Profiling the genome-wide landscape of tandem repeat expansions

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

March 2019

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 and are not profiled by existing genome-wide tools. We present GangSTR, a novel algorithm for genome-wide genotyping of both short and expanded TRs. GangSTR extracts information from paired-end reads into a unified model to estimate maximum likelihood TR lengths. We validate GangSTR on real and simulated data and show that GangSTR outperforms alternative methods in both accuracy and speed. We apply GangSTR to a deeply sequenced trio to profile the landscape of TR expansions in a healthy family and validate novel expansions using orthogonal technologies. Our analysis reveals that healthy individuals harbor dozens of long TR alleles not captured by current genome-wide methods. GangSTR will likely enable discovery of novel disease-associated variants not currently accessible from NGS.

1 Introduction

Next-generation sequencing (NGS) has the potential to profile nearly all genetic variants simultaneously in a single assay. Indeed, whole exome sequencing (WES) and whole genome sequencing (WGS) have successfully identified single nucleotide polymorphisms (SNPs) and small indels contributing to a range of phenotypes, including Mendelian diseases [47], cancer [3], and complex traits [5]. Recently, several studies have demonstrated the power of NGS to genotype more complex structural variants (SVs) and revealed a contribution to a variety of traits including gene expression [8], cancer [45], and autism spectrum disorder [7]. Despite this progress, NGS pipelines struggle with highly repetitive regions of the genome, which are still routinely filtered from most studies.

Here, we focus on short tandem repeats (STRs) with motif lengths of 1-6bp and variable number tandem repeats (VNTRs) with motif lengths of up to 20bp, which we collectively refer to as TRs. TRs have been implicated in dozens of disorders [29], such as Huntington's Disease and Fragile X Syndrome, which together affect millions of individuals worldwide [17, 33, 35]. In most cases, the pathogenic mutation is an expansion of the number of repeats. Importantly, known pathogenic TRs represent just a small fraction of the more than one million TRs in the human genome [43]. Recently, thousands of TRs have been shown to play a role in gene regulation [13, 34] and it is becoming increasingly clear that TRs across the genome are likely to have widespread contributions to complex polygenic traits [32, 14, 15]. In these cases, smaller expansions or contractions may subtly increase or decrease risk for a trait, similar to effect sizes observed for point mutations, and work together to modulate an individual's disease risk [41].

Over the last several years, we and others have developed a series of tools for genome-wide genotyping of STRs [12, 44, 16, 23] from short reads or targeted genotyping of VNTRs [4] from both short and long reads. These tools primarily rely on identifying reads that completely enclose the repeat of interest. While most TRs in the human genome can theoretically be spanned by 100bp reads [27], in practice repeats longer than around 70bp are difficult or impossible to genotype due to an insufficient number of enclosing reads. Notably, in

*Correspondence should be addressed to Email: mgymrek@ucsd.edu. [1]Department of Electrical and Computer Engineering, [2]Department of Medicine and [3] Department of Computer Science and Engineering, University of California San Diego, 9500 Gilman Drive, MC 0639, La Jolla, CA, 92093, USA

our recent genome-wide analysis [36] using HipSTR [44], more than 150,000 STRs were filtered because they showed strong departure from genotype frequencies expected Hardy-Weinberg equilibrium, in part because of dropout of long alleles. The list of filtered TRs includes most known pathogenic TR expansions, for which even normal alleles typically exceed the length of short reads [15]. Thus existing NGS pipelines provide an incomplete picture of genome-wide variation at TRs.

Recently, several methods have been developed to analyze expanded TRs from NGS, but all face limitations that do not allow for unbiased genome-wide analysis of TR lengths. exSTRa [39] classifies a repeat as "expanded" vs. "normal" but requires a control cohort and does not estimate repeat length, which is often informative of disease severity or age of onset [15] and is required for performing genome-wide association studies. STRetch [9] can perform genome-wide expansion identification but does not analyze short TRs, is limited to motifs of up to 6bp, and is computationally expensive. Tredparse [38] models multiple aspects of paired end reads but cannot estimate repeat lengths longer than the sequencing fragment length. ExpansionHunter [11] produces accurate genotypes across a range of repeat lengths except when both alleles are close to or longer than the sequencing read length. Finally, Tredparse and ExpansionHunter have been primarily designed for targeted analysis of known pathogenic expansions and do not scale genome-wide.

Long read technologies have recently been applied to genotype long and complex repeats, such as the CCG repeat implicated in Fragile X Syndrome [26] and a complex pentamer repeat implicated in myoclonus epilepsy [18]. While long reads offer a potential solution to genome-wide TR analysis, NGS remains the gold standard for diagnostic sequencing and population-wide studies due to its low cost and substantially higher throughput [30]. Furthermore, the low per-base accuracy and high indel rate of long read technologies present major challenges to accurate quantification of repeat counts, especially for TRs with short motif lengths. Thus, we focus here on the challenge of comprehensive TR genotyping from short reads.

Here, we present GangSTR, a novel method for genome-wide analysis of TRs from NGS data. GangSTR relies on a general statistical model incorporating multiple properties of paired-end reads into a single maximum likelihood framework capable of genotyping both normal length and expanded repeats. We extensively benchmark GangSTR against existing methods on both simulated and real datasets harboring a range of allele lengths and show that GangSTR is both faster and more accurate than existing solutions. Finally, we apply GangSTR to genotype TRs using high-coverage NGS from a trio family to evaluate Mendelian inheritance and validate novel repeat expansions using orthogonal long read and capillary electrophoresis data. Altogether, our analyses demonstrate GangSTR's ability to robustly genotype a range of TR classes, which will likely enable identification of novel pathogenic expansions as well as genome-wide association studies of TR variation in large cohorts.

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

2 MATERIALS AND METHODS

2.1 Overview of the GangSTR model

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. Its 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 most of existing tools have primarily focused on repeat-enclosing reads, other pieces of information, such as fragment length, coverage, and existence of partially enclosing reads, are all functions of repeat number. Recent 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). The underlying genotype is represented as a tuple $\langle A, B \rangle$, where A and B are the repeat lengths of the two alleles of an individual. 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. Two types of probabilities are computed for each read pair: the

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

Tool	Enclosing	FRR	Spanning	Off-target FRR	Estimates # rpts.	Genome-wide	Estimation limit
LobSTR [12]	X				X	X	< Read Length
HipSTR [44]	X				X	X	< Read Length
STRetch [9]		X		X	X	X	Only reports expanded TRs
exSTRa [40]		X		X		X	Does not estimate TR length
Tredparse [38]	X	X	X		X		< Fragment Length
ExpansionHunter [11]	X	X		X		X	Poor performance when both alleles long
GangSTR	X	X	X	X	X	X	Not limited by fragment or read length

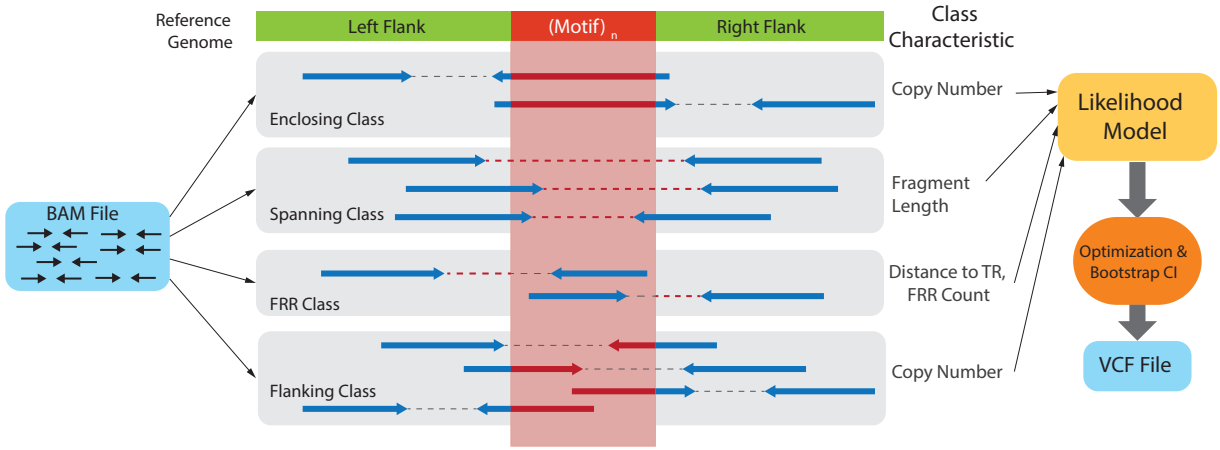

Figure 1: **Schematic of GangSTR method.** Paired end reads from an input set of alignments are separated into various read classes, each of which provides information about the length of the TR in the region. This information is used to find the maximum likelihood diploid genotype and confidence interval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

class probability, which is the probability of observing a read pair of a given class given the true genotype, and the read probability, which gives the probability of observing a particular characteristic of the read pair. A different characteristic is modeled for each class (Figure 2).

2.2 Computation of Log Likelihood

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

$$\begin{aligned}\mathcal{L}(\langle A,B\rangle) &= logP(R;\langle A,B\rangle)\\ &= \overbrace{log\prod_{r\in R}P(r;\langle A,B\rangle)}^{\mathcal{L}_P}+\overbrace{logP(|FRR|;\langle A,B\rangle)}^{\mathcal{L}_N}\end{aligned} \tag{1}$$

Where $\mathcal{L}(\langle A,B\rangle)$ corresponds to the total log likelihood of genotype $\langle A,B\rangle$, which consists of term $\mathcal{L}_P$ combining the contribution of each read pair r from the set of informative read pairs R, and term $\mathcal{L}_N$ which models the total number of FRR reads.

2.2.1 Read pair term

The first term in (1) is calculated by extracting characteristics from every informative read pair, where the specific characteristic modeled depends on the class of the read. Each read pair is assigned to one or more

classes. If a read pair belongs to multiple classes (for example, a read pair can be both spanning and flanking), it appears once in each class for its contribution to both likelihood classes. The read pair term is computed as follows:

$$\begin{aligned}\mathcal{L}_P &= log \prod_{r \in R} P(r; \langle A, B \rangle) \\ &= \sum_{r \in R} log P(r; \langle A, B \rangle) \\ &= \sum_{r \in R} log \sum_{\mathcal{C}_j \in \mathbb{C}} P(r, \mathcal{C}_j; \langle A, B \rangle) \end{aligned} \tag{2}$$

where $\mathbb{C} = \{\mathcal{C}_j\} = \{enclosing, spanning, FRR, flanking\}$ is the set of all informative read classes. Every informative read pair r belongs to a class of informative reads, we denote this class by $C(r)$. The value of $P(r, \mathcal{C}_j; \langle A, B \rangle)$ is set to 1 if $C(r) = \mathcal{C}_j$ and 0 otherwise. We thus simplify the term for each read pair:

$$\begin{aligned}\mathcal{L}_p &= \sum_{r \in R} log P(r, C(r); \langle A, B \rangle) \\ &= \sum_{r \in R} log \underbrace{P(r|C(r_i); \langle A, B \rangle)}_{\text{Read Probability}} \underbrace{P(C(r); \langle A, B \rangle)}_{\text{Class Probability}} \end{aligned} \tag{3}$$

Finally, in a diploid model we assume each read pair is equally likely to originate from allele A or B:

$$\begin{aligned}\mathcal{L}_p = \sum_{r \in R} log \frac{1}{2} \Big\{ & P(r|C(r); A) P(C(r); A) \\ & + P(r|C(r); B) P(C(r); B) \Big\} \end{aligned} \tag{4}$$

2.2.2 Class probability

The class probability, $P(C(r); A)$, models the relative abundance of different classes of informative reads for an underlying repeat length A. We use the schematic in Figure 2 to describe how class probabilities are modeled. We consider a repeat with A copies of a motif of size m bp plus F bp of flanking region on either side. Denote the starting position of each read in a pair relative to the beginning of this region as $S1$ and $S2$, where each read in the pair has length r. Then we can define class probabilities as:

$$\begin{aligned} P(\mathcal{C} = Enclosing; A) &= P(S_2 < F, S_2 + r > F + Am) \\ P(\mathcal{C} = Spanning; A) &= P(S_1 < F, S_2 > F + Am - r) \\ P(\mathcal{C} = FRR; A) &= P(S_1 \leq F, F \leq S_2 \leq F + Am - r) \\ P(\mathcal{C} = Flanking; A) &= P(S_1 < F, S_1 + r < F + Am, S_1 + r > F) \end{aligned}$$

Class probabilities capture changes in the relative abundance of each class as a function of TR length (Supplementary Figure 1). Closed form solutions to compute class probabilities are given in the Supplementary Note.

2.2.3 Read probability

The read probability, $P(r|C(r); A)$, models a separate informative characteristic for each class of informative read pairs as a function of repeat length A.

The number of repeats observed in enclosing reads (parameter n in Figure 2A) can trivially estimate repeat size. However, errors introduced during PCR can alter the number of repeats observed. We model the size of PCR errors using a geometric distribution with default parameter $p = 0.9$ as suggested by HipSTR [44] (Figure 2B).

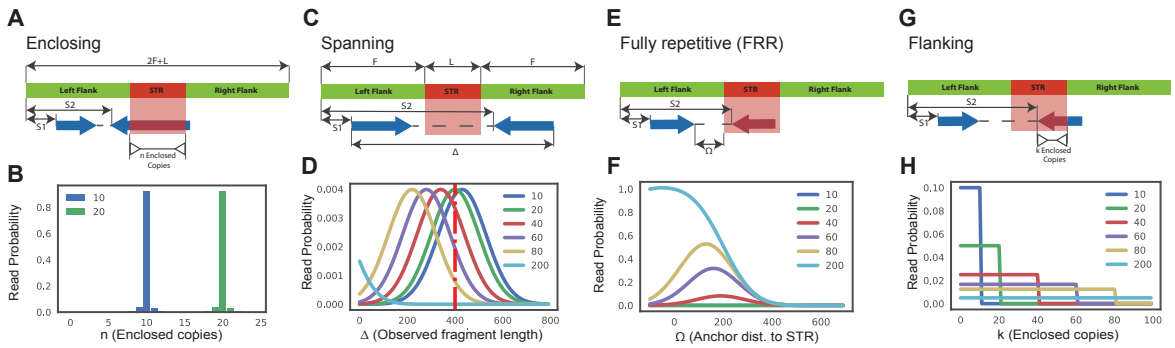

Figure 2: **Four classes of informative read pairs.** A. Enclosing class: characteristic n corresponds to the number of repeat copies enclosed in the read. B. n is modeled for different repeat length accounting for errors introduced during PCR. C. Spanning class: characteristic Δ denotes the observed fragment length for a read pair. D. Δ is modeled for different repeat lengths. Longer repeats give shorter observed fragment lengths. The red vertical dashed line gives the mean actual fragment length. E. Fully Repetitive Read (FRR) class: characteristic Ω is the distance of the non-repetitive read from the repeat region. F. Ω is modeled for different repeat lengths. Longer repeats give shorter observed Ω values. G. Flanking class: characteristic k shows the number of copies extracted from the flanking read. H. k is modeled for different repeat lengths. $S1$ and $S2$ give the start coordinates of each read in the pair relative to the beginning of the first flanking region. For A, C, E, and G, F shows the length (bp) of the flanking region and the repeat is L bp long (A copies of a repeat of length m). For B, D, F, and H, each color denotes a different repeat length (blue=10 copies, green=20 copies, red=40 copies, purple=60 copies, gold=80 copies, light blue=200 copies).

Spanning read pairs have one mate aligned to either side of the TR. In a sample with a TR expansion, the spanning read pair's apparent fragment length based on mapped read positions (parameter Δ in Figure 2C) will shrink compared to the actual fragment length by an amount corresponding to the size of the TR expansion. We thus model observed fragment length as a normal distribution where the mean is a function of repeat length (Figure 2D).

Fully repetitive reads (FRRs) often have an anchor mate that maps in the flanking region before or after the TR. The distance of the anchor from the TR locus (parameter Ω in Figure 2E) is modeled as a function of TR length, with smaller Ω values indicating longer TRs (Figure 2F).

Flanking reads partially cover the TR. The number of repeats in a flanking read (parameter k in Figure 2G) indicates that one allele is at least of size k. For a TR with length A, flanking reads are equally likely to exhibit a number of repeats k ranging from 1 to A (Figure 2H).

Closed form solutions to compute each class probability are given in the Supplementary Note.

2.2.4 Repetitive Read Count Term

The $\mathcal{L}_N$ term in (1) assigns a likelihood to the total number of observed fully repetitive reads. We use a Poisson distribution with parameter λ to model the expected number of observed FRR reads, which is linearly related to the size of alleles A and B. Assuming uniform average coverage C_v, read length r, and motif length m, we can calculate λ using (5). The unit step function $u(.)$ ensures alleles shorter than the read length have 0 expected FRR reads.

$$\lambda = u(A - \frac{r}{m}) \cdot \frac{C_v(A \cdot m - r)}{2r} + u(B - \frac{r}{m}) \cdot \frac{C_v(B \cdot m - r)}{2r} \tag{5}$$

Then we compute the $\mathcal{L}_N$ term as:

$$\begin{aligned} \mathcal{L}_N &= logP(|FRR|; \langle A, B \rangle) \\ &= log\frac{e^{-\lambda}\lambda^{|FRR|}}{|FRR|!} \\ &= -\lambda + |FRR| \cdot log\lambda - log\big(|FRR|!\big) \end{aligned}$$

We use Stirling's approximation to calculate $\log(|FRR|!)$ for large $|FRR|$ values:

$$log(n!) \approx \left(n + \frac{1}{2}\right) \cdot log(n) - n + \frac{1}{2} log(2\pi) \tag{6}$$

2.3 Local Realignment

For enclosing, flanking, and FRR reads GangSTR must obtain accurate counts of the number of repeats contained in each read. For reads fully enclosing the TR plus a minimum of 20bp on either end, repeat count is extracted from the CIGAR score present in the BAM files. Reads starting or ending closer to the TR boundaries or that are fully repetitive are prone to alignment errors and are subject to stringent local realignment. Similar to Tredparse [38], we create artificial reference sequences consisting of flanking region (of size of the read length) on either side and different numbers of repeats, starting from the longest stretch of perfect copies of the repeat and ending with the 2 + 1.1 times the total number of copies of the motif seen in the read. Each read is realigned to the candidate sequences and the reference with the highest realignment score is used to determine the number of repeat copies and the class of the read (flanking, enclosing, or FRR). Realignment is performed using an efficient implementation of the Smith-Waterman algorithm [48].

2.4 Retrieving reads mapped to off-target regions

For large expansions some fragments consist entirely of the repeat and may not map to the correct genomic region (off-target). To rescue these reads, we scan a predefined set of off-target regions for additional FRR reads. While in some cases these off-target FRRs cannot be uniquely mapped, our genome-wide analysis below suggests expansions of most TR motifs are rare, and thus most off-target FRRs of the same motif likely originate from the same locus.

To identify off-target regions for each pathogenic TR, we simulated reads for expanded alleles and aligned them back to the reference genome (see simulation settings below). We extracted positions of reads mapped outside of the simulated region (5000bp on either side of the TR). We merged off-target regions within 30bp of each other and expanded the final merged regions by 10bp on either side. The GangSTR implementation allows users to choose whether or not to include off-target FRRs in the maximum likelihood calculation.

2.5 Optimization

For each TR, GangSTR determines the possible range of repeat lengths from oberved reads. Minimum and maximum counts are determined by enclosing and flanking reads if present. If FRRs are observed, the maximum count is leniently set to a value with mean expected FRR count 5 times the observed count.

By default, GangSTR uses an exhaustive grid search over all possible allele pairs and returns the maximum likelihood diploid genotype. To speed up optimization for TRs with a large range of possible alleles, GangSTR also implements an efficient multi-step optimization procedure. To account for the irregularity of the likelihood surface, we perform a modular optimization procedure with each step searching a different range of allele lengths. First, any enclosing allele a with support of two or more reads is added to the list of potential alleles. In the second step, each potential enclosing allele, a, is used to perform 1-dimensional optimization of the likelihood function to find allele b, were $< a, b >$ minimizes the likelihood function. Next, multiple rounds of 2-dimensional optimization are performed to find $< c, d >$ genotypes that minimize the likelihood function. In each round the optimizer uses a different initial point which helps prevent reporting local optima. Any potential allele from each step, a, b, c, d, is added to the list of potential alleles. In the final step we compare the likelihood from any combination of two alleles in this list, to find the maximum likelihood genotype. All 1 and 2-dimensional optimization is performed using the COBYLA algorithm [31] implemented in the NLopt library [20].

2.6 Quality metrics

GangSTR reports three separate quality metrics to accommodate a range of downstream applications.

2.6.1 Bootstrap confidence intervals and standard errors

In each bootstrap round, GangSTR resamples the set of informative reads (with replacement) to create a bootstrap sample and performs the above optimization procedure on this set of read pairs. The number of

bootstrap samples, N_b, is set by the user. GangSTR records all bootstrap estimates in separate lists for shorter and longer alleles. These lists are then sorted and used to find the confidence interval at the desired level of significance and standard errors on allele lengths.

2.6.2 Genotype likelihoods and quality score

Let L equal the sum of likelihoods for each possible genotype and L_{ML} be the likelihood of the maximum likelihood genotype. GangSTR returns a quality score $Q = \frac{L_{ML}}{L}$. This is equivalent to a posterior probability of the maximum likelihood genotype assuming a uniform prior. This value is most informative for short allele lengths where repeat unit resolution can be achieved. For TR expansions with larger standard errors, the posterior probability of any particular genotype will be low and expansion probabilities are more informative.

2.6.3 Expansion probability

Given a user-specified repeat number expansion threshold X, GangSTR computes the probabilities of no expansion (P_0), a heterozygous expansion above the threshold (P_1), or a homozygous expansion above the threshold (P_2) as:

$$\begin{aligned} P_0 &= \Sigma_{\langle A,B\rangle \in G s.t. A<X, B<X} L(\langle A,B\rangle)/L \\ P_1 &= \Sigma_{\langle A,B\rangle \in G s.t. A<X, B\geq X} L(\langle A,B\rangle)/L \\ P_2 &= \Sigma_{\langle A,B\rangle \in G s.t. A\geq X, B\geq X} L(\langle A,B\rangle)/L \end{aligned} \tag{7}$$

where G is the set of all possible diploid genotypes, $L(\langle A,B\rangle)$ is the likelihood of genotype $\langle A,B\rangle$, and L is as defined above.

2.7 Benchmarking using simulated reads

Reads were simulated using wgsim (https://github.com/lh3/wgsim). Unless otherwise specified, we used parameters mean fragment length (-d) 500, standard deviation of fragment length (-s) 100, and read length (-1 and -2) 150. Mutation rate (-r), fraction of indels (-R) and probability of indel extension (-X) were all set to 0, and base error rate (-e) was set to 0.005. The number of simulated reads (-N) was calculated using the following formula $N = \frac{C(2F+Am)}{2r}$, where C is the average coverage, set to 40x. F is the length of the simulated flanking region around the TR, set to 10,000bp. 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. Simulated genotypes for each pathogenic TR were selected such that the shorter allele covers the normal or premutation range, while the longer allele could be either normal, premutation, or pathogenic (Supplementary Table 1).

Reads were aligned to the hg38 reference genome using BWA-MEM [24] with parameter -M. GangSTR v2.3 was run using the disease-specific reference files for each TR available on the GangSTR website with --coverage set to the simulated coverage level and with the --targeted option. Tredparse v0.7.8 was run with with --cpus 6, --useclippedreads, and --tred appropriately set for each disease locus. ExpansionHunter v2.5.5 was used with and --read-depth preset to the simulated coverage level.

2.8 Quantifying genotyping performance with RMSE

Root mean square error (RMSE) was used to compare estimated vs. expected repeat allele lengths. We denote the diploid genotype of sample i with $\langle x_1^i, x_2^i\rangle$. For each diploid genotype, we ordered the two alleles by length such that $x_1^i \leq x_2^i$. Then to compare estimated $X = \{\langle x_1^1, x_2^1\rangle, \langle x_1^2, x_2^2\rangle ... \langle x_1^n, x_2^n\rangle\}$ and expected $Y = \{\langle y_1^1, y_2^1\rangle, \langle y_1^2, y_2^2\rangle ... \langle y_1^n, y_2^n\rangle\}$ genotypes, RMSE is defined as:$\sum_{i=1}^{n}\sum_{j=1}^{2}\frac{(y_j^i - x_j^i)^2}{2n}$.

2.9 Analysis of genomes and exomes with validated expansions

Whole genome sequencing datasets for samples with previously validated repeat expansions were obtained from the European Genome-Phenome Archive (dataset ID: EGAD00001003562). GangSTR v2.3 was run using the disease-specific reference files for each TR with option --targeted. For Fragile X Syndrome, --ploidy was set to 1 for males and 2 for females. ExpansionHunter v2.5.5 was run using the set of off-target regions given in the GangSTR reference files for Huntington's Disease, and with their published off-target regions for Fragile X Syndrome. Tredparse v.0.7.8 was run using default parameters and --tred set to HD or FXS for Huntington's or Fragile X Syndrome, respectively.

Whole exome sequencing datasets for Huntington's Disease patients were obtained from dbGaP accession phs000371.v1.p1. Fastq files were aligned to the hg19 reference genome using BWA-MEM [24]. PCR duplicates were removed using the samtools [25] rmdup command. Validated repeat lengths were obtained from data fields HDCAG1 and HDCAG2 in table pht002988.v1.p1.c1. We inferred fragment length mean and standard deviation per sample after removing read pairs mapping more than 1kb apart. GangSTR v2.3 was run with --insert-mean and --insert-sdev set to the values computed for each sample. We additionally used parameters --nonuniform and --targeted. ExpansionHunter v2.5.5 was run with --read-depth set to the mean coverage at the TR plus surrounding region. Tredparse v0.7.8 was run with options --useclippedreads and --tred HD.

2.10 Constructing a genome-wide repeat reference panel

Tandem Repeats Finder [6] was used to create a panel of repetitive regions with motifs up to 20bp in the hg19 and hg38 reference genomes using parameters matching weight=2, mismatch penalty=5, indel score=17, match probability=80, and indel probability=10. We required a minimum score threshold of 24 to ensure at least 12bp matching the motif for each TR and removed TRs with reference lengths greater than 1000bp.

This initial panel was subject to multiple filters to avoid imperfect or complex TR regions that cannot be accurately genotyped. First, motifs formed by homopolymer runs (i.e., "AAAA") or by combining smaller sub-motifs (i.e., "ATAT" is made of 2×"AT") were discarded. Based on thresholds used in previous TR references [43], we required TRs with motif size 2 or 3 to have at least 5 or 4 copies in tandem, respectively, and larger motifs to have at least 3 copies. To avoid errors in the local realignment step of GangSTR, all repeating regions were trimmed until they no longer contained any imperfections in their first and last three copies of the motif. We removed TRs within 50bp of another TR as these regions tend to be low complexity and result in low quality calls. Next we discarded remaining TRs that do not consist of perfect repetitions of the motif. Finally, we manually added disease associated TRs to ensure notation is consistent with other methods (e.g. [11, 38]).

2.11 Run time evaluation

All timing and memory experiments were tested in a Linux environment running Centos 7.4.1708 on a server with 28 cores (Intel® Xeon® CPU E5-2660 v4 @ 2.00GHz) and 125 GB RAM and were performed on a single core. Each experiment was run 5 times and the mean value was reported. Tredparse was evaluated on all available TRs ("treds") since it does not allow specifying a subset of TRs for analysis. For all timing analyses we used the --skip-unaligned option for ExpansionHunter, which improved run time. For scalability tests, we randomly chose varying sized sets of TRs from the genome-wide reference. Timing was performed with the UNIX time command and the sum of the sys and user times was reported. Memory usage was measured using the UNIX top command. Virtual memory was measured every 0.1 seconds and the maximum value was reported.

2.12 Genome-wide TR analysis in a CEU trio

Whole genome sequencing data (BAM files) for the CEU trio consisting of NA12878, NA12891, and NA12892 were obtained from the European Nucleotide Archive (ENA accession: PRJEB3381).

GangSTR v2.3 was run on each family member (NA12878, NA12891, NA12892) using the hg19_ver13 reference available on the GangSTR website with default parameters. We supplied an --str-info-file with the expansion threshold for each TR set to the read length of 101bp. HipSTR v.0.6.2 was run on NA12878 with non-default parameters: --lib-from-samp --def-stutter-model --max-str-len 1200 --min-reads 15 --output-filters. STRetch v0.4.0 was run on NA12878 using the GangSTR reference (limited to motifs up to 6bp) as input regions and with no control genomes specified.

We used our filtering tool, DumpSTR (see Code Availability), to filter GangSTR and HipSTR calls. DumpSTR has various recommended filtering settings depending on the downstream application. For example, for applications where precise estimation of TR length is important, more stringent quality filters should be applied vs. for applications targeted at identifying whether a TR is expanded or not. Thus we applied two filter levels referenced in the results as level 1 and level 2.

First, level 1 filters were used to filter out TRs that could not be reliably called. For HipSTR level 1 filtering, we applied dumpSTR options: --max-call-DP 1000 --min-supp-reads 1, which removes calls with abnormally high coverage or calls with no supporting reads, respectively. For GangSTR level 1 filtering, we

applied dumpSTR options: --max-call-DP 1000 --min-call-DP 20 --filter-spanbound-only --filter-badCI, which removes calls with abnormally high coverage, calls where only spanning or bounding reads were found, or calls for which the maximum likelihood genotype falls outside of the 95% bootstrap confidence interval. We additionally filtered regions overlapping annotated segmental duplications in hg19 (UCSC Genome Browser [21] track hg19.genomicSuperDups table).

Second, level 2 filters were used to further restrict to TRs with high confidence length estimates to compare HipSTR vs. GangSTR concordance. For HipSTR level 2 filtering, we applied additional options: --min-call-DP 10 --min-call-Q 0.9 --max-call-flank-indel 0.15 --max-call-stutter 0.15 as recommended on the HipSTR website. For GangSTR level 2 filtering, we applied additional options: –min-call-Q 0.9 –min-total-reads 50.

Mendelian inheritance was determined using two metrics. First, we used maximum likelihood estimates for each sample at each locus to determine whether the child genotype could be explained by parental genotypes. Second, in a less stringent analysis, we determined whether reported confidence intervals were consistent with Mendelian inheritance. Let child, mother, and father confidence intervals be denoted as $(c_1^l - c_1^h, c_2^1 - c_2^h)$, $(m_1^l - m_1^h, m_2^l - m_2^h)$, and $(f_1^l - f_1^h, f_2^l - f_2^h)$, where superscripts 1 and 2 denote the short and long allele at each diploid genotype and subscripts 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 $c_1^l - c_1^h$ overlapped either maternal confidence interval and $c_2^1 - c_2^h$ overlapped either paternal confidence interval, or vice versa.

2.13 Validating GangSTR using long reads

Oxford Nanopore Technologies (ONT) data for NA12878 was obtained from the Nanopore WGS Consortium (https://github.com/nanopore-wgs-consortium/NA12878). Pacific Biosciences (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).

For each repeat, we used the Pysam (https://github.com/pysam-developers/pysam) python wrapper around htslib and samtools [25] 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.

2.14 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).

2.15 Code availability

GangSTR is freely available at https://github.com/gymreklab/GangSTR. The dumpSTR filtering tool is available at https://github.com/gymreklab/STRTools.

3 RESULTS

3.1 GangSTR outperforms existing TR expansion genotypers

We first evaluated GangSTR's performance by benchmarking against Tredparse [38] and ExpansionHunter [11], two alternative methods for genotyping repeat expansions. We focused on these methods since they output estimated repeat number at both normal and expanded TRs and do not require a control cohort as input (Table 1). We simulated reads for a set of 14 well-characterized repeats involved in repeat expansion disorders. 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 (Supplementary Table 1). In each case, paired-end 150bp reads were simulated to a target of 40-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 (Methods).

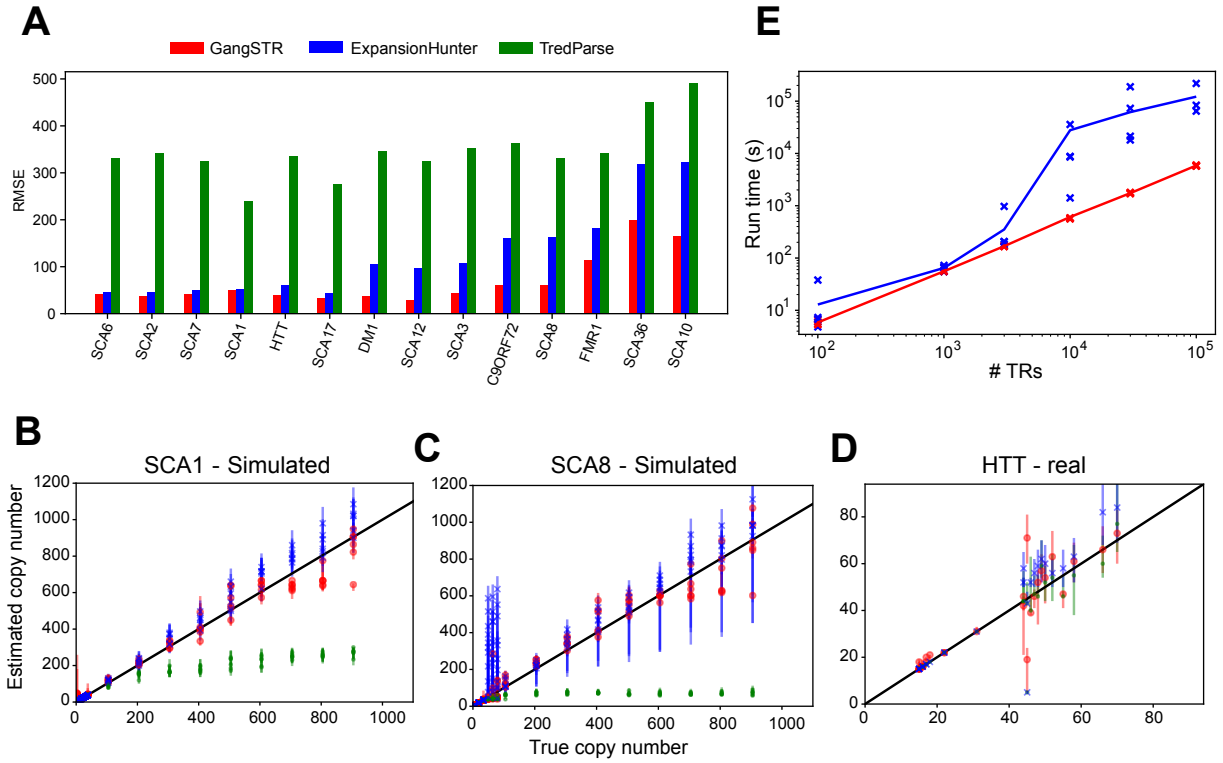

Figure 3: **Evaluation of TR genotypers on real and simulated data at pathogenic repeat expansions.** A. RMSE for each simulated locus. HTT=Huntington's Disease; SCA=spinocerebellar ataxia; DM=Myotonic Dystrophy; C9ORF72=amyotrophic lateral sclerosis/frontotemporal dementia. TRs are sorted from left to right by ascending length of the pathogenic allele. B. Comparison of true vs. estimated repeat number for each simulated genotype for HTT. Black solid line gives the diagonal. C. Comparison of true vs. estimated repeat number for each simulated genotype for SCA8. D. Comparison of true vs. estimated repeat number for HTT using real WGS data. E. Run time vs. number of TRs. The x-axis shows the number of TRs in the reference set used. The y-axis shows running time in seconds. Lines give mean value across 5 runs. Points ("x") give raw data values for each run. For two runs with 10^5 TRs ExpansionHunter did not run to completion. In all panels, red=GangSTR; blue=ExpansionHunter; green=Tredparse.

GangSTR genotypes showed the most robust performance compared to other tools across a wide range of repeat lengths, with the smallest RMSE for all TRs tested (Figure 3A, Supplementary Figure 2). At TRs for which the normal range allele is below the read length (SCA6, SCA2, SCA7, SCA1, HTT, and SCA17), both ExpansionHunter and GangSTR accurately predicted the lengths of both alleles (Figure 3B). However, GangSTR demonstrated a distinct advantage over ExpansionHunter in genotyping TRs for which both the normal and pathogenic allele were close to or longer than the read length, where ExpansionHunter estimates become unstable (Figure 3A, C). Tredparse performed well at short alleles but consistently underestimated alleles longer than the fragment length (Figure 3B, C) which accounts for its inflated RMSE results.

We performed additional simulations at the Huntington's Disease locus to test the effects of sequencing parameters on each tool's performance. GangSTR and ExpansionHunter both improved significantly as a function of coverage and read length, whereas Tredparse was relatively unaffected (Supplementary Figures 3,4). Performance of all tools was mostly consistent across mean fragment lengths (Supplementary Figure 5).

We then tested GangSTR's performance on real NGS data from individuals with validated pathogenic repeat expansions (Methods). Notably, only a small number of such samples are available. Thus tests on real data were limited to two TRs 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 PCR-free WGS data [11]. All tools performed well on the HTT TR (Figure 3D). GangSTR showed the smallest overall error ($\text{RMSE}_{\text{GANGSTR}}$=7.9; $\text{RMSE}_{\text{TREDPARSE}}$=8.3; $\text{RMSE}_{\text{EXPANSIONHUNTER}}$=10.1) with a small bias in ExpansionHunter for overestimating repeat lengths. Performance was notably worse for all tools at FMR1 (Supplementary Figure 6; $\text{RMSE}_{\text{GANGSTR}}$=29.3; $\text{RMSE}_{\text{TREDPARSE}}$=34.8; $\text{RMSE}_{\text{EXPANSIONHUNTER}}$=27.3). Notably, the FMR1 TR has 100% GC content and very few reads mapping directly to the TR could be identified. This highlights a major challenge in calling GC-rich TRs that are still not sequenced well even with PCR-free protocols.

We additionally tested each tool on 200 whole exome sequencing datasets from patients with validated Huntington's Disease expansions (Methods, Supplementary Figure 7). GangSTR again showed the smallest error ($\text{RMSE}_{\text{GANGSTR}}$=5.4; $\text{RMSE}_{\text{TREDPARSE}}$=96.4; $\text{RMSE}_{\text{EXPANSIONHUNTER}}$=9.1). Notably, ExpansionHunter gave biased estimates, presumably due to uneven coverage profiles in exomes. Tredparse again underestimated calls for alleles approaching the fragment length (mean=200bp).

Finally, we evaluated computational performance of each tool on various sets of input TRs. We first used the 14 pathogenic TRs to time each tool. GangSTR performed the fastest (mean=15.4s), with ExpansionHunter showing similar run time (mean=16.1s). Tredparse was significantly slower (mean=82.4). We then performed additional evaluation of the scalability of GangSTR and ExpansionHunter by testing on input TR sets ranging from 100 to 100,000 TRs (Figure 3E). GangSTR run time scaled linearly with reference size as expected, whereas ExpansionHunter run time grew super-linearly. Notably, ExpansionHunter only finished on 3 out of 5 runs with 10,000 TRs and would not run to completion on larger TR sets, potentially due to stalling at problematic loci. We additionally tested the maximum memory requirement of each method. GangSTR memory usage stayed relatively constant at under 1GB, whereas ExpansionHunter memory usage grew linearly with the number of TRs in the reference set (Supplementary Figure 8).

3.2 Genome-wide TR profiling

We next evaluated GangSTR's utility for genome-wide TR genotyping. To this end, we used Tandem Repeats Finder [6] to construct a set of all STRs (motif length 2-6bp) and short VNTRs (motif length 7-20bp) in the human reference genome (Methods). In total, we identified 829,231 TRs (780,328 autosomal) in hg19 with a mean length of 15.6bp. Of these, 5,828 are found in coding regions (Figure 4A), most of which have lengths that are multiples of 3bp.

We used our genome-wide panel to genotype autosomal repeats using GangSTR on WGS with 30x coverage for a trio of European descent consisting of the highly characterized NA12878 individual and her parents (NA12891 and NA12892). After filtering low quality loci (Methods, level 1 filters), an average of 727,255 TRs were genotyped per sample. As expected, most alleles matched the reference (Supplementary Figure 9) with a bias toward calling alleles shorter than the reference. Both alleles for the majority of TRs (>99%) had maximum likelihood lengths less than the read length of 101bp (Supplementary Figure 10). To evaluate GangSTR calls, we determined whether estimated genotypes followed patterns expected based on the trio family structure (Methods). Overall, 98.9% of TRs followed Mendelian inheritance when considering maximum likelihood genotypes. For 99.9% of TRs, 95% confidence intervals were consistent with Mendelian

inheritance (Methods). These values changed to 90.6% and 99.3% respectively after removing TRs that were homozygous reference in all samples. 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 4B, Supplementary Figure 11).

We evaluated GangSTR's utility for genome-wide TR profiling by benchmarking against HipSTR using the same reference TR set. After removing low quality loci from each dataset, (Methods, level 1 filters) GangSTR produced calls at 63,052 TRs that could not be reliably genotyped by HipSTR (Figure 4C). The set of TRs not genotyped by HipSTR tended to be longer than average (20.8bp vs. 15.0bp in hg19, one-sided t-test $p < 10^{-200}$) and includes 8/14 known pathogenic TRs analyzed in Figure 3), demonstrating the limitations of relying on enclosing reads. Notably, 11,596 TRs were not called by GangSTR but were present in HipSTR output. These primarily consist of repeats with SNPs or indels in or near the TR sequence which did not pass GangSTR's stringent local realignment process. After applying stringent recommended quality filters for each tool (Methods, level 2 filters), TRs called by both tools showed extremely high concordance (>99%) (Figure 4D) with strong correlation between allele lengths reported by each (Pearson r=0.99; $p < 10^{-200}$; n=651,264), demonstrating that GangSTR can robustly genotype both STRs previously analyzed using HipSTR as well as long TRs previously excluded from genome-wide analyses.

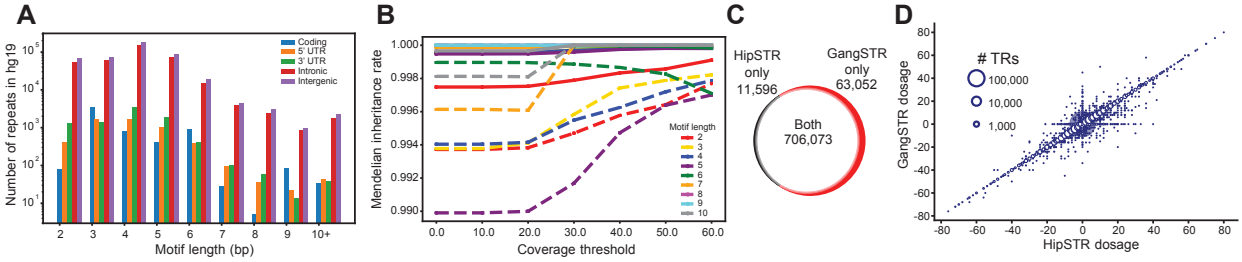

Figure 4: **Genome-wide TR genotyping.** A. Composition of TRs in the hg19 reference genome. The x-axis gives the motif length and the y-axis (log10 scale) gives the number of TRs in the genome. Colored bars represent TRs overlapping various genomic annotations (blue=coding, orange=5' UTR, green=3' UTR, red=intronic, purple=intergenic). 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 TRs, computed based on 95% confidence intervals as described in Methods. Dashed lines are computed after excluding loci where all three samples were homozygous for the reference allele. C. Overlap between TRs genotyped by HipSTR and GangSTR. D. Comparison of HipSTR and GangSTR genotypes. The x-axis and y-axis show the sum of the two allele lengths genotyped by HipSTR and GangSTR in bp relative to the hg19 reference genome (dosage), respectively. The size of the bubble represents the number of points at that coordinate.

3.3 Genome-wide detection of novel TR expansions

We next evaluated whether GangSTR could identify novel repeat expansions in a healthy genome (NA12878). GangSTR identified 69 TRs passing level 1 filters predicted to have at least one allele longer than the read length (101bp) with greater than 80% probability (see Expansion Probabilities defined in Methods) (Supplementary Table 2). Of these, 58 showed evidence of expansions in one or both parents. Long repeats were highly enriched for repeats with motif $AAAG_n$ (27 TRs, one-sided Fisher's exact test $p = 2.7 * 10^{-19}$) and related motifs of the form A_nG_m (Supplementary Table 3). 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 [2, 1, 46].

For comparison, we applied STRetch [9], an alternative tool for detecting repeat expansions, using the GangSTR reference TR set of TRs restricted to motif lengths up to 6bp (808,868 total TRs). STRetch leverages a modified reference genome containing decoy repeat sequences to identify potentially expanded TRs. It only attempts to genotype TRs with candidate expansions and thus is unsuitable for unbiased genome-wide TR genotyping. After filtering for segmental duplications (Methods), STRetch returned results

for 45 TRs (Supplementary Table 4). Notably, STRetch took approximately 157 CPU-hours (6.5 days) compared to 16.6 CPU-hours for GangSTR on a single genome. TRs genotyped by both GangSTR and STRetch showed concordant repeat number estimates (Pearson r=0.61, $p = 4.9 * 10^{-5}$, n=37, Figure 5A, Supplementary Table 4). However only 5 of the 69 TRs with alleles longer than 101bp reported by GangSTR were genotyped by STRetch. Overall these results show that GangSTR provides a more comprehensive analysis of genome-wide TR variation.

To validate putative expansions identified by GangSTR, we examined long read data from WGS for NA12878 generated using Pacific Biosciences (PacBio) [28] and Oxford Nanopore Technologies (ONT) [19]. For each of the 69 TRs 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). In 60/68 cases with supporting reads from PacBio, at least one read showed evidence of an allele >101bp (51/68 for ONT) (Supplementary Table 2). ONT showed less evidence of expansions, perhaps due to a deletion bias. Both long read technologies exhibit high error rates at homopolymer runs [42], resulting in messy sequence within repeats themselves (Figure 5B)

Finally, for a subset of 12 candidate expansions, we additionally performed capillary electrophoresis to measure TR lengths (Methods, Supplementary Table 5). Capillary results showed evidence of long alleles for the majority (10/12) of TRs (Figure 5C,D, Supplementary Figure 12). Notably, expanded TRs proved difficult to amplify and capillary results in many cases did not clearly indicate two distinct allele lengths. Further, in some cases GangSTR, PacBio, and ONT gave discordant results, with either strikingly different repeat lengths or an ambiguous signal that could not be resolved using capillary electrophoresis (Supplementary Figure 12). Still, the majority of long TRs identified by GangSTR were validated by at least one of these orthogonal technologies. Taken together, these results demonstrate GangSTR's ability to identify novel expanded TRs from genome-wide data and highlight the challenges in precisely validating TR lengths at these loci.

4 DISCUSSION

4.1 A unified framework for genotyping a wide range of TRs

Our study presents GangSTR, a novel tool for genotyping TRs from NGS data. GangSTR is a flexible tool that can be used for a variety of applications, including genome-wide TR genotyping, targeted detection of TR expansions at known pathogenic loci, and genome-wide discovery of novel TR expansions. We show that GangSTR outperforms existing tools in both speed and accuracy in a range of settings using simulated and real NGS datasets. We applied GangSTR genome-wide to genotype hundreds of thousands of TRs in a deeply sequenced healthy trio. We identified dozens of long repeat alleles which were confirmed by orthogonal long read and capillary electrophoresis technologies.

GangSTR outperforms state of the art methods for characterizing TR expansions from NGS (Figure 3). Our targeted simulation analyses demonstrate that GangSTR produced accurate TR length estimates in a range of settings, including unexpanded genotypes and genotypes that are either heterozygous or homozygous for long alleles. GangSTR's advantage becomes more pronounced for TRs with longer normal-length alleles. ExpansionHunter [11] does not accurately genotype TRs heterozygous for two long alleles since its model is primarily based on sequencing coverage. Our model overcomes this limitation by incorporating orthogonal information available from spanning read pairs. While Tredparse [38] similarly models observed fragment lengths for spanning read pairs, it does not analyze read pairs where both reads are mapped to off-target regions, and cannot genotype TRs longer than the fragment length.

Beyond TR expansions implicated in Mendelian disorders, mounting evidence suggests that thousands of TRs genome-wide contribute to polygenic phenotypes such as gene expression [13]. Accurate genome-wide TR genotyping will be critical for performing association studies to identify these TRs and quantify their contribution to common disease. GangSTR extends our existing methods for genome-wide TR genotyping to accommodate repeats longer than the read length and identifies tens of thousands of TRs that were missed by HipSTR [44].

Genome-wide analysis additionally allows for identifying novel pathogenic TR expansions or expansions present in healthy genomes. While existing tools allow for this, they do not produce genome-wide TR length estimates. STRetch [10] identifies novel expansions, but requires a time-consuming and memory intensive step to realign raw reads to a modified reference sequence containing decoy regions. Due to compute requirements. performing realignment is often not feasible to implement in high-throughput pipelines. Additionally, STRetch only identifies a subset of TR alleles that are expanded from the repeat sequence, and thus cannot

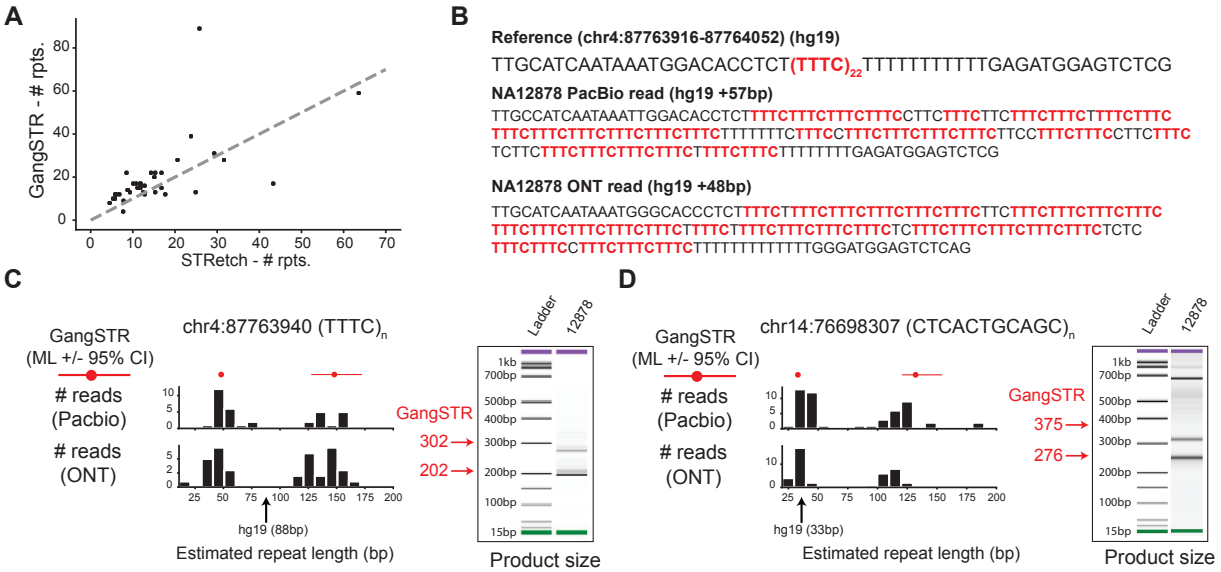

Figure 5: **Discovery and validation of genome-wide TR expansions.** A. Comparison of STRetch and GangSTR estimated repeat lengths. The x-axis shows the estimated repeat number returned by STRetch. The y-axis shows the estimated repeat number of the longest of two alleles reported as the maximum likelihood genotype by GangSTR. Only TRs called by both tools and passing all GangSTR filters are shown. The gray dashed line shows the diagonal. B. 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. C, D. For each of the TRs 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 TR 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.

be used to obtain accurate diploid TR lengths. exSTRa [39] can also be used to find novel expansions, but requires a matched control cohort to identify expansions and reports only expansion status, rather than TR length estimates. On the other hand, GangSTR generates unbiased TR length estimates genome-wide, which can be used in diverse downstream applications such as association testing or discovery of Mendelian disease loci. Further, GangSTR is far more efficient, taking around 16.5 CPU-hours to run on a single genome compared to days for competing methods.

4.2 Remaining challenges in TR genotyping

Genome-wide TR genotyping still faces several important limitations. First, all tools described here, including GangSTR, require a TR panel based on the reference genome as input. Thus they are not able to genotype TRs that are not properly assembled in the reference genome. Additionally, TRs with complex structures, such as sequence imperfections, highly repetitive flanking regions, or multiple different adjacent repeating motifs, are ambiguous to define and their boundaries depend highly on the choice of parameters used to create the reference. Complex TRs are a source of errors in GangSTR genotypes. Our realignment step relies on aligning reads to an artificial reference created for each possible TR allele by stitching together perfect repetitions of the TR motif. Because of this design choice, repetitive motifs in the flanking regions surrounding a TR locus can reduce robustness of the realignment step. We attempt to filter most of these regions from our reference set to avaoid TRs that cannot be reliably called. A more complex model is required

to account for these regions.

Most previous tools focused on STRs with motifs up to 6bp. Here, we have expanded our reference to include VNTRs with motifs up to 20bp. This limit can theoretically be expanded. However, longer motifs tend to have more complex imperfections. Additionally, several aspects of GangSTR's model rely on identifying several copies of a repeat unit in a single read (e.g. enclosing and flanking reads). Thus accuracy is likely to decrease slightly at longer motifs.

Second, due to a lack of large ground truth datasets our validation experiments relied heavily on simulated data. These simulations assume uniform coverage and do not capture many error modes present in real data such as PCR, GC biases, or DNA degradation.

Third, some TRs are still not adequately captured by short reads. For example, TRs in regions with extremely high GC content are often very poorly covered due to biases induced by PCR and other sequencing steps. Furthermore, TRs with highly repetitive flanking regions are still inaccessible due to poor sequence alignment of anchoring or spanning reads. Additionally, while GangSTR can genotype TRs well beyond the fragment length, it still produces noisy estimates at extremely long TRs (e.g. thousands of bp), especially when both alleles are long. We suspect this is primarily due to variance in FRR coverage which grows linearly with total repeat length. While some of these challenges may be overcome with improved modeling techniques, some TRs are likely to remain out of reach using NGS.

Finally, 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 [37, 22], rendering the notion of a "correct" genotype meaningless. Indeed, for several loci we saw evidence of a spectrum of repeat numbers in all technologies tested. GangSTR could be extended in the future to incorporate somatic mosaicism into its model.

Some of the limitations mentioned above could be overcome using long read technologies such as PacBio or ONT. However, we focused on Illumina short reads here as Illumina is rapidly becoming the clinical standard and remains unmatched in cost and accuracy. It is likely that hybrid approaches combining both short and long read data will provide the greatest accuracy.

5 CONCLUSION

Overall, GangSTR allows rapid and accurate genotyping of both short and expanded TRs and can be readily applied to large NGS cohorts to enable novel genetic discoveries across a broad range of applications.

6 ACKNOWLEDGEMENTS

We thank Vineet Bafna, Vikas Bansal, Egor Dolzhenko, and Alon Goren for helpful comments on the method and manuscript. We thank Harriet Dashnow for advice on running STRetch.

Whole exome sequencing for Huntington's Disease patients was obtained from DbGaP accession phs000371.v1.p1, which was supported by the Massachusetts Huntington's Disease Center without Walls (NS16367) and the PREDICT-HD study of the Huntington Study Group (NS40068).

6.1 Funding

This work was supported in part by the National Institutes of Health [DP5OD024577, R01HG010149 to M.G.]. 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 [ACI-1548562].

6.1.1 Conflict of interest statement.

None declared.

References

[1] S A Ahrendt, P A Decker, K Doffek, B Wang, L Xu, M J Demeure, J Jen, and D Sidransky. Microsatellite instability at selected tetranucleotide repeats is associated with p53 mutations in non-small cell lung cancer. *Cancer Res.*, 60(9):2488–2491, May 2000.

[2] Albino Bacolla, Jacquelynn E Larson, Jack R Collins, Jian Li, Aleksandar Milosavljevic, Peter D Stenson, David N Cooper, and Robert D Wells. Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties. *Genome Res.*, 18(10):1545–1553, October 2008.

[3] M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, M. C. Wendl, J. Kim, B. Reardon, P. K. Ng, K. J. Jeong, S. Cao, Z. Wang, J. Gao, Q. Gao, F. Wang, E. M. Liu, L. Mularoni, C. Rubio-Perez, N. Nagarajan, I. Cortes-Ciriano, D. C. Zhou, W. W. Liang, J. M. Hess, V. D. Yellapantula, D. Tamborero, A. Gonzalez-Perez, C. Suphavilai, J. Y. Ko, E. Khurana, P. J. Park, E. M. Van Allen, H. Liang, M. S. Lawrence, A. Godzik, N. Lopez-Bigas, J. Stuart, D. Wheeler, G. Getz, K. Chen, A. J. Lazar, G. B. Mills, R. Karchin, L. Ding, S. J. Caesar-Johnson, J. A. Demchok, I. Felau, M. Kasapi, M. L. Ferguson, C. M. Hutter, H. J. Sofia, R. Tarnuzzer, Z. Wang, L. Yang, J. C. Zenklusen, J. J. Zhang, S. Chudamani, J. Liu, L. Lolla, R. Naresh, T. Pihl, Q. Sun, Y. Wan, Y. Wu, J. Cho, T. DeFreitas, S. Frazer, N. Gehlenborg, G. Getz, D. I. Heiman, J. Kim, M. S. Lawrence, P. Lin, S. Meier, M. S. Noble, G. Saksena, D. Voet, H. Zhang, B. Bernard, N. Chambwe, V. Dhankani, T. Knijnenburg, R. Kramer, K. Leinonen, Y. Liu, M. Miller, S. Reynolds, I. Shmulevich, V. Thorsson, W. Zhang, R. Akbani, B. M. Broom, A. M. Hegde, Z. Ju, R. S. Kanchi, A. Korkut, J. Li, H. Liang, S. Ling, W. Liu, Y. Lu, G. B. Mills, K. S. Ng, A. Rao, M. Ryan, J. Wang, J. N. Weinstein, J. Zhang, A. Abeshouse, J. Armenia, D. Chakravarty, W. K. Chatila, I. de Bruijn, J. Gao, B. E. Gross, Z. J. Heins, R. Kundra, K. La, M. Ladanyi, A. Luna, M. G. Nissan, A. Ochoa, S. M. Phillips, E. Reznik, F. Sanchez-Vega, C. Sander, N. Schultz, R. Sheridan, S. O. Sumer, Y. Sun, B. S. Taylor, J. Wang, H. Zhang, P. Anur, M. Peto, P. Spellman, C. Benz, J. M. Stuart, C. K. Wong, C. Yau, D. N. Hayes, J. S. Parker, M. D. Wilkerson, A. Ally, M. Balasundaram, R. Bowlby, D. Brooks, R. Carlsen, E. Chuah, N. Dhalla, R. Holt, S. J. M. Jones, K. Kasaian, D. Lee, Y. Ma, M. A. Marra, M. Mayo, R. A. Moore, A. J. Mungall, K. Mungall, A. G. Robertson, S. Sadeghi, J. E. Schein, P. Sipahimalani, A. Tam, N. Thiessen, K. Tse, T. Wong, A. C. Berger, R. Beroukhim, A. D. Cherniack, C. Cibulskis, S. B. Gabriel, G. F. Gao, G. Ha, M. Meyerson, S. E. Schumacher, J. Shih, M. H. Kucherlapati, R. S. Kucherlapati, S. Baylin, L. Cope, L. Danilova, M. S. Bootwalla, P. H. Lai, D. T. Maglinte, D. J. Van Den Berg, D. J. Weisenberger, J. T. Auman, S. Balu, T. Bodenheimer, C. Fan, K. A. Hoadley, A. P. Hoyle, S. R. Jefferys, C. D. Jones, S. Meng, P. A. Mieczkowski, L. E. Mose, A. H. Perou, C. M. Perou, J. Roach, Y. Shi, J. V. Simons, T. Skelly, M. G. Soloway, D. Tan, U. Veluvolu, H. Fan, T. Hinoue, P. W. Laird, H. Shen, W. Zhou, M. Bellair, K. Chang, K. Covington, C. J. Creighton, H. Dinh, H. Doddapaneni, L. A. Donehower, J. Drummond, R. A. Gibbs, R. Glenn, W. Hale, Y. Han, J. Hu, V. Korchina, S. Lee, L. Lewis, W. Li, X. Liu, M. Morgan, D. Morton, D. Muzny, J. Santibanez, M. Sheth, E. Shinbrot, L. Wang, M. Wang, D. A. Wheeler, L. Xi, F. Zhao, J. Hess, E. L. Appelbaum, M. Bailey, M. G. Cordes, L. Ding, C. C. Fronick, L. A. Fulton, R. S. Fulton, C. Kandoth, E. R. Mardis, M. D. McLellan, C. A. Miller, H. K. Schmidt, R. K. Wilson, D. Crain, E. Curley, J. Gardner, K. Lau, D. Mallery, S. Morris, J. Paulauskis, R. Penny, C. Shelton, T. Shelton, M. Sherman, E. Thompson, P. Yena, J. Bowen, J. M. Gastier-Foster, M. Gerken, K. M. Leraas, T. M. Lichtenberg, N. C. Ramirez, L. Wise, E. Zmuda, N. Corcoran, T. Costello, C. Hovens, A. L. Carvalho, A. C. de Carvalho, J. H. Fregnani, A. Longatto-Filho, R. M. Reis, C. Scapulatempo-Neto, H. C. S. Silveira, D. O. Vidal, A. Burnette, J. Eschbacher, B. Hermes, A. Noss, R. Singh, M. L. Anderson, P. D. Castro, M. Ittmann, D. Huntsman, B. Kohl, X. Le, R. Thorp, C. Andry, E. R. Duffy, V. Lyadov, O. Paklina, G. Setdikova, A. Shabunin, M. Tavobilov, C. McPherson, R. Warnick, R. Berkowitz, D. Cramer, C. Feltmate, N. Horowitz, A. Kibel, M. Muto, C. P. Raut, A. Malykh, J. S. Barnholtz-Sloan, W. Barrett, K. Devine, J. Fulop, Q. T. Ostrom, K. Shimmel, Y. Wolinsky, A. E. Sloan, A. De Rose, F. Giuliante, M. Goodman, B. Y. Karlan, C. H. Hagedorn, J. Eckman, J. Harr, J. Myers, K. Tucker, L. A. Zach, B. Deyarmin, H. Hu, L. Kvecher, C. Larson, R. J. Mural, S. Somiari, A. Vicha, T. Zelinka, J. Bennett, M. Iacocca, B. Rabeno, P. Swanson, M. Latour, L. Lacombe, B. Tetu, A. Bergeron, M. McGraw, S. M. Staugaitis, J. Chabot, H. Hibshoosh, A. Sepulveda, T. Su, T. Wang, O. Potapova, O. Voronina, L. Desjardins, O. Mariani, S. Roman-Roman, X. Sastre, M. H. Stern, F. Cheng, S. Signoretti, A. Berchuck, D. Bigner, E. Lipp, J. Marks, S. McCall, R. McLendon, A. Secord, A. Sharp, M. Behera, D. J. Brat, A. Chen, K. Delman, S. Force, F. Khuri, K. Magliocca,

S. Maithel, J. J. Olson, T. Owonikoko, A. Pickens, S. Ramalingam, D. M. Shin, G. Sica, E. G. Van Meir, H. Zhang, W. Eijckenboom, A. Gillis, E. Korpershoek, L. Looijenga, W. Oosterhuis, H. Stoop, K. E. van Kessel, E. C. Zwarthoff, C. Calatozzolo, L. Cuppini, S. Cuzzubbo, F. DiMeco, G. Finocchiaro, L. Mattei, A. Perin, B. Pollo, C. Chen, J. Houck, P. Lohavanichbutr, A. Hartmann, C. Stoehr, R. Stoehr, H. Taubert, S. Wach, B. Wullich, W. Kycler, D. Murawa, M. Wiznerowicz, K. Chung, W. J. Edenfield, J. Martin, E. Baudin, G. Bubley, R. Bueno, A. De Rienzo, W. G. Richards, S. Kalkanis, T. Mikkelsen, H. Noushmehr, L. Scarpace, N. Girard, M. Aymerich, E. Campo, E. Gine, A. L. Guillermo, N. Van Bang, P. T. Hanh, B. D. Phu, Y. Tang, H. Colman, K. Evason, P. R. Dottino, J. A. Martignetti, H. Gabra, H. Juhl, T. Akeredolu, S. Stepa, D. Hoon, K. Ahn, K. J. Kang, F. Beuschlein, A. Breggia, M. Birrer, D. Bell, M. Borad, A. H. Bryce, E. Castle, V. Chandan, J. Cheville, J. A. Copland, M. Farnell, T. Flotte, N. Giama, T. Ho, M. Kendrick, J. P. Kocher, K. Kopp, C. Moser, D. Nagorney, D. O'Brien, B. P. O'Neill, T. Patel, G. Petersen, F. Que, M. Rivera, L. Roberts, R. Smallridge, T. Smyrk, M. Stanton, R. H. Thompson, M. Torbenson, J. D. Yang, L. Zhang, F. Brimo, J. A. Ajani, A. M. A. Gonzalez, C. Behrens, J. Bondaruk, R. Broaddus, B. Czerniak, B. Esmaeli, J. Fujimoto, J. Gershenwald, C. Guo, A. J. Lazar, C. Logothetis, F. Meric-Bernstam, C. Moran, L. Ramondetta, D. Rice, A. Sood, P. Tamboli, T. Thompson, P. Troncoso, A. Tsao, I. Wistuba, C. Carter, L. Haydu, P. Hersey, V. Jakrot, H. Kakavand, R. Kefford, K. Lee, G. Long, G. Mann, M. Quinn, R. Saw, R. Scolyer, K. Shannon, A. Spillane, J. Stretch, M. Synott, J. Thompson, J. Wilmott, H. Al-Ahmadie, T. A. Chan, R. Ghossein, A. Gopalan, D. A. Levine, V. Reuter, S. Singer, B. Singh, N. V. Tien, T. Broudy, C. Mirsaidi, P. Nair, P. Drwiega, J. Miller, J. Smith, H. Zaren, J. W. Park, N. P. Hung, E. Kebebew, W. M. Linehan, A. R. Metwalli, K. Pacak, P. A. Pinto, M. Schiffman, L. S. Schmidt, C. D. Vocke, N. Wentzensen, R. Worrell, H. Yang, M. Moncrieff, C. Goparaju, J. Melamed, H. Pass, N. Botnariuc, I. Caraman, M. Cernat, I. Chemencedji, A. Clipca, S. Doruc, G. Gorincioi, S. Mura, M. Pirtac, I. Stancul, D. Tcaciuc, M. Albert, I. Alexopoulou, A. Arnaout, J. Bartlett, J. Engel, S. Gilbert, J. Parfitt, H. Sekhon, G. Thomas, D. M. Rassl, R. C. Rintoul, C. Bifulco, R. Tamakawa, W. Urba, N. Hayward, H. Timmers, A. Antenucci, F. Facciolo, G. Grazi, M. Marino, R. Merola, R. de Krijger, A. P. Gimenez-Roqueplo, A. Piche, S. Chevalier, G. McKercher, K. Birsoy, G. Barnett, C. Brewer, C. Farver, T. Naska, N. A. Pennell, D. Raymond, C. Schilero, K. Smolenski, F. Williams, C. Morrison, J. A. Borgia, M. J. Liptay, M. Pool, C. W. Seder, K. Junker, L. Omberg, M. Dinkin, G. Manikhas, D. Alvaro, M. C. Bragazzi, V. Cardinale, G. Carpino, E. Gaudio, D. Chesla, S. Cottingham, M. Dubina, F. Moiseenko, R. Dhanasekaran, K. F. Becker, K. P. Janssen, J. Slotta-Huspenina, M. H. Abdel-Rahman, D. Aziz, S. Bell, C. M. Cebulla, A. Davis, R. Duell, J. B. Elder, J. Hilty, B. Kumar, J. Lang, N. L. Lehman, R. Mandt, P. Nguyen, R. Pilarski, K. Rai, L. Schoenfield, K. Senecal, P. Wakely, P. Hansen, R. Lechan, J. Powers, A. Tischler, W. E. Grizzle, K. C. Sexton, A. Kastl, J. Henderson, S. Porten, J. Waldmann, M. Fassnacht, S. L. Asa, D. Schadendorf, M. Couce, M. Graefen, H. Huland, G. Sauter, T. Schlomm, R. Simon, P. Tennstedt, O. Olabode, M. Nelson, O. Bathe, P. R. Carroll, J. M. Chan, P. Disaia, P. Glenn, R. K. Kelley, C. N. Landen, J. Phillips, M. Prados, J. Simko, K. Smith-McCune, S. VandenBerg, K. Roggin, A. Fehrenbach, A. Kendler, S. Sifri, R. Steele, A. Jimeno, F. Carey, I. Forgie, M. Mannelli, M. Carney, B. Hernandez, B. Campos, C. Herold-Mende, C. Jungk, A. Unterberg, A. von Deimling, A. Bossler, J. Galbraith, L. Jacobus, M. Knudson, T. Knutson, D. Ma, M. Milhem, R. Sigmund, A. K. Godwin, R. Madan, H. G. Rosenthal, C. Adebamowo, S. N. Adebamowo, A. Boussioutas, D. Beer, T. Giordano, A. M. Mes-Masson, F. Saad, T. Bocklage, L. Landrum, R. Mannel, K. Moore, K. Moxley, R. Postier, J. Walker, R. Zuna, M. Feldman, F. Valdivieso, R. Dhir, J. Luketich, E. M. M. Pinero, M. Quintero-Aguilo, C. G. Carlotti, J. S. Dos Santos, R. Kemp, A. Sankarankuty, D. Tirapelli, J. Catto, K. Agnew, E. Swisher, J. Creaney, B. Robinson, C. S. Shelley, E. M. Godwin, S. Kendall, C. Shipman, C. Bradford, T. Carey, A. Haddad, J. Moyer, L. Peterson, M. Prince, L. Rozek, G. Wolf, R. Bowman, K. M. Fong, I. Yang, R. Korst, W. K. Rathmell, J. L. Fantacone-Campbell, J. A. Hooke, A. J. Kovatich, C. D. Shriver, J. DiPersio, B. Drake, R. Govindan, S. Heath, T. Ley, B. Van Tine, P. Westervelt, M. A. Rubin, J. I. Lee, N. D. Aredes, and A. Mariamidze. Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell*, 173(2):371–385, Apr 2018.

[4] Mehrdad Bakhtiari, Sharona Shleizer-Burko, Melissa Gymrek, Vikas Bansal, and Vineet Bafna. Targeted genotyping of variable number tandem repeats with adVNTR, 2017.

[5] S. Benonisdottir, A. Oddsson, A. Helgason, R. P. Kristjansson, G. Sveinbjornsson, A. Oskarsdottir, G. Thorleifsson, O. B. Davidsson, G. A. Arnadottir, G. Sulem, B. O. Jensson, H. Holm, K. F. Alexandersson, L. Tryggvadottir, G. B. Walters, S. A. Gudjonsson, L. D. Ward, J. K. Sigurdsson, P. D. Iordache,

M. L. Frigge, T. Rafnar, A. Kong, G. Masson, H. Helgason, U. Thorsteinsdottir, D. F. Gudbjartsson, P. Sulem, and K. Stefansson. Epigenetic and genetic components of height regulation. *Nat Commun*, 7:13490, 11 2016.

[6] G Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, 27(2):573–580, January 1999.

[7] W. M. Brandler, D. Antaki, M. Gujral, M. L. Kleiber, J. Whitney, M. S. Maile, O. Hong, T. R. Chapman, S. Tan, P. Tandon, T. Pang, S. C. Tang, K. K. Vaux, Y. Yang, E. Harrington, S. Juul, D. J. Turner, B. Thiruvahindrapuram, G. Kaur, Z. Wang, S. F. Kingsmore, J. G. Gleeson, D. Bisson, B. Kakaradov, A. Telenti, J. C. Venter, R. Corominas, C. Toma, B. Cormand, I. Rueda, S. Guijarro, K. S. Messer, C. M. Nievergelt, M. J. Arranz, E. Courchesne, K. Pierce, A. R. Muotri, L. M. Iakoucheva, A. Hervas, S. W. Scherer, C. Corsello, and J. Sebat. Paternally inherited cis-regulatory structural variants are associated with autism. *Science*, 360(6386):327–331, 04 2018.

[8] C. Chiang, A. J. Scott, J. R. Davis, E. K. Tsang, X. Li, Y. Kim, T. Hadzic, F. N. Damani, L. Ganel, S. B. Montgomery, A. Battle, D. F. Conrad, and I. M. Hall. The impact of structural variation on human gene expression. *Nat. Genet.*, 49(5):692–699, May 2017.

[9] H. Dashnow, M. Lek, B. Phipson, A. Halman, S. Sadedin, A. Lonsdale, M. Davis, P. Lamont, J. S. Clayton, N. G. Laing, D. G. MacArthur, and A. Oshlack. STRetch: detecting and discovering pathogenic short tandem repeat expansions. *Genome Biol.*, 19(1):121, 08 2018.

[10] Harriet Dashnow, Monkol Lek, Belinda Phipson, Andreas Halman, Mark Davis, Phillipa Lamont, Joshua Clayton, Nigel Laing, Daniel MacArthur, and Alicia Oshlack. STRetch: detecting and discovering pathogenic short tandem repeats expansions, 2017.

[11] Egor Dolzhenko, Joke J F A van Vugt, Richard J Shaw, Mitchell A Bekritsky, Marka van Blitterswijk, Giuseppe Narzisi, Subramanian S Ajay, Vani Rajan, Bryan R Lajoie, Nathan H Johnson, Zoya Kingsbury, Sean J Humphray, Raymond D Schellevis, William J Brands, Matt Baker, Rosa Rademakers, Maarten Kooyman, Gijs H P Tazelaar, Michael A van Es, Russell McLaughlin, William Sproviero, Aleksey Shatunov, Ashley Jones, Ahmad Al Khleifat, Alan Pittman, Sarah Morgan, Orla Hardiman, Ammar Al-Chalabi, Chris Shaw, Bradley Smith, Edmund J Neo, Karen Morrison, Pamela J Shaw, Catherine Reeves, Lara Winterkorn, Nancy S Wexler, US–Venezuela Collaborative Research Group, David E Housman, Christopher W Ng, Alina L Li, Ryan J Taft, Leonard H van den Berg, David R Bentley, Jan H Veldink, and Michael A Eberle. Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Res.*, 27(11):1895–1903, November 2017.

[12] Melissa Gymrek, David Golan, Saharon Rosset, and Yaniv Erlich. lobSTR: A short tandem repeat profiler for personal genomes. *Genome Res.*, 22(6):1154–1162, June 2012.

[13] Melissa Gymrek, Thomas Willems, Audrey Guilmatre, Haoyang Zeng, Barak Markus, Stoyan Georgiev, Mark J Daly, Alkes L Price, Jonathan K Pritchard, Andrew J Sharp, and Yaniv Erlich. Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.*, 48(1):22–29, January 2016.

[14] Anthony J Hannan. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. *Trends Genet.*, 26(2):59–65, February 2010.

[15] Anthony J Hannan. Tandem repeats mediating genetic plasticity in health and disease. *Nat. Rev. Genet.*, 19(5):286–298, May 2018.

[16] Gareth Highnam, Christopher Franck, Andy Martin, Calvin Stephens, Ashwin Puthige, and David Mittelman. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res.*, 41(1):e32, January 2013.

[17] Jessica Hunter, Oliver Rivero-Arias, Angel Angelov, Edward Kim, Iain Fotheringham, and Jose Leal. Epidemiology of fragile X syndrome: a systematic review and meta-analysis. *Am. J. Med. Genet. A*, 164A(7):1648–1658, July 2014.

[18] H. Ishiura, K. Doi, J. Mitsui, J. Yoshimura, M. K. Matsukawa, A. Fujiyama, Y. Toyoshima, A. Kakita, H. Takahashi, Y. Suzuki, S. Sugano, W. Qu, K. Ichikawa, H. Yurino, K. Higasa, S. Shibata, A. Mitsue, M. Tanaka, Y. Ichikawa, Y. Takahashi, H. Date, T. Matsukawa, J. Kanda, F. K. Nakamoto, M. Higashihara, K. Abe, R. Koike, M. Sasagawa, Y. Kuroha, N. Hasegawa, N. Kanesawa, T. Kondo, T. Hitomi, M. Tada, H. Takano, Y. Saito, K. Sanpei, O. Onodera, M. Nishizawa, M. Nakamura, T. Yasuda, Y. Sakiyama, M. Otsuka, A. Ueki, K. I. Kaida, J. Shimizu, R. Hanajima, T. Hayashi, Y. Terao, S. Inomata-Terada, M. Hamada, Y. Shirota, A. Kubota, Y. Ugawa, K. Koh, Y. Takiyama, N. Ohsawa-Yoshida, S. Ishiura, R. Yamasaki, A. Tamaoka, H. Akiyama, T. Otsuki, A. Sano, A. Ikeda, J. Goto, S. Morishita, and S. Tsuji. Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nat. Genet.*, 50(4):581–590, Apr 2018.

[19] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. The oxford nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol.*, 17(1):239, November 2016.

[20] Steven G Johnson. The nlopt nonlinear-optimization package, 2014.

[21] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at UCSC. *Genome Res.*, 12(6):996–1006, June 2002.

[22] Cara Kraus-Perrotta and Sarita Lagalwar. Expansion, mosaicism and interruption: mechanisms of the CAG repeat mutation in spinocerebellar ataxia type 1. *Cerebellum Ataxias*, 3:20, November 2016.

[23] Snædís Kristmundsdóttir, Brynja D Sigurpálsdóttir, Birte Kehr, and Bjarni V Halldórsson. popSTR: population-scale detection of STR variants. *Bioinformatics*, September 2016.

[24] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. March 2013.

[25] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009.

[26] E. W. Loomis, J. S. Eid, P. Peluso, J. Yin, L. Hickey, D. Rank, S. McCalmon, R. J. Hagerman, F. Tassone, and P. J. Hagerman. Sequencing the unsequenceable: expanded CGG-repeat alleles of the fragile X gene. *Genome Res.*, 23(1):121–128, Jan 2013.

[27] S. Mallick, H. Li, M. Lipson, I. Mathieson, M. Gymrek, F. Racimo, M. Zhao, N. Chennagiri, S. Nordenfelt, A. Tandon, P. Skoglund, I. Lazaridis, S. Sankararaman, Q. Fu, N. Rohland, G. Renaud, Y. Erlich, T. Willems, C. Gallo, J. P. Spence, Y. S. Song, G. Poletti, F. Balloux, G. van Driem, P. de Knijff, I. G. Romero, A. R. Jha, D. M. Behar, C. M. Bravi, C. Capelli, T. Hervig, A. Moreno-Estrada, O. L. Posukh, E. Balanovska, O. Balanovsky, S. Karachanak-Yankova, H. Sahakyan, D. Toncheva, L. Yepiskoposyan, C. Tyler-Smith, Y. Xue, M. S. Abdullah, A. Ruiz-Linares, C. M. Beall, A. Di Rienzo, C. Jeong, E. B. Starikovskaya, E. Metspalu, J. Parik, R. Villems, B. M. Henn, U. Hodoglugil, R. Mahley, A. Sajantila, G. Stamatoyannopoulos, J. T. Wee, R. Khusainova, E. Khusnutdinova, S. Litvinov, G. Ayodo, D. Comas, M. F. Hammer, T. Kivisild, W. Klitz, C. A. Winkler, D. Labuda, M. Bamshad, L. B. Jorde, S. A. Tishkoff, W. S. Watkins, M. Metspalu, S. Dryomov, R. Sukernik, L. Singh, K. Thangaraj, S. Paabo, J. Kelso, N. Patterson, and D. Reich. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624):201–206, Oct 2016.

[28] Alice McCarthy. Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chem. Biol.*, 17(7):675–676, July 2010.

[29] S. M. Mirkin. Expandable DNA repeats and human disease. *Nature*, 447(7147):932–940, Jun 2007.

[30] M. O. Pollard, D. Gurdasani, A. J. Mentzer, T. Porter, and M. S. Sandhu. Long reads: their purpose and place. *Hum. Mol. Genet.*, 27(R2):R234–R241, Aug 2018.

[31] Michael JD Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in optimization and numerical analysis*, pages 51–67. Springer, 1994.

[32] Maximilian O Press, Keisha D Carlson, and Christine Queitsch. The overdue promise of short tandem repeat variation for heritability. *Trends Genet.*, 30(11):504–512, November 2014.

[33] Tamara Pringsheim, Katie Wiltshire, Lundy Day, Jonathan Dykeman, Thomas Steeves, and Nathalie Jette. The incidence and prevalence of huntington's disease: A systematic review and meta-analysis. *Mov. Disord.*, 27(9):1083–1091, 2012.

[34] Javier Quilez, Audrey Guilmatre, Paras Garg, Gareth Highnam, Melissa Gymrek, Yaniv Erlich, Ricky S Joshi, David Mittelman, and Andrew J Sharp. Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.*, 44(8):3750–3762, 2016.

[35] Luis Ruano, Claudia Melo, M Carolina Silva, and Paula Coutinho. The global epidemiology of hereditary ataxia and spastic paraplegia: a systematic review of prevalence studies. *Neuroepidemiology*, 42(3):174–183, March 2014.

[36] Shubham Saini, Ileena Mitra, and Melissa Gymrek. A reference haplotype panel for genome-wide imputation of short tandem repeats, 2018.

[37] Meera Swami, Audrey E Hendricks, Tammy Gillis, Tiffany Massood, Jayalakshmi Mysore, Richard H Myers, and Vanessa C Wheeler. 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(16):3039–3047, 2009.

[38] Haibao Tang, Ewen F Kirkness, Christoph Lippert, William H Biggs, Martin Fabani, Ernesto Guzman, Smriti Ramakrishnan, Victor Lavrenko, Boyko Kakaradov, Claire Hou, Barry Hicks, David Heckerman, Franz J Och, C Thomas Caskey, J Craig Venter, and Amalio Telenti. Profiling of Short-Tandem-Repeat disease alleles in 12,632 human whole genomes. *Am. J. Hum. Genet.*, 101(5):700–715, November 2017.

[39] Rick M. Tankard, Mark F Bennett, Peter Degorski, Martin B. Delatycki, Paul J. Lockhart, and Melanie Bahlo. Detecting tandem repeat expansions in cohorts sequenced with short-read sequencing data. *bioRxiv*, 2018.

[40] Rick M Tankard, Martin B Delatycki, Paul J Lockhart, and Melanie Bahlo. Detecting known repeat expansions with standard protocol next generation sequencing, towards developing a single screening test for neurological repeat expansion disorders, 2017.

[41] P. M. Visscher, N. R. Wray, Q. Zhang, P. Sklar, M. I. McCarthy, M. A. Brown, and J. Yang. 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.*, 101(1):5–22, Jul 2017.

[42] Jason L Weirather, Mariateresa de Cesare, Yunhao Wang, Paolo Piazza, Vittorio Sebastiano, Xiu-Jie Wang, David Buck, and Kin Fai Au. Comprehensive comparison of pacific biosciences and oxford nanopore technologies and their applications to transcriptome analysis. *F1000Res.*, 6:100, February 2017.

[43] Thomas Willems, Melissa Gymrek, Gareth Highnam, 1000 Genomes Project Consortium, David Mittelman, and Yaniv Erlich. The landscape of human STR variation. *Genome Res.*, 24(11):1894–1904, November 2014.

[44] Thomas Willems, Dina Zielinski, Jie Yuan, Assaf Gordon, Melissa Gymrek, and Yaniv Erlich. Genome-wide profiling of heritable and de novo STR variations. *Nat. Methods*, April 2017.

[45] K. O. Wrzeszczynski, V. Felice, M. Shah, S. Rahman, A. K. Emde, V. Jobanputra, M. O Frank, and R. B. Darnell. Whole Genome Sequencing-Based Discovery of Structural Variants in Glioblastoma. *Methods Mol. Biol.*, 1741:1–29, 2018.

[46] L Xu, J Chow, J Bonacum, C Eisenberger, S A Ahrendt, M Spafford, L Wu, S M Lee, S Piantadosi, M S Tockman, D Sidransky, and J Jen. Microsatellite instability at AAAG repeat sequences in respiratory tract cancers. *Int. J. Cancer*, 91(2):200–204, January 2001.

[47] Y. Yang, D. M. Muzny, J. G. Reid, M. N. Bainbridge, A. Willis, P. A. Ward, A. Braxton, J. Beuten, F. Xia, Z. Niu, M. Hardison, R. Person, M. R. Bekheirnia, M. S. Leduc, A. Kirby, P. Pham, J. Scull, M. Wang, Y. Ding, S. E. Plon, J. R. Lupski, A. L. Beaudet, R. A. Gibbs, and C. M. Eng. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N. Engl. J. Med.*, 369(16):1502–1511, Oct 2013.

[48] Mengyao Zhao, Wan-Ping Lee, Erik P Garrison, and Gabor T Marth. Ssw library: an simd smith-waterman c/c++ library for use in genomic applications. *PloS one*, 8(12):e82138, 2013.