Title

LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads.

Journal

Genome Biology, 25(1)

Authors

Ziaei Jam, Helyaneh

Zook, Justin

Javadzadeh, Sara

et al.

Publication Date

2024-07-04

DOI

10.1186/s13059-024-03319-2

Genome Biology

SHORT REPORT

Open Access

LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads

Helyaneh Ziaei Jam[1], Justin M. Zook[2], Sara Javadzadeh[1], Jonghun Park[1], Aarushi Sehgal[1] and Melissa Gymrek[1,3*]

*Correspondence: mgymrek@ucsd.edu

[1] Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA
[2] Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Dr, Gaithersburg, MD, USA
[3] Department of Medicine, University of California San Diego, La Jolla, CA, USA

Abstract

Tandem repeats are frequent across the human genome, and variation in repeat length has been linked to a variety of traits. Recent improvements in long read sequencing technologies have the potential to greatly improve tandem repeat analysis, especially for long or complex repeats. Here, we introduce LongTR, which accurately genotypes tandem repeats from high-fidelity long reads available from both PacBio and Oxford Nanopore Technologies. LongTR is freely available at https://github.com/gymrek-lab/longtr and https://zenodo.org/doi/10.5281/zenodo.11403979.

Keywords: Tandem repeats, Long reads, Microsatellites

Background

Tandem repeats (TRs), including short tandem repeats (STRs; repeat unit 1–6bp) and variable number tandem repeats (VNTRs; repeat unit 7+bp), refer to regions of the genome that consist of adjacent repeated units. TRs are a large source of genetic variation in humans [1] and are implicated in a growing list of Mendelian and complex traits [2]. In the last decade, multiple tools have been developed to estimate the repeat length and/or sequence of TRs using short reads (e.g., [3–6]), but certain repeats such as highly complex TRs have remained intractable. Long-read sequencing technologies offer a promising solution. However, tools designed for short reads are ineffective on long reads given the considerable differences including in read length, base calling accuracy, error profiles at STRs (Additional file 1: Fig. S1), and paired-end vs. single-end format.

Results and discussion

Here, we introduce LongTR, which extends the HipSTR [3] method originally developed for short read STR analysis in order to genotype STRs and VNTRs from accurate long reads available from both PacBio [7] and Oxford Nanopore Technologies [8] (ONT). LongTR takes as input sequence alignments for one or more samples and a reference set of TRs and outputs the inferred sequence and length of each allele at each locus. It uses a

clustering strategy combined with partial order alignment to infer consensus haplotypes from error-prone reads, leveraging read phase information when available, followed by sequence realignment using a hidden Markov model which is used to score each possible diploid genotype at each locus (the " Methods" section; Additional file 1: Fig. S2). Unlike other existing long read TR genotypers, LongTR supports multi-sample calling, employs a technology-specific homopolymer error model, and outputs genotype quality scores.

We ran LongTR to genotype repeats in a reference set of 937,122 human TRs from Project Adotto [9] using 30× PacBio HiFi reads from HG002 (Data Availability). To evaluate the accuracy of genotype calls, we extracted alleles for genotyped TRs from the haplotype-resolved genome assembly of HG002 [10], which was generated using multiple technologies and orthogonal computational methods and is thus treated as a ground truth here. LongTR outputs inferred allele sequences in addition to repeat length, enabling sequence level comparisons of the alleles to the assembly. Of 814,319 repeats for which exactly two alleles were extracted from the assembly, LongTR showed 84.7% sequence concordance and 98.5% length concordance allowing for 1bp differences. To further evaluate the accuracy of LongTR, we performed TR genotyping in an Ashkenazi trio (HG002, HG003, and HG004) using the same reference set and determined the Mendelian inheritance (MI) of TR genotypes. Overall, LongTR showed 86% MI at sites where at least one trio member was not homozygous for the reference allele. Mendelian consistency monotonically increases with genotype quality scores computed by LongTR, which can be used to filter low quality calls (Fig. 1a). Homopolymer repeats showed the lowest consistency with MI of 78.3%. We trained a PacBio HiFi homopolymer error model at 840,248 homopolymer repeats from the HipSTR reference, which contains homopolymers with more precise boundaries than the Adotto set, by comparing observed repeat lengths at HiFi reads to those obtained from the HG002 assembly (Additional file 1: Table S1). In the Ashkenazi trio we observed an MI of 83% and 81% for LongTR with and without error modeling (Additional file 1: Fig. S3). Furthermore, we observed a 13% increase in concordance of LongTR alleles with the HG002 assembly when using the homopolymer error model.

Multiple methods for genotyping TRs from long reads have been developed over the last several years [11–14]. We first focused on benchmarking LongTR against TRGT [15], a recently developed TR genotyper that outperforms previous methods on PacBio HiFi reads. Similar to LongTR, TRGT outputs inferred sequences in addition to estimated repeat length. We ran TRGT on HG002 using the same sequencing data and reference set as LongTR. Both tools had similar running time, with TRGT taking 436 min and LongTR taking 428 min to finish. TRGT and LongTR genotyped 99.83% and 99.18% of the repeats respectively. At sites called by both methods, 86.0% of alleles have identical lengths and 98.5% differ in length by at most a single copy number. TRGT calls in the Ashkenazi trio showed reduced Mendelian inheritance rates compared to LongTR (79% for TRGT vs. 86% for LongTR).

LongTR and TRGT showed similar length concordance with the HG002 assembly, with LongTR performing slightly better (97.8% for TRGT and 98.5% for LongTR allowing for 1bp differences). LongTR showed further gains when evaluating sequence concordance. In both cases, the advantage of LongTR over TRGT was highest at long (>500bp) repeats (Fig. 1b). We identified multiple scenarios in which LongTR calls

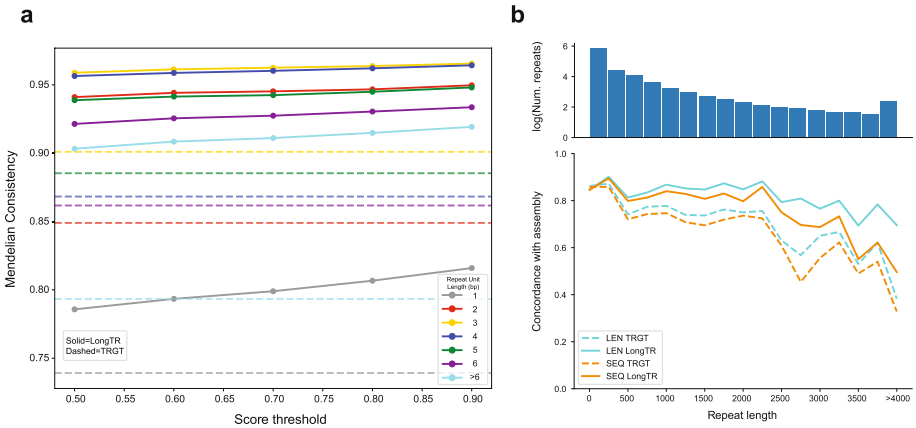

Fig. 1 **a** Assessing Mendelian consistency of TR calls in an Ashkenazi trio using PacBio HiFi reads. The *x*-axis gives the LongTR score threshold to include calls, and the *y*-axis gives the percentage of TRs for which genotypes in the trio follow Mendelian consistency. Trio-TR pairs for which all members were called as homozygous for the reference allele were excluded. Dashed = TRGT; solid = LongTR. Note TRGT does not report a quality score and thus a single horizontal line is shown. Color indicates the size of the repeat unit (in bp) considered. **b** Concordance of TR genotypes obtained from PacBio HiFi with assembly alleles in HG002. TRs were binned by length of the repeat (in bp, bin size = 250bp) in GRCh38. The *x*-axis shows the TR length, and the *y*-axis shows the percent of alleles that match the assembly. Blue lines show when only length is considered. Orange lines show when both length and sequence are considered. Dashed = TRGT; solid = LongTR. The top panel shows the number of repeats in each bin, on a logarithmic scale

match the assembly and TRGT does not, including regions with a high number of truncated reads or regions called as large insertions by TRGT that had low read support. Furthermore, LongTR detected 514 TRs with large structural deletions that remove the entire repeat, resulting in a null allele, whereas these cases are not reported by TRGT (Additional file 1: Fig. S4). Overall, these results suggest both tools perform similarly at TRs in regions that are easier to genotype, whereas LongTR obtains higher quality genotypes at longer or structurally complex regions. Notably, we observed multiple instances where both LongTR and TRGT reported identical genotypes that differed from the assembly. These cases usually occurred in complex genomic regions such as large insertions or segmental duplications or regions of high homozygosity. While complex regions tend to be more accurate with assembly-based approaches, regions of high homozygosity are known to pose challenges to diploid assemblies [16] and thus likely represent assembly errors rather than errors from LongTR or TRGT (Additional file 1: Fig. S5).

We next benchmarked LongTR against two alternative methods for genotyping TRs from long reads, Straglr [14] and Tandem-genotypes [13]. Both comparisons were done using 30× PacBio HiFi reads for HG002. Unlike LongTR and TRGT, these methods output estimated repeat copy number but not inferred allele sequences at each locus. We ran Straglr and LongTR to genotype repeats from the HipSTR reference set, which contains repeats with simpler motif structure, as Straglr automatically infers the motif sequence and more complex repeats caused large discrepancies between LongTR and Straglr calls. Out of 20,592 repeats on chromosome 21, LongTR genotyped 20,158 and Straglr genotyped 13,705. From 13,655 repeats genotyped by both, they matched on both alleles for 47% of the repeats and matched for one allele for another 5% allowing for 1bp off. For benchmarking LongTR against Tandem-genotypes, we used the Adotto

reference set as input to both tools. Out of 12,802 repeats on chromosome 21, Tandem-genotypes and LongTR genotyped 97% and 96% of regions respectively. Regions skipped by LongTR lacked a sufficient number of high-quality reads and were filtered. Since Tandem-genotypes only reports the copy number, we checked if the difference between copy number of LongTR alleles and Tandem-genotypes was less than 1. Out of 12,232 repeats genotyped by both, LongTR and Tandem-genotypes agreed on both alleles for 75% of TRs and agreed on only one allele in 24% of repeats. Overall, these comparisons suggest modest concordance between Straglr and Tandem-genotypes vs. LongTR which may be driven in part by differences in how alleles are reported (imprecise copy number count vs. allele sequences).

We sought to further assess the ability of LongTR to genotype longer repeats. First, we compared LongTR genotypes in HG002 to those from adVNTR [17], a tool specifically designed for genotyping VNTRs, on a reference set of 10,186 autosomal gene-proximal VNTRs (Data Availability). When allowing for differences in up to a single copy number (the "Methods" section), LongTR and adVNTR showed 96% concordance. adVNTR required approximately 23h to genotype HG002, compared to 1.5h for LongTR on this repeat set. Second, we determined whether LongTR could identify large expansions in HiFi reads obtained from patients harboring long pathogenic alleles implicated in Huntington's disease ($n=4$) and Fragile X Syndrome ($n=3$). LongTR correctly identified expansions in *HTT* and *FMR1*, including alleles consisting of up to several thousands bp for both loci (Additional file 1: Table S2). Allele sequences reported by LongTR match repeat unit copy number counts for these reads reported on the dataset website for the tested samples (Data Availability, Additional file 1: Fig. S6, visualized using TRviz [18]).

We next evaluated the ability of LongTR to genotype TRs in a separate long read technology, using ONT's recently released Duplex reads (average length 27 kb compared to 15–20kb for PacBio HiFi reads) available for HG002 [19]. Overall, we observed high concordance between genotypes obtained from phased ONT Duplex and PacBio HiFi (90% allowing for 1bp off; Fig. 2a) with the latter showing higher concordance with the assembly (99% for PacBio HiFi vs 88% for ONT, allowing for 1bp off; $n=798{,}291$ TRs called for both ONT and PacBio HiFi). We found that phasing ONT reads improved performance, with 87% and 89% concordance against the assembly for unphased and phased reads, respectively, across 12,503 TRs on Chromosome 21. We further evaluated ONT using haplotagged Simplex SUP reads on Chromosome 21 and found that SUP and Duplex reads showed similar concordance with the assembly (88.9% for Duplex vs. 88.7% for SUP).

We identified several large repeat expansions uniquely detected by the ONT and not PacBio HiFi data, which were consistent with the alleles in the assembly. These repeats are enriched in the GIAB set of hard-to-map regions [21] (Fisher's exact test two-sided $p=1.75e-292$) (Data Availability). Examining these expansions showed that discrepancies often occurred in regions where few or no PacBio HiFi reads aligned to or spanned the insertion, leading LongTR to inaccurately genotype the locus (Fig. 2b). Our observations suggest that the longer read lengths of ONT enhance the detection accuracy particularly for regions with large insertions compared to the reference.

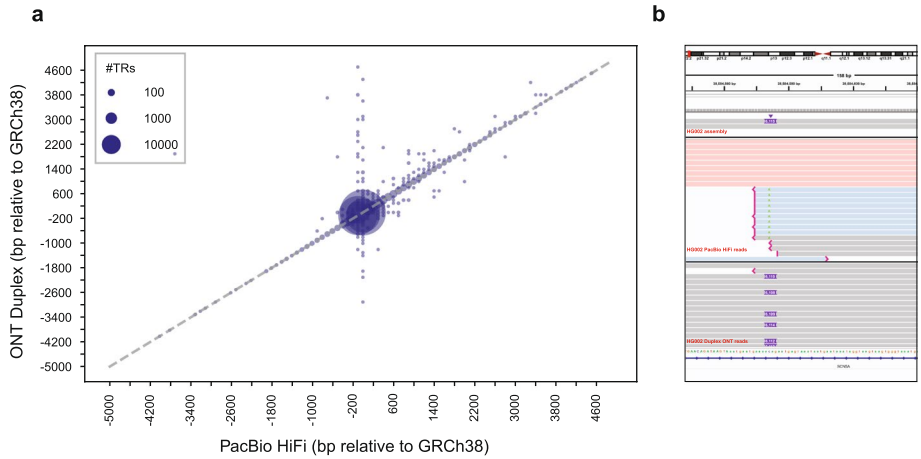

Fig. 2 **a** Comparison of LongTR genotypes on ONT Duplex vs. PacBio HiFi data. For each call, we computed the average of the length of each allele relative to the GRCh38 reference. The *x*-axis gives the calls using PacBio data, and the *y*-axis gives the calls using ONT Duplex data. Bubble size scales with the number of calls at each coordinate. **b** IGV [20] screenshot comparing PacBio HiFi reads vs. ONT Duplex reads at a TR genotyped by LongTR. The top window shows the assembly alignment, the middle window shows aligned HiFi reads, and the bottom window shows aligned ONT Duplex reads at a (AGTAAATAATG)n VNTR. All data is aligned to GRCh38. Red and blue denote PacBio HiFi reads from the two haplotypes of HG002 based on haplotag information. Gray reads have no haplotag information. HiFi reads were clipped at the large repeat insertion, resulting in an incorrect genotype call

Finally, we compared TR genotypes obtained from HipSTR on Illumina short reads and LongTR on PacBio HiFi reads for HG002. For this analysis, we used the hg38 reference set of 1,638,945 STRs available from HipSTR's website, of which only 1.6% are longer than 100bp. Of 1,556,278 STRs that were genotyped by both methods, 88% were concordant, increasing to 97% if allowing for 1bp length difference (Additional file 1: Fig. S7a). HipSTR reported homozygous reference for all repeats with length above 250bp (the Illumina read length) and no read distribution information, indicating genotypes for longer repeats are not reliable. Concordance (by allowing 1bp off) decreases with increasing length of the repeat (Additional file 1: Fig. S7b).

Conclusions

Overall, LongTR provides accurate sequence-resolved TR genotyping from long reads for nearly 99% of TRs in mappable regions of the genome and outperforms existing methods for this task. While the majority of TRs can be resolved, we identified multiple TRs with large, complex insertions relative to the reference that are challenging to span even with long reads and may in some cases be misrepresented by reference assemblies. These cases may represent the limits of mapping-based approaches. Future work is needed to incorporate alternative approaches, such as pangenome or assembly-based methods, that do not suffer from these limitations. We envision these improvements will enable systematic incorporation of TRs into genome-wide analyses for a range of applications.

Methods

Overview of the LongTR method

LongTR extends HipSTR [3], which was originally developed to analyze short reads, to genotype both VNTRs and STRs using accurate long reads. Here, we use accurate long reads to refer to PacBio HiFi and ONT Duplex reads, each of which have been shown to have per-base error rates comparable to those of Illumina reads [22, 23]. Like HipSTR, LongTR begins with aligned reads for one or more samples and the reference coordinates for a predefined set of TRs. Predefined TRs may be simple repeats (e.g., $[AC]_n$) but can also comprise more complex TRs with multiple distinct repeat units (e.g., $[AC]_n[GT]_k[T]_l$). Then, for each TR, it extracts reads encompassing the repeat, infers candidate TR haplotypes, and uses a hidden Markov model to realign all the reads overlapping a repeat region to the candidate haplotypes. Finally, it outputs a VCF file containing each individual's TR genotypes and corresponding quality scores (Additional file 1: Fig. S2). Below, we discuss key steps where LongTR differs from the HipSTR method to enable analysis of TRs from long reads.

Haplotype identification

LongTR uses a new method for the identification of candidate haplotypes from input reads of samples at each TR. It starts by trimming all reads aligned to the target TR of interest to include the repeat plus a user-defined window of context sequence. It then iterates over all trimmed sequences and includes any sequence supported by a sufficient number of reads by one or more samples (at least two reads and more than 20% of reads in a single sample, or more than 5% of all reads across all samples) as a candidate haplotype.

In some scenarios, typically when the repeat is long or the region is complex with multiple insertions and deletions, reads from a haplotype fail to meet the criteria set in the first stage, resulting in their exclusion from the set of haplotypes. To address this, we perform a second iteration during which we identify samples for which over 25% of aligned reads lack a corresponding representative haplotype and form additional candidate haplotypes using excluded reads for each of these samples. These previously excluded reads for each sample are then sorted by length of the sequence aligned to the repeat region, after which a greedy clustering algorithm is applied to form the initial sequence clusters. In this method, the first cluster is formed by designating the first sequence as its centroid. Starting with the second sequence S, we evaluate whether there exists a centroid C within the centroids set for which the edit distance between S and C is below a given threshold T. T is initially set to a small number (10). We then refine the initial sequence clusters through the following steps: (1) a consensus sequence for each cluster is generated using partial order alignment [24] (POA) on read sequences within each cluster. The consensus is used to update the cluster centroid. (2) After updating the centroids for all clusters, we again sort the clusters in order of the length of their revised centroids. (3) We iterate through the clusters, merging two clusters whenever the edit distance between their updated centroids is below T. (4) This process is repeated until no further clusters can be merged. (5) After cluster refinement, we only include clusters with a number of reads above $min(10, 0.1 \times n_s)$ where n_s is the number of excluded reads per

sample s. Finally, we check if the total number of reads in all remaining clusters are more than $0.8 \times n_s$; otherwise, we will increase T (attempting in order the following values: 20, 50, 80, 100, 150, 200, 300, 400, 500, 600, 700) to relax the constraints on sequence similarity for both adding sequences to existing clusters and merging cluster and repeat the steps above. Centroid sequences of final clusters are added to the set of potential haplotypes. Edit distance is calculated using the Needleman-Wunsch algorithm with parameters gap_score = 1, match_score = 0, mismatch_score = 1. We optimized the algorithm by computing a lower bound of total edit distance at each row of the dynamic programming table $dp[n][m]$, where n and m are the lengths of the sequences. For each cell in a row i, we iterate over columns j and compute $\min(\mathrm{dp}[i, \mathrm{j}] + \mathrm{abs}\big((n-m) - \big(i-j\big)\big))$. This value shows the minimum number of insertions or deletions needed to reach cell $dp[n, m]$. If the minimum value per row exceeds the threshold T, computation stops.

Alignment of reads to candidate haplotypes

The original HipSTR approach employs a rigid alignment technique for aligning repeat sequences to candidate haplotypes, operating under the assumption that errors occur in multiples of the repeat unit length and happen only once within the repeat. However, these assumptions do not hold true for long reads (Additional file 1: Fig. S1), likely due to the fact that errors are driven by other processes than PCR. Instead, numerous errors can be present at different positions of the repeat rather than being strictly tied to the repeat unit size (Additional file 1: Fig. S1), with single base insertions or deletions most prevalent. To allow for a more flexible error model, we used a hidden Markov model approach based on that of Dindel [25] to align reads to each candidate haplotype.

LongTR models errors in homopolymer repeats using a geometric distribution as follows:

$$p(errorsize = l) = \begin{cases} 1 - i - d, l = 0 \\ i \times \rho(1-\rho)^{l-1}, l > 0 \\ d \times \rho(1-\rho)^{-l-1}, l < 0 \end{cases}$$

where ρ controls the size of error, and i and d are the probability of error increasing or decreasing the length of repeat respectively. To obtain the values for i, d, and ρ, we used the ALLREADS format field in LongTR output to extract the observed read lengths at each locus and used assembly genotypes as ground truth to compute the base pair differences s with the actual genotype. Then, $\rho = \frac{\#InDels}{\sum abs(s)}$, $i = \frac{\#insertion}{\#reads}$, $d = \frac{\#deletions}{\#reads}$ were computed for homopolymers falling in each specific length range.

Phasing information

We leverage HipSTR's existing option to use read phase information from haplotagged reads (and renamed the option from --10x-bams to --phased-bam) to accurately identify sample haplotypes. To consider phasing information, LongTR requires that at least one read from each haplotype be present, and unphased reads constitute no more than 20% of the total sample's reads.

Genotyping

LongTR iterates over all pairs of haplotypes (H_i, H_j) (including homozygous pairs) and calculates a score for each possible genotype $G(H_i, H_j)$ based on all observed reads R using the following formula:

$$S(G(H_i, H_j)) = \prod_{r \in R} P(h_r = 1) * S_{rH_i} + P(h_r = 2) * S_{rH_j}$$

where S_{rH_i} is the alignment score from aligning read r to haplotype H_i and $P(h_r = 1)$ is the probability that read r is generated from the first haplotype. When read phase information is available, $P(h_r = k)$ is set to 0 or 1 based on the haplotag field. Otherwise, we set the probability to come from haplotype 1 vs. 2 as equally likely. The score for each haplotype pair is then normalized by the sum of scores for all possible genotypes and the result is reported as the quality score of the genotype in LongTR output, defined as Q.

Implementation

LongTR is implemented in C++. It leverages the HTSlib [26] library to read directly from cloud addresses or URLs, which can avoid the costly step of downloading large sequence alignment files.

Evaluating LongTR

LongTR v1.0 was run using non-default parameters --min-reads 4, --haploid-chrs chrX,chrY, --max-tr-len 10,000, --skip-assembly, --indel-flank-len 25, and --phased-bam to genotype repeats with at least 4 overlapping reads, consider chromosome X and chromosome Y as haploid chromosomes, genotype repeats with length up to 10,000bp, skipping assembly of flanking regions, considering InDels up to 25bp around the repeat as InDels affecting the repeat size, and leveraging haplotag information when genotyping, respectively. This set of parameters was used for all the following experiments, unless otherwise stated. To evaluate repeat allele sequences returned by LongTR, the HG002 assembly v1.0.1 [10] (Data Availability), with higher accuracy at homopolymers, was mapped to the GRCh38 reference genome using minimap2 [27] v2.26-r1175. For each repeat on autosomal chromosomes, allele sequences from the maternal and paternal haplotypes that completely span the repeat were extracted and compared to allele sequences reported in the output VCF file of LongTR. For Mendelian consistency analysis on autosomal chromosomes, LongTR was used to perform joint genotyping in all three samples. We only considered a repeat in the Mendelian consistency analysis if (1) all samples were successfully genotyped at that repeat and (2) the genotype for at least one sample was not homozygous for the reference allele. The minimum quality score reported by LongTR among all three samples is considered the assigned score for that repeat.

Benchmarking against TRGT

TRGT v0.5.0 was used to genotype TRs from PacBio HiFi reads using non-default parameters --karyotype XY, --flank-len 25, and --max-depth 10,000. For Mendelian consistency analysis on autosomal chromosomes, TRGT was run separately on each

sample. We ran both LongTR and TRGT on the same computer clusters running Rocky Linux 9. HPC hardware specification is as follows: CPU: (Intel(R) Xeon(R)) Platinum 8358 CPU @ 2.60 GHz) with 1 TB RAM. All analyses were performed on a single core with 4 GB memory. Each experiment was run 5 times and the mean value was reported. Timing was performed with the UNIX time command and the sum of the sys and user times was reported.

Benchmarking against Straglr

We ran Straglr v1.5.0 using the non-default parameters --min_support 4, --genotype_in_size, --min_str_len 1, and --max_str_len 1000 to genotype repeats with at least 4 overlapping reads and report genotypes in terms of allele sizes instead of copy numbers, genotype homopolymers, and genotype repeats with motif length up to 1000bp.

Benchmarking against Tandem-genotypes

Input sequencing data should be aligned using LAST [28] prior to TR genotyping with Tandem-genotypes. The aligned BAM file for HG002 was converted to fastq format using samtools [26] and then realigned using LAST v1542 to the GRCh38 reference genome as described in the Tandem-genotypes documentation. Tandem-genotypes v1.9.1 was then run using non-default parameters -o2 and -u1 to report 2 alleles per repeat and genotype homopolymers.

Benchmarking against adVNTR

For evaluation of adVNTR and LongTR on VNTRs, adVNTR v1.5.0 was run with non-default parameters --accuracy-filter, --pacbio, and --log-pacbio-reads. The reference set of VNTRs was downloaded from the adVNTR GitHub (see Data Availability). Since adVNTR represents alternative alleles as integer multiples of the consensus repeat unit, direct allele length comparison was not possible. Therefore, we computed the concordance between LongTR and adVNTR copy number estimates allowing for one copy number difference to accommodate the complex VNTRs consisting of multiple motifs with different sequences. This analysis was done on autosomal chromosomes.

Evaluation on ONT reads

Oxford Nanopore Duplex data for HG002 (Data Availability) was aligned to GRCh38 using minimap2 v2.26-r1175. We used WhatsHap [29] v2.2 to haplotag sequencing data with small variants called by DeepVariant [30] as input (Data Availability). The Simplex SUP dataset was already aligned and phased (Data Availability). LongTR was then run on the Adotto reference restricting to autosomal chromosomes.

Comparison to short read STR calls

HipSTR v0.7 was used to genotype STRs from 250bp paired-end PCR-free Illumina reads for HG002 (Data Availability) with non-default parameters --min-reads 4, --def-stutter-model, and --haploid-chrs chrY,chrX to genotype repeats with at least 4 overlapping reads, use default values for the stutter error model, and to consider chromosome Y and X as haploid chromosomes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03319-2.

Additional file 1. Contains supplementary figures S1-S7, tables S1-S2 and their legends.

Additional file 2. Review history.

Acknowledgements
Certain commercial equipment, instruments, or materials are identified to specify adequately experimental conditions or reported results. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology nor does it imply that the equipment, instruments, or materials identified are necessarily the best available for the purpose. We thank Eric Mendenhall for helpful comments on the manuscript.

Review history
The review history is available as Additional file 2.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
H.Z.J. developed LongTR, led analyses, and wrote the manuscript. J.M.Z. helped perform comparison of LongTR and TRGT genotypes to the HG002 assembly and helped supervise ONT analyses. S.J. and J.P. helped with adVNTR analyses. A.S. helped with testing and development of LongTR. M.G. supervised development of LongTR and analyses and wrote the manuscript.

Funding
This work was supported in part by NIH grants 1R01HG010149 (M.G.) and U01DA051234 (M.G.).

Availability of data and materials
LongTR is implemented as a C++ package and is accessible on GitHub (https://github.com/gymrek-lab/LongTR) [31] under GNU General Public License v2.0. An archived version of LongTR is available on Zenodo (https://doi.org/10.5281/zenodo.11403979) [32].
Project Adotto Tandem-Repeat Regions and Annotations [33]: https://zenodo.org/record/8329210/files/adotto_repeats.hg38.bed.gz?download=1
PacBio HiFi reads aligned to GRCh38 for HG002, HG003, and HG004 [34]:
https://downloads.pacbcloud.com/public/revio/2022Q4/
PacBio HiFi reads for patients with pathogenic expansions and metadata [34]:
https://downloads.pacbcloud.com/public/dataset/RepeatExpansionDisorders_NoAmp.
HG002 assembly [10]:
https://s3-us-west-2.amazonaws.com/human-pangenomics/T2T/HG002/assemblies/hg002v1.0.1.fasta.gz
adVNTR reference set of repeats [17]:
https://drive.google.com/file/d/1DetpBQySPNe2YAJa4FsjHn9qiRNS3wEV/view
Oxford Nanopore Duplex data for HG002 [19]:
https://human-pangenomics.s3.amazonaws.com/index.html?prefix=submissions/0CB931D5-AE0C-4187-8BD8-B3A9C9BFDADE--UCSC_HG002_R1041_Duplex_Dorado/Dorado_v0.1.1/stereo_duplex/
Oxford Nanopore aligned and haplotagged Simplex SUP data for HG002 [35]:
https://42basepairs.com/download/s3/ont-open-data/giab_2023.05/analysis/variant_calling/hg002_sup_all/hg002.haplotagged.bam
DeepVariant SNP calls used by WhatsHap [34]:
https://downloads.pacbcloud.com/public/revio/2022Q4/HG002-rep3/analysis/HG002.m84005_220827_014912_s1.GRCh38.deepvariant.phased.vcf.gz
GIAB set of difficult to map regions [21]:
https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.3/GRCh38@all/Union/GRCh38_alllowmapandsegdupregions.bed.gz
Illumina reads for HG002 [36]:
https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/HG002_NA24385_son/NIST_Illumina_2x250bps/novoalign_bams/.
HipSTR hg38 reference TR set [3]:
https://github.com/HipSTR-Tool/HipSTR-references/raw/master/human/hg38.hipstr_reference.bed.gz

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Samples HG002, HG003, and HG004 are part of the Personal Genome Project collection and have been consented for commercial use and for public posting of Personally Identifying Genetic Information (PIGI), which allows open-access, public posting of extensive genetic data.

Competing interests
The authors declare that they have no competing interests.

Received: 30 January 2024 Accepted: 21 June 2024
Published online: 04 July 2024

References

1. Ziaei Jam H, Li Y, DeVito R, Mousavi N, Ma N, Lujumba I, et al. A deep population reference panel of tandem repeat variation. Nat Commun. 2023;14:6711.
2. Hannan AJ. Tandem repeats mediating genetic plasticity in health and disease. Nat Rev Genet. 2018;19:286–98.
3. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods. 2017. https://doi.org/10.1038/nmeth.4267.
4. Kristmundsdottir S, Eggertsson HP, Arnadottir GA, Halldorsson BV. popSTR2 enables clinical and population-scale genotyping of microsatellites. Bioinformatics. 2020;36:2269–71.
5. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res. 2019;47:e90.
6. Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Narzisi G, et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. Genome Res. 2017;27:1895–903.
7. Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37:1155–62.
8. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol. 2016;17:239.
9. English AC, Dolzhenko E, Ziaei Jam H, McKenzie SK, Olson ND, De Coster W, et al. Analysis and benchmarking of small and large genomic variants across tandem repeats. Nat Biotechnol. 2024. https://doi.org/10.1038/s41587-024-02225-z.
10. Rhie A, Nurk S, Cechova M, Hoyt SJ, Taylor DJ, Altemose N, et al. The complete sequence of a human Y chromosome. Nature. 2023;621:344–54.
11. Ren J, Gu B, Chaisson MJP. Vamos: Variable-number tandem repeats annotation using efficient motif sets. Genome Biol. 2023;24:175.
12. Bolognini D, Magi A, Benes V, Korbel JO, Rausch T. TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data. Gigascience 2020;9. https://doi.org/10.1093/gigascience/giaa101.
13. Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. Genome Biol. 2019;20:58.
14. Readman C, Indhu-Shree R-B, Jan MF, Inanc B. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. Genome Biol. 2021;22:224.
15. Dolzhenko E, English A, Dashnow H, De Sena BG, Mokveld T, Rowell WJ, et al. Characterization and visualization of tandem repeats at genome scale. Nat Biotechnol. 2024. https://doi.org/10.1038/s41587-023-02057-3.
16. Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, et al. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 2020;30:1291–305.
17. Bakhtiari M, Park J, Ding Y-C, Shleizer-Burko S, Neuhausen SL, Halldórsson BV, et al. Variable number tandem repeats mediate the expression of proximal genes. Nat Commun. 2021;12:2075.
18. Park J, Kaufman E, Valdmanis PN, Bafna V. TRviz: a Python library for decomposing and visualizing tandem repeat sequences. Bioinform Adv. 2023;3:vbad058.
19. Koren S, Bao Z, Guarracino A, Ou S, Goodwin S, Jenike KM, et al. Gapless assembly of complete human and plant chromosomes using only nanopore sequencing. BioRxivorg. 2024. https://doi.org/10.1101/2024.03.15.585294.
20. IGV: Integrative genomics viewer n.d. https://www.igv.org/ (Accessed 2 Jan 2024).
21. Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, et al. PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions. Cell Genom 2022;2. https://doi.org/10.1016/j.xgen.2022.100129.
22. Oxford Nanopore technologies. Oxford Nanopore Technologies n.d. https://nanoporetech.com/platform/accuracy (Accessed 7 Jan 2024).
23. PacBio revio. PacBio 2022. https://www.pacb.com/revio/ (Accessed 7 Jan 2024).
24. Lee C. Generating consensus sequences from partial order multiple sequence alignment graphs. Bioinformatics. 2003;19:999–1008.
25. Albers CA, Lunter G, MacArthur DG, McVean G, Ouwehand WH, Durbin R. Dindel: accurate indel calls from short-read data. Genome Res. 2011;21:961–73.
26. Bonfield JK, Marshall J, Danecek P, Li H, Ohan V, Whitwham A, et al. HTSlib: C library for reading/writing high-throughput sequencing data. Gigascience 2021;10. https://doi.org/10.1093/gigascience/giab007.
27. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
28. Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. Adaptive seeds tame genomic sequence comparison. Genome Res. 2011;21:487–93.
29. Martin M, Ebert P, Marschall T. Read-based phasing and analysis of phased variants with WhatsHap. Methods Mol Biol. 2023;2590:127–38.
30. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7.
31. Ziaei Jam H, Zook JM, Javadzadeh S, Park J, Sehgal A, Gymrek M. LongTR. GitHub 2023. https://github.com/gymrek-lab/LongTR (Accessed 2024).

32. Ziaei Jam H, Zook JM, Javadzadeh S, Park J, Sehgal A, Gymrek M. LongTR. Zenodo 2024. https://zenodo.org/doi/10.5281/zenodo.11403979 (Accessed 2024).
33. English A. Project adotto tandem-repeat regions and annotations 2022. https://doi.org/10.5281/ZENODO.6930201.
34. Datasets - PacBio - Highly accurate long-read sequencing. PacBio 2020. https://www.pacb.com/connect/datasets/ (Accessed 4 June 2024).
35. Oxford Nanopore Technologies. Sequencing Genome in a Bottle samples 2023. https://doi.org/10.5281/ZENODO.8363974.
36. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016;3:160025.

Publisher's Note