Read-level genotyping of short tandem repeats using long reads and single-nucleotide variation with STRkit

David R Lougheed, Tomi Pastinen and Guillaume Bourque

Genome Res. published online January 15, 2026
Access the most recent version at doi:10.1101/gr.280766.125

Title: Read-level genotyping of short tandem repeats using long reads and single-nucleotide variation with STRkit

Authors: David R Lougheed[1,2]*, Tomi Pastinen[3], Guillaume Bourque[1,2,4]*

[1] Canadian Centre for Computational Genomics, Montréal, QC, Canada H3A 0G1
[2] Department of Human Genetics, McGill University, Montréal, QC, Canada H3A 0G1
[3] Genomic Medicine Center, Children's Mercy Hospital and Research Institute, KC, MO, USA 64108
[4] Victor Phillip Dahdaleh Institute of Genomic Medicine at McGill University, Montréal, QC, Canada H3A 0G1

*Corresponding authors: David R Lougheed (david.lougheed@mail.mcgill.ca) and Guillaume Bourque (guil.bourque@mcgill.ca)

Abstract

Variation in short tandem repeats (STRs) is implicated in Mendelian disease and complex traits, but can be difficult to resolve with short-read genome sequencing. We present STRkit, a software package for genotyping STRs using long read sequencing (LRS) that uses proximate single-nucleotide variants to improve genotyping accuracy without a priori haplotype information. We show that STRkit has unique strengths versus other methods: it can use data from both major LRS technologies (Pacific Biosciences HiFi [PB] and Oxford Nanopore [ONT]) to output both allele- and read-level copy number and sequence, performs best in benchmarking with F1 scores of 0.9631 and 0.9544 with PB and ONT data respectively, achieves higher rates of Mendelian consistency than other genotyping tools, and is open source software. STRkit's features open up new possibilities for association testing, assessing patterns of STR inheritance, and better understanding the functional effects of these notable repeat elements.

Keywords

long reads, short tandem repeats, genotyping, Mendelian diseases, association studies, software tool

INTRODUCTION

Short tandem repeats (STRs), or microsatellites, are repetitive genomic elements found in both eukaryotes and prokaryotes where a small motif, from 1 or 2 to 6-13 base pairs long (International Human Genome Sequencing Consortium 2001, Ellegren 2004, Chiu *et al.* 2021), is repeated several times in a row. These regions cover around 5% of the human genome (Horton *et al.* 2023), are often multi-allelic, and are significant contributors to total genomic variation (Ellegren 2004, English *et al.* 2024, Liao *et al.* 2023, Hannan *et al.* 2018). STRs typically have high rates of mutation, usually via stepwise repeat expansion or contraction (Fan and Chu 2007, Ellegren 2004).

Variation in STRs is associated with over 60 inherited and sporadic diseases in humans (Gall-Duncan *et al.* 2022). In Huntington's disease, fragile X syndrome (FXS), and others, copy number expansion beyond a threshold causes the disease, and further expansion is associated with phenotypic severity (Igarashi *et al.* 1992; Duyao *et al.* 1993; Brinkman *et al.* 1997; Matsuura *et al.* 2000; Blauw *et al.* 2012; Bragg *et al.* 2017). The relationship between copy number and disease severity is not always linear; for example, individuals with between 85 and 89 CCG copies (in the same repeat that causes FXS with ≥ 200 copies) have a higher risk of developing fragile X-associated

primary ovarian insufficiency (FXPOI) than those with more or fewer repeats in the FXPOI-pathogenic repeat range of 55 to 199 copies (Allen *et al.* 2021). STR motif composition can also affect phenotype: insertion of a stretch of a non-canonical motif into a non-coding STR in the *DAB1* gene is responsible for a form of spinocerebellar ataxia (Seixas *et al.* 2017). Beyond expansion detection, precisely determining copy number and motif composition in disease-causing STR expansions may be required to fully quantify their relationship with phenotype (Rajan-Babu *et al.* 2024).

Other phenotypic relationships with STRs are not limited to a single locus: autism is associated with a genome-wide increase in rare STR expansions (Trost *et al.* 2020) and small *de novo* STR mutations (Mitra *et al.* 2021). Expansion and somatic instability in STRs feature in several cancers (Erwin *et al.* 2022, Yamamoto and Imai 2019). STRs are implicated in gene expression and regulation (Gymrek *et al.* 2016; Hannan 2018; Cui *et al.* 2025), including by acting as binding sites for transcription factors (Horton *et al.* 2023). Direct STR genotyping at a genome scale is thus valuable; STRs cannot always be imputed using SNVs (Gymrek *et al.* 2016, Saini *et al.* 2018), as SNV-STR linkage disequilibrium (LD) is lower than SNV-SNV pair LD (Willems *et al.* 2014, Press *et al.* 2018). Examination of STR variation in the human genome may help to address the “missing heritability” problem, explaining a portion of complex trait heritability not yet identified (Hannan 2018).

However, STRs are difficult to characterize with short read sequencing (SRS) technologies due to often long repeat tracts and lack of flanking sequence within reads

(Narzisi & Schatz 2015, Mousavi *et al.* 2019), affecting mappability to a reference genome. Some SRS STR genotyping tools can estimate larger-than-read-length allele sizes (Dolzhenko *et al.* 2017, Dolzhenko *et al.* 2019, Mousavi *et al.* 2019) but cannot capture variation in motif structure nor precisely genotype these long expansions (Mousavi *et al.* 2019). By contrast, modern long read sequencing (LRS) technologies like Oxford Nanopore's nanopore sequencing (ONT) with newer chemistry, or Pacific Biosciences (PacBio) high-fidelity circular consensus sequencing (HiFi), can give accurate reads of sizes in the tens of kilobases (Gustafson *et al.* 2024; Tanudisastro *et al.* 2024; Wenger *et al.* 2019). These reads span pathogenic expansions that exceed SRS read length (Hannan 2018) and include more flanking sequence. These flanking sequences may include single-nucleotide variants (SNVs) that could be used to cluster STR-spanning reads into alleles, as SNVs are the most reliably-called form of variation with LRS (Wenger *et al.* 2019, Gustafson *et al.* 2024).

Several LRS STR genotyping tools have been recently developed, including Straglr (Chiu *et al.* 2021), TRGT (Dolzhenko *et al.* 2024), LongTR (Ziaei Jam *et al.* 2024), and STRdust (De Coster *et al.* 2024). These methods take standard BAM/CRAM alignments as input and output variants as VCF files. LongTR and TRGT have been tested on whole-genome catalogs of STRs, meaning they should be suitable for association testing and novel expansion discovery, among other whole-genome-scale applications. However, they do not take full advantage of LRS data, or have additional restrictions which limit their use – TRGT can use proximate SNVs to aid genotyping, but is license-restricted to only work with PacBio data; Straglr does not output consensus sequences

for alleles; and neither LongTR nor STRdust output read-level data, limiting their use for motif composition analysis or detecting within-allele somatic instability.

Here, we aim to develop and evaluate a new STR genotyping software, STRkit, which uses STR-proximate SNV information contained in long reads to accurately genotype and locally phase STRs, while giving researchers as much data as possible for downstream proband, trio, and population-level analyses, including STR consensus sequences and read-level STR copy number and sequence data. By using standard benchmarking tools and publicly available data from multiple long-read sequencing platforms, as well as a set of thirty trios sequenced using PacBio HiFi technology, we will evaluate our new software's effectiveness in capturing STR variation, including pathogenic expansions, and compare its genotyping and computational performance against other long-read STR genotyping tools.

RESULTS

The STRkit software package

We created an STR genotyping method, STRkit, as an easily-installable Python package. Our algorithm takes aligned long-read sequencing (LRS) BAM or CRAM files as input, with an optional realignment step for mis-mapped reads (**Figure 1A** and **Methods**). It then determines STR allele copy number and optionally genotypes proximate SNVs, whose locations are extracted from a user-provided VCF catalog file from, e.g., dbSNP (Sherry *et al.* 2001). For diploid alleles, depending on what SNV data are present, STRkit either uses a Gaussian mixture model allele-calling approach with copy numbers, segregates alleles based on SNV haplotype groups, or uses a mix of both information sources. STRkit can also genotype haploid alleles for, e.g., loci on the sex chromosomes of human XY or XO individuals. If two STRs have common SNV haplotype groups, they will be locally phased. If a phased alignment file is provided, STRkit can use read haplotype data instead. Genotypes are then reported in either VCF, TSV, or JSON formats. JSON output can be explored in an included web-based visualization tool (**Figure 1B**). Our software package also includes a Mendelian inheritance (MI) analysis tool, for benchmarking callers or detecting candidate *de novo* mutations. We compared the features of STRkit against those of other LRS STR genotyping packages to highlight its unique features, including the MI tool, SNV incorporation and output during STR genotyping, and read-level copy number output (**Table 1**).

Table 1: Feature comparison of long-read sequencing STR genotyping tools

Feature	STRkit	LongTR	TRGT	Straglr	STRdust	Notes
Copy number output	✓	-	✓	✓	-	
Allele size confidence intervals	✓	-	✓	✓	-	
Allele consensus sequence output	✓	✓	✓	-	✓	Straglr v1.5.2+ VCFs use symbolic alternate alleles.
Beyond read-length genotyping	-	-	-	Partial	-	Straglr can use reads that cover only part of a repeat to support alleles, but does not have a size model which incorporates in-repeat reads like some short-read STR genotypers.
De novo proximate SNV phasing	✓ + output	-	✓	-	-	STRkit and TRGT can use heterozygous SNVs to cluster STR alleles; STRkit can output them to VCF.
Existing phased SNV incorporation	-	✓✓	-	-	-	
Haplotagged alignment file support	✓	✓	✓	-	✓	All but Straglr can use phased read data from, e.g., WhatsHap (Martin *et al*. 2016) to call STRs.
Methylation handling	-	-	✓✓	-	-	
ONT read support	✓	✓	Explicitly forbidden	✓	✓	**TRGT:** Theoretically works on ONT data, but forbidden by software license.
Read-level data	✓	-	Partial	✓	✓	**STRkit:** JSON output with read-level peak ID + sequence data; **TRGT:** overlapping reads in BAM; **Straglr:** TSV output with read-level copy numbers; **STRdust:** read-level sequence output in VCF.
Mendelian inheritance calculation tool	✓✓	-	-	-	-	STRkit includes a tool to output loci which do not respect Mendelian inheritance in a set of trio VCFs.
Free and open-source software license	Yes (GPLv3)	Yes (GPLv2)	-	Yes (GPLv3)	Yes (MIT)	TRGT's license restricts it to be only used with PacBio sequencing data, and the software cannot be forked and subsequently re-distributed.
Multi-threading/processing	✓	-	✓	✓	✓	

Comparison of features of STRkit with those of recent long-read sequencing STR genotyping tools: LongTR, TRGT, Straglr and STRdust. Features with double check marks are unique to that particular tool.

Precise STR genotyping with long reads using proximate SNVs

To evaluate the performance of STRkit, we first obtained PacBio HiFi reads at ~32× coverage, ONT R10 simplex reads which we subsetted to ~32× coverage, and ONT R10 duplex reads at ~12× coverage for the Genome-in-a-Bottle (GIAB) HG002 individual from the GIAB Ashkenazi trio. We then aligned these reads to the HG38 reference genome (see **Methods**). Next, we genotyped ~900 000 short tandem repeat regions using STRkit running in two modes: one mode incorporating SNV data to help call and locally phase alleles, and one mode without SNVs or other haplotype tag data.

Our benchmark STR regions are derived from a ~1.7 million tandem repeat benchmark variant set developed for HG002 (English *et al.* 2024) and released as an official benchmark by the GIAB consortium. The full benchmark covers ~8.1% of the HG38 human reference genome, but not all of these benchmark regions are STRs; some have homopolymers or larger TRs, which we excluded to provide more precise estimates of performance (see **Methods**). We also removed regions with more than one tandem repeat because the tool we used to evaluate performance on the benchmark, Truvari (English *et al.* 2022), cannot handle proximate unphased variants. Our final set of 914 676 benchmark STRs covers ~0.87% of HG38.

To evaluate STRkit (with and without SNV incorporation) against other tools, we also genotyped our benchmark set of STR loci using LongTR (Ziaei Jam *et al.* 2024), Straglr (Chiu *et al.* 2021), STRdust (De Coster *et al.* 2024), and TRGT (Dolzhenko *et al.* 2024). All tools were configured to not use any prior phasing data from read haplotype tags or

158 phased sample SNVs. We then assessed their respective performance using Truvari
159 and Laytr (English *et al*. 2024); except for Straglr, whose output is not compatible with
160 Truvari. These results are summarized in **Table 2**. With HiFi data, STRkit with SNV
161 incorporation performed best overall in most metrics (accuracy = 0.9767, F1 = 0.9633),
162 although TRGT achieved marginally better recall (0.9697 versus 0.9552 for STRkit).
163 STRkit without SNVs performed second best in terms of precision, accuracy, and F1
164 score with HiFi. With ONT simplex data, we see similar rankings in caller performance,
165 with slightly worse absolute performance across the board versus HiFi. Here, STRkit
166 with SNVs (accuracy = 0.9709, F1 = 0.9544) performs best, followed by STRkit without
167 SNVs (accuracy = 0.9641, F1 = 0.9440), LongTR (accuracy = 0.9613, F1 = 0.9354),
168 and STRdust (accuracy = 0.8237, F1 = 0.7581). With low-coverage ONT duplex data,
169 we again lose absolute performance in metrics versus both higher-coverage sequencing
170 technology configurations, with the same tool ranking as with ONT simplex: STRkit with
171 SNV incorporation performs best (accuracy = 0.9435, F1 = 0.9051), followed by STRkit
172 without SNVs (accuracy = 0.9373, F1 = 0.8952) and then the other tools.
173
174 Truvari also reports a compound variant quality score, called the “TruScore”, combining
175 overlap, size similarity, and sequence similarity between a given call and the benchmark
176 into an overall score (English *et al.* 2022). With HiFi data, STRkit with SNVs achieved
177 the best average TruScore of 98.85, closely followed by LongTR and TRGT with 98.80,
178 STRkit without SNVs with 98.62 and beating STRdust’s average score (95.72) by a
179 larger margin. With ONT simplex and duplex data, we observe the same relative
180 ordering in TruScore.

Table 2: Performance metrics for long-read STR genotypers on the HG002 benchmark subset

Panel A: HG002, PacBio HiFi, ~32× coverage

Caller	Precision	Recall	Accuracy	F1	Avg. TruScore	% called
STRkit	0.9717	0.9545	0.9766	0.9631	98.85	99.88%
STRkit (-SNV)	0.9701	0.9487	0.9742	0.9593	98.62	99.87%
LongTR	0.9570	0.9541	0.9738	0.9555	98.80	99.63%
Straglr	*N/A*					97.18%
STRdust	0.7911	0.8790	0.8862	0.8327	95.72	99.84%
TRGT	0.9474	0.9675	0.9722	0.9574	98.80	99.87%

Panel B: HG002, ONT R10 Simplex, ~32× coverage

Caller	Precision	Recall	Accuracy	F1	Avg. TruScore	% called
STRkit	0.9594	0.9494	0.9709	0.9544	98.44	99.90%
STRkit (-SNV)	0.9488	0.9394	0.9641	0.9440	98.14	99.88%
LongTR	0.9312	0.9397	0.9613	0.9354	98.24	99.70%
Straglr	*N/A*					96.86%
STRdust	0.6701	0.8727	0.8237	0.7581	94.80	99.63%

Panel C: HG002, ONT R10 Duplex, ~12× coverage

Caller	Precision	Recall	Accuracy	F1	Avg. TruScore	% called
STRkit	0.9098	0.9004	0.9435	0.9051	96.03	99.78%
STRkit (-SNV)	0.8980	0.8924	0.9373	0.8952	95.66	99.76%
LongTR	0.8682	0.8893	0.9296	0.8786	95.71	99.59%
Straglr	*N/A*					96.86%
STRdust	0.7059	0.8460	0.8443	0.7697	94.11	99.63%

Performance metrics for long-read STR genotypers on an STR-only subset of HG002 variants regions from the Genome-in-a-Bottle tandem repeats v1.0 benchmark, as measured by Truvari: precision, recall, accuracy, F1 score, and Truvari's TruScore allele similarity metric. All metrics except TruScore are only for variants with an allele size difference of ≥5 base pairs versus HG38. Straglr VCFs are not compatible with Truvari, and TRGT cannot use ONT data due to a licensing restriction. Underlines indicate the best value for a metric within the sequencing technology.

Using the Laytr tool, we stratified performance metrics (F1 score, precision, recall) by the locus-maximum absolute change (Δ) in allele size from the reference genome (**Supplemental Fig. S1**). Most STR Δ allele sizes are small (95.5% of allele size changes are within 50 base pairs (bp) of the HG38 reference genome) and that is where STRkit with HiFi data shows a slight but measurable improvement over other tools in terms of F1 score and precision. With ONT simplex and duplex data for small and intermediate allele size changes versus the reference (within 200 bp of HG38), STRkit performed markedly better versus other tools. With larger Δ allele size, all tools performed worse with both sequencing technologies, with TRGT appearing to perform the best with these loci when it could be used (PacBio only). It was with these larger Δ allele sizes where STRkit's SNV incorporation helps its genotyping performance the most, although only in the case of the low-coverage ONT duplex data did STRkit with SNVs perform the best overall with medium-large Δ allele sizes (300-1000 bp).

Our evaluation of STRkit without the SNV incorporation step confirmed that using SNV data improves STR genotyping. All metrics output by Truvari for STRkit improved when the SNV step was included, for both HiFi and both types of ONT data (**Table 2**). For most loci in the call-set (~68-78% depending on individual and technology) STRkit used information from at least one heterozygous SNV for clustering reads into alleles (**Supplemental Table S1**).

Separately, we confirmed the accuracy of STRkit's proximate SNV calling and local phasing step by comparing SNV output to a phased small variant benchmark set

published by the Genome-in-a-Bottle consortium (Wagner *et al.* 2022, see **Methods**). Across the trio, we found an incorrect call rate of 0.008% (HiFi 32×), 0.025% (ONT simplex 32×), or 0.126% (ONT duplex 12×), considering only SNVs shared between the STRkit output and the benchmark. To look for phasing errors, we took phase sets with more than one SNV and compared phased SNV calls with the benchmark variant set. Here, we found a phase-set error rate (with one or more "flipped" SNVs) of 0.31% (HiFi), 0.34% (ONT simplex), and 0.51% (ONT duplex). These error rates are detailed in **Supplemental Table S2**.

High-quality genotypes track STR allele inheritance in trios

STRkit also includes a tool for calculating the rate of Mendelian inheritance (MI) of alleles from parents to offspring, in terms of copy number and allele sequence/length. Loci are considered to follow MI if allele sequences from the child can be found in the parents. This process also identifies loci that do not meet expectations of MI in the trio – sites of potential *de novo* mutation. Versions of this metric have been used to evaluate other STR genotyping software, e.g., GangSTR (Mousavi *et al.* 2019). On the Ashkenazi trio with HiFi data, STRkit achieves a sequence-wise (seq.) MI rate of 97.56% (97.98% in terms of seq. length, 99.04% in terms of seq. length ±1 base pair). Without SNVs, STRkit achieves a lower seq. MI rate of 93.64% (95.58% seq. len., 98.42% ±1bp). LongTR and TRGT again perform almost as well as STRkit with HiFi data; LongTR achieves a seq. MI rate of 94.52% (94.85% seq. len., 97.54% ±1bp), and TRGT achieves a seq. MI rate of 94.77% (97.19% seq. len., 98.98% ±1bp). On the same trio with ONT simplex data, rates of Mendelian inheritance are slightly lower

across the board, with STRkit (with SNVs) once again performing the best in seq. MI rate (94.44%), seq. length MI (94.99%), and seq. len. MI ±1bp (97.96%), closely followed by LongTR and STRkit without SNVs. With both sequencing technologies, incorporating SNVs to the read clustering process improves allele sequence quality with STRkit. With both HiFi and ONT simplex data, Straglr and STRdust achieve significantly lower rates of MI versus other tools. These results are summarized in **Supplemental Table S3**.

We then stratified sequence-wise MI by reference genome locus length (**Figure 2**). The bulk of loci are small, which is again where STRkit has the clearest advantage over other methods when incorporating SNVs with both HiFi and ONT simplex. Above ~200 base pairs in length, the rates drop off, as STRs become harder to sequence accurately and are more likely to accumulate mutations in their passage from parent to child.

Finally, to test the callers' performance beyond the benchmark trio, we examined MI in a set of 30 trios sequenced using PacBio HiFi data from the Genomic Answers for Kids cohort (Cohen *et al.* 2022). Most (26/30) trios tested were sequenced at low depth (<10×; parental coverage as low as 3×), two at medium depth (10-15×), and two at high depth (~37× in the proband, ~22× in the parents). Here, we see that STRkit (with or without SNVs) and TRGT outperform LongTR, STRdust, and Straglr (**Supplemental Fig. S2**), with TRGT appearing to perform slightly better than STRkit with copy number-wise MI in the two high-coverage trios, and no significant differences between STRkit and TRGT with low-coverage trios or with sequence/sequence-length-wise MI.

Computational performance of SNV-informed STR genotyping

Most genotyping tools tested (LongTR, STRkit, and TRGT) can analyze the filtered HG002 benchmark catalog in under a day measured in core-minutes using ~32× coverage PacBio HiFi / ONT simplex data, or ~12× ONT R10 duplex data (**Supplemental Table S4**). STRkit is slower than TRGT and LongTR, but is on the same order of magnitude (in the hundreds of core-minutes irrespective of technology); in contrast, Straglr is in the thousands of core-minutes, and STRdust is the slowest overall, requiring thousands of core-minutes for HiFi/low-coverage ONT and tens of thousands of core-minutes for the 32× ONT simplex samples. Where possible, we used 8 cores of an AMD EPYC 7532 processor, although LongTR does not support using multiple CPU cores. With HiFi data, TRGT was the fastest overall. For both types of ONT data, LongTR was fastest overall, although STRkit was the fastest program supporting multi-core parallelization. STRkit consumes more memory than TRGT or LongTR with HiFi data on average, but less than STRdust and Straglr, and the SNV incorporation step increases its memory usage. LongTR is the most memory-efficient STR caller for the GIAB trio samples sequenced with both ONT sequencing datasets.

Read-level genotyping of targeted long-read sequencing data captures known instability in a pathogenic expansion

We ran STRkit and other tandem repeat callers on targeted CCS sequencing data of pathogenic repeat expansions in the *HTT* and *FMR1* genes (**Supplemental Table S5** and **Methods**). We ran STRkit using its targeted sequencing mode, without SNV incorporation as the targeted sequencing data lacks sufficient flanking sequence to

make use of proximate SNVs. The callers successfully detected the repeat expansions in almost all cases, except for one *HTT* expansion missed by Straglr, and one *FMR1* expansion missed by STRdust. The genotypes overall were highly concordant.

Many of the tools tested also provide some degree of read-level output, allowing further expansion analysis. STRkit and Straglr both output read-level copy numbers, which we can use to examine an instance of somatic instability in this dataset. De Luca *et al.* (2021) found that one of the expansion samples tested, NA20253, has three copy number peaks, i.e., mosaic alleles, which they measured at 107, 134, and 175 CAG repeats, with the first being the most prominent. **Figure 3** shows the read-level data for this expanded allele from STRkit's output, along with annotations showing the peaks found by De Luca *et al.* (2021); we observe three read peaks which mirror those found using their triplet repeat-primed PCR assay, but slightly shifted towards a higher copy number. Evidence of these three peaks is also visible in the read-level copy number data output by Straglr.

STRkit also can output read-level motif-sized *k*-mer data, which can be used to explore non-canonical motif content in expansions. We used this output to visualize canonical and non-canonical motifs in *FMR1* expansion-containing samples, showing small differences in non-canonical motif content between samples (**Supplemental Fig. S3**).

DISCUSSION

Here, we present STRkit, a new tool for calling short tandem repeat genotypes from long read sequencing (LRS) data. Our tool has a unique combination of features, filling a niche in the STR genotyping analytical space: an STR caller which can operate at a genome-wide scale with both major LRS technologies, precisely genotype both sequence and copy number, optionally incorporate proximate SNVs into the calling process and use them to phase STRs locally, and report read-level data (**Table 1**). Both STR copy number and sequence are phenotypically relevant properties (Olson *et al.* 2023); however, LongTR and STRdust only output sequence, and Straglr (as of v1.5.2) outputs only copy number. Of the tested tools, all are LRS technology-independent except TRGT, which can only be used with PacBio data due to a licensing restriction. Technology-independent tools are critical, because they enable repeatable research and community-inclusive development, allow for confirmation of findings across technologies, and allow the best genotyping approach to be evaluated and applied regardless of data source – for example the cross-technology benchmarking we performed to evaluate STRkit. Our benchmarking, except for a set of trios used in our Mendelian inheritance analysis, also used open data, tools, and scripts (see **Methods**).

STRkit, when incorporating proximate SNVs into the genotyping process, demonstrates state-of-the-art performance in several benchmark metrics with both PacBio HiFi data and modern nanopore sequencing data (ONT R10 simplex and duplex). While performance gains are modest with HiFi data versus the next best-performing tool, LongTR (**Table 2**), maximizing precision and accuracy has direct benefits for genome-

wide analysis, where statistical power in correlation tests would be reduced by lower-quality genotypes. At similar depths of sequencing coverage, HiFi data outperforms ONT data, a pattern which has been observed in other long-read variant call benchmarking (Harvey *et al.* 2023). With ONT data, our performance gains were larger, achieving a ~2-3% higher F1 score (for high-coverage simplex and low-coverage duplex, respectively) than LongTR and narrowing the performance gap between ONT and HiFi data for calling STRs. Our SNV incorporation process allows clustering STR-overlapping reads into alleles without relying solely on copy number. This aids our genotyping approach to a small but measurable degree (**Table 2**), and for most loci we can incorporate SNV information into the STR calling process (**Supplemental Table S1**), although it may not be the primary explanation for STRkit's improved benchmark metrics versus other tools, especially with ONT data (**Table 2**, **Supplemental Fig. S1**). SNV data captured by STRkit may prove useful for more than genotyping; in downstream analysis, they could aid, for example, trio phasing or STR mutation modeling, although our current approach only calls heterozygous STR-proximate SNVs. When we assessed our SNV calls against a small variant benchmark for this trio, we found very low error rates in terms of incorrect calls or phasing errors (**Supplemental Table S2**), supporting their useability in downstream analysis. STRkit can also function without SNV data, which can be useful for targeted sequencing datasets where SNV-containing flanking regions are not present, or for reference genomes (human or otherwise) where an SNV catalog is not available. It also may be helpful for discovering low-coverage alleles (e.g., certain expansions), to avoid discarding reads without sufficient flanking information.

When metrics are faceted by allele size difference (Δ allele size) versus the reference genome, tools generally performed well until Δ allele size reached ~2.5k base pairs or longer (**Supplemental Fig. S1**), except for a noticeable dip in genotyping quality in the 3-500 base pair range. Manual inspection indicates that some of these may be regions with non-canonical motifs or non-tandem-repeat sequences embedded within a tandem repeat locus, which may impact STRkit's modeling of candidate repeat tract sequences or result in reads being filtered out to a greater degree versus other tools; future work could implement alternate allele clustering systems which do not assume an entirely tandem repeat sequence structure.

Of the five callers, STRkit, LongTR, and TRGT best accommodated the large catalog of ~900 000 loci, with relatively quick runtimes (in the 100s of core-minutes) and low memory requirements (**Supplemental Table S4**). Our general approach to genotyping does add a computational overhead, as it is slower to genotype the benchmark catalog versus LongTR and TRGT. We believe that this is due to our additional read-level data collection and copy-number determination approach; future performance improvements may be achievable from optimizing these procedures. STRdust and Straglr both had significantly longer runtimes (1000s of core-minutes), and Straglr used an order of magnitude more memory than the other tools. STRkit is the only one of the three most performant tools (in terms of both genotype quality and computation speed) which can report full read-level copy number data, meaning it may be useful for genome-wide somatic instability detection. LongTR's lack of parallel processing support means it

avoids overhead, at the cost of longer per-sample runtime, whereas other methods were evaluated using eight CPU cores, requiring inter-process communication. STRkit's shared SNV context, used for local phasing of STR genotypes, results in slower-than-linear performance gains with more CPU cores, trading off additional computation for increased genotype quality. Reducing the overhead of this inter-process communication could yield future performance improvements.

In our trio analysis for Mendelian inheritance (MI), we quantified the proportion of STR loci where both alleles in the HG002 child were found in the parents. STRkit had the best performance across the four versions of the metric we tested in this trio (**Supplemental Table S3**). We also analyzed rates of MI in a set of 30 trios sequenced with HiFi, most of which were sequenced at low depths of coverage (**Supplemental Fig. S2**); low coverage almost certainly resulted in allelic dropout, but in general STRkit and TRGT outperform the other tools tested here. MI is limited as a benchmarking metric, however, because systematic errors such as off-by-one errors which affect both parent and child may not be adequately considered. Despite this, our high MI rates, when combined with high absolute scores on the benchmark (which are less affected by these forms of systematic error, since they are compared to a truth set), indicate that STRkit should be the most accurate currently available tool for genome-scale STRs genotyping, and potentially for detecting the most common category of *de novo* STR mutation events: ±1 repeat unit changes. However, without a more complete mutation detection procedure with statistical confidence measures to control for false discovery, we cannot claim that all loci which do not respect MI are true *de novo* mutations. Future

benchmarking could also include more robust trio-aware approaches using, e.g., trio assemblies or trio-aware phasing data, or could involve developing simulated benchmark data, for example by incorporating an STR error model into a simulation tool like PBSIM3 (Ono *et al.* 2022).

So far, we have discussed genome-wide benchmarking results from ~900 000 tandem repeat loci. Very few STR loci have known pathogenic expanded forms – Gall-Duncan *et al.* (2022) reported 63 as of 2021. STRkit provides a unique feature-set useful for close examination of specific expansion loci. To demonstrate this, we used its targeted sequencing mode (which changes read-resampling behavior for use with targeted sequencing) to genotype STRs with known expansions in the *HTT* or *FMR1* genes (**Supplemental Table S5**), where, like other tools, it correctly detects the expanded allele. Unlike other tools, STRkit can both generate a whole-allele consensus sequence and report read-level copy number and sequence data, including motif-sized *k*-mer distributions (**Figure 1B**, **Supplemental Fig. S3**), at the same time. STRkit does not directly detect somatic instability, unlike a tool such as prancSTR (Sehgal *et al.* 2024); however, read-level copy number data (reported by STRkit and Straglr) replicate a multi-peak HTT mosaic expansion previously captured using a repeat-primed PCR technique by De Luca *et al.* 2021 (**Figure 3**). The peaks visible in the sequencing data are slightly offset from the reported peaks in this expansion, possibly due to the passage of the sampled cell line. This mosaic expansion revealed itself only with high-depth targeted sequencing and analysis of read-level data, demonstrating that read-level data can be crucial for characterizing expanded alleles.

Beyond somatic instability detection, there are other STR genotyping features STRkit does not yet implement. Methylation reporting using HiFi's base methylation data, for example, remains exclusive to TRGT. Our calling approach also has limitations: reads are primarily clustered based on motif copy number if SNV or haplotype information is not available; this is useful for obtaining copy number confidence intervals but may miss more subtle allelic sequence variation. In the absence of SNV data, our tool may also miss variation when faced with large allele differences versus the HG38 reference genome, to a greater degree than other methods (**Supplemental Fig. S1**). Every tool tested here, with the exception of Straglr, cannot genotype alleles beyond the length of a long read (**Table 1**); in the vast majority of cases, this will not pose issues, but this can cause reduced sequencing coverage or complete allele dropout with some very large expansions which have been observed in, e.g., individuals affected by SCA10 (Matsuura *et al.* 2000). This can be mitigated through the use of ultra-long-read sequencing technology, e.g., ultra-long nanopore reads (Jain *et al.* 2018); long-read genotyping tools could also employ longer-than-read-length allele-calling methods similar to those used by short-read STR genotyping tools like GangSTR (Mousavi *et al.* 2019) or ExpansionHunter (Dolzhenko *et al.* 2019). Finally, STRkit, like most of the other tools tested, is a catalog-based genotyper – it requires a list of loci with reference genome coordinates and a motif or IUPAC-coded motif pattern. This limits what can be genotyped to only what is already present in the reference, and may not function well with non-reference motifs within known loci or may miss certain disease-causing expansions completely (Tanudisastro *et al.* 2024). Annotating an assembly or using a

443 *de novo* approach such as ExpansionHunter Denovo (Dolzhenko *et al.* 2020) or Straglr
444 without a provided catalog can find repeats that STRkit and others cannot.
445
446 Short tandem repeats (STRs) cover around 5% of the human genome, are implicated in
447 genetic disorders and transcriptional regulation, and are major contributors to overall
448 genomic variation. STRs are important to genotype directly, as they cannot always be
449 imputed with single-nucleotide variation (SNVs), but these elements are difficult to
450 resolve with traditional sequencing methods and complex to analyze – they are multi-
451 allelic and defined by both their length and motif composition, rather than a simple
452 binary genotype. STRkit uses accurate long-read data and a strategy which
453 incorporates both SNVs and copy-number genotyping into the STR genotyping process
454 to achieve better performance in several metrics and producing local STR-SNV phasing
455 data. The tool also includes read-level copy number and motif composition data for STR
456 loci, while many other methods do not, enabling analysis of intra-allele variation in STR
457 copy number and motif composition. The approach fills a niche in the STR genotyping
458 software ecosystem and opens up new possibilities for association testing, assessing
459 patterns of STR inheritance, and better understanding the functional effects of these
460 repeat elements.

METHODS

STR genotyping approach

The general genotyping approach used in STRkit is as follows:

1. The user provides a BAM or CRAM-formatted long-read sequencing (LRS) alignment file, generated using an aligner such as minimap2 (Li 2018), as well as a BED-formatted catalog of STRs to genotype. Optionally, they also provide a VCF with a list of SNVs to genotype if proximate to an STR.
2. For each STR in the catalog, we optionally re-determine the copy number of the STR in the reference genome to allow for imperfect reference motif copies and adjust the catalog coordinates to correct for small boundary errors.
3. The set of all reads spanning each STR region (including flanking sequences on either side for realignment) is extracted using the *PySAM* library (v0.22.0; https://github.com/pysam-developers/pysam).
4. Low-quality or badly-aligned reads are filtered out. If realignment is enabled, badly-aligned reads may be realigned to the reference genome.
5. For each read, a series of candidate STR tracts of varying copy number, starting with a candidate closest in length to the aligned read's STR segment size, are generated using the cataloged motif. Each of these candidate STR tracts are aligned to the read using the *parasail* library's semi-global alignment algorithm (Daily 2016), incorporating 5' and 3' flanking sequences to 'anchor' the STR candidate and to allow for complete STR tract deletion. The best-scoring alignment candidate is kept for each read in question. After we have finished this

process, we have approximate copy numbers for each read. *If the read-level* k-*mer functionality is enabled, motif-sized* k-*mers are computed for each read.*

6. **If an SNV VCF catalog is provided:** For each read, SNVs that differ from the reference are collected; these are then filtered to those which occur in many reads, are of high base-level quality, and can split the reads into two groups, i.e., likely a heterozygous SNV and not a sequencing error.
7. 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. This process adds weight to reads with longer STR tracts. Given random read localization on a reference and a read size distribution, longer STR alleles are less likely to be entirely 'captured' by a read; in this process, allele size bias is corrected for, and long expansion tracts can be better captured. Reads (with associated copy numbers and SNV calls) are then resampled a number of times (defaulting to 100), according to the re-sampling weight.
8. From here, multiple possibilities are available to call allele peaks:
    a. **If the locus is haploid (e.g., on the Y sex chromosome of an XY individual):** A Gaussian distribution is fitted to the copy numbers.
    b. **If there is haplotype information in the alignment file and *--use-hp* is specified:** The haplotype information for the reads is used if available. Otherwise, one of the next three methods is used.
    c. **If there is a lot of informational SNVs for the locus:** SNVs are used to split the reads into allele peaks, since SNV assessment is generally very accurate in NGS data. A distance matrix is calculated, and an

agglomerative clustering function from the *scikit-learn* library (Pedregosa *et al.* 2011) is applied to group the reads into alleles.

d. **If there is some SNV information available:** SNV data are used, but copy number is incorporated into a compound distance function, where SNV differences are weighed more heavily (5:1 or 10:1) versus copy number differences. Agglomerative clustering is used here as well.

e. **If no or limited SNV information is available:** A Gaussian Mixture Model (GMM) from *scikit-learn* is applied to each re-sampling to derive estimates of allele copy numbers via a bootstrap process. Each allele starts with a two-peak GMM; if one peak is assigned very little explanatory weight after fitting, we switch to a single-peak GMM (i.e., a homozygous model).

If the organism is diploid, reads are assigned to peaks based on SNV genotypes (if available) or bootstrapped estimates of peak mean, standard deviation, and weight – the complete set of parameters characterizing a peak in a GMM. If the peaks overlap to a high degree and no informative SNV genotypes are available, reads are instead randomly assigned to peaks.

9. Final copy number genotype estimates, consensus sequences, and copy number confidence intervals are computed. *If the* k*-mer functionality is enabled, motif-sized* k*-mers are computed for each allele.*

10. A final report is generated with STR genotypes, parameters used, peak assignments (and peak assignment method used) for each read, copy numbers determined for each read, standard deviations for peaks, and optionally:

    a. Read-level SNV calls.

b. Read- and/or peak-level motif-size *k*-mer quantification.

Benchmarking genome-scale tandem repeat genotyping

We obtained unaligned HiFi LRS data for the HG002 individual from PacBio at https://downloads.pacbcloud.com/public/revio/2022Q4/. We obtained ONT R10 stereo duplex LRS data for HG002 from the Human Pangenomics Project at https://human-pangenomics.s3.amazonaws.com/index.html?prefix=submissions/0CB931D5-AE0C-4187-8BD8-B3A9C9BFDADE--UCSC_HG002_R1041_Duplex_Dorado/Dorado_v0.1.1/stereo_duplex/. We obtained ONT R10 simplex LRS data for HG002 from EPI2ME at https://42basepairs.com/browse/s3/ont-open-data/giab_2025.01/. We aligned all these read-sets to the UCSC HG38 analysis set reference genome (https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/analysisSet/hg38.analysisSet.fa.gz), using minimap2 v2.28, with the *map-hifi* and *lr:hq* presets for the HiFi and ONT R10 (simplex and duplex) data, respectively.

English *et al.* (2024) have published a genome-wide tandem repeat benchmark via the Genome-in-a-Bottle consortium (GIAB) for the HG002 Ashkenazim individual. We obtained version 1.0 of this benchmark from the NCBI at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/TandemRepeats_v1.0/GRCh38/ and corresponding benchmark region annotations from Zenodo at https://doi.org/10.5281/zenodo.6930201. This benchmark includes 1 784 804 regions, some with multiple tandem repeat loci defined in the same

region. The authors extended the Truvari software (English *et al.* 2022) to support benchmarking region-based STR calls. For benchmarking, we used Truvari version 5.3.0. We filtered out some complex or non-STR tandem repeat definitions from their benchmark region catalog, using the following steps:

1. Motifs were normalized (e.g., CACACA → CA).
2. Homopolymers and tandem repeats with motif length >10bp were removed.
3. Benchmark regions with more than one tandem repeat were removed, since unphased proximate haplotypes cannot be correctly evaluated with Truvari.

Additionally, homopolymers were filtered from the HG002 benchmark variant VCF to mitigate erroneous false negatives when homopolymers were present near other STR regions. To combine and visualize the benchmarking results after running Truvari, we used Laytr (commit *9cc1910*; https://github.com/ACEnglish/laytr). Applying this filtering, we ended up with a catalog of 914 676 loci.

For benchmarking STRkit, we chose four recently-developed general-purpose LRS STR genotyping tools, all of which take standard BAM/CRAM-formatted alignment files and can output VCFs, for use with Truvari. The set of tools, with their versions used for benchmarking, are as follows:

- LongTR v1.2 (Ziaei Jam *et al.* 2024)
- Straglr v1.5.5 (Chiu *et al.* 2021) with TRF v4.09.1 (Benson 1999)
- STRkit v0.23.0 (*our method, benchmarked with and without SNV incorporation*)
- STRdust (De Coster *et al.* 2024) v0.11.4 (commit *19c6ecd*)
- TRGT v4.0.0 (Dolzhenko *et al.* 2024)

The scripts that we used to benchmark these tools, alongside the parameters used for benchmarking, can be found in the tool batch scripts located at https://github.com/davidlougheed/strkit_paper/tree/main/2_giab_calls. This folder also contains scripts used to generate **Supplemental Fig. S1** and compute the summary data found in **Table 2** and supplemental tables.

To calculate the approximate STRkit SNV call and phasing error rates, we compared a) all SNV calls with a phased version of the small variant benchmark set for the HG002 individual (Wagner *et al.* 2022; obtained from https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NISTv4.2.1/GRCh38/SupplementaryFiles/HG002_GRCh38_1_22_v4.2.1_benchmark_hifiasm_v11_phasetransfer.vcf.gz), and b) multi-SNV phase sets with phased SNVs in this benchmark set. SNV errors were counted as either a “false heterozygote” (called by STRkit as heterozygous, but homozygous in the benchmark set) or “other errors” (non-matching calls). A phase set was counted as having an error if at least one SNV in the phase set had a flipped genotype versus the benchmark.

Assessing computational performance

All STR genotyping tools were run on the Digital Research Alliance of Canada’s Narval cluster, using 8 cores of an AMD EPYC 7532 processor, or 1 core where parallel processing was not supported (i.e., LongTR). Compute time was normalized into CPU-minutes by multiplying runtime by the number of cores used.

598 **Calculating Mendelian inheritance in call sets**

599 We obtained unaligned HiFi reads for the Ashkenazi trio from PacBio at
600 https://downloads.pacbcloud.com/public/revio/2022Q4/ and unaligned ONT R10 simplex
601 reads for this trio from EPI2ME at https://42basepairs.com/browse/s3/ont-open-
602 data/giab_2025.01/. We aligned both readsets to the UCSC HG38 analysis set
603 reference genome using minimap2 v2.28 with the *map-hifi* and *lr:hq* presets for HiFi and
604 ONT data, respectively.
605
606 We used trio data for 30 trios (with a mixture of low and high sequencing coverage)
607 from the Genomic Answers for Kids (GA4K) data to perform additional Mendelian
608 inheritance analysis. GA4K data are available under controlled access via dbGaP and
609 hosted on the AnVIL platform under the accession number phs002206.v5.p1.
610
611 In STRkit, we include a module named *strkit mi* to calculate rates of trio Mendelian
612 inheritance (MI) in autosomal loci from the output of all STR genotyping tools we
613 benchmarked. STRkit can calculate MI in terms of binary 'yes/no' copy number
614 inheritance, inheritance ±1 repeat unit, parent-offspring 95% confidence interval
615 overlap, exact sequence inheritance, sequence length, and sequence length ± 1 base
616 pair. For MI calculation, the tool performs the following steps:

617 1. For each locus in the STR callset for the child in the trio, check if a
618 corresponding call can be found in both parent callsets. If not, skip the locus.
619 2. If the locus is in a user-specified exclusion set, e.g., for removing known *de novo*
620 mutation, skip it.

3. Add this locus to the set $\mathbb{S}$ 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 $\mathbb{S}$.)
5. If the calls for the locus are consistent with Mendelian inheritance (for each version of the MI metric $\mathtt{X} \in \{\mathtt{strict}, \mathtt{\pm 1\ repeat}, \mathtt{95\%\ CI}, \mathtt{sequence}, \mathtt{seq.\ len.}, \mathtt{seq.\ len.\ \pm 1\ base\ pair}\}$), add one to a counter c_X of consistent loci for metric X. Some versions of the metric are not supported by some of the genotyping methods; LongTR and STRdust do not provide copy numbers or 95% copy number confidence intervals, and Straglr does not provide 95% copy number confidence intervals or consensus allele sequences.
6. After all loci have been processed, return the total rate of MI as $c/|\mathbb{S}|$.

Genotyping expansions from targeted sequencing data

We accessed a dataset of Pacific Biosciences targeted sequencing of expansions at https://downloads.pacbcloud.com/public/dataset/RepeatExpansionDisorders_NoAmp/. The same STR genotyping tool versions used in our benchmarking were compared with this targeted disease expansion sequencing dataset. We specified sex chromosome configurations for each sample to genotype the *FMR1* locus. Batch scripts with specific parameters, as well as the script used to generate **Figure 3**, are available at https://github.com/davidlougheed/strkit_paper/tree/main/4_pathogenic_exp. Coriell genotypes were manually collected from https://www.coriell.org/ Oct 4, 2022.

Software availability

The source code for the STRkit package is available on GitHub at: https://github.com/davidlougheed/strkit/ and in Zenodo under the DOI https://doi.org/10.5281/zenodo.12689906, and is obtainable as a Python package in the Python Package Index (PyPI) under the name *strkit*. The analysis scripts from this manuscript are available in GitHub at https://github.com/davidlougheed/strkit_paper/. STRkit version 0.23.0, used for the analyses in this manuscript, is available in **Supplemental Code**.

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

ACKNOWLEDGEMENTS

We would like to thank Rob Eveleigh and Mathieu Bourgey for providing valuable instruction on long read data and benchmarking approaches, José Héctor Gálvez López for his important feedback on our genotyping approach, Elia Afanasiev for his help increasing computational performance and for proofreading the manuscript, Boriana Apostolov and Stephen C. Lougheed for proofreading the manuscript, Nicolas Buitrago for his work on the Rust *parasail* bindings, and Adam C. English for his work and support with the Truvari and Laytr tools. We would like to acknowledge Calcul Quebec, SecureData4Health and the Digital Research Alliance of Canada for computing resources.

This work was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2024-05329) and the Canadian Institutes of Health Research (PJT-191707). G.B. is supported by a Canada Research Chair Tier 1 award and a FRQ-S, Distinguished Research Scholar award.

Author Contributions

D.R.L. developed the software package and performed the benchmarking analyses. G.B. and D.R.L. conceived of the project. T.P. gave guidance and insight during the development of the project. G.B. supervised the project. All authors read and approved the final manuscript.

REFERENCES

Allen EG, Charen K, Hipp HS, Shubeck L, Amin A, He W, Nolin SL, Glicksman A, Tortora N, McKinnon B, et al. 2021. Refining the risk for fragile X–associated primary ovarian insufficiency (FXPOI) by FMR1 CGG repeat size. *Genet Med* **23**: 1648–1655.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Blauw HM, van Rheenen W, Koppers M, Van Damme P, Waibel S, Lemmens R, van Vught PWJ, Meyer T, Schulte C, Gasser T, et al. 2012. NIPA1 polyalanine repeat expansions are associated with amyotrophic lateral sclerosis. *Human Molecular Genetics* **21**: 2497–2502.

Bragg DC, Mangkalaphiban K, Vaine CA, Kulkarni NJ, Shin D, Yadav R, Dhakal J, Ton M-L, Cheng A, Russo CT, et al. 2017. Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in TAF1. *Proc Natl Acad Sci U S A* **114**: E11020–E11028.

Brinkman RR, Mezei MM, Theilmann J, Almqvist E, Hayden MR. 1997. The likelihood of being affected with Huntington disease by a particular age, for a specific CAG size. *Am J Hum Genet* **60**: 1202–1210.

Chiu R, Rajan-Babu I-S, Friedman JM, Birol I. 2021. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. *Genome Biology* **22**: 224.

Cohen ASA, Farrow EG, Abdelmoity AT, Alaimo JT, Amudhavalli SM, Anderson JT, Bansal L, Bartik L, Baybayan P, Belden B, et al. 2022. Genomic answers for children: Dynamic analyses of >1000 pediatric rare disease genomes. *Genet Med* **24**: 1336–1348.

Cui Y, Arnold FJ, Li JS, Wu J, Wang D, Philippe J, Colwin MR, Michels S, Chen C, Sallam T, et al. 2025. Multi-omic quantitative trait loci link tandem repeat size variation to gene regulation in human brain. *Nat Genet* 1–10.

699 Daily J. 2016. Parasail: SIMD C library for global, semi-global, and local pairwise
700 sequence alignments. *BMC Bioinformatics* **17**: 81.

701 De Coster W, Hoijer I, Bruggeman I, D'Hert S, Melin M, Ameur A, Rademakers R. 2024.
702 Visualization and analysis of medically relevant tandem repeats in nanopore
703 sequencing of control cohorts with pathSTR. *Genome Res* gr.279265.124.

704 De Luca A, Morella A, Consoli F, Fanelli S, Thibert JR, Statt S, Latham GJ, Squitieri F.
705 2021. A Novel Triplet-Primed PCR Assay to Detect the Full Range of Trinucleotide
706 CAG Repeats in the Huntingtin Gene (HTT). *International Journal of Molecular*
707 *Sciences* **22**: 1689.

708 Dolzhenko E, Bennett MF, Richmond PA, Trost B, Chen S, van Vugt JJFA, Nguyen C,
709 Narzisi G, Gainullin VG, Gross AM, et al. 2020. ExpansionHunter Denovo: a
710 computational method for locating known and novel repeat expansions in short-read
711 sequencing data. *Genome Biology* **21**: 102.

712 Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, Emig-
713 Agius D, Gross A, Narzisi G, Bowman B, et al. 2019. ExpansionHunter: a sequence-
714 graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics*
715 **35**: 4754–4756.

716 Dolzhenko E, English A, Dashnow H, De Sena Brandine G, Mokveld T, Rowell WJ,
717 Karniski C, Kronenberg Z, Danzi MC, Cheung WA, et al. 2024. Characterization and
718 visualization of tandem repeats at genome scale. *Nat Biotechnol* 1–9.

719 Dolzhenko E, van Vugt JJFA, Shaw RJ, Bekritsky MA, van Blitterswijk M, Narzisi G,
720 Ajay SS, Rajan V, Lajoie BR, Johnson NH, et al. 2017. Detection of long repeat
721 expansions from PCR-free whole-genome sequence data. *Genome Res* **27**: 1895–
722 1903.

723 Duyao M, Ambrose C, Myers R, Novelletto A, Persichetti F, Frontali M, Folstein S, Ross
724 C, Franz M, Abbott M. 1993. Trinucleotide repeat length instability and age of onset
725 in Huntington's disease. *Nat Genet* **4**: 387–392.

726 Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev*
727 *Genet* **5**: 435–445.

728 English AC, Dolzhenko E, Ziaei Jam H, McKenzie SK, Olson ND, De Coster W, Park J,
729 Gu B, Wagner J, Eberle MA, et al. 2024. Analysis and benchmarking of small and
730 large genomic variants across tandem repeats. *Nat Biotechnol* 1–12.

731 English AC, Menon VK, Gibbs RA, Metcalf GA, Sedlazeck FJ. 2022. Truvari: refined
732 structural variant comparison preserves allelic diversity. *Genome Biology* **23**: 271.

733 Erwin GS, Gürsoy G, Al-Abri R, Suriyaprakash A, Dolzhenko E, Zhu K, Hoerner CR,
734 White SM, Ramirez L, Vadlakonda A, et al. 2022. Recurrent repeat expansions in
735 human cancer genomes. *Nature* **613**: 96.

736 Fan H, Chu J-Y. 2007. A Brief Review of Short Tandem Repeat Mutation. *Genomics*
737 *Proteomics Bioinformatics* **5**: 7–14.

738 Gall-Duncan T, Sato N, Yuen RKC, Pearson CE. 2022. Advancing genomic
739 technologies and clinical awareness accelerates discovery of disease-associated
740 tandem repeat sequences. *Genome Res* **32**: 1–27.

741 Gustafson JA, Gibson SB, Damaraju N, Zalusky MPG, Hoekzema K, Twesigomwe D,
742 Yang L, Snead AA, Richmond PA, De Coster W, et al. 2024. High-coverage
743 nanopore sequencing of samples from the 1000 Genomes Project to build a
744 comprehensive catalog of human genetic variation. *Genome Res*.

745 Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, Daly MJ, Price AL,
746 Pritchard JK, Sharp AJ, et al. 2016. Abundant contribution of short tandem repeats
747 to gene expression variation in humans. *Nat Genet* **48**: 22–29.

748 Hannan AJ. 2018. Tandem repeats mediating genetic plasticity in health and disease.
749 *Nat Rev Genet* **19**: 286–298.

750 Harvey WT, Ebert P, Ebler J, Audano PA, Munson KM, Hoekzema K, Porubsky D, Beck
751 CR, Marschall T, Garimella K, et al. 2023. Whole-genome long-read sequencing
752 downsampling and its effect on variant-calling precision and recall. *Genome Res* **33**:
753 2029–2040.

754 Horton CA, Alexandari AM, Hayes MGB, Marklund E, Schaepe JM, Aditham AK, Shah
755 N, Suzuki PH, Shrikumar A, Afek A, et al. 2023. Short tandem repeats bind
756 transcription factors to tune eukaryotic gene expression. *Science* **381**: eadd1250.

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**: 2300–2302.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Jain C, Rhie A, Hansen NF, Koren S, Phillippy AM. 2022. Long-read mapping to repetitive reference sequences using Winnowmap2. *Nat Methods* **19**: 705–710.

Kalman L, Johnson MA, Beck J, Berry-Kravis E, Buller A, Casey B, Feldman GL, Handsfield J, Jakupciak JP, Maragh S, et al. 2007. Development of genomic reference materials for Huntington disease genetic testing. *Genet Med* **9**: 719–723.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Liao W-W, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ, et al. 2023. A draft human pangenome reference. *Nature* **617**: 312–324.

Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, Schöenhuth A, Marschall T. 2016. WhatsHap: fast and accurate read-based phasing. 085050. https://www.biorxiv.org/content/10.1101/085050v2 (Accessed January 27, 2025).

Matsuura T, Yamagata T, Burgess DL, Rasmussen A, Grewal RP, Watase K, Khajavi M, McCall AE, Davis CF, Zu L, et al. 2000. Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. *Nat Genet* **26**: 191–194.

Mitra I, Huang B, Mousavi N, Ma N, Lamkin M, Yanicky R, Shleizer-Burko S, Lohmueller KE, Gymrek M. 2021. Patterns of de novo tandem repeat mutations and their role in autism. *Nature* **589**: 246–250.

Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. 2019. Profiling the genome-wide landscape of tandem repeat expansions. *Nucleic Acids Res* **47**: e90.

Narzisi G, Schatz MC. 2015. The Challenge of Small-Scale Repeats for Indel Discovery. *Frontiers in Bioengineering and Biotechnology* **3**: 8.

785 Olson ND, Wagner J, Dwarshuis N, Miga KH, Sedlazeck FJ, Salit M, Zook JM. 2023.
786 Variant calling and benchmarking in an era of complete human genome sequences.
787 *Nat Rev Genet* **24**: 464–483.

788 Ono Y, Hamada M, Asai K. 2022. PBSIM3: a simulator for all types of PacBio and ONT
789 long reads. *NAR Genom Bioinform* **4**: lqac092.

790 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M,
791 Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine Learning in
792 Python. *Journal of Machine Learning Research* **12**: 2825–2830.

793 Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D,
794 Dijamco J, Nguyen N, Afshar PT, et al. 2018. A universal SNP and small-indel
795 variant caller using deep neural networks. *Nat Biotechnol* **36**: 983–987.

796 Press MO, McCoy RC, Hall AN, Akey JM, Queitsch C. 2018. Massive variation of short
797 tandem repeats with functional consequences across strains of Arabidopsis thaliana.
798 *Genome Research* **28**: 1169.

799 Rajan-Babu I-S, Dolzhenko E, Eberle MA, Friedman JM. 2024. Sequence composition
800 changes in short tandem repeats: heterogeneity, detection, mechanisms and clinical
801 implications. *Nat Rev Genet* **25**: 476–499.

802 Saini S, Mitra I, Mousavi N, Fotsing SF, Gymrek M. 2018. A reference haplotype panel
803 for genome-wide imputation of short tandem repeats. *Nat Commun* **9**: 4397.

804 Sehgal A, Jam HZ, Shen A, Gymrek M. 2024. Genome-wide detection of somatic
805 mosaicism at short tandem repeats. *Bioinformatics* **40**: btae485.

806 Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001.
807 dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**: 308–
808 311.

809 Tanudisastro HA, Deveson IW, Dashnow H, MacArthur DG. 2024. Sequencing and
810 characterizing short tandem repeats in the human genome. *Nat Rev Genet* **25**: 460–
811 475.

Trost B, Engchuan W, Nguyen CM, Thiruvahindrapuram B, Dolzhenko E, Backstrom I, Mirceta M, Mojarad BA, Yin Y, Dov A, et al. 2020. Genome-wide detection of tandem DNA repeats that are expanded in autism. *Nature* **586**: 80–86.

Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, Stankovic A, Kovacevic V, Yoo B, Miller N, et al. 2022. Benchmarking challenging small variants with linked and long reads. *Cell Genom* **2**: 100128.

Wenger AM, Peluso P, Rowell WJ, Chang P-C, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* **37**: 1155–1162.

Willems T, Gymrek M, Highnam G, Mittelman D, Erlich Y. 2014. The landscape of human STR variation. *Genome Res* **24**: 1894–1904.

Yamamoto H, Imai K. 2019. An updated review of microsatellite instability in the era of next-generation sequencing and precision medicine. *Seminars in Oncology* **46**: 261–270.

Ziaei Jam H, Zook JM, Javadzadeh S, Park J, Sehgal A, Gymrek M. 2024. LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads. *Genome Biology* **25**: 176.

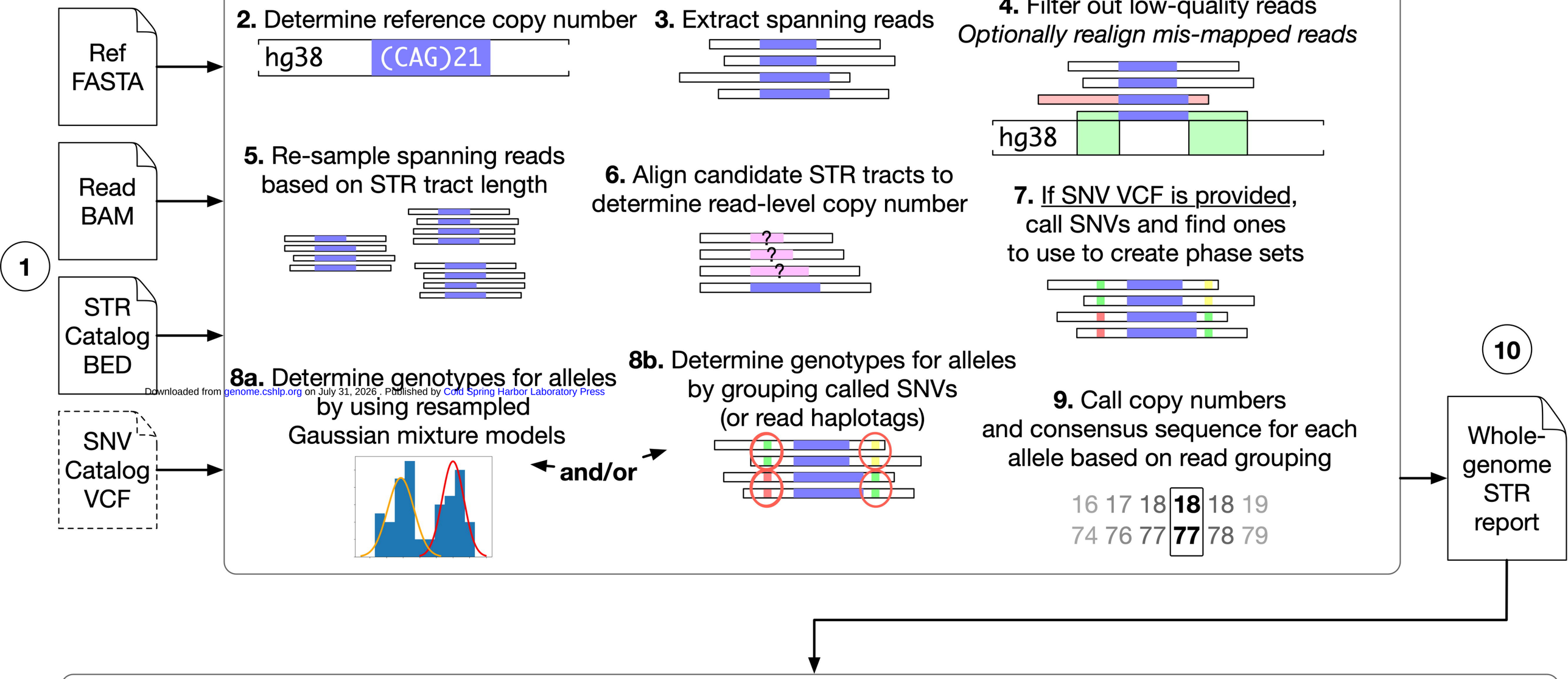
FIGURE LEGENDS

Figure 1: Flowchart of the STRkit genotyping and visualization workflow. **A.** The genotyping workflow (also see **Methods**). Our approach includes an optional SNV incorporation step, which can find heterozygous SNVs proximate to STRs and use them to cluster reads to call STR alleles. **B.** Users can explore the generated report in the STRkit visualization tool. The read count histogram shows the read-level distribution of copy numbers for the locus. Here, a pathogenic expansion in the *HTT* gene is shown. The *k*-mer distribution plots show motif-sized *k*-mer sequence diversity among read STR sequences for each allele peak.

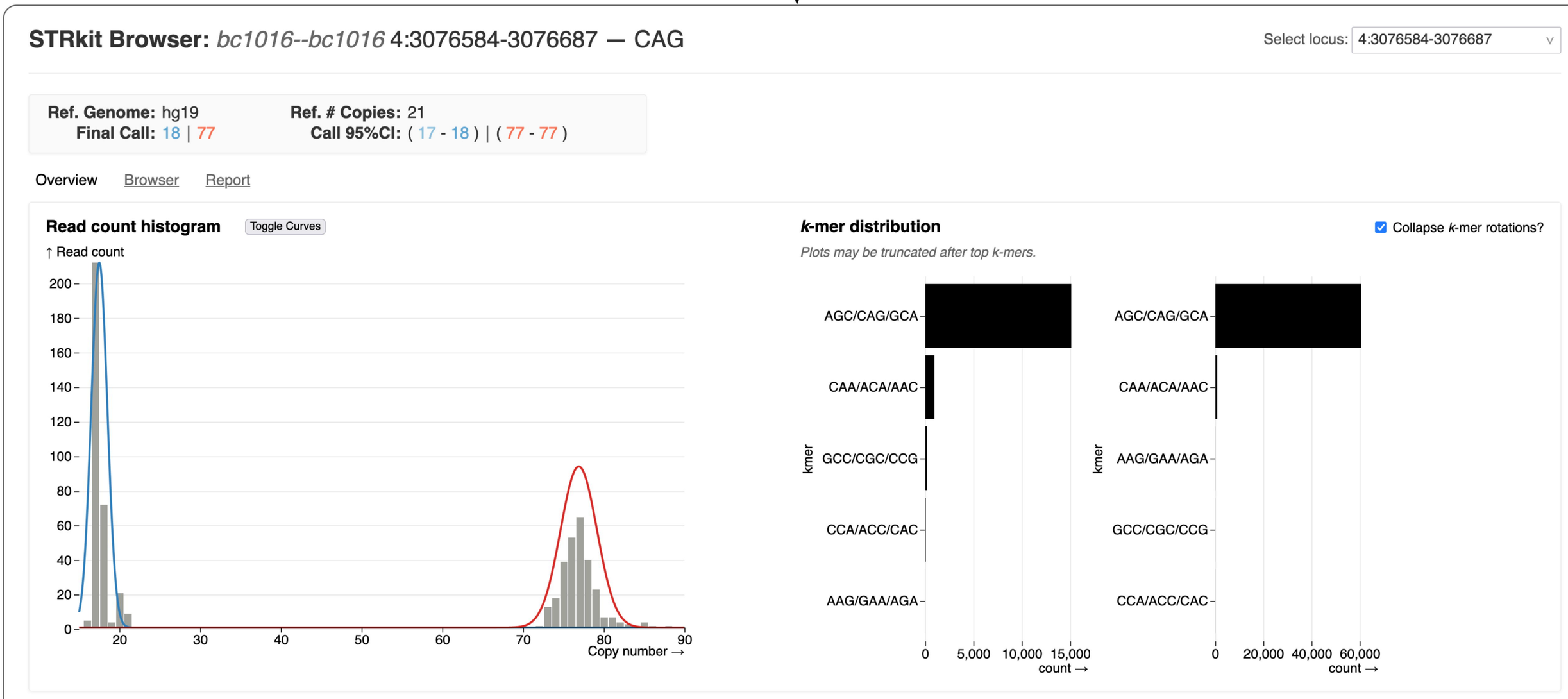
Figure 2: Rates of sequence Mendelian inheritance (MI), and number of trio-called loci, by reference genome locus length. Expectations of MI are met if both child allele sequences can be seen in the two parents (one in the mother and one in the father). The vast majority of short tandem repeat regions in the catalog are below ~200 base pairs long, where STRkit (with SNV incorporation) achieves the highest rate.

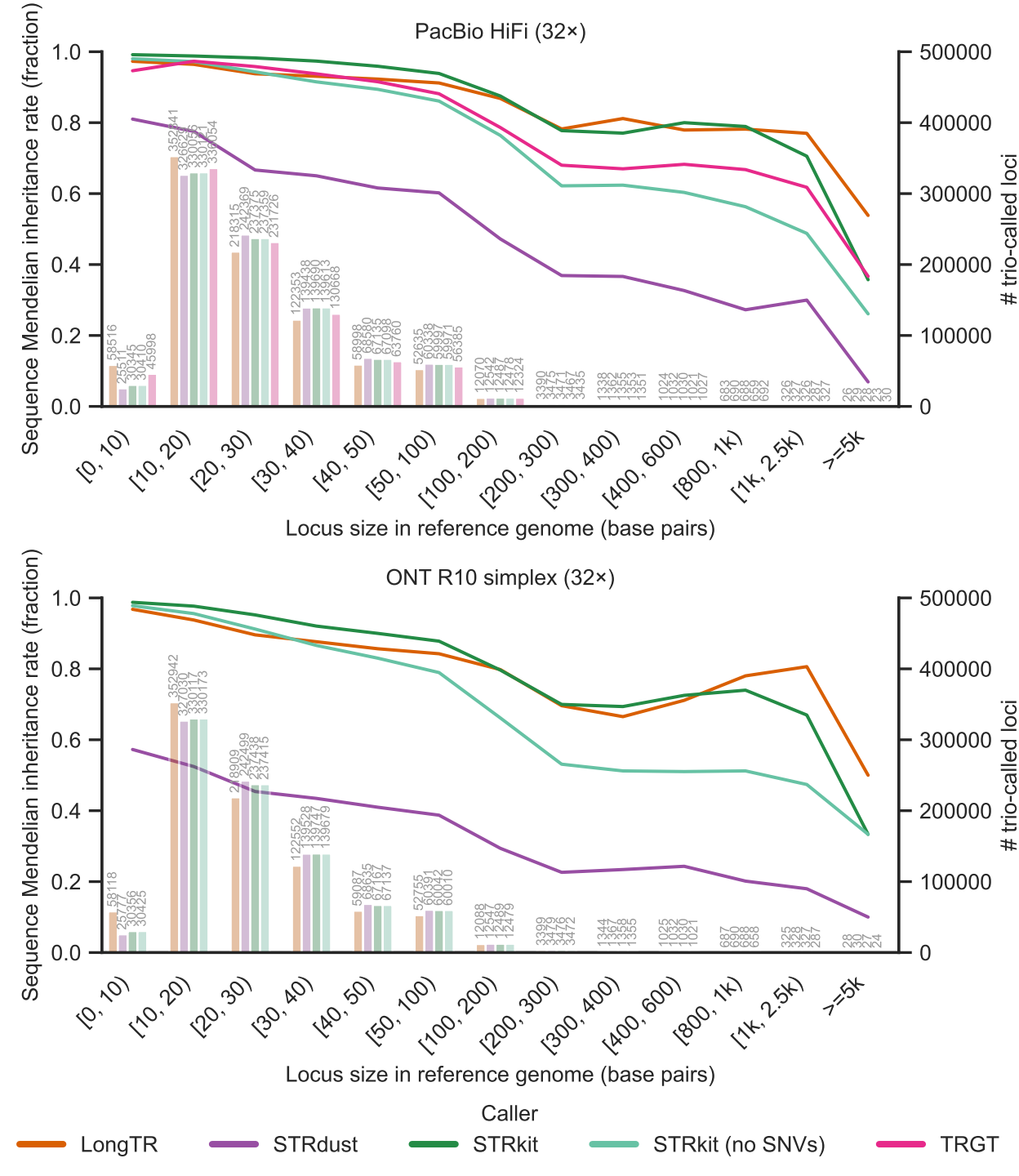
Figure 3: *HTT* HD-causing STR allele copy number distribution in targeted sequencing of the NA20253 sample. Vertical lines show the three expansion copy number peaks found by De Luca *et al.* (2021), which fall near read peaks in read-level output – two mosaic alleles, at 134 and 175 repeats, with the main allele peak (with the majority of reads) at 107 repeats.

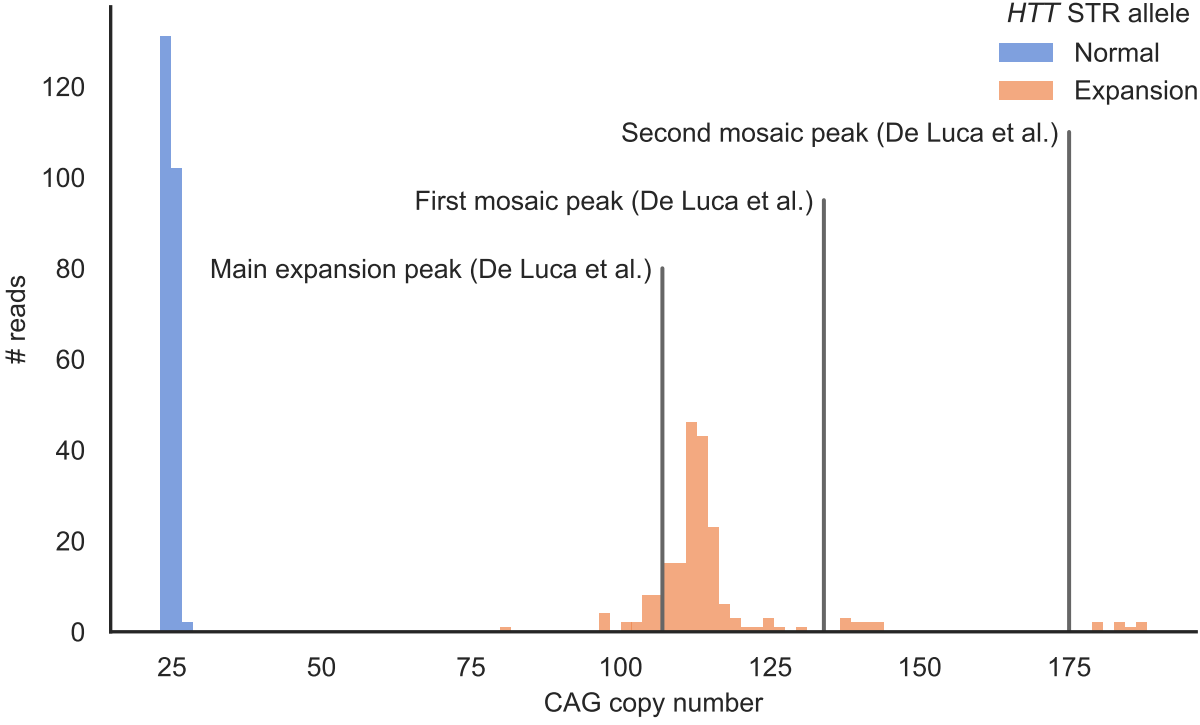
A



B




PacBio HiFi (32×)
Sequence Mendelian inheritance rate (fraction)
trio-called loci
1.0
0.8
0.6
0.4
0.2
0.0
500000
400000
300000
200000
100000
0
58516
25511
30345
30410
45998
352441
326065
330121
330173
330054
218315
242369
237375
237439
231726
122353
139438
139690
139613
130668
58998
69590
67093
67137
63760
52635
60338
59997
59971
56385
12070
12542
12487
12478
12324
3390
3475
3467
3472
3435
1338
1362
1355
1353
1351
1024
1032
1036
1021
1027
683
690
688
689
692
326
327
326
287
327
26
28
23
23
30
[0, 10)
[10, 20)
[20, 30)
[30, 40)
[40, 50)
[50, 100)
[100, 200)
[200, 300)
[300, 400)
[400, 600)
[800, 1k)
[1k, 2.5k)
>=5k
Locus size in reference genome (base pairs)
ONT R10 simplex (32×)
Sequence Mendelian inheritance rate (fraction)
trio-called loci
1.0
0.8
0.6
0.4
0.2
0.0
500000
400000
300000
200000
100000
0
58118
25777
30356
30425
352942
327017
330121
330173
248909
242499
237438
237415
122552
139528
139747
139679
59087
69035
67137
52755
60391
60042
60016
12088
12547
12489
12479
3399
3479
3476
3472
1344
1367
1358
1355
1025
1032
1036
1021
687
690
688
688
325
328
327
287
28
30
23
24
[0, 10)
[10, 20)
[20, 30)
[30, 40)
[40, 50)
[50, 100)
[100, 200)
[200, 300)
[300, 400)
[400, 600)
[800, 1k)
[1k, 2.5k)
>=5k
Locus size in reference genome (base pairs)
Caller
LongTR
STRdust
STRkit
STRkit (no SNVs)
TRGT


HTT STR allele
Normal
Expansion
Second mosaic peak (De Luca et al.)
First mosaic peak (De Luca et al.)
Main expansion peak (De Luca et al.)
reads
0
20
40
60
80
100
120
25
50
75
100
125
150
175
CAG copy number