typically sequence variants of the overlapping TRs and ended up being filtered out from the catalog with insufficient (<400) occurrences in the cohort.

lines 178 & 191: I feel the qualifier 'small' may not be appropriate for a set of 1.4 million or 1.6 million loci. Maybe this could be rephrased.

Author's response: The number of nested loci has been reduced to ~100 thousand. This is a small (2.1%) subset of the total number of approximately five million loci in the catalog.

line 210: The description of criterion (d) is unclear to me: could the authors clarify how a SD of 100 or more indicates limited variability?

Author's response: We thank the reviewer for picking this up. After inspection, we agree that a SD of 100 bp is indeed too big for collecting loci of limited variability. We have reduced the SD threshold to 20 bp (line 218). This reduces the number of identified loci by almost 10-fold. As a consequence, the variability of allele sizes is more constrained and the analysis is more successful in detecting potential errors of the reference genomes in these loci (lines 222-224).

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3 (Remarks to the Author):

The work by Chiu, Birol et al. describes constructing a tandem repeat catalog of the human genome, obtained by analyzing long-read sequencing data from hundreds of human donors. This could become a useful resource and, at a technical level, the execution of the analyses seems sound. However, the catalog as provided (in vast BED files that, even in their compressed form, are >3GB, i.e. larger than the human genome sequence itself) is too large to be useful to most readers, including this reviewer. I wasn't even able to perform basic due diligence by checking on specific TRs whose properties I know about.

Author's response: We thank our reviewer for pointing out the technical difficulty of downloading and using the catalogs. The revised catalog file prepared after new filtering (lines 173-177) is about 800Mb in compressed format. In addition, we indexed the BED files with Tabix so that loci-of-interest can readily be extracted using the command: `tabix <catalog.gz> chromosome:start-end`.

The authors' criterion for inclusion in this catalog are extremely permissive – far too permissive, I would say. Even trivially short TRs like "AGAG" (that would have been apparent from genome sequences long before the current work) meet the authors' criteria, and in fact it sounds like they comprise the vast majority of the 18 million elements. This renders the resulting catalog enormous and unwieldy – I couldn't even uncompress it on my computer. And it is therefore hard for many readers to use effectively, and was hard for me to perform basic due diligence as a reviewer. This took a long time, delayed my review, and I was still unsuccessful in the end.

Author's response: The revised catalog contains about 5 million loci (see Author's Response to Reviewer #1's first major issue and manuscript lines 173-177). As mentioned above, the indexed BED files should be easily interrogated for specific regions-of-interest using the tabix utility. The uncompressed file is 18G in size.

If this is to become a useful resource, I strongly suggest making available additional files that contain subsets of the catalog that are much more likely to be of interest in common use cases. The one of clearest interest would be one containing just the polymorphic tandem repeats (as the authors note, "the overwhelming majority of STRs genotyped in our samples are non-polymorphic"). And perhaps another limited to the TRs that affect gene coding sequences, or that have at least a threshold number of repeats (such as 10).

Author's response: Two abridged versions of the revised catalog have been prepared: one with TRs only in genic regions (~2.4 million loci) and another with TRs that have more than one allele genotyped in our cohort (~2.1 million loci). Both abridged catalogs are available from the Zenodo repository (lines 834-836).

In its current state, the catalog is dominated by trivially short repeats that were apparent long before the current work and that make the useful parts of the catalog effectively inaccessible to most readers.

Author's response: ~70% of the revised catalog has 5 or fewer repeat copies, compared to 89% of the original catalog. We hope that the new catalog, together with the accompanying abridged versions, will become a useful and accessible resource for the research community.

Minor comments:

The manuscript would ideally include some summary descriptive figures or tables reporting how many of the TRs contain 10, 20, 30, 40 or more repeats, and how many contain 2, 3, 4, 5 or more alleles. This is much more interesting than the "18 million" number that is emphasized in the abstract – a number that, again, arises mainly from trivial short repeats that were apparent long before the current work.

Author's response: We thank our reviewer for this suggestion. We have added three figures describing the motif length distribution of TRs in the final catalog (Fig. 1d), the cumulative distribution of repeat length measured in repeat copies (Fig. 1e), and the cumulative distribution of number of alleles per locus (Fig. 1f) of all loci in the revised catalog. Discussion of these figures has been added to the text. (lines 187-190)

Some very straightforward analyses would greatly expand interest to readers. For example,

- what features of a TR (number of copies, length of motif, genic/nongenic, sequence content) affect the likelihood that it is polymorphic in human populations?

Author's response: Our analysis showed that number of copies or repeat length contributes more to polymorphism (37% of STRs with alleles of >1 length detected, line 309) in our data than sequence content (14%, line 380). We also added an analysis comparing TR lengths within and outside of subtelomeric regions because we observed some highly polymorphic loci in the subtelomeres (lines 332-337, supplementary Fig. 8) that have also been reported in another study. Our results indicate that TRs in subtelomeres are more polymorphic than those outside, and this effect is more apparent for VNTRs than STRs (lines 325-332, supplementary Fig. 7).

- which specific TRs show the greatest differentiation in length across human populations?

Author's response: We did not identify any TRs that demonstrate statistically significant length differentiation across the five superpopulations represented in this study.

Reviewer #3 (Remarks to the Author):

It is helpful that the giant catalog files can be now navigated with tabix (they were impossible to evaluate on the first submission) – though like most readers, I don't have tabix on my laptop, and it took quite some time and help to figure out how to install it. It is worth thinking about how to make the catalog easy for most readers to access.

Once I had figured out tabix, this did allow me (finally!) to do some due diligence on the study's results for well-known repeats. And overall, the results look sensible. So, that was reassuring w.r.t. quality of the catalog.

There really does need to be some kind of README file explaining the various fields of the catalog (did I miss this?). For example, it is not clear how impure repeats were handled – I am guessing that perhaps the reported “copy number” includes inexact instances of the motif that interrupt the repeat (unless they are close to the end), but I couldn't find where this is stated or explained. For the polymorphic repeats (a small fraction of the total) with impurities/interruptions, it would be so much more helpful to have the catalog include the full sequences of the observed alleles (including where the interruptions are, and thus how long the “pure” section of the repeat is) and the frequencies/counts of these alleles, instead of just the light summary numbers (copy number, etc.). I continue to think that the catalog is dominated by short and somewhat trivial repeats (making it unnecessarily large) and would be more useful if it contained more information (such as precise allele sequences) about the subset of (generally longer) repeats that are polymorphic.

For the commonly polymorphic repeats (a small fraction of the total), I would strongly suggest including a VCF with the sequences of the individual alleles that were ascertained.

Author's response: We are glad that the reviewer managed to interrogate the new version of the catalog. A README explaining the columns of the catalog is indeed present in the Github (code) repository, but we agree that a README should be added to the Zenodo (data) repository as well. This has been done, describing the various files in addition to the columns within the catalog (<https://zenodo.org/records/15237646>). The reviewer's guess is right – the copy numbers reported in the catalog are derived by dividing the length of the allele sequence (repeat size) by the length of the consensus motif (lines 629-630). Therefore, impurities and interruptions may inflate the actual copy number of the consensus motif within the allele.

As per request by the reviewer, we have now added a VCF file (<https://zenodo.org/records/15237646>) cataloging polymorphic loci that have at least two alleles different in size and at least two observations in the samples we analyzed. For each locus, we extracted sequences of all alleles from the diploid assemblies, binned them to identify unique sequences that are different from the reference, and reported them in the ALT field of the VCF record. We also reported the count of each ALT allele in the custom INFO field ALT_COUNT. The reference allele (REF_SEQ) and its count (REF_COUNT) observed in our samples are also reported. For each allelic sequence presented, we capitalized the bases comprising the canonical motif while leaving the remainder (impurities, interruptions, or assembly errors) in lowercase. This was done to aid users in visually discerning the impurities of the repeat sequences. We also added an INFO field called MEDIAN_PURITY to report the median purity (fraction of summed lengths of canonical repeats over the entire allele) of all alleles genotyped. We chose to report the median instead of the mean because the former is a more robust metric not affected by outliers. Also, it is well-documented (and demonstrated in this manuscript) that repeat expansions, particularly in disease-associated loci, may comprise variant motifs or changed proportions of interruptions. We added this information to help users assess the qualities of our genotypes and perform filtering of the cataloged loci based on purity if so desired. The VCF is compressed and indexed by Tabix to facilitate interrogation of specific genomic regions.