nature biotechnology

Article

https://doi.org/10.1038/s41587-023-02057-3

Characterization and visualization of tandem repeats at genome scale

Received: 11 May 2023

Accepted: 6 November 2023

Published online: 02 January 2024

Check for updates

Egor Dolzhenko[1,12], Adam English[2,12], Harriet Dashnow[3,12], Guilherme De Sena Brandine[1], Tom Mokveld[1], William J. Rowell[1], Caitlin Karniski[1], Zev Kronenberg[1], Matt C. Danzi[4], Warren A. Cheung[5], Chengpeng Bi[5], Emily Farrow[5], Aaron Wenger[1], Khi Pin Chua[1], Verónica Martínez-Cerdeño[6,7,8], Trevor D. Bartley[6,7], Peng Jin[9], David L. Nelson[10], Stephan Zuchner[4], Tomi Pastinen[5], Aaron R. Quinlan[3], Fritz J. Sedlazeck[2,10,11,12] & Michael A. Eberle[1,12]

Tandem repeat (TR) variation is associated with gene expression changes and numerous rare monogenic diseases. Although long-read sequencing provides accurate full-length sequences and methylation of TRs, there is still a need for computational methods to profile TRs across the genome. Here we introduce the Tandem Repeat Genotyping Tool (TRGT) and an accompanying TR database. TRGT determines the consensus sequences and methylation levels of specified TRs from PacBio HiFi sequencing data. It also reports reads that support each repeat allele. These reads can be subsequently visualized with a companion TR visualization tool. Assessing 937,122 TRs, TRGT showed a Mendelian concordance of 98.38%, allowing a single repeat unit difference. In six samples with known repeat expansions, TRGT detected all expansions while also identifying methylation signals and mosaicism and providing finer repeat length resolution than existing methods. Additionally, we released a database with allele sequences and methylation levels for 937,122 TRs across 100 genomes.

Tandem repeats (TRs) are regions of the genome consisting of exact or near-exact repetitions of DNA sequence motifs. Many subtypes of TRs have been defined, including homopolymers (1 base pair (bp) motifs), short tandem repeats (STRs; 2–6-bp motifs) and variable number tandem repeats (VNTRs; >6-bp motifs). TRs contribute a substantial fraction of genetic variation in a typical human genome and are estimated to account for over 70% of structural variants (SVs) longer than 50 bp (ref. 1). TR expansions have been linked to over 50 monogenic disorders, such as Huntington's disease[2], amyotrophic lateral sclerosis[3] and Fragile X syndrome[4]. Lengths of many TRs are correlated with gene expression[5], and, recently, de novo TR expansions have been associated with cancer[6,7] and some neurodevelopmental and psychiatric disorders[8,9]. Furthermore, somatic mosaicism of TRs associated with rare disease can affect the age of onset, severity and progression of disease[10–12].

[1]Pacific Biosciences of California, Menlo Park, CA, USA. [2]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [3]Departments of Human Genetics and Biomedical Informatics, University of Utah, Salt Lake City, UT, USA. [4]Dr. John T. Macdonald Foundation Department of Human Genetics and John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL, USA. [5]Genomic Medicine Center, Children's Mercy Kansas City, Kansas City, MO, USA. [6]Institute for Pediatric Regenerative Medicine, Shriner's Hospital for Children and UC Davis School of Medicine, Sacramento, CA, USA. [7]Department of Pathology & Laboratory Medicine, UC Davis School of Medicine, Sacramento, CA, USA. [8]MIND Institute, UC Davis School of Medicine, Sacramento, CA, USA. [9]Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA. [10]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. [11]Department of Computer Science, Rice University, Houston, TX, USA. [12]These authors contributed equally: Egor Dolzhenko, Adam English, Harriet Dashnow, Fritz J. Sedlazeck, Michael A. Eberle. e-mail: meberle@pacificbiosciences.com

Despite this correlation between TR length and phenotype, TRs have been understudied owing to the difficulty in developing accurate, high-throughput, genome-wide assays[13]. Additionally, although many bi-allelic variants can be studied indirectly through linkage disequilibrium with single-nucleotide polymorphisms (SNPs), TRs are more likely to be missed in these studies because their hypervariability will tend to reduce the correlation with individual variants[14].

A variety of assays have been designed to profile different features of the repeat sequence, including length, specific sequence interruptions, methylation and mosaicism. Southern blot and polymerase chain reaction (PCR)-based assays enable a lower-throughput profiling of repeat lengths at a limited number of loci[15,16] and detection of repeat interruptions[17]. Recently, informatics methods have been developed to resolve some TRs in short-read sequencing data[18–25]. These methods make it possible to study repeats at the genome scale; however, they are less accurate when the repeat is near or larger than the length of the sequencing reads (typically 150 bp for short reads). Many known repeats are only pathogenic when their size reaches several hundreds of base pairs[26], meaning that short-read sequencing often cannot determine a pathogenic repeat's exact length and sequence composition. For example, it is not possible to use short-read sequences to reliably distinguish between premutations (165–600 bp) and full expansions (>600 bp) of the *FMR1* repeat[27]. In such cases, secondary orthogonal testing, such as repeat-primed PCR or Southern blot, is required to determine the length and, thus, pathogenicity of the repeat. This is a limitation when assessing an individual's genome for premutation or full mutation risk alleles.

Because of the length limitations and high structural complexity of many TR regions, many short-read STR or TR callers, such as GangSTR[22] and ExpansionHunter[20,23], focus on TRs that consist of nearly perfect stretches of motif copies. In contrast, long-read sequencing is particularly well suited for comprehensive repeat analysis because it can capture the entirety of the repeat sequence. However, computational methods for analysis of TRs in long reads must also cope with error patterns of the long-read sequencing technologies and the high structural complexity of repeat regions. Recently, a few long-read methods for TR analysis were introduced[28–30]. However, these tools focus only on a few loci or measure the overall repeat length, making them suitable only for profiling STRs consisting of nearly perfect repetitions of motif sequence. Thus, there is a need for general purpose methods for TR analysis capable of profiling both simple STRs and more complex VNTRs. In addition to the basic repeat length genotyping, a comprehensive analysis of TRs requires tools that can characterize repeat allele sequences as well as profile and visualize mosaicism. Mosaicism is an inherent feature of cancer-associated genome instability and certain pathogenic repeat expansions. These capabilities are necessary to fully explore the mechanisms and impact of TR mutations on disease phenotype.

The high accuracy of PacBio HiFi long-read sequencing makes it possible to comprehensively characterize both germline and somatic variation of TRs across the genome[31,32]. Furthermore, this technology enables CpG methylation profiling of TR regions, providing the potential to simultaneously assess genetic and epigenetic mutations of TR regions to reveal hidden biological patterns. In particular, the association between repeat length and methylation status can be leveraged to detect highly methylated pathogenic expansions. For example, individuals with reduced methylation of the *FMR1* repeat have been observed to have a reduced Fragile X syndrome phenotype[33].

Here we describe the Tandem Repeat Genotyping Tool (TRGT), a method for analysis of repeats in long-read data, as well as a companion method for Tandem Repeat Visualization (TRVZ). TRGT makes it possible to analyze structurally complex TRs that cannot be accurately represented by other available methods. It reports the full-length allele sequences and mean methylation levels of simple and compound TRs within each specified locus. Additionally, TRVZ affords a visual inspection of repeat alleles, which can be used for assessing any important repeat regions[34].

Results

Accurate TR variant calling from PacBio HiFi sequencing data

TRGT is designed for analysis of repeat alleles in HiFi sequencing data across a user-provided list of repeat regions. The input for TRGT consists of a BAM file with aligned HiFi reads and a list of repeat definitions (Fig. 1a). In brief, TRGT works by locating the repeat flanks in each read overlapping the repeat (Fig. 1b), clustering the reads to determine the consensus sequence for each repeat allele (Fig. 1b) and then using the repeat structure defined for each locus (Fig. 1c) to locate the boundaries of motif copies within each allele (Fig. 1d). Although the structure of simple repeats is defined by specifying the repeating motif, more complex repeats are defined by hidden Markov models (HMMs) (Fig. 1c), following previous work that demonstrated suitability of HMMs for modeling TRs[35]. The output of TRGT consists of a VCF file[36] with annotated repeat allele sequences (Fig. 1e) and their methylation levels. TRVZ is a companion tool to visualize the reads aligned to the repeat alleles (Fig. 1f) and can be used to determine the accuracy of the genotype results returned by TRGT.

To assess the accuracy of TRGT genotype calls, we used TRGT to genotype 937,122 repeat regions (Supplementary Fig. 1) spanning 122 Mbps of the reference genome[37] in 36× whole-genome HiFi sequencing data of HG002 from the Genome in a Bottle. We then compared the resulting repeat alleles to a recent state-of-the-art assembly of the same sample from the Telomere-to-Telomere (T2T) Consortium[38–40]. Compared to this assembly, 98.06% of the alleles either agreed exactly or had, at most, a single base pair difference (Fig. 2c). To further assess the accuracy of the genotypes, we calculated the Mendelian consistency of repeat lengths across the family trio consisting of HG002, HG003 and HG004 samples (Fig. 2a) where the missing genotypes were counted as errors. Overall, TRGT showed a Mendelian consistency rate of 89.00% across all repeats, and most of the errors (86.57%) corresponded to cases where the number of repeats differed by one between a parent and child (Fig. 2b). Ignoring such 'off-by-one' calls results in a Mendelian consistency rate of 98.38%. As expected, homopolymers and dinucleotide repeats were more error prone (82.95% exact and 98.38% off-by-one consistency) compared to repeats with longer motifs (97.29% exact and 98.75% off-by-one consistency). If we exclude all repeats genotyped as homozygous reference in all family members, the consistency rate drops to 85.62% (97.89% ignoring off-by-one errors). Some drop in accuracy is expected, of course, because all excluded repeats are consistent by definition. TRGT completed the analysis across these 30-fold sequencing coverage datasets in 35 min using 32 CPU cores.

In addition to TRGT, we evaluated two other methods used for profiling TRs in long-read sequencing data—tandem-genotypes[29] and Straglr[30]—and also one method designed for short reads, GangSTR[22] (which was evaluated on the corresponding short-read data). The same repeat catalog was used for all methods. Because these tools were designed to estimate repeat lengths and not repeat sequence or mosaicism, we assessed them by measuring length-based Mendelian consistency. Mendelian consistency was 86.44% (96.08% allowing for off-by-one calls), 70.05% (90.34%) and 63.89% (72.90%), respectively, for tandem-genotypes, Straglr and GangSTR compared to 89.00% (98.38%) for TRGT (Supplementary Fig. 2). Thus, TRGT offers an improvement in accurately measuring TR length. Furthermore, TRGT can assess sequence context, repeat methylation and mosaicism and facilitate repeat visualization via TRVZ.

Next, we assessed TRGT's ability to detect mosaic expansions where, instead of a single expanded allele, we observe reads supporting a distribution of allele sizes falling within a certain size range. For this analysis, we focused on the *FMR1* repeat region in the NA07537 sample, which was sequenced to nearly 500-fold HiFi read depth using the NoAmp targeted assay[41]. TRGT estimated that the length of the

https://doi.org/10.1038/s41587-023-02057-3

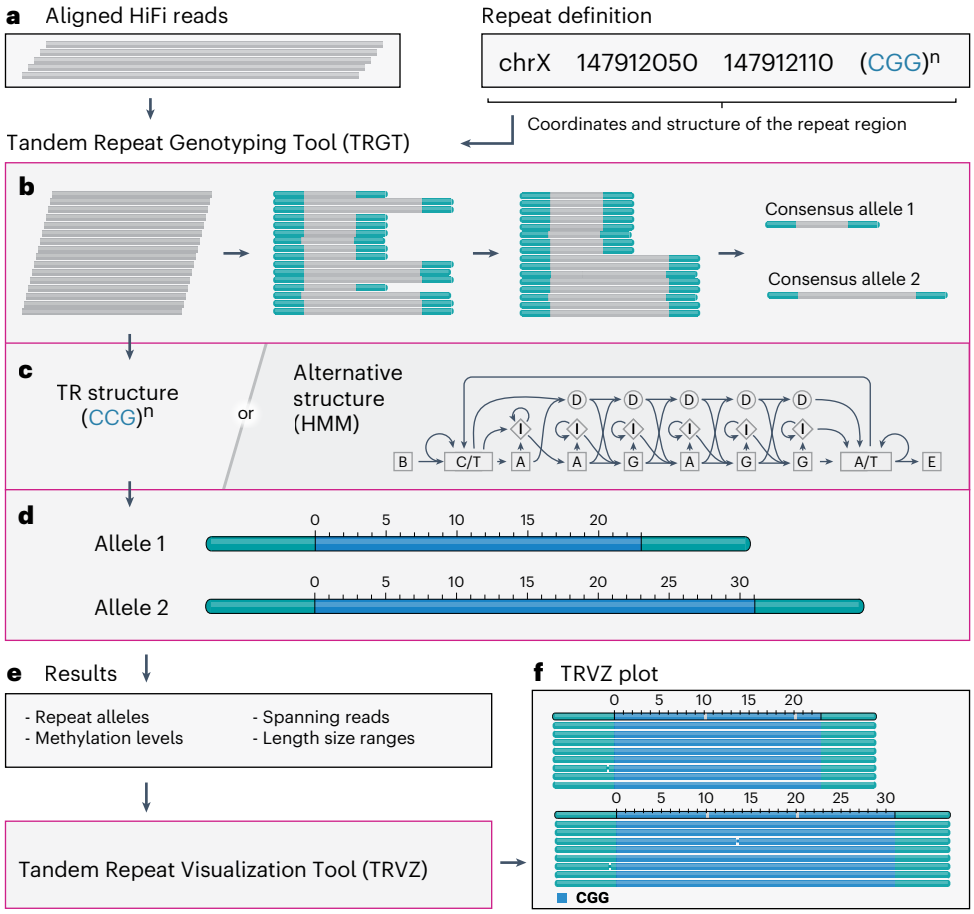

Fig. 1 | An overview of TRGT and TRVZ. a, Input to TRGT consists of HiFi reads and a list of repeat definitions. **b**, TRGT determines consensus repeat alleles. **c**, TRGT uses the pre-specified structure of the TR region to locate individual motif copies in each repeat allele. **d**, More complex repeat regions are specified with HMMs. **e**, Overview of key fields in TRGT's output. **f**, TRVZ generates plots that display repeat alleles and reads aligning to them, with optional methylation.

mosaic expansion ranges from 813 bp to 1,204 bp, which was concordant with the previous studies of this sample[42]. To assess TRGT's ability to accurately determine mosaicism at lower sequencing depths, we subsampled these data to depths ranging from 10-fold to 100-fold (100 replicates were performed at each depth). We then measured the proportion of the expanded alleles observed in the original sample captured by the corresponding TRGT's allele size interval. On average, over 75% of the expanded alleles were captured at 15-fold sequencing depth or higher, and, as expected, the confidence intervals were centered at 1,000 bp—the point estimate of the expansion size (Fig. 2d,e). Additionally, we characterized mosaicism across our genome-wide repeat catalog. This analysis revealed that 99.47% of repeat alleles show no or very little evidence of mosaicism (Supplementary Fig. 4) compared to the *FMR1* expansion. Furthermore, this *FMR1* expansion has a higher mosaicism score than 99.63% of repeat alleles that exceed it in length (Supplementary Fig. 4), indicating that mosaic expansions, such as the *FMR1* repeat in NA07537, are rare in most human genomes.

Population analysis of TRs

To study the genome and population-wide variability of the 937,122 TRs, we built the TRGTdb database (Methods) from a collection of 100 HiFi samples from the Human Pangenome Reference Consortium (HPRC) (Supplementary Table 1). We measured the length polymorphism of a repeat by computing the number of its alleles with distinct lengths per 100 samples (length polymorphism score). To reduce conflation between true alleles and technical artifacts (for example, one-off errors in homopolymer regions; Fig. 2b), we only used alleles appearing at frequency above 1% (that is, observed at least three times). We observed that 31.23% of the repeat loci showed no evidence of recurrent mutations (11.35% were mono-allelic and 19.88% were bi-allelic), whereas 68.77% were multi-allelic. Of the multi-allelic loci, 66.30% had 3–5 alleles, 26.87% had 6–10 alleles and 6.83% had more than 10 alleles (Fig. 3a). Additionally, we evaluated Hardy–Weinberg equilibrium (HWE) using the TRTools package[43] and found that over 99% of repeats passed previously used HWE thresholds[44] (Supplementary Fig. 6).

We investigated variations in the composition of the repeat sequences. For this, we compared the composition of two repeat alleles by calculating one minus the Jaccard index between the corresponding sets of high-frequency k-mers that we call the composition difference score (Methods). For example, the composition difference score (CDS) of repeat alleles $(CAG)_{10}$ and $(CAG)_{100}$ is 0.0 because of their identical composition despite the significant length difference. In contrast, although alleles $(CAG)_{10}$ and $(CAA)_{10}$ have the same length, their CDS is 1.0. To measure the degree of sequence composition polymorphism of a repeat, we calculate the mean of CDS for all pairs of repeat alleles. We refer to this value as the composition polymorphism score (CPS) of the repeat. The CPS was below 0.01 for over 98.75% of repeats, and only 0.09% of repeats had CPSs above 0.2 (Fig. 3b). This distribution indicates that TRs tend to have very homogenous sequence composition.

Given this collection of samples, we next characterized the variation of known pathogenic repeats. To compare the length and composition polymorphism of known pathogenic repeats in HPRC samples relative to our genome-wide repeat catalog, we calculated *z*-scores for the length and CPSs for 56 known pathogenic repeats relative to the

https://doi.org/10.1038/s41587-023-02057-3

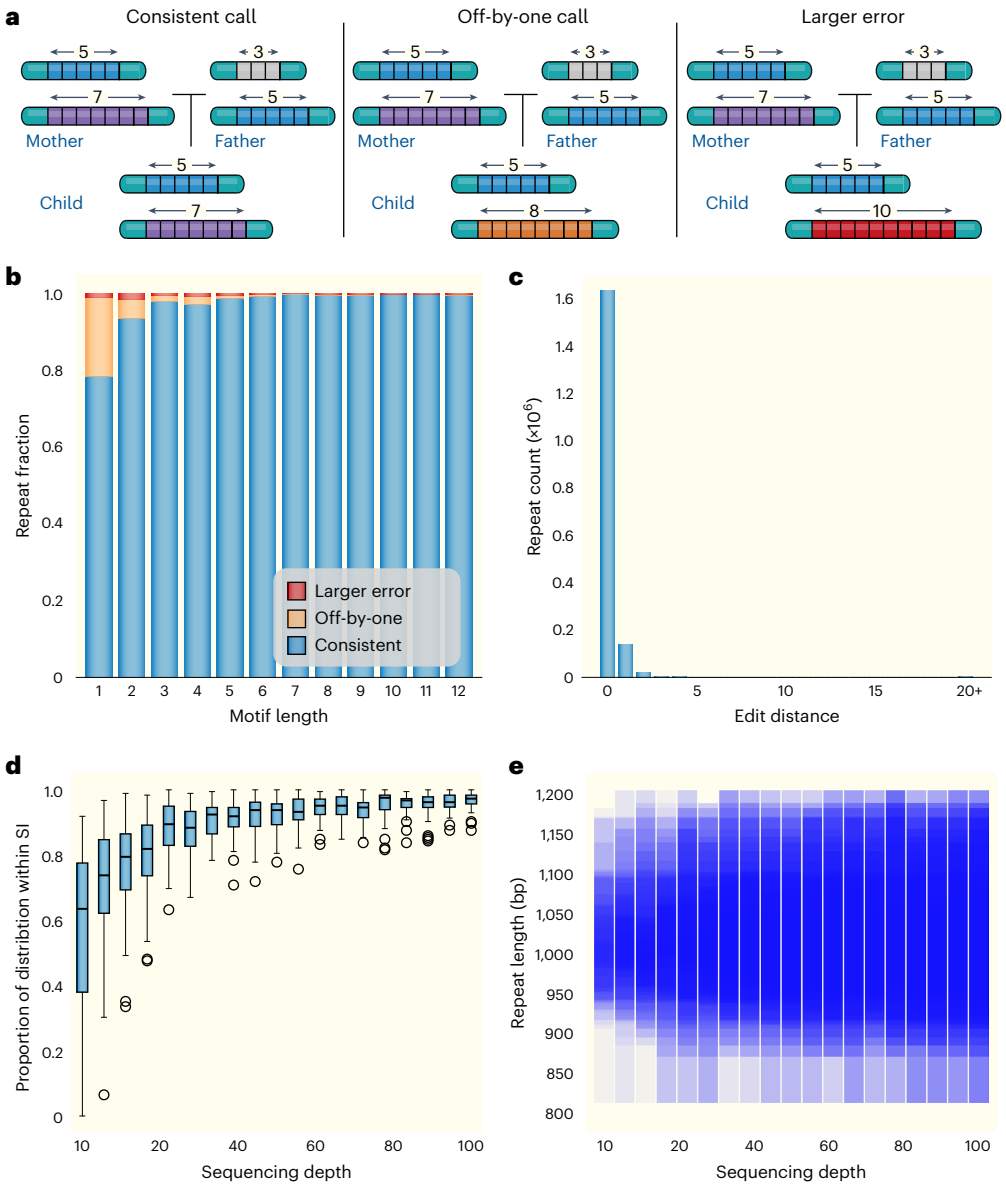

Fig. 2 | TRGT benchmarks. a, Examples of a consistent genotype, an off-by-one error and a larger error. **b**, A histogram stratifying the distribution of Mendelian errors by motif length. **c**, Edit distances between repeat alleles estimated by TRGT and an HG002 genome assembly. **d**, The proportion of the expanded *FMR1* repeat distribution captured by TRGT's size intervals from subsampled 500-fold depth NoAmp targeted sequence data (using $n = 100$ replicates for each depth) (the center line is at the median; the box extends from the first to the third quartile; and the whiskers extend to the farthest data point within 1.5× of the interquartile range from the box). **e**, Density of TRGT's size intervals.

corresponding genome-wide distributions (Fig. 3c). Consistent with our expectations, coding pathogenic repeats exhibited little polymorphism (Fig. 3c). In contrast, non-coding repeats tended to have higher length polymorphism compared to other repeats across the genome. Furthermore, *STARD7*, *YEATS2*, *RFC1* and *BEAN1* were the only pathogenic loci to exhibit substantial composition polymorphism. This is consistent with the fact that pathogenic expansions of these repeats correspond to changes in the motif sequence. Our analysis suggests that studies focused on identifying pathogenic expansions may prioritize non-coding repeats with polymorphic length or composition in addition to coding repeats.

We also profiled CpG methylation by using TRGT to estimate the mean methylation level of each repeat allele. The resulting distribution of methylation levels was consistent with the expected human genome methylation profile: CpG denser regions had markedly lower methylation compared to the CpG sparser regions (Fig. 3d).

We next focused our analysis on TR loci that overlap CpG islands and annotated each by their intersection with promoters[45,46]. In total, 9,821 TR loci overlap CpG islands, and 2,671 overlap promoters. The average methylation levels of 1,425,694 TR alleles overlapping CpG islands have a bimodal distribution (Fig. 3e). The lower peak of the distribution can be partially explained by CpG island TR alleles overlapping promoters. These findings confirm previous observations on the relationship among CpG islands, promoters and methylation[47,48]. When considering all TR alleles that fall within the top third of the average methylation range (corresponding to methylation levels between 0.68 and 1.0), we found that only 0.5% overlap with CpG islands. Moreover, we identified 2,315 alleles originating from 552 loci that overlap promoters and exhibit an average methylation level greater than 0.68. Among these 552 loci, 317 had more than two observed alleles with higher average methylation.

https://doi.org/10.1038/s41587-023-02057-3

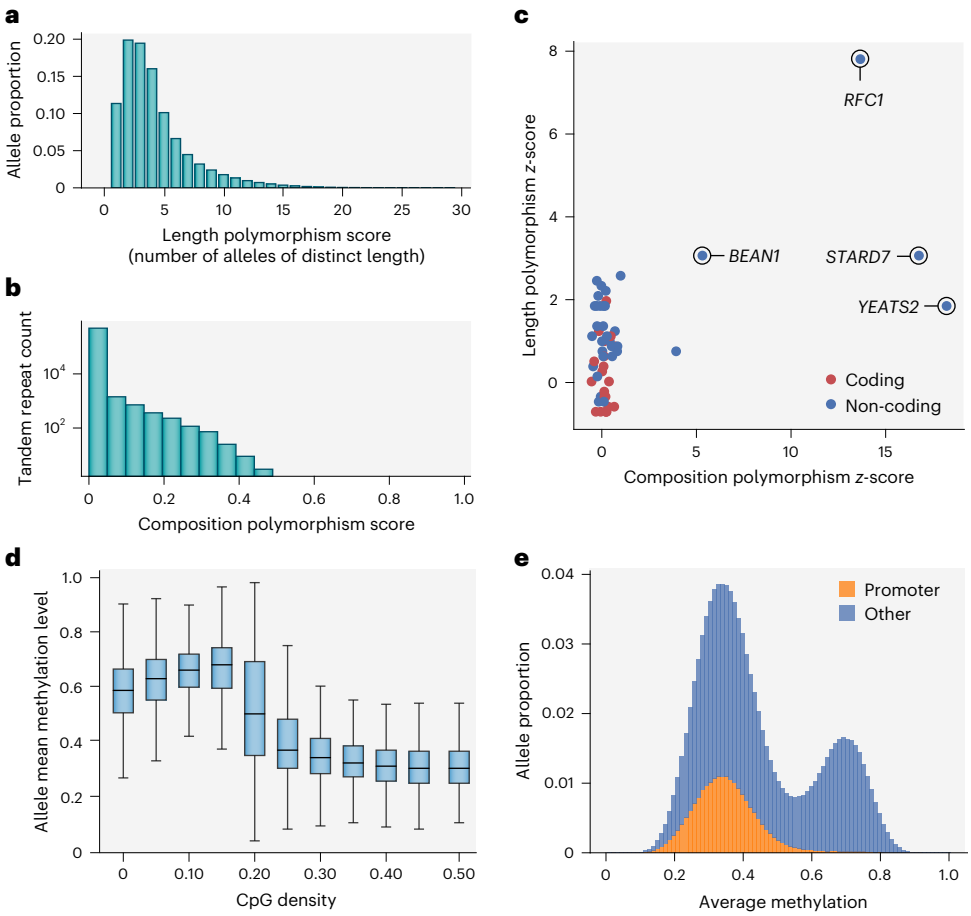

Fig. 3 | Genetic and epigenetic variation of n = 937,122 TR regions across 100 HPRC samples. a, Distribution of length polymorphism scores defined as the number of alleles of distinct length per 100 samples. **b**, Distribution of allele CPSs. **c**, Length and composition *z*-scores for known pathogenic repeats. **d**, Distribution of allele mean methylation levels stratified by CpG density (the center line is at the median; the box extends from the first to the third quartile; and the whiskers extend to the farthest data point within 1.5× of the interquartile range from the box). **e**, Mean methylation levels of TRs overlapping CpG islands.

We next analyzed the genomes of six individuals sequenced at Children's Mercy Kansas City who were previously identified to carry repeat expansions at known pathogenic loci in one of four genes: TRGT correctly identified the expansions in each sample, calling an *FMR1* 350 bp premutation and 1 Kbp full expansion (117 and 300 CGG motifs, respectively), a *DMPK* expansion spanning over 5 Kbp, two *STARD7* expansions each spanning over 1 Kbp and an *ATXN10* expansion >4 Kbp (Supplementary Table 2 and Supplementary Fig. 3). Compared to the previously applied testing that only sized the repeat to broad ranges, TRGT identified the size of the repeats to nearly base pair resolution.

Detailed characterization of *RFC1* repeat region

A repeat region within the *RFC1* gene located at chr4:39348424–39348479 (hg38) was recently associated with cerebellar ataxia, neuropathy and vestibular areflexia syndrome (CANVAS)[49,50]. Unlike most other pathogenic repeats, *RFC1* repeat alleles are known to have heterogeneous sequence composition consisting of stretches of AAAAG, AAAGG and other motifs. CANVAS itself has been linked to bi-allelic expansions consisting of either AAGGG or ACAGG motifs. Within TRGT, this repeat is described by an HMM whose topology is defined by the manually curated motif sequences (Fig. 4a). The HMM makes it possible to segment the sequence of each allele into a set of regions spanned by each motif. For example, the short allele of *RFC1* repeat in the HG04228 sample consists of a stretch of the AAAAG motif, whereas the long allele consists of a 700-bp stretch of the AAGGG motif (Fig. 4b). To investigate the population-level structure of *RFC1*, we used TRGT to analyze this repeat in the 100 HPRC samples. We first summarized the composition of each allele by computing the fraction of its sequence spanned by each constituent motif (Fig. 4c). This allowed us to group the alleles into five composition clusters (Fig. 4c,d). The alleles in each cluster are characterized by the presence of a relatively long stretch of one of the five motifs (AAAAG, AAGAG, AAAGGG, AAGGG and AAAGG). The largest cluster consisted of alleles composed of the AAAAG motif. The alleles within this cluster could be further subdivided into two groups: short alleles spanning fewer than 200 bp and longer expansions spanning more than 300 bp (Fig. 4e). Another cluster consisted of three alleles containing long stretches of the pathogenic AAGGG motif (Fig. 4c,d), which is consistent with the high carrier frequency of this expansion[49–52].

The *FMR1* repeat in HPRC samples and expansion carriers

We analyzed the CGG repeat in the promoter region of the *FMR1* gene associated with Fragile X syndrome[53]. *FMR1* alleles containing between 55 and 200 CGGs are called premutations and have been linked with Fragile X-associated ataxia syndrome and Fragile X-associated primary ovarian insufficiency[53]. Alleles with 200 or more CGGs are called full expansions and cause Fragile X syndrome. Full expansions have been associated with heavy CpG methylation as well as mosaicism, meaning that the exact length of the expanded repeat can vary across the cells[53]. In addition to the overall length, AGG interruptions within the repeat sequence have been associated with increased stability of the repeat and reduced risk of a parent with a premutation passing a full expansion to their child[54].

https://doi.org/10.1038/s41587-023-02057-3

Fig. 4 | Genetic variation of *RFC1* repeat alleles. a, An HMM representing the population structure of the *RFC1* TR derived from a priori known motifs. **b**, A TRVZ plot depicting both alleles of the *RFC1* repeat in the HG04228 sample. **c**, A heat map depicting the span of each motif (columns) on each allele (rows); each cluster of alleles is associated with the color of its dominant motif. **d**, An example allele from each cluster. **e**, Lengths of alleles belonging to each cluster.

The overall distribution of *FMR1* allele lengths (Fig. 5a) was similar to that reported previously, with a mean size of approximately 30 repeats[55]. Our analysis also identified two *FMR1* alleles of premutation length in HG04184 (58 motifs) and HG00438 (60 motifs) samples (Fig. 5b and Fig. 5c, respectively).

As most of the X chromosome in males is transcriptionally active, we expected *FMR1* methylation to be low. Indeed, most reads spanning this repeat in the HG01099 male sample were almost completely devoid of methylation (Fig. 5d,e). We observed the same low methylation pattern in all other male samples that we analyzed (Fig. 5f). Next, we analyzed the HG03831 sample derived from a female donor. Both alleles of this repeat spanned significantly fewer than 55 CGG copies and two AGG interruptions, conclusively indicating that this individual is not a carrier. We observed a bimodal methylation pattern of each allele consistent with chromosome X inactivation (Fig. 5g). Summarizing the distribution of median methylation levels for each read in this (Fig. 5h) and all other female samples (Fig. 5i) confirms the bimodal nature of *FMR1* methylation in females.

Finally, we analyzed the *FMR1* repeat in three brain samples previously determined to have the premutation (two samples) or full mutation (one sample). TRGT estimated the samples with expected premutations to have an *FMR1* repeat spanning 67 and 123 copies of the repeat motif (Fig. 5j,k), consistent with the established range for premutations. Interestingly, one of the premutation samples did not contain the stabilizing AGG interruptions (Fig. 5j), signaling an increased risk of transmitting a full expansion to children. The sample with expected full mutation was estimated to contain 325 motif copies and exhibited a strong degree of mosaicism, with repeat lengths ranging from 200 to 366 (Fig. 5l). Notably, all samples showed the expected methylation patterns (Fig. 5m). The female premutation sample exhibited bimodal methylation, whereas the male premutation sample was lowly methylated. In contrast, the male sample with the full expansion was highly

https://doi.org/10.1038/s41587-023-02057-3

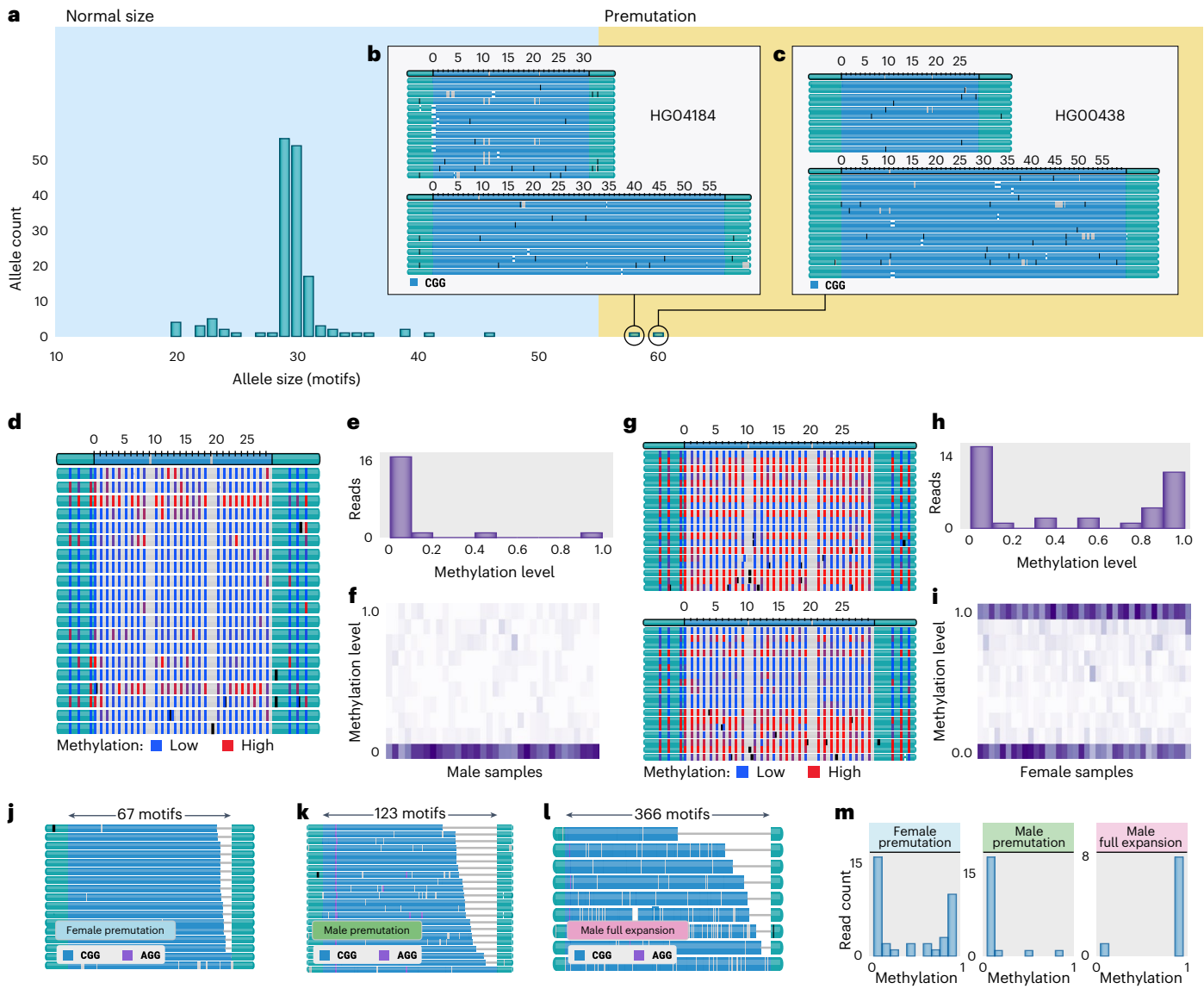

Fig. 5 | Genetic and epigenetic variation of *FMR1* repeat. a, Distribution of *FMR1* allele sizes in 100 HPRC samples. **b**,**c**, TRVZ plots of *FMR1* repeat in the HG04184 (**b**) and HG00438 (**c**) samples, respectively, showing premutation alleles. **d**, TRVZ plot of *FMR1* repeat in the HG01099 male sample displaying CpG methylation. **e**, Distribution of median methylation levels for HG01099 reads spanning *FMR1* repeat. **f**, Distributions of median methylation levels for *FMR1* reads across all male samples. **g**, TRVZ plot of *FMR1* repeat in HG03831 female sample displaying CpG methylation. **h**, Distribution of median methylation levels for HG03831 reads spanning *FMR1* repeat. **i**, Distributions of median methylation levels for *FMR1* reads across all female samples. **j**, Premutation repeat allele from a prefrontal cortex sample from a female donor (short allele not shown). **k**, Premutation repeat allele from a prefrontal cortex sample from a male donor. **l**, Fully expanded repeat allele from a prefrontal cortex sample from a male donor. **m**, Methylation profile of prefrontal cortex samples.

methylated, characteristic of Fragile X syndrome. These results demonstrate the utility of TRGT and TRVZ to accurately identify and visualize complex TRs alongside patterns of mosaicism and methylation across different samples and tissue sources.

Discussion

Here we describe a software tool, TRGT, to quantify TRs from HiFi sequencing data and demonstrate that it can accurately characterize both known pathogenic repeats and a genome-wide catalog of almost 1 million TRs. About 98.06% of the allele sequences generated by TRGT for the HG002 sample either agreed exactly or had, at most, a single base pair difference with the assembly. In addition to accurately genotyping TRs, we have included two companion methods to increase the utility of TRGT. TRVZ allows users to visualize the read-level evidence supporting the genotype calls made by TRGT, and TRGTdb builds a database of TRs that can be used to annotate sample-specific variant calls relative to a population. These companion methods will help researchers and clinical laboratories annotate and visually inspect the genotypes made by TRGT.

Compared to current methods for testing known pathogenic repeats, TRGT combined with HiFi reads can, as a single test, deliver many features that match or even surpass the performance of general wet-lab-based testing protocols. For example, TRGT estimates an exact motif count of a repeat. This is especially critical for assessing individuals carrying pathogenic repeats. For affected individuals, TRGT provided a count of the numbers of repeats and indicated the range of mosaicism. In contrast, other established sequencing or wet lab methods, such as repeat-primed PCR and Southern blot, merely provide size ranges or an average repeat length. Additionally, TRGT quantifies both the size and the motif sequence of repeats, which is

critical to interpreting loci such as *RFC1* and *SAMD12* (ref. 56). Finally, because TRGT also reports the average methylation from HiFi sequence reads, users can get the repeat length, sequence context and methylation status from a single sequencing experiment. Although TRGT can provide all of this information as a single test, it should be noted that, because it requires spanning reads, it may fail to call variants in low-sequencing-depth samples or regions. Furthermore, TRGT performance might be impacted by suboptimal read length. Efforts are currently underway to improve TRGT's ability to identify pathogenic repeat expansions with lower sequencing coverage or in cases when reads do not fully span the repeat, as can happen for particularly large expansions.

Most of the known repeats become pathogenic at sizes beyond what can be resolved with short reads alone[57]. For example, the sizes of pathogenic *FMR1* expansions (spanning over 200 repeats or 600 bp) are consistently underestimated even when using state-of-the-art short-read repeat callers[27]. Additionally, TRs with high mutation rates are unlikely to exhibit strong linkage disequilibrium with surrounding SNPs[14]. This means that SNP-based studies will be less likely to detect these TR risk alleles through association. Conversely, a genome-wide catalog of TRs genotyped with TRGT, possibly sequenced at lower depths, will greatly improve the power to detect TRs associated with complex traits. Because TRs may be more likely to have epistatic interactions, association studies that include accurate genotyping of all variants, including TRs, may help explain much of the missing heritability[13].

Although TRGT includes many features absent from other TR-specific variant callers, there are areas for continued development. Although we cataloged almost 1 million repeats across 100 samples, there is a need to extend this database to more TRs and include more samples of diverse ancestry. With a more complete and diverse database, we can perform a systematic analysis of repeat length and sequence context as well as methylation levels. This database can be leveraged to identify whether TRs in a sample are significantly expanded relative to the population in much the same way that frequency databases, such as gnomAD[58], are used to annotate SNPs or insertions and deletions (INDELs). We are continuing to extend our repeat catalog to include a more complete representation of all variable repeats, including ones that may not show up as repetitive in the current reference genomes. Furthermore, our analysis shows (Supplementary Fig. 7; see also Supplementary Fig. 5) that TRGT can reliably recover repeat alleles with length up to and exceeding 1 Kb at 10-fold sequencing depth. However, for analysis of known pathogenic repeats that exhibit significant mosaicism and expand to much longer lengths, we recommend depths of 20-fold at the minimum. Although TRGT relies on spanning reads and is, thus, limited to detecting repeat alleles up to 10 Kb in length, we are expecting to remove this limitation in future versions of the tool. TRGT has been specifically designed to work with PacBio HiFi data using algorithms and thresholds optimized for quality profile and length of HiFi reads.

Here we show that the tools TRGT, TRVZ and TRGTdb can highlight many important properties that are observed in known pathogenic TRs, including hypermethylation and variability in the repeat sequence. This demonstrates a significant advance in the tools available for unraveling the often under-explored complexity of TRs. Combined, these tools will enable researchers to gain novel insights into many aspects of the evolution, genetic diversity and medical implications of TRs.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-023-02057-3.

References

1. English, A. et al. Benchmarking of small and large variants across tandem repeats. Preprint at *bioRxiv* https://doi.org/10.1101/2023.10.29.564632 (2023).
2. Caron, N. S., Wright, G. E. B. & Hayden, M. R. Huntington disease. In GeneReviews® (eds. Adam, M. P. et al.) (Univ. Washington, 1998).
3. Siddique, N. & Siddique, T. Amyotrophic lateral sclerosis overview. In GeneReviews® (eds. Adam, M. P. et al.) (Univ. Washington, 2001).
4. Hunter, J. E., Berry-Kravis, E., Hipp, H. & Todd, P. K. *FMR1* disorders. In GeneReviews® (eds. Adam, M. P. et al.) (Univ. Washington, 1998).
5. Gymrek, M. et al. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48**, 22–29 (2016).
6. Erwin, G. S. et al. Recurrent repeat expansions in human cancer genomes. *Nature* **613**, 96–102 (2023).
7. Li, K., Luo, H., Huang, L., Luo, H. & Zhu, X. Microsatellite instability: a review of what the oncologist should know. *Cancer Cell Int.* **20**, 16 (2020).
8. Trost, B. et al. Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* **586**, 80–86 (2020).
9. Mojarad, B. A. et al. Genome-wide tandem repeat expansions contribute to schizophrenia risk. *Mol. Psychiatry* **27**, 3692–3698 (2022).
10. Morales, F. et al. Somatic instability of the expanded CTG triplet repeat in myotonic dystrophy type 1 is a heritable quantitative trait and modifier of disease severity. *Hum. Mol. Genet.* **21**, 3558–3567 (2012).
11. Morales, F. et al. Longitudinal increases in somatic mosaicism of the expanded CTG repeat in myotonic dystrophy type 1 are associated with variation in age-at-onset. *Hum. Mol. Genet.* **29**, 2496–2507 (2020).
12. Overend, G. et al. Allele length of the *DMPK* CTG repeat is a predictor of progressive myotonic dystrophy type 1 phenotypes. *Hum. Mol. Genet.* **28**, 2245–2254 (2019).
13. Press, M. O., Carlson, K. D. & Queitsch, C. The overdue promise of short tandem repeat variation for heritability. *Trends Genet.* **30**, 504–512 (2014).
14. Payseur, B. A., Place, M. & Weber, J. L. Linkage disequilibrium between STRPs and SNPs across the human genome. *Am. J. Hum. Genet.* **82**, 1039–1050 (2008).
15. Zhou, Y. et al. Robust fragile X $(CGG)_n$ genotype classification using a methylation specific triple PCR assay. *J. Med. Genet.* **41**, e45 (2004).
16. Tarleton, J. Detection of FMR1 trinucleotide repeat expansion mutations using Southern blot and PCR methodologies. In *Neurogenics: Methods and Protocols* (ed. Potter, N. T.) 29–39 (Springer, 2003).
17. Rajan-Babu, I. S., Law, H. Y., Yoon, C. S., Lee, C. G. & Chong, S. S. Simplified strategy for rapid first-line screening of fragile X syndrome: closed-tube triplet-primed PCR and amplicon melt peak analysis. *Expert Rev. Mol. Med.* **17**, e7 (2015).
18. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: a short tandem repeat profiler for personal genomes. *Genome Res.* **22**, 54–62 (2012).
19. Willems, T. et al. Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* **14**, 590–592 (2017).
20. Dolzhenko, E. et al. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27**, 1895–1903 (2017).
21. Dashnow, H. et al. STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.* **19**, 121 (2018).

22. Mousavi, N., Shleizer-Burko, S., Yanicky, R. & Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res.* **47**, e90 (2019).
23. Dolzhenko, E. et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).
24. Dolzhenko, E. et al. ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol.* **21**, 102 (2020).
25. Dashnow, H. et al. STRling: a k-mer counting approach that detects short tandem repeat expansions at known and novel loci. *Genome Biol.* **23**, 257 (2022).
26. Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19**, 286–298 (2018).
27. Ibañez, K. et al. Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study. *Lancet Neurol.* **21**, 234–245 (2022).
28. Giesselmann, P. et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nat. Biotechnol.* **37**, 1478–1481 (2019).
29. Mitsuhashi, S. et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. *Genome Biol.* **20**, 58 (2019).
30. Chiu, R., Rajan-Babu, I. S., Friedman, J. M. & Birol, I. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biol.* **22**, 224 (2021).
31. Wenger, A. M. et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.* **37**, 1155–1162 (2019).
32. Coster, W. D., De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing [Internet]. *Nat. Rev. Genet.* **22**, 572–587 (2021).
33. Oostra, B. A. & Willemsen, R. *FMR1*: a gene with three faces. *Biochim. Biophys. Acta* **1790**, 467–477 (2009).
34. Roy, S. et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines: a joint recommendation of the Association for Molecular Pathology and the College of American Pathologists. *J. Mol. Diagn.* **20**, 4–27 (2018).
35. Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V. & Bafna, V. Targeted genotyping of variable number tandem repeats with adVNTR. *Genome Res.* **28**, 1709–1719 (2018).
36. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics.* **27**, 2156–2158 (2011).
37. English, A. Project Adotto Tandem-Repeat Regions and Annotations. *Zenodo* https://doi.org/10.5281/zenodo.7013709 (2022).
38. Zook, J. M. et al. A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.* **38**, 1347–1355 (2020).
39. Wang, T. et al. The Human Pangenome Project: a global resource to map genomic diversity. *Nature* **604**, 437–446 (2022).
40. Rautiainen, M. et al. Telomere-to-telomere assembly of diploid chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
41. Tsai, Y. C. et al. Amplification-free, CRISPR–Cas9 targeted enrichment and SMRT sequencing of repeat-expansion disease causative genomic regions. Preprint at *bioRxiv* https://doi.org/10.1101/203919 (2017).
42. Grosso, V. et al. Characterization of *FMR1* repeat expansion and intragenic variants by indirect sequence capture. *Front. Genet.* **12**, 743230 (2021).
43. Mousavi, N. et al. TRTools: a toolkit for genome-wide analysis of tandem repeats. *Bioinformatics* **37**, 731–733 (2020).
44. Ziaei Jam, H. et al. A deep population reference panel of tandem repeat variation. *Nat. Commun.* **14**, 6711 (2023).
45. Dreos, R., Ambrosini, G., Cavin Périer, R. & Bucher, P. EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Res.* **41**, D157–D164 (2013).
46. Karolchik, D. et al. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493–D496 (2004).
47. Vavouri, T. & Lehner, B. Human genes with CpG island promoters have a distinct transcription-associated chromatin organization. *Genome Biol.* **13**, R110 (2012).
48. Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA* **99**, 3740–3745 (2002).
49. Rafehi, H. et al. Bioinformatics-based identification of expanded repeats: a non-reference intronic pentamer expansion in *RFC1* causes CANVAS. *Am. J. Hum. Genet.* **105**, 151–165 (2019).
50. Cortese, A. et al. Biallelic expansion of an intronic repeat in *RFC1* is a common cause of late-onset ataxia. *Nat. Genet.* **51**, 649–658 (2019).
51. Akçimen, F. et al. Investigation of the *RFC1* repeat expansion in a Canadian and a Brazilian ataxia cohort: identification of novel conformations. *Front. Genet.* **10**, 1219 (2019).
52. Fan, Y. et al. No biallelic intronic AAGGG repeat expansion in *RFC1* was found in patients with late-onset ataxia and MSA. *Parkinsonism Relat. Disord.* **73**, 1–2 (2020).
53. Hagerman, R. J. et al. Fragile X syndrome. *Nat. Rev. Dis. Primers* **3**, 17065 (2017).
54. Yrigollen, C. M. et al. AGG interruptions and maternal age affect *FMR1* CGG repeat allele stability during transmission. *J. Neurodev. Disord.* **6**, 24 (2014).
55. Huang, W. et al. Distribution of fragile X mental retardation 1 CGG repeat and flanking haplotypes in a large Chinese population. *Mol. Genet. Genomic Med.* **3**, 172–181 (2015).
56. Depienne, C. & Mandel, J. L. 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? *Am. J. Hum. Genet.* **108**, 764–785 (2021).
57. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet.* **17**, 507–572 (2016).
58. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

Methods

TRGT

TRGT performs TR genotyping using HiFi reads that overlap each repeat. The input to TRGT consists of a BAM file[59] with aligned HiFi reads and a file with repeat definitions. The output consists of a VCF file containing full-length repeat allele sequences and their methylation levels as well as a BAM file with portions of HiFi reads that span each repeat. Analysis of each repeat region proceeds as follows:

1. TRGT locates reads that span a given repeat region, and these reads are assigned to each allele by clustering. To cluster the reads, TRGT first calculates the edit distances between all pairs of reads and then performs agglomerative clustering using Ward linkage[60]. It then filters out any cluster containing fewer than 10% of the total number of spanning reads. For diploid repeats, TRGT assigns the two largest clusters to each allele. For haploid repeats, the largest cluster is assigned to the allele.
2. To determine the consensus sequence of each repeat allele, TRGT selects a read of the median length from the corresponding cluster of reads and uses it as the alignment backbone. All reads in the cluster are aligned against this backbone sequence. The consensus sequence is then determined by scanning the backbone and incorporating bases by performing a majority vote on the alignment operations. For example, if most read alignments contain a sequence insertion at some position of the backbone, then this sequence insertion is incorporated into the consensus.
3. TRGT next annotates occurrences of individual repeat motifs within the sequence of each consensus allele. Different annotation algorithms are used depending on the type of repeat. For example, simple TRs that can be described by repetitions of one or multiple fixed-sized motifs are annotated using a fast algorithm based on finding the longest path in an acyclic graph. More complex repeats are annotated using HMMs. These annotation methods are described below.
4. The methylation level of each repeat allele is set equal to the mean methylation level of all CpGs contained within the allele in all reads that support the allele.

Annotation of simple TR regions

We define 'simple TR regions' as regions whose population structure can be described as a series of TRs, possibly separated by interrupting sequences. The structure of such regions is described by an expression $(m_1)^{n_1}s_1(m_2)^{n_2}\cdots s_{k-1}(m_k)^{n_k}$ where m_i is the motif of i-th TR, n_i is the (allele-specific) motif count of i-th TR and s_i is a possibly empty sequence separating TRs i and $i+1$. Given a query allele sequence (Supplementary Fig. 8a), the segmentation algorithm proceeds as follows. First, we create a graph whose nodes correspond to matches between the query sequence and motifs m_i and interrupting sequences s_i (Supplementary Fig. 8b). Then, we create a directed edge from node m_i to node x if x is the next occurrence (in the topological order induced by the query sequence) of m_i, s_i or m_{i+1} (Supplementary Fig. 8c). Nodes s_i are connected using the same rule. We then determine a path that spans the largest number of bases. This path can be determined by calculating the longest path in a directed acyclic graph, which covers the largest number of bases terminating at each node (Supplementary Fig. 8d). This path corresponds to the segmentation of the original query sequence (Supplementary Fig. 8e,f).

Annotation of complex TR regions

Certain TR regions cannot be represented by the expressions introduced in the previous section. We call such regions complex. Following previous work[35], we use HMMs to model the structure of these repeats. TRGT can synthesize HMMs that model sequences that correspond to runs of a specified set of motifs. These runs can occur in an arbitrary order. *RFC1* (Fig. 4) is one example of such repeats. HMMs of this family all have similar topology (Supplementary Fig. 9): the customary start and end states (Supplementary Fig. 9a,b); a pair of silent states delineating the start and end of each motif run (Supplementary Fig. 9c,d); a pair of states delineating the start and end of each repeat motif (Supplementary Fig. 9e,f); and, finally, a block of states representing the motif occurrence sequence consisting of states corresponding to matches/mismatches and INDELs of motif bases. TRGT can also accommodate HMMs with other topologies. In this case, it requires that the HMM specification includes a list of edges that connect the terminal nodes of each motif as well as the sequence or label of each motif.

TRVZ tool

TRVZ is a companion visualization tool for TRGT allowing users to view selected repeats of interest. The input to TRVZ consists of files generated by TRGT. The output is an image in svg, pdf or png file formats. TRVZ generates a read pileup plot corresponding to each repeat allele (Supplementary Fig. 10). The top track of each allele plot shows the consensus sequence determined by TRGT (Supplementary Fig. 10a). The consensus is annotated according to its alignment to the perfect repeat of the same length. The solid color corresponds to matches, gray blocks to mismatches, horizontal lines to deletions and vertical lines to insertions in the allele sequence relative to the perfect repeat. For example, two AGG interruptions that are typically present in the sequence of non-expanded *FMR1* repeats will result in two mismatches (Supplementary Fig. 10) because this repeat is defined as $(CGG)^n$. The tracks below the top track correspond to the alignments of HiFi reads to each repeat allele (Supplementary Fig. 10b).

TRGT database

VCF entries are useful for representing variation but can be difficult to leverage for programmatic queries of the data. To normalize the data contained within VCFs, we consider each VCF entry to contain information that can be split into three tables: Locus, Allele and Sample. The Locus information corresponds to the VCF entry's CHROM and POS columns, which represent a location in the reference. The Allele information corresponds to the VCF entry's ALT, QUAL, FILTER and (generally) INFO columns, which represent variation observed at a Locus. The Sample information corresponds to the VCF entry's FORMAT and SAMPLE columns, which represent descriptions of Alleles observed in a sample at a Locus. VCF information is extracted and held in-memory as three Pandas DataFrames (one for each table) before being saved on-disk using Apache Parquet. Apache Parquet is an efficiently compressed, column-oriented file format. To store information across multiple runs of TRGT, all Loci and Alleles are consolidated into a single table and stored in their own Parquet file. However, each Sample is stored in its own Parquet file. These files are organized within a directory representing the database. By storing each table separately, the genotype information can be removed via deletion of Samples' Parquet files. Full de-identification can be achieved with a TRGTdb command for removing allele sequences, randomizing allele numbers or shuffling genotypes across samples. On average, a single sample from the 100-sample HPRC TRGTdb has an on-disk storage size of 11.4 Mb using TRGTdb compared to individual bgzip compressed VCFs requiring 92.3 Mb, an 87.6% decrease.

To assist users with creating a TRGTdb, command line tools are distributed as part of the TRGT package. Command line tools for 'standard' queries are included, such as allele counts, the number of monozygotic reference sites and per-locus genotype information. The outputs of these queries can be saved in tab-delimited, comma-separated, Parquet or Joblib formats. Finally, to assist users in creation of custom queries, a TRGTdb Python API is also distributed. Full documentation on the TRGTdb tool is available online[61]. Annotation of TR loci within the TRGTdb against UCSC genome tracks was performed using PyRanges[62]. All analyses performed with TRGTdb can be recreated by following the Jupyter Notebook tutorials[63].

Genome-wide repeat catalog

The genome-wide repeat catalog was derived from version 0.2 of the Adotto catalog[1,64]. The catalog was built from the consolidation of six sources of TR regions in GRCh38: Genome In a Bottle repeats[65]; SimpleRepeats track from the UCSC table browser[66]; Ensembl repeats[67]; adVNTR repeat database[68]; and a catalog of polymorphic repeats from Illumina[69]. Regions were filtered to those between 10 bp and 50 kbp in span before being expanded by ±25 bp and merging overlapping regions. The GRCh38 reference sequence spanned by the resulting set of putative TR regions was re-annotated with TandemRepeatFinder (TRF) version 4.09, and the motif from the longest-spanning annotation was added to the merged region's coordinates. Regions without a TRF annotation were removed from the set of TRs. Finally, 86 haplotype-resolved long-read assemblies were gathered from three projects[70–72] and aligned/variant called with minimap2/paftools[73]. We extracted regions containing non-SNP variants (INDELs ≥1 bp, SVs). The final catalog contains 937,122 TR regions spanning 122 Mbps.

TR benchmark

To assess TRGT's sensitivity to expanded pathogenic STR loci, we ran it on whole-genome sequencing (WGS) of six individuals with orthogonally confirmed clinical assays. These individuals were enrolled in the Genomic Answers for Kids program[74]. Samples were collected and sequenced on PacBio HiFi Sequel II and IIe systems as previously described[75]. Sex was inferred using Somalier[76] and then provided to TRGT using the –karyotype flag. TRGT version 0.5.0 was run at known pathogenic loci, using pathogenic_repeats.hg38.bed (commit b10e7f5). Expansions identified by TRGT were further visualized using TRVZ version 0.5.0 (Supplementary Fig. 3). Orthogonal clinical testing was performed by triplet-primed PCR or Southern blot as part of clinical care (Supplementary Table 2). The subsampling analysis was performed by randomly selecting reads from the original BAM file to achieve the desired depth and then applying TRGT/TRVZ to the resulting subsampled BAM file.

We compared TRGT calls to those made from a high-quality assembly. We compared TRs to the HG002 diploid genome assembly as follows: (1) we extracted sequences of all repeat alleles from the HG002 VCF file generated by TRGT; (2) we added a 250-bp flanking sequence to both sides of each allele (extracted from the GRCh38 reference genome) and mapped the resulting sequences to the paternal and maternal contigs of HG002 assembly with minimap2; and (3) we picked the top-scoring assignment of alleles to paternal contigs for each TR. The benchmarks used Straglr version 1.4.1 (ref. 30), GangSTR version 2.5.0 (ref. 22) and tandem-genotypes version 1.9.0 (ref. 29). TRGT was run with default parameters; tandem-genotypes was run with parameters '-o2–min-unit=1'; Straglr was run with parameters '–min_str_len 1–max_str_len 1000–max_num_clusters 2'; and GangSTR was run with default parameters. Mendelian consistency analysis was performed by genotyping the repeats in the HG002, HG003 and HG004 family trio with each method and then comparing the lengths of repeats in the child to those of their parents (Fig. 2a). Fractional lengths were rounded to the nearest integer.

TR composition analysis

To study the variation in sequence composition of TR alleles, we first defined the CDS that compares sequences of two alleles. Then, we used CDSs to define the CPS that measures the variation in sequence composition of a TR across a given set of samples. The CDS between alleles a_1 and a_2 is defined by:

$$CDS(a_1, a_2, k, n) = 1 - JaccardIndex(S(a_1, k, n), S(a_2, k, n)),$$

where $S(a_i,k,n)$ is the set of k-mers of length k present in the allele a_i that appear at least n times in at least one repeat allele. The Jaccard index between two sets is defined as the size of the intersection of these sets divided by the size of their union. For our analyses, we used k-mers of length 5 that appear five or more times in at least one allele ($k = 5$ and $n = 5$). We then defined CPS for a TR as the mean of CDSs calculated for all pairs of alleles.

Genotyping TRs using flanking SNPs

Sometimes, it can be difficult to resolve variation in TR regions based on the repeat sequence alone. One example is measuring methylation of homozygous repeats: if a repeat is homozygous, the reads and their methylation levels cannot be assigned to alleles based only on the repeat sequence. Another example is genotyping repeats with mosaic alleles. Such alleles give rise to reads supporting a range of repeat lengths, making it difficult to determine their allele of origin. We propose using SNPs surrounding the repeat to overcome these issues. These flanking SNPs provide independent evidence that allows us to assign reads to alleles and subsequently genotype repeats and determine their allele-specific methylation.

For modeling purposes, we associate each read r spanning the repeat with a vector of 1 s and 0 s indicating presence or absence of each SNP that the read overlaps. That is, $r[k] = 1$ if the read r contains k-th SNP and $r[k] = 0$ otherwise. A local haplotype is similarly defined as a vector of 0 s and 1 s. The genotype consists of a pair of local haplotypes $G = (H, H')$. We evaluate the posterior probability of the genotype G given the set of observed HiFi reads R following a well-known model[77] for genotyping SNPs: $P(G|R) \sim P(R|G)\,P(G)$, where $P(R|G)$ is the likelihood of observing reads R given the genotype G, and $P(G)$ is the prior probability of the genotype G. Furthermore, $P(R|G) = \Pi P(r|G = (H, H')) = \Pi[P(r|H) + P(r|H')]/2$, where the product is taken over all reads $r \in R$. Here, $P(r|H) = \Pi P(k|r, H)$ where $P(k|r, H) = p$ if $r[k] = H[k]$ and $P(k|r, H) = 1 - p$ otherwise. The genotype probabilities $P(G)$ can be estimated by genotyping repeats in control cohorts. Using this model, TRGT determines the most likely genotype $G = (H, H')$ and the corresponding assignment of each read r to either H or H'. Finally, TRGT calculates the consensus sequence for each repeat allele from the reads assigned to the corresponding local haplotype.

Sequencing of the Genome in a Bottle reference samples

The DNA was sheared with Megaruptor 3 to target size of 15–20 kb. SMRTbell prep kit 3.0 was used for sequencing library preparation. The sequencing was performed on the Revio system using Revio sequencing plate (PacBio, 102-587-400) and Revio SMRT Cell tray (102-202-200) with 24-h movies. The HiFi reads were generated on the Revio system and then aligned to GRCh38 reference with pbmm2 version 1.9.0 (ref. 78).

Data analysis and visualization

Data analysis and visualization were performed with Python 3.10, Jupyter Notebook 7.0.4, Matplotlib 3.8.0 and Pandas 2.1.1 (refs. 79–81).

Institutional review board approval

The institutional review board of Children's Mercy Kansas City (study no. 1120514) approved this study.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

PacBio Revio sequencing of HG002, HG003 and HG004 samples has been deposited to the Sequence Read Archive (SRA)[82]. Version 0.7 of the HG002 assembly from the Telomere-to-Telomere Consortium was downloaded from GitHub[40,83]. The data created as part of Genomic Answers for Kids are available through NIH/NCBI dbGAP, accession number phs002206 (ref. 84). Human Pangenome Reference Consortium data are available at the SRA under BioProject ID PRJNA850430 (ref. 85) and the AWS Registry of Open Data[86]. The short-read data for HG002, HG003 and HG004 are available from the 1000 Genomes Phase 3 Reanalysis with DRAGEN 3.5 and 3.7 within the AWS Registry of Open

Data[87]. TRGT repeat catalogs and TRGTdb for 100 HPRC samples have been deposited into a dedicated Zenodo repository[88].

Code availability

The source code of TRGT, TRVZ and TRGTDB is available on GitHub[64].

References

59. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
60. Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).
61. TRGTdb tutorial. https://github.com/ACEnglish/trgt/blob/main/tdb_tutorial.md
62. Stovner, E. B. & Sætrom, P. PyRanges: efficient comparison of genomic intervals in Python. *Bioinformatics* **36**, 918–919 (2020).
63. ACEnglish/trgt. https://github.com/ACEnglish/trgt/tree/main/notebooks
64. Dolzhenko, E. et al. TRGT: tandem repeat genotyper. *Github* https://github.com/PacificBiosciences/trgt/ (2023).
65. Index of /ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/LowComplexity. https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/genome-stratifications/v3.0/GRCh38/LowComplexity/
66. Table Browser. https://genome.ucsc.edu/cgi-bin/hgTables
67. Repeats. http://useast.ensembl.org/info/genome/genebuild/assembly_repeats.html
68. Bakhtiari, M., Park, J., Javadzadeh, S., Homer, N. & De Coster, W. A tool for genotyping Variable Number Tandem Repeats (VNTR) from sequence data. *Github* https://github.com/mehrdadbakhtiari/adVNTR (2023).
69. Qiu, Y. J., Deshpande, V., Avdeyev, P., Dolzhenko, E. & Eberle, M. A. Illumina/RepeatCatalogs. *Github* https://github.com/Illumina/RepeatCatalogs (2023).
70. Lucas, J., Li, H. & Jeltje human-pangenomics/HPP_Year1_Assemblies. Assemblies from HPP Year 1 production. *Github* https://github.com/human-pangenomics/HPP_Year1_Assemblies (2023).
71. Ebert, P. et al. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science* **372**, eabf7117 (2021).
72. Garg, S. et al. Chromosome-scale, haplotype-resolved assembly of human genomes. *Nat. Biotechnol.* **39**, 309–312 (2021).
73. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
74. Cohen, A. S. A. et al. Genomic answers for children: dynamic analyses of >1000 pediatric rare disease genomes. *Genet. Med.* **24**, 1336–1348 (2022).
75. Cheung, W. A. et al. Direct haplotype-resolved 5-base HiFi sequencing for genome-wide profiling of hypermethylation outliers in a rare disease cohort. *Nat. Commun.* **14**, 3090 (2023).
76. Pedersen, B. S. et al. Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Med.* **12**, 62 (2020).
77. Li, R. et al. SNP detection for massively parallel whole-genome resequencing. *Genome Res.* **19**, 1124–1132 (2009).
78. Töpfer, A. et al. PacificBiosciences/pbmm2. A minimap2 frontend for PacBio native data formats. *Github* https://github.com/PacificBiosciences/pbmm2 (2023).
79. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
80. Granger, B. E. & Perez, F. Jupyter: thinking and storytelling with code and data. *Comput. Sci. Eng.* **23**, 7–14 (2021).
81. pandas-dev/pandas: Pandas. *Zenodo* https://doi.org/10.5281/zenodo.10045529 (2023).
82. *Homo sapiens* (human): WGS of GIAB HG002-4 trio with PacBio HiFi. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1028149 (2023).
83. Hansen, N. F., Phillippy, A., Koren, S. & Walenz, B. Telomere-to-telomere consortium HG002 'Q100' project. *Github* https://github.com/marbl/hg002 (2023).
84. Genomic Answers for Kids (GA4K). dbGaP. https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs002206.v4.p1
85. *Homo sapiens*: Human Pangenome Reference Consortium (HPRC). https://www.ncbi.nlm.nih.gov/bioproject/730823 (2021).
86. Human PanGenomics Project. https://registry.opendata.aws/hpgp-data/
87. 1000 Genomes Phase 3 Reanalysis with DRAGEN 3.5 and 3.7. https://registry.opendata.aws/ilmn-dragen-1kgp/
88. Dolzhenko, E. & English, A. Repeat catalogs for TRGT. *Zenodo* https://doi.org/10.5281/zenodo.8329210 (2023).

Acknowledgements

We would like to thank M. Gymrek, I. Deveson and the anonymous reviewers for helping us to substantially improve the manuscript and TRGT. We are grateful to the Telomere-to-Telomere Consortium, the Human Pangenome Reference Consortium and the Genome in a Bottle Consortium for releasing datasets essential for this study. We would also like to acknowledge many TRGT users who provided valuable feedback that helped us to substantially improve the tool. We thank generous donors to the Genomic Answers for Kids program at Children's Mercy Kansas City. A.E. was supported by grant HHSN268201800002I. H.D. was supported by grants K99HG012796 and 5T32HG008962-07. P.J. was supported by grants NS111602, HD104458 and HD104463. D.L.N. was supported by grants HD104463, NS051630 and HD103555. S.Z. was supported by grant 2R01NS072248. T.P. was supported by grant UL1TR002366. A.R.Q. was supported by grant R01HG010757. F.J.S. was supported by grants 1U01HG011758-01, 3OT2OD002751 and 1UG3NS132105-01.

Author contributions

E.D. and M.A.E. devised and implemented the initial versions of TRGT and TRVZ. A.E. and F.J.S. implemented TRGTdb. H.D. performed analysis of samples with known expansions, in collaboration with W.A.C., C.B., E.F. and T.P. H.D., W.J.R., Z.K. and A.W. guided the development of TRGT. G.D.S.B., E.D., H.D. and M.C.D. performed benchmarking analyses. T.M. and G.D.S.B. contributed major improvements to the TRGT source code. E.D., H.D., A.E., G.D.S.B. and T.M. performed TR analyses in the HPRC samples. V.M.-C., T.D.B., P.J. and D.L.N. generated sequencing from prefrontal cortex samples of individuals with *FMR1* expansions. M.A.E., F.J.S., A.R.Q., T.P. and S.Z. provided guidance and supervision. E.D., A.E., H.D., F.J.S. and M.A.E. wrote the manuscript, with assistance from C.K., K.P.C., W.J.R., Z.K., A.W. and A.R.Q. All authors read and approved the manuscript.

Competing interests

E.D., G.D.S.B., T.M., W.J.R., C.K., Z.K., K.P.C., A.W. and M.A.E. are employees and shareholders of Pacific Biosciences. F.J.S. received research support from Illumina, Pacific Biosciences, Nanopore and Genentech. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-02057-3.

Correspondence and requests for materials should be addressed to Michael A. Eberle.

Peer review information *Nature Biotechnology* thanks Ira Deveson, Melissa Gymrek and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Michael A Eberle

Last updated by author(s): Nov 1, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	PacBio HiFi data was generated on Revio and Sequel systems
Data analysis	The HiFi reads were aligned to GRCh38 genome assembly with pbmm2 v1.9.0 and then analyzed with TRGT and TRVZ (source code deposited to https://github.com/PacificBiosciences/trgt/). The subsequent data analysis was performed with Python 3.10, Jupyter Notebook v7.0.4, Matplotlib v3.8.0, Pandas v2.1.1, Pyranges v0.0.129, Somalier v0.2.17, karyoploteR v1.28.0, Mosdepth v0.3.3. TandemRepeatFinder v4.09 and minimap2 v2.26 was used to generate the repeat catalog. We ran TRGT v0.5.0 (trgt --genome {genome} --repeats {repeats} --reads {bam} --output-prefix {sample} --karyotype {karyotype} --threads {threads}), tandem-genotypes 1.9.0 (tandem-genotypes -o2 --mint-unit=l {catalog} {maf} > {output}), Straglr vl.4.1(straglr.py {barn} {genome} {sample} --min_support 2 --loci {catalog} --min_str_len 1 --max_num_clusters 2 --nprocs {threads}), and GangSTR v2.5.0 (gangstr --barn {bam} --ref {genome} --regions {catalog} --out {sample}}. Hardy-Weinberg equilibrium analysis was performed using code from the TRtools v5.0.1 package.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

PacBio Revio sequencing of HG002, HG003, and HG004 samples has been deposited to SRA. Version 0.7 of the HG002 assembly from the "Telomere-to-Telomere" (T2T) Consortium was downloaded from GitHub. The data created as part of Genomic Answers for Kids is available through NIH/NCBI dbGAP, accession: phs002206. Human Pangenome Reference Consortium (HPRC) data is available at NCBI SRA under the BioProject IDs PRJNA850430 and AWS Registry of Open Data. The short-read data for HG002, HG003, HG004 is available from the 1000 Genomes Phase 3 Reanalysis with DRAGEN 3.5 and 3.7 within the AWS Registry of Open Data. TRGT repeat catalogs and TRGTdb for 100 HPRC samples have been deposited into a dedicated Zenodo repository.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	The sex was known from the The International Genome Sample Resource (https://www.internationalgenome.org/) and was supplied as a parameter to TRGT in order to determine the the number of repeat alleles reported for sex chromosomes.
Reporting on race, ethnicity, or other socially relevant groupings	Socially relevant groupings were not used during the course of this analysis.
Population characteristics	Population characteristics were not used during the course of this analysis.
Recruitment	No participants were recruited specifically for this study.
Ethics oversight	The institutional review board of Children's Mercy Kansas City (Study#l1120514) approved this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. The analysis focused on 100 HPRC samples to capture the most common repeats alleles.
Data exclusions	No data was purposefully excluded from the analysis.
Replication	We used TRGT to independently analyze 937,122 tandem repeats across the genome. Furthermore we confirmed that TRGT was able to successfully call known repeat expansions in six independent samples.
Randomization	We did not perform randomization since the main purpose of this work was to introduce TRGT and use it describe variation in repeat regions across the genome.
Blinding	Blinding was not performed, since our analysis had no defined sample groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.