Profiling the genome-wide landscape of tandem repeat expansions

Nima Mousavi[1,2,3], Sharona Shleizer-Burko[2,3] Melissa Gymrek[1,2,3*]

[1] Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA USA
[2] Department of Medicine, University of California San Diego, La Jolla, CA USA
[3] Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA USA

* Correspondence should be addressed to mgymrek@ucsd.edu

Abstract

Tandem Repeat (TR) expansions have been implicated in dozens of genetic diseases, including Huntington's Disease, Fragile X Syndrome, and hereditary ataxias. Furthermore, TRs have recently been implicated in a range of complex traits, including gene expression and cancer risk. While the human genome harbors hundreds of thousands of TRs, analysis of TR expansions has been mainly limited to known pathogenic loci. A major challenge is that expanded repeats are beyond the read length of most next-generation sequencing (NGS) datasets. We present GangSTR, a novel algorithm for genome-wide profiling of both normal and expanded TRs. GangSTR extracts information from paired-end reads into a unified model to estimate maximum likelihood TR lengths. We validated GangSTR on real and simulated TR expansions and show that GangSTR outperforms alternative methods. We applied GangSTR to more than 150 individuals to profile the landscape of TR expansions in a healthy population and validated novel expansions using orthogonal technologies. Our analysis revealed that each individual harbors dozens of TR alleles longer than standard read lengths and identified hundreds of potentially mis-annotated TRs in the reference genome. GangSTR is packaged as a standalone tool that will likely enable discovery of novel pathogenic variants not currently accessible from NGS.

Introduction

Next-generation sequencing (NGS) has the potential to profile all genetic variants simultaneously in a single test. However, most variant discovery pipelines have focused on single nucleotide variants (SNVs) or short indels which are easiest to genotype. Tandem repeat (TR) variants, such as short tandem repeats (STRs; motif length 1-6bp) and variable number tandem repeats (VNTRs; motif length >6bp) have been implicated in dozens of disorders that collectively affect millions of individuals worldwide[1–3]. In most cases, the pathogenic mutation is an expansion of the number of repeats. Additional pathogenic repeat expansions continue to be discovered[4], usually through cumbersome mapping and long-read sequencing efforts. Despite their importance, most clinically relevant TR variants are largely missing from standard NGS pipelines due to the bioinformatics challenges they present.

Over the last several years, we and others have developed a series of tools for genome-wide genotyping of STRs[5–8] or targeted genotyping of VNTRs[9] from short reads with accuracy comparable to the gold standard capillary electrophoresis technique. These tools primarily rely on identifying reads that completely enclose the repeat of interest. While most TRs in the human genome can be spanned by 100bp reads[10], most known pathogenic TR expansions exceed this range. Furthermore, the reliance on enclosing reads induces a strong bias toward calling short alleles. Thus, existing tools either completely ignore or produce erroneous genotypes at these key loci.

Recently, several methods have been developed that can handle repeats longer than the read length[11–14]. However, these face important limitations. First, several[13,14] only attempt to classify repeats as “expanded” vs. “normal” and do not return estimates of the actual repeat count, which is often informative of disease severity or age of onset[15]. An additional challenge is that these tools require a “control” population for comparison which is not always available. Second, existing methods are primarily built for analyzing whole genome sequencing (WGS), and do not handle other types of data such as whole exome sequencing well. Third, these tools have focused mainly on tri- or tetra-nucleotide expansion disorders and have mostly ignored longer TRs. Finally, all of these methods have been designed to target several dozen known pathogenic loci, and do not trivially scale to genome-wide unbiased analyses.

Here, we present GangSTR, a novel method for simultaneously genotyping both normal and expanded TRs from NGS data. GangSTR relies on a general statistical model incorporating multiple properties of paired-end reads into a single maximum likelihood framework. Unlike previous tools, GangSTR is built to profile genome-wide repeats rather than restricting to a targeted set of known disease-associated TRs and thus can be used to discover novel pathogenic variants. We extensively benchmarked GangSTR against existing methods on both simulated and real datasets harboring normal alleles and pathogenic expansions. We then applied GangSTR to genotype TRs using high-coverage NGS from a trio family to evaluate Mendelian inheritance and validated novel repeat expansions using orthogonal long read and capillary electrophoresis data. Finally, we applied GangSTR to 150 whole genomes to identify and characterize long TRs in a healthy population. Altogether, our analyses demonstrate GangSTR's ability to identify genome-wide repeat expansions which will likely allow for discovery of novel pathogenic loci not currently accessible from short reads using existing tools.

GangSTR is packaged as an open-source tool at https://github.com/gymreklab/GangSTR.

Results

A novel method for genotyping short and expanded TRs from short reads

GangSTR is an end-to-end method that takes sequence alignments and a reference set of TRs as input and outputs estimated diploid repeat lengths. It's core component is a maximum likelihood framework incorporating various sources of information from short paired-end reads into a single model that is applied separately to each TR in the genome.

Multiple aspects of paired-end short reads can be informative of the length of a repetitive region. Reads that completely enclose a repeat trivially allow determination of the repeat number by simply counting the observed number of repeats. While existing tools have primarily focused on repeat-enclosing reads, other pieces of information, such as insert size, coverage, and existence of partially enclosing reads, are all functions of repeat number. New tools for targeted genotyping of expanded STRs utilize various combinations of these information sources (**Table 1**).

GangSTR incorporates each of these informative aspects of paired-end read alignments into a single joint likelihood framework (**Figure 1**). We define four classes of paired-end reads: *enclosing* read pairs ("E") consist of at least one read that contains the entire TR plus non-repetitive flanking region on either end; *spanning* read pairs ("S") originate from a fragment that completely spans the TR, such that each read in the pair maps on either end of the repeat; *flanking* read pairs ("F") contain a read that partially extends into the repetitive sequence of a read; and *fully repetitive* read pairs ("FRR") contain at least one read consisting entirely of the TR motif. The underlying genotype is represented as a tuple <A, B>, where A and B are the repeat lengths of the two alleles of an individual. Two types of probabilities are computed for each read pair: the *class probability*, which is the probability of seeing a read pair of a given class given the true genotype (**Supplementary Figure 1**), and the *read probability* (**Supplementary Figures 2-5**), which gives the probability of observing a particular characteristic of the read pair.

A different characteristic is modeled for each class. For "E" read pairs, read probabilities model the observed TR count, accounting for errors introduced during PCR as was done in previous tools including lobSTR[5] and HipSTR[6]. For "S" read pairs, read probabilities model the fragment length as a Gaussian distribution. Shorter observed fragment lengths than expected are indicative of a repeat expansion compared to the reference genome, whereas longer fragment lengths may indicate a contraction (**Supplementary Figure 4**). For "FRR" read pairs, read probabilities model the distance of the non-repetitive mate pair to the TR (**Supplementary Figure 5**). Additionally, "F" read pairs boost the likelihood of repeat lengths at least as long as the observed number of copies of the motif.

In addition to characteristics discussed above, our model contains a term for the total number of "FRR" reads. This term assigns a probability to an underlying genotype <A, B> based on the expected number of sequenced reads that are fully repetitive ("FRR"), which should grow with the length of the TR. To calculate this probability, we assume a uniform coverage and model the number of "FRR" reads using a Poisson distribution with parameter linearly related to the size of A and B alleles (details in **Supplementary Note**).

Reads that contain evidence of large expansions often greatly deviate from the reference genome and thus are misaligned. In order to identify and extract these reads, we perform local

realignment (similar to TredParse[12], details in **Supplementary Note)** on suspected reads aligned to the vicinity of the TR locus and their mate-pairs. We refer to informative read pairs that are aligned near the TR locus (or have a mate-pair aligned to the area around the TR locus) as “on-target” reads. Most of the “E”, “S”, “F”, and “FRR” read pairs are on-target. However, for large expansions some fragments consist entirely of the repeat and may not align to the correct genomic region (“off-target”). In order to rescue these reads, we scan a predefined set of off-target regions (**Methods**) which allows us to expand the genotyping range beyond the fragment length. While these “off-target” FRRs cannot be uniquely mapped to a specific genomic region, our genome-wide analysis below shows large TR expansions of motifs involved in repeat disorders are rare, and thus most off-target FRRs of the same motif in a given genome are likely to originate from the same TR locus. The GangSTR implementation allows user to choose whether or not to include off-target FRRs in the maximum likelihood calculation.

Tool	**Enclosing Reads**	**FRR**	**Spanning Reads**	**Off-target FRR**	**Genome-wide discovery**	**Estimation Limit**
LobSTR	✦					<Read length
HipSTR	✦				✦	<Read length
ExpansionHunter	✦	✦		✦		Not limited by fragment or read length
Tredparse	✦	✦	✦			<Fragment length
GangSTR	✦	✦	✦	✦	✦	Not limited by fragment or read length

Table 1: Classes of read pairs and features used by existing tools for genotyping TRs from short reads.

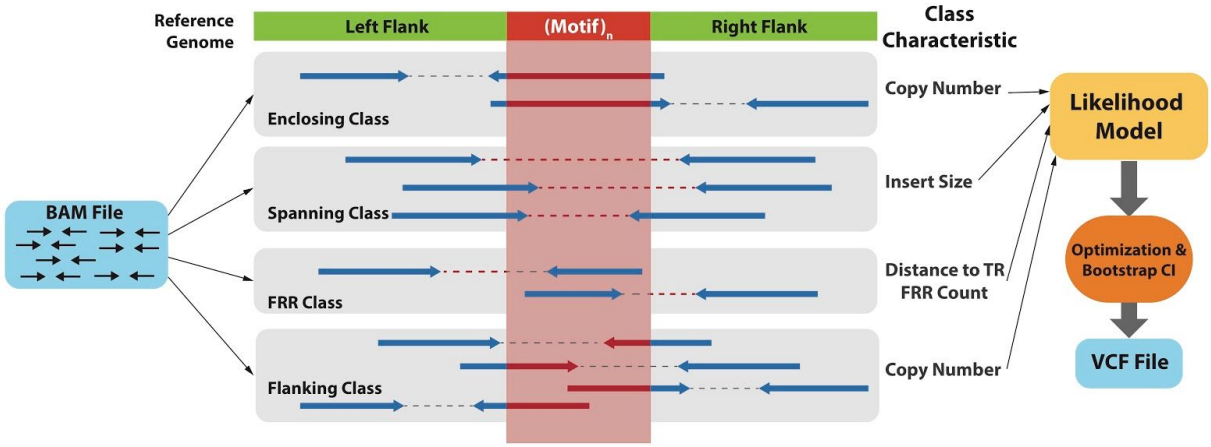

Figure 1 Schematic of GangSTR method. Paired end reads from an input set of alignments are separated into various read classes, each of which provides information about the length of the TR in the region. This information is used to find the maximum likelihood diploid genotype and confidence interval on the repeat length. Results are reported in a VCF file.

The likelihood model computes the probability of the observed reads given a true underlying diploid genotype:

$$logL(<A,B>) = log\prod_i P(r_i;<A,B>) \; + \; logP(|FRR|;<A,B>) = L_P + L_N$$

$$\rightarrow L_P = \sum_i logP(r_i;<A,B>)$$

$$= \sum_i log\{P(r_i|c_i;<A,B>)P(c_i;<A,B>)\}$$

$$\rightarrow L_N = logP(|FRR|;<A,B>)$$

Where $logL(<A,B>)$ corresponds to the total log likelihood given underlying genotype <A,B>, which consists of term L_P combining the contribution of each informative read pair r_i, and term L_N corresponding to the total number of "FRR" reads.. The calculation of L_P is divided for different classes of read pairs to account for differences in the characteristics modeled for each class, where c_i gives the class of each read pair. L_N is calculated using a Poisson model for the number of "FRR" reads based on A, B, and average coverage. The log likelihood function is maximized over all possible diploid genotypes by performing a hybrid optimization approach combining grid search and the Constrained Optimization by Linear Approximation (COBYLA) method. Confidence intervals are determined by repeating the maximum likelihood procedure

on bootstrapped samples of paired-end reads. Full details of the likelihood model and implementation are given in the **Supplementary Note** and **Supplementary Figures 1-5.**

GangSTR outperforms existing TR expansion genotypers

We first evaluated GangSTR's performance by benchmarking against Tredparse[12] and ExpansionHunter[11], two alternative methods for genotyping repeat expansions, using simulated reads for a set of 10 well-characterized repeats involved in trinucleotide pathogenic repeat expansions (**Supplementary Table 1**). Since almost all known repeat expansion disorders follow an autosomal dominant inheritance pattern, we simulated individuals heterozygous for one normal range allele and a second allele that varied along the range of normal and pathogenic repeat counts (**Methods**). In each case, paired-end 100bp reads were simulated to a target of 50-fold coverage, a standard setting for clinical-grade whole genomes. Performance at each locus was measured as the root mean square error (RMSE) between true vs. observed alleles.

GangSTR genotypes had the smallest RMSE for all loci tested (**Figure 2A-C, Supplementary Figure 6)** and was most robust to different ranges of genotypes and experimental parameters. All tools performed similarly for cases where both alleles were shorter than the read length. Tredparse consistently underestimated alleles longer than the fragment length (**Figure 2B,C**). ExpansionHunter performed well in most ranges but produced poor repeat estimates when both alleles were longer than the read length. We performed additional simulations at the Huntington's Disease locus to test the effects of sequencing parameters on each tool's performance. As expected, all tools improved with longer read length (**Supplementary Figure 7**). GangSTR and ExpansionHunter both improved significantly as a function of coverage, whereas Tredparse unexpectedly had worse performance at higher coverages (**Supplementary Figure 8**). In concordance with the fact that Tredparse is limited by fragment length, its performance increased as a function of fragment length, whereas other tools were unaffected by fragment length variation (**Supplementary Figure 9**).

We then tested GangSTR's performance on NGS data from individuals with validated pathogenic repeat expansions (**Methods**). Unfortunately, only a small number of such samples are available. Thus tests on real data were limited to two loci implicated in Huntington's Disease (HTT) and Fragile X Syndrome (FMR1) with sufficient sample sizes. We first genotyped the HTT

and FMR1 loci in 14 and 25 samples respectively with available WGS data[11]. All tools performed well on the HTT locus (**Figure 2D**). GangSTR showed the smallest overall error ($RMSE_{GANGSTR}$=3.09; $RMSE_{TREDPARSE}$=8.30; $RMSE_{EXPANSIONHUNTER}$=10.13) with a small bias in ExpansionHunter for overestimating repeat lengths. Performance was notably worse for all tools at FMR1 (**Supplementary Figure 10;** $RMSE_{GANGSTR}$=30.07; $RMSE_{TREDPARSE}$=34.84; $RMSE_{EXPANSIONHUNTER}$=27.36), presumably since the expanded repeat has 100% GC content and is much longer than the HTT locus.

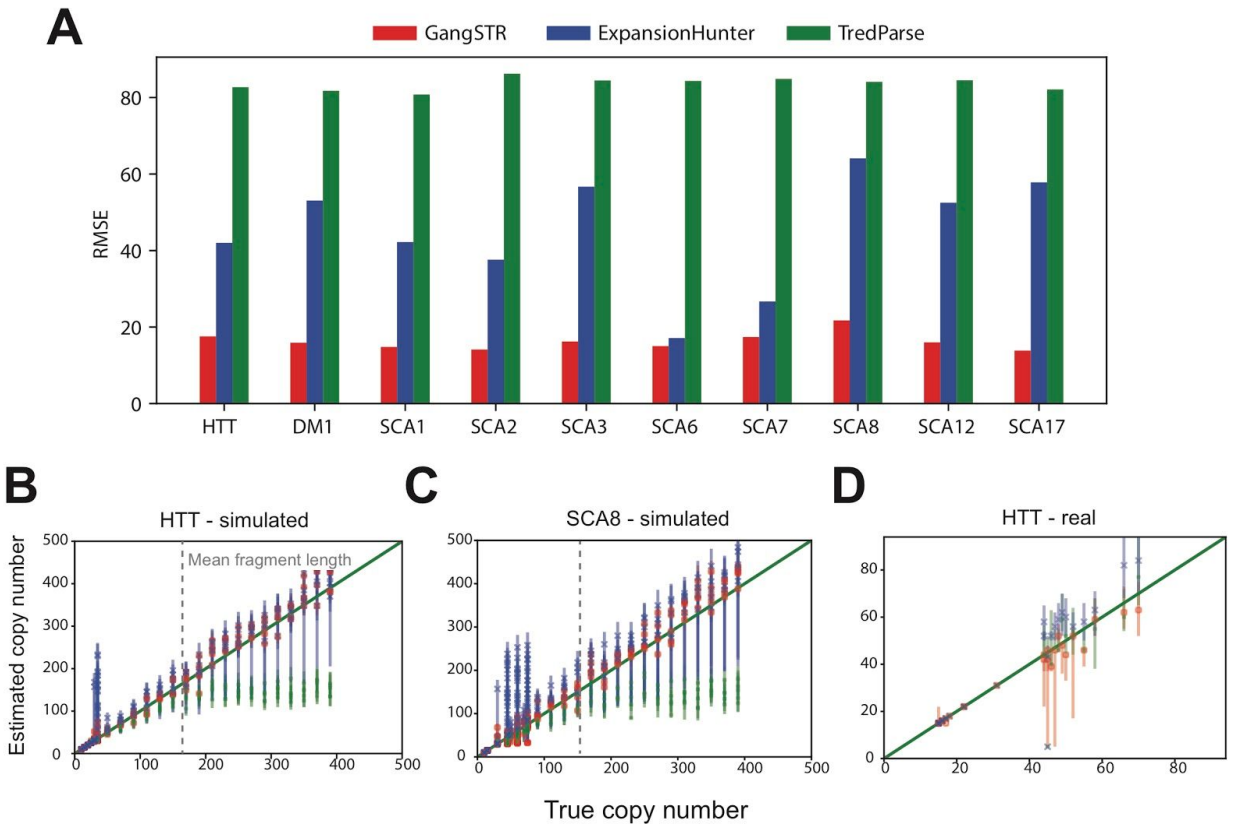

Figure 2: Evaluation of genotypers on real and simulated data at pathogenic repeat expansions. A. RMSE for each simulated locus. HTT=Huntington's Disease; SCA=spinocerebellar ataxia. DM=Myotonic Dystrophy. **B. Comparison of true vs. estimated repeat number for each simulated genotype for HTT.** Dashed gray line gives the mean fragment length. Green solid line gives the diagonal. **C. Comparison of simulated genotypes at SCA8. D. Comparison of true vs. estimated repeat number for HTT using real WGS data.** In all panels, red=GangSTR; blue=ExpansionHunter; green=Tredparse.

We additionally tested each tool on 200 whole exome sequencing datasets from patients with validated Huntington's Disease expansions (**Methods, Supplementary Figure 11**). GangSTR

again showed the smallest error ($RMSE_{GANGSTR}$=4.61; $RMSE_{TREDPARSE}$=96.1; $RMSE_{EXPANSIONHUNTER}$=8.29). Notably, ExpansionHunter, which relies on an underlying model of uniform sequence coverage, gave biased estimates, presumably due to uneven coverage profiles in exomes. As for whole genomes, Tredparse underestimated calls for loci with repeats approaching the fragment length (mean=200bp).

Finally, we evaluated computational performance of each tool on the 10 target loci used in the simulation experiments above on five real whole genomes (**Methods, Supplementary Table 2**). GangSTR gave 13x speedup over Tredparse and 49x speedup over ExpansionHunter using default parameters with a single core. Running ExpansionHunter with a pre-specified genome-wide coverage level rather than calculating coverage on the fly had comparable run time to GangSTR.

Genome-wide detection of TR expansions

Encouraged by our performance at a targeted set of known pathogenic loci, we evaluated whether GangSTR could be used to identify novel expansions by profiling TRs genome-wide. To this end, we used Tandem Repeats Finder[16] to construct a set of all STRs (motif length 2-6bp) and short VNTRs (motif length 7-15bp) in the human reference genome (**Methods**). In total, we identified approximately 580,000 candidate loci with a mean length of 19bp in hg19. Of these, 4,424 are found in coding regions (**Figure 3A**), which primarily contain TRs with motif lengths that are multiples of 3bp.

We used our genome-wide panel to genotype repeats using GangSTR on 30X WGS for a trio of European descent consisting of the highly characterized NA12878 individual and her parents (NA12891 and NA12892). After filtering out low quality loci (**Methods**), an average of 489,246 TRs were profiled per sample. To evaluate GangSTR calls, we determined whether genotypes followed patterns expected based on the trio family structure (**Methods**). Overall, 98.7% of calls were consistent with Mendelian inheritance. The quality of calls steadily increased as a function of the minimum number of observed reads at the locus and was mostly consistent across repeats with different motif lengths (**Figure 3B**). As expected, most repeats matched the reference allele (**Figure 3C**) and the majority (99.99%) were short enough to be completely enclosed by 101bp reads (**Figure 3D**). We noticed an excess of alleles with maximum likelihood lengths close to or slightly longer than the read length. While the size of confidence intervals

tended to increase linearly with allele size, confidence interval sizes increased sharply for alleles estimated to be between 101-125bp (**Supplementary Figure 12**). A similar increase in confidence interval size was observed near the read length for simulated TRs (**Figure 2B, C; Supplementary Figure 6**), suggesting that we tend to produce less precise allele sizes in this range, presumably due to difficulty identifying enclosing and FRR reads for these alleles.

GangSTR identified 174 TRs in NA12878 for which the maximum likelihood length estimate for at least one allele was longer than the read length (**Supplementary Table 3**). Of these, 88% of loci that had calls in all family members were consistent with Mendelian inheritance. Many of the remaining loci showed evidence of expansion in one or both parents. We further identified a high-confidence subset of 65 of these with at least one allele called above 125bp and thus less likely to be due to a read length artifact. Long repeats were highly enriched for repeats with motif $AAAG_n$ (87 repeats, one-sided Fisher's exact test $p=1.67\times10^{-94}$) or $AAAGG^n$ (26 repeats, one-sided Fisher's exact test $p=2.16\times10^{-48}$) (**Supplementary Table 4**). Most other expansions had related motifs of the form A_nG_m. This finding is concordant with previous reports that AAG, AAAG, and AAGG repeats exhibit strong base-stacking interactions that simultaneously promote expansions through replication slippage and protect the resulting secondary structure from DNA repair[17–19].

To further validate GangSTR calls, we examined long read data from WGS for NA12878 generated using Pacific Biosciences (PacBio)[20] and Oxford Nanopore Technologies[21,22] (ONT). For a subset of 10 candidate expansions, we additionally performed capillary electrophoresis to measure TR lengths (**Methods, Supplementary Table 5**). For each of the 174 loci with at least one allele longer than the read length, we extracted regions of PacBio and ONT reads overlapping the TR and determined the repeat length supported by each read (**Methods**). The majority of loci showed evidence of expansion using each technology (166/174 in PacBio and 126/174 in ONT; **Supplementary Table 5**) and were concordant with GangSTR predictions and capillary electrophoresis results (**Figure 3E, F; Supplementary Figure 13**). ONT showed less evidence of expansions, perhaps due to a deletion bias. Both long read technologies exhibit high error rates at homopolymer runs[23], resulting in messy sequence within repeats themselves (**Figure 3E**). Notably, naive comparison of repeat lengths across technologies was difficult and highlights a need for further methods development. While in most cases all three methods (GangSTR, PacBio, ONT) gave concordant results, several loci showed either strikingly different

repeat lengths or no clear signal, which could not be resolved using capillary electrophoresis (**Supplementary Figure 14**). Large differences between technologies may potentially be due to biases introduced during PCR amplification or could represent true somatic variation due to extremely high instability at some repeats.

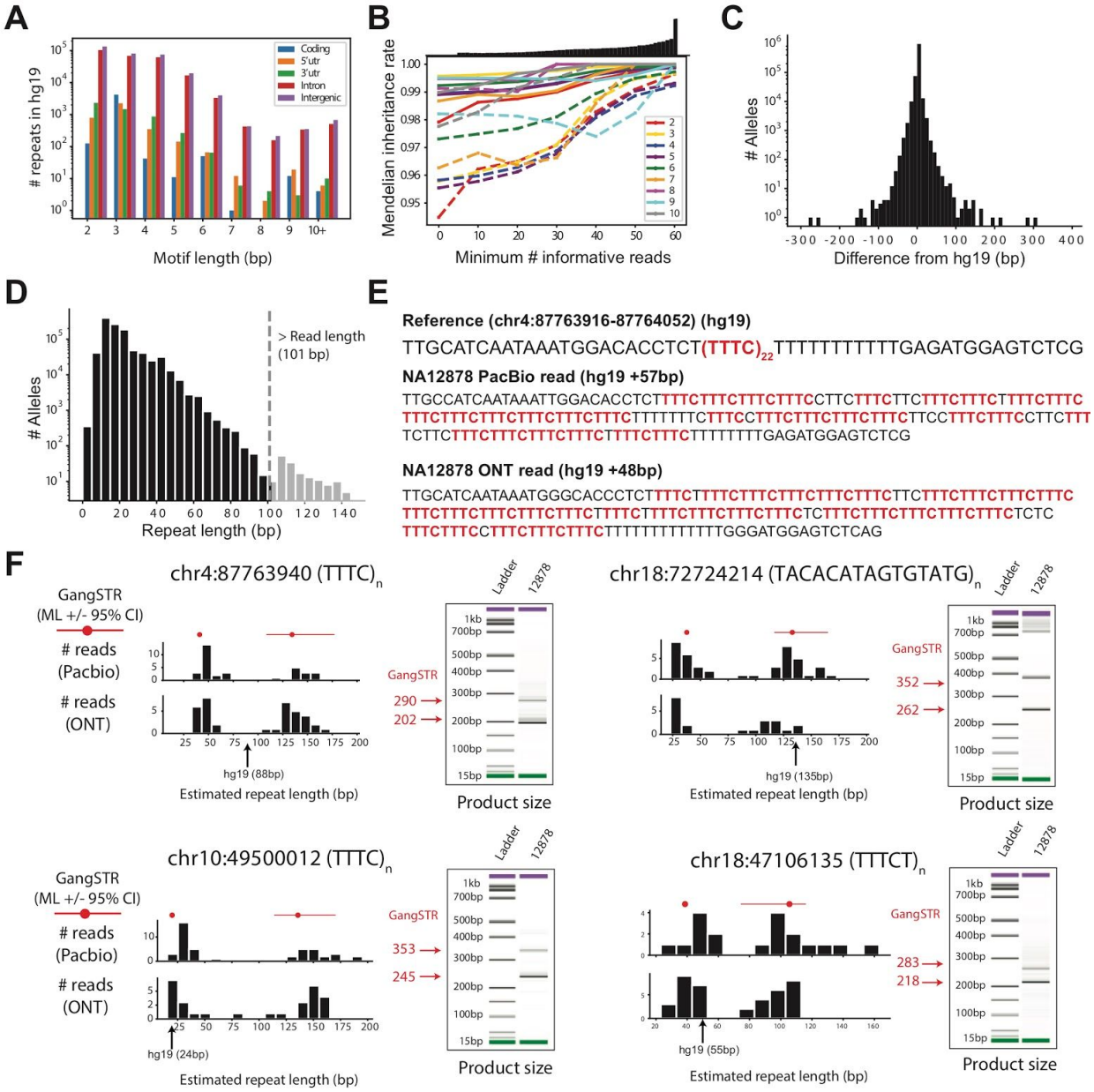

Figure 3: Discovery and validation of genome-wide TR expansions. A. Composition of TRs in the hg19 reference genome. The x-axis gives the motif length and the y-axis ($\log_{10}$ scale) gives the number of TR occurrences in the genome. Colored bars represent TRs

overlapping various genomic annotations. **B. Mendelian inheritance of GangSTR genotypes in a CEU trio as a function of the number of informative read pairs.** Colors denote repeat lengths. Solid lines give mean Mendelian inheritance rate across all loci, computed as described in **Methods**. Dashed lines are computed after excluding loci where all three samples were homozygous for the same allele. **C. Distribution of repeat lengths in NA12878 compared to the hg19 reference.** Y-axis is on a $\log_{10}$ scale. **D. Distribution of total repeat lengths in NA12878.** Y-axis is on a $\log_{10}$ scale. Gray bars to the right of the dashed line indicate alleles longer than the read length of 101bp. **E. Example sequence at a candidate TR expansion.** The reference sequence and representative reads from PacBio (top) and ONT (bottom) for NA12878 are shown for a locus where GangSTR predicted a 48bp expansion from the reference genome. Instances of the repeat motif are shown in red. **F. Validation of candidate expansions.** For each of the four loci shown, left plots compare GangSTR genotypes to those predicted by long reads. Red dots give the maximum likelihood repeat lengths predicted by GangSTR and red lines give the 95% confidence intervals for each allele. Black histograms give the distribution of repeat lengths supported by PacBio (top) and ONT (bottom) reads. The black arrow denotes the length in hg19. The right plots show PCR product sizes for each locus as estimated using capillary electrophoresis. Left bands show the ladder and right bands show product sizes in NA12878. Green and purple bands show the lower and upper limits of the ladder, respectively. Red arrows and numbers give product sizes expected for the two alleles called by GangSTR.

The landscape of TR expansions in a healthy population

We applied GangSTR to determine the frequency of long repeat alleles in a healthy population. For this, we focused on 150 whole genomes sequenced to approximately 45X using paired-end 150bp reads consisting of individuals of European, Asian, and African descent. After filtering (**Methods**), we genotyped an average of 515,384 loci in each sample and identified on average 51.9 TRs per genome with at least one allele longer than 100bp and 6 TRs with at least one allele longer than the read length of 150bp (**Supplementary Table 6**), consistent with our findings in NA12878 above. For the analyses below, we identify “long” alleles as those longer than 100bp, as this read length is commonly used in available NGS datasets.

Allele frequencies at known pathogenic loci determined by GangSTR matched closely to previously reported frequencies and recapitulated known differences between population groups (**Figure 4A-B**). For example, at the CAG repeat implicated in Huntington’s Disease, most alleles consisted of 17-30 repeats, similar to normal allele ranges reported previously (normal alleles from dbGaP samples phs000371.v1.p1). Notably, African and Asian samples had shorter average repeat lengths (mean 18.5 repeats in Europeans vs. 17.5 and 18.1 in Asians and Africans, respectively), consistent with previously reported frequencies[24–27] and the greater prevalence of

Huntington's Disease in Europeans[2]. Similarly, CTG repeat numbers at the DM1 locus in East Asian samples matched allele frequencies reported previously in Asian populations[28].

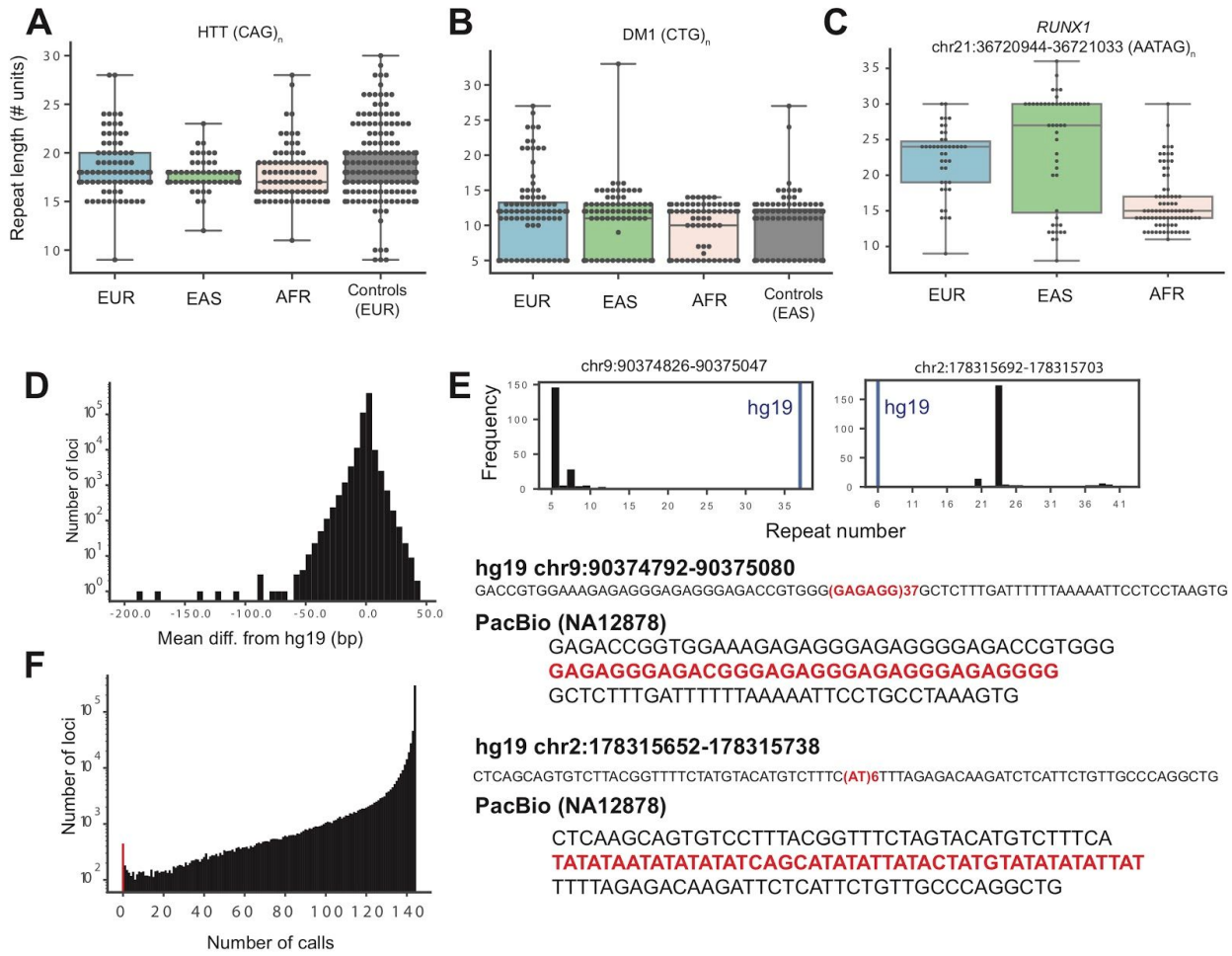

Figure 4: Genome-wide repeat analysis in a healthy population. A. Comparison of allele frequencies at the Huntington's Disease locus estimated by GangSTR to previously reported frequency spectra. Gray gives distribution of the normal allele from Huntington's Disease patients reported in dbGaP dataset phs000371.v1.p. **B. Comparison of allele frequencies at the DM1 locus.** Gray gives distribution for Chinese samples reported in Ambrose, *et al*.[28]. For **A-C** blue=European (EUR), green=East Asian (EAS), red=African (AFR), and gray=previously reported frequencies. Dots give individual alleles. Boxes give the interquartile range, lines give medians, and whiskers extend to the data extreme value. **D. Example locus with high frequency of long alleles and high variability across populations.** The AATAG repeat in an intron of *RUNX1* was identified by GangSTR to have significantly different allele sizes across populations. **D. Distribution of mean allele lengths for each locus relative to hg19.** Y-axis is on a $\log_{10}$ scale. **E. Example loci with observed repeat lengths showing strong deviations from the hg19 allele.** Left: an example locus for which all samples showed large deletions from hg19. Right: an example locus for which all samples showed expansions compared to hg19. Histograms give allele frequencies across all samples. Blue bars give the hg19 allele length. Bottom: sequence structure of these loci in the

hg19 reference genome and in representative reads from NA12878 PacBio data. The repeat motif is shown in red. **F. Distribution of call rate across all profiled repeats.** Y-axis is on a $\log_{10}$ scale. Red bar denotes several hundred loci for which no individual could be genotyped.

We next analyzed the frequency of long repeat alleles genome-wide. Of loci called in at least 50 samples, 95 had at least 10% frequency of alleles >100bp. Of these, 20 had significantly different repeat lengths between populations (Bonferroni-adjusted ANOVA $p<0.01$). For example, an AATAG repeat in an intron of *RUNX1* showed widely varying allele length distributions (mean 22.1, 24.1, and 16.2 in Europeans, East Asians, and Africans, respectively) (**Figure 4C, Supplementary Table 7**). Similar to expansions identified in NA12878, repeats with long alleles were strongly enriched for "AAAG" (n=24; one-sided Fisher's exact test $p=7.6\times10^{-19}$) and "AAAGG" (n=11;one-sided Fisher's exact test $p=3.5\times10^{-20}$) and similar motifs.

Finally, we wondered whether genome-wide analysis of long TRs could identify poorly annotated regions in the human reference genome. We analyzed the mean length difference from the reference allele across all repeats analyzed (**Figure 4D**). For 94% of loci analyzed, the mean repeat length across the 150 samples was within 5bp of the reference allele. However, many repeats showed strong deviations from the reference in all samples. For example, the GAGAGG repeat at chr9:90374826-90375047 was shorter than the hg19 sequence by nearly 200bp in all samples analyzed (**Figure 4E**). Similarly, an AT repeat at chr2:178315692-178315703 was greater than the reference by an average of 36bp. Notably, more than 300 repeats in our reference did not have any genotypes passing quality filters (**Figure 4F**). In many cases these repeats tended to have highly repetitive flanking regions with sequence similar to the repeat motif. These repeats likely represent regions where the reference sequence is poorly annotated or which are intractable for calling with short reads.

Discussion

Our study presents GangSTR, a novel tool for genome-wide genotyping of both short and expanded TRs from NGS data. Our results on simulated and real datasets show that GangSTR can genotype repeat lengths at known pathogenic TR loci with greater accuracy than existing tools. Importantly, GangSTR can be applied to both whole genome sequencing or targeted sequencing experiments and does not require a matched control cohort. Furthermore, we

demonstrate our ability to detect novel repeat expansions by applying GangSTR genome-wide to a deeply sequenced whole genome (NA12878) and validating candidate expansions using orthogonal long read and capillary technologies. Applying GangSTR to a healthy cohort identified dozens of expanded repeats, mostly of the form $AAAG_n$, suggesting these repeats are particularly unstable. Finally, our genome-wide analysis highlights thousands of TRs that are potentially beyond the scope of short reads or are misannotated in the human reference genome.

GangSTR greatly expands the repertoire of repeats that can be profiled using high-throughput sequencing experiments. There is a growing interest in the role of TRs in single-gene disorders (*e.g.* inherited or *de novo* STR expansions), cancer[29], and in complex traits[15,30–32] such as gene expression[33] or DNA methylation[34]. While NGS can theoretically capture most TRs, genome-wide studies so far have been limited to repeats that are shorter than the read length[10,33,35]. On the other hand, based on known pathogenic repeats identified to date, longer repeats are considerably more likely to have phenotypic consequences. Existing methods for longer repeats have so far used a targeted approach restricted to known pathogenic loci. GangSTR merges methods for genotyping both short and long repeats into a unified statistical model capable of genotyping the vast majority of genomic TRs using a single tool. Thus GangSTR will facilitate efforts to identify novel functional repeats in both Mendelian and complex diseases.

Our study faces several important limitations. First, while GangSTR uses a single model to capture all repeat lengths, it does not yet incorporate some recent improvements for genotyping short repeats implemented in HipSTR[6] (*e.g.*, local haplotype reassembly and phasing, per-locus stutter models). However, it would be straightforward to import these techniques in future releases.Additionally, GangSTR currently cannot handle repeats with complex repeat structures consisting of multiple distinct motifs. Finally, several thousand loci in our reference are still largely inaccessible or difficult to genotype accurately. These mainly include TRs with 100% GC content, TRs with many imperfections in the repeat sequence itself, or loci with low complexity flanking regions. A more advanced model for TRs that allows for imperfections would facilitate genotyping these complex loci.

We focused on Illumina short read data here as it is rapidly becoming the clinical standard and remains unmatched in cost and accuracy. Some limitations could be overcome using long read

technologies such as PacBio or ONT. However, genotyping repeats from long reads is not trivial due to the high indel error rate and represents an area for future methods development. It is likely that hybrid approaches combining both short and long read data will provide the greatest accuracy. Notably, for some repeats we could not obtain reliable genotypes using any technology, including short reads, long reads, or PCR methods. This may be due to a combination of difficulty amplifying highly repetitive regions, difficulty sequencing complex repeats, or high error rates in long read data. Additionally, some unstable repeats may exhibit high rates of somatic variation[36,37], rendering the notion of a "correct" genotype meaningless. Indeed, for several loci we saw evidence for a spectrum of repeat numbers in all technologies tested. GangSTR could be extended in the future to incorporate somatic mosaicism into its model.

Overall, GangSTR allows accurate detection of repeat expansions genome-wide and can be readily applied to large NGS cohorts to enable novel genetic discoveries across a broad range of applications.

Methods

Benchmarking using simulated reads

Reads were simulated using wgsim (https://github.com/lh3/wgsim) with mean insert size (-d) 500, standard deviation of insert size (-s) 100, and read length (-1 and -2) 100. Mutation rate (-r), fraction of indel (-R) and probability of indel extension (-X) were all set to 0.0001, and base error rate (-e) of 0.001 was used. The number of simulated reads (-N) was calculated using the following formula:

$$N = \frac{C \,.\, (2F + A.m)}{2r}$$

Where C is the average coverage, F is the length of the simulated flanking region around the locus, A is the number of copies of the motif of length m present in the simulated sample (simulated allele), and r is the read length. The range of genotypes for each disorder was selected such that one allele only covers normal or pre-risk range, while the other allele can be either normal, pre-risk, or pathogenic.

Reads were aligned to the hg38 reference genome using BWA-MEM[38] with parameter -M. GangSTR was run using the disease-specific reference files for each locus given on the

GangSTR website and with parameters --frrweight 0.25 --enclweight 1.0 --spanweight 1.0 --flankweight 1.0 --ploidy 2 --numbstrap 50 --minmatch 4 --minscore 80 --useofftarget with --coverage pre-set. We used Tredparse v0.7.8 with --cpus 6 and --tred appropriately set for each disease locus. ExpansionHunter v2.5.1 was used with --skip-unaligned and --read-depth preset to the simulated coverage.

Identifying off-target regions

Off-target regions corresponding to each motif were identified by creating artificial fully repetitive read pairs and performing alignment using BWA-MEM[38] with parameter -M. The resulting alignment positions were clustered to identify off-target regions.

Quantifying genotyping performance with RMSE

Root mean square error (RMSE) was used to compare estimated vs. expected repeat allele lengths (**Figure 2**, **Supplementary Figures 6-11**). For each diploid genotype (x_1, x_2), we ordered the two alleles by length such that $x_1 \leq x_2$. Then to compare estimated $X = \{(x_{11}, x_{12}), (x_{21}, x_{22}) \ldots (x_{n1}, x_{n2})\}$ and expected $Y = \{(y_{11}, y_{12}), (y_{21}, y_{22}) \ldots (y_{n1}, y_{n2})\}$ genotype estimates, RMSE was defined as: $\sqrt{(\sum_{j=1}^{n} \sum_{i=1}^{2} (y_{ij} - x_{ij})^2 / 2n)}$.

Constructing a genome-wide repeat reference panel

Tandem Repeats Finder[16] was used to create an initial panel of repetitive regions with motifs up to 15bp in the hg19 and hg38 reference genomes. Matching weight 2, mismatch penalty 5, indel score 17, match probability 80, and indel probability 10 were used as parameters. A minimum score threshold of 24 ensured at least 12bp matching for any repetitive region. The length of the repeating region was capped at 1,000 bp.

This initial panel was further subject to multiple filters to avoid imperfect repeat regions that contain mismatches, insertions, or deletions from the repeat motif. First, motifs that are formed by homopolymer runs (i.e., “AAAA”) or by combining smaller sub-motifs (i.e., “ATAT” is made of 2×“AT”) are discarded from the reference. To avoid errors in the local realignment step of GangSTR, all repeating regions are trimmed until they no longer contain any imperfections in their first and last four copies of the motif. Next we require that the trimmed repeating region is a

perfect repetition of the motif. This step ensures there are no errors in longer STRs that may pass the trimming step. Finally, we set a threshold of at least four surviving copies for motifs of length 2-8bp and at least three copies for motifs of length greater than 8bp.

Mendelian inheritance analysis

GangSTR was run on each family member (NA12878, NA12891, NA12892) using the hg19_ver8 reference available on the GangSTR website with parameters: --frr weight 0.25 --enclweight 1.0 --spanweight 1.0 --flankweight 1.0 --ploidy 2 --numbstrap 50 --minmatch 5 --minscore 80 --genomewide. GangSTR genotypes for sample were filtered to exclude (1) calls for which only spanning reads were observed, since these gave unreliable genotypes (2) calls for which either allele of the maximum likelihood genotype was not contained in the 95% confidence interval obtained from bootstrapping estimates and (3) loci overlapping segmental duplications (UCSC Genome Browser[39] hg19.genomicSuperDups table). Let child, mother, and father confidence intervals be denoted as $(c_{1l} - c_{1h},\ c_{21} - c_{2h})$, $(m_{1l} - m_{1h},\ m_{2l} - m_{2h})$, and $(f_{1l} - f_{1h}, f_{2l} - f_{2h})$, where "1" and "2" denote the short and long allele at each diploid genotype and "l" and "h" represent the low and high end of the confidence interval for each allele. A locus was considered to follow Mendelian inheritance if $c_{1l} - c_{1h}$ overlapped either maternal confidence interval and $c_{21} - c_{2h}$ overlapped either paternal confidence interval, or vice versa.

Validating GangSTR using long reads

For each repeat, we used the Pysam (https://github.com/pysam-developers/pysam) python wrapper around htslib and samtools[40] to identify overlapping PacBio or ONT reads and extract the portion of the read overlapping the repeat +/- 50bp. We estimated the repeat length by taking the difference in length between the reference sequence and the number of bases of each read aligned in that region based on the CIGAR score.

GangSTR analysis in a healthy control population

Each sample was genotyped separately using GangSTR with the hs37_ver8 reference available on the GangSTR website using parameters --frr weight 0.25 --enclweight 1.0 --spanweight 1.0 --flankweight 1.0 --ploidy 2 --numbstrap 5 --minmatch 5 --minscore 80 --genomewide. Resulting genotypes were filtered to exclude calls made using less than 10 informative read pairs, calls made using only spanning read pairs, or calls for which the maximum likelihood genotype was

not within the reported 95% confidence interval. Loci overlapping annotated segmental duplications were removed from downstream analyses.

Experimental validation of repeat lengths

Candidate TRs with long alleles identified in NA12878 were PCR amplified using GoTaq (Promega #PRM7123) with primers shown in **Supplementary Table 5**. PCR products were purified using NucleoSpin® Gel and PCR Clean-up (Macherey-Nagel #740609) and analyzed with capillary electrophoresis using an Agilent 2100 Bioanalyzer and an Agilent DNA 1000 kit (#5067-1504).

Datasets

Whole genome sequencing for samples with previously validated repeat expansions were obtained from the European Genome-Phenome Archive (dataset ID: EGAD00001003562). Analysis of exome sequencing for Huntington's Disease patients is based on study data downloaded from the dbGaP web site under phs000371.v1.p1. Whole genome sequencing data for the CEU trio consisting of NA12878, NA12891, and NA12892 was obtained from the European Nucleotide Archive (ENA accession: PRJEB3381). ONT data for NA12878 was obtained from the Nanopore WGS Consortium (https://github.com/nanopore-wgs-consortium/NA12878). PacBio data for NA12878 was obtained from the Genome in a Bottle website (ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/NA12878_PacBio_MtSinai). Whole genome sequencing for 150 control samples of European, African, and East Asian origin were downloaded from ENA study PRJEB20654. Unless otherwise specified, all coordinates given are using the hg19 reference genome.

Acknowledgements

Research reported in this publication was supported in part by the Office Of The Director, National Institutes of Health under Award Number DP5OD024577. M.G. was supported in part by NIH/NIMH grant R01 MH113715. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) comet resource at the San Diego Supercomputing Center through allocations ddp268 and csd568. XSEDE is supported by National Science Foundation

grant number ACI-1548562. We thank Vineet Bafna, Vikas Bansal, and Alon Goren for helpful comments on the method and manuscript.

Author Contributions

M.G. conceived the method, helped design and perform analyses, and drafted the initial manuscript. N.M. designed and wrote the GangSTR algorithm and performed analyses. S.S.B. designed and performed experimental validation of TR genotypes. All authors have read and approved the final manuscript.

Competing and Financial Interests

The authors have no competing financial interests to disclose.

References

1. Hunter, J. *et al.* Epidemiology of fragile X syndrome: a systematic review and meta-analysis. *Am. J. Med. Genet. A* **164A,** 1648–1658 (2014).
2. Pringsheim, T. *et al.* The incidence and prevalence of Huntington's disease: A systematic review and meta-analysis. *Mov. Disord.* **27,** 1083–1091 (2012).
3. Ruano, L., Melo, C., Silva, M. C. & Coutinho, P. The global epidemiology of hereditary ataxia and spastic paraplegia: a systematic review of prevalence studies. *Neuroepidemiology* **42,** 174–183 (2014).
4. Ishiura, H. *et al.* Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat. Genet.* **50,** 581–590 (2018).
5. Gymrek, M., Golan, D., Rosset, S. & Erlich, Y. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.* **22,** 1154–1162 (2012).
6. Willems, T. *et al.* Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods* (2017). doi:10.1038/nmeth.4267

7. Highnam, G. *et al.* Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res.* **41,** e32 (2013).
8. Kristmundsdóttir, S., Sigurpálsdóttir, B. D., Kehr, B. & Halldórsson, B. V. popSTR: population-scale detection of STR variants. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw568
9. Bakhtiari, M., Shleizer-Burko, S., Gymrek, M., Bansal, V. & Bafna, V. Targeted Genotyping of Variable Number Tandem Repeats with adVNTR. (2017). doi:10.1101/221754
10. Willems, T. *et al.* The landscape of human STR variation. *Genome Res.* **24,** 1894–1904 (2014).
11. Dolzhenko, E. *et al.* Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.* **27,** 1895–1903 (2017).
12. Tang, H. *et al.* Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes. *Am. J. Hum. Genet.* **101,** 700–715 (2017).
13. Dashnow, H. *et al.* STRetch: detecting and discovering pathogenic short tandem repeats expansions. (2017). doi:10.1101/159228
14. Tankard, R. M., Delatycki, M. B., Lockhart, P. J. & Bahlo, M. Detecting known repeat expansions with standard protocol next generation sequencing, towards developing a single screening test for neurological repeat expansion disorders. (2017). doi:10.1101/157792
15. Hannan, A. J. Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.* **19,** 286–298 (2018).
16. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
17. Bacolla, A. *et al.* Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. *Genome Res.* **18,** 1545–1553 (2008).

18. Ahrendt, S. A. *et al.* Microsatellite instability at selected tetranucleotide repeats is associated with p53 mutations in non-small cell lung cancer. *Cancer Res.* **60,** 2488–2491 (2000).

19. Xu, L. *et al.* Microsatellite instability at AAAG repeat sequences in respiratory tract cancers. *Int. J. Cancer* **91,** 200–204 (2001).

20. McCarthy, A. Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chem. Biol.* **17,** 675–676 (2010).

21. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36,** 338–345 (2018).

22. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.* **17,** 239 (2016).

23. Weirather, J. L. *et al.* Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res.* **6,** 100 (2017).

24. Barron, L. H. *et al.* A study of the Huntington's disease associated trinucleotide repeat in the Scottish population. *J. Med. Genet.* **30,** 1003–1007 (1993).

25. Baine, F. K. *et al.* Huntington disease in the South African population occurs on diverse and ethnically distinct genetic haplotypes. *Eur. J. Hum. Genet.* **21,** 1120–1127 (2013).

26. Novelletto, A. *et al.* Analysis of the trinucleotide repeat expansion in Italian families affected with Huntington disease. *Hum. Mol. Genet.* **3,** 93–98 (1994).

27. Wang, C. K. *et al.* DNA Haplotype Analysis of CAG Repeat in Taiwanese Huntington's Disease Patients. *Eur. Neurol.* **52,** 96–100 (2004).

28. Ambrose, K. K. *et al.* Analysis of CTG repeat length variation in the gene in the general population and the molecular diagnosis of myotonic dystrophy type 1 in Malaysia. *BMJ*

Open **7,** e010711 (2017).

29. Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and characterization of microsatellite instability across 18 cancer types. *Nat. Med.* (2016). doi:10.1038/nm.4191
30. Press, M. O., Carlson, K. D. & Queitsch, C. The overdue promise of short tandem repeat variation for heritability. *Trends Genet.* **30,** 504–512 (2014).
31. Saini, S., Mitra, I. & Gymrek, M. A reference haplotype panel for genome-wide imputation of short tandem repeats. (2018). doi:10.1101/277673
32. Hannan, A. J. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends Genet.* **26,** 59–65 (2010).
33. Gymrek, M. *et al.* Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.* **48,** 22–29 (2016).
34. Quilez, J. *et al.* Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.* **44,** 3750–3762 (2016).
35. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* (2016). doi:10.1038/nature18964
36. Swami, M. *et al.* Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Hum. Mol. Genet.* **18,** 3039–3047 (2009).
37. Kraus-Perrotta, C. & Lagalwar, S. Expansion, mosaicism and interruption: mechanisms of the CAG repeat mutation in spinocerebellar ataxia type 1. *Cerebellum Ataxias* **3,** 20 (2016).
38. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
39. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006

(2002).

40. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).