

A comprehensive tandem repeat catalog of the human genome

Readman Chiu^{1*}, Indhu-Shree Rajan-Babu², Inanc Birol^{1,2}, Jan M. Friedman^{2,4}

¹Canada's Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC V5Z 4S6, Canada

²Department of Medical Genetics, University of British Columbia, Vancouver, BC V5Z 4H4, Canada

⁴BC Children's Hospital Research Institute, Vancouver, BC V5Z 4H4, Canada

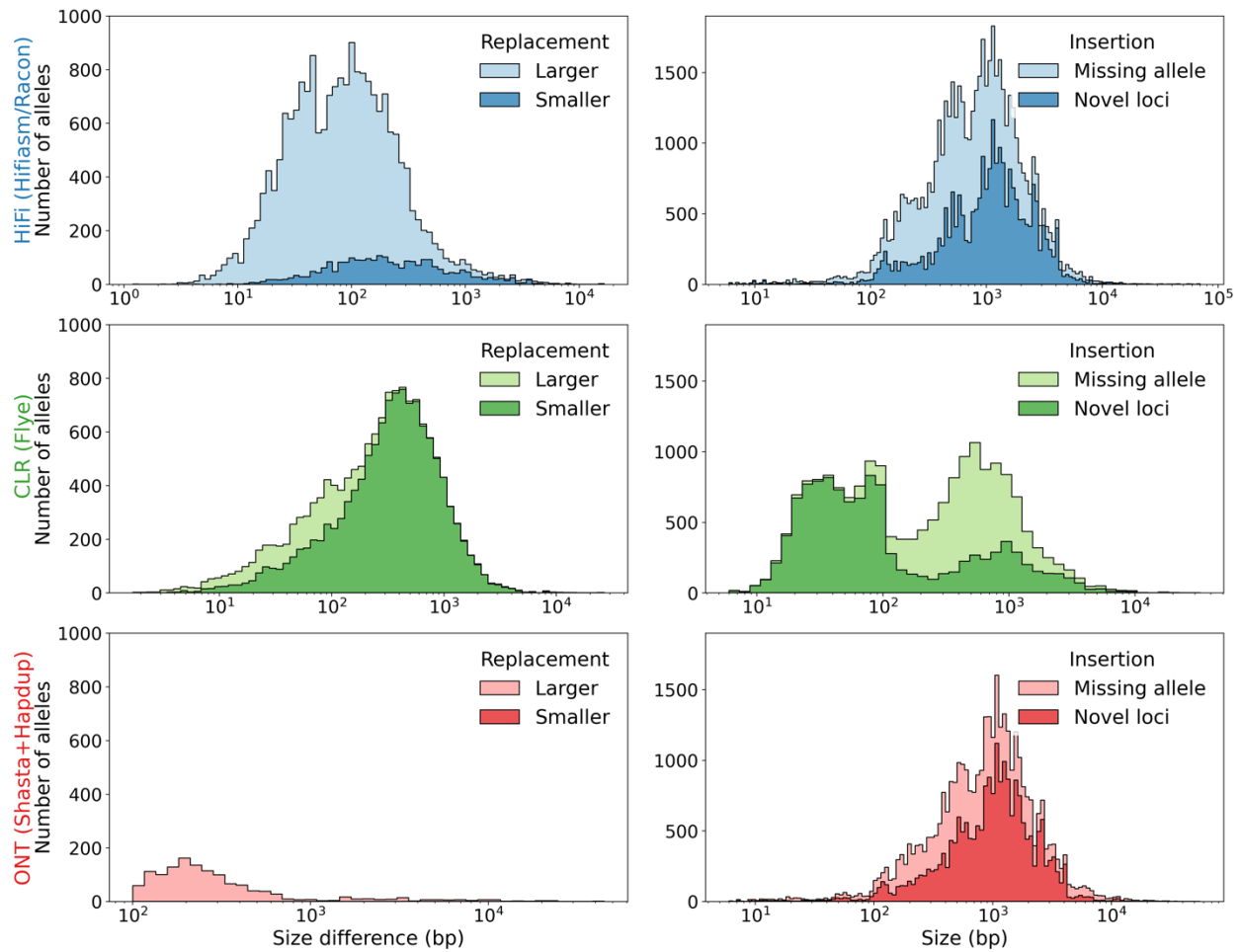
Supplementary Information

Table of Contents

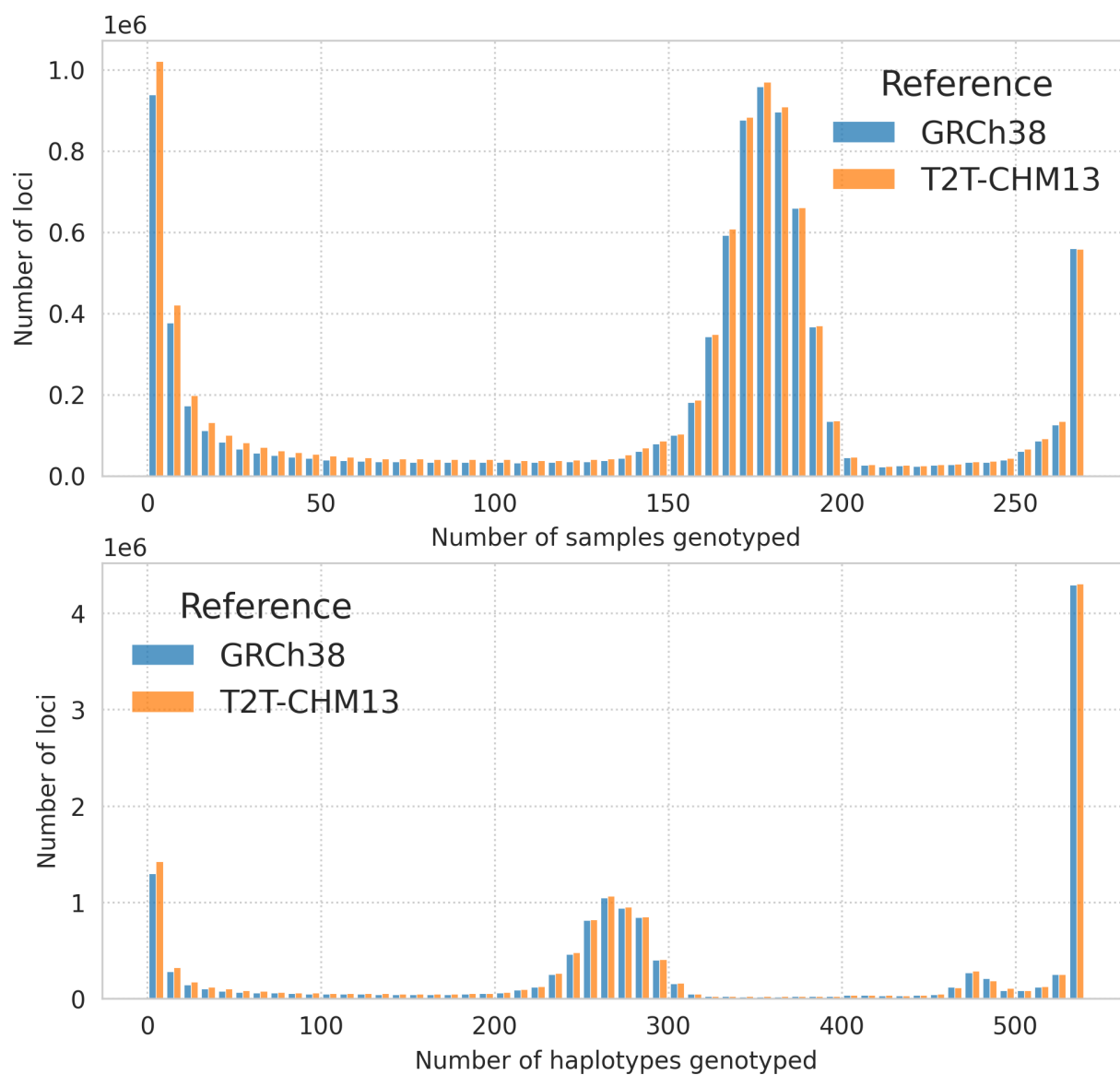
<i>Supplementary Figures</i>	3
Supplementary Figure 1. Size distributions of edited alleles performed on assembly-based genotypes.	3
Supplementary Figure 2. Number of genotypes collected per locus.	4
Supplementary Figure 3. Density and coverage of T2T-CHM13 catalog.	5
Supplementary Figure 4. Agreement between assembly-based and read-based genotypes.	6
Supplementary Figure 5. Correlation between assembly-based and TRGT genotypes of TRs with small size differences (20-100 bp) from GRCh38.	7
Supplementary Figure 6. Size variability of polymorphic trinucleotide repeats.....	8
Supplementary Figure 7. Comparison of size variability of TRs within and outside of subtelomeric regions.	9
Supplementary Figure 8. Size variability of two highly polymorphic subtelomeric loci.	10
Supplementary Figure 9. Multiple sequence alignment of translated amino sequences of distinct assembled STRs at chr9:35,906,548-35,906,626 in exon 1 of <i>HRCT1</i>	12
Supplementary Figure 10. Reconstruction of topmost polymorphic tandem repeats in coding sequences by assembler.....	13
Supplementary Figure 11. Phylogenetic analysis of the seven loci which exhibit association between TR size and motif sequence.	14
Supplementary Figure 12. Genotyping discrepancy at the (ATTCT) _n STR in ATXN10 in HG01123.	16
Supplementary Figure 13. Five unannotated STRs in genic loci with alleles larger than reference by at least 200 bp.	17
<i>Supplementary Tables</i>	18
Supplementary Table1. Online locations of the sequence data used in this study.	18

Supplementary Table 2. Average number of Straglr-based edits made to assembly-based genotypes per sample by sequencing technology and assembler.	18
--	----

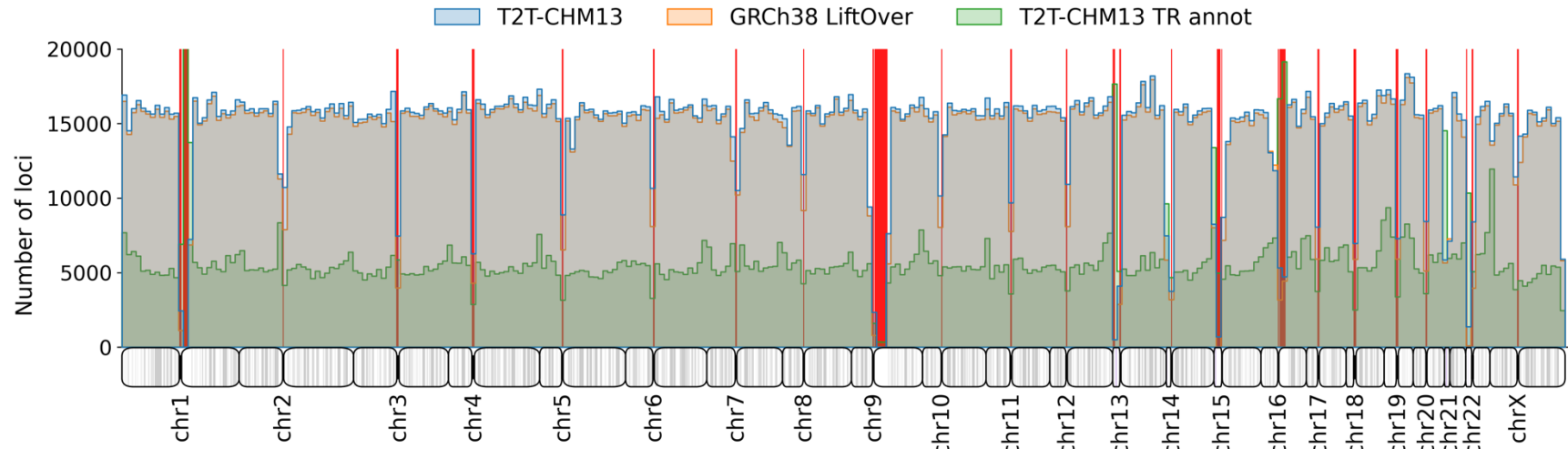
Supplementary Figures



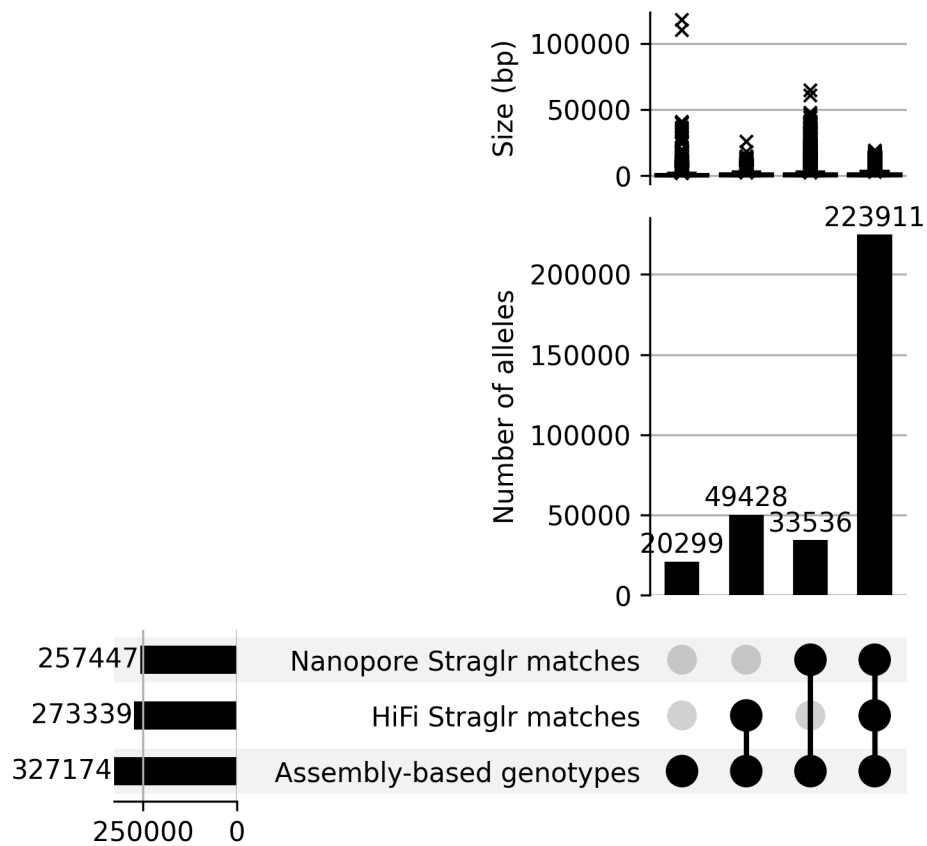
Supplementary Figure 1. Size distributions of edited alleles performed on assembly-based genotypes. Each row shows edits performed on data coming from different sequencing platforms with assembler software indicated (blue = PacBio HiFi; green = PacBio CLR; red = ONT). (Left) Alleles differentiated by whether replacements (Straglr-based) are larger (light-color) or smaller (dark-color) than original (assembly-based) genotypes. (Right) Inserted alleles (Straglr-based genotypes) composed of missing assembly haplotypes (light-color) or loci (dark-color).



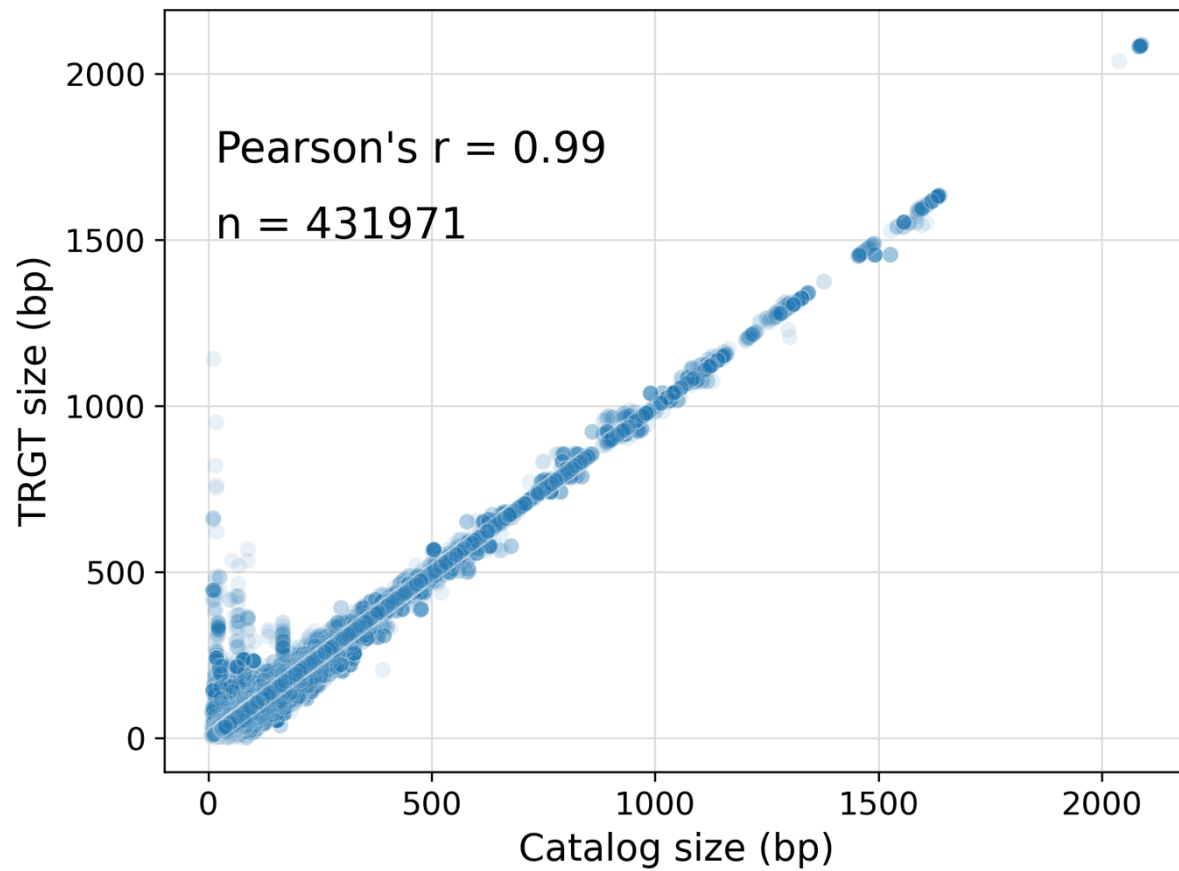
Supplementary Figure 2. Number of genotypes collected per locus. Number of loci (Y-axis) with genotypes from X(-axis) number of samples (top) or haplotypes (bottom).



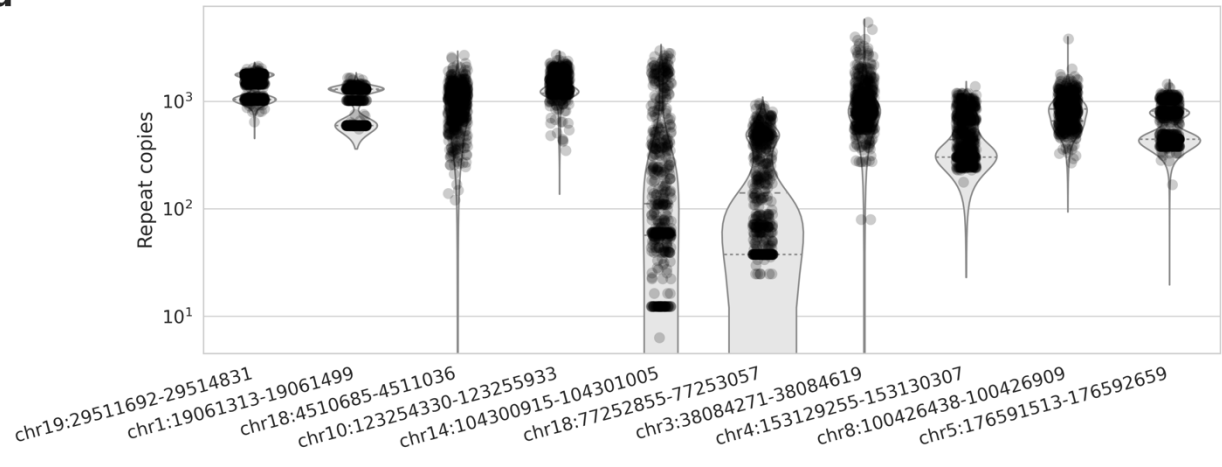
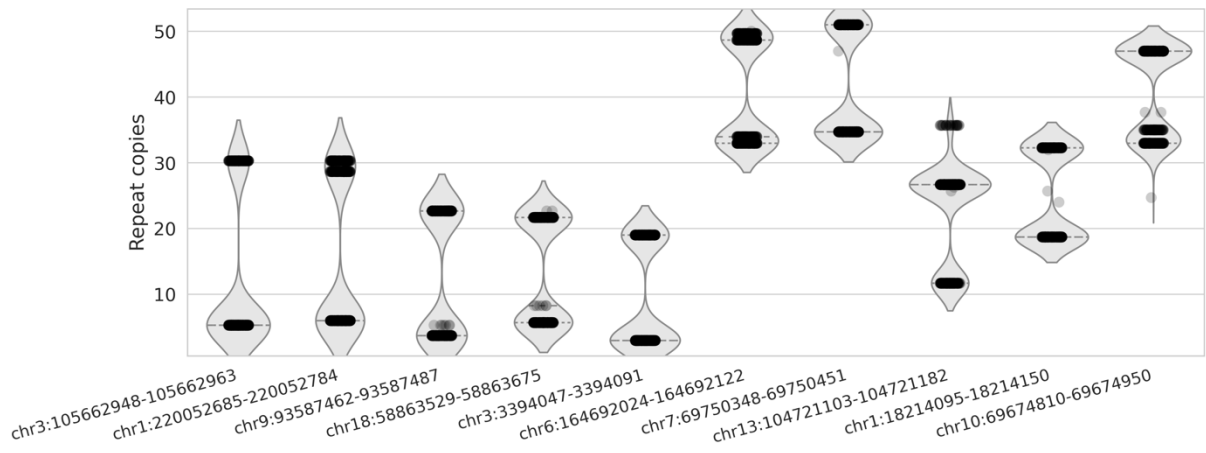
Supplementary Figure 3. Density and coverage of T2T-CHM13 catalog. Red bars indicate T2T-CHM13 heavy repeat regions: regions $\geq 1\text{Mb}$ after merging TRs $\leq 100\text{bp}$ apart. Scale of X-axis: 1 unit = 10Mb



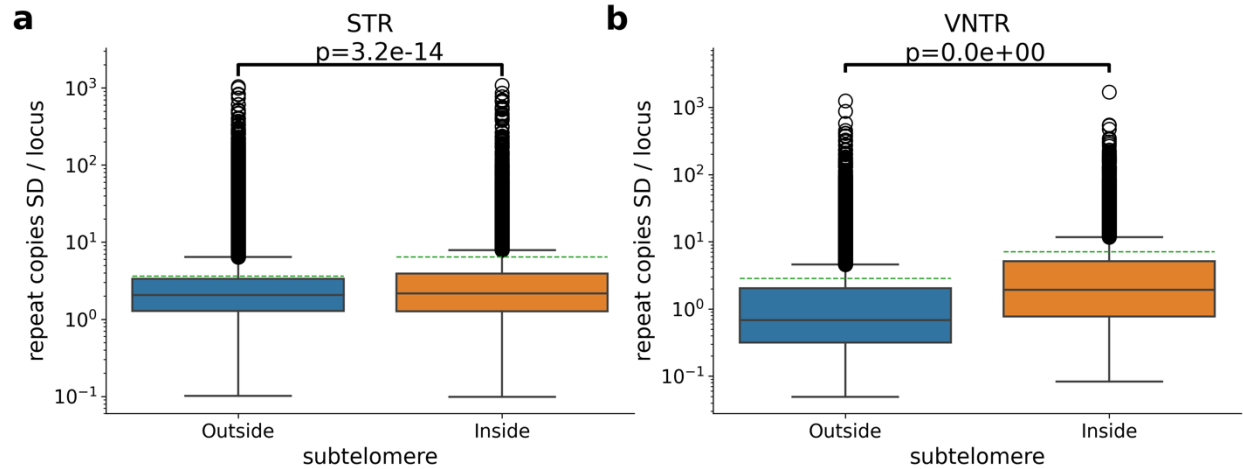
Supplementary Figure 4. Agreement between assembly-based and read-based genotypes. Overlaps (number of alleles) between TR genotypes in HPRC samples with Nanopore sequencing data reported by 1) Straglr on Nanopore sequencing reads; 2) Straglr on HiFi sequencing reads; and 3) assembly-based genotyping. (top) Number of alleles in each intersection; (bottom) Size (bp) distribution of alleles within each intersection below.



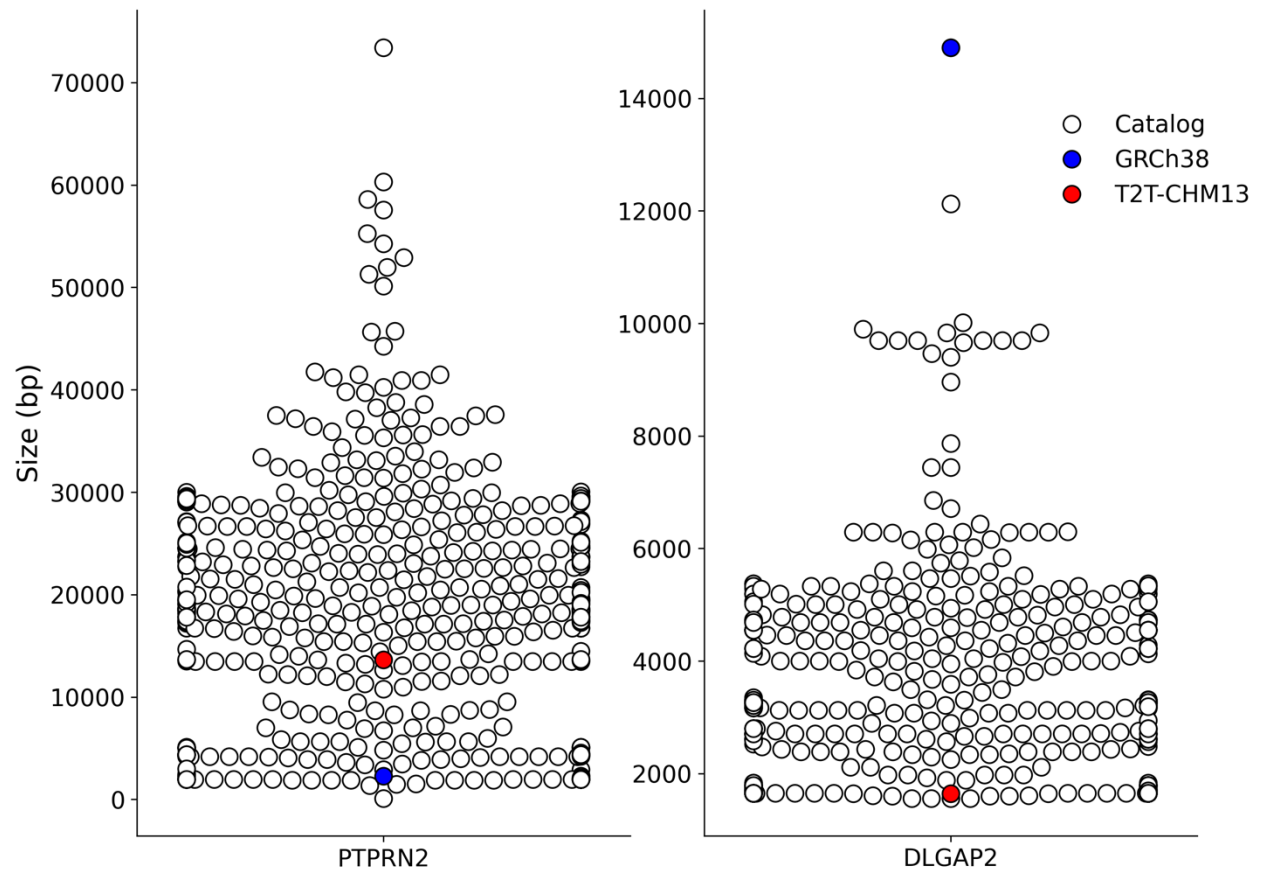
Supplementary Figure 5. Correlation between assembly-based and TRGT genotypes of TRs with small size differences (20-100 bp) from GRCh38.

a**b**

Supplementary Figure 6. Size variability of polymorphic trinucleotide repeats. a. Top-ten loci with highest IQR in repeat copies of all trinucleotide repeats in the catalog. Y-axis in logarithmic scale. b. Top-ten loci with highest IQR in repeat copies of trinucleotide repeats with ≤ 5 distinct repeat copies.



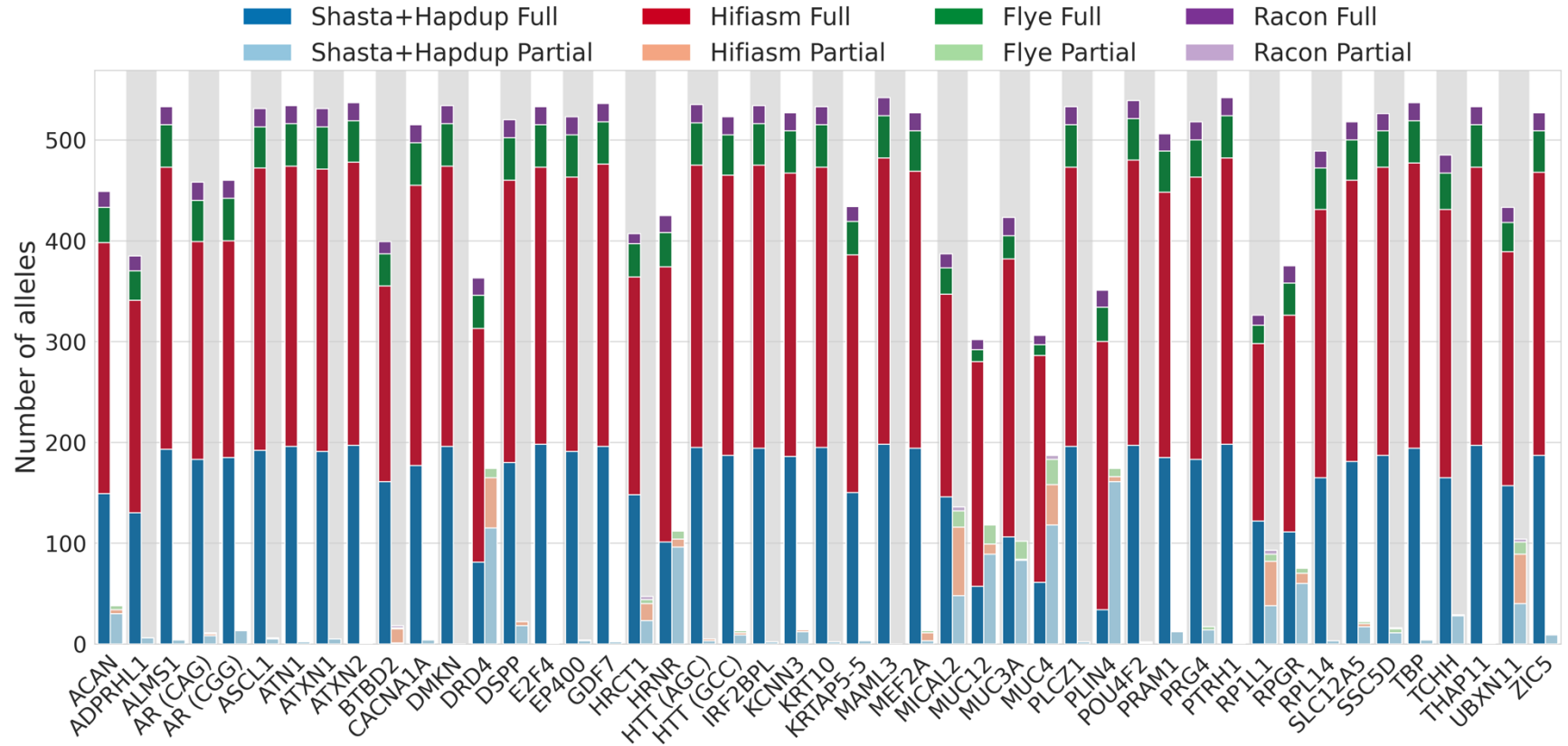
Supplementary Figure 7. Comparison of size variability of TRs within and outside of subtelomeric regions. Y-axis (logarithmic scale): standard deviation (SD) in repeat copies for each TR with at least ten distinct repeat copies. Boxplots show the medians (horizontal line within box), interquartile ranges (IQR, box), minimum and maximum values (whiskers) that are defined as 1.5x of IQR. Dotted green line: mean SD of distribution. Distributions are compared using the two-sided Mann-Whitney test with Bonferroni correction. P-values (p) are indicated in plots. **a.** STR. $N_{\text{TRs Outside}} = 137,264$; $N_{\text{TRs Inside}} = 10,216$; **b.** VNTR. $N_{\text{TRs Outside}} = 27,655$; $N_{\text{TRs Inside}} = 7,451$.



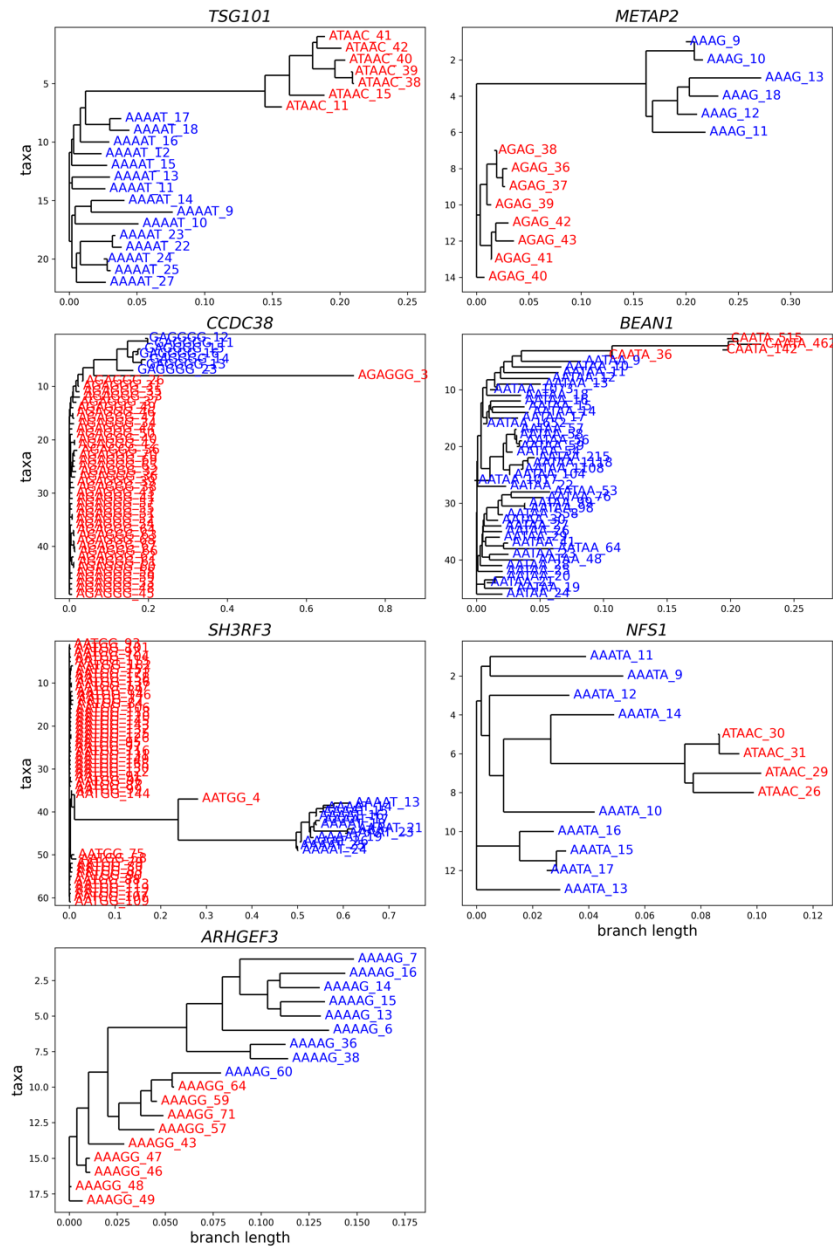
Supplementary Figure 8. Size variability of two highly polymorphic subtelomeric loci. Each open circle represents a haplotype allele. Sizes of the two loci, PTPRN2 (chr7:158,146,915-158,149,202) and DLGAP2 (chr8:1,393,013-1,407,915), in the reference genomes GRCh38 (blue) and T2T-CHM13 (red) are included to indicate the contrasting choice of allele sizes for incorporation into the reference sequences.



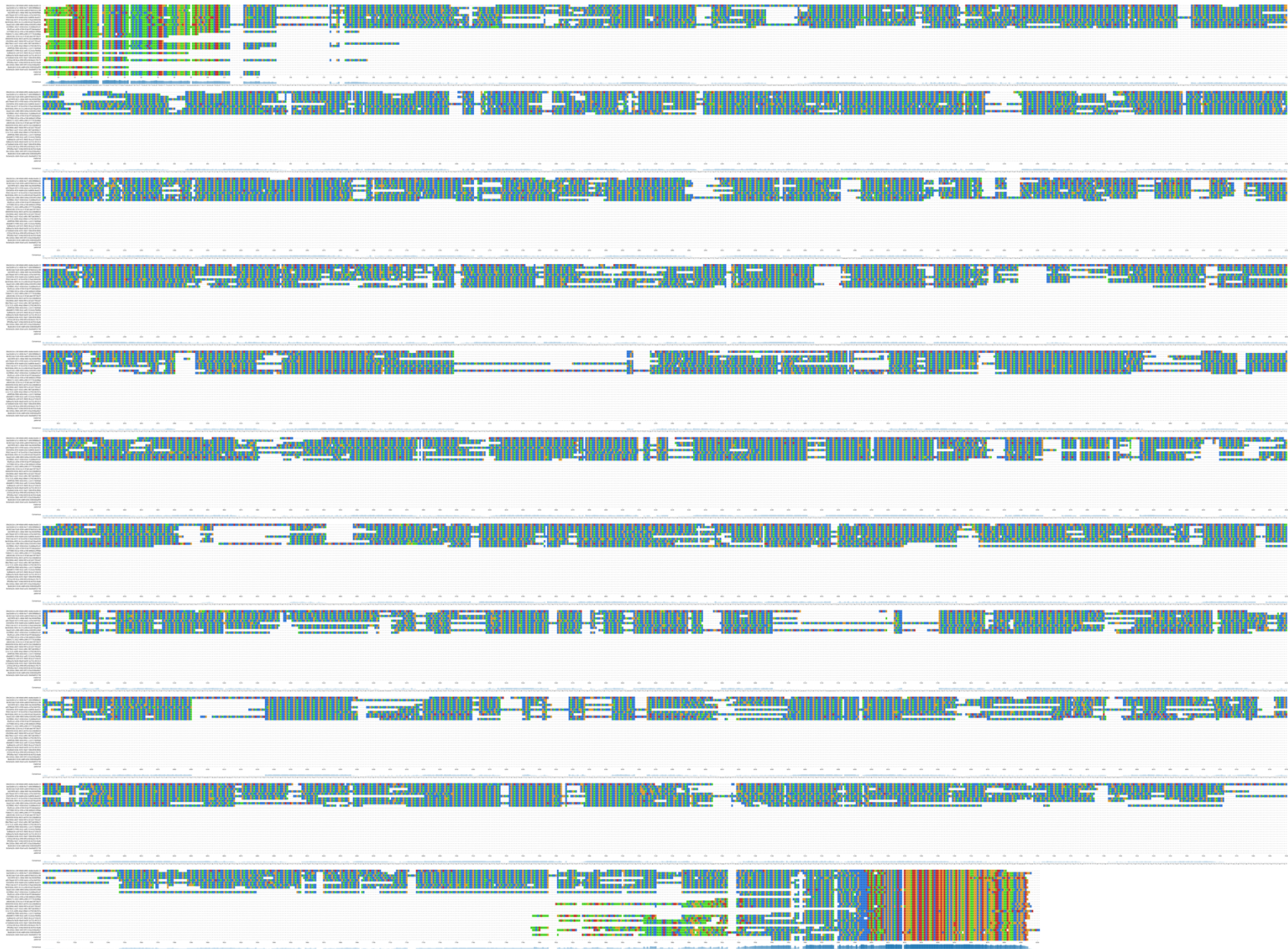
Supplementary Figure 9. Multiple sequence alignment of translated amino sequences of distinct assembled STRs at chr9:35,906,548-35,906,626 in exon 1 of *HRCT1*. Each sequence may be detected in multiple samples in our study cohort and only one representative (labeled by sample name and haplotype) was chosen for alignment.



Supplementary Figure 10. Reconstruction of topmost polymorphic tandem repeats in coding sequences by assembler. “Full” (i.e. maintenance of opening reads frame, left stacked bars of each gene) and “Partial” (right stacked bars of each gene) labels assigned to reconstruction results determined by whether or not premature stop codons was introduced to translated amino sequence and hence leading to mismatch of downstream sequence.



Supplementary Figure 11. Phylogenetic analysis of the seven loci which exhibit association between TR size and motif sequence. Red = long alleles. Blue = short alleles. Each allele is labeled with the motif sequence followed by the repeat copies in that allele.



Supplementary Figure 12. Genotyping discrepancy at the (ATTCT)_n STR in ATXN10 in HG01123. Multiple sequence alignment of Nanopore reads (labeled by read names) supporting a ~6.2kb allele (top 10 rows), a 65 bp allele (middle 18 rows), and the parental alleles (last two rows, labeled “maternal” and “paternal”) assembled by trio Hifiasm in the HPRC assembly of HG01123. The ATTCT repeat tract is manifested as mostly blue (color of nucleotide T) with specks of green (color of nucleotide A) and orange (color of nucleotide C). 100 bp of flanking sequences are included as shown on both ends of the alignments.

Supplementary Tables

Consortium	Location
HPRC	https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=working/
HGSVC2	https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/release/v1.0/
1000G	
ONT	https://s3.amazonaws.com/1000g-ont/index.html?prefix=FIRST_100_FREEZE/

Supplementary Table1. Online locations of the sequence data used in this study.

Sequencing Technology	Assembler	GRCh38		T2T-CHM13	
		Edits	Novel loci	Edits	Novel loci
PacBio HiFi	Hifiasm (HPRC), Racon (HGSVC2)	298	83	236	30
PacBio CLR	Flye	1,043	264	639	209
Nanopore	Shasta+Hadup	156	106	117	58

Supplementary Table 2. Average number of Straglr-based edits made to assembly-based genotypes per sample by sequencing technology and assembler. “Edits” = sum of corrections and missing haplotypes. “Novel loci” = loci from Straglr results that do not overlap loci from assembly-based genotypes.