Using High-Fidelity Long-Read Sequencing to Better Detect and Understand Short Tandem Repeat Variation in Humans

David R. Lougheed

Department of Human Genetics, Faculty of Medicine and Health Sciences,

McGill University, Montreal, Quebec, Canada

December 2022

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science.

Abstract

Variation in short tandem repeats (STRs) is implicated in many diseases, often in the form of STR expansion disorders, and complex traits. Traditional PCR-based STR genotyping has low throughput, and recent advances in genomic sequencing technologies have resulted in the discovery of additional STR-phenotype associations. Short-read sequencing struggles to resolve variation in long STR tracts, making long reads an attractive prospect for whole-genome STR genotyping, especially with a technology such as circular consensus sequencing (CCS) which boasts both long reads and low error rates. However, current long read STR genotyping methods have been designed with high-error long reads in mind, and do not take full advantage of the sequencing accuracy possible with CCS.

To address this gap, I develop a new long-read STR genotyping tool, called `STRkit`. `STRkit` uses a dynamic programming algorithm to realign candidate tandem repeats to each read and determine STR motif copy number, with several filtering steps to eliminate or re-process low-quality reads and alignments. From here, the software applies a Gaussian mixture model tuned for high-fidelity long reads to determine the most likely genotype in terms of copy number for the locus. Using `STRkit`, I show that CCS reads outperform nanopore-sequenced long reads and paired-end short reads for STR genotyping on a whole-genome benchmark. I also demonstrate that `STRkit` outperforms other long-read STR genotyping software when given either CCS or nanopore-sequenced long reads and can detect pathogenic STR expansions.

Overall, I show that high-fidelity long reads are useful for resolving STR copy number variation of varying magnitudes, especially when using appropriate STR genotyping software. For STR genotyping, one long read yields one copy number, which may in the future help analyze STR motif or copy number mosaicism and facilitate disentangling non-identical long STR allele genotypes.

Résumé

La variation des répétitions en tandem courtes (STR) est impliquée dans de nombreuses maladies, souvent sous la forme de troubles associés à l'expansion des STR. Le génotypage traditionnel des STR basé sur la PCR a un faible débit mais les progrès récents des technologies de séquençage génomique ont permis de découvrir d'autres associations STR-phénotype. Cela dit, le séquençage à lecture courte a du mal à résoudre la variation dans les longs segments STR. Dans ce contexte, les technologies de lectures longues, telle que le séquençage par consensus circulaire (CCS) qui se targue à la fois de lectures longues et de faibles taux d'erreur, apporte une perspective attrayante pour le génotypage STR du génome entier. Cependant, les méthodes actuelles de génotypage STR avec des de données de lecture longue ont été conçues en tenant compte d'un haut taux d'erreur et ne tirent pas pleinement parti de la précision de séquençage possible avec le CCS.

Dans le cadre de cette thèse, j'ai développé un nouvel outil de génotypage STR qui s'applique aux lectures longues, appelé STRkit. STRkit utilise un algorithme de programmation dynamique pour réaligner les répétitions en tandem candidates sur chaque lecture et déterminer le nombre de copies du motif STR, avec plusieurs étapes de filtrage pour éliminer ou retraiter les lectures et les alignements de faible qualité. À partir de là, le logiciel applique un modèle de mélange gaussien adapté aux longues lectures haute-fidélité CCS pour déterminer le génotype le plus probable en termes de nombre de copies pour chaque locus. En utilisant STRkit, je montre que les lectures

CCS sont plus performantes que les longues lectures de type Nanopore et les lectures courtes en paires pour le génotypage STR sur un génome entier de référence. Je démontre également que STRkit surpasse d'autres logiciels de génotypage STR à lecture longue lorsqu'il reçoit des lectures longues CCS ou de type Nanopore et peut détecter des expansions STR pathogéniques.

Dans l'ensemble, je montre que les lectures longues à haute-fidélité sont utiles pour résoudre la variation du nombre de copies de STR de différentes magnitudes, surtout lorsqu'on utilise un logiciel de génotypage STR approprié. Pour le génotypage STR, chaque lecture longue donne un nombre de copies, ce qui pourrait à l'avenir aider à analyser le motif STR ou le mosaïcisme du nombre de copies et faciliter le démêlage des génotypes d'allèles STR longs non-identiques.

Table of Contents

List of Abbreviations

ALS	Amyotrophic lateral sclerosis
bp	Base pairs (unit of measure)
CANVAS	Cerebellar ataxia with neuropathy and bilateral vestibular areflexia syndrome
CCS	Circular consensus sequencing
CE	Capillary electrophoresis
`EHDn`	`ExpansionHunter Denovo` (STR expansion detection software)
eSTR	Expression(-associated) Short Tandem Repeat
eQTL	Expression Quantitative Trait Locus
FMR1	The Fragile X messenger ribonucleoprotein 1 gene, encoding FMRP.
FMRP	Fragile X messenger ribonucleoprotein, encoded by FMR1.
FXPOI	Fragile X-associated primary ovarian insufficiency
FXS	Fragile X syndrome
FXTAS	Fragile X-associated tremor/ataxia syndrome.
GeM-HD	Genetic Modifiers of Huntington's Disease Consortium
GIAB	The Genome-in-a-Bottle consortium.
GMM	Gaussian mixture model
GRC	Genome Reference Consortium
GRCh38	Genome Reference Consortium Human Build 38
HD	Huntington's disease
HiFi	Pacific Biosciences' HiFi read technology.

HTT	The gene encoding the Huntingtin protein.
HMM	Hidden Markov model
IGSR	International Genome Sample Resource
Indel	Insertion/deletion
IUPAC	International Union of Pure and Applied Chemistry
ME	Mendelian inheritance error
ONT	Oxford Nanopore Technologies
ONT-UL	Oxford Nanopore Technologies ultra-long reads.
PCR	Polymerase chain reaction
SCA	Spinocerebellar ataxia
SNP	Single-nucleotide polymorphism
SSR	Simple sequence repeat
STR	Short tandem repeat
T2T	Telomere-to-Telomere
TR	Tandem repeat
UCSC	University of California, Santa Cruz
UTR	Untranslated region (of a gene)
VNTR	Variable-number tandem repeat

List of Figures

List of Tables

Acknowledgements

I would like to thank my supervisor, Prof. Guillaume Bourque, for his mentorship and support over the past few years, as well as my supervisory committee members Prof. Ziv Gan-Or and Prof. Ioannis Ragoussis for their advice and guidance.

I would also like to thank Mathieu Bourgey and Rob Eveleigh for their helpful suggestions in our regular “long read” meetings, while they lasted, and Prof. Tomi Pastinen for his insights into long read sequencing technologies.

Finally, I would like to thank my family for their love, support, and proof-reading, and my friends for making the pandemic bearable.

Format of the Thesis

This document is structured as a traditional thesis.

Contribution of Authors

Chapter 1: I wrote this chapter and did the primary literature review. My supervisor, Prof. Guillaume Bourque, was of great help with finding relevant literature especially at the beginning of my project, and of helping me keep on top of the frequent updates in the field of short tandem repeat genetics.

Chapter 2: The concept for this project came from discussions with my supervisor. I wrote the entirety of the STR genotyping algorithm presented here; however, it is built using free and open-source libraries developed by innumerable contributors. For the benchmarking performed in this project, I wrote workflows which used a great deal of public data in this project. For this, I acknowledge the substantial contributions to public genomic data availability by the Genome-in-a-Bottle Consortium (GIAB), the Human Genome Structural Variant Consortium, and Pacific Biosciences.

Chapter 3: All results, figures, and tables presented here are my own, created from the synthesis of the STR genotyping approach I describe, and benchmarking datasets derived from public data.

Chapter 4: All discussion and figures are my own work.

Chapter 5: All conclusions and ideas for future directions are my own work.

Chapter 1 Background

1.1 Short tandem repeats: definition and structure

Short tandem repeats (STRs), alternatively known as microsatellites or short sequence repeats (SSRs), are repetitive DNA elements which make up around 3% of the human genome (Lander *et al.*, 2001), although ancestral STR-derived sequences may comprise as much as ~6.8% of the genome (Shortt *et al.*, 2020). Structurally, they consist of a motif of anywhere from 1 or 2 to between 6 and 13 base pairs (bp) long, repeated multiple times. The number of times that a motif is repeated is known as the "copy number". Definitions vary slightly; Shortt *et al.* (2020) include motifs from 1-6 bp long in their definition, whereas Chiu *et al.* (2021) exclude homopolymers (i.e., 1 bp motifs); Lander *et al.* (2001) expand the definition to include motifs up to 13 bp long.

STRs are of interest due to their unique mutation mechanisms, patterns, high rates of mutation and polymorphism at some loci (Ellegren 2004; Weber and Wong 1993), and because of their association or causal relationship with complex and Mendelian traits and disorders (Gall-Duncan *et al.*, 2021), as well as DNA methylation and expression (Pretto *et al.*, 2014; Gymrek *et al.*, 2016). STRs have also been used in population genetics to examine population structure (Kang *et al.*, 2010).

1.2 STR diversity: polymorphism, mutation, linkage, and imputation

STRs on average mutate at a much higher rate than the rest of the genome and can be highly polymorphic; these polymorphisms occur most frequently in the form of changes in copy number (Ellegren, 2004), meaning many different length polymorphisms can be found for some STRs in the human population. Most STR copy number mutations are small, with an overall bias towards expansion (Ellegren, 2000; Mitra *et al.*, 2021), which suggests a stepwise model for copy number mutation. Two main mechanisms of STR copy number mutation have been proposed: recombination and slippage (Gemayel *et al.*, 2010), although the latter is favoured (Ellegren, 2004). Slippage occurs when, during DNA replication, one of either the template or elongating strand dissociates from the other strand and mis-pairs in a way which either adds or removes whole motif copies (Gemayel *et al.*, 2010); usually, only one or two copies are added or subtracted (Ellegren, 2000). STR mutation rates are heterogeneous; some loci mutate at a much higher rate (Ellegren *et al.* 2000, Gemayel *et al.*, 2010). These rates correlate positively with allele size; long STR alleles tend to be more unstable (Ellegren, 2004).

Due to their elevated mutation rate and tendency towards multi-allelic polymorphism, STRs can be useful for studying population history and structure, especially in closely-related populations (Kang *et al.*, 2010). Assessing STR mutation rate is important for this: calculating divergence time using a particular set of loci with population history models requires specifying the mutation rates of the loci as a parameter (Goldstein *et al.*, 1994; Zhivotovsky *et al.*, 2004).

STR polymorphism also has implications for genome-wide association studies (GWAS) which have been used to associate regions of the genome with complex phenotypes that are otherwise difficult to ascertain. GWAS typically use single nucleotide polymorphisms (SNPs) to find associations between genomic regions and a phenotype, relying on linkage disequilibrium with surrounding DNA to capture more variation (Uffelmann *et al.*, 2021). Due to the tendency for STRs to have a high mutation rate, only a subset of STRs can be imputed using nearby SNPs, and SNP-STR imputation power is reduced with multi-allelic (versus bi-allelic) STR loci (Saini *et al.*, 2018). Genotyping STRs directly, then, could capture a different set of linked genomic variation, and could be incorporated into GWAS to produce more powerful association tests. Hannan (2018) suggests that, indeed, STR variation itself may be implicated in the "missing heritability" problem, explaining some portion of complex trait heritability that we may have not yet elucidated.

1.3 STR-associated disease and phenotype

Many STR-associated disorders in humans are associated specifically with repeat expansion alleles in key loci. Expansions are STR alleles with a copy number significantly above what is observable in most or all wild-type alleles in the population distribution and can be in coding or untranslated regions of genes, or in intergenic regions; as a result, there are various mechanisms through which expansions can cause disease (Gall-Duncan *et al.*, 2021). Examples of expansion-associated diseases

include Huntington's disease (HD), caused by an expansion in a coding portion of the HTT gene; fragile X syndrome (FXS) and other fragile X-related clinical phenotypes, caused by variable-length expansions in the 5' untranslated region (UTR) of the FMR1 gene (Allen *et al.*, 2021), and various forms of spinocerebellar ataxia (SCA; Brouwer *et al.*, 2009). Disorders such as HD, which are typically hereditary, are also often characterized by a phenomenon called "anticipation", where successive generations have progressively lengthened repeat tracts and often either switch from a non-disordered "pre-mutation" state to a disordered "mutation" state or experience worse symptoms as continued expansion occurs within the pathogenic repeat count range (Gall-Duncan *et al*., 2021). Copy number and age of onset are negatively correlated in many expansion disorders, including HD (Igarashi *et al.*, 1992; Duyao *et al.*, 1993; Matsuura *et al.*, 2000; Blauw *et al.*, 2012; Bragg *et al.*, 2017). Repeat count genotyping in these key disease-associated loci is therefore critical to assessing risk of pre-mutation to full-mutation expansion in progeny and understanding disease genotype-phenotype relationships.

Fragile X-associated primary ovarian insufficiency (FXPOI) is an expansion disorder which has a more complex non-linear relationship between repeat expansion size and phenotype, and further highlights how studying STR copy number, and not just binary expansion status, is important. This disorder can occur when an allele of an STR with a CCG motif in the 5' UTR of the FMR1 gene (GRCh38 coordinates: `chrX:147912037-147912111`) has between 55 and 199 copies; however, peak risk occurs at an intermediate copy number of 85-89 (Allen *et al.*, 2021). FMR1 expansions in this same

copy number range also can cause fragile X-associated tremor/ataxia syndrome (FXTAS; Leehey, 2009). With a copy number of ≥200, this same repeat expansion causes a third disease, Fragile X syndrome (FXS), and is linked to nearby DNA methylation (Sutcliffe *et al.*, 1992; Alisch *et al*., 2013).

Classical examples of expansion disorders such as HD or FXS result from expanded alleles in these specific critical loci, but phenotypes can be associated with the presence of expansions in a more complex fashion. Trost *et al.* (2020) found that rare tandem repeat expansions are more common in autistic children, suggesting a contribution of STRs to this complex phenotype. These expansions are not the only type of STR variation associated with autism; Mitra *et al.* (2021) found a link between autism and an increased quantity of small *de novo* STR mutation. Variation in tandem repeats has also been associated with DNA transcription through multiple different possible mechanisms (Hannan, 2018). Gymrek *et al.* (2016) found thousands of STRs which function as expression quantitative trait loci (eQTL), i.e., their copy number correlates with the expression of a particular transcript; they called these loci “eSTRs”.

1.4 Beyond simple sequence repeats: interruption, instability, and mosaicism

An STR is defined by the presence of a repeating short motif, but there are a range of other factors which make precise definitions of specific loci more complicated. It may be hard to define the exact boundaries of STRs within the genome, due to imperfect motif copies, and there may be interruptions or multiple motifs in the same locus. These pattern disruptions can modulate an expansion disorder's phenotype (Ishiura *et al.* 2018, Corbett *et al.* 2019, Rafehi *et al.* 2019; Gall-Duncan *et al.*, 2021).

In HD, the associated STR region in the HTT gene is usually found in a form described by the pattern $(CAG)_nCAACAG$ (Genetic Modifiers of Huntington's Disease [GeM-HD] Consortium, 2019; GRCh38 coordinates: `chr4:3074877-3074940`). Despite this trinucleotide expansion occurring in a coding region, where both CAA and CAG code for glutamine, the presence of one or two CAA interruptions at the end of the STR ($[CAACAG]_{\{1\text{-}2\}}$) is correlated with a delayed age of onset (GeM-HD, 2019). Thus, the number of uninterrupted CAG repeats is more powerfully associated with disease phenotype than number of glutamines the STR tract encodes; this strange association may give insights into the mechanism of HD.

Age of onset in HD is associated with HTT expansion length, but this factor does not explain all variation. Swami *et al.* (2009) found that somatic instability in the expansion also contributes to age of onset – detecting this instability and representing the expansion requires capturing more than just a single copy number for the expanded

allele. In the FMR1 gene, researchers have also observed copy number instability. Pretto *et al.* (2014) found that mosaicism around the pre-mutation/full-mutation boundary of 200 repeats correlated with an increased expression of the gene product FMRP and a reduction in phenotype severity. They also discovered a positive correlation between methylation mosaicism and clinical outcome. Seixas *et al.* (2017) found a non-reference ATTTC STR insertion which segregates with SCA, while STR alleles of a similar length with only the ATTTT motif present were non-pathogenic.

1.5 Genotyping STRs, past and present

1.5.1 Traditional STR genotyping approaches: usage and limitations

PCR with capillary electrophoresis is a common and effective way of genotyping STRs of interest, with Willems *et al.* (2017) describing it as the "gold standard" for STR genotyping. This, combined with certain highly polymorphic marker STRs, or "STRPs" (STR polymorphisms), provides a powerful tool for linkage-mapping genes of interest (Weber and Wong, 1993; Ellegren, 2004). However, it requires unique primer pairs in flanking regions for targeted loci and is low throughput – whole-genome surveying is effectively impossible, although multi-locus panels are available (Kedzierska *et al.*, 2018). Willems *et al.* (2017) found that between technical replicates, PCR genotypes were internally consistent around 98.5% of the time; while this is therefore not a perfect tool, the method sets a target for accuracy and repeatability when using DNA sequencing technologies for STR genotyping. For their study on mosaicism of an STR

expansion in the FMR1 gene, Pretto *et al.* (2014) used both PCR with electrophoresis and DNA Southern blot analysis.

1.5.2 Genotyping STRs with genomic sequencing technologies

In the past decade, the newfound affordability of short-read whole-genome sequencing (WGS), together with algorithms for repeat genotyping, has opened new avenues in STR analysis, genome-wide profiling, and association testing. `LobSTR` (Gymrek *et al.*, 2012) was an early tool in short-read STR genotyping and enabled high-throughput genotyping of 10s of 1000s of loci at a time, so long as the alleles in question were smaller than the length of a short read. This tool was later used to predict causal associations between the length of some specific STRs and gene expression (Gymrek *et al.*, 2016). More recent work based on genome-wide STR surveying include the discovery of increases in small *de novo* STR mutation discussed in section 1.3 (Mitra *et al.*, 2021), which used a framework built on the `GangSTR` genotyping method (Mousavi *et al.*, 2019), and an association found by Tazelaar *et al.* (2020) between expansions in the ATXN1 gene and amyotrophic lateral sclerosis (ALS).

The three STR genotyping methods mentioned above rely on catalogues of positions of STRs in the genome. These catalogues can be incomplete, which I discuss further in section 1.6.1. To address this limitation, Dolzhenko *et al.* (2020) introduced the `ExpansionHunter Denovo` (`EHDn`) software, which works with paired short-read sequencing data; Rafehi *et al.* (2019) used this tool to find an expanded repeat in the

RFC1 gene associated with CANVAS, and Trost *et al.* (2020) used it to discover an increase in rare expansions in autistic individuals. `EHDn` cannot assess copy number; instead, it reports significant increase in intra-repeat reads, i.e., read pairs in which one read entirely consists of repetitive DNA from an expanded STR allele. This mechanism, while clever, highlights a drawback in short read sequencing: despite high read accuracy and throughput, read size prevents precise STR copy number genotyping.

Long read technologies such as Oxford Nanopore (ONT) ultra-long reads (ONT-UL; Jain *et al.*, 2018) and Pacific Biosciences (PacBio)'s SMRT sequencing provide some advantages in sequencing longer STRs; the read lengths of these technologies allow for sequencing the entire length of these long repetitive regions. For example, Chaisson *et al.* (2015) used SMRT sequencing to resolve gaps enriched in STRs in the GRCh37 reference genome. Multiple groups of researchers have created STR genotyping software for long reads; these include `PacmonsTR` (Ummat & Bashir, 2014), `RepeatHMM` (Liu *et al.*, 2017 & 2020), `Tandem-genotypes` (Mitsuhashi *et al.*, 2019), and `Straglr` (Chiu *et al.*, 2021). Despite ONT-UL and SMRT sequencing technologies' ability to span entire STR alleles with a single read, they suffer from high error rates (Wenger *et al.*, 2019), which likely affect the genotyping of STR repeat sequences. More recently, PacBio's high fidelity circular consensus sequencing (CCS) long-read technology has shown error rates more comparable to short read sequencing with small insertions and deletions such as those found in STR loci (Wenger *et al.*, 2019) while maintaining a longer read size. This technology may allow for better resolving of STR

variation in longer STR alleles while maintaining the ability, like short read sequencing, to genotype small changes in copy number.

Despite excellent performance from CCS when assessing different types of DNA variation, the technology is not without limitations. Currently, CCS is expensive, and because a single read is the product of several combined 'subreads', storing the original subread data is space-intensive. CCS reads also have challenges accurately capturing homopolymer sequences, which are long continuous sequences of the same base pair (Wenger *et al.*, 2019); this may affect genotyping of STRs with motifs that contain a homopolymer, e.g., $(AAAT)_n$. Liu *et al.* (2017) found that CCS reads reduced performance versus SMRT sequencing when STR genotyping using their tool, `RepeatHMM`, albeit with an older version of the CCS protocol.

1.5.3 Limitations with existing sequencing-based STR genotyping methods

All STR genotyping methods which use sequencing data have limitations. Some of these come from the sequencing method used; for example, as I have already discussed, there are types of repetitive DNA patterns short reads cannot easily resolve. Other limitations stem from implementation; for example, models which aim to reduce the impact of error from noisy long reads may do so at the cost of computational performance and, potentially, sensitivity. In Table 1, I show a list of long read STR genotyping software and some of their known limitations. Most of these methods, as well as all short read STR genotyping approaches which are not specific to expansions

and which resolve copy number, rely on user-provided STR genotype catalogues referring to coordinates in a particular reference genome. I discuss the pitfalls of relying on catalogues in section 1.6.1.

Long read STR genotyper	Sequencing method†	Limitations
PacmonsTR (Ummat & Bashir, 2014)	PacBio non-CCS	Relies on deprecated aligner (BLASR), which no longer appears on PacBio's GitHub page
`RepeatHMM` (Liu *et al.*, 2019 & 2020)	PacBio non-CCS & ONT with configurable hidden Markov model	Computationally slower and may miss more calls than other approaches (Chiu *et al.*, 2021); limited support for heterogeneous repeat expansions; written in deprecated Python version (v2.7); limited to what is in catalogue
`Tandem-genotypes` (Mitsuhashi *et al.* 2019)	Any	Requires realignment using LAST, which is a slow step (Chiu *et al.* 2021); does not check motif composition of insertions; uses crude k-means prediction for genotype calls; limited to what is in catalogue
`Straglr` (Chiu *et al.* 2021)	Any	Uses `TRF` to determine read copy number, which may fail with disruptions or short alleles; limited support for heterogeneous repeat expansions
Miscellaneous ONT-specific callers (Giesselmann *et al.*, 2019; Fang *et al.*, 2022)	ONT	Use raw ONT signal data – not portable across sequencing technologies.
†: CCS = circular consensus sequencing; ONT = Oxford Nanopore Technologies		

Table 1: Long read STR genotyping methods and their limitations.

1.6 Evaluating sequencing-based STR genotyping methods

To evaluate STR genotyping approaches in a systematic way, individuals with 'ground truth' STR genotypes at known genomic coordinates are required to compare software output against; a list of these coordinates mapping to sequences in a reference genome serve as an "STR catalogue" specifying regions for software to genotype.

1.6.1 Reference genomes and genome-wide STR locus cataloguing

The UCSC genome browser (Kent *et al.*, 2002) includes tracks of STRs ('simple repeats') generated with Tandem Repeats Finder (`TRF`; Benson, 1999) by running the program on a reference genome. The cataloguing of TRs using a reference genome is naturally limited by what sequences are contained in the reference; several parts of the latest Genome Reference Consortium reference genome, GRCh38, including many repetitive regions, are not fully resolved (Nurk *et al.* 2022). Another limitation of catalogue-based approaches comes from treating STRs as loci with a single motif; as I have discussed in section 1.4, some diseases are associated with the presence or expansion of a non-reference STR motif, in addition to or replacing the motif occurring in the reference genome.

Improvements in cataloguing STRs genome-wide will likely come with the new CHM13 telomere-to-telomere reference genome, which includes an enormous 112.9% increase in simple sequence repeat DNA (Nurk *et al.* 2022). The creation of STR databases and

curation approaches such as Illumina's `STR-finder` (Dolzhenko *et al.* 2019), which specifically searches for polymorphic STRs, can also help more comprehensively record STRs of interest in the human population.

1.6.2 The Genome-in-a-Bottle Consortium small variant benchmark

The Genome in a Bottle consortium (GIAB) has published high-depth sequencing data for multiple individuals, including two child/mother/father trios: one Ashkenazim (HG002-004) and one Han Chinese (HG005-HG007) (Zook *et al.*, 2016). To better resolve trio genomic variation, these data include read-sets from a range of sequencing technologies at high depths of coverage, which have led to the production of benchmark variant callsets for the trios (Zook *et al.*, 2019; Wagner *et al.* A, 2022; Wagner *et al.* B, 2022). These benchmarks consist of a subset of variation in these individuals validated using multiple sequencing technologies and variant calling software. Among these benchmark datasets are a set of small SNVs and insertion/deletion (indel) variants (Zook *et al.*, 2019; Wagner *et al.* A, 2022); some of these indels are copy number changes in STRs, although the tooling the consortium used to resolve indels in tandem repeats does not perfectly resolve all copy number variation (Wenger *et al.*, 2019). The variant sets also include some instances of probable *de novo* STR mutation.

1.6.3 Using Mendelian inheritance patterns to evaluate STR genotyping

Trio datasets such as those from GIAB provide another means through which genotype call quality can be assessed, using the pattern of Mendelian inheritance. The proportion of loci that are consistent with Mendelian inheritance in a callset, combined with the knowledge of instances of true *de novo* mutation from a benchmark "ground truth" variant set, allow estimates of reliability of genotyping software – given a set of alleles one knows are transmitted from parent to child, one can assess how frequently an STR genotyping method generates the same copy number allele for the child and parent. Mendelian inheritance tracing has been used by Gymrek *et al.* (2012) and Mousavi *et al.* (2019) to assess the performance of their respective STR genotyping software, and by Niehus *et al.* (2021) to do the same for their genomic deletion-identifying software.

The inverse of the rate of Mendelian inheritance can be called 'Mendelian inheritance error', or ME, which STR genotyping methods should aim to get as close to 0 as possible with the sequencing data given (Figure 1). This assessment method alleviates some of the "burden of truth" from the variant truth set itself, since it is only required for knowing which loci are truly Mendelian-inconsistent in the trio, rather than precise genotypes; one expects that few *de novo* mutations are present in the benchmark set.

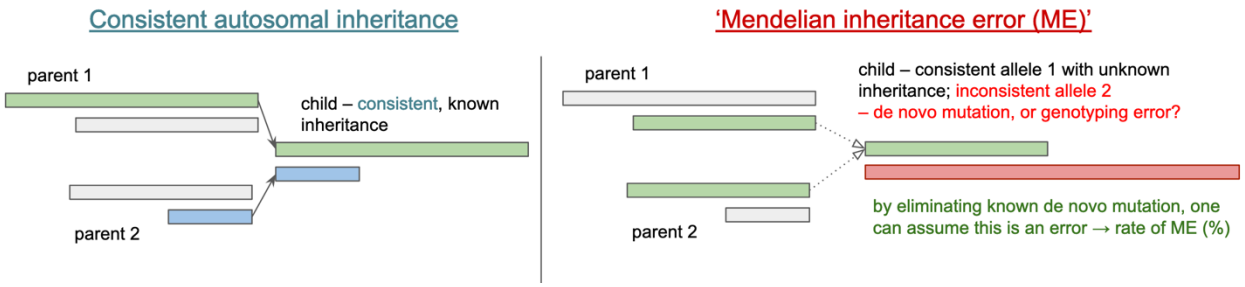

Figure 1: Assessing the rate of violation of Mendelian inheritance as a metric for genotyping quality for a set of calls for autosomal STR loci.

1.6.4 A public dataset of PCR-derived STR genotypes

Payseur, Place, and Weber (2008) published a dataset of 721 polymorphic STR (STRP) genotypes, in the form of PCR fragment sizes, for multiple HapMap project individuals. These loci were taken from the 5 centimorgan Marshfield STRP panel (Ghebranious *et al.*, 2003). Of the individuals genotyped, two have been sequenced using both high-coverage Illumina short read technology (Byrska-Bishop *et al.*, 2022), and PacBio CCS at a roughly 5-7X coverage (Ebert *et al.*, 2021). Using a technique developed by Pemberton *et al.* (2009), these PCR product sizes can be back-converted to copy number changes relative to a reference genome and can serve as an *in-vitro*-sourced benchmark callset for STR genotyping. This approach was used by Saini *et al.* (2018) to evaluate STR genotype imputation using single-nucleotide polymorphisms (SNPs).

1.6.5 Targeted CCS sequencing and a public pathogenic expansion dataset

I have discussed possible benchmark datasets for STR genotyping; these sets are likely comprised of mostly wild-type polymorphism and small *de novo* mutations. However, performance on small variants does not necessarily correspond to performance in disease-causing expansions. Pacific Biosciences (PacBio) has published a targeted sequencing dataset with eight samples: four with expansions in the HTT gene, three with expansions in the FMR1 gene, and one control sample (accessed from https://downloads.pacbcloud.com/public/dataset/RepeatExpansionDisorders_NoAmp/ July 1, 2022). These individuals' HTT and FMR1 genes are sequenced with PacBio "HiFi No-Amp" targeted sequencing technology to a high depth of coverage.

1.7 Objectives and hypothesis

In this project, I have two major aims. The first is to evaluate STR genotyping approaches for both short- and long-read technologies and compare their strengths, with a focus on examining the potential of CCS data for genome-wide STR genotyping. To evaluate these approaches, I will use the ground-truth datasets I have discussed to benchmark genotyping performance. The second aim is to develop a new STR genotyping method, designed specifically to take advantage of CCS' high sequencing accuracy, and compare it against other genotypers to assess its performance using datasets with ground-truth STR genotypes available. By optimizing STR genotyping for high quality reads, I expect that some of the error compensation approaches used in

other methods can be eliminated, resulting in reduced computation time and improved sensitivity to small copy number changes. In the above sections, I have emphasized how ascertaining an STR's copy number can be important to understanding its causative effects or associations. My goal is to address some drawbacks in other long read STR genotyping approaches which hinder this, while maintaining the advantages long reads offer for resolving long, complex alleles. Given the realized potential of CCS in other areas like indel and structural variant detection (Wenger *et al.*, 2019) and existing work by Liu *et al.* (2017) and Chiu *et al.* (2021), CCS reads should scale to whole-genome STR profiling. I do not aim to perform a comprehensive benchmarking of all STR genotyping and expansion detection approaches, nor solve all issues discussed in section 1.5.3, e.g., addressing every drawback of STR catalogues and existing reference genomes.

Chapter 2 Materials and Methods

2.1 Reference genome and tandem repeat catalog

I used the GRCh38 reference genome, and its corresponding catalogue of short tandem repeats as generated by the Tandem Repeats Finder program (Benson, 1999) available from the UCSC genome browser (Kent *et al.* 2002), as discussed in section 1.6.1.

2.2 Benchmarking and validation datasets

I used multiple different datasets and DNA sequencing technologies to build an overall picture of genotyping method performance across a range of STR sizes (Table 2).

Dataset	STRs captured	Purpose	Methods section	Data citations
PCR-derived STR genotype set	Polymorphic 15-115bp alleles; 3 & 4 nt. motifs	Comparison benchmark of wet lab-validated genotypes	2.2.1	Payseur *et al.* (2008); Pemberton *et al. (2009);* Fairley *et al.* (2020)
Ashkenazim trio STR genotypes from small variant benchmark set (GIAB)	Whole-genome STR set; non-pathogenic alleles (0-300bp)	Scaled-up comparison: Test whole-genome genotype capability and throughput at various depths of coverage.	2.2.2	Wagner *et al.* (2022A)
PacBio HiFi targeted pathogenic expansion sequencing	Pathogenic expanded alleles in HTT and FMR1	Validation of STRkit's performance with disease alleles of interest	2.2.6	Accessed from pacbcloud.com

Table 2: Benchmarking datasets used in the thesis.

2.2.1 Generating a PCR-genotyped small STR truth set

I accessed PCR and capillary electrophoresis-derived STR genotype data published by Payseur, Place, and Weber (2008), described in section 1.6.4. I downloaded 30X coverage Illumina and ~5x coverage PacBio CCS sequencing data from the International Genome Sample Resource (IGSR) portal (Fairley *et al.*, 2020) for the NA19238 individual from the Yoruba in Ibadan (YRI) population. Pemberton *et al.* (2009) published a dataset of PCR primer sequences for Marshfield STRPs. They used these sequences to convert PCR fragment sizes to repeat counts through UCSC's In-Silico PCR tool (Kent *et al.*, 2002) on the HG17 reference genome. Overlap between the STR genotype and primer sequence datasets allowed for the re-alignment of PCR primers for 155 STRPs onto the GRCh38 reference genome with in-silico PCR, giving expected PCR fragment sizes for this more up-to-date reference genome. Using these results and the formula published by Pemberton *et al.* (2009), I calculated relative copy number genotypes at these 155 loci in NA19238 to serve as a truth set.

To reduce bias from non-repeat indels within the sequences flanked by the PCR primers, I used variant callsets generated using 30X coverage Illumina short read sequencing data and accessed from the IGSR portal to filter out loci for which a sample had at least one non-repeat indel within the bounds of the PCR primers.

2.2.2 Creating a genome-wide STR benchmark call set

To benchmark genotyping methods and compare short and long high-accuracy reads, I used version 4.2.1 of the small variant benchmark set for the Ashkenazim trio (Wagner *et al.*, 2022 A) to create a whole-genome STR copy number callset. The resulting genotypes are expressed as a change in copy number relative to the GRCh38 reference genome. To create this truth set, I identified all variants in the GIAB benchmark set which overlapped a tandem repeat in the Tandem Repeats Finder (`TRF`) annotation track from the UCSC Genome Browser (Kent *et al.* 2002) for GRCh38. I then applied the following filtering and transformation steps to get a final catalogue of benchmark-quality short tandem repeat genotypes for the Ashkenazim trio:

1. Sex chromosomes were removed to focus on only diploid loci and avoid having to incorporate karyotypic sex metadata.
2. STRs in centromeres or segmental duplications, as defined by tracks in the UCSC genome browser, were removed – reducing low-confidence sequencing regions.
3. STRs with `TRF` quality scores below 90 (out of a possible 100) were removed, to keep only a set of confidently-identified tandem repeats.
4. Homopolymers were removed.
5. STRs where `TRF` reported differences between motif size and repeat period were removed, to eliminate more complex repeat patterns.
6. Variants in the short variant benchmark set which did not overlap an STR in the `TRF` catalogue were discarded.

7. Variants in the benchmark set which were not indel variants or did not contain at least an entire insertion or removal of the catalogued STR motif were removed.
8. Variants in the benchmark set whose indel variants consisted of less than 80% copies of the catalogued STR motif (by base content) were removed.
9. Indel sizes were converted to a copy number change relative to reference using the size of the motif as defined by `TRF`. Indels only contributed to copy number change if they mapped to within the boundaries of the STR as listed by `TRF`.
10. Mendelian inheritance consistency was assessed for the trio.

2.2.3 Calculating the root-mean-squared error (RMSE) of genotyping output

I calculated root-mean-squared error (RMSE) for overall genotyping performance of a given genotyper, as well as RMSE binned by allele size, using Equation 1.

$$RMSE_L = \sqrt{\sum_{l \in L} (c_{l1} - t_{l1})^2 + (c_{l2} - t_{l2})^2}$$

Equation 1: RMSE equation, where $\langle c_{l1}, c_{l2} \rangle$ are the sorted called copy numbers (relative to reference) for a bi-allelic locus (relative to reference, and $\langle t_{l1}, t_{l2} \rangle$ are the sorted true relative copy numbers, given by the truth set, for each locus l in called locus set L.

2.2.4 Calculating the binary accuracy of genotyping output

I calculated binary accuracy for overall genotyping performance, as well as accuracy binned by allele size, using Equation 2.

$$Acc_L = \frac{\sum_{l \in L} \sum_{i=1}^{2} \begin{cases} 1, & c_{li} - t_{li} < 0.5 \\ 0, & c_{li} - t_{li} \geq 0.5 \end{cases}}{|L|}$$

Equation 2: Binary accuracy equation, where $\langle c_{l1}, c_{l2} \rangle$ are the sorted called copy numbers (relative to reference) for a bi-allelic locus (relative to reference, and $\langle t_{l1}, t_{l2} \rangle$ are the sorted true relative copy numbers for each locus l in a called locus set L.

2.2.5 Generating subsampled alignments for GIAB trio individuals

To generate different depths of genomic coverage of alignment files from the Ashkenazim trio, I used the following steps:

1. I calculated the average depth of coverage for each alignment using the `samtools depth -a` command.
2. I determined the fraction of reads needed for each desired genomic coverage level in the sub-sampling analyses (4, 6, 8, 10, 15, 20, 25, 30, 40, 50, 60x, the latter two only if available).
3. I used these fractions as parameters for the `samtools view --subsample` command to generate each sub-sampled alignment file.

The following aligners were used for each sequencing technology by the original creators of the data:

- Illumina 2x150bp and 2x250bp short read technologies: NovoAlign v1.15.1
- ONT-UL: `minimap2` v2.17

  except for use with the `Tandem-genotypes` method: `LAST` version 1418
- PacBio CCS: `pbmm2` versions 1.1.0 and 1.2.0

  except for use with the `Tandem-genotypes` method: `LAST` version 1418

2.2.6 A PacBio HiFi Targeted Expansion Sequencing Dataset

I accessed a dataset of Pacific Biosciences targeted sequencing of expansions from a repository in their GitHub organization: https://github.com/PacificBiosciences/apps-scripts. The example dataset itself can be found at https://downloads.pacbcloud.com/public/dataset/RepeatExpansionDisorders_NoAmp/

2.3 Comparisons to existing STR genotyping software

2.3.1 Inclusion criteria for existing methods

To select STR genotyping methods to benchmark CCS reads and my approach against, I followed these criteria of inclusion:

- The method must support an alignment format which can be generated with readily-accessible software.
- The method must be performant enough to genotype the entirety of our chosen locus catalog in a 'tractable' amount of time (in this case, <7 days).
- The method must be able to genotype 'non-expanded' loci – it cannot just be an expansion detection tool; it should be able to resolve more subtle variation.
- It must support genotyping arbitrary loci provided via a catalog file, rather than requiring pre-made models.
- It must support all motif sizes, not just trinucleotide STRs.

2.3.2 Parameter settings for comparison

For the comparison of STR genotypers, the tools were run with the following parameters (only listed if they were non-standard, i.e., not specifying reference genome/alignment file/etc.):

- `ExpansionHunter`: N/A
- `GangSTR`: N/A (default bootstrap iterations: 100)
- `RepeatHMM`: `--SplitAndReAlign 0`; `--SeqTech Pacbio` or `Nanopore`; `--MinSup 4`; `--RepeatTime 3`
- `Straglr`: `--min_cluster_size 2`; `--min_support 4`; `--max_num_clusters 2`

- `STRkit`: `--min-reads 4`; `--min-allele-reads 2`; `--num-bootstrap 100`
  - If PacBio CCS: `--realign` and `-hq`
- `Tandem-genotypes`: `--far=70`; `--output=2`

2.3.3 Generating LAST alignments for use with the `Tandem-genotypes` tool

The program `Tandem-genotypes` (Mitsuhashi *et al.*, 2019) requires as input read alignments generated by `LAST` (Kiełbasa *et al.*, 2011). This aligner has many parameters which affect alignment time and quality. To generate read sets to use re-align with `LAST`, I used Picard (Broad Institute, 2019) to convert BAM files into FASTQ files. For use with `Tandem-genotypes`, I followed recommendations from the tool's authors, and ran LAST version 1418 with the following parameters:

- `last-train`: `-Q0`
- `lastal`: default parameters

2.3.4 Benchmarking hardware and language versions

I ran all STR genotypers written in Python with Python v3.8. Methods written in C++ were compiled with GCC 9.3.0. I ran genotyping software on Calcul Québec's Béluga cluster using 1 core of an Intel Gold 6148 processor and as much memory as required.

2.4 Creating an STR calling toolkit

To address limitations described in section 1.5.3 and make my genotyping approach and downstream analysis code available to other researchers, I created an STR genotyping and analysis software package, `STRkit`, written in the Python programming language and released as free and open-source software accessible at https://github.com/davidlougheed/strkit/ or in the PyPI Python package repository as '`strkit`'.

2.4.1 Genotyping approach

I outline the general genotyping approach used in `STRkit` below. An implementation of the approach is included in the `STRkit` package as a sub-command; '`strkit call`'.

1. An alignment software such as `minimap2` (Li, 2018) or `pbmm2` (a wrapper for `minimap2` implemented by Pacific Biosciences for use with their sequencing technologies) is used to generate an aligned BAM file.
2. For each STR in a provided catalogue, a corresponding DNA sequence from a user-provided reference genome is extracted, including flanking DNA, and the reference copy number is determined using a variation on the method described in step 6. Instead of aligning with both the 5' and 3' flanking sequence included at the same time, the 5' flank is included and the 3' boundary is allowed to expand; then, this process is repeated with the 3' flank and the 5' boundary. In this way,

the STR region is permitted to expand in either direction in the event of slight misalignment or a disagreement between the catalogue and `STRkit`'s method.

3. For each STR in the catalogue, all reads which overlap the STR region (with flanking sequences on either side of a user-configurable length, 70bp by default) are extracted using the `PySAM` library (version 0.19.1; https://github.com/pysam-developers/pysam). During this process, some filtering is applied:
    a. Supplementary reads are skipped; a flag is set if a primary and supplementary alignment for the read both map to the STR region (i.e., the read has a chimeric alignment, which may occur with large expansions.)
    b. Reads with no successful alignment are skipped.
4. If a read is soft-clipped in the STR region and the user has provided the `--realign` command-line flag, a local realignment to reference is attempted, using the `parasail` library's semi-global alignment algorithm (Daily, 2016).
5. If, at this point, a read cannot not fully *encompass* the STR tract and flanking region, it is discarded.
6. For each encompassing read, a series of candidate STR tracts of varying length, starting with a tract closest to the reference STR size ± any insertions or deletions in the alignment, are generated from the motif provided in the catalogue. Each of these candidate STR tracts are realigned using `parasail`, incorporating the 5' and 3' flanking sequences.
7. The best-scoring copy number and alignment is kept for the read in question, resulting in a read → copy number map once every encompassing read's STR tract copy number has been counted. If a user chooses to do so, motif-sized k-

mers are tabulated for each read, which can later be amalgamated into peak-level k-mer counts.

8. For each read, a re-sampling weight is generated, corresponding to the estimated inverse probability of observing a read encompassing an STR tract of the same size (see section 2.4.2 for an explanation of what this accomplishes).
9. Read copy numbers are resampled many times to generate bootstrap re-samplings, according to the re-sampling weight calculated in the above step.
10. A Gaussian Mixture model (GMM) the *scikit-learn* library (Pedregosa *et al.*, 2011), initialized via the K-means++ algorithm (Arthur and Vassilvitskii, 2007), is trained on each re-sampling of read copy numbers to derive estimates for the short and long allele copy numbers.
11. Final genotype estimates and 95% confidence intervals for the short and long allele copy number are computed from the distribution of resampled copy numbers from step 9.
12. Reads are assigned to allele peaks based on bootstrapped estimates of peak mean, standard deviation, and weight – the complete set of parameters characterizing a peak in a GMM.
13. If a user chooses to do so, motif-sized k-mers are tabulated for each peak using the collected read-level k-mer data.

2.4.2 Correcting for STR allele size bias

As STR allele size increases, it becomes less likely that a read of a size from a fixed distribution (i.e., from a given technology) will encompass the entire allele, introducing an allele bias to a set of reads for a particular locus. In non-targeted (whole-genome) sequencing, read size is independent of a given STR allele's size. One can estimate the probability of a randomly selected read spanning an entire STR tract plus pre-specified flanking region, given that it overlaps the STR and there is sufficient depth of coverage of the locus. In Equation 3, $\bar{m}$ is the average length of all reads overlapping the STR region (a surrogate for overall read size distribution), and *t* is the size of the STR region, with flanking sequence, in a particular read. Reads can then be re-sampled, weighted by the inverse of this probability.

$$\hat{p}(span \mid overlap) = \frac{\bar{m} - t + 1}{\bar{m} + t - 2}$$

Equation 3: Approximate probability of encountering a read large enough to span a given STR tract, used to mitigate the effects of large allele drop-out.

In targeted sequencing, the same assumption that a realized read length is independent from STR tract size no longer holds since reads may always span only a targeted locus plus some surrounding sequence. In "targeted mode", I instead treat the current read size as representative of all reads for that STR to perform re-weighting.

2.4.3 Visualizing calls in a web application

As part of creating a visualization tool for STR genotypes, I used the `igv.js` library (Robinson *et al.*, 2022) to display reads with their repeat counts in a genome browser context, and the `Observable Plot` library to display bar plots and histograms (https://github.com/observablehq/plot). The visualization tool is included as part of the `STRkit` package and available under the sub-command `strkit visualize`.

2.4.4 Mendelian inheritance error calculator

An estimate of rates of deviation from Mendelian inheritance serves as a measure of genotyping method reliability and sensitivity (see section 1.6.3). In `STRkit`, I include a module named `strkit mi` to calculate rates of Mendelian inheritance error (ME) from the output of all STR genotyping methods included in the benchmarking sections of this project, for both binary 'yes/no' inheritance observations and parent-offspring 95% confidence interval overlap. For results obtained with this calculator, see section 3.3.4. To measure ME in trio genotyping calls, I perform the following steps:

1. For each locus in the STR callset for the child in the trio, check if a corresponding call can be found in both parent callsets. If not, skip the locus.
2. If the locus is in a user-specified exclusion set, e.g., for removing known *de novo* mutation, skip it.
3. Add this locus to the set S of seen loci for this trio.

4. If (slightly different from step 1) a call *failed* in any of the three individuals, skip the locus (after it has been added to set S.)
5. If the calls for the locus are consistent with Mendelian inheritance, add one to a counter c of consistent loci. Additionally, keep track of counters for binned allele size, with bins of size 10 bp., binning based on reference allele size.
6. After all loci have been processed, return the total rate of ME as $1 - (c)/|S|$ and, in the same fashion, calculate rates of ME for each allele size bin.

Chapter 3 Results

3.1 Selected STR genotyping software for benchmarking

The final set of STR genotyping software included in the benchmark, as chosen using the criteria listed in section 2.3.1, is as follows:

Short read methods:

- `ExpansionHunter` v5.0.0 (Dolzhenko *et al.*, 2017 & 2019)
- `GangSTR` v2.5.0 (Mousavi *et al.*, 2019)

Long read methods:

- `RepeatHMM` v2.0.3 (Liu *et al.*, 2017 & 2020)
- `Straglr` v1.3.0 (Chiu *et al.*, 2021)
- `Tandem-genotypes` v1.9.0 (Mitsuhashi *et al.*, 2019)
    - `LAST` alignment was performed with version 1418
- `STRkit` v0.7.0 (currently unpublished)

3.2 STRkit genotype calls correlate strongly with PCR product sizes

To validate that STR genotyping methods capture true STR variation, we examined correlations between STR genotypes calculated from sequencing data from the 1000

Genomes and Human Genome Structural Variant consortiums and our PCR-derived truth set, previously discussed in section 1.6.4.

Using the method outlined in section 2.2.1, I generated a truth set of PCR-derived genotypes for one individual (NA19238) and 155 polymorphic STR loci. This dataset does not capture the full range of STR variation; rather, it focuses just on tri- and tetra-nucleotide repeats, and all alleles are smaller than 115 bp (Figure 2).

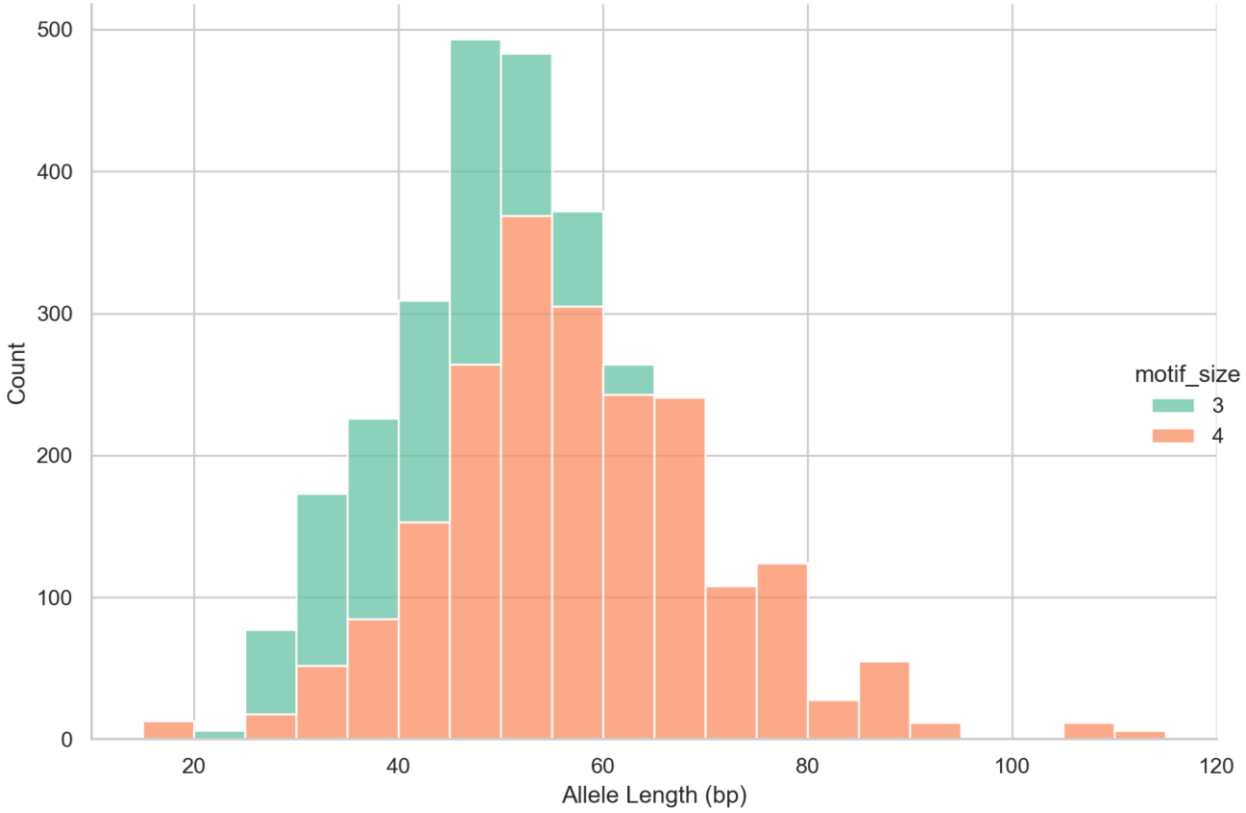

Figure 2: Distribution of allele and motif sizes in the PCR-derived STR genotype truth set. These loci are known to be polymorphic (Payseur, Place, and Weber, 2008).

I then compared calls from STR genotyping software and copy numbers derived from these PCR product sizes. At the same depth of coverage, long read methods show a stronger correlation with the truth set than short read methods and call a greater portion of the provided catalogue. GangSTR only achieved a correlation coefficient (r^2) of 0.19 before author-recommended filtering steps, which remove many erroneous calls and improve correlation (Figure 3). Among long read methods, `Tandem-genotypes` and our method, `STRkit`, achieve the highest $r^2 = 0.88$.

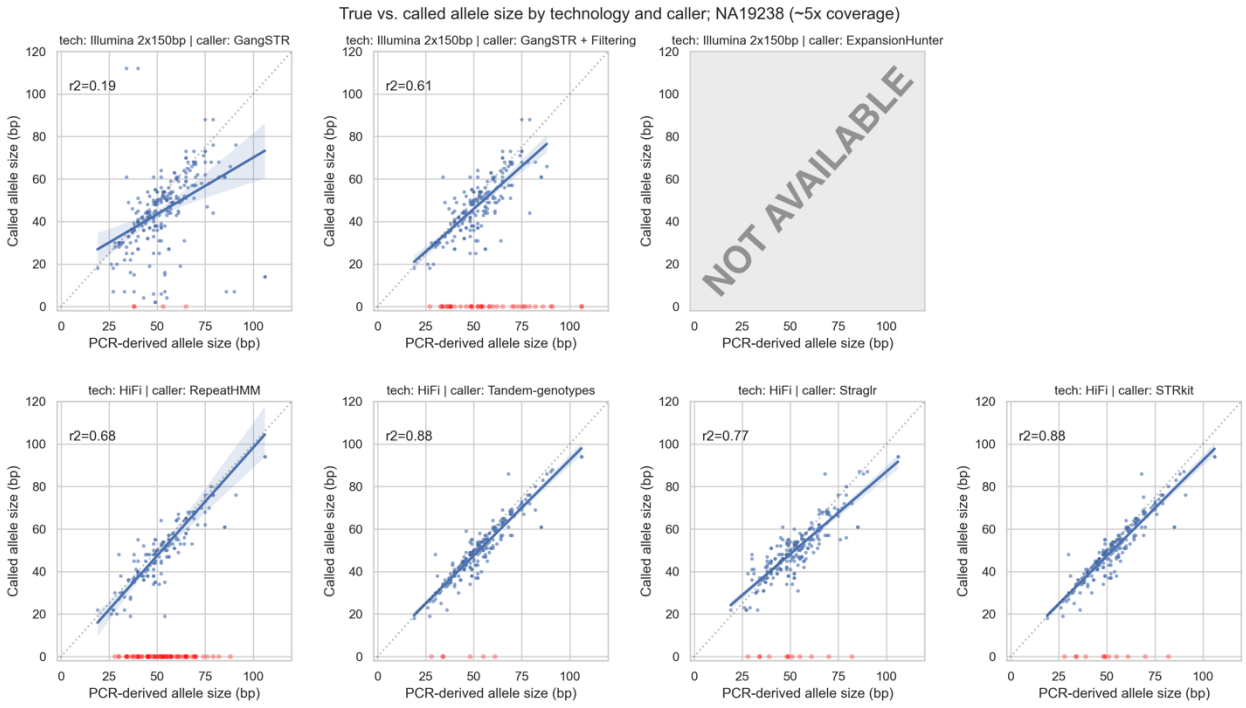

Figure 3: Correlations between a PCR-derived small STR genotype truth set and calls made from low-coverage genomic data for the NA19238 sample. A low-coverage Illumina dataset was created by subsampling high-coverage 30X sequencing data.

3.3 STRkit outperforms other STR genotyping methods on a whole-genome benchmark set

3.3.1 Creating a genome-wide STR benchmarking dataset

The set of alleles used in section 3.2 is small ($n = 310$) and does not capture the full range of STR variation in the genome; most alleles are less than 70 base pairs long (Figure 2). To create a benchmark containing genome-wide variation, I used the method outlined in section 2.2.2 to generate a set of STR copy number genotypes, relative to reference genome GRCh38, from the Ashkenazim GIAB trio. The final variant set contains 36113 loci and a variety of allele sizes (Figure 4), as well as various motif sizes, compositions, and complexity. In the HG002 individual, 13435 of these variants are homozygous non-reference, and 22678 are heterozygous (Table 3).

	HG002 (Child)	HG003 (Father)	HG004 (Mother)
Heterozygous	22678	23539	23615
Homozygous alternate	13435	12574	12498

Table 3: Breakdown of locus zygosity in my STR benchmark set, created from the Genome in a Bottle small variant benchmark set for an Ashkenazim trio.

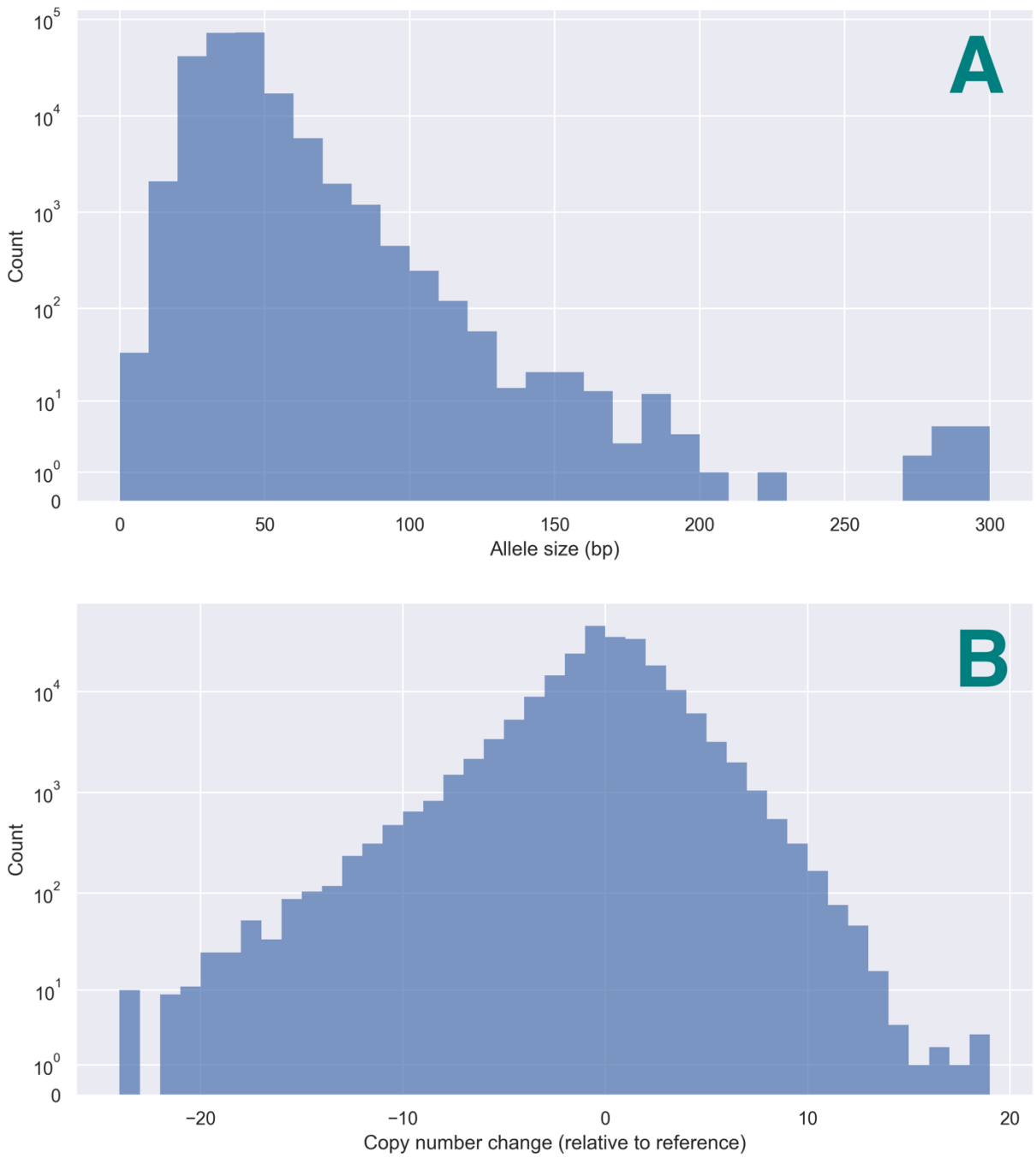

Figure 4: Distribution of STR allele sizes (A) and copy number change relative to the GRCh38 reference genome (B) found in the Genome-in-a-Bottle small variant benchmark for an Ashkenazim trio.

3.3.2 STRkit minimizes error and maximizes accuracy on a high-coverage genome-wide STR benchmark

Using the genome-wide STR benchmark truth set from section 3.3.1, I compared `STRkit`'s genotyping output with other long- and short-read genotyping software capable of whole-genome STR profiling (listed in section 3.1). I subsampled all alignment files across all sequencing technologies for individuals in the benchmark trio to 40-fold average genomic read depth using the approach described in section 2.2.5. I calculated two comparison metrics for 10bp-wide allele size bins: root-mean-squared error (RMSE; section 2.2.3) and binary accuracy (section 2.2.4) of allele copy number relative to the reference genome.

In this benchmark evaluation, genotypes from `STRkit` and `Tandem-genotypes` had the lowest average error for most size bins (Figure 5 A); `STRkit` had the lowest overall error by a small margin (Appendix A, Table S5). `STRkit` was most accurate (Figure 5 B). CCS reads gave the lowest error across the allele size spectrum; short read methods show substantially increased rates of error and lower accuracy as allele size increases. ONT-UL reads have a high baseline error rate and low accuracy across allele size. `RepeatHMM` failed to finish in the time allocated to it (5 days) when running with ONT-UL reads. Figure 5 C shows the proportion of the benchmark truth set catalogue called; the drop-off visible for `GangSTR` with 150bp Illumina reads above ~80 bp is caused by filtering steps recommended by the authors of the tool to remove

inaccurate calls. With 250bp Illumina reads, `GangSTR` starts to miss longer alleles even without filtering. `RepeatHMM` and `Straglr` miss some shorter alleles with CCS data.

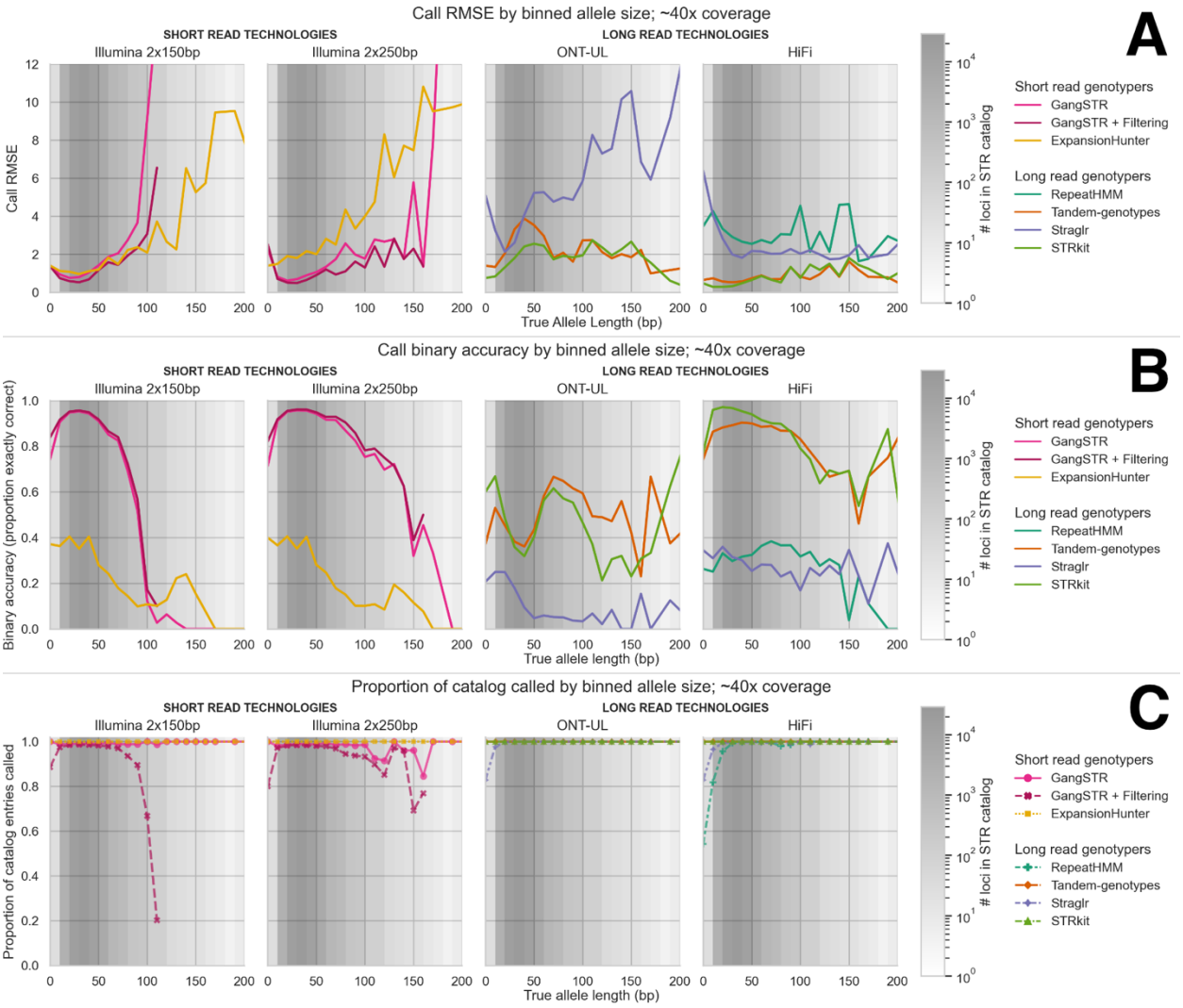

Figure 5: Benchmarking results for a whole-genome STR truth set derived from the Ashkenazim trio GIAB short variant benchmark, using sequencing data at 40x genome-average coverage for the HG002 sample. Shown is a comparison of RMSE (A), binary accuracy relative to reference (B), and proportion of catalogued loci called (C) by allele length and method. Bins are shaded by log-density of alleles in the truth set.

3.3.3 STRkit outperforms other long read STR genotypers when assessing intra-locus copy number difference and classifying locus zygosity

While the results in section 3.3.2 show overall rates of genotyping error, they only indirectly capture how well the different STR genotyping methods can distinguish between alleles within a locus – i.e., correctly predict zygosity. To interrogate this aspect of genotyping more thoroughly, I looked at the RMSE of estimated allele copy number difference with high-coverage sequencing (Figure 6 A) and performance on the zygosity classification task (Figure 6 B) in the whole-genome benchmark.

CCS reads with either `STRkit` or `Tandem-genotypes` achieve the lowest overall intra-allele copy number error, meaning they more precisely capture the difference in copy number between the shorter and longer STR allele within a locus (Figure 6 A). This is true even for relatively short alleles (~50bp).

When assessing zygosity, `RepeatHMM` and `Straglr` significantly overestimate and indeed almost exclusively predict homozygosity, discarding intra-allele variation when reporting final genotypes. Inversely, `Tandem-genotypes` over-predicts heterozygotes. `STRkit` over-predicts heterozygotes to a lesser degree and outperforms all other long read STR genotypers with this classification task, achieving comparable performance to short read methods.

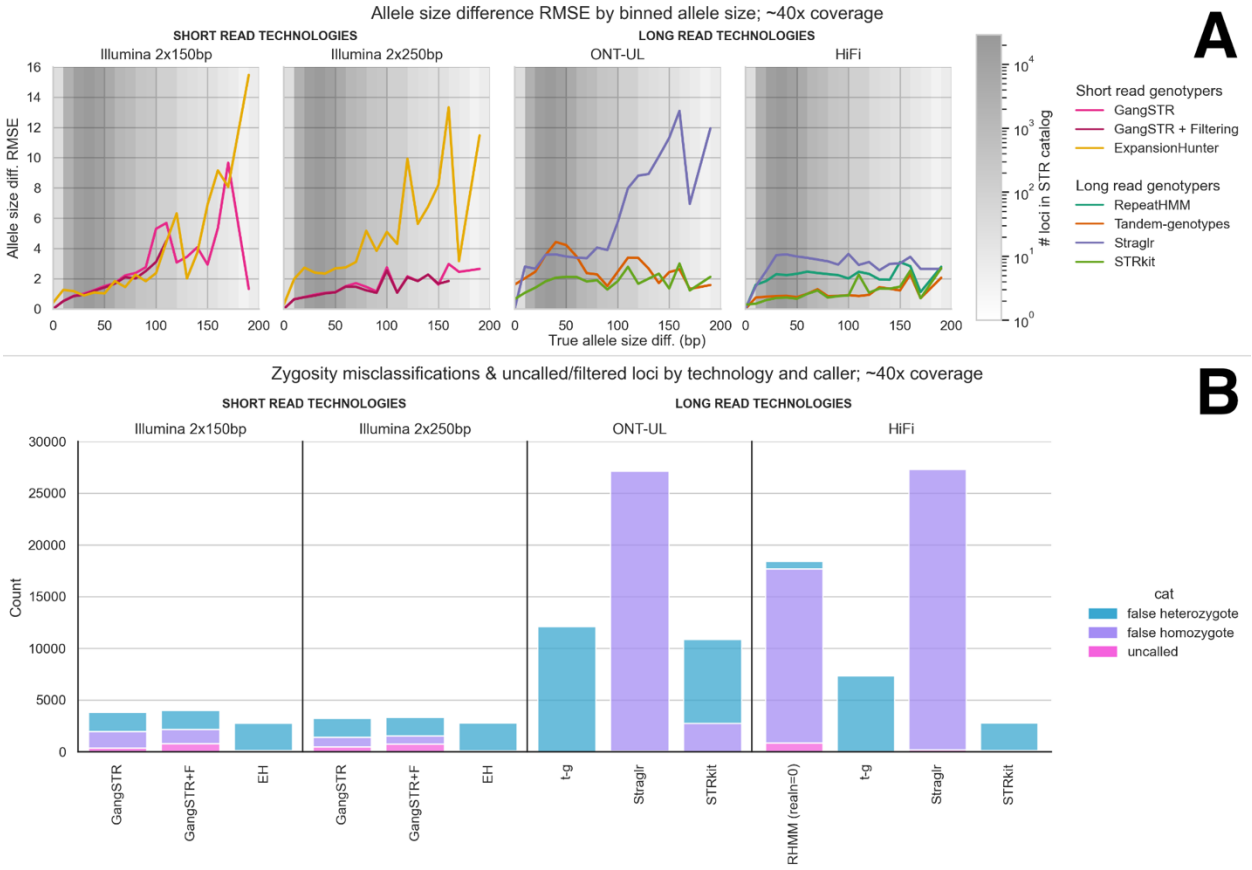

Figure 6: Comparison of STR genotyping performance when assessing copy number distance between alleles within a locus at 40x sequencing coverage for the HG002 sample. (A) Average allele size difference RMSE by allele size (10bp bins). (B) Zygosity classification performance, where locus calls are treated as homozygous if they predict the same two copy numbers (or copy numbers within 0.5 repeats of each other in the case of `Straglr`, because it counts fractional repeats.)

3.3.4 STRkit improves tracing of parent-child allele inheritance in long read data

Ascertaining *de novo* mutation has implications for population genetics and complex trait association studies. While results for the GIAB truth set from sections 3.3.2 and 3.3.3 should correlate with sensitivity to *de novo* mutation, I wanted to assess this sensitivity more directly. To do this, I included a Mendelian error calculator in `STRkit` (see section 2.4.4 for the approach) to quantify how well a given sequencing technology/STR genotyper pairing follows allele transmission from parents to offspring. I used this calculator to compare rates of Mendelian inheritance error (ME) on the Genome-in-a-Bottle-derived whole-genome benchmark callset; a comparison by read technology and binned allele size using high-coverage sequencing is shown in Figure 7. Only loci which were consistent with Mendelian inheritance in the original trio ground-truth callset were included in this comparison. Some results for `Tandem-genotypes` are missing due to the `LAST` alignment software not completing in the time allocated to it (5 days) for the HG002 individual.

In this comparison, `STRkit` with CCS reads achieves low rates of ME across the spectrum of locus sizes, with a similar error rate to `ExpansionHunter` with short reads (Figure 7). With CCS reads, `RepeatHMM` and `Straglr` perform poorly, with around a 64% and 75% rate of overall ME, respectively. All methods perform relatively poorly with ONT-UL reads, with `STRkit` performing best of those available. `GangSTR` performs significantly worse than `ExpansionHunter` with 150bp reads as allele size grows.

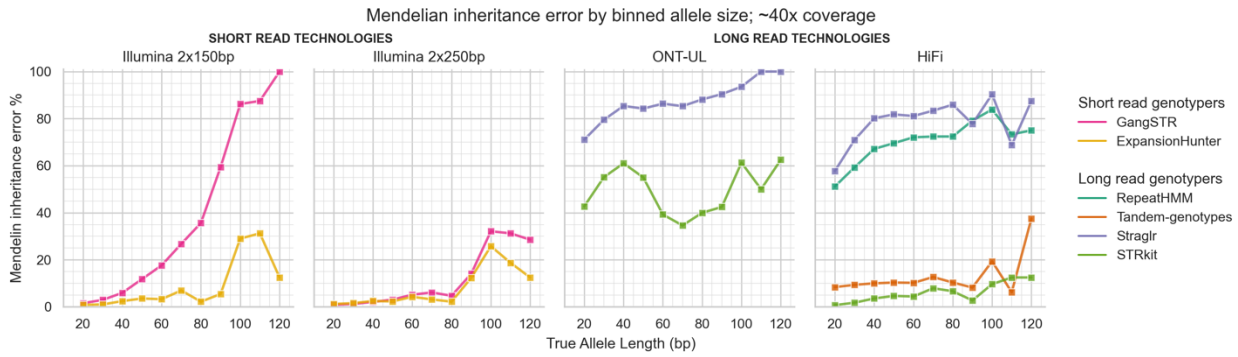

Figure 7: Mendelian inheritance error in the Ashkenazim trio by reference locus size (i.e., the size of the locus in the GRCh38 reference genome; 10 bp. bins) at 40x average depth of sequencing coverage.

3.3.5 STR genotyping using long reads tolerates low sequencing depth

The genome-wide benchmarking results examined in sections 3.3.2 and 3.3.4 are from high-coverage (40x) sequencing data. Sequencing at this depth of coverage is not always feasible due to cost, and existing sequencing datasets may not be available at this depth. To examine how STR genotyping methods perform at lower read depths, I ran the genotypers on subsampled alignments of individuals in the Genome-in-a-Bottle-derived STR benchmark. For the description of how subsampled alignments were created, see section 2.2.5. I also tested runtimes for each method with each subsampled alignment to examine how sequencing technology, coverage depth, and algorithmic approach affect allele genotyping throughput.

Short read methods were affected more severely by loss of coverage, i.e., more sequencing coverage was required to achieve maximal STR genotyping performance with short reads in the whole-genome STR benchmark (Figure 8 A, B, C). At all depths of coverage, `STRkit` with CCS reads achieved the lowest overall genotyping error (Figure 8 A) and highest accuracy (Figure 8 B). `STRkit` performs the best at all coverage levels in terms of Mendelian inheritance error (ME) in both long-read technologies (Figure 8 C), although `Tandem-genotypes` calls were not available for higher levels of coverage with ONT-UL reads due to alignment timeout. For short reads, `ExpansionHunter` outperforms `GangSTR`, especially as coverage decreases. Rates of ME with `STRkit` on CCS reads continued to improve as coverage increases beyond 30x, where other metrics do not improve much past this threshold.

Short read STR genotyping software showed much higher computational throughput in terms of number of loci processed per second (Figure 8 D). Within long read methods, `STRkit` shows slightly worse computational throughput than `Straglr`, while producing better genotyping results in this benchmark. `RepeatHMM` has extremely low allele throughput, which was noted by the `Straglr` authors as well in their testing of other methods (Chiu *et al.*, 2021), and did not finish in the time I was able to allocate to it when given Oxford Nanopore ultra-long reads. `Tandem-genotypes` was fastest for long reads, with the caveat that the alignment software required (`LAST`) is much slower than `minimap2`, which was used for other alignments, also noted by Chiu *et al.* (2021).

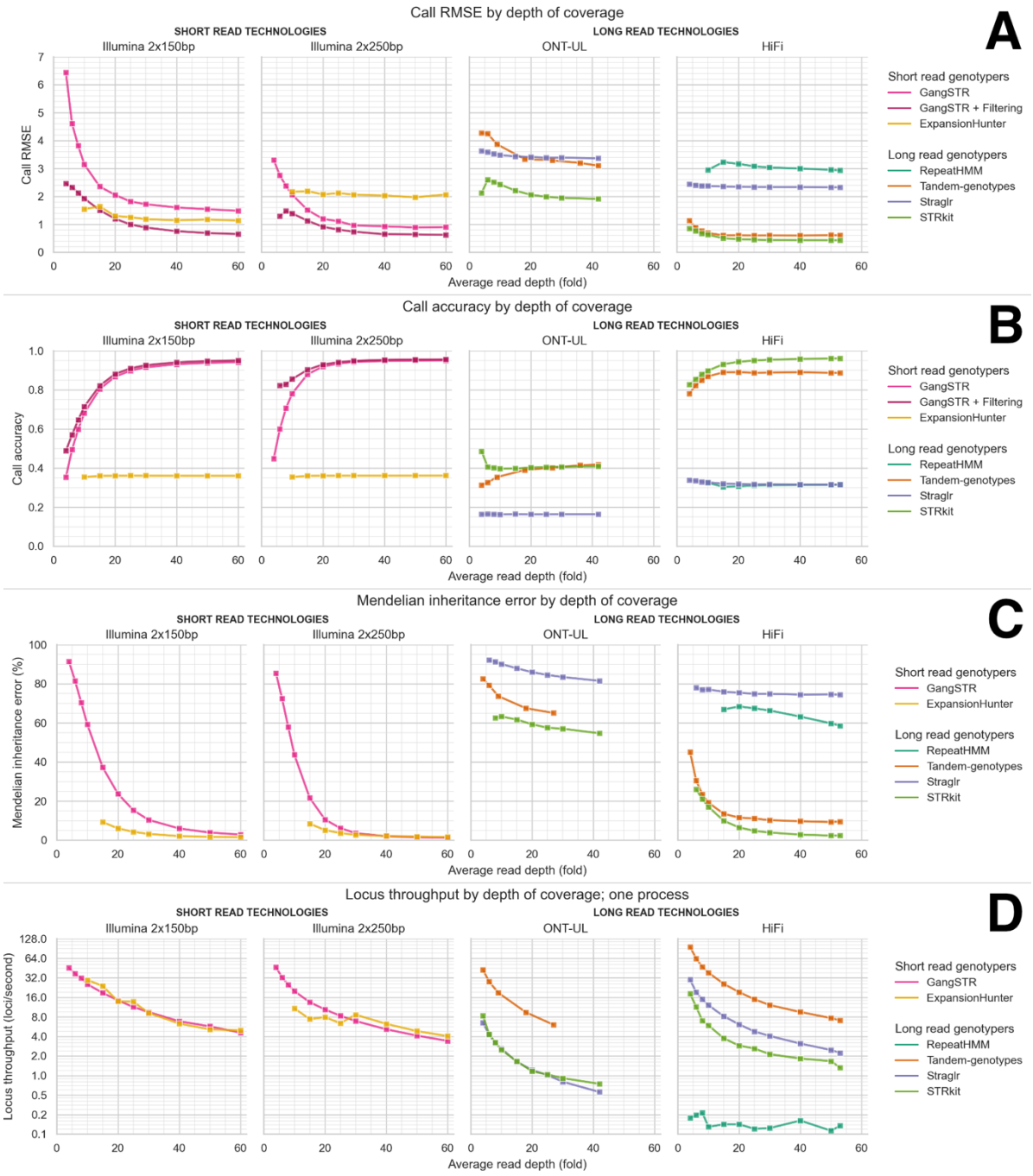

Figure 8: STR genotyping performance in terms of error (A), accuracy (B), ME (C), and computational throughput (D) by average depth of sequencing coverage for the HG002 sample. Alignment files were subsampled to multiple different coverage levels. I excluded points from sub-figures A-C if more than 50% of the catalogue was not called.

3.4 STRkit detects pathogenic expansions in targeted CCS data

In sections 3.2 and 3.3, I have examined STRkit's genotyping capability with genome-wide STR variation, and, by proxy in section 3.3.4, the method's potential for small *de novo* mutation detection. Until now, genotyping of large pathogenic alleles has not been assessed. Most known STR-caused diseases are expansion disorders (section 1.3), but almost all genomic copy number variation within the GIAB benchmark is within 20 repeat units of the reference genome (Figure 4: Distribution of STR allele sizes (A) and copy number change relative to the GRCh38 reference genome (B) found in the Genome-in-a-Bottle small variant benchmark for an Ashkenazim trio. B). To validate `STRkit`'s ability to genotype pathogenic loci, I obtained a public targeted expansion sequencing dataset published by PacBio (section 2.2.6) and corresponding expansion repeat counts provided from the Coriell Institute (https://www.coriell.org/; last accessed Oct. 4, 2022). In all samples except the control, `STRkit` found an expanded allele. In most cases, `STRkit`'s reported genotypes fell within the range reported by the institute or another source (Table 4), with some longer expansions showing greater repeat counts in sequencing data versus what is reported for the sample. There is visible mosaicism in the expanded HTT allele of sample NA20253 (Figure 9 A), which De Luca *et al.* (2021) also found using a repeat-primed PCR technique. There is also likely mosaicism in the NA14044 sample (Figure 9 B), noted by Chiu *et al*. (2021).

	Expanded Allele Genotype (from Coriell§ except where noted)		**Adjusted*† `STRkit` Genotype**		**Notes**
Sample	HTT	FMR1	HTT*	FMR1†	
NA13505	22/**50**	No expansion	22/**51**	30	
NA13509	15/**70**	"	15/**75**	30/31	
NA20253	22/96-103[a]	"	22/**114**	20	Some mosaicism visible and validated with PCR by De Luca *et al.* (2021) (Figure 9 A)
NA14044	19/**250**	"	19/**962**	30	Large range of copy number in expanded visible in reads (Figure 9 B)
NA13664	No expansion	28±3/**49±3**	16/17	31/**54**	"Upper limit of normal"
NA06896	"	23/95-140	12/20	23/**187**	'Pre-mutation' expansion
NA07537	"	28-29/**>200**	12/17	29/**339**	
HEK293	N/A (control)	N/A (control)	17/18	35/35	Control – no known expansions

§ Accessed from https://www.coriell.org/ Oct 4, 2022.
* The `TRF` catalog includes a tailing CAACAG as part of the HTT repeat, so I subtracted 2 from reported repeat counts for comparison against the Coriell-reported genotypes.
† The `TRF` catalog includes an interrupting section (4 amino acids) in the FMR1 repeat, so I subtracted 4 from reported repeat counts.
[a] Mean PCR genotype across 10 volunteer laboratories from Kalman *et al.* 2007.

Table 4: Expanded HTT and FMR1 alleles as genotyped by `STRkit`.

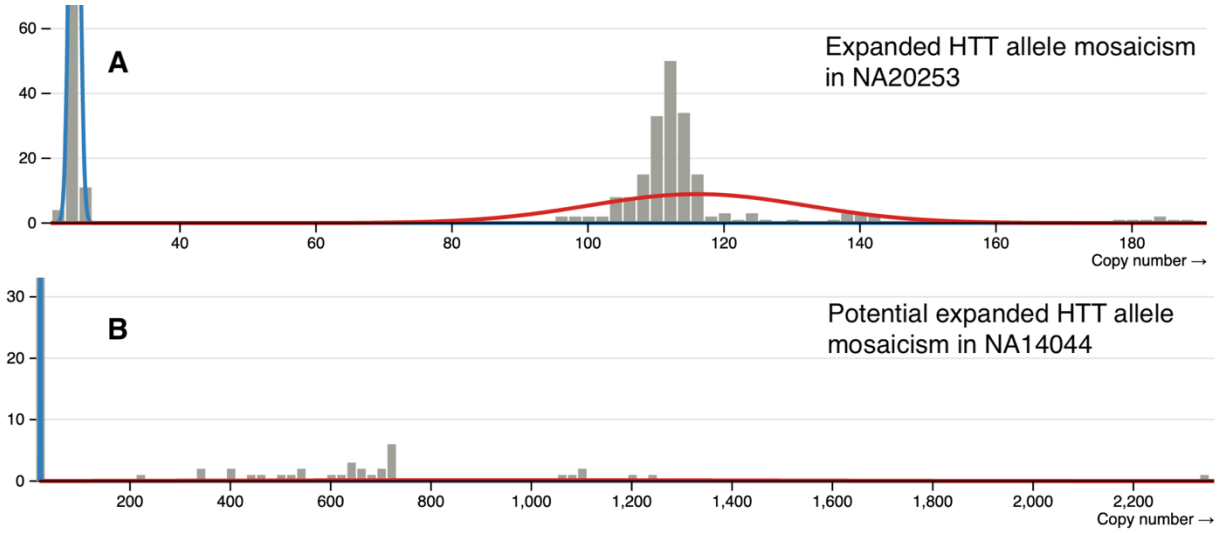

Figure 9: Instances of mosaicism in expanded HTT alleles captured by targeted CCS. NA20253 (A) has three visible peaks, at ~110, ~140, and ~180. The red and blue lines are `STRkit`'s best-guess peak calls; their poor fit is a result of the wide spread of expanded alleles within the sample.

3.5 Read-level visualization of copy number with `STRkit` reveals sequencing noise and STR instability

Visualizations of read-level copy numbers, such as the histograms in Figure 9, can help researchers understand the extent of somatic instability of an STR within a particular sample, which has been implicated in, e.g., age of onset in Huntington's disorder (Swami *et al.*, 2009). To facilitate this and the analysis of results of STR genotyping more generally, `STRkit` includes a graphical visualization tool with a web-based interface that can show read-level copy number data for a given locus, as well as motif repeat k-mers and run parameters (Figure 10). I used this tool to visualize the extent of instability in targeted sequencing data in two expanded HTT alleles (Figure 9). To better show repeat expansions, I contributed code to the `igv.js` library (Robinson *et al.*, 2022) to show base-pair insertion counts and allow dynamic colouring of reads based on a property such as read-level STR copy number (Figure 10 B). Halman, Dolzhenko, and Oshlack (2022) developed a tool with similar goals, `STRipy`, to visualize short-read STR genotyping results from `ExpansionHunter`, although their method also includes links to literature about locus-associated disease and pathogenic expansion threshold information.

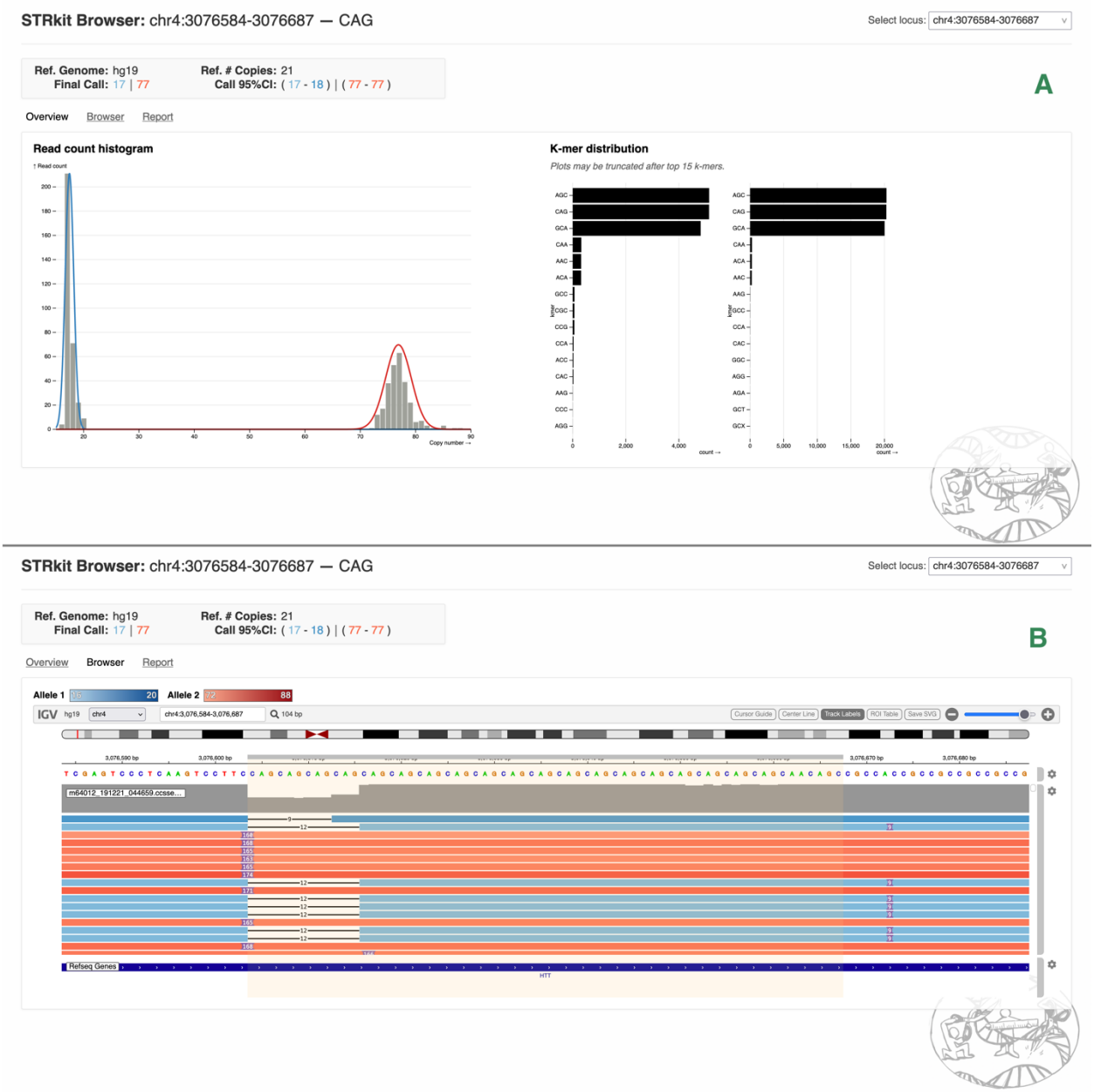

Figure 10: Web user interface for `STRkit`'s "visualize" functionality, showing data from an expansion in HTT sequenced using CCS from section 3.4. (A) The overview section, with a histogram of repeat counts by read, and a distribution plot of repeat motif sequences found in the alleles. (B) An `igv.js` genome browser, showing reads with expansion insertions (purple boxes).

Chapter 4 Discussion

4.1 Comparing STR genotyping methods

In this project, I set out to evaluate the potential of CCS long reads for genotyping STRs and develop a genotyping method which maximally takes advantage of the potential of these high-fidelity long reads. I designed a software package called `STRkit` for this purpose, with an approach that is more sensitive to small copy number changes, building on the foundations of existing short- and long-read STR genotyping software and trying to address problems encountered in my own and others' use of these packages, such as `RepeatHMM`'s runtime and `Tandem-genotypes`' costly realignment process (Chiu *et al.* 2021).

To benchmark STR genotyping methods, including `STRkit`, I first wanted to show that they capture real STR variation. Using a truth set of small polymorphic STRs created using the method described in section 2.2.1, I compared correlations between PCR product size-derived STR genotypes and copy number output from `STRkit` and other genotyping methods given low-coverage sequence data from the NA19238 individual. All methods show a correlation between PCR and sequenced genotypes, with `STRkit` and `Tandem-genotypes` achieving the highest correlations ($r^2 = 0.88$) at an average genomic depth of coverage of 5x (Figure 3). This result indicates that genomic sequencing, especially with these methods, captures real variation in repeat count validated with traditional laboratory techniques; in this case, using capillary

electrophoresis to assess PCR product sizes. Normally, PCR product sizes may not correlate exactly with copy numbers, since other non-STR copy number change indel variants may occur within the bounds of the PCR primer pairs. To eliminate this bias, I used variant callsets created from high-coverage short read sequencing data to skip loci with non-STR indels in these regions, as outlined in section 2.2.1.

The PCR-genotyped benchmark STR set contains only short STRs (Figure 2); all alleles in the set are below 115 base pairs in length. A single Illumina 150bp short read can encompass any allele in the dataset with flanking sequence information. Pathogenic repeat expansions can be anywhere from ~100bp (e.g., at the lower end of the pathogenic HTT repeat expansion) to thousands of base pairs in, for example, SCA10 (McFarland *et al.*, 2015). The precise repeat counts of expanded alleles are often correlated with phenotype (Allen *et al.* 2021; Gall-Duncan *et al.*, 2021). I thus decided to augment my testing with a much larger genome-wide truth set derived from a deeply-sequenced trio of individuals with an available benchmark variant callset, published by the Genome in a Bottle consortium.

In this genome-wide benchmark, CCS reads showed the greatest potential for genotyping STR alleles in the 0-200 base pair size range; this is most apparent in `STRkit` and `Tandem-genotypes`' low average root-mean-squared error (RMSE) rates: while the error rate increased with allele size, it did so at a much lower rate than with short read genotyping (Figure 5 A). `STRkit` with CCS reads achieved the highest binary accuracy in the high-coverage sequencing benchmark (Figure 5 B, Table S5),

which I attribute to its sensitive peak calling approach. Here, `STRkit` is the only software using CCS reads to achieve parity with short read methods in terms of accuracy in smaller alleles, while also achieving overall better performance with longer alleles versus short read methods. `RepeatHMM` and `Straglr` both perform poorly in my RMSE and accuracy comparisons; when accounting for the results from my zygosity classification task (Figure 6 B), I surmise that these poor results are due to calling almost all loci as homozygotes, thereby effectively taking the average of two different copy numbers and missing small copy number heterozygosity in many loci. In both Illumina short read technologies, one can observe that error rate and accuracy both worsen as alleles expand past ~1/2 read length for the two read sizes (75bp and 125 bp thresholds for 150bp and 250bp reads, respectively) – this is an inflection point at which a short read which overlaps an STR allele is less likely to fully encompass it than not. `Straglr` shows poor performance with very small alleles in both long read technologies; when I investigated their method, I found that they used Tandem Repeats Finder (`TRF`) internally for finding STRs within long reads. For novel STR discovery, this makes sense; however, for STR genotyping, it can cause problems, as demonstrated here. `TRF` requires a certain score threshold before reporting a repeat, which small alleles (or, in the extreme case, the complete absence of an allele) would not reach, resulting in the call being missed or misreported. `STRkit`'s copy number assessment incorporates match scores for flanking regions (see section 2.4.1 for the full algorithm), which allows for correct genotyping of extremely small STR alleles, or even the complete deletion of an STR tract.

The presence of many heterozygotic loci in the benchmark dataset allowed me to evaluate the peak calling aspect of STR genotyping, i.e., determining alleles from a pool of read-level copy number data. Quantified intra-locus allele copy number difference error (Figure 6 A) shows that CCS reads and associated STR genotyping methods are the best of the technologies tested at differentiating allele copy number within a locus. In my zygosity classification task (Figure 6 B), `STRkit` stands out as comparable to short read methods such as `ExpansionHunter`, where other long read methods misclassify more often, with `RepeatHMM` and `Straglr` showing extreme bias towards calling homozygotes and Tandem`-genotypes` showing a tendency to over-call heterozygotes. Performance in this task has implications for measuring true STR variation within populations, as well as assessing STR mutation rate. It follows that methods which miss variation will under-report genetic diversity of STRs and mis-represent population allele distributions, if used for this purpose. In long read methods, the task of allele calling is somewhat independent of read-level copy number assessment; choices made by software authors here can affect the error trade-off between over- and under-reporting homozygotes. For example, I attribute `RepeatHMM`'s tendency to over-call homozygotes to its error model, which was designed for high-error long reads, and as a result may 'smooth over' true variation when given lower-error reads. In their discussion of `Straglr`, Chiu *et al.* (2021) note that their results may be improved by preventing the Gaussian mixture model (GMM) implementation from wide distributions into a single copy number; my results confirm that `Straglr` tends to group disparate loci together

when using CCS data and could benefit from a more sensitive GMM when using such data, such as the one I used in `STRkit`.

Methods which struggled to determine zygosity correctly and had high intra-locus allele error also performed poorly in my Mendelian inheritance error (ME) benchmark (Figure 7, Figure 8 C). The root causes here are likely the same: variation is missed, meaning peak calling models frequently end up taking the average of two copy numbers when reporting alleles. Since only one allele of a locus from a parent is transmitted to a child, a parent's average copy number within a polymorphic locus will frequently differ from the child's, resulting in a high rate of ME for callers which over-call homozygosity. `STRkit` performs well in the ME benchmark, showing lower error than other methods with CCS data and `GangSTR` with short reads (Figure 7). Performance was comparable to `ExpansionHunter` with shorter alleles and appeared to be better than any short read method with longer alleles, although allele sample size is too low to say this conclusively. An expanded dataset with additional trios could help elucidate this.

Rates of ME have implications for using STR genotyping software to detect likely *de novo* mutation: false positives when detecting *de novo* mutation should naturally be fewer if inherited alleles are correctly traced parent to child. *De novo* mutation detection is relevant to medical genomics and population genetics studies; Mitra *et al.* (2021) used sequencing-based STR genotyping to calculate the contribution of small *de novo* mutation events to autism, and better-resolved mutation rates can be used in population genetics modeling (Goldstein *et al.*, 1995; Zhivotovsky *et al.* 2004). Naturally, the more

accurately an STR genotyper can trace the inheritance and mutation of alleles, the better it should be at estimating genome-wide and locus-specific mutation rates. As a metric, ME has limitations; it cannot quantify systematic bias or genotyping errors which are ‘inherited’ alongside the allele itself. Consider a “pathologically bad” STR genotyper which reports a copy number of 0 for every allele it receives: this would yield a perfect 0% Mendelian error. However, when examined in context with other benchmarking results here, this concern is assuaged: all callers which perform well in terms of ME also achieve low error and high accuracy in the other GIAB benchmark, indicating that methods with low ME are correctly tracing parent-child allele inheritance.

So far, I have discussed three different metrics for assessing genotyping performance: absolute genotyping error, accuracy, and Mendelian inheritance error. All three vary as a function of average genomic read depth (Figure 8 A, B, C); for both ONT-UL and CCS data, `STRkit` achieves the highest accuracy and lowest absolute genotyping error at all coverage levels. As one expects, performance tends to improve as coverage increases, and more data are available. The point at which performance asymptotes is sequencing technology-dependent; long read methods appear less affected by a reduction in base-wise average coverage, and performance gain seems to asymptote at a lower coverage level. For multi-base DNA features like STRs, there is a difference between base-level average coverage as I use it here and “feature-level” average coverage; I discuss the differences between these and the difficulty of comparing coverages across sequencing technologies in section 4.3.

Coverage significantly affects allele throughput and thus computational runtime (Figure 8 D). There is also quite a large performance spread between STR genotyping methods; `Tandem-genotypes` is the fastest overall, followed by both short-read genotyping tools. I believe this to be a combination of implementation choice and an inherent property of the data they process; both `ExpansionHunter` and `GangSTR` are implemented in C++, which is generally faster than Python. The performance gain of `Tandem-genotypes`, on the other hand, likely results from its approach; it does not examine the sequence itself, and simply reports the overall size change of an aligned region relative to the reference, rounded to a whole copy number change. In contrast, `RepeatHMM`, `Straglr`, and `STRkit` model STRs at a motif and copy number level, which should improve their ability to properly count copy number if small indel errors are present in individual motif copies. Uniquely among long-read STR genotypers, and inspired by short read methods, `STRkit` also re-samples read counts multiple times to calculate a genotype confidence interval. `Tandem-genotypes` requires realignment with the `LAST` aligner, which Chiu *et al.* (2021) found to be around half as fast as `minimap2`; this offsets some of the throughput it gains versus other long read methods. `STRkit` and `Straglr` both support parallelization for individual samples whereas `Tandem-genotypes` does not, so `STRkit` should be comparable to or faster than `Tandem-genotypes` for, e.g., small batches of samples where multiple cores can be allocated to each sample, when also accounting for alignment time. `RepeatHMM`'s extremely poor performance may stem from its implementation of a complex Hidden Markov model for determining copy number; these models should be more error-

resistant for high error reads, but for CCS data they appear to be less relevant and computationally intractable at the whole-genome scale.

4.2 `STRkit`'s genotyping performance and limitations

Despite its moderate computational speed, `STRkit`'s improvement in genotyping quality over other methods at all depths of coverage with CCS makes this sequencing technology more powerful for STR genotyping. `STRkit` includes some unique features which may contribute to this performance; one such feature is STR tract re-weighting, described in section 2.4.2. As an STR tract increases in size, it becomes less likely that a read from a given technology will encompass the entire tract, with or without flanking sequences on either side. This leads to sampling bias if an STR genotyping approach uses only reads which span an entire STR locus, as all existing long-read STR genotyping methods discussed here do. A naïve use of read-level data to assess true allele size, mosaicism, or STR instability will place too much weight on short STR tracts, and not enough on longer ones (e.g., expansions), following this sampling bias; to counter this, `STRkit` samples longer alleles more frequently during the bootstrapping steps. Another feature to which I attribute `STRkit`'s accuracy is its handling of reference copy number. Most other STR calling methods (`ExpansionHunter`, `GangSTR`, `Straglr`, `Tandem-genotypes`) use the catalogue coordinates to calculate the repeat count of the STR in the reference genome, but their own method to calculate repeat count in samples; this creates a disconnect between the reference genome and the sample in cases of inexact repetition of motifs or fuzzy boundaries of STRs within

the reference sequence and can lead to systematic bias in reported relative copy number for these types of loci. I believe this to be the cause of `ExpansionHunter`'s poor accuracy on the GIAB benchmark (Figure 5 B). `STRkit` uses the same algorithmic approach to count copies in the reference and in reads from a sample, which should reduce this category of bias. This algorithm allows for some flexibility in catalogue coordinates to count imperfect copies beyond what the catalogue may include, which also captures slight misalignments of motif copy insertions or deletions to just outside catalogued STR boundaries.

`STRkit` did not solve all issues with existing STR genotyping methods, as discussed in section 1.5.3; for example, it uses a catalogue-based approach, which generally depends on a complete reference genome to identify as many loci as possible and may miss loci which do not occur in the reference genome at all or may contain interruptions and non-reference repeat motifs; in contrast, `Straglr` has a mode which uses repeat-like insertions to detect expansions without provided reference coordinates. I expect that with gradually increasing usage and annotation of the new Telomere-to-Telomere CHM13 reference genome newer, more complete STR catalogues can be deployed, which `STRkit` can then take advantage of. With non-reference repeat motifs, `STRkit` has some limited support for matching via IUPAC codes (e.g., `N` representing any of `ATGC`), but this requires knowledge *a priori* of what form a non-reference locus may take. `STRkit` still tolerates some interruption, because the alignment method penalizes the number of mismatches linearly and thus can compensate for error if the overall

alignment still scores well; however, if an interrupting motif is extremely disparate from the one in the catalogue, it may result in missed or inaccurate genotype calls. The latest versions of the `ExpansionHunter` tool (Dolzhenko *et al.* 2019) include a more advanced pattern-matching system which supports more forms of interrupted STRs, albeit requiring *a priori* knowledge to an even greater degree than using IUPAC codes with `STRkit`.

4.3 Benchmarking limitations

With an extensive comparison of existing STR genotyping methods and the development of a new one, I showed that CCS reads outperform short read sequencing for STR genotyping in many situations. However, short reads still serve as a useful STR genotyping resource. They are shown here to be very accurate for small alleles, and still useful for evaluating copy number in longer alleles in an approximate fashion (Figure 5), while being significantly more cost-effective than CCS as of time of writing – a factor which is not captured in the benchmark. `GangSTR` and `ExpansionHunter`'s more advanced algorithmic approaches allow genotyping beyond read length by combining data from multiple read pairs to find the most likely allele copy numbers, and given equal time, both tools can process more loci than any long read approach (Figure 8 D). Nevertheless, short reads fundamentally limit what STR variation can be captured and genotyped precisely, especially with, for example, non-reference motif interruptions (Dolzhenko *et al.*, 2020), mosaicism, or somatic repeat instability – as multiple reads from a potentially heterogeneous STR tract must be combined.

The PCR-derived genotypes I used for the first benchmark do not capture this limitation of short reads, although they do reveal the poor performance of short read technologies at low coverage. As shown in section 3.2, all alleles in this benchmark fit within the span of a single short read. To capture a more representative set of alleles, I expanded the scope of validation and method comparison using small variant benchmark sets from the GIAB consortium; these data have their own caveats. Because alleles were extracted from a set of so-called “small” variants (i.e., indels not large enough to be classified as ‘structural variants’ by the consortium), the resulting STR genotype set is potentially missing larger STR expansions. Wenger *et al.* (2019) suggest that the GIAB tooling does not perfectly capture tandem repeat variation; however, Wagner *et al.* (2022 A) presented a new version of the GIAB small variant benchmark, which I used here, which claims to improve TR genotype calls.

One overall limitation of the benchmarking I performed is the low number of individuals; for the PCR benchmark, one individual was used, and for the Genome-in-a-Bottle benchmark, I calculated a truth set for a single trio for use with Mendelian inheritance error calculation, and for just the offspring, HG002, for other benchmarking. Expanding the benchmark to more individuals would give more insight into rarer alleles such as uncommon expansions or deletions; for the PCR benchmark, this would require sequencing data for additional individuals to be made available in the IGSR portal. For the GIAB benchmark, there is a second trio available which I did not incorporate but which could be included in a future study. My benchmarking also did not examine

performance with specific motif compositions or regions in the genome; for example, to see if the struggles of CCS with homopolymers (Wenger *et al.*, 2019) extend to homopolymer-containing STR motifs.

Another limitation of benchmarking in this project is the use of a single aligner for each sequencing technology. Rajan-Babu *et al.* (2021) showed that choice of aligner impacts STR genotyping in short reads, and a method such as Winnowmap2 (Jain *et al.* 2022), which purports to be more sensitive when aligning long reads to repetitive DNA, may improve calling across many methods tested here.

My benchmarking did not address the identification of pathogenic expansions within whole-genome STR genotype sets. I used exclusively high-coverage targeted sequencing for validating pathogenic expansion genotyping by `STRkit` (section 3.4). The whole-genome benchmark did, however, contain alleles in the pathogenic size range of some expansion disorders. In whole genome sequencing contexts, lower coverage may affect the ability of various methods to separate expansions from sequencing noise or genotyping error, and sequencing technologies themselves may behave differently between targeted and whole-genome sequencing protocols.

Comparing across depths of coverage between sequencing technologies is difficult in general, and in both benchmark datasets there is limited interpretational use in comparing short and long read technologies at the same depth of coverage, beyond looking at overall trends in the performance/coverage curves shown in Figure 8. For

example, at four-fold depth of coverage, one expects four reads on average to overlap a given single base pair regardless of sequencing method. However, with long reads, depth of coverage for the entire STR element is greater, i.e., there is a greater probability that a longer read spans the element (Figure 11). This probability is a function of the sequencing technology-specific read length distribution and allele size. Single-ended long reads cannot be easily used for STR genotype assessment if they terminate in the middle of a perfect repeat; an STR genotyper can use a single allele-spanning read to compute a copy number estimate, whereas a non-spanning read can only produce a lower bound on copy number. For example, given a simple sequence repeat of $(CAG)_n$ and an alignment like what is shown in Figure 11 B, if a read terminates inside the STR, giving a DNA pattern of $(CAG)_{n-m}$, it cannot be known precisely how many repeats m exist between the read terminus and the true end of the repeat region. Further complicating coverage comparisons, each CCS read consists of data from multiple subreads (Wenger *et al.*, 2019). This non-equivalence of quantifications of depth of coverage between technologies explains part of the steeper drop-off in performance as coverage decreases in short-read STR genotyping methods as compared to long-read methods (Figure 8), while illustrating a key benefit of accurate long reads for STR genotyping and somatic instability assessment: allele-spanning reads are abundant for all but the longest alleles, and one spanning read yields one copy number.

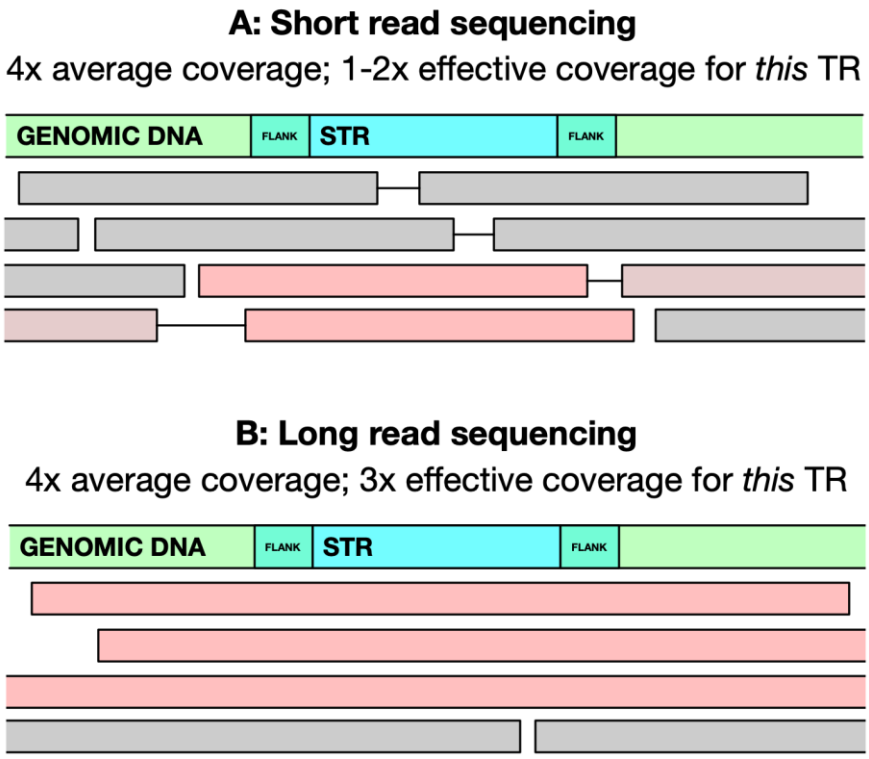

Figure 11: Nonequivalence of depth of coverage for STRs across short and long-read technologies. Reads which span the STR region are shown in pink.

Chapter 5 Conclusions and Future Directions

Here, I present `STRkit`, an STR genotyping toolkit designed to work with high-fidelity long read technologies such as CCS and provide tools for researchers to explore STR variation at a genome-wide scale. I designed a workflow to generate an STR benchmarking dataset using Genome-in-a-Bottle's public small variant benchmark data for an Ashkenazim trio and evaluated my method and others on these data and other public data. I found the CCS technology to be a viable candidate for genome-wide STR profiling when paired with the right software, outperforming Oxford Nanopore ultra-long reads across both benchmarks, and beating short read technologies with longer STR alleles. High accuracy long read technologies such as CCS should facilitate replication studies and new discoveries of associations between STR copy number and disease phenotype. I show that `STRkit` achieves low error and high genotyping accuracy on my benchmark datasets, comparing favourably to both other long- and short-read STR genotyping approaches, and is capable of genotyping pathogenic expansions implicated in tens of diseases affecting many people globally (Gall-Duncan *et al.*, 2021). My method, with CCS data, achieves performance parity with paired-end short reads in zygosity classification tasks and tracing Mendelian inheritance, outperforming all existing long read STR genotyping approaches and demonstrating capabilities applicable to future *de novo* STR mutation detection studies.

Currently, it is up to users of `STRkit` to perform downstream analyses with the generated callsets; however, in a more complete version of the software I aim to include

analyses pipelines to automate tasks that I foresee to be of interest: trio *de novo* variant detection, case-control cohort comparisons, and a genome and population-wide STR copy number distribution database for mutation and expansion detection. By incorporating these as software functions, I aim to assist researchers in more efficiently answering research questions related to STRs. To further this goal, I may need to push the genotyping accuracy of `STRkit` even higher – at genome-wide scales, greater statistical confidence becomes mandatory to mitigate the multiple testing problem. One way which I could augment the capabilities of my method using high-accuracy long reads would be to utilize genomic variation surrounding STR loci of interest to better resolve haplotypes and phase STR alleles with other forms of variation; allele peaks could then be better separated when close in copy number. `STRkit`'s computational performance could also be improved to allow for larger studies; short read STR genotypers outperform most existing long read ones due to a mix of programming language choice, implementation decisions, and limitations from the sequencing data itself; re-implementing parts of `STRkit` in a faster language could make it competitive with `GangSTR` and `ExpansionHunter` in terms of genotyping throughput.

I have implemented a visual motif composition comparison tool in `STRkit`'s `visualize` function (section 3.5), to contrast motif compositions between alleles in a specified locus. An extension of this could add statistical tests for case/control analyses of motif composition differences, paralleling my proposed case/control STR copy number analysis pipeline; this is another task where low-error long reads should solve limitations imposed by short read sequencing. STR motif compositional changes in a

non-coding part of the DAB1 gene cause spinocerebellar ataxia (Seixas *et al.*, 2017); this discovery could potentially be replicated using CCS and this proposed feature.

Another potential extension for `STRkit` is mosaicism detection, which could use an approach like the one employed by Chiu *et al.* (2021) with `Straglr`. Figure 9A shows an example of HTT mosaicism captured by CCS, confirmed by De Luca *et al.* (2021) with PCR analysis, in a locus of a size that should be difficult to resolve with short reads given my benchmark findings. Currently, `STRkit` calls either one or two peaks from read data for all autosomes. In `Straglr`, any number of peaks may be called; implementing a similar feature for `STRkit` could allow for automatic detection of loci with more than one expansion peak and potentially, using CCS, replicate findings such as Swami *et al.* (2009)'s implication of somatic expansion instability in Huntington's disease age of onset, or Pretto *et al.* (2014)'s discovery of an association between mosaicism in FMR1 and FXPOI phenotype, found using traditional techniques. This potential capability, and the realized potential shown in my other findings, stems directly from error rate and read length. When reads span an entire expansion, as long reads can, one read yields one copy number – but a complex expansion allele need not be represented entirely by a single copy number, opening the door for new types of genomic analyses and discoveries.

Chapter 6 References

1. Alisch, R. S., Wang, T., Chopra, P., Visootsak, J., Conneely, K. N., & Warren, S. T. (2013). Genome-wide analysis validates aberrant methylation in fragile X syndrome is specific to the FMR1 locus. *BMC Medical Genetics*, *14*(1), 18. https://doi.org/10.1186/1471-2350-14-18
2. Allen, E. G., Charen, K., Hipp, H. S., Shubeck, L., Amin, A., He, W., Nolin, S. L., Glicksman, A., Tortora, N., McKinnon, B., Shelly, K. E., & Sherman, S. L. (2021). Refining the risk for fragile X–associated primary ovarian insufficiency (FXPOI) by FMR1 CGG repeat size. *Genetics in Medicine*, *23*(9), Article 9. https://doi.org/10.1038/s41436-021-01177-y
3. Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, 1027–1035.
4. Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573
5. Blauw, H. M., van Rheenen, W., Koppers, M., Van Damme, P., Waibel, S., Lemmens, R., van Vught, P. W. J., Meyer, T., Schulte, C., Gasser, T., Cuppen, E., Pasterkamp, R. J., Robberecht, W., Ludolph, A. C., Veldink, J. H., & van den Berg, L. H. (2012). NIPA1 polyalanine repeat expansions are associated with amyotrophic lateral sclerosis. *Human Molecular Genetics*, *21*(11), 2497–2502. https://doi.org/10.1093/hmg/dds064

6. Brouwer, J. r., Willemsen, R., & Oostra, B. a. (2009). The FMR1 gene and fragile X-associated tremor/ataxia syndrome. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, *150B*(6), 782–798. https://doi.org/10.1002/ajmg.b.30910

7. Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Eichler, E. E., Korbel, J. O., Lee, C., Marschall, T., Devine, S. E., … Zody, M. C. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, *185*(18), 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

8. Chaisson, M. J. P., Huddleston, J., Dennis, M. Y., Sudmant, P. H., Malig, M., Hormozdiari, F., Antonacci, F., Surti, U., Sandstrom, R., Boitano, M., Landolin, J. M., Stamatoyannopoulos, J. A., Hunkapiller, M. W., Korlach, J., & Eichler, E. E. (2015). Resolving the complexity of the human genome using single-molecule sequencing. *Nature*, *517*(7536), 608–611. https://doi.org/10.1038/nature13907

9. Chiu, R., Rajan-Babu, I.-S., Friedman, J. M., & Birol, I. (2021). Straglr: Discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biology*, *22*(1), 224. https://doi.org/10.1186/s13059-021-02447-3

10. Corbett, M. A., Kroes, T., Veneziano, L., Bennett, M. F., Florian, R., Schneider, A. L., Coppola, A., Licchetta, L., Franceschetti, S., Suppa, A., Wenger, A., Mei, D., Pendziwiat, M., Kaya, S., Delledonne, M., Straussberg, R., Xumerle, L., Regan, B., Crompton, D., … Gecz, J. (2019). Intronic ATTTC repeat expansions in STARD7

in familial adult myoclonic epilepsy linked to chromosome 2. *Nature Communications*, *10*, 4920. https://doi.org/10.1038/s41467-019-12671-y

11. Daily, J. (2016). Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. *BMC Bioinformatics*, *17*(1), 81. https://doi.org/10.1186/s12859-016-0930-z

12. De Luca, A., Morella, A., Consoli, F., Fanelli, S., Thibert, J. R., Statt, S., Latham, G. J., & Squitieri, F. (2021). A Novel Triplet-Primed PCR Assay to Detect the Full Range of Trinucleotide CAG Repeats in the Huntingtin Gene (HTT). *International Journal of Molecular Sciences*, *22*(4), Article 4. https://doi.org/10.3390/ijms22041689

13. Dolzhenko, E., Bennett, M. F., Richmond, P. A., Trost, B., Chen, S., van Vugt, J. J. F. A., Nguyen, C., Narzisi, G., Gainullin, V. G., Gross, A. M., Lajoie, B. R., Taft, R. J., Wasserman, W. W., Scherer, S. W., Veldink, J. H., Bentley, D. R., Yuen, R. K. C., Bahlo, M., & Eberle, M. A. (2020). ExpansionHunter Denovo: A computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biology*, *21*(1), 102. https://doi.org/10.1186/s13059-020-02017-z

14. Dolzhenko, E., Deshpande, V., Schlesinger, F., Krusche, P., Petrovski, R., Chen, S., Emig-Agius, D., Gross, A., Narzisi, G., Bowman, B., Scheffler, K., van Vugt, J. J. F. A., French, C., Sanchis-Juan, A., Ibáñez, K., Tucci, A., Lajoie, B. R., Veldink, J. H., Raymond, F. L., … Eberle, M. A. (2019). ExpansionHunter: A sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics (Oxford, England)*, *35*(22), 4754–4756. https://doi.org/10.1093/bioinformatics/btz431

15. Dolzhenko, E., van Vugt, J. J. F. A., Shaw, R. J., Bekritsky, M. A., van Blitterswijk, M., Narzisi, G., Ajay, S. S., Rajan, V., Lajoie, B. R., Johnson, N. H., Kingsbury, Z., Humphray, S. J., Schellevis, R. D., Brands, W. J., Baker, M., Rademakers, R., Kooyman, M., Tazelaar, G. H. P., van Es, M. A., … Eberle, M. A. (2017). Detection of long repeat expansions from PCR-free whole-genome sequence data. *Genome Research*, *27*(11), 1895–1903. https://doi.org/10.1101/gr.225672.117

16. Duyao, M., Ambrose, C., Myers, R., Novelletto, A., Persichetti, F., Frontali, M., Folstein, S., Ross, C., Franz, M., Abbott, M., Gray, J., Conneally, P., Young, A., Penney, J., Hollingsworth, Z., Shoulson, I., Lazzarini, A., Falek, A., Koroshetz, W., … MacDonald, M. (1993). Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nature Genetics*, *4*(4), Article 4. https://doi.org/10.1038/ng0893-387

17. Ebert, P., Audano, P. A., Zhu, Q., Rodriguez-Martin, B., Porubsky, D., Bonder, M. J., Sulovari, A., Ebler, J., Zhou, W., Mari, R. S., Yilmaz, F., Zhao, X., Hsieh, P., Lee, J., Kumar, S., Lin, J., Rausch, T., Chen, Y., Ren, J., … Eichler, E. E. (2021). Haplotype-resolved diverse human genomes and integrated analysis of structural variation. *Science (New York, N.Y.)*, *372*(6537), eabf7117. https://doi.org/10.1126/science.abf7117

18. Ellegren, H. (2000). Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature Genetics*, *24*(4), Article 4. https://doi.org/10.1038/74249

19. Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics*, *5*(6), Article 6. https://doi.org/10.1038/nrg1348

20. Fairley, S., Lowy-Gallego, E., Perry, E., & Flicek, P. (2020). The International Genome Sample Resource (IGSR) collection of open human genomic variation resources. *Nucleic Acids Research*, *48*(D1), D941–D947. https://doi.org/10.1093/nar/gkz836

21. Fang, L., Liu, Q., Monteys, A. M., Gonzalez-Alegre, P., Davidson, B. L., & Wang, K. (2022). DeepRepeat: Direct quantification of short tandem repeats on signal data from nanopore sequencing. *Genome Biology*, *23*(1), 108. https://doi.org/10.1186/s13059-022-02670-6

22. Gall-Duncan, T., Sato, N., Yuen, R. K. C., & Pearson, C. E. (2021). Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences. *Genome Research*. https://doi.org/10.1101/gr.269530.120

23. Gemayel, R., Vinces, M. D., Legendre, M., & Verstrepen, K. J. (2010). Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics*, *44*, 445–477. https://doi.org/10.1146/annurev-genet-072610-155046

24. Genetic Modifiers of Huntington's Disease Consortium. (2019). CAG Repeat Not Polyglutamine Length Determines Timing of Huntington's Disease Onset. *Cell*, *178*(4), 887-900.e14. https://doi.org/10.1016/j.cell.2019.06.036

25. Ghebranious, N., Vaske, D., Yu, A., Zhao, C., Marth, G., & Weber, J. L. (2003). STRP Screening Sets for the human genome at 5 cM density. *BMC Genomics*, *4*(1), 6. https://doi.org/10.1186/1471-2164-4-6

26. Giesselmann, P., Brändl, B., Raimondeau, E., Bowen, R., Rohrandt, C., Tandon, R., Kretzmer, H., Assum, G., Galonska, C., Siebert, R., Ammerpohl, O., Heron, A., Schneider, S. A., Ladewig, J., Koch, P., Schuldt, B. M., Graham, J. E., Meissner, A., & Müller, F.-J. (2019). Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. *Nature Biotechnology*, *37*(12), Article 12. https://doi.org/10.1038/s41587-019-0293-x

27. Goldstein, D. B., Linares, A. R., Cavalli-Sforza, L. L., & Feldman, M. W. (1995). An Evaluation of Genetic Distances for Use with Microsatellite Loci. *Genetics*, *139*(1), 463–471.

28. Gymrek, M., Golan, D., Rosset, S., & Erlich, Y. (2012). lobSTR: A short tandem repeat profiler for personal genomes. *Genome Research*, *22*(6), 1154–1162. https://doi.org/10.1101/gr.135780.111

29. Gymrek, M., Willems, T., Guilmatre, A., Zeng, H., Markus, B., Georgiev, S., Daly, M. J., Price, A. L., Pritchard, J. K., Sharp, A. J., & Erlich, Y. (2016). Abundant contribution of short tandem repeats to gene expression variation in humans. *Nature Genetics*, *48*(1), Article 1. https://doi.org/10.1038/ng.3461

30. Halman, A., Dolzhenko, E., & Oshlack, A. (2022). STRipy: A graphical application for enhanced genotyping of pathogenic short tandem repeats in sequencing data. *Human Mutation*, *43*(7), 859–868. https://doi.org/10.1002/humu.24382

31. Hannan, A. J. (2018). Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews Genetics*, *19*(5), Article 5. https://doi.org/10.1038/nrg.2017.115

32. Igarashi, S., Tanno, Y., Onodera, O., Yamazaki, M., Sato, S., Ishikawa, A., Miyatani, N., Nagashima, M., Ishikawa, Y., & Sahashi, K. (1992). Strong correlation between the number of CAG repeats in androgen receptor genes and the clinical onset of features of spinal and bulbar muscular atrophy. *Neurology*, *42*(12), 2300–2302. https://doi.org/10.1212/wnl.42.12.2300

33. *Index of /public/dataset/RepeatExpansionDisorders_NoAmp*. (n.d.). Retrieved July 1, 2022, from https://downloads.pacbcloud.com/public/dataset/RepeatExpansionDisorders_NoAmp/

34. Ishiura, H., Doi, K., Mitsui, J., Yoshimura, J., Matsukawa, M. K., Fujiyama, A., Toyoshima, Y., Kakita, A., Takahashi, H., Suzuki, Y., Sugano, S., Qu, W., Ichikawa, K., Yurino, H., Higasa, K., Shibata, S., Mitsue, A., Tanaka, M., Ichikawa, Y., … Tsuji, S. (2018). Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. *Nature Genetics*, *50*(4), Article 4. https://doi.org/10.1038/s41588-018-0067-2

35. Jain, C., Rhie, A., Hansen, N. F., Koren, S., & Phillippy, A. M. (2022). Long-read mapping to repetitive reference sequences using Winnowmap2. *Nature Methods*, *19*(6), 705–710. https://doi.org/10.1038/s41592-022-01457-8

36. Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with

ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345. https://doi.org/10.1038/nbt.4060

37. Kalman, L., Johnson, M. A., Beck, J., Berry-Kravis, E., Buller, A., Casey, B., Feldman, G. L., Handsfield, J., Jakupciak, J. P., Maragh, S., Matteson, K., Muralidharan, K., Richie, K. L., Rohlfs, E. M., Schaefer, F., Sellers, T., Spector, E., & Richards, C. S. (2007). Development of genomic reference materials for Huntington disease genetic testing. *Genetics in Medicine*, *9*(10), Article 10. https://doi.org/10.1097/GIM.0b013e318156e8c1
38. Kang, L., Li, S., Gupta, S., Zhang, Y., Liu, K., Zhao, J., Jin, L., & Li, H. (2010). Genetic structures of the Tibetans and the Deng people in the Himalayas viewed from autosomal STRs. *Journal of Human Genetics*, *55*(5), Article 5. https://doi.org/10.1038/jhg.2010.21
39. Kedzierska, K. Z., Gerber, L., Cagnazzi, D., Krützen, M., Ratan, A., & Kistler, L. (2018). SONiCS: PCR stutter noise correction in genome-scale microsatellites. *Bioinformatics*, *34*(23), 4115–4117. https://doi.org/10.1093/bioinformatics/bty485
40. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102
41. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P., & Frith, M. C. (2011). Adaptive seeds tame genomic sequence comparison. *Genome Research*, *21*(3), 487–493. https://doi.org/10.1101/gr.113985.110
42. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K.,

Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., … The Wellcome Trust: (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), Article 6822. https://doi.org/10.1038/35057062

43. Leehey, M. A. (2009). Fragile X-associated tremor/ataxia syndrome: Clinical phenotype, diagnosis, and treatment. *Journal of Investigative Medicine: The Official Publication of the American Federation for Clinical Research*, *57*(8), 830–836. https://doi.org/10.2310/JIM.0b013e3181af59c4
44. Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191
45. Liu, Q., Tong, Y., & Wang, K. (2020). Genome-wide detection of short tandem repeat expansions by long-read sequencing. *BMC Bioinformatics*, *21*(21), 542. https://doi.org/10.1186/s12859-020-03876-w
46. Liu, Q., Zhang, P., Wang, D., Gu, W., & Wang, K. (2017). Interrogating the “unsequenceable” genomic trinucleotide repeat disorders by long-read sequencing. *Genome Medicine*, *9*(1), 65. https://doi.org/10.1186/s13073-017-0456-7
47. Matsuura, T., Yamagata, T., Burgess, D. L., Rasmussen, A., Grewal, R. P., Watase, K., Khajavi, M., McCall, A. E., Davis, C. F., Zu, L., Achari, M., Pulst, S. M., Alonso, E., Noebels, J. L., Nelson, D. L., Zoghbi, H. Y., & Ashizawa, T. (2000). Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nature Genetics*, *26*(2), Article 2. https://doi.org/10.1038/79911
48. McFarland, K. N., Liu, J., Landrian, I., Godiska, R., Shanker, S., Yu, F., Farmerie, W. G., & Ashizawa, T. (2015). SMRT Sequencing of Long Tandem Nucleotide

Repeats in SCA10 Reveals Unique Insight of Repeat Expansion Structure. *PLOS ONE*, *10*(8), e0135906. https://doi.org/10.1371/journal.pone.0135906

49. Mitra, I., Huang, B., Mousavi, N., Ma, N., Lamkin, M., Yanicky, R., Shleizer-Burko, S., Lohmueller, K. E., & Gymrek, M. (2021). Patterns of de novo tandem repeat mutations and their role in autism. *Nature*, *589*(7841), 246–250. https://doi.org/10.1038/s41586-020-03078-7

50. Mitsuhashi, S., Frith, M. C., Mizuguchi, T., Miyatake, S., Toyota, T., Adachi, H., Oma, Y., Kino, Y., Mitsuhashi, H., & Matsumoto, N. (2019). Tandem-genotypes: Robust detection of tandem repeat expansions from long DNA reads. *Genome Biology*, *20*(1), 58. https://doi.org/10.1186/s13059-019-1667-6

51. 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. https://doi.org/10.1093/nar/gkz501

52. Niehus, S., Jónsson, H., Schönberger, J., Björnsson, E., Beyter, D., Eggertsson, H. P., Sulem, P., Stefánsson, K., Halldórsson, B. V., & Kehr, B. (2021). PopDel identifies medium-size deletions simultaneously in tens of thousands of genomes. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-020-20850-5

53. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence

of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

54. Payseur, B. A., Place, M., & Weber, J. L. (2008). Linkage Disequilibrium between STRPs and SNPs across the Human Genome. *The American Journal of Human Genetics*, *82*(5), 1039–1050. https://doi.org/10.1016/j.ajhg.2008.02.018

55. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

56. Pemberton, T. J., Sandefur, C. I., Jakobsson, M., & Rosenberg, N. A. (2009). Sequence determinants of human microsatellite variability. *BMC Genomics*, *10*(1), 612. https://doi.org/10.1186/1471-2164-10-612

57. *Picard*. (2019). [Java]. Broad Institute. https://github.com/broadinstitute/picard

58. Pretto, D., Yrigollen, C. M., Tang, H.-T., Williamson, J., Espinal, G., Iwahashi, C. K., Durbin-Johnson, B., Hagerman, R. J., Hagerman, P. J., & Tassone, F. (2014). Clinical and molecular implications of mosaicism in FMR1 full mutations. *Frontiers in Genetics*, *5*, 318. https://doi.org/10.3389/fgene.2014.00318

59. Quilez, J., Guilmatre, A., Garg, P., Highnam, G., Gymrek, M., Erlich, Y., Joshi, R. S., Mittelman, D., & Sharp, A. J. (2016). Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Research*, *44*(8), 3750–3762. https://doi.org/10.1093/nar/gkw219

60. Rajan-Babu, I.-S., Peng, J. J., Chiu, R., Birch, P., Couse, M., Guimond, C., Lehman, A., Mwenifumbo, J., van Karnebeek, C., Friedman, J., Adam, S., Souich, C. D., Elliott, A., Lehman, A., Mwenifumbo, J., Nelson, T., van Karnebeek, C., Friedman, J., Li, C., … CAUSES Study. (2021). Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions. *Genome Medicine*, *13*(1), 126. https://doi.org/10.1186/s13073-021-00932-9

61. Robinson, J. T., Thorvaldsdóttir, H., Turner, D., & Mesirov, J. P. (2022). *igv.js: An embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV)* (p. 2020.05.03.075499). bioRxiv. https://doi.org/10.1101/2020.05.03.075499

62. 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), Article 1. https://doi.org/10.1038/s41467-018-06694-0

63. Seixas, A. I., Loureiro, J. R., Costa, C., Ordóñez-Ugalde, A., Marcelino, H., Oliveira, C. L., Loureiro, J. L., Dhingra, A., Brandão, E., Cruz, V. T., Timóteo, A., Quintáns, B., Rouleau, G. A., Rizzu, P., Carracedo, Á., Bessa, J., Heutink, P., Sequeiros, J., Sobrido, M. J., … Silveira, I. (2017). A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of DAB1, Mapping to SCA37, Causes Spinocerebellar Ataxia. *The American Journal of Human Genetics*, *101*(1), 87–103. https://doi.org/10.1016/j.ajhg.2017.06.007

64. Shortt, J. A., Ruggiero, R. P., Cox, C., Wacholder, A. C., & Pollock, D. D. (2020). Finding and extending ancient simple sequence repeat-derived regions in the human genome. *Mobile DNA*, *11*, 11. https://doi.org/10.1186/s13100-020-00206-y

65. Sutcliffe, J. S., Nelson, D. L., Zhang, F., Pieretti, M., Caskey, C. T., Saxe, D., & Warren, S. T. (1992). DNA methylation represses FMR-1 transcription in fragile X syndrome. *Human Molecular Genetics*, *1*(6), 397–400. https://doi.org/10.1093/hmg/1.6.397

66. Swami, M., Hendricks, A. E., Gillis, T., Massood, T., Mysore, J., Myers, R. H., & Wheeler, V. C. (2009). Somatic expansion of the Huntington's disease CAG repeat in the brain is associated with an earlier age of disease onset. *Human Molecular Genetics*, *18*(16), 3039–3047. https://doi.org/10.1093/hmg/ddp242

67. Tazelaar, G. H. P., Boeynaems, S., De Decker, M., van Vugt, J. J. F. A., Kool, L., Goedee, H. S., McLaughlin, R. L., Sproviero, W., Iacoangeli, A., Moisse, M., Jacquemyn, M., Daelemans, D., Dekker, A. M., van der Spek, R. A., Westeneng, H.-J., Kenna, K. P., Assialioui, A., Da Silva, N., Povedano, M., … van Es, M. A. (2020). ATXN1 repeat expansions confer risk for amyotrophic lateral sclerosis and contribute to TDP-43 mislocalization. *Brain Communications*, *2*(2), fcaa064. https://doi.org/10.1093/braincomms/fcaa064

68. Trost, B., Engchuan, W., Nguyen, C. M., Thiruvahindrapuram, B., Dolzhenko, E., Backstrom, I., Mirceta, M., Mojarad, B. A., Yin, Y., Dov, A., Chandrakumar, I., Prasolava, T., Shum, N., Hamdan, O., Pellecchia, G., Howe, J. L., Whitney, J., Klee, E. W., Baheti, S., … Yuen, R. K. C. (2020). Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature*, *586*(7827), Article 7827. https://doi.org/10.1038/s41586-020-2579-z

69. Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., & Posthuma, D. (2021). Genome-wide association

studies. *Nature Reviews Methods Primers*, *1*(1), Article 1. https://doi.org/10.1038/s43586-021-00056-9

70. Ummat, A., & Bashir, A. (2014). Resolving complex tandem repeats with long reads. *Bioinformatics (Oxford, England)*, *30*(24), 3491–3498. https://doi.org/10.1093/bioinformatics/btu437

71. Wagner, J., Olson, N. D., Harris, L., Khan, Z., Farek, J., Mahmoud, M., Stankovic, A., Kovacevic, V., Yoo, B., Miller, N., Rosenfeld, J. A., Ni, B., Zarate, S., Kirsche, M., Aganezov, S., Schatz, M. C., Narzisi, G., Byrska-Bishop, M., Clarke, W., … Zook, J. M. (2022). Benchmarking challenging small variants with linked and long reads. *Cell Genomics*, *2*(5), 100128. https://doi.org/10.1016/j.xgen.2022.100128

72. Wagner, J., Olson, N. D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., Hwang, Y.-C., Gupta, R., Wenger, A. M., Rowell, W. J., Khan, Z. M., Farek, J., Zhu, Y., Pisupati, A., Mahmoud, M., Xiao, C., Yoo, B., Sahraeian, S. M. E., Miller, D. E., … Sedlazeck, F. J. (2022). Curated variation benchmarks for challenging medically relevant autosomal genes. *Nature Biotechnology*, *40*(5), Article 5. https://doi.org/10.1038/s41587-021-01158-1

73. Weber, J. L., & Wong, C. (1993). Mutation of human short tandem repeats. *Human Molecular Genetics*, *2*(8), 1123–1128. https://doi.org/10.1093/hmg/2.8.1123

74. Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C.-S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome.

Nature Biotechnology, *37*(10), Article 10. https://doi.org/10.1038/s41587-019-0217-9

75. Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., & Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. *Nature Methods*, *14*(6), Article 6. https://doi.org/10.1038/nmeth.4267

76. Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., Weng, Z., Liu, Y., Mason, C. E., Alexander, N., Henaff, E., McIntyre, A. B. R., Chandramohan, D., Chen, F., Jaeger, E., Moshrefi, A., Pham, K., Stedman, W., Liang, T., … Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific Data*, *3*, 160025. https://doi.org/10.1038/sdata.2016.25

Appendix A: Supplementary data

Tech	Genotyper	Acc.	Acc. (95% CI)	RMSE	N
Ill. 150bp	GangSTR	93.1%	95.6%	1.74	77814
	GangSTR + Filter	94.1%	96.4%	0.78	76980
	ExpansionHunter	36.1%	36.6%	1.17	78578
Ill. 250bp	GangSTR	94.8%	95.8%	1.07	77618
	GangSTR + Filter	95.3%	96.2%	0.65	77092
	ExpansionHunter	36.2%	36.6%	2.04	78574
ONT-UL	RepeatHMM	N/A	N/A	N/A	N/A
	Tandem-genotypes	41.8%	N/A	3.10	78578
	Straglr	16.4%	N/A	3.37	78464
	STRkit	40.9%	81.3%	1.91	78578
HiFi	RepeatHMM	31.4%	N/A	3.00	76856
	Tandem-genotypes	89.0%	N/A	0.60	78572
	Straglr	31.6%	N/A	2.33	78198
	STRkit	**95.8%**	**97.9%**	**0.44**	78578

Table S5: Accuracy and root-mean-squared error (RMSE) by sequencing technology and STR genotyping software at 40-fold average depth of coverage.