Title
Methods for studying the genome-wide landscape of tandem repeats

Author
Mousavi, Nima

Publication Date
2021

UNIVERSITY OF CALIFORNIA SAN DIEGO

Methods for studying the genome-wide landscape of tandem repeats

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Intelligent Systems, Robotics, and Control)

by

Nima Mousavi

Committee in charge:

Professor Melissa Gymrek, Chair
Professor Siavash Mirarab, Co-Chair
Professor Vineet Bafna
Professor Vikas Bansal
Professor Niema Moshiri

2021

The dissertation of Nima Mousavi is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my family,

their support and patience gave me the space that I needed to grow.

EPIGRAPH

There is a place where the sidewalk ends,
And before the street begins,
And there the grass grows soft and white,
And there the sun burns crimson bright,
And there the moon-bird rests from his flight
To cool in the peppermint wind.

—Shel Silverstein

TABLE OF CONTENTS

Dissertation Approval Page iii

Dedication iv

Epigraph v

Table of Contents vi

List of Figures ix

List of Tables x

Acknowledgements xi

Vita xii

Abstract of the Dissertation xiii

Chapter 1 Introduction 1
1.1 Introduction to Tandem Repeats 1
1.2 Genotyping of Tandem Repeats 2
1.3 Outline 2

Chapter 2 Profiling the genome-wide landscape of tandem repeat expansions . . 4
2.1 Introduction 5
2.2 Materials and Methods 8
2.2.1 Overview of the GangSTR model 8
2.2.2 Computation of Log Likelihood 9
2.2.2.1 Read pair term 10
2.2.2.2 Class probability 12
2.2.2.3 Read probability 12
2.2.2.4 Repetitive Read Count Term 14
2.2.3 Local Realignment 15
2.2.4 Retrieving reads mapped to off-target regions 15
2.2.5 Optimization 16
2.2.6 Quality metrics 17
2.2.6.1 Bootstrap confidence intervals and standard errors 17
2.2.6.2 Genotype likelihoods and quality score 17
2.2.6.3 Expansion probability 17
2.2.7 Benchmarking using simulated reads 18
2.2.8 Quantifying genotyping performance with RMSE 19

2.2.9 Analysis of genomes and exomes with validated expansions 19
2.2.10 Constructing a genome-wide repeat reference panel . . . 20
2.2.11 Run time evaluation 20
2.2.12 Genome-wide TR analysis in a CEU trio 21
2.2.13 Validating GangSTR using long reads 23
2.2.14 Experimental validation of repeat lengths 23
2.2.15 Code availability 24
2.3 Results 24
2.3.1 GangSTR outperforms existing TR expansion genotypers 24
2.3.2 Genome-wide TR profiling 26
2.3.3 Genome-wide detection of novel TR expansions 28
2.4 Discussion 30
2.4.1 A unified framework for genotyping a wide range of TRs 30
2.4.2 Remaining challenges in TR genotyping 32
2.5 Supplementary Note: GangSTR Model 34
2.5.1 Class probabilities 34
2.5.1.1 Class Probability of Enclosing Reads 34
2.5.1.2 Class Probability of Spanning Reads 35
2.5.1.3 Fragment Length Distribution 35
2.5.1.4 Class Probability of Flanking Reads 37
2.5.1.5 Class Probability of FRRs 38
2.5.2 Read probabilities 39
2.5.2.1 Enclosing Reads 40
2.5.2.2 Flanking Reads 40
2.5.2.3 Spanning Reads 41
2.5.2.4 Fully Repetitive Reads (FRRs) 41

Chapter 3 TRTools: a toolkit for genome-wide analysis of tandem repeats 61
3.1 Introduction 62
3.2 Features and Methods 62
3.2.1 DumpSTR 64
3.2.2 MergeSTR 64
3.2.3 Statistics and QC utilities 65
3.2.4 Python library for data analysis 66
3.3 Discussion 66
3.4 Supplementary Material 67
3.4.1 Datasets 67
3.4.2 TR Harmonizer implementation details 68
3.4.3 Commands for generating figures 71
3.4.4 Supplementary Tables 76

Chapter 4 Applications of Methods for Genotyping and Filtering of Tandem Repeats 78
4.1 Methods for Identifying de-novo Tandem Repeat mutations in individuals affected by Autism Spectrum Disorder 79
4.1.1 Improvements to GangSTR: method for genome-wide genotyping of tandem repeats 79
4.1.2 Evaluating de-novo variant discovery using simulation . . 80
4.2 Method for creating ensemble Tandem Repeats Callset 81
4.2.1 Graph-based Ensemble Variant Caller 82
4.2.1.1 Locus Merger 83
4.2.1.2 Caller QC 83
4.2.1.3 Allele Merger 84
4.2.1.4 Call Merger 84
4.2.2 Validation 85
4.2.3 Future Works 85

Chapter 5 Conclusion 91

Bibliography 92

LIST OF FIGURES

Figure 2.1: Schematic of GangSTR method 10
Figure 2.2: Four classes of informative read pairs 13
Figure 2.3: Evaluation of TR genotypers on real and simulated data at pathogenic repeat expansions 43
Figure 2.4: Genome-wide TR genotyping 44
Figure 2.5: Discovery and validation of genome-wide TR expansions 45
Figure 2.6: Class probabilities as a function of TR length 46
Figure 2.7: Comparison of true vs. estimated repeat number on simulated data for different loci 47
Figure 2.8: Estimation accuracy for simulated samples of HTT vs. read length . . 48
Figure 2.9: Estimation accuracy for simulated samples of HTT vs. average coverage 49
Figure 2.10: Estimation accuracy for simulated samples of HTT vs. fragment length 50
Figure 2.11: Comparison of true vs. estimated repeat number using real HTT exome data 51
Figure 2.12: Running time of GangSTR and ExpansionHunter vs. reference size . . 52
Figure 2.13: Peak memory usage by GangSTR and ExpansionHunter vs. reference size 53
Figure 2.14: Distribution of repeat lengths in NA12878 compared to the hg19 reference. Y-axis is on a log10 scale. 53
Figure 2.15: Distribution of total repeat lengths in NA12878 54
Figure 2.16: Mendelian inheritance of GangSTR genotypes in a CEU trio as a function of the number of informative read pairs 55
Figure 2.17: Discovery and validation of genome-wide TR expansions 56
Figure 2.17: Discovery and validation of genome-wide TR expansions (continued). . 57

Figure 3.1: TRTools visualizations 63

Figure 4.1: Evaluation of MonSTR using simulated data 87
Figure 4.2: Pipeline for creating ensemble TR callset using 1000 Genomes dataset. 88
Figure 4.3: Pointers to the next processed VCF record in each TR VCF file. . . . 88
Figure 4.4: Example of a trivial graph showing a 1-1-1 mapping between alleles from different callers. 89
Figure 4.5: Example of a non-trivial connected component in allele graph. 89
Figure 4.6: Peaks from capillary electrophoresis correspond to TR calls made on a trio of samples. 90

LIST OF TABLES

Table 2.1: Classes of read pairs and features used by existing tools for genotyping TRs from short reads. 9
Table 2.2: Target pathogenic repeats used in benchmarking experiments 44
Table 2.3: Candidate TRs long alleles (>101bp) in NA12878 58
Table 2.4: Enrichment of motifs with long alleles in NA12878 59
Table 2.5: Comparison of STRetch and GangSTR output 59
Table 2.6: Primers for capillary electrophoresis validation. 60

Table 3.1: Summary of current TRTools utilities 65
Table 3.2: TR calling methods currently supported by TRTools 76

ACKNOWLEDGEMENTS

Chapter 2, in full, is a reprint of the material as it appears in Nucleic Acids Research 2019 by Nima Mousavi, Sharona Shleizer-Burko, Richard Yanicky, and Melissa Gymrek. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Bioinformatics 2020 by Nima Mousavi, Jonathan Margoliash, Neha Pusarla, Shubham Saini, Richard Yanicky, and Melissa Gymrek. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 4, in part, is a reprint of the material as it appears in Nature 2021 by Ileena Mitra, Bonnie Huang, Nima Mousavi, Nichole Ma, Michael Lamkin, Richard Yanicky, Sharona Shleizer-Burko, Kirk E. Lohmueller, and Melissa Gymrek. The dissertation author was one of the co-authors and investigative collaborators of this paper.

Chapter 4, in part, is unpublished material by Nima Mousavi, Nichole Ma, Helyaneh Ziaei Jam, Bonnie Huang, Mikhail Maksimov, Jonghun Park, Yunjiang Qiu, Egor Dolzhenko, Vineet Bafna, and Melissa Gymrek. The dissertation author was the primary investigator and author of this paper.

VITA

2011-2015	B. S. in Electrical Engineering , Sharif University of Technology, Tehran
2015-2019	M. S. in Electrical Engineering (Intelligent Systems, Robotics, and Control), University of California San Diego
2015-2021	Ph. D. in Electrical Engineering (Intelligent Systems, Robotics, and Control), University of California San Diego

PUBLICATIONS

Mousavi, N., Shleizer-Burko, S., Yanicky, R., & Gymrek, M. (2019). “Profiling the genome-wide landscape of tandem repeat expansions”. Nucleic Acids Research, 47(15), e90-e90.

Mousavi, N., Margoliash, J., Pusarla, N., Saini, S., Yanicky, R., & Gymrek, M. (2021). “TRTools: a toolkit for genome-wide analysis of tandem repeats”. Bioinformatics, 37(5), 731-733.

Mitra, I., Huang, B., Mousavi, N., Ma, N., Lamkin, M., Yanicky, R., ... & Gymrek, M. (2021). “Patterns of de novo tandem repeat mutations and their role in autism”. Nature, 589(7841), 246-250.

Saini, S., Mitra, I., Mousavi, N., Fotsing, S. F., & Gymrek, M. (2018). “A reference haplotype panel for genome-wide imputation of short tandem repeats”. Nature communications, 9(1), 1-11.

ABSTRACT OF THE DISSERTATION

Methods for studying the genome-wide landscape of tandem repeats

by

Nima Mousavi

Doctor of Philosophy in Electrical Engineering (Intelligent Systems, Robotics, and Control)

University of California San Diego, 2021

Professor Melissa Gymrek, Chair
Professor Siavash Mirarab, Co-Chair

Tandem Repeats (TRs) are a class of genetic variants formed by motifs of 1-20 nucleotides repeating in tandem. Previous studies show that expansion at specific TR loci is the leading cause of dozens of Mendelian disorders such as Huntington's disease and Fragile X syndrome. Furthermore, copy numbers at TR loci are correlated with complex traits such as gene expression. Tandem repeats are highly mutable and therefore a great subject to study genetic diversity. However, current bioinformatics pipelines are often incapable of processing these loci accurately. Challenges in sequencing, alignment, and interpretation have led to TR loci being overlooked in many studies. We have created a method for genome-wide genotyping of TRs and a toolkit for processing, filtering, and quality control of TR callsets. These methods have allowed us and the community to study repeat expansions on a genome-wide scale. In addition, we have applied our work to study de-novo variants contributing to Autism Spectrum Disorder risk and have found multiple candidate TRs. Another application of our methods is the novel tool for creating an ensemble callset of TRs across a large population. Our efforts in creating methods and applying them to various applications have allowed us to gain a better understanding of TRs and their genetic diversity on a population scale.

Chapter 1

Introduction

1.1 Introduction to Tandem Repeats

Tandem repeats (TRs) are a class of repetitive genetic variants that consist of motifs of 1-20 nucleotides repeating in tandem. There are more than a 1.5 million human TR loci [WZY$^+$17b] which form more than 3% of the human genome [SMS03].

Several coding and non-coding tandem repeats throughout the human genome are known to cause disorders. Pathogenic TRs mediate pathogenicity through multiple pathways such as epigenetic dysregulation (Friedreich ataxia and fragile X syndrome), toxic RNA gain of function (myotonic dystrophies and spinocerebellar ataxias), and change of protein function in polyalanine disorders and polyglutamine diseases (such as oculopharyngeal muscular atrophy and Huntington disease) [Han18b]. In addition to pathogenic TRs, TR variations are associated with complex traits such as gene expression [FMW$^+$19].

TRs have a much higher mutation rate compared to other types of genetic variants such as single-nucleotide variants and short insertions and deletions [Han18b]. Majority of TR variation is mediated through polymerase slippage during cell replication [Rya19]. These slippage events can cause the TR region to gain (expansion) or lose (contraction)

TR motifs. Owing to high variation rate, TRs often have an extended range of possible alleles (multiallelic regions) [Han18b].

Therefore, TRs are a highly variable class of genetic variations that contribute to disease and complex traits in humans. However, the complex nature of these regions renders them a complicated subject to study. I discuss some of these challenges in the next section.

1.2 Genotyping of Tandem Repeats

The standard of care in genotyping disease causing TR loci is usually experimental methods such as repeat primed PCR, long-range PCR, or southern blot [CMM$^+$04]. While these methods are highly accurate and provide context in genotyping specific TR loci, they often have high cost and are not easily scalable. To tackle this problem, several bioinformatics methods have been proposed to allow high-throughput genotyping of TRs using sequencing datasets [MSBYG19, DvVS$^+$17, WZY$^+$17a, KEAH19]. These methods rely on short-read whole genome sequencing (WGS) datasets, which allow accurate characterization of the genome with low error rates (0.24% error rate) [PGB$^+$18]. In addition to short-read sequencing, long-read approaches have also been used to genotype tandem repeats [DRDCB$^+$19, LZW$^+$17, GYM$^+$20]. However, long-read technologies have a relatively higher error rate and are not as widely available in clinical and research settings due to their higher cost [MSBYG19]. As a result, my work has been focused on using short-read sequencing datasets to create genome-wide genotype profiles of TRs.

1.3 Outline

Chapter 2 describes the statistical method for genome-wide genotyping of tandem repeats. Chapter 3 describes a toolkit for post processing, filtering, and quality control of tandem repeat callsets. Chapter 4 examines multiple applications of applying the methods described in chapters 2 and 3 to identify de-novo TR mutations in ASD affected individuals and to create a merged TR callset by combining information from several different TR callers.

Chapter 2

Profiling the genome-wide landscape of tandem repeat expansions

Most of this chapter was first published as:

Mousavi, N., Shleizer-Burko, S., Yanicky, R., & Gymrek, M. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic acids research. (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.

2.1 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 [YMR^{+}13], cancer [BTPP^{+}18], and complex traits [BOH^{+}16]. 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 [CSD^{+}17], cancer [WFS^{+}18], and autism spectrum disorder [BAG^{+}18]. 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 [Mir07], such as Huntington's Disease and Fragile X Syndrome, which together affect millions of individuals worldwide [HRAA^{+}14, PWD^{+}12, RMSC14]. 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

[WGH^{+}14]. Recently, thousands of TRs have been shown to play a role in gene regulation [GWG^{+}16, QGG^{+}16] and it is becoming increasingly clear that TRs across the genome are likely to have widespread contributions to complex polygenic traits [PCQ14, Han10, Han18b]. 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 [VWZ^{+}17]. These studies apply linear or logistic regression models at each TR in the genome to test for association between TR copy number and phenotype across a cohort of samples.

Over the last several years, we and others have developed a series of tools for genome-wide genotyping of STRs [GGRE12, WZY^{+}17b, HFM^{+}13, KSKH16] from short reads or targeted genotyping of VNTRs [BSBG^{+}17] 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 [MLL^{+}16], in practice repeats longer than around 70bp are difficult or impossible to genotype due to an insufficient number of enclosing reads. Notably, in our recent genome-wide analysis [SMG18] using HipSTR [WZY^{+}17b], 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 [Han18b]. 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 [TBD^{+}18] 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 [Han18b] and is required for performing genome-wide association studies. STRetch [DLP$^+$18] can perform genome-wide expansion identification but does not analyze short TRs, is limited to motifs of up to 6bp, and is computationally expensive. Tredparse [TKL$^+$17] models multiple aspects of paired end reads but cannot estimate repeat lengths longer than the sequencing fragment length. ExpansionHunter [DvVS$^+$17] 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 [LEP$^+$13] and a complex pentamer repeat implicated in myoclonus epilepsy [IDM$^+$18]. While long reads offer a potential solution to genome-wide TR analysis, NGS remains the gold standard for diagnostic sequencing and population-wide studies due to its low cost and substantially higher throughput [PGM$^+$18]. 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.

2.2 Materials and Methods

2.2.1 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 the 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 2.1).

GangSTR incorporates each of these informative aspects of paired-end read alignments into a single joint likelihood framework (Figure 2.1). 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

Table 2.1: Classes of read pairs and features used by existing tools for genotyping TRs from short reads.

Method	Enclosing	FRR	Spanning	Off-target FRR	Estimates # rpts.	Genome-wide	Estimation limit
LobSTR [GGRE12]	X					X	<Read length
HipSTR [WZY$^+$17b]	X						<Read length
STRetch [DLP$^+$18]		X		X	X	X	Only reports expanded TRs
exSTRa [TDLB17]		X		X		X	Does not estimate TR length
Tredparse [TKL$^+$17]	X	X	X		X		<Fragment length
ExpansionHunter [DvVS$^+$17]	X	X		X	X		Poor performance when both alleles long
GangSTR	X	X	X	X	X	X	Not limited to fragment or read length

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 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).

2.2.2 Computation of Log Likelihood

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

$$\begin{aligned}\mathcal{L}(\langle A,B\rangle) &= logP(R;\langle A,B\rangle)\\ &= \overbrace{log\prod_{r\in R}P(r;\langle A,B\rangle)}^{\mathcal{L}_P}+\overbrace{logP(|FRR|;\langle A,B\rangle)}^{\mathcal{L}_N}\end{aligned} \tag{2.1}$$

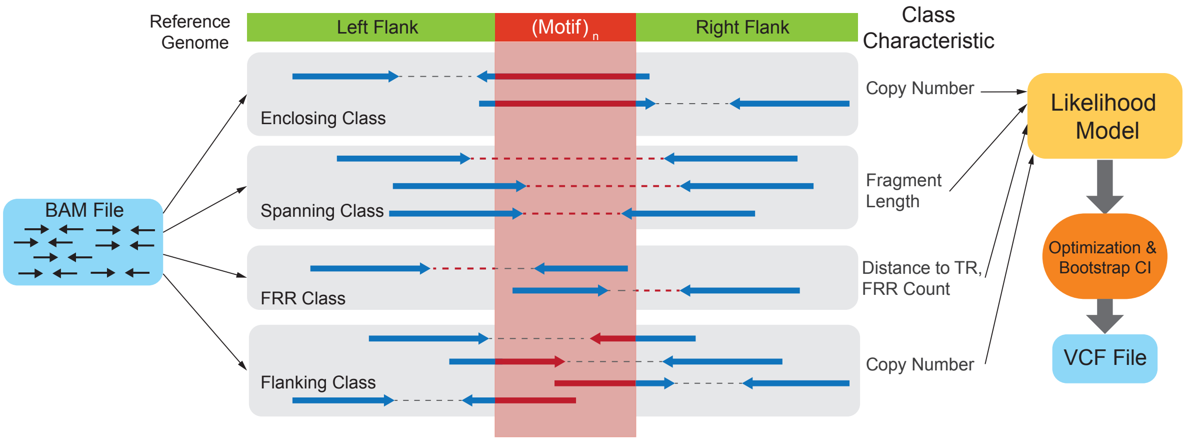

Figure 2.1: Schematic of GangSTR method. Paired end reads from an input set of alignments are separated into various read classes, each of which provides information about the length of the TR in the region. This information is used to find the maximum likelihood diploid genotype and confidence interval on the repeat length. Results are reported in a VCF file.

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.

2.2.2.1 Read pair term

The first term in (2.1) 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:

$$\begin{aligned}
\mathcal{L}_P &= log\prod_{r\in R} P(r;\langle A,B\rangle) \\
&= \sum_{r\in R} logP(r;\langle A,B\rangle) \\
&= \sum_{r\in R} log \sum_{\mathcal{C}_j\in\mathbb{C}} P(r,\mathcal{C}_j;\langle A,B\rangle) \qquad (2.2)
\end{aligned}$$

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:

$$\begin{aligned}
\mathcal{L}_p &= \sum_{r\in R} logP(r,C(r);\langle A,B\rangle) \\
&= \sum_{r\in R} log \underbrace{P(r|C(r_i);\langle A,B\rangle)}_{\text{Read Probability}} \underbrace{P(C(r);\langle A,B\rangle)}_{\text{Class Probability}} \qquad (2.3)
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{L}_p = \sum_{r\in R} log\frac{1}{2}\Big\{&P(r|C(r);A)P(C(r);A) \\
&+P(r|C(r);B)P(C(r);B)\Big\} \qquad (2.4)
\end{aligned}$$

2.2.2.2 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 2.2 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 (Figure). Closed form solutions to compute class probabilities are given in the Supplementary Note.

2.2.2.3 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 2.2A) 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 [WZY+17b] (Figure 2.2B).

Spanning read pairs have one mate aligned to either side of the TR. In a sample

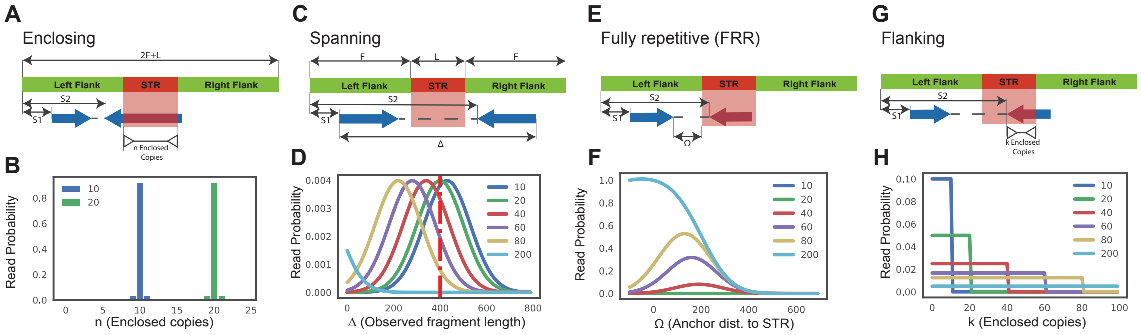

Figure 2.2: Four classes of informative read pairs. A. Enclosing class: characteristic n corresponds to the number of repeat copies enclosed in the read. B. n is modeled for different repeat length accounting for errors introduced during PCR. C. Spanning class: characteristic Δ denotes the observed fragment 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).

with a TR expansion, the spanning read pair's apparent fragment length based on mapped read positions (parameter Δ in Figure 2.2C) 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 2.2D).

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 2.2E) is modeled as a function of TR length, with smaller Ω values indicating longer TRs (Figure 2.2F).

Flanking reads partially cover the TR. The number of repeats in a flanking read (parameter k in Figure 2.2G) 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 2.2H).

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

2.2.2.4 Repetitive Read Count Term

The $\mathcal{L}_N$ term in (2.1) 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 (2.5). The unit step function $u(.)$ ensures alleles shorter than the read length have 0 expected FRR reads.

$$\lambda = u(A - \frac{r}{m}) \cdot \frac{C_v(A \cdot m - r)}{2r} + u(B - \frac{r}{m}) \cdot \frac{C_v(B \cdot m - r)}{2r} \tag{2.5}$$

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:

$$log\big(n!\big) \approx \left(n + \frac{1}{2}\right) \cdot log(n) - n + \frac{1}{2}log(2\pi) \tag{2.6}$$

2.2.3 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 [TKL$^+$17], 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 [ZLGM13].

2.2.4 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.

2.2.5 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 [Pow94] implemented in the NLopt library [Joh14].

2.2.6 Quality metrics

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

2.2.6.1 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.

2.2.6.2 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.

2.2.6.3 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:

$$\begin{aligned}
P_0 &= \Sigma_{\langle A,B\rangle \in G s.t. A<X, B<X} L(\langle A,B\rangle)/L \\
P_1 &= \Sigma_{\langle A,B\rangle \in G s.t. A<X, B\geq X} L(\langle A,B\rangle)/L \\
P_2 &= \Sigma_{\langle A,B\rangle \in G s.t. A\geq X, B\geq X} L(\langle A,B\rangle)/L
\end{aligned} \tag{2.7}$$

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.

2.2.7 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 (Table 2.2).

Reads were aligned to the hg38 reference genome using BWA-MEM [Li13] 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.

2.2.8 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:$\sum_{i=1}^{n} \sum_{j=1}^{2} \frac{(y_j^i - x_j^i)^2}{2n}$.

2.2.9 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 [Li13]. PCR duplicates were removed using the samtools [LHW$^+$09] 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.

2.2.10 Constructing a genome-wide repeat reference panel

Tandem Repeats Finder [Ben99] 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 [WGH+14], 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. [DvVS+17, TKL+17]).

2.2.11 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.

2.2.12 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_ver13.1 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. For both filter levels, we additionally filtered regions overlapping annotated segmental duplications in hg19 (UCSC Genome Browser [KSF$^+$02] track hg19.genomicSuperDups table) and regions that overlapped more than one other TR in the raw TR set from Tandem Repeats Finder [Ben99] that was used to create the reference panel.

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.

2.2.13 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 [LHW+09] 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.

Assembled paternal and maternal haplotypes were extracted from the TrioCanu [KRW+18] assembly of NA12878 (data availability:
https://obj.umiacs.umd.edu/marbl_publications/triobinning/). Contigs were aligned to the hg19 reference genome using Minimap2 [Li18] with recommended settings for full genome assembly alignment (options: -c --cs -ax asm5). The length of each TR was estimated by counting the difference in length between the reference sequence and aligned assembled haplotypes in the +/- 50bp window around each TR using Pysam
(https://github.com/pysam-developers/pysam) as described above.

2.2.14 Experimental validation of repeat lengths

Candidate TRs with long alleles identified in NA12878 were PCR amplified using GoTaq (Promega #PRM7123) with primers shown in Table 2.6. 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).

2.2.15 Code availability

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

2.3 Results

2.3.1 GangSTR outperforms existing TR expansion genotypers

We first evaluated GangSTR's performance by benchmarking against Tredparse [TKL+17] and ExpansionHunter [DvVS+17], 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 2.1). 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 (Table 2.2). 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).

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 2.3A, Figure 2.7). 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 2.3B). 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 2.3A, C). Tredparse performed well at short alleles but consistently underestimated alleles longer than the fragment length (Figure 2.3B, 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 (Figures 2.8, 2.9). Performance of all tools was mostly consistent across mean fragment lengths (Figure 2.10).

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 [DvVS+17]. All tools performed well on the HTT TR (Figure 2.3D). 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 markedly worse for all tools at FMR1 (Figure 2.3E; $\text{RMSE}_{\text{GANGSTR}}$=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, Figure 2.11). GangSTR again showed the smallest error ($\mathrm{RMSE}_{\mathrm{GANGSTR}}$=5.4; $\mathrm{RMSE}_{\mathrm{TREDPARSE}}$=96.4; $\mathrm{RMSE}_{\mathrm{EXPANSIONHUNTER}}$=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 2.12). 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 (Figure 2.13).

2.3.2 Genome-wide TR profiling

We next evaluated GangSTR's utility for genome-wide TR genotyping. To this end, we used Tandem Repeats Finder [Ben99] 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 2.4A), 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 673,252 TRs were genotyped per sample. As expected, most alleles matched the reference (Figure 2.14) 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 (Figure 2.15). 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 2.4B, Figure 2.16).

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 43,571 TRs that could not be reliably genotyped by HipSTR (Figure 2.4C). Of these, 7 are known pathogenic TRs analyzed in Figure 2.3, demonstrating the limitations of relying on enclosing reads. Notably, 1,880 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 2.4D) with strong correlation between allele lengths reported by each (Pearson r=0.99; $p < 10^{-200}$; n=542,467), demonstrating that GangSTR can robustly genotype both STRs previously analyzed using HipSTR as well as long TRs previously excluded from genome-wide analyses.

2.3.3 Genome-wide detection of novel TR expansions

We next evaluated whether GangSTR could identify novel repeat expansions in a healthy genome (NA12878). GangSTR identified 56 TRs predicted to have at least one allele longer than the read length (101bp) with greater than 80% probability (see Expansion Probabilities defined in Methods) (Table 2.3). Of these, 46 showed evidence of expansions in one or both parents. Long repeats were highly enriched for repeats with motif $AAAG_n$ (17 TRs, one-sided Fisher's exact test $p = 1.2 * 10^{-10}$) and related motifs of the form A_nG_m (Table 2.4). 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 [BLC+08, ADD+00, XCB+01].

For comparison, we applied STRetch [DLP+18], 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 (Table 2.5)). 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=0.68, $p = 1.5 * 10^{-5}$, n=33, Figure 2.5A, Table 2.5). However only 4 of the 56 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) [McC10] and

Oxford Nanopore Technologies (ONT) [JOPA16]. For each of the 56 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 53/56 cases with supporting reads from PacBio, at least one read showed evidence of an allele >101bp (46/56 for ONT) (Table 2.3). ONT showed less evidence of expansions, perhaps due to a deletion bias. Both long read technologies exhibit high error rates at homopolymer runs [WdCW+17], resulting in messy sequence within repeats themselves (Figure 2.5B).

In addition to using raw long reads for comparison, we extracted repeat regions from error-corrected phased haplotype assemblies of NA12878 generated using TrioCanu [KRW+18]. We used the phased assemblies to estimate diploid repeat lengths at each candidate expansion (Table 2.3). Overall, repeat lengths reported by GangSTR are similar to those extracted from haplotype-resolved assemblies (Pearson r=0.84, p=9.0e-13 for the smaller allele for each genotype, and Pearson r=0.76, p=1.1e-9 for the larger allele). Notably, several experimentally validated expansions reported by GangSTR (see below) are not supported by assembled haplotypes (Figure 2.5C, Table 2.3), even when they were evident in raw reads. These TRs may represent regions that could not be fully phased by assembly methods, and highlight a current limitation of long read assemblies at highly variable repeat regions.

Finally, for a subset of 11 candidate expansions, we additionally performed capillary electrophoresis to measure TR lengths (Methods, Table 2.6). Capillary results showed evidence of long alleles for the majority (9/11) of TRs (Figure 2.5C,D, Supplementary Figure 12). Notably, expanded TRs proved difficult to amplify and capillary results in some 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

(Figure 2.17). 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.

2.4 Discussion

2.4.1 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 2.3). 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 [DvVS+17] 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 [TKL+17] 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 [GWG^{+}16]. 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 [WZY^{+}17b].

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 [DLP^{+}17] 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 be used to obtain accurate diploid TR lengths. exSTRa [TBD^{+}18] 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.

2.4.2 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 [SHG+09, KPL16], 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.

2.5 Supplementary Note: GangSTR Model

2.5.1 Class probabilities

Class probability describes the probability of a read pair belonging to a specific class, considering uniform coverage. For any value of underlying allele length A, this probability can give an intuition for the relative abundance of different classes of reads (Supplementary Figure 1).

Derivation of class probabilities for each read pair class are given below. Notation corresponds to that used in the main text and depicted in Figure 2.

2.5.1.1 Class Probability of Enclosing Reads

Without loss of generality, we assume the first mate in the pair is enclosing. The calculation is similar for the other mate (Equation (2.8)). Assuming uniformity of coverage, we use a uniform distribution to find the probability of a TR region being enclosed by a read.

$$\begin{aligned} P(c_i = E; A) &= P(S_1 < F, S_1 + r > F + A \cdot m) \\ &= P(F + A \cdot m - r < S_1 < F) \\ &= \frac{(F) - (F + A \cdot m - r)}{2F + A \cdot m - 2r} \\ &= \frac{(r - A \cdot m)}{2F + A \cdot m - 2r} \end{aligned} \tag{2.8}$$

2.5.1.2 Class Probability of Spanning Reads

2.5.1.3 Fragment Length Distribution

We model the observed fragment length δ with a limited Gaussian random variable Δ with the following distribution:

$$f_{\Delta}(\delta) = \frac{1}{C\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(\delta-\mu)^2} \qquad ; r \leq \delta \leq \infty \tag{2.9}$$

In this equation μ is average fragment length, σ is the standard deviation of the fragment length distribution, C is a normalization constant to account for limited range of δ, and r is the read length.

Integration of this probability density function arises several times throughout the rest of this document. We compute these integrals using a helper Gaussian distribution X:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad ; -\infty \leq x \leq \infty \tag{2.10}$$

and it's cumulative density function (CDF):

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx \tag{2.11}$$

1) For $r \leq a, b \leq \infty$:

$$\begin{aligned}\int_a^b f_{\Delta}(\delta)d\delta &= F_{\Delta}(b) - F_{\Delta}(a) \\ &= \frac{1}{C}\{F_X(b) - F_X(a)\}\end{aligned}$$

2) For $r \leq a, b \leq \infty$:

$$\begin{aligned}\int_a^b (\delta-\mu) f_\Delta(\delta) d\delta &= \frac{1}{C\sqrt{2\pi}\sigma}\int_a^b (\delta-\mu) e^{-\frac{1}{2\sigma^2}(\delta-\mu)^2} d(\delta-\mu) \\ &= -\frac{\sigma}{C\sqrt{2\pi}}\{e^{-\frac{1}{2\sigma^2}(b-\mu)^2} - e^{-\frac{1}{2\sigma^2}(a-\mu)^2}\} \\ &= -\frac{\sigma^2}{C}\{f_X(b) - f_X(a)\}\end{aligned}$$

A read pair is classified as spanning if it's two mates are mapped in the flanking region before and after the TR locus.

$$\begin{aligned}P(c_i = S; A) &= P(S_1 < F, S_2 > F + A\cdot m - r) \\ &= \int_{2r}^{2F+A\cdot m} P(S_1 < F, S_2 > F + A\cdot m - r | \Delta = \delta) f_\Delta(\delta) d\delta \\ &= \int_{2r}^{2F+A\cdot m} P(S_1 < F, S_1 + \Delta - r > F + A\cdot m - r | \Delta = \delta) f_\Delta(\delta) d\delta \\ &= \int_{2r}^{2F+A\cdot m} P(F + A\cdot m - \Delta < S_1 < F) | \Delta = \delta) f_\Delta(\delta) d\delta \qquad (2.12) \\ &= \int_{2r}^{2F+A\cdot m} \frac{(F) - (F + A\cdot m - \delta)}{2F + A\cdot m - 2r} u(\delta - A\cdot m) f_\Delta(\delta) d\delta \qquad (2.13) \\ &= \int_{max\{2r, A\cdot m\}}^{2F+A\cdot m} \frac{\delta - A\cdot m}{2F + A\cdot m - 2r} f_\Delta(\delta) d\delta\end{aligned}$$

Step function $u(.)$ is introduced in (2.13) to satisfy the condition in (2.12), $x + A\cdot m - \Delta < x$, which simplifies to $\Delta > A\cdot m$. This condition is then imposed in the integral limit. We continue the calculation using the helper integrals from Section 2.5.1.3.

$$\begin{aligned}
P(c_i = S; A) &= \int_{max\{2r, A\cdot m\}}^{2F+A\cdot m} \frac{(\delta-\mu)+\mu-A\cdot m}{2F+A\cdot m-2r} f_\Delta(\delta) d\delta \\
&= \frac{1}{2F+A\cdot m-2r}\Big\{(\mu-A\cdot m)\int_{max\{2r,A\cdot m\}}^{2F+A\cdot m} f_\Delta(\delta)d\delta \\
&+ \int_{max\{2r,A\cdot m\}}^{2F+A\cdot m}(\delta-\mu)f_\Delta(\delta)d\delta\Big\} \\
&= \frac{\mu-A\cdot m}{C(2F+A\cdot m-2r)}\Big[F_X\big(2F+A\cdot m\big)-F_X\big(max\{2r,A\cdot m\}\big)\Big] \\
&- \frac{\sigma^2}{C(2F+A\cdot m-2r)}\Big[f_X\big(2F+A\cdot m\big)-f_X\big(max\{2r,A\cdot m\}\big)\Big] \\
&= \frac{1}{C(2F+A\cdot m-2r)}\Bigg\{(\mu-A\cdot m)\Big[F_X\big(2F+A\cdot m\big)-F_X\big(max\{2r,A\cdot m\}\big)\Big] \\
&-\sigma^2\Big[f_X\big(2F+A\cdot m\big)-f_X\big(max\{2r,A\cdot m\}\big)\Big]\Bigg\}
\end{aligned}$$

2.5.1.4 Class Probability of Flanking Reads

Without loss of generality, we assume the first mate in the pair is flanking. The calculation is similar for the other mate. Assuming uniform coverage, we use a uniform distribution to find the probability of observing a flanking read.

$$\begin{aligned}
P(c_i = F; A) &= P\big(S_1 < F, S_1 + r < F + A\cdot m, S_1 + r > F\big) \\
&= P\big(F - r < S_1 < min\{F, F + A\cdot m - r\}\big) \\
&= \frac{min\{F + A\cdot m - r, F\} - (F - r)}{2F + A\cdot m - 2r} \\
&= \frac{F + min\{A\cdot m - r, 0\} - F + r}{2F + A\cdot m - 2r} \\
&= \frac{min\{A\cdot m, r\}}{2F + A\cdot m - 2r} \qquad (2.14)
\end{aligned}$$

2.5.1.5 Class Probability of FRRs

$$\begin{aligned}
P(c_i = FRR; A) &= P(S_1 \leq F, F \leq S_2 \leq F + A \cdot m - r) \\
&= \int_{2r}^{2F+A\cdot m} P(S_1 \leq F, F \leq S_1 + \Delta - r \leq F + A \cdot m - r | \Delta = \delta) f_\Delta(\delta) d\delta \\
&= \int_{2r}^{2F+A\cdot m} P(S_1 \leq x, S_1 \leq F + A \cdot m - \delta, S_1 \geq x + r - \delta) f_\Delta(\delta) d\delta \qquad (2.15)
\end{aligned}$$

We combine the inequalities describing S_1 in (2.15) to derive conditions that need to hold for this integral to have non-zero value.

- $\left.\begin{array}{c} S_1 \geq F + r - \delta \\ S_1 \leq F + A \cdot m - \delta \end{array}\right\} \Rightarrow F + A \cdot m - \delta \geq F + r - \delta \Rightarrow A \cdot m \geq r$

  ⇒ This condition is the clear condition underlying presence of FRR reads. Smaller TR lengths have 0 probability of having an FRR read.

- $\left.\begin{array}{c} S_1 \geq F + r - \delta \\ S_1 \leq F \end{array}\right\} \Rightarrow F \geq F + r - \delta \Rightarrow \delta \geq r$

  ⇒ The lower limit of the integral is $\delta \geq 2r$, hence this condition is satisfied for the range of possible δ values.

Since there are two upper bounds for S_1 in (2.15), we need to consider two different scenarios:

- $x \leq F + A \cdot m - \delta \Rightarrow \delta \leq A \cdot m$

  Therefore, for $2r \leq \delta \leq A \cdot m$; $A \cdot m \geq 2r$, integrand is simplified to:

  $\Rightarrow P(S_1 \leq F, S_1 \leq F + A \cdot m - \delta, S_1 \geq F + r - \delta) = P(F + r - \delta \leq S_1 \leq F)$

  For $A \cdot m < 2r$, this part has no contribution.

- $F > F + A\cdot m - \delta \Rightarrow \delta > A\cdot m$

  Similarly, for $A\cdot m \leq \delta \leq 2F + A\cdot m$, integrand is simplified to:

  $\Rightarrow P(S_1 \leq F, S_1 \leq F + A\cdot m - \delta, S_1 \geq F + r - \delta) = P(F + r - \delta \leq S_1 \leq F + A\cdot m - \delta)$

Continuing integration for $A\cdot m \geq 2r$:

$$\begin{aligned}
P(c_i = FRR; A) &= \int_{2r}^{A\cdot m} P(F + r - \delta \leq S_1 \leq F) f_\Delta(\delta) d\delta \\
&+ \int_{A\cdot m}^{2F+A\cdot m} P(F + r - \delta \leq S_1 \leq F + A\cdot m - \delta) f_\Delta(\delta) d\delta \\
&= \int_{2r}^{A\cdot m} \frac{(F) - (F + r - \delta)}{2F + A\cdot m - 2r} f_\Delta(\delta) d\delta \\
&+ \int_{A\cdot m}^{2F+A\cdot m} \frac{(F + A\cdot m - \delta) - (F + r - \delta)}{2F + A\cdot m - 2r} f_\Delta(\delta) d\delta \\
&= \int_{2r}^{A\cdot m} \frac{(\delta - \mu) + (\mu - r)}{2F + A\cdot m - 2r} f_\Delta(\delta) d\delta \\
&+ \int_{A\cdot m}^{2F+A\cdot m} \frac{A\cdot m - r}{2F + A\cdot m - 2r} f_\Delta(\delta) d\delta \\
&= \frac{1}{C(2F + A\cdot m - 2r)} \Bigg\{ -\sigma^2 \big[f_X(A\cdot m) - f_X(2r)\big] \\
&+ (\mu - r)\big[F_X(A\cdot m) - F_X(2r)\big] \\
&+ (A\cdot m - r)\big[F_X(2F + A\cdot m) - F_X(A\cdot m)\big] \Bigg\} \qquad ; A\cdot m \geq 2r
\end{aligned}$$

The result is similar for $A\cdot m < 2r$, except the first two terms are zero in this case:

$$P(c_i = FRR; A) = \frac{A\cdot m - r}{C(2F + A\cdot m - 2r)} \Big\{ F_X(2F + A\cdot m) - F_X(A\cdot m) \Big\} \quad ; A\cdot m < 2r \tag{2.16}$$

2.5.2 Read probabilities

For each class of informative reads, the read probability describes the distribution of the informative characteristic of the class, given an underlying allele A (Figure 2). The details of read probability for each class of informative reads is presented in the following sections.

2.5.2.1 Enclosing Reads

Enclosing reads contain the whole repeating region, as well as flanking regions before and after. Therefore, the number of copies can be directly extracted after performing the local realignment step..

The HipSTR stutter model [WZY+17a] explains the distribution of the number of repeat copies in enclosing reads. Equation (2.17) shows the probability of a read with r_i copies having an error of length δ copies compared to the underlying true number of copies A. In this model, u and d correspond to the probability of stutter adding or removing copies of the motif, and ρ_s is the parameter of the geometric distribution that governs the number of stutter deviations from true number of copies A.

$$P(r_i - A = \delta | c_i = E; A) = \begin{cases} 1-u-d & \delta = 0 \\ u\rho_s\,(1-\rho_s)^{\delta-1} & \delta > 0 \\ d\rho_s\,(1-\rho_s)^{-\delta-1} & \delta < 0 \end{cases} \tag{2.17}$$

2.5.2.2 Flanking Reads

Flanking reads with n copies of the motif imply that one of the alleles has at least n copies of the motif. We use a uniform distribution (similar to [?]) to model the distribution of reads in the flanking class:

$$P(r_i = n | c_i = F; A) = \begin{cases} \frac{1}{A} & n \leq A \\ 0 & n > A \end{cases} \tag{2.18}$$

2.5.2.3 Spanning Reads

Fragments that completely span the TR region can create spanning read pairs. Spanning read pairs consist of two mates that are mapped to the flanking region before and after the TR. During alignment, spanning reads originating from an expanded TR allele experience a decrease in the observed fragment length (Figure 2C-D). Therefore, the distribution of fragment lengths for spanning reads is similar to the fragment length distribution in section 2.5.1.3, with a decrease in average fragment length by an amount equal to the size of expansion. If the reference has $\boldsymbol{R}$ copies of an $\boldsymbol{m}$ base pair motif, we can describe the class probability of spanning reads with the following Gaussian distribution:

$$P(r_i|c_i = S) \sim N(\mu - (A - R) \cdot m, \sigma) \tag{2.19}$$

2.5.2.4 Fully Repetitive Reads (FRRs)

FRR reads are extracted from both on and off target regions to create the repetitive read count term in the likelihood model. Here we discuss another informative aspect of FRR reads, the distance of an anchored mate to the repeat region (Figure 2).

Anchored FRRs are read pairs that contain one read completely consisting of repeats, while the other mate pair is mapped to the flanking region before or after the TR. Using the fragment length distribution (see 2.5.1.3), we model the distance of the anchor read from the repeat region (shown by $\boldsymbol{\Omega}$) to obtain read probability of this class of reads. We use the notation from section 2.5.1.3 to derive the read probability of anchored FRR reads.

$$
\begin{aligned}
P(r_i|c_i = FRR; A) &= P(\Omega + 2r < \Delta < \Omega + r + L) && (2.20)\\
&= F_\Delta(\Omega + r + L) - F_\Delta(\Omega + 2r) && (2.21)\\
&= \frac{1}{C}[F_X(\Omega + r + L) - F_X(\Omega + 2r)] && (2.22)
\end{aligned}
$$

On the other hand, fragments that originate from within the repeating region generate FRR read pairs (both mates repetitive). These read pairs do not have an anchor, and are most likely aligned to one of the off-target regions associated with the TR. These read pairs contribute to both FRR count term (adding two FRR reads) and read pair term (FRR class probability computed for $\Omega = -r$).

Chapter 2, in full, is a reprint of the material as it appears in Nucleic Acids Research 2019. Mousavi, Nima, Sharona Shleizer-Burko, Richard Yanicky, and Melissa Gymrek. The dissertation author was the primary investigator and author of this paper.

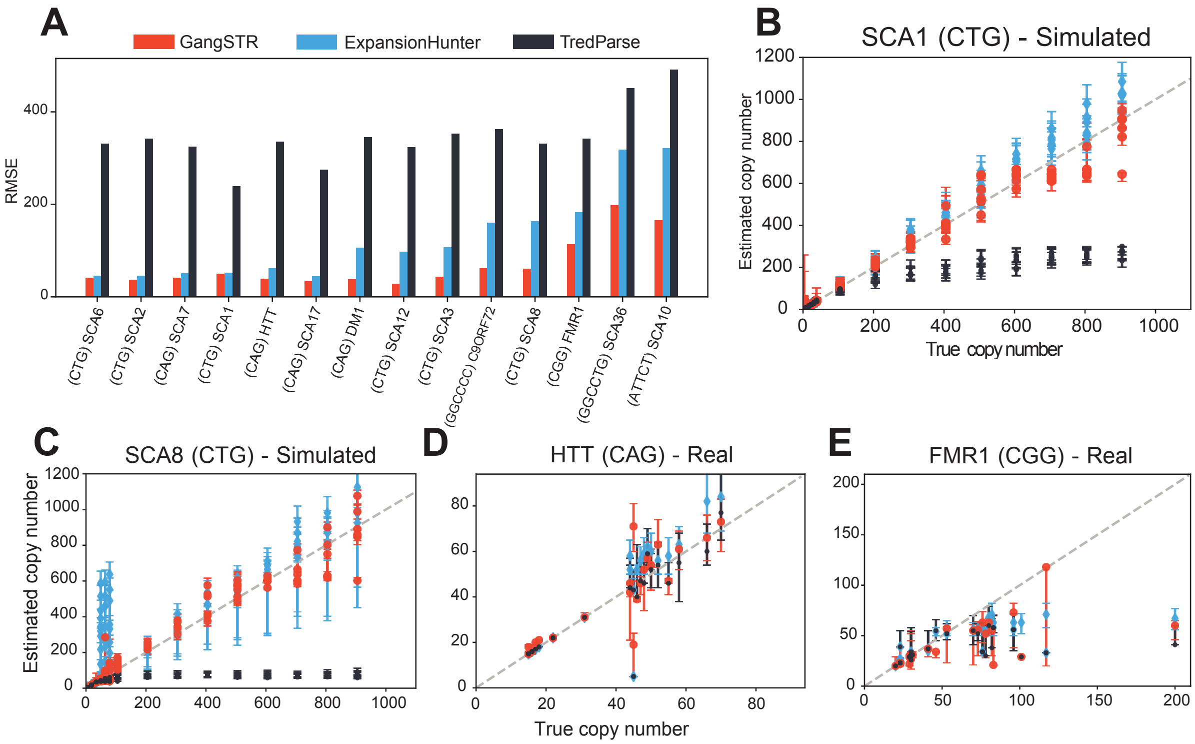

Figure 2.3: Evaluation of TR genotypers on real and simulated data at pathogenic repeat expansions. A. RMSE for each simulated locus. HTT=Huntington's Disease; SCA=spinocerebellar ataxia; DM=Myotonic Dystrophy; C9ORF72=amyotrophic lateral sclerosis/frontotemporal dementia; FMR1=Fragile X Syndrome. TRs are sorted from left to right by ascending length of the pathogenic allele. The motif for each locus is specified in parentheses. B. Comparison of true vs. estimated repeat number for each simulated genotype for SCA1. Gray dashed 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. Comparison of true vs. estimated repeat number for FMR1 using real WGS data. In all panels, red=GangSTR; blue=ExpansionHunter; black=Tredparse.

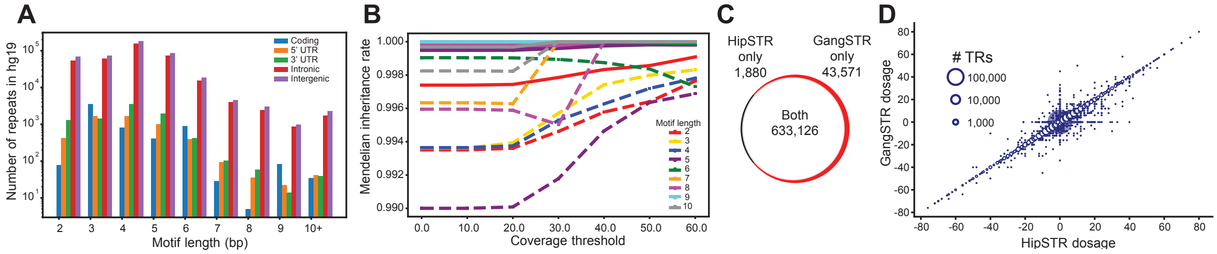

Figure 2.4: Genome-wide TR genotyping. A. Composition of TRs in the hg19 reference genome. The x-axis gives the motif length and the y-axis (log10 scale) gives the number of TRs in the genome. Colored bars represent TRs overlapping various genomic annotations (blue=coding, orange=5' UTR, green=3' UTR, 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.

Table 2.2: Target pathogenic repeats used in benchmarking experiments. Simulated samples have one allele from "Simulation repeat range" column and the other allele covers the range (5, 1005) with step size 100. Simulation repeat range is given in terms of repeat copy number.

Abbreviation	Disease	Gene	Motif	Repeat location	Pathogenic cutoff	Simulation repeat range
SCA6	Spinocerebellar ataxia 6	CACNA1A	CAG	chr19:13207859-13207897 (hg38) chr19:13318673-13318711 (hg19)	20 (60 bps)	[4, 7, 10, 13, 16, 19]
SCA2	Spinocerebellar ataxia 2	ATXN2	CAG	chr12:111598951-111599019 (hg38) chr12:112036755-112036823 (hg19)	33 (99 bps)	[2, 8, 14, 20, 26, 32]
SCA7	Spinocerebellar ataxia 7	ATXN7	CAG	chr3:63912686-63912715 (hg38) chr3:63898362-63898391 (hg19)	34 (102 bps)	[3, 9, 15, 21, 27, 33]
SCA1	Spinocerebellar ataxia 1	ATXN1	CAG	chr6:16327636-16327722 (hg38) chr6:16327867-16327953 (hg19)	39 (117 bps)	[3, 10, 17, 24, 31, 38]
HTT	Huntington's Disease	HTT	CAG	chr4:3074877-3074933 (hg38) chr4:3076604-3076660 (hg19)	40 (120bps)	[4, 11, 18, 25, 32, 39]
SCA17	Spinocerebellar ataxia 17	TBP	CAG	chr6:170561908-170562021 (hg38) chr6:170870996-170871109 (hg19)	43 (123 bps)	[2, 10, 18, 26, 34, 42]
DM1	Myotonic Dystrophy 1	DMPK	CTG	chr19:45770205-45770264 (hg38) chr19:46273463-46273522 (hg19)	50 (150 bps)	[4, 13, 22, 31, 40, 49]
SCA12	Spinocerebellar ataxia 12	PPP2R2B	CAG	chr5:146878729-146878758 (hg38) chr5:146258292-146258321 (hg19)	51 (153 bps)	[5, 14, 23, 32, 41, 50]
SCA3	Spinocerebellar ataxia 3	ATXN3	CAG	chr14:92071011-92071034 (hg38) chr14:92537355-92537378 (hg19)	60 (120bps)	[4, 15, 26, 37, 48, 59]
C9ORF72	Amyotrophic Lateral Sclerosis (ALS)	C9ORF72	GGCCCC	chr9:27573529-27573546 (hg38) chr9:27573527-27573544 (hg19)	31 (186 bps)	[5, 10, 15, 20, 25, 30]
SCA8	Spinocerebellar ataxia 8	ATXN8OS	CTG	chr13:70139384-70139428 (hg38) chr13:70713516-70713560 (hg19)	80 (240 bps)	[4, 19, 34, 49, 64, 79]
FMR1	Fragile X syndrome	FMR1	CGG	chrX:147912051-147912110 (hg38) chrX:146993569-146993628 (hg19)	200 (600 bps)	[4, 43, 82, 121, 160, 199]
SCA36	Spinocerebellar ataxia 36	NOP56	GGCCTG	chr20:2652734-2652757 (hg38) chr20:2633380-2633403 (hg19)	650 (3900 bps)	[4, 133, 262, 391, 520, 649]
SCA10	Spinocerebellar ataxia 10	ATXN10	ATTCT	chr22:45795355-45795424 (hg38) chr22:46191235-46191304 (hg19)	800 (4000 bps)	[4, 163, 322, 481, 640, 799]

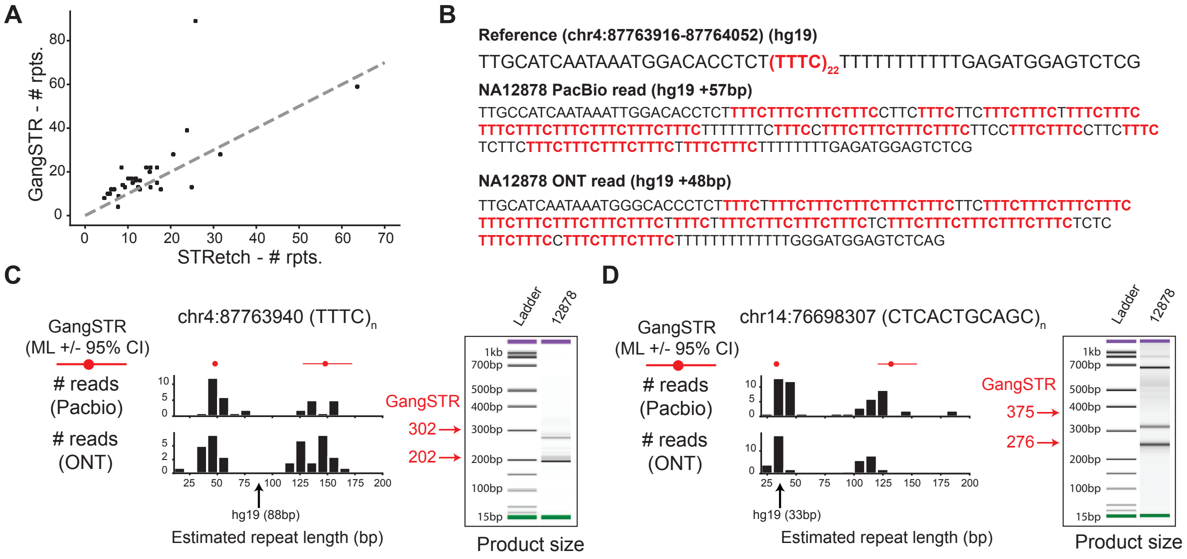

Figure 2.5: Discovery and validation of genome-wide TR expansions. A. Comparison of STRetch and GangSTR estimated repeat lengths. The x-axis shows the estimated repeat number returned by STRetch. The y-axis shows the estimated repeat number of the longest of two alleles reported as the maximum likelihood 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.

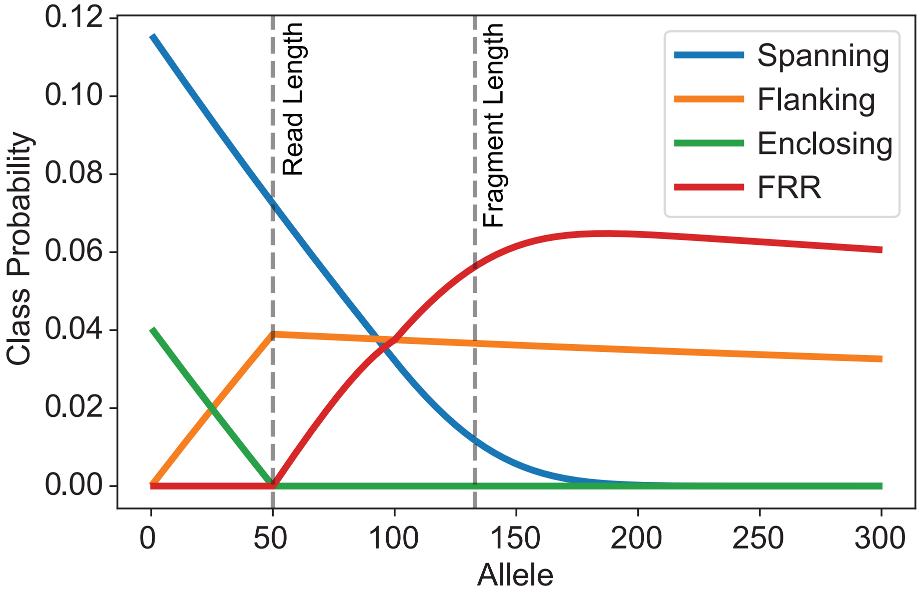

Figure 2.6: Class probabilities as a function of TR length. The x-axis shows the allele length in number of repeats. The y-axis shows the probability that a read mapped to the TR region would be from each class. Results were calculated using a 3bp long repeat unit, read length=100bp, and fragment length=400bp. Blue=spanning reads, green=flanking reads, red=enclosing reads, and purple=FRR reads.

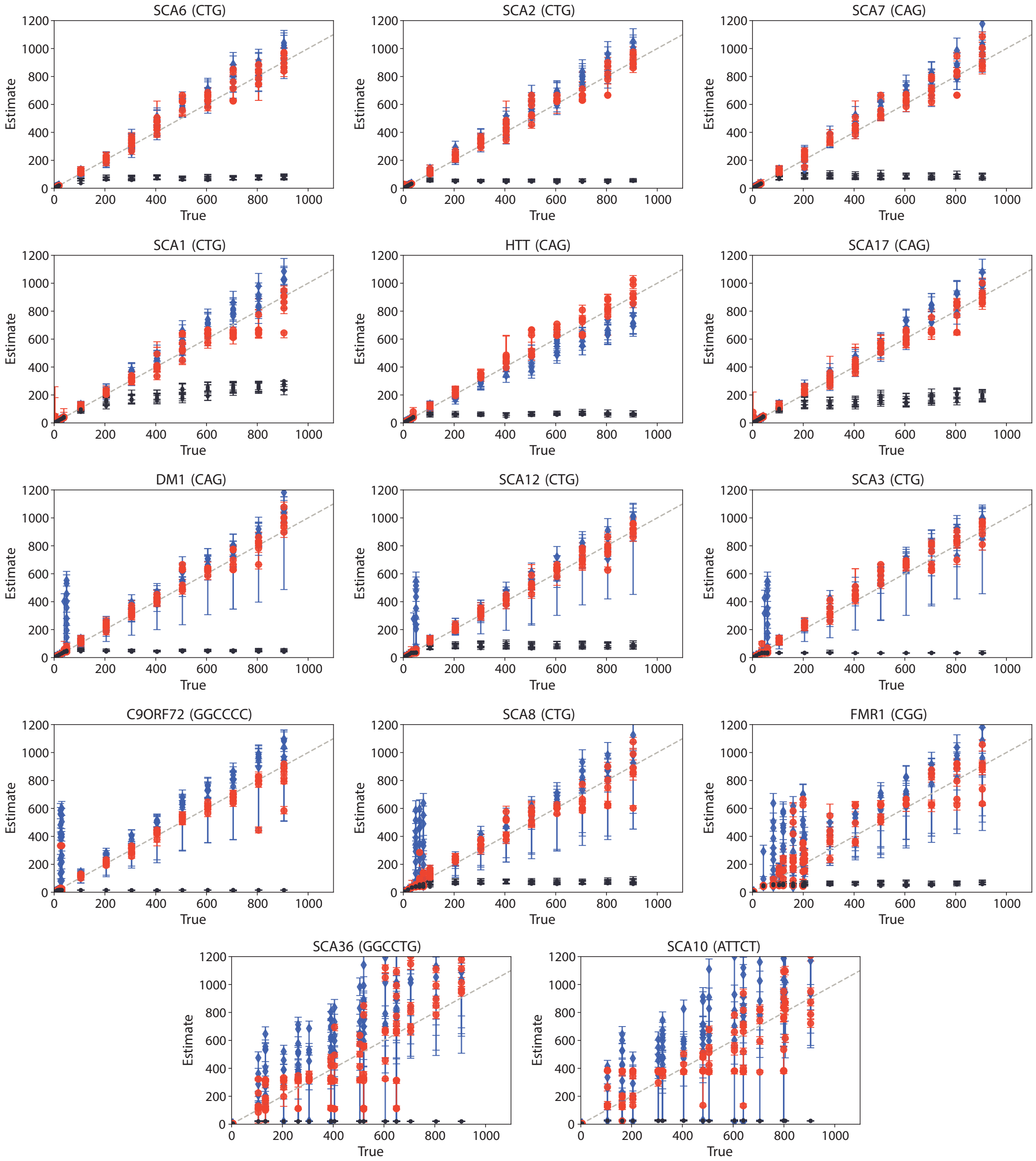

Figure 2.7: Comparison of true vs. estimated repeat number on simulated data for different loci The x-axis shows the simulated allele length in number of repeats. The y-axis shows the estimated allele length in number of repeats. The accuracy (root mean square error) for each panel is plotted in Figure 3A. The motif for each locus is specified in parentheses in plot title. In all panels, red=GangSTR; blue=ExpansionHunter; black=Tredparse.

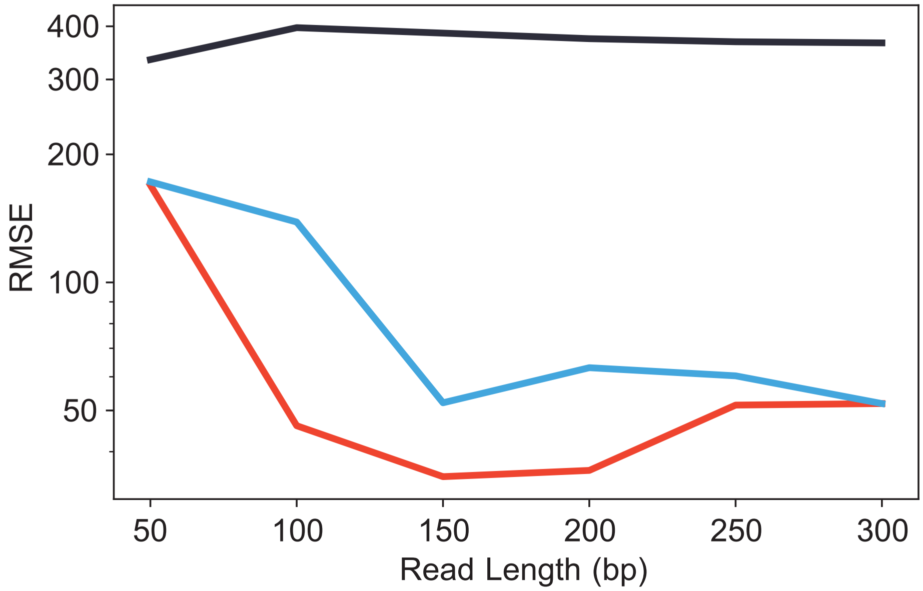

Figure 2.8: Estimation accuracy for simulated samples of HTT vs. read length Root Mean Square Error (RMSE) for estimation of simulated samples of the HTT locus with different read lengths. Red=GangSTR; blue=ExpansionHunter; black=Tredparse.

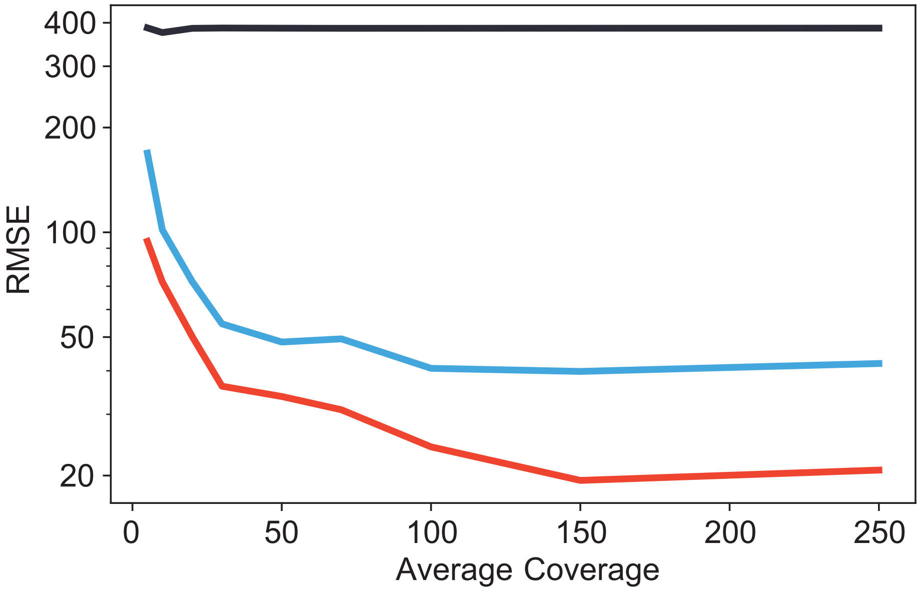

Figure 2.9: Estimation accuracy for simulated samples of HTT vs. average coverage Root Mean Square (RMSE) for estimation of simulated samples of the HTT locus with different coverages. Red=GangSTR; blue=ExpansionHunter; black=Tredparse.

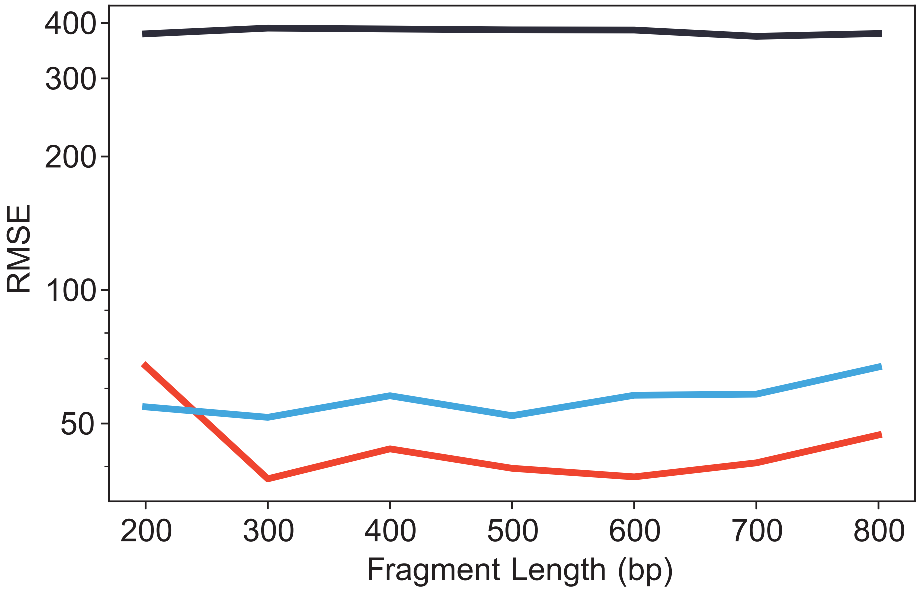

Figure 2.10: Estimation accuracy for simulated samples of HTT vs. fragment length Root Mean Square (RMSE) for estimation of simulated samples of HTT locus with different fragment lengths. Red=GangSTR; blue=ExpansionHunter; black=Tredparse.

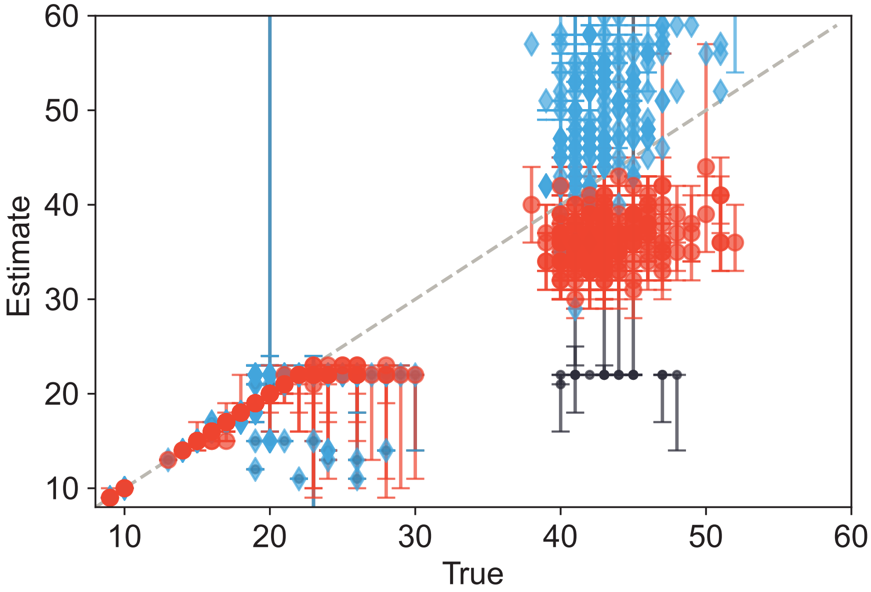

Figure 2.11: Comparison of true vs. estimated repeat number using real HTT exome data The x-axis shows the experimentally validated allele length in number of repeats. The y-axis shows the estimated allele length in number of repeats. Gray dashed line gives the diagonal. red=GangSTR; blue=ExpansionHunter; black=Tredparse.

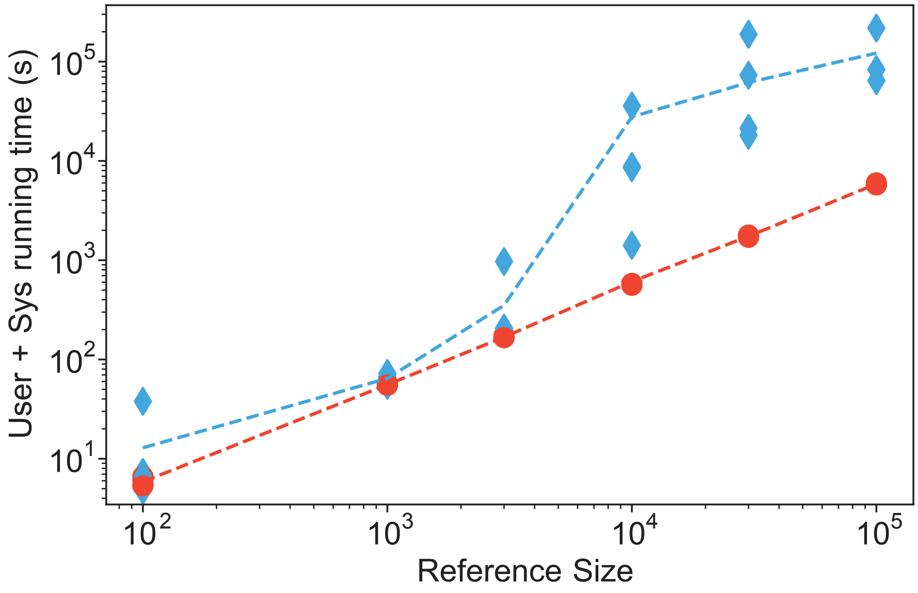

Figure 2.12: Running time of GangSTR and ExpansionHunter vs. reference size The x-axis shows the number of TRs in the reference set used. The y-axis shows running time (User + Sys) in seconds. Lines give mean value across 5 runs. Points (“x”) give raw data values for each of 5 runs. For two runs with 10^5 TRs ExpansionHunter did not run to completion. Red=GangSTR, blue=ExpansionHunter.

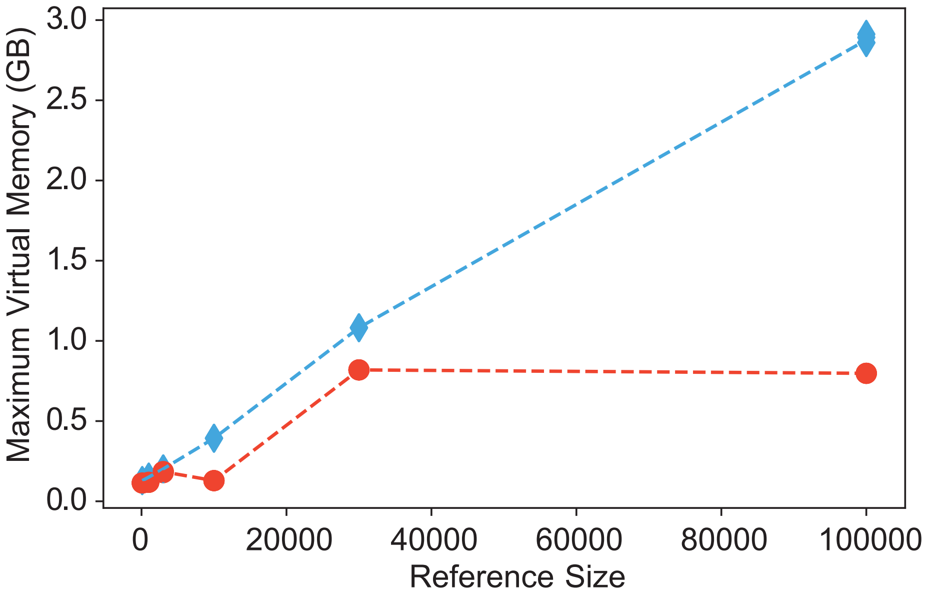

Figure 2.13: Peak memory usage by GangSTR and ExpansionHunter vs. reference size The x-axis shows the number of TRs in the reference set used. The y-axis shows maximum virtual memory usage in gigabytes. Lines give mean value across 5 runs. Points ("x") give raw data values for each of 5 runs. Red=GangSTR, blue=ExpansionHunter.

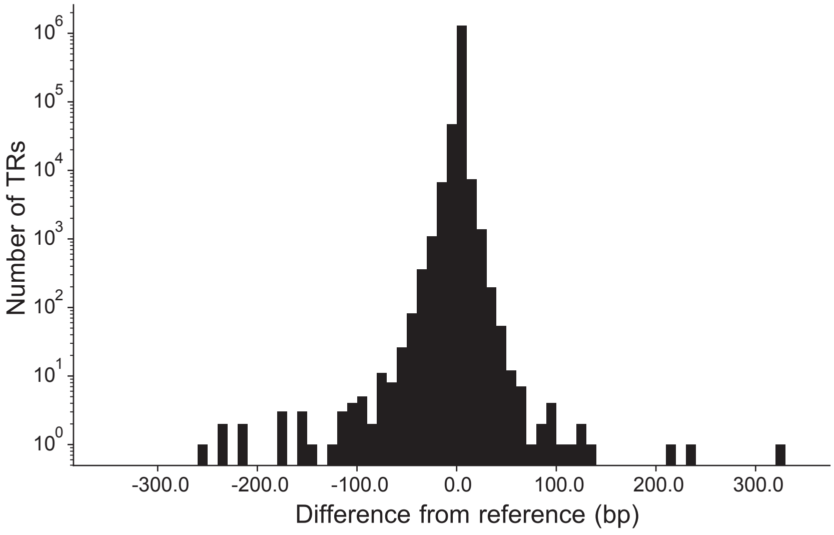

Figure 2.14: Distribution of repeat lengths in NA12878 compared to the hg19 reference. Y-axis is on a log10 scale.

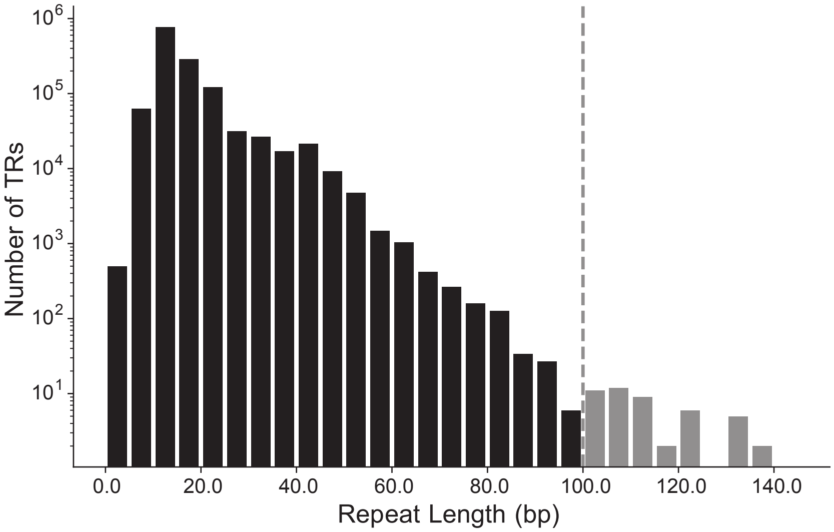

Figure 2.15: Distribution of total repeat lengths in NA12878. Y-axis is on a log10 scale. Gray bars to the right of the dashed line indicate alleles longer than the read length of 101bp.

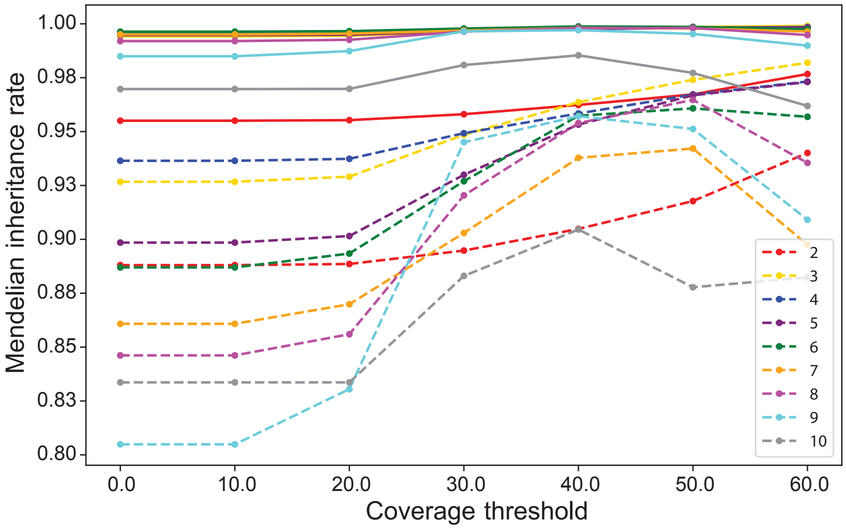

Figure 2.16: Mendelian inheritance of GangSTR genotypes in a CEU trio as a function of the number of informative read pairs Mendelian inheritance of GangSTR genotypes in a CEU trio as a function of the number of informative read pairs. Colors denote repeat lengths. Solid lines give mean Mendelian inheritance rate across all TRs, computed using maximum likelihood GangSTR genotypes as described in Methods. Dashed lines are computed after excluding loci where all three samples were homozygous for the reference allele.

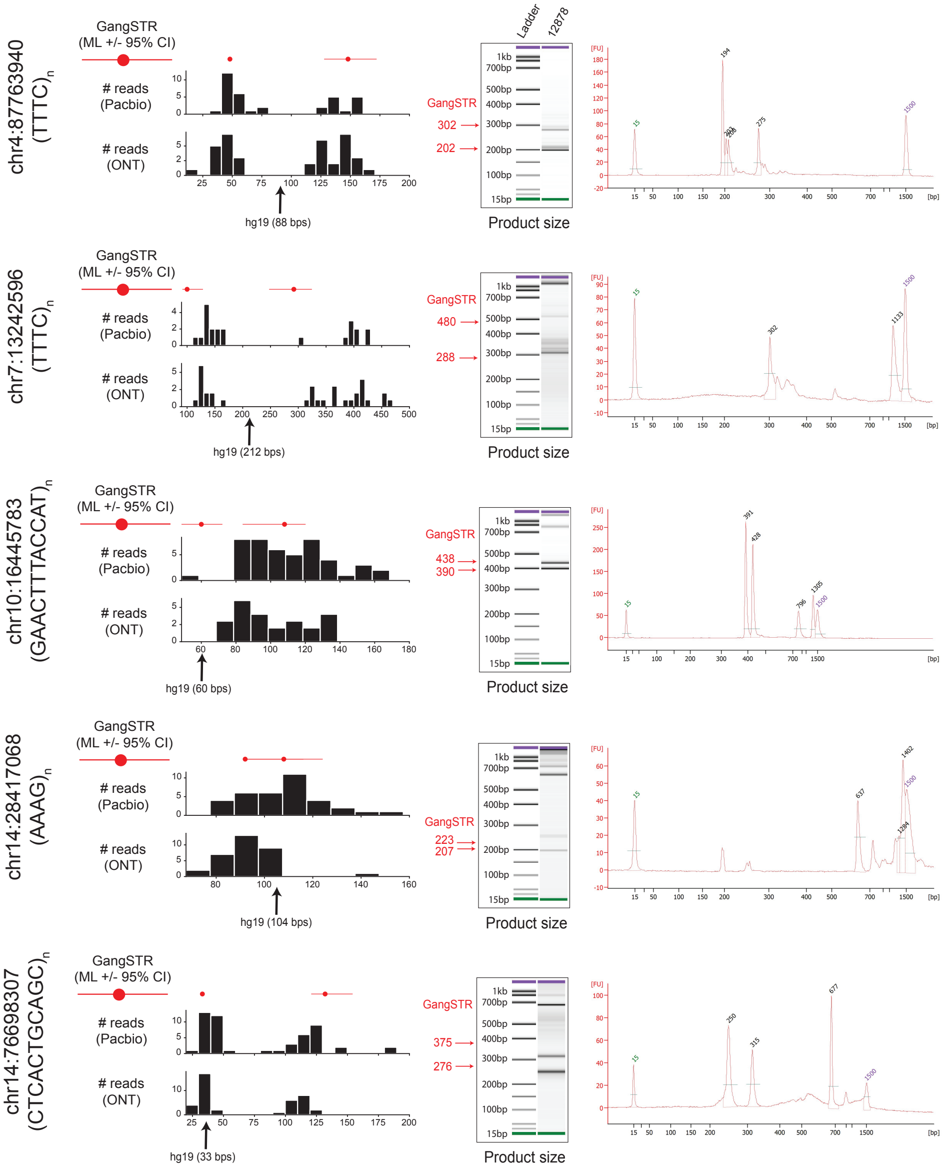

Figure 2.17: Discovery and validation of genome-wide TR expansions. For each of the 9 TRs shown, left plots compare GangSTR genotypes to those predicted by long reads. Red dots give the maximum likelihood repeat lengths predicted by GangSTR and red lines give the 95% confidence intervals for each allele. 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 middle 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. Right plots give the capillary electrophoresis traces produced by the Agilent Bioanalyzer.

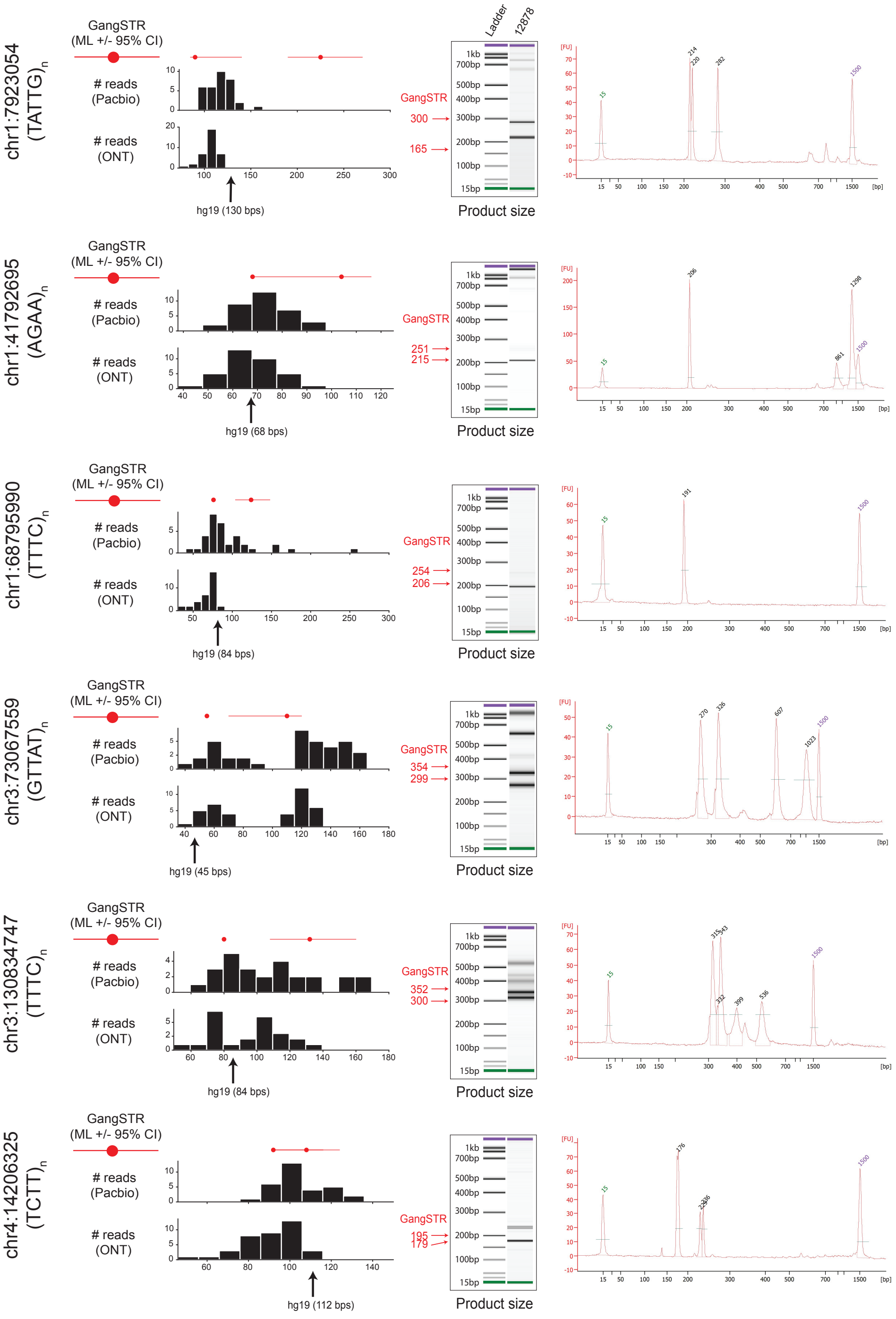

Figure 2.17: Discovery and validation of genome-wide TR expansions (continued).

Table 2.3: Candidate TRs long alleles (>101bp) in NA12878. Description of columns: Refcopy: Number of copies of the motif in hg19. GangSTR: Maximum likelihood diploid repeat copy number returned by GangSTR. P(het): Posterior probability that the genotype is heterozygous for one allele greater than 101bp. P(hom): Posterior probability that the genotype is homozygous for both alleles greater than 101bp. MI: Indicates whether confidence intervals in the trio are consistent with Mendelian inheritance. NA indicates one or more parents failed filtering steps so Mendelian inheritance couldn't be determined. Parent: Lists which parents show evidence (>80% posterior probability) of an expansion. PacBio: Maximum allele length supported by PacBio reads for NA12878. "-" indicates no PacBio reads were found in the region. ONT: Maximum allele length supported by ONT reads for NA12878. "-" indicates no ONT reads were found in the region. Assembly: Diploid repeat copy number from maternal and paternal TrioCanu assembly (in that order).

Coord (hg19)	Refcopy	Motif	GangSTR	P(het)	P(hom)	MI	Parent	PacBio	ONT	Assembly
chr1:2897558	4	AACAGGAGGTCTGGT	1,7	1.00	0.00	True	NA	147	108	2.9,6.0
chr1:7923054	26	AATAC	18,45	0.93	0.07	True	NA12891, NA12892	154	123	22.4,22.4
chr1:22720748	13	AAAG	16,30	1.00	0.00	True	NA12891, NA12892	163	87	21.5,15.8
chr1:35267829	44	AAACC	20,90	0.28	0.72	True	NA12891, NA12892	377	275	44.8,46.0
chr1:41792695	17	AAAG	17,26	1.00	0.00	True	NA12891	225	93	16.5,18.8
chr1:61347419	3	AAAG	3,33	1.00	0.00	True	NA12891, NA12892	64	23	2.8,2.8
chr1:64329379	19	AATAC	17,22	1.00	0.00	True	NA12892	231	109	19.8,17.8
chr1:68795990	21	AAAG	19,31	1.00	0.00	True	NA12892	363	85	19.0,18.8
chr1:154098099	29	AAGG	26,26	0.74	0.26	NA	NA12891	241	217	NA,NA
chr1:208680308	6	AATGTGGTATATATACAT	4,5	0.67	0.29	NA	NA	168	141	4.9,5.9
chr10:16445783	5	AAAGTTCATGGT	5,9	1.00	0.00	True	NA12891, NA12892	169	135	6.9,9.8
chr10:99438638	60	AC	46,59	0.73	0.27	True	NA12891, NA12892	147	168	51.5,NA
chr10:125413213	15	AAAGG	14,21	1.00	0.00	True	NA	192	134	13.8,21.4
chr11:17574076	3	ACACAGGACAGGTGGGGG	6,6	0.13	0.87	NA	NA	557	495	23.9,24.0
chr11:31932832	26	AAAGG	20,49	0.58	0.42	True	NA12891, NA12892	280	199	35.2,26.8
chr11:107461059	53	AG	47,62	0.77	0.23	True	NA12891, NA12892	223	89	NA,46.5
chr12:15314073	22	AAAG	22,28	0.98	0.02	True	NA12891, NA12892	211	127	27.8,22.8
chr12:117836405	20	AAAAT	9,21	1.00	0.00	True	NA12891	138	111	9.0,19.6
chr13:29027163	23	AAAGG	16,23	0.88	0.12	True	NA12891, NA12892	205	105	19.6,20.8
chr13:44716269	35	AGCCG	14,22	1.00	0.00	True	NA12891	182	133	16.2,21.8
chr13:87882390	3	AAAG	3,33	1.00	0.00	True	NA12891, NA12892	92	60	5.0,2.8
chr13:96047512	16	AAAGG	17,24	1.00	0.00	True	NA12891	155	113	NA,21.6
chr14:28417068	26	AAAG	23,27	0.80	0.20	True	NA12891, NA12892	216	143	24.8,24.8
chr14:76698307	3	ACTGCAGCCTC	3,12	1.00	0.00	True	NA12891, NA12892	535	125	9.9,2.9
chr15:54367612	5	AAGCTCCGGCTCACTGC	4,5	0.88	0.11	True	NA	189	94	5.1,5.0
chr15:61429219	20	AAAG	21,26	1.00	0.00	True	NA12892	143	112	20.8,NA
chr15:90651456	13	AAAAT	8,21	1.00	0.00	NA	NA	114	119	7.8,NA
chr16:3899380	55	AC	36,55	0.97	0.03	True	NA12891, NA12892	584	124	46.0,46.0
chr16:50509578	53	AAAG	17,39	1.00	0.00	True	NA12891, NA12892	169	90	20.8,18.2
chr16:58865692	10	AAGGAGGG	7,14	1.00	0.00	True	NA12891, NA12892	140	110	13.8,12.9
chr17:32835083	21	AAAGG	18,27	0.92	0.08	True	NA12891, NA12892	146	146	22.8,16.0
chr19:39720793	15	AAAGG	20,26	0.55	0.45	True	NA12891, NA12892	218	275	NA,NA
chr2:54425083	18	AATAC	17,24	1.00	0.00	True	NA12892	156	140	16.8,16.8
chr2:163609414	25	AAAAG	67,89	0.00	1.00	True	NA12891, NA12892	426	287	51.0,51.6
chr21:36720944	18	AATAG	19,37	0.78	0.22	True	NA12891, NA12892	182	156	28.8,23.8
chr22:47769363	3	AAGGGAGGCCAGGAGGAG	3,6	1.00	0.00	True	NA12891	117	113	2.9,5.9
chr3:5830605	11	AAATGCACAGGAAT	6,16	0.61	0.39	True	NA12891, NA12892	230	180	11.9,10.9
chr3:73067559	9	AACAT	11,22	1.00	0.00	True	NA12892	165	135	23.8,NA
chr3:86384908	19	AAAAT	18,22	0.70	0.30	NA	NA	162	120	19.0,19.0
chr3:130834747	21	AAAG	20,33	1.00	0.00	True	NA12892	167	135	20.8,20.2
chr4:14206325	28	AAAG	23,27	0.90	0.10	True	NA12891, NA12892	203	115	24.8,24.0
chr4:21716410	61	AAAAT	18,48	0.71	0.29	True	NA12891, NA12892	293	287	42.8,52.8
chr4:87763940	22	AAAG	12,37	1.00	0.00	True	NA12892	303	168	12.8,12.2
chr4:90302001	3	AAAG	3,26	1.00	0.00	True	NA	129	48	6.5,5.5
chr5:75792512	3	AAAG	3,28	1.00	0.00	True	NA12891, NA12892	67	18	4.8,2.8
chr5:157994659	20	AAAG	17,26	1.00	0.00	True	NA12891	143	107	20.8,19.8
chr6:128925487	18	AGAGCGGG	5,17	1.00	0.00	True	NA12891, NA12892	515	161	15.9,17.9
chr7:2852271	12	ACATC	18,39	0.78	0.22	True	NA12891, NA12892	188	181	27.8,25.8
chr7:6460939	18	AGCGCGGGAGGCGCAGGC	4,6	0.99	0.00	True	NA12892	652	431	24.7,NA
chr7:13242596	53	AAAG	25,73	0.60	0.40	True	NA12891, NA12892	428	469	NA,35.8
chr7:105084942	6	AACACCTATAGC	3,8	0.88	0.00	True	NA	544	318	NA,NA
chr7:127898719	17	AAAG	22,26	0.89	0.11	NA	NA12892	184	126	29.0,29.8
chr7:134201476	15	AAAAG	15,22	1.00	0.00	True	NA12891	156	115	15.4,15.0
chr8:119927182	13	AGAGAGCG	11,20	0.90	0.10	True	NA12891, NA12892	263	136	13.9,13.9
chr8:130361920	15	AAAAT	16,21	0.88	0.12	NA	NA12892	141	116	20.8,19.8
chr8:140126207	5	AAGACGACTCCACCCCACAG	3,6	1.00	0.00	NA	NA	133	121	3.0,5.0

Table 2.4: Enrichment of motifs with long alleles in NA12878. All motifs found at least twice in long alleles (>101bp) in NA12878 are shown. P-values were computed using a one-sided Fisher's exact test.

Motif	Num. TRs	P-val
AAAG	17	1.23e-10
AAAGG	6	6.15e-08
AATAC	3	1.51e-05
AAAAT	5	9.15e-02
AAAAG	2	4.10e-01
AC	2	9.41e-01

Table 2.5: Comparison of STRetch and GangSTR output. Description of columns: STRetch copy num: Estimated repeat copy number returned by STRetch. GangSTR gt: Maximum likelihood genotype (in terms of repeat copy number) returned by GangSTR. GangSTR filter: Locus-level GangSTR filters. QEXP: Expansion probability returned by GangSTR, which gives the probability of no expansion, a heterozygous expansion, or a homozygous expansion based on comparison to a predefined threshold. In this case the threshold was set to the read length of 101bp.

Chrom	STR Pos (hg19)	Motif	hg19 copy num	STRetch copy num	GangSTR gt	GangSTR filter	QEXP
chr20	49282720	AAAAG	7	23.90	.	NOCALL	.
chr3	121505017	AAAAG	8	24.90	13,13	PASS	1.00,0.00,0.00
chr4	36460354	ACACAT	7	17.70	12,12	PASS	0.98,0.02,0.00
chr7	2852270	ACATC	11	23.80	18,39	PASS	0.00,0.78,0.22
chr6	72287642	AGAGAT	3	12.80	.	SpanBoundOnly	.
chr3	47788930	AAATAT	6	12.40	10,13	PASS	1.00,0.00,0.00
chr3	123773204	AAAAAT	3	7.80	8,9	PASS	1.00,0.00,0.00
chr4	30718515	AGC	11	20.60	12,28	PASS	1.00,0.00,0.00
chr1	208440810	AGC	8	14.30	9,22	PASS	1.00,0.00,0.00
chr16	57893926	AAAAT	13	16.80	15,15	PASS	1.00,0.00,0.00
chr16	65292358	ACATAT	3	6.10	4,12	PASS	1.00,0.00,0.00
chr2	112925446	AACAT	9	12.80	.	SpanBoundOnly	.
chr4	10768264	AAAAG	14	17.80	.	SpanBoundOnly	.
chr10	49500011	AAAG	5	8.50	7,22	PASS	0.87,0.13,0.00
chr16	17564764	CCG	4	8.70	.	LowCallDepth, SpanBoundOnly	.
chr17	15378610	AAAAG	9	11.80	13,16	PASS	1.00,0.00,0.00
chr2	17512498	ACACAT	3	5.30	5,10	PASS	1.00,0.00,0.00
chr3	188449305	AAAAC	3	5.80	4,12	PASS	1.00,0.00,0.00
chr5	11846	ACCCCG	3	5.30	.	LowCallDepth	.
chr5	176523778	AGG	3	7.70	4,4	PASS	1.00,0.00,0.00
chr6	65241442	AAAAT	9	11.80	9,15	PASS	1.00,0.00,0.00
chr7	134201475	AAAAG	14	16.80	15,22	PASS	0.00,1.00,0.00
chr8	87100322	AAAAT	6	8.80	9,14	PASS	1.00,0.00,0.00
chr1	234275697	AAAAT	9	10.80	14,17	PASS	0.89,0.11,0.00
chr1	41410562	AAAGC	10	11.80	.	SpanBoundOnly	.
chr1	59446780	AAAG	41	43.30	15,17	PASS	1.00,0.00,0.00
chr1	85576133	AAAT	7	9.30	11,13	PASS	1.00,0.00,0.00
chr10	106199462	AAC	8	11.10	14,15	PASS	1.00,0.00,0.00
chr10	99438637	AC	59	63.60	46,59	PASS	0.00,0.73,0.27
chr11	62855843	AACATC	3	4.50	8,8	PASS	1.00,0.00,0.00
chr12	79152282	AATGT	11	12.80	11,12	PASS	1.00,0.00,0.00
chr13	113588869	AGC	7	10.10	9,17	PASS	1.00,0.00,0.00
chr15	75186380	AC	33	37.60	.	LowCallDepth, SpanBoundOnly	.
chr18	4057004	AC	27	31.60	24,28	PASS	1.00,0.00,0.00
chr19	53608032	AACAT	4	5.80	5,10	PASS	1.00,0.00,0.00
chr2	163609413	AAAAG	24	25.80	67,89	PASS	0.00,0.00,1.00
chr2	206247807	AGAT	13	15.30	14,22	PASS	0.89,0.11,0.00
chr21	34315567	AAAAG	7	8.80	9,14	PASS	1.00,0.00,0.00
chr22	22928364	AAAG	27	29.30	23,31	PASS	0.00,0.75,0.25
chr3	139970426	AAAGG	10	11.80	13,17	PASS	0.92,0.08,0.00
chr5	168182394	AAAG	13	15.30	13,13	PASS	1.00,0.00,0.00
chr6	43120472	AACAT	11	12.80	11,16	PASS	1.00,0.00,0.00
chr7	105372194	AAAAG	10	11.80	11,17	PASS	0.94,0.06,0.00
chr7	55955293	CCG	12	15.10	9,20	PASS	1.00,0.00,0.00
chr8	136094609	ATCCC	5	6.80	6,12	PASS	1.00,0.00,0.00

Table 2.6: Primers for capillary electrophoresis validation.

Chrom	STR Pos (hg19)	Forward Primer	Reverse Primer
chr1	7923054	CAATAAGGCCTACCCTGACG	GGGCAACAAGAGCAAAACTT
chr1	41792695	AGCTGCTTGAGAAGCTGAGG	CCCCATGGCTTTAACTCACT
chr1	68795990	TTCCTCTCCCCAACACTTTTT	TGAGCCTCAGGAGATTGTTG
chr3	73067559	GGTTGACAGCGGGATTTAAG	GAGCCATGGACACATCACTG
chr3	130834747	TGGCGAGGTATTGTGGTAGA	TGACGAGTTAATGGGTGCAG
chr4	14206325	ACAAACTTCTATGGGCTCGAT	CCTGGGCAAAGAGAGTGAAA
chr4	87763940	AGCTGTCCTGAGTTGCATCA	GACTGAGGCAGGAGAAATGC
chr7	13242596	GCATTTTCCTGATGGCTAAA	TTAGCCGGGTGTGGTAGC
chr10	16445783	TGCCCAATAAGTATGAGAAGAACA	AAGTTCAAAAGGCCAGACCA
chr14	28417068	CTGGGCGATAGAGCAAGACT	CCCTCATACCAAAGTGAACAAA
chr14	76698307	ATAGAGTGCAGTGGGGCAAA	GAGCCCAAGAGTTCAACACC

Chapter 3

TRTools: a toolkit for genome-wide analysis of tandem repeats

Most of this chapter was first published as:

Mousavi, N., Margoliash, J., Pusarla, N., Saini, S., Yanicky, R., & Gymrek, M. TRTools: a toolkit for genome-wide analysis of tandem repeats. Bioinformatics. (2021).

Abstract: A rich set of tools have recently been developed for performing genome-wide genotyping of tandem repeats (TRs). However, standardized tools for downstream analysis of these results are lacking. To facilitate TR analysis applications, we present TRTools, a Python library and suite of command line tools for filtering, merging, and quality control of TR genotype files. TRTools utilizes an internal harmonization module making it compatible with outputs from a wide range of TR genotypers.
Availability: TRTools is freely available at https://github.com/gymreklab/TRTools.
Documentation: Detailed documentation is available at https://trtools.readthedocs.io.

3.1 Introduction

Tandem repeats (TRs) represent one of the largest sources of human genetic variation and are well-known to affect many human phenotypes [Han18a]. Improvements in sequencing technology and bioinformatics algorithms have led to the recent development of a rich set of tools for performing genome-wide analysis of TR variation [WZY^{+}17a, MSBYG19, KEAH19, ea17, BSBG^{+}18]. These tools take aligned sequencing reads as input and output Variant Call Format (VCF) files containing estimates of TR copy number at one or more genomic TRs. The resulting VCF files may be used for a wide variety of downstream applications. However, before doing so it is usually necessary to perform filtering, quality control (QC), and merging of files across samples. While utilities exist for performing such manipulations on VCF files containing SNP variants, these tools often do not handle multi-allelic TRs and are not designed to compute TR-specific statistics. Further, different TR genotypers use different allele annotations, complicating the use of downstream tools.

Here, we present TRTools, an open-source toolkit for performing analyses on TR genotypes. TRTools provides utilities for filtering, merging, comparing, and performing QC on TR VCF files. It may be used to analyze either short tandem repeats (STRs; repeat units 1-6bp) or variable number tandem repeats (VNTRs; repeat units >6bp) collectively referred to here as TRs. It is currently compatible with five genotypers (GangSTR, HipSTR, ExpansionHunter, PopSTR2, and adVNTR, summarized in 3.2) and can easily be extended to handle VCFs from additional tools.

3.2 Features and Methods

TRTools consists of a suite of command-line utilities and a corresponding Python library for performing common operations on TR genotypes, including filtering, callset

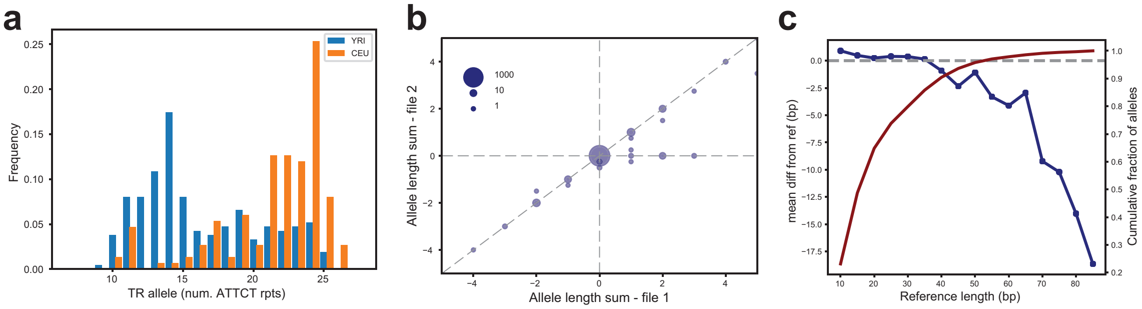

Figure 3.1: TRTools visualizations. (a) Allele frequency distribution at an example pentanucleotide TR output by statSTR based on GangSTR genotypes for two sample sets (YRI population consisting of Yorubans from Nigeria and CEU population of Northwestern European descent). (b) Example TR genotype comparison output by compareSTR. The plot compares genotypes (in terms of number of repeats difference from hg19) from HipSTR (x-axis) to those from ExpansionHunter (y-axis) on 5,000 tetranucleotide TRs. Bubble sizes give the number of calls included in each point. (c) Example reference bias plot output by qcSTR using popSTR2 genotypes. The plot shows the average deviation of TR alleles called vs. the reference length of the TR (in bp). The red line shows the cumulative percentage of allele calls below each reference length threshold.

comparisons, and other workflows. It parses VCF files using the PyVCF [Cas12] library and implements a “TR harmonizer” module that converts VCF formats from each tool to a standardized representation Supplementary Material). This harmonization step enables downstream operations to proceed agnostic of the original tool used to produce the genotypes. For all utilities described below, the --vcftype argument may be used to specify the genotyping tool used. If not specified, the type is automatically inferred. In the following sections, we summarize the current functionality available in TRTools. Utilities are summarized in Table 3.1. Each utility described below is available as a standalone command line tool within the TRTools package.

3.2.1 DumpSTR

dumpSTR is a tool for filtering TR VCF files. It performs call-level filtering (e.g., minimum call depth, minimum call quality) and locus-level filtering (e.g., minimum call rate or deviation from Hardy-Weinberg Equilibrium). dumpSTR is specially built to handle VCF FORMAT and INFO fields unique to TR genotypers. Unlike standard VCF filtering tools, it also computes locus-level metrics such as heterozygosity and Hardy-Weinberg Equilibrium based on TR allele lengths. It takes as input a VCF file and outputs a new VCF with locus-level filters annotated in the FILTER column and call-level filters annotated in the FORMAT field for each call.

```
dumpSTR --vcf VCF --out OUTPREFIX \
  [--vcftype={eh|gangstr|hipstr|popstr|advntr}] \
  [filters]
```

3.2.2 MergeSTR

mergeSTR is a method for merging VCF files generated by TR genotyping methods. While methods for merging VCF files currently exist [Li11], TR VCFs have unique characteristics that call for a specialized merging tool. TRs are often multi-allelic, and VCFs generated using different sample sets may contain different alternate allele sets. Further, existing tools may normalize TR alleles to remove redundant sequence when merging, which can interfere with downstream analysis of TR lengths. (For example, BCFtools [Li11] normalizes REF=CAG, ALT=CAGCAG to REF=C, ALT=CAGC, which is not desirable in a TR analysis). mergeSTR takes two or more VCFs generated by the same TR genotyper as input and outputs a merged VCF file containing all of the samples included in the input VCFs.

```
mergeSTR --vcfs VCF1,VCF2[,...],VCFn \
  --out OUTPREFIX
```

3.2.3 Statistics and QC utilities

TRTools provides a suite of statistics and QC utilities to allow fast high-level checks of TR runs.

statSTR allows users to compute locus level statistics on multi-sample TR VCFs, such as the mean allele length, allele frequency distributions, and call rate. It outputs a tab-delimited file listing user-specified statistics for each TR. statSTR additionally allows outputing plots of allele frequency distributions at specific TRs (Fig. 3.1a).

```
statSTR --vcf VCF --out OUTPREFIX [statistics]
```

compareSTR allows users to compare calls from two VCF files. These can be generated by the same or different tools. This allows users to compare calls across platforms or for different runtime options. Fig. 3.1b shows an example plot created by compareSTR comparing two call sets.

```
compareSTR --vcf1 VCF1 --vcf2 VCF2 \
  [--vcftype1 VCFTYPE] [--vcftype2 VCFTYPE] \
  --out OUTPREFIX [options] \
```

Table 3.1: Summary of current TRTools utilities

Command	Description
dumpSTR	Filter a TR genotype dataset
mergeSTR	Merge two or more VCFs generated by a TR genotyper
statSTR	Generate per-locus statistics from a VCF of TR genotypes
compareSTR	Compare two TR genotype call sets
qcSTR	Output quality control plots for a TR genotype call set

qcSTR automatically generates plots for performing quality control of TR genotype datasets. For example, Fig. 3.1c shows a plot demonstrating an expected deletion bias at long alleles based on popSTR2 genotypes.

```
qcSTR --vcf VCF --out OUTPREFIX [options]
```

Additional use cases for each utility using output from each supported TR genotyping tool are provided in the TRTools documentation.

3.2.4 Python library for data analysis

To enable researchers to leverage TRTools features in their own custom tools, we have packaged it as a Python library. The underlying functionality for operations such as harmonizing VCF records across TR genotypers or performing string manipulations on TR sequences can be accessed by importing the library into a Python script.

```
import vcf, trtools.utils.tr_harmonizer as trh
reader = vcf.Reader(open(”my.vcf”))
vcftype = trh.InferVCFType(reader)
rec = reader.next()
trrecord = trh.HarmonizeRecord(vcftype, rec)
trrecord.GetAlleleFrequencies(uselength=True)
# {10: 0.2, 15: 0.8} dict of num. rpts.->freq
```

3.3 Discussion

Quality control and filtering are crucial steps for nearly any genome- or population-scale analysis. TRTools meets a pressing need for standardized tools for performing these tasks on TR datasets, which are not handled well by mainstream tools. This toolkit currently supports five major TR genotypers. It can easily be extended to additional TR genotyping methods for either short or long reads as long as they are compatible with the VCF standard and report precise repeat copy numbers. Improved handling of imprecise repeat copy numbers and more complex repeat sequences reported by error-prone long reads is a topic of future development of TRTools. Finally, TRTools can incorporate additional utilities as the community continues to develop standards for TR analysis.

3.4 Supplementary Material

3.4.1 Datasets

BAM files containing reads from high-coverage whole genome sequencing datasets for the 1000 Genomes Project [FLGPF20] were accessed through the European Nucleotide Archive accession number PRJEB31736. They were processed using GangSTR [MSBYG19] v2.4.2.12 with non-default parameter --grid-threshold 250 using the TR reference file hg38_ver17.bed.gz available on the GangSTR website (https://github.com/gymreklab/gangstr). Allele frequencies for a pentanucleotide repeat in the promoter of RUNX1 (hg38 chr21:35348646-35348646) for samples from the YRI (Yoruba in Ibadan, Nigeria) and CEU (Utah Residents with Northern and Western European Ancestry) populations are shown in Fig. 3.1a in the main text.

Whole genome sequencing (BAM file aligned to hg19) for Platinum Genomes sample NA12881 was downloaded from dbGaP (accession phs001224.v1.p1). A single chromosome (chromosome 10) was extracted using samtools [LHW$^+$09]. TRs were genotyped using ExpansionHunter [ea17] v3.2.0 and HipSTR [WZY$^+$17a] v0.6.2. For both, we used the GangSTR version 16 reference subsetted to the first 5,000 tetranucleotides as the input set of TRs. ExpansionHunter was run with parameter -a path-aligner. HipSTR was otherwise run with default parameters. We used dumpSTR to filter each callset and compareSTR to generate the bubble plot in Fig. 3.1b using the options shown below

A VCF file generated by popSTR2 [KEAH19] on Platinum Genomes samples NA12891, NA12892, and NA12878 (dbGaP phs001224.v1.p1) was obtained from the PopSTR authors. This file was used to generate the reference bias plot shown in Fig. 3.1c in the main text.

3.4.2 TR Harmonizer implementation details

The TRHarmonizer Python library provides a uniform interface for accessing VCFs created by different tandem repeat (TR) genotypers. This library is the shared basis for all the command-line tools in the TRTools package. It is designed to cleanly handle differences in how different genotypers represent alleles, quality-scores and other metadata describing TR genotypes. This allows coding against a uniform interface while analyzing genetic variation at TRs regardless of which genotyper was used. The TRHarmonizer library also allows third parties to leverage the harmonization functionality outside of the command-line tools provided in TRTools.

A major challenge in analyzing TR genotypes is that alleles are represented differently in VCF outputs of different genotypers. The example below for chr21:47251618 (hg19) genotyped in Platinum Genomes sample NA12878 shows the different ways reference and alternate alleles are specified in VCFs by the genotypers which TRTools currently supports.

- adVNTR*, GangSTR, HipSTR
  - REF: AGTTAGTTAGTTAGTT
  - ALT: AGTTAGTTAGTTAGTTAGTT
- ExpansionHunter
  - REF: A
  - ALT: <STR5>

  - INFO: REF=4;RU=AGTT

- PopSTR
  - REF: AGTTAGTTAGTTAGTT
  - ALT: <5>
  - INFO: Motif=AGTT

* Note that while this TR was not called by AdVNTR because its motif is too short, AdVNTR output represents alleles in the same format as HipSTR and GangSTR.

Furthermore, consider the example at chr21:16402147:

- adVNTR, GangSTR, HipSTR
  - REF: AAATAAATAAATAAATAAAT
  - ALT: AAATAAATAAATAAAT
- ExpansionHunter
  - REF: A
  - ALT: <STR4>
  - INFO: REF=5;RU=AAAT
- PopSTR
  - REF: AAATAAATAAATAAATAAATAATAAA
  - ALT: <5.5>
  - INFO: Motif=AAAT

Here, popSTR's representation of alleles changes to specify impurities and partial repeats.

The key function of the TRHarmonizer module, HarmonizeRecord, takes as input a PyVCF [Cas12] record (a PyVCF.model._Record object) and a VCF type (one of: "advntr", "eh", "gangstr", "hipstr" or "popstr", corresponding to the supported genotypers) and outputs a TRRecord object (analogous to PyVCF.model._Record) storing alleles and other metadata in a standardized format. This allows downstream analyses to proceed agnostic of the genotyper which created the record. The TRRecord stores allele length genotypes as the number of copies of the motif corresponding to that length. This number is a float to allow for impurities and partial repeats. For genotypers which infer sequence alleles, the record additionally stores the sequence of the allele in all uppercase. In addition to alleles, a TRRecord also provides a uniform method for accessing the TR motif, per-sample quality scores and other metadata supplied by the underlying genotyper. The main text and TRTools documentation show examples of how to use the TRHarmonizer interface from Python.

TRHarmonizer is designed to be lightweight, and as such there are similar yet more complex use-cases that TRHarmonizer intentionally does not support. It does not have any insight into sequencing technologies which produce data that is later processed by TR genotypers into VCFs. As such it relies on the alleles, calls and associated quality scores output by the genotypers, each of which use their own models to compute quality scores. TRTools makes no attempt to modify those scores based on sequencing errors or other sources of error.

TRHarmonizer also does not handle differences in variant coordinates, whether due to differences in choice of variant reference set or differences between calling algorithms. Note that this is only relevant to compareSTR, as that is the only one of our tools designed to process TRs from multiple VCFs produced by different genotypers simultaneously. The types of differences related to variant coordinates that TRHarmonizer does not handle includes:

- Repeat regions which some callers choose to represent as a single variant and other callers represent as multiple variants

- Overlapping variants of different lengths due to decisions about whether to phase the repeat variant with other nearby variants

- Overlapping variants of different lengths due to different choices as to which parts of a locus constitute impure repeats and which constitute flanking regions

Rather, TRHarmonizer restricts itself to comparing variants called by different callers whose reference alleles start and end at the same base pairs. Handling different variant representations is a complex problem that has been the subject of significant work [CBG+15, Kru] and is best handled by haplotype comparison tools which have been tailored to the specific use-case at hand.

Finally, TRHarmonizer can be readily extended to support any TR genotyping tool built on top of any sequencing or genotyping technology as long as the tool produces a valid VCF file representing each TR as a distinct record in the VCF. Supporting additional tools simply requires adding a short function to the TRHarmonizer module converting records to the standardized format described above.

3.4.3 Commands for generating figures

The following code snippets show the commands used to generate the figures in the main text.

Listing 3.1: Code to generate Fig. 3.1a

```
#!/bin/bash
# YRIVCF and CEUVCF were generated by GangSTR v2.4.2.12
REGION=chr21:35348646-35348646
```

```
tabix --print-header $YRIVCF $REGION | bgzip -c > yri_runx1.vcf.gz
tabix --print-header $CEUVCF $REGION | bgzip -c > ceu_runx1.vcf.gz
tabix -p vcf yri_runx1.vcf.gz
tabix -p vcf ceu_runx1.vcf.gz

# Merge
mergeSTR --vcfs yri_runx1.vcf.gz,ceu_runx1.vcf.gz --out yri_ceu_runx1
bgzip -f yri_ceu_runx1.vcf
tabix -p vcf -f yri_ceu_runx1.vcf.gz

# Get sample lists
bcftools query -l yri_runx1.vcf.gz > yri_samples.txt
bcftools query -l ceu_runx1.vcf.gz > ceu_samples.txt

# StatSTR
# Compute stats separately on YRI and CEU samples
statSTR \
    --vcf yri_ceu_runx1.vcf.gz \
    --samples yri_samples.txt,ceu_samples.txt \
    --sample-prefixes YRI,CEU \
    --region $REGION \
    --out yri_ceu_runx1 \
    --afreq --use-length --plot-afreq
# Output file yri_ceu_runx1-chr21-35348646.pdf shown in Figure
```

Listing 3.2: Code to generate Fig. 3.1b

```
#!/bin/bash

SAMPLE=NA12881
# $SAMPLE-hipstr.vcf.gz and $SAMPLE-eh-path.vcf.gz generated
# by calling HipSTR and ExpansionHunter on the same TR reference

# Filter
dumpSTR \
    --vcf $SAMPLE-hipstr.vcf.gz \
    --hipstr-min-call-Q 0.9 \
    --hipstr-min-call-DP 10 \
    --hipstr-max-call-DP 1000 \
    --hipstr-max-call-flank-indel 0.15 \
    --hipstr-max-call-stutter 0.15 \
    --hipstr-min-supp-reads 2 \
    --out $SAMPLE-hipstr.filtered
cat $SAMPLE-hipstr.filtered.vcf | vcf-sort | \
    bgzip -c > $SAMPLE-hipstr.filtered.vcf.gz
tabix -p vcf $SAMPLE-hipstr.filtered.vcf.gz

dumpSTR \
    --vcf $SAMPLE-EH-path.vcf \
    --vcftype eh \
    --eh-min-call-LC 50 \
    --out $SAMPLE-eh-path.filtered
# Edit sample name to be same
```

```
cat $SAMPLE-eh-path.filtered.vcf | sed 's/NA12881.chr10/NA12881/' | \
    vcf-sort | bgzip -c > $SAMPLE-eh-path.filtered.vcf.gz
tabix -p vcf $SAMPLE-eh-path.filtered.vcf.gz

# Add contigs to EH
zcat $SAMPLE-hipstr.filtered.vcf.gz | grep congi > hg19_contigs.txt
bcftools annotate -h hg19_contigs.txt \
    $SAMPLE-eh-path.filtered.vcf.gz | \
    bgzip -c > $SAMPLE-eh-path-contigs.filtered.vcf.gz
tabix -p vcf -f $SAMPLE-eh-path-contigs.filtered.vcf.gz

# Make bubbles plot and compare
compareSTR \
    --vcf1 $SAMPLE-eh-path-contigs.filtered.vcf.gz \
    --vcf2 $SAMPLE-hipstr.filtered.vcf.gz \
    --vcftype1 eh \
    --vcftype2 hipstr \
    --out eh-path-hipstr \
    --bubble-min -5 --bubble-max 5
# Output file eh-path-hipstr-bubble-periodALL.pdf shown in Figure
```

Listing 3.3: Code to generate Fig. 3.1c

```
#!/bin/bash
qcSTR --vcf $popSTR_vcf --out popstr_qc
# Output file popstr_qc-diffref-bias.pdf shown in Figure
```

Table 3.2: TR calling methods currently supported by TRTools. *These tools may be run on Illumina data that is not PCR-free, but may have reduced accuracy on those datasets.

Method (Version tested)	Supported TR classes	Num. TRs in reference	Supported sequencing technologies	Use case notes
AdVNTR [BSBG+18] (v1.3.3)	Repeat unit length 6-100bp.	158,522 (genic hg19)	Illumina or PacBio	Designed for targeted genotyping of VNTRs on a single sample at a time. Only handles repeats shorter than the read length. Infers allele lengths by default. May alternatively identify putative frameshift mutations within VNTRs. May be run on large panels of TRs but is compute-intenstive.
Exp. Hunter [ea17] (v3.2.2)	Designed for STRs (typically with repeat unit length ≤6bp). Can handle complex repeat structures specified by regular expressions (e.g. (CAG)*(CCG)*).	25 (hg19)	PCR-free* Illumina	Designed for targeted genotyping of repeat expansions at known pathogenic TRs but may be run genome-wide on both short and expanded TRs using a custom TR panel. Can handle repeats with complex structures such as interruptions or nearby repeats.
GangSTR [MSBYG19] (v2.4.4)	Designed for STRs or VNTRs with repeat unit length 1-20bp.	829,233 (hg19_ver_13_1, excludes homopolymers)*	Paired-end PCR-free* Illumina	Designed for genome-wide genotyping of short or expanded TRs. Infers allele lengths. Allows multi-sample calling.
HipSTR [WZY+17a] (v0.6.2)	Repeat unit length 1-9bp.	1,620,030 (hg19)	Illumina	Designed for genome-wide genotyping of STRs shorter than the read length. Can phase repeats with SNPs. Allows multi-sample calling.
PopSTR2 [KEAH19]	Repeat unit length 1-6bp.	540,1401 (hg38)	Illumina	Designed for genome-wide genotyping of short or expanded TRs. Allows multi-sample calling.

3.4.4 Supplementary Tables

Table 3.2 shows TR callings methods currently supported by TRTools. Since each of these tools take as input a list of TRs to genotype, they can also be used on custom panels of TR loci. Tool information and reference panel numbers shown above are based on downloads from the github repository of each tool as of July 2, 2020.

Chapter 3, in full, is a reprint of the material as it appears in Bioinformatics 2020. Mousavi, Nima, Jonathan Margoliash, Neha Pusarla, Shubham Saini, Richard Yanicky, and Melissa Gymrek. The dissertation author was one of the primary investigators and authors of this paper.

Chapter 4

Applications of Methods for Genotyping and Filtering of Tandem Repeats

Chapters 2 and 3 describe the methods that we have developed for creating, filtering, and post-processing of genome-wide tandem repeat (TR) callsets. In this chapter, I describe our approach to applying these methods to solve biological and population genetics problems.

Section 4.1 describes my methodological contributions to Mitra et al. ([MHM$^+$21]). The goal of my work was to lay the foundations for a pipeline for discovery and evaluation of de-novo variants that contribute to Autism Spectrum Disorder (ASD) risk. To this end, I have made updates to our method for genome-wide genotyping of TRs, GangSTR [MSBYG19], to facilitate discovery of de-novo variants. Furthermore, I have created the simulated data for the simulation platform to validate our calls.

Section 4.2 explains our ongoing work on creating an ensemble TR callset by combining information from multiple different TR variant callers. This callset requires us to create a new method that utilizes the strength of each caller (as determined through statistical measures and model limitations) to identify the most confident TR call at each locus.

We apply this method to samples of the 1000 genomes project [FLGPF20] to facilitate downstream population genetics analyses in the future.

4.1 Methods for Identifying de-novo Tandem Repeat mutations in individuals affected by Autism Spectrum Disorder

Most of this section was first published as part of:

Mitra, I., Huang, B., Mousavi, N., Ma, N., Lamkin, M., Yanicky, R., Shleizer-Burko, S., Lohmueller, K., & Gymrek, M. Patterns of de novo tandem repeat mutations and their role in autism. Nature. (2021)

Autism Spectrum Disorder (ASD) are a group of neurological disorders with early age of onset. These disorders are characterized by a heterogeneous set of symptoms such as restricted interest, lack of communication, or impaired socialization [RSVG14]. Family studies have shown a major contribution of genetic factors to ASD risk. These genetic factors can be inherited or de-novo [MHM+21]. De-novo variants are variants that are only present in the offspring, and are absent in both parents. Specifically, contribution of de-novo variants to ASD risk has been estimated to be around 30% of the simplex cases (only a single child affected by ASD per family) [IOS+14].

In [MHM+21], Mitra et al. set out to study the contribution of de-novo TR variants. In this work, I contributed the following methodological work.

4.1.1 Improvements to GangSTR: method for genome-wide genotyping of tandem repeats

Autism Spectrum Disorder (ASD) has a strong male bias. The ratio of male to female individuals affected by ASD is between 4:1 and 3:1 [LHM17]. Due to this disparity across sexes we have to study sex chromosomes with additional rigor. The original version of GangSTR, our method for genome-wide genotyping of TRs described in Section 2, did not support genotyping in sex chromosomes. In order to be able to study contribution de-novo variant on sex chromosomes to ASD risk, I developed support for sex chromosomes in GangSTR. In addition, the downstream method for identifying de-novo variants, MonSTR [MHM+21], requires genotype likelihoods and read support for each call to be reported in the output. Read support for each calls shows how many specific reads supported each individual call, which in concert with genotype likelihoods, were used for filtering purposes in MonSTR. I further improved GangSTR to report these values for each TR call.

4.1.2 Evaluating de-novo variant discovery using simulation

The following paragraphs from Mitra et al. [MHM+21] describe my contribution to our pipeline for evaluating de-novo TR calls using a simulated dataset.

We created 78 quad families with 100 TR loci randomly selected from TRs passing all filters described above in the SSC cohort. One simulated quad family consists of the father, mother, child with known mutation (proband), and child with no mutation (control). We tested the ability of our entire pipeline to genotype TRs with GangSTR and call de novo mutations with MonSTR. To test the effect of depth of coverage, we generated datasets with 1–50× mean coverage with a mutation size of +1 or -1 repeat unit changes in the proband. To test the effect of TR mutation size, we generated WGS data with 40× coverage and mutations in probands ranging from -10 to 30 repeat unit changes. Contraction mutations that would have resulted in negative repeat copy numbers were excluded. For both tests, we simulated data under three scenarios: (1) both parents with homozygous reference TR genotypes, (2) one parent heterozygous, (3) both parents heterozygous (Figure ??).

WGS data were simulated using ART_illumina [HLMM12] v2.5.8 with non-default parameters -ss HS25 (HiSeq 2500 simulation profile), -l 150 (150 bp reads), -p (paired-end reads), -f coverage (coverage was set as described above), -m 500 (mean fragment size) and -s 100 (standard deviation of fragment size). ART_illumina was applied to fasta files generated from 10-kb windows surrounding each TR locus, applying any mutations as described above. The resulting fastq files were aligned to the hg38 reference genome using bwa mem [Li13] v0.7.12-r1039 with non-default parameter -R "@RG\tID:sample_id\tSM:sample_id\", which sets the read group tag ID and sample name to sample_id for each simulated sample. TRs were genotyped from aligned reads jointly across all members of the same family with GangSTR using identical settings to those applied to SSC data.

We tested three mutation-calling settings: a naive mutation-calling method based on hard genotype calls, MonSTR using default parameters, and MonSTR using an identical set of filters as applied to SSC data. We found overall all methods perform similarly well above 30× coverage. At lower coverage, MonSTR's model-based method achieves reduced sensitivity but greater specificity compared to a naive mutation-calling pipeline (Figure ??)."

4.2 Method for creating ensemble Tandem Repeats Callset

Tandem Repeat (TR) variant calling methods have created a fast and inexpensive way of analyzing the genetic makeup of individuals at Tandem Repeat loci using sequencing data. Several methods for genotyping TRs using short-read sequencing data have been published in the past few years [MSBYG19, ea17, WZY$^+$17a, BSBG$^+$18]. Each method focuses on a different approach for calling TRs, and as a result, each method has advantages and weaknesses. Using our knowledge on pros and cons of each TR variant calling method, we have created a graph-based method for merging callsets generated by each caller to create an ensemble callset.

Creating this merged ensemble callset will produce a dataset that is more accurate than each individual variant caller, as demonstrated by experimental validation and Mendelian study. After phasing the variants and combining the TR callset with a single nucleotide variant dataset, we can create a SNV-TR haplotype panel, similar to the work by Saini et al. [SMG18]. Such haplotype panel can be used to impute TR calls into abundantly available SNV datasets at a very low computational cost.

Figure 4.2 shows the pipeline for generating ensemble callset and variant phasing. The first step of generating the merged callset is running each of the variant callers on all of the samples from the 1000 Genomes dataset. We perform variant calling using HipSTR [WZY$^+$17a], GangSTR [MSBYG19], ExpansionHunter [ea17], and AdVNTR [BSBG$^+$18]. The variant calling is performed on 2504 unrelated samples from phase three of the 1000 Genomes project [FLGPF20]. For validation purposes, we genotype 698 additional related samples [BBEZ$^+$21] as well.

The output Variant Call Format (VCF) file from each caller is then used by our novel method for generating ensemble callsets. The following sections describe the details of our approach.

4.2.1 Graph-based Ensemble Variant Caller

In this section we discuss the algorithmic details of the graph-based ensemble TR variant caller (TR-ensemble in Figure 4.2).

The first step in merging TR callsets is to identify overlapping calls that can be merged using this method. To do so, we process all VCF files that are input to our method one variant at a time. At any given point during running time, we can identify the next variant to be processed from each VCF file with a pointer (Figure 4.3). Since we are starting from sorted VCF files, we can process the variants in the same order. Among the four next variants, the variant with the smallest genomic position is named minimum variant. We evaluate all next variants to find the variants that overlap with minimum variant and have the same canonical motif. This group of overlapping variants with the same motif are called a record cluster. We have to merge each record cluster separately and report a merged TR call (see Section 4.2.1.1). If the record cluster only includes the minimum variant, there are no overlapping variants to merge and we report the minimum variant to the output. The pointers in VCF files that contributed variants to the record cluster are all incremented and the process is repeated with the new set of four next variants.

4.2.1.1 Locus Merger

Variants that overlap the minimum variant and share canonical repeat motif form a record cluster. This section describes how we merge the variants in a record cluster. As a first step, we create a locus graph based on all of the calls in the record cluster. Figures 4.4 and 4.5 show examples of record cluster graph. In these figures, each node corresponds to an allele. Alleles of equal length are connected by edges. This allows us to identify which alleles across different callers have the same length. The annotation for each node shows the origin of the called allele (methods: eh(ExpansionHunter), hipstr(HipSTR), gangstr(GangSTR)) and the length difference from the reference allele. $*$ denotes the reference allele.

4.2.1.2 Caller QC

After identifying all of the calls from each caller at each locus, we can compute statistics such as accordance with Hardy Weinberg Equilibrium, call rate, Mendelian inheritance rate, etc. to facilitate scoring how much we trust each caller at a specific locus. This information will be useful when merging alleles in the next step.

4.2.1.3 Allele Merger

We are able to find the calls that are compatible with each other (have the same length) across different methods using the annotation described in the previous section. Each connected component in record cluster graphs corresponds to a group of compatible alleles across different methods. If there is no ambiguity in the calls made by different callers, we have only 1-1 mappings in each connected component (see Figure 4.4). Alternatively, Figure 4.5 shows an example of an ambiguous and non-trivial connected component. In this example, all alleles have lengths equal to the reference allele. However, there are two different nodes corresponding to HipSTR calls, only one of which matches the sequence of the reference allele (showed by *). The other node denotes an allele with the same length as the reference allele, but with a different sequence. This is a fairly common occurrence as HipSTR is the only method capable of making sequence level calls (other methods can only make length calls). This advantage of the HipSTR method means, in this connected component, we trust the calls from HipSTR more than other methods.

We create a mapping between input alleles in each connected components and a merged call to facilitate finding the ensemble calls. We use information that we have on advantages of each method (HipSTR is the only method capable of making sequence level calls, ExpansionHunter and GangSTR are capable of calling variants longer than read length, and AdVNTR is the only method genotyping the longer motifs) as well as QC metrics (Section 4.2.1.2) to create this mapping.

4.2.1.4 Call Merger

Using the mapping between input calls and output merged calls created in the Allele Merger step (Section 4.2.1.3), we can find the merged TR corresponding to any call. This process is performed by iterating over the calls for each sample and identifying the output merged call. Appropriate measures of confidence (which methods agree with the final call, etc.) are reported with the merged call.

4.2.2 Validation

We use capillary electrophoresis to experimentally validate a subset of the calls. Figure 4.2 shows how our validation ties in with generation of the ensemble TR callset. For each locus we experimentally validate calls made by the ensemble caller. A subset of samples included in our validation plan are form parent child trios [BBEZ+21]. This allows us to have an orthogonal method of validating the calls. Figure 4.6 shows an example validating calls in a trio.

4.2.3 Future Works

This work is an ongoing effort in gaining a better understanding of the genome-wide landscape of TRs. The merged ensemble callset has been created using calls from the 1000 genomes dataset [BBEZ+21]. This callset includes a total of 858,565 loci. 74 loci have undergone experimental validation, and more will be added to our validated dataset. After generating the merged ensemble callset, we want to study population specific alleles and selection patterns. Furthermore, by phasing the variants and using a SNV dataset, we can create a SNV-TR haplotype panel that can be used for imputing TRs into widely available SNV callsets.

Chapter 4, in part, is a reprint of the material as it appears in Nature 2021. Mitra, Ileena, Bonnie Huang, Nima Mousavi, Nichole Ma, Michael Lamkin, Richard Yanicky, Sharona Shleizer-Burko, Kirk E. Lohmueller, and Melissa Gymrek. The dissertation author was one of the co-authors and investigative collaborators of this paper.

Chapter 4, in part, is unpublished material by Nima Mousavi, Nichole Ma, Helyaneh Ziaei Jam, Bonnie Huang, Mikhail Maksimov, Jonghun Park, Yunjiang Qiu, Egor Dolzhenko, Vineet Bafna, and Melissa Gymrek. The dissertation author was the primary investigator

and author of this paper.

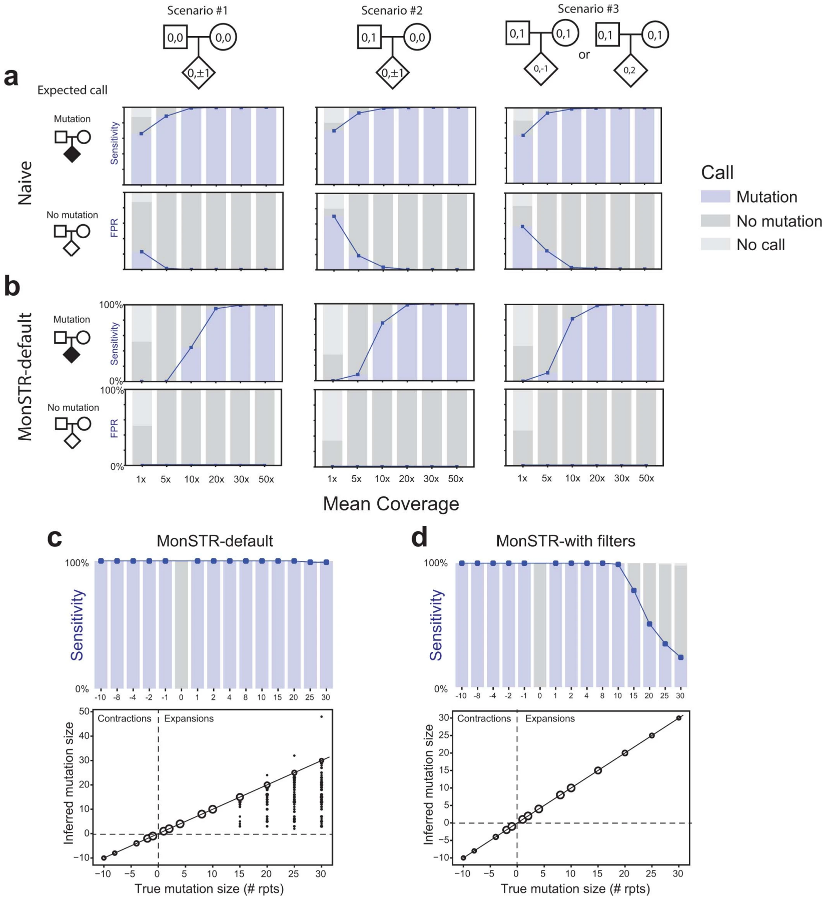

Figure 4.1: a, Evaluation of a naive TR mutation-calling method. WGS was simulated for probands with mutations and controls with no mutation under three different scenarios for a range of mean sequencing coverages (Methods of [MHM$^+$21]). Top plots show the sensitivity (blue line). Bottom plots show the false positive rate (FPR). Shaded bars show the percent of transmissions called as mutation (blue), no mutation (dark grey), or no call (light ray). b, Evaluation of MonSTR's default model-based method. Plots are the same as in a. but based on MonSTR's default model (Supplementary Methods of [MHM$^+$21]). Note FPR lines are not visible because all are at 0%. c, Evaluation of TR mutation calling using default model-based MonSTR settings as a function of mutation size. The top plot is the same as in a, b, and shows the sensitivity to detect mutations as a function of their size. The bottom plot compares the estimated called mutation size (y-axis) compared to the true simulated mutation size (x-axis). Bubble sizes show the number of mutation calls represented at each point. d, Evaluation of TR mutation calling as a function of mutation size after quality filtering. Plots are same as in c, but using the stringent quality filters in MonSTR applied to analyze the SSC cohort. Compared to default settings, sensitivity is decreased especially for larger expansions but inferred mutation sizes are unbiased. All plots are based on simulation of 100 randomly chosen TR loci (Methods). c, d, show results for scenario #1. Figure from Mitra et al. [MHM$^+$21].

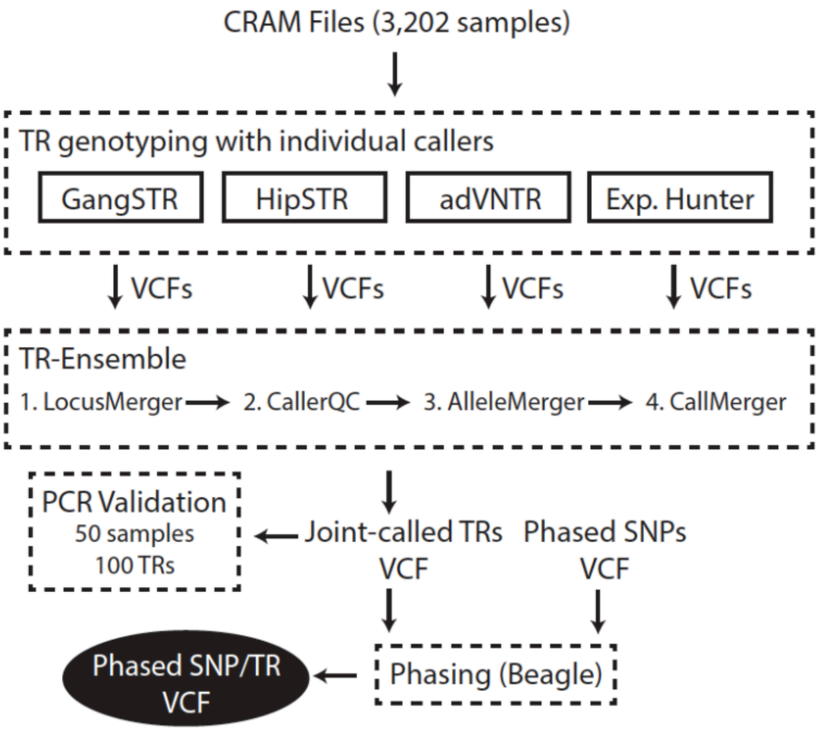

Figure 4.2: Pipeline for creating ensemble TR callset using 1000 Genomes dataset.

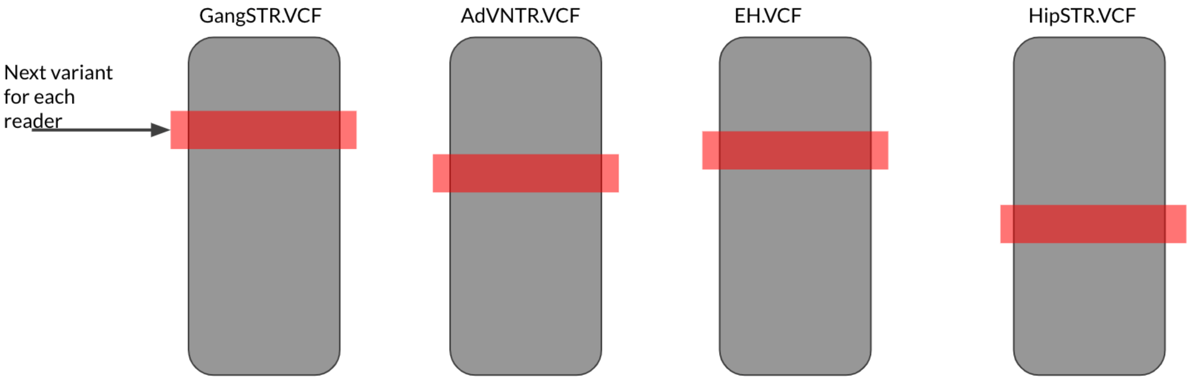

Figure 4.3: Pointers to the next processed VCF record in each TR VCF file.

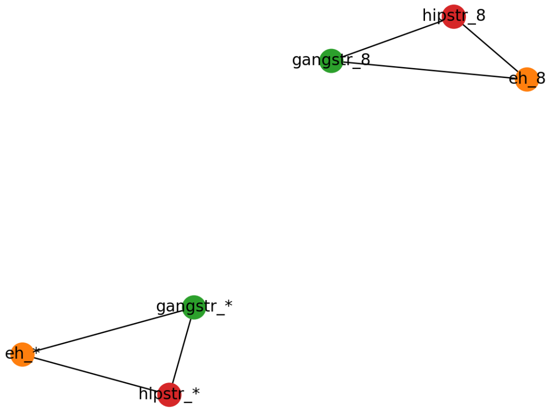

Figure 4.4: Example of a trivial graph showing a 1-1-1 mapping between alleles from different callers.

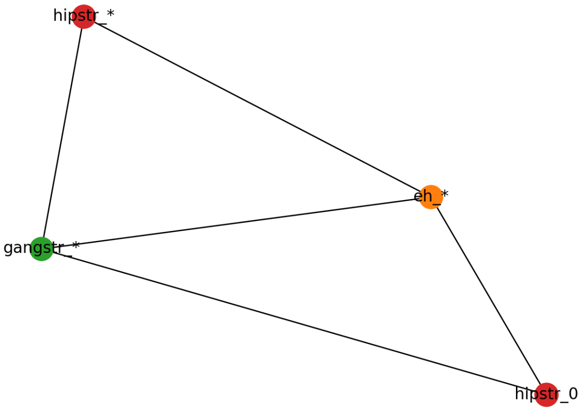

Figure 4.5: Example of a non-trivial connected component in allele graph.

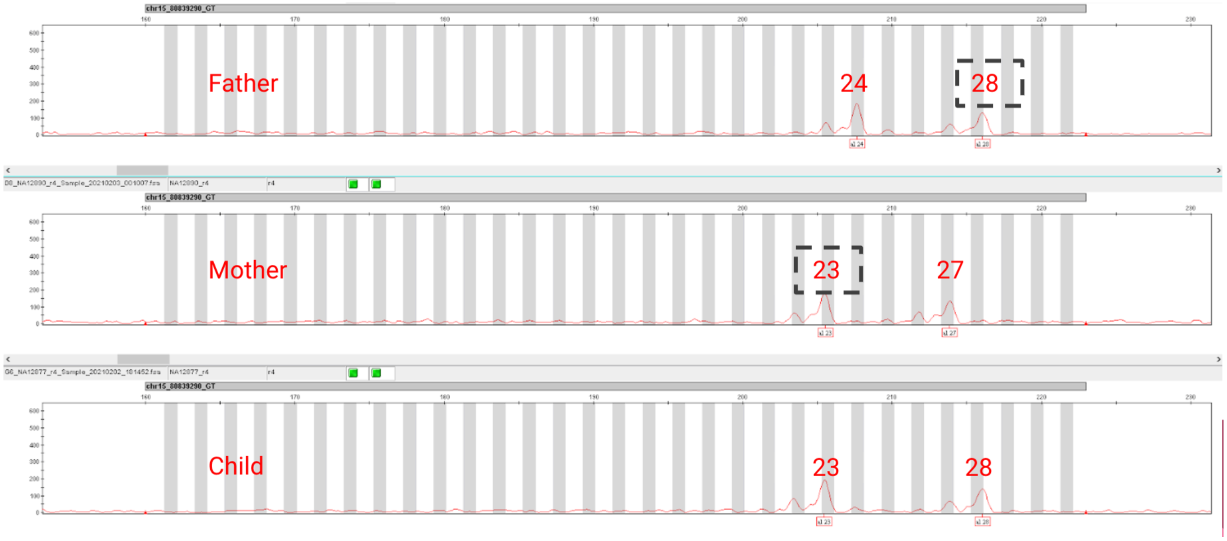

Figure 4.6: Peaks from capillary electrophoresis correspond to TR calls made on a trio of samples.

Chapter 5

Conclusion

I have presented a set of tools for processing tandem repeat (TR) variants in the human genome. Chapter 2 describes our method for genome-wide genotyping of TRs. Chapter 3 explains our toolkit for post-processing, quality control, and filtering of TR callsets. Chapter 4 provides two examples of applying these tools to solve biological (contribution of de-novo TR variant to ASD) and statistical problems (creating an ensemble TR callset across a population-scale dataset).

Bibliography

[ADD$^+$00] 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.

[BAG$^+$18] 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.

[BBEZ$^+$21] Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, Anna O Basile, Haley J Abel, Allison A Regier, André Corvelo, Wayne E Clarke, Rajeeva Musunuri, Kshithija Nagulapalli, et al. High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios. bioRxiv, 2021.

[Ben99] G Benson. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res., 27(2):573–580, January 1999.

[BLC$^+$08] 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.

[BOH$^+$16] 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.

[BSBG$^+$17] Mehrdad Bakhtiari, Sharona Shleizer-Burko, Melissa Gymrek, Vikas Bansal, and Vineet Bafna. Targeted genotyping of variable number tandem repeats with adVNTR, 2017.

[BSBG$^+$18] M. Bakhtiari, S. Shleizer-Burko, M. Gymrek, V. Bansal, and V. Bafna. Targeted genotyping of variable number tandem repeats with advntr. Genome Res, 28:1709–1719, 2018.

[BTPP$^+$18] 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.

[Cas12] J. Casbon. Pyvcf - a variant call format parser for python. 2012.

[CBG+15] John G Cleary, Ross Braithwaite, Kurt Gaastra, Brian S Hilbush, Stuart Inglis, Sean A Irvine, Alan Jackson, Richard Littin, Mehul Rathod, David Ware, et al. Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. BioRxiv, page 023754, 2015.

[CMM+04] Claudia Cagnoli, Chiara Michielotto, Tohru Matsuura, Tetsuo Ashizawa, Russell L Margolis, Susan E Holmes, Cinzia Gellera, Nicola Migone, and Alfredo Brusco. Detection of large pathogenic expansions in frda1, sca10, and sca12 genes using a simple fluorescent repeat-primed pcr assay. The Journal of Molecular Diagnostics, 6(2):96–100, 2004.

[CSD+17] 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.

[DLP+17] 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.

[DLP+18] 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.

[DRDCB+19] Arne De Roeck, Wouter De Coster, Liene Bossaerts, Rita Cacace, Tim De Pooter, Jasper Van Dongen, Svenn D'Hert, Peter De Rijk, Mojca Strazisar, Christine Van Broeckhoven, et al. Nanosatellite: accurate characterization of expanded tandem repeat length and sequence through whole genome long-read sequencing on promethion. Genome biology, 20(1):1–16, 2019.

[DvVS+17] 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.

[ea17] Dolzhenko et al. Detection of long repeat expansions from pcr-free whole-genome sequence data. Genome Res., 27:1895–1903, 2017.

[FLGPF20] Susan Fairley, Ernesto Lowy-Gallego, Emily Perry, and Paul Flicek. The international genome sample resource (igsr) collection of open human genomic variation resources. Nucleic Acids Research, 48(D1):D941–D947, 2020.

[FMW+19] Stephanie Feupe Fotsing, Jonathan Margoliash, Catherine Wang, Shubham Saini, Richard Yanicky, Sharona Shleizer-Burko, Alon Goren, and Melissa Gymrek. The impact of short tandem repeat variation on gene expression. Nature genetics, 51(11):1652–1659, 2019.

[GGRE12] 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.

[GWG+16] 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.

[GYM+20] Devika Ganesamoorthy, Mengjia Yan, Valentine Murigneux, Chenxi Zhou, Minh Duc Cao, Tania PS Duarte, and Lachlan JM Coin. High-throughput multiplexed tandem repeat genotyping using targeted long-read sequencing. F1000Research, 9(1084):1084, 2020.

[Han10] Anthony J Hannan. Tandem repeat polymorphisms: modulators of disease susceptibility and candidates for 'missing heritability'. Trends Genet., 26(2):59–65, February 2010.

[Han18a] A. J. Hannan. Tandem repeats mediating genetic plasticity in health and disease. Nat. Rev. Genet, 19:286–298, 2018.

[Han18b] Anthony J Hannan. Tandem repeats mediating genetic plasticity in health and disease. Nat. Rev. Genet., 19(5):286–298, May 2018.

[HFM^{+}13] 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.

[HLMM12] Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. Art: a next-generation sequencing read simulator. Bioinformatics, 28(4):593–594, 2012.

[HRAA^{+}14] 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.

[IDM^{+}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.

[IOS^{+}14] Ivan Iossifov, Brian J O'roak, Stephan J Sanders, Michael Ronemus, Niklas Krumm, Dan Levy, Holly A Stessman, Kali T Witherspoon, Laura Vives, Karynne E Patterson, et al. The contribution of de novo coding mutations to autism spectrum disorder. Nature, 515(7526):216–221, 2014.

[Joh14] Steven G Johnson. The nlopt nonlinear-optimization package, 2014.

[JOPA16] 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.

[KEAH19] S. Kristmundsdottir, H. P. Eggertsson, G. A. Arnadottir, and B. V. Halldorsson. popSTR2 enables clinical and population-scale genotyping of microsatellites. Bioinformatics, 2019.

[KPL16] 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.

[Kru] Peter Krusche. Haplotype comparison tools.

[KRW$^+$18] Sergey Koren, Arang Rhie, Brian P Walenz, Alexander T Dilthey, Derek M Bickhart, Sarah B Kingan, Stefan Hiendleder, John L Williams, Timothy PL Smith, and Adam M Phillippy. De novo assembly of haplotype-resolved genomes with trio binning. Nature biotechnology, 36(12):1174, 2018.

[KSF$^+$02] 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.

[KSKH16] Snædís Kristmundsdóttir, Brynja D Sigurpálsdóttir, Birte Kehr, and Bjarni V Halldórsson. popSTR: population-scale detection of STR variants. Bioinformatics, September 2016.

[LEP$^+$13] 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.

[LHM17] Rachel Loomes, Laura Hull, and William Polmear Locke Mandy. What is the male-to-female ratio in autism spectrum disorder? a systematic review and meta-analysis. Journal of the American Academy of Child & Adolescent Psychiatry, 56(6):466–474, 2017.

[LHW$^+$09] 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.

[Li11] H. Li. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27:2987–2993, 2011.

[Li13] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. March 2013.

[Li18] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, 34(18):3094–3100, 2018.

[LZW$^+$17] Qian Liu, Peng Zhang, Depeng Wang, Weihong Gu, and Kai Wang. Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. Genome medicine, 9(1):1–16, 2017.

[McC10] Alice McCarthy. Third generation DNA sequencing: pacific biosciences' single molecule real time technology. Chem. Biol., 17(7):675–676, July 2010.

[MHM$^+$21] Ileena Mitra, Bonnie Huang, Nima Mousavi, Nichole Ma, Michael Lamkin, Richard Yanicky, Sharona Shleizer-Burko, Kirk E Lohmueller, and Melissa Gymrek. Patterns of de novo tandem repeat mutations and their role in autism. Nature, 589(7841):246–250, 2021.

[Mir07] S. M. Mirkin. Expandable DNA repeats and human disease. Nature, 447(7147):932–940, Jun 2007.

[MLL$^+$16] 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.

[MSBYG19] N. Mousavi, S. Shleizer-Burko, R. Yanicky, and M. Gymrek. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res, 47, 2019.

[PCQ14] 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.

[PGB$^+$18] Franziska Pfeiffer, Carsten Gröber, Michael Blank, Kristian Händler, Marc Beyer, Joachim L Schultze, and Günter Mayer. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific reports, 8(1):1–14, 2018.

[PGM$^+$18] 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.

[Pow94] 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.

[PWD$^+$12] 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.

[QGG$^+$16] 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.

[RMSC14] 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.

[RSVG14] Rasim O Rosti, Abdelrahim A Sadek, Keith K Vaux, and Joseph G Gleeson. The genetic landscape of autism spectrum disorders. Developmental Medicine & Child Neurology, 56(1):12–18, 2014.

[Rya19] Calen P Ryan. Tandem repeat disorders. Evolution, medicine, and public health, 2019.

[SHG$^+$09] 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.

[SMG18] Shubham Saini, Ileena Mitra, and Melissa Gymrek. A reference haplotype panel for genome-wide imputation of short tandem repeats, 2018.

[SMS03] Subbaya Subramanian, Rakesh K Mishra, and Lalji Singh. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. Genome biology, 4(2):1–10, 2003.

[TBD$^+$18] 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.

[TDLB17] 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.

[TKL$^+$17] 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.

[VWZ$^+$17] 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.

[WdCW$^+$17] 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.

[WFS$^+$18] 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.

[WGH$^+$14] 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.

[WZY$^+$17a] T. Willems, D. Zielinski, J. Yuan, A. Gordon, M. Gymrek, and Y. Erlich. Genome-wide profiling of heritable and de novo str variations. Nat. Methods, 14:590–592, 2017.

[WZY+17b] 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.

[XCB+01] 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.

[YMR+13] 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.

[ZLGM13] 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.