A comparison of software for analysis of rare and common short tandem repeat (STR) variation using human genome sequences from clinical and population-based samples.

John W. Oketch (1), Louise V. Wain (2,3), Edward J. Hollox (1*)

1. Department of Genetics and Genome Biology, University of Leicester, Leicester UK
2. Department of Health Sciences, University of Leicester, Leicester UK
3. National Institute for Health Research, Leicester Respiratory Biomedical Research Centre, Glenfield Hospital, Leicester UK

* Corresponding author:
Dr E Hollox,
Department of Genetics and Genome Biology,
University of Leicester,
Leicester
UK
ejh33@le.ac.uk

Keywords

Short tandem repeat, STR, Genome-In-a-bottle, sequencing, genotyping, forensics

Abstract

Short tandem repeat (STR) variation is an often overlooked source of variation between genomes. STRs comprise about 3% of the human genome sequence, are highly polymorphic, and are known to cause not only Mendelian disease, but extensively affect gene expression. Nevertheless, their contribution to common disease is not well-understood, but recent software tools designed to genotype STRs using short read sequencing data are beginning to address this. Here, we compare software that genotypes common STRs and detect rarer STR expansions genome-wide, with the aim of applying them to population-scale genome sequences. By using the Genome-In-A-Bottle (GIAB) consortium and 1000 Genomes Project sequencing data, we can compare performance in terms of sequence length, depth, computing resources needed, genotyping accuracy and number of STRs genotyped. To ensure broad applicability of our findings, we also measure performance against a set of clinical samples with known STR expansions, and a set of STRs commonly used for forensic identification. By using analysis of Mendelian inheritance patterns and comparison with capillary electrophoresis genotypes, we find that both HipSTR and GangSTR perform well in genotyping common STRs, with GangSTR outperforming HipSTR for genotyping call rate and memory usage. Analysis for expanded STRs showed ExpansionHunter denovo (EHdn) and STRetch outperformed GangSTR, but EHdn used considerably less processor time and memory compared to STRetch. Analysis on shared genomic sequence data provided by the GIAB consortium allows future performance comparisons of new software approaches on a common set of data, facilitating comparisons and allowing researchers to choose the best software that fulfils their needs.

Introduction

Over the past two decades, most research on the contribution of genetic variation in the human genome to disease has focused on single nucleotide variation. Short tandem repeat (STR) variation has been generally overlooked, as it is not readily assayed by chip-based hybridisation approaches. STRs are short DNA sequence motifs, typically 2-6bp in size, that are repeated five or more times in tandem. They are distributed throughout the human genome, comprising about 3% of the genome sequence, and have a mutation rate generally much higher than single nucleotide variants (around 2x10-3 per locus per generation for STRs compared to 10-8 for single nucleotide variants) (Brinkmann et al., 1998; Burgarella and Navascués, 2011; Willems et al., 2014; Gymrek, 2017). Because of this high mutation rate, they are frequently polymorphic with multiple alleles in a population.

STRs cause a variety of severe rare monogenic diseases that are inherited in a Mendelian manner (Hannan, 2018). These include triplet-repeat expansion diseases, such as Huntington's disease and Myotonic dystrophy, caused by expansion of 3 bp repeat STRs in the coding region of *HTT*, and *DMPK* genes respectively. Some disorders also occur with other repeat motifs in both coding and in non-coding regions (Hannan, 2018; Paulson, 2018; Lieberman et al., 2019; van Kuilenburg et al., 2019). Emerging evidence shows STRs to play a role in modulating disease risk in several complex diseases such as autism and neurodegenerative disorders (Hannan, 2020; Trost et al., 2020).

STRs can affect the function of genes in two main ways, either by directly altering the coding region of genes or by affecting expression levels as part of enhancer elements. For example, in addition to STRs encoding coding CAG-repeats leading to polyglutamine tracts in genes such as *HTT*, polymorphic STRs also encode polyalanine tracts of varying length in a variety of genes (Lavoie et al., 2003). It is well established that STR variation is associated with expression levels of genes both at a single gene level and at a genome-wide level (Contente et al., 2002; Gymrek et al., 2016; Fotsing et al., 2019). It is unclear the extent to which STRs are responsible for genemowide association signals detected using SNPs. Linkage disequilibrium between STRs and SNPs varies, but is generally lower than between two SNPs at the same recombination distance because the high mutation rate of STRs breaks down linkage disequilibrium. This is shown in attempts to impute STR length genotypes from SNP genotypes, as biallelic STRs can generally be reliably imputed but multiallelic STRs are much less reliably imputed (Saini et al., 2018).

Given the potential importance of STRs in human disease, genome-wide studies examining STR variation are clearly warranted. The first generation of STR calling software, including LobSTR (Gymrek et al., 2012), RepeatSeq (Highnam et al., 2013) and HipSTR (Willems et al., 2017), aimed to genotype STRs using reads that completely covered the STR. This led to the limitation that STRs longer than the sequencing read length could not be genotyped. The second generation of STR calling software uses information from the distance between the paired sequencing reads, in addition to direct sequence information across the repeat, to genotype the locus. This attempts to capture information on STRs that are longer than the

sequencing read length, and is particularly effective at identifying large expansions at an STR (Bahlo et al., 2018).

Two recent pieces of software are focused on calling large expansions: Expansion Hunter de Novo (EHdn) (Dolzhenko et al., 2020) and STRetch (Dashnow et al., 2018). Both use sequence reads that span STRs and mate pair distance information to identify large expansions, but the key difference between the two is that EHdn does not require a predefined catalogue of STRs to genotype, unlike STRetch. We also compare two pieces of software that focus on calling all known STRs across the genome, HipSTR (Willems et al., 2017) and GangSTR (Mousavi et al., 2019). These both require a predefined catalogue of known STRs but are distinguished by the fact that GangSTR incorporates mate pair distance information as well as STR-spanning sequence reads but HipSTR does not. GangSTR also attempts to identify large expansions, and was also compared against EHdn and STRetch. The software tools also differ in the repeat unit sizes they can detect. GangSTR and EHdn can genotype repeats (longer than read length) of 1-20 bp repeat units, while HipSTR and STRetch are restricted to 1bp-6bp repeat units.

Our motivation for this study was to identify software to genotype common STRs and identify rarer STR expansions on a large dataset suitable for biobank-scale data analysis. We assess four software tools in their ability to genotype known common STRs and identify rarer STR expansions, and their applicability to studies involving large datasets, with a view to identifying an approach to investigate the role of STRs in human diseases. We use different sequence coverages, read lengths and genomes from different populations to estimate practical parameters with broad applicability, such as run time, memory usage and proportion of STRs called. In most analyses, establishing a truth set based on real data, rather than simulated data, is challenging, and therefore measurement of error rate is not possible. We use both Mendelian inheritance in parent-offspring trios and comparison with capillary electrophoresis to measure the genotyping error rate of different approaches. Our genome data includes those from the Genome in a Bottle consortium and 1000 Genomes project, collections of publicly available samples analysed by different methods ensuring data can be cross validated as new technical approaches become available (Zook et al., 2016).

Methods

Samples and datasets analysed

An overview of the approaches used in this paper is shown in Figure 1.
The four tools were benchmarked by analysing three different datasets. Firstly, seven gold standard human genome sequences from the Genome in a Bottle consortium (https://github.com/genome-in-a-bottle/giab_data_indexes), secondly, four child-parent trios from the 1000 Genomes project (https://www.internationalgenome.org/data) and thirdly, 95 samples with known trinucleotide repeat mutations obtained from the Coriell institute (https://ega-archive.org/datasets/EGAD00001003562). The accession numbers for samples obtained from the Genome in a Bottle(GIAB) and the 1000G Project are listed in **Table 1**. The Coriell institute samples are available at the European Genome-phenome Archive database. These sequences had been aligned to the GRCh38 human genome assembly. The collated dataset was subsequently analysed as described below.
95 whole genomes including previously validated disease-causing STR expansions obtained from Coriell Institute (Dolzhenko et al., 2017). These samples have been validated with one the following repeat expansions: Fragile X Syndrome (FMR1), Huntington disease (HTT), Friedreich's ataxia (FXN), Myotonic Dystrophy (DM1), Spinocerebellar Ataxia 1/3 (ATXN1/3), Spinal and Bulbar Muscular Atrophy (SBMA). The samples were sequenced at 2x150 bp reads on Illumina HiSeqX and repeat expansions detected using ExpansionHunter and standard PCR techniques (Dolzhenko et., al 2017). These sequences were aligned to the human GRCh37 human genome assembly.

Alignment and bam preparation

Preliminary processing was done to index the human reference genome (hg38). All cram files were converted to bam files using samtools v1.9 (Li et., al 2009). The seven GIAB bam files were downsampled from their original coverage of either 300x or 100x to a final coverage of ~30x. In summary, a total of 5/7 GIAB samples had an intial genome coverage of ~300x and 2/7 had a genome coverage of ~100x. The 300x genomes were downsampled to 100x and further all samples downsampled to ~30x genome coverage. The bam files were refined to remove sequence duplicates and the quality of mapping was assessed using qualimap tool v2.2.1 (Okonechnikov et., al 2016). The downsampled 30x genomes were collated with 12 30x genomes obtained from the 1000G Project. Overall, we generated 3 datasets: 300x (n = 5), 100x (n = 7) and 30x (n =19) genomes, consisting of 6 child-parent trios (Table 1). All the bam files were sorted and indexed before calling for STR genotypes. The final 30x coverage was chosen to assess the applicability of STR genotyping tools on large cohorts of genomes sequenced at 30x genome coverage. STR genotyping was performed at the three levels of genome coverage and calls compared across coverage and between the tools.

Table 1 - Samples used in this study

Sample ID	Population	Trio	Collection	Average fold sequence coverage	Paired end Read length (bp)	Accession number
HG002	Askanazi	Son	GIAB	30, 100, 300	150	NA24385
HG003	Askanazi	Father	GIAB	30, 100, 300	150	NA24149
HG004	Askanazi	Mother	GIAB	30, 100, 300	150	NA24143
HG005	Chinese	Son	GIAB	30, 100, 300	250	NA24631
HG006	Chinese	Father	GIAB	30, 100	150	NA24694
HG007	Chinese	Mother	GIAB	30, 100	150	NA24691
NA12878	CEU	-	GIAB	30, 100, 300	150	NA12878
HG00403	CHS	Father	1000G	30	150	HG00403
HG00404	CHS	Mother	1000G	30	150	HG00404
HG00405	CHS	Son	1000G	30	150	HG00405
NA18485	YRI	Son	1000G	30	150	NA18485
NA18487	YRI	Father	1000G	30	150	NA18487
NA18489	YRI	Mother	1000G	30	150	NA18489
HG01500	IBS	Father	1000G	30	150	HG01500
HG01501	IBS	Mother	1000G	30	150	HG01501
HG01502	IBS	Son	1000G	30	150	HG01502
NA06984	CEU	Father	1000G	30	150	NA06984
NA06989	CEU	Mother	1000G	30	150	NA06989
NA12329	CEU	Daughter	1000G	30	150	NA12329

GIAB=Genome in a Bottle
1000G=1000 Genomes Project

STR calling

GangSTR v.2.5.4 was run using default parameters. The raw STR calls were filtered using the DumpSTR tool as recommended by the authors (Mousavi et al 2019). HipSTR v.0.6.2 was run using some non-default parameters (see supplementary methods). We used reference panels recommended by the software providers (https://github.com/HipSTR-Tool/HipSTR and https://github.com/gymreklab/GangSTR), which are of very different sizes, with GangSTR listing 832379 STR loci and HipSTR listing 1,638,944 loci. These catalogues have been previously tested and provided by each tool.

A STRetch custom reference panel was created using the STRetch pipeline (https://github.com/Oshlack/STRetch/wiki/Reference-Data). STRetch was run using a whole genome sequencing pipeline (STRetch_wgs_bam_pipeline.groovy) with the default settings. In this study, samples were compared against each other at each loci at the different levels of genome coverage (i.e. 30x,100x and 300x). The putative expanded loci (p value <0.05) were tabulated for each sample. STR calls were filtered to remove 1 bp repeat unit that are prone to sequencing errors.

EHdn does not require a predefined catalogue of STR genomic coordinates, instead EHdn scans for expanded repeats longer than the read length or fragment length (mate-pair length) by searching for two types of STR reads: (a) Anchored In Repeat Reads (IRR) whose mates confidently align to genomic regions adjacent to the repeat region and (b) IRR where both mates bind within an STR (Dolzhenko et. al 2020). The read counts are compared and Z-scores are calculated for a case sample with counts that exceed the mean distribution. Putative expanded loci with z-score greater than 1.0 are reported. To explore STR expansion profiles in each sample, the sample to be examined was treated as a case sample and the rest of the samples as controls, and performed outlier analysis. Putative expanded loci were recorded for each sample.

To assess the sensitivity of GangSTR, EHDn and STRetch at detecting STR expansions, the three tools were run on samples with known clinical STR expansions. EHdn was run in a case-control mode as opposed to outlier analysis. Samples that harbour the same mutations were treated as cases and samples with alternate expansions used as controls. For STRetch, samples were used as a control to each other and the depth normalised read counts compared to identify significantly expanded loci in each sample relative to the rest. STRetch was run on the GRCh37 genome reference panel of STR genomic coordinates containing known STR-disease loci provided with the tool. GangSTR was run using GangSTR's SRCh37 latest STR panel, retrieved from GangSTR's website. GangSTR was run with the same parameters described above. However, for Fragile X Syndrome and Spinal and Bulbar Muscular Atrophy, GangSTR was run by specifying the ploidy of the X chromosome for male and female.

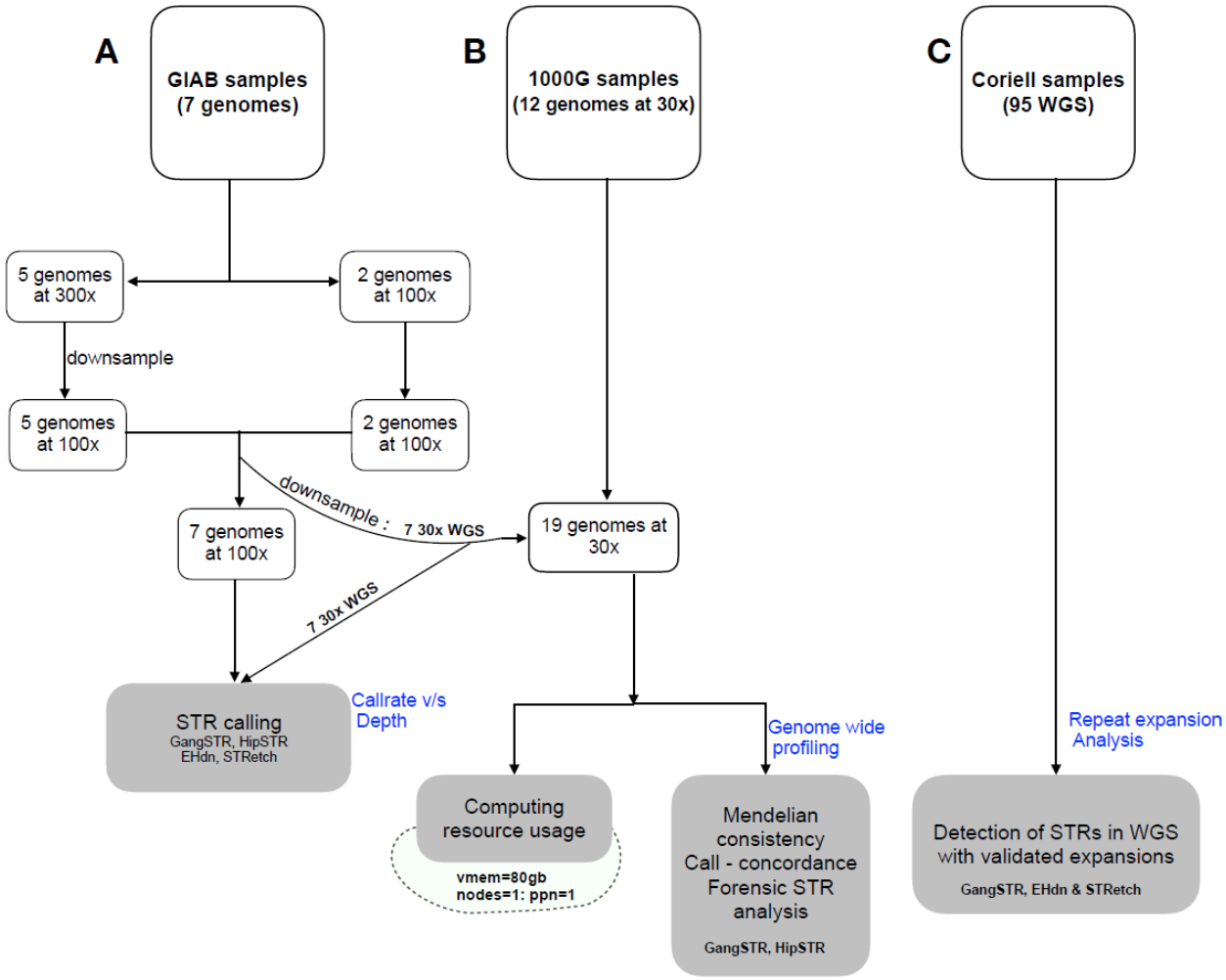
Data availability

Genotype call vcf files for GangSTR and HipSTR are available for the Genome In a bottle samples are at https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/

Genotype call vcf files for GangSTR and HipSTR are available for the 1000 Genomes samples used are at https://doi.org/10.25392/leicester.data.19804360.v1

Figure 1 - Flowchart of analyses and samples used in this study

a) Genome in a bottle (GIAB) samples
b) 1000 Genomes project samples
c) Known clinical samples available from Coriell



Results

Computing resource usage

For large scale studies involving thousands of samples, the relative processing time of different software tools can become an important aspect of software choice. To evaluate this, we recorded the time and memory taken to process each bam file from five samples sequenced at 30x coverage, using a single core of an Intel Xeon Skylake CPU running at 2.6GHz clock speed with 80Gb RAM available (Table 2). GangSTR took between 10 and 13 hours, by far the longest time per sample, considerably longer than its comparator software HipSTR, despite calling fewer STRs.. This is presumably due to the extra computational time needed to incorporate mate pair distances into the genotype calls.

STRetch took between 7 and ~9 hours to process a single bam file, compared with EHdn which took much less time, around 30 minutes. However, STRetch is the only software tested with a multithreading option, and increasing the core number from 1 to 28 reduces the time to about 1.5 hours per sample. STRetch also was notable for its high memory usage - typically around 50Gb - in contrast to the other three software tools using ~1Gb or less.

Table 2. Comparison of processing times and memory usage of STR-calling software

	CpuTime (hours:minutes:seconds)				RAM (Gb)			
Sample	gang STR	HipSTR	STRetch	EHdn	gangSTR	HipSTR	STRetch	EHdn
NA18485	13:14:10	02:38:53	08:09:04	00:31:39	0.27	1.2	49.8	0.57
NA18487	10:55:06	02:42:26	07:58:50	00:31:09	0.20	1.1	50.3	0.48
NA06984	10:33:40	02:32:12	07:44:31	00:31:35	0.23	0.84	48.5	0.58
NA06989	10:53:39	02:45:10	07:05:29	00:28:24	0.46	1.1	49.0	0.57
NA12329	12:10:21	03:07:52	09:15:38	00:30:53	0.24	0.93	48.8	0.51

Number of common STR loci genotyped

We compared GangSTR and HipSTR, the two software tools that aim to genotype all STRs across the genome. Both rely on a reference panel of STR loci provided by the user, and the number of the loci in that reference panel provide the upper limit on the number of called STRs. Therefore our results are presented as a percentage of STRs called using that particular reference panel.

Table 3 shows several broad patterns common to both GangSTR and HipSTR. Firstly, both software tools call a high proportion of the STRs in their respective catalogues across all samples, with GangSTR calling a higher proportion than HpSTR, although the different sizes of their catalogues must be taken into account. Secondly, more calls are made from 100x coverage data than 30x. The increase is small across most samples, but is large (over 10 percentage points) for HG005, where both tools call fewer than 90% of STRs in the 30x coverage data. HG005 is the only sample in our dataset that is sequenced using 250bp sequence reads rather than 150bp.

Taken together, 30x 150bp is not only sufficient but appears optimal for STR calling, at least using the measures in table 2, there is little benefit in increasing sequencing depth, and using 250bp reads is detrimental to STR calling, at least for the analysis parameters we used.

Table 3 - Number of loci called as percentage of total in catalogue for GangSTR and HipSTR and call concordance as function of sequence depth

	GangSTR					HipSTR				
Sample	Called at 30x (%)	Called at 100x (%)	Calls in both (%)	Identical by allele length	Identical by allele sequenc	Called at 30x (%)	Called at 100x (%)	Calls in both (%)	Identical by allele length	Identical by allele sequenc
HG002	96.8	98.3	96.7	99.2	NA	91.2	91.2	87.8	99.2	98.9
HG003	96.8	98.3	96.7	98.9	NA	91.1	91.1	87.7	99.2	98.8
HG004	96.9	97.8	96.7	99.1	NA	90.6	90.6	87.2	99.1	98.8
HG005	87.7	97.9	87.6	99.8	NA	71.8	71.7	68.3	99.8	99.6
HG006	96.6	98.2	94.5	99.2	NA	90.9	91.0	85.6	99.2	98.9
HG007	96.8	97.7	96.6	99.1	NA	90.7	90.7	85.4	99.2	98.9
NA12878	96.9	97.8	96.7	98.1	NA	90.2	90.1	84.8	99.1	98.8

Genotype call concordance of common STRs

To more fairly compare the number of STRs genotyped and accuracy of HipSTR and GangSTR, we called STRs using both software tools and the same STR catalogue of 811899 STRs, included in both reference panels. Over 90% of STRs in the catalogue were called by both GangSTR and HipSTR, with GangSTR calling slightly more STRs (Figure 2a). We compared the actual genotype calls made on the STRs called by both HipSTR vs GangSTR by plotting each allele call for each on a scatterplot, normalised against the allele in the reference genome (Figure 2b). This shows that 87.6% of the calls were identical between HipSTR and GangSTR, and of those calls 95.0% were identical with the reference genome. There was no strong bias in under- or over-calling repeat number made by HipSTR or GangSTR.

Accuracy of common STR calls assessed using Mendelian inheritance

Accuracy of the genotypes made by GangSTR and HipSTR can be measured using the expectation of Mendelian inheritance of alleles, whereby we would expect to see one allele of a child's genotype in their mother and one allele in their father. Any allele in a child not observed in one of the parents is due either to a genotyping error in the child or the parent, or a de novo mutation in the child. Assuming a rate of de novo mutations at STR loci as 10-4 per generation, (Willems et al., 2016), the contribution to observed inheritance errors of de novo mutation is minimal and therefore differences in inconsistencies can be attributed to differences in STR genotyping accuracy.

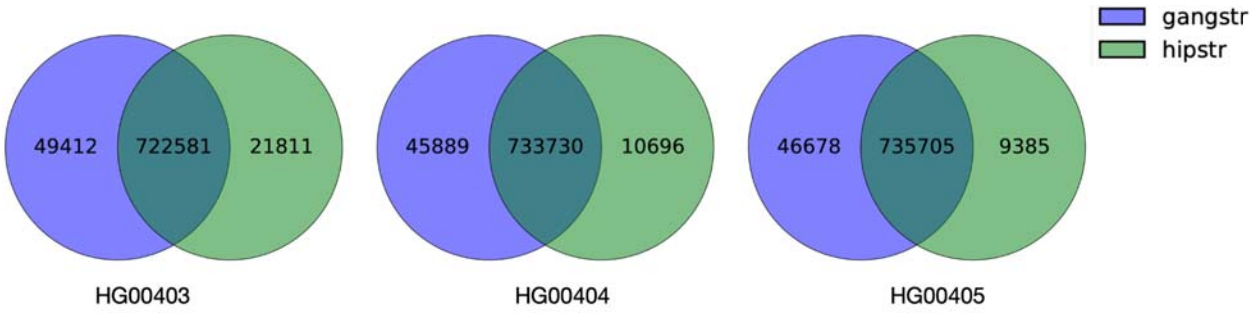
Analysis of Mendelian inconsistencies using bcftools software across five mother-father-child trios shows very little difference in genotyping accuracy between HipSTR and GangSTR (table 4). Comparison between the Ashkenazi trio sequenced at 100x coverage and 30x coverage shows that increased sequencing coverage does not improve genotyping accuracy.

Figure 2 - Comparison of STR calls from HipSTR and GangSTR

d) Comparison of the numbers of loci called by HipSTR (green) and GangSTR (blue) on three individuals of the CHS trio sequenced at 30x coverage.

e) Comparison of the genotypes of the STRs called by both GangSTR and HipSTR. X axis represents the call made by GangSTR in comparison to the reference sequence (0=same as reference sequence, -20 represents 20 fewer repeat units than the reference sequence. The Y axis represents the call made by HipSTR.

a)



b)

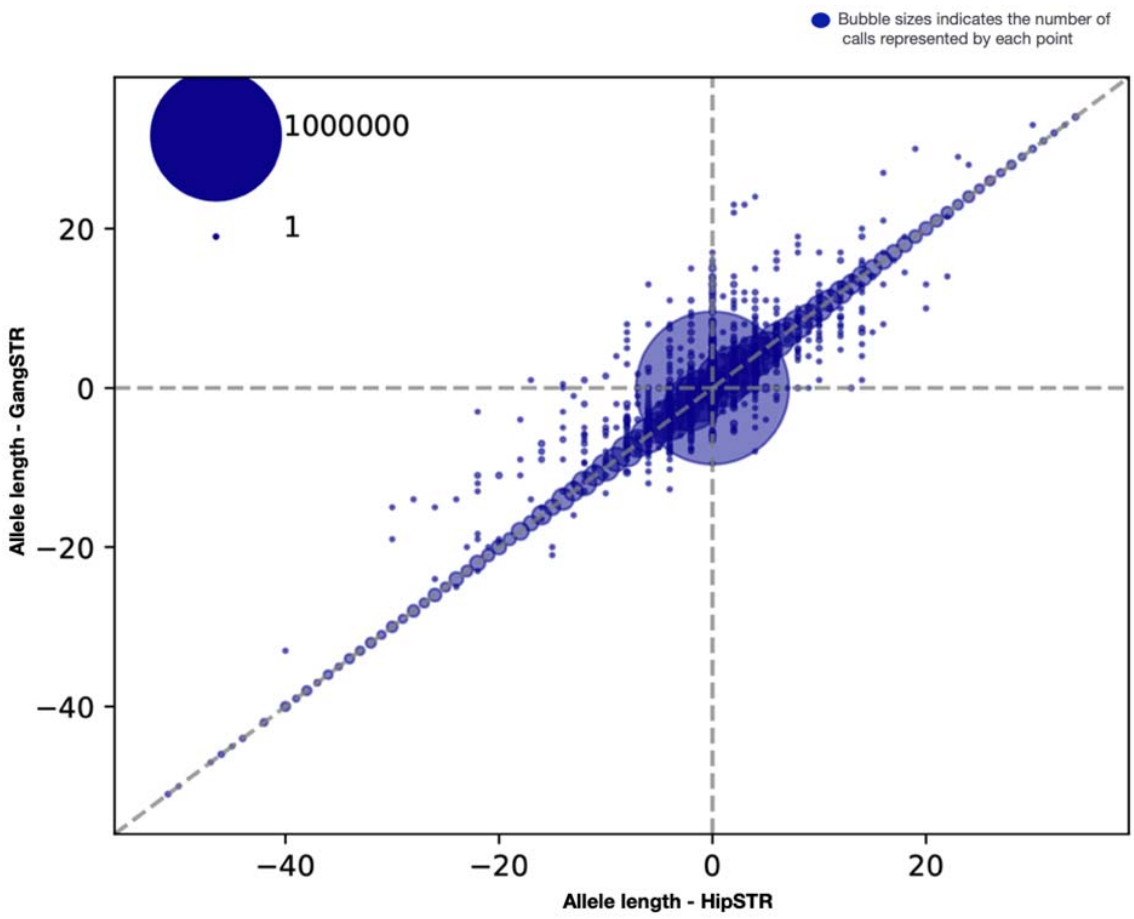


Table 4 - Mendelian inheritance of HipSTR and GangSTR

Parent-offspring Trio	Coverage	GangSTR Mendelian consistent (%)	HipSTR Mendelian consistent (%)
Ashkenazi HG002,HG003,HG004	100x	99.0	99.0
Ashkenazi HG002,HG003,HG004	30x	99.3	99.1
YRI: NA18485, NA18489, NA18487	30x	99.1	99.1
CEU: NA06984, NA06989, NA12329	30x	99.8	99.9
CHS: HG00403, HG00404, HG00405	30x	99.9	99.9
IBS: HG01500, HG01501, HG01502	30x	99.2	96.6

In order to investigate the possible reasons for erroneous genotype calls, we stratified the genotypes that showed Mendelian inconsistencies by repeat unit size (figure 3). By comparing against the distribution of STRs stratified by repeat unit size in the reference list, It is clear that, for both HipSTR and GangSTR, there is overrepresentation of 2bp repeat unit STRs in the incorrect genotype calls. Therefore for both approaches, 2bp repeat STR genotypes have the highest error rate, over twice the error rate of other STRs.

Genotyping of known forensic STR loci

Thirteen STRs, known as core Combined DNA Index System (CODIS) STRs, are long STRs that show extensive variation between individuals and are used across many different forensic STR panels for identification of individuals from genomic DNA (Butler, 2006; Brearley et al., 2020). Analysis of the performance of GangSTR and HipSTR in calling these STRs in comparison to capillary electrophoresis (CE) data is useful both for practical forensic analysis, but also as a measure of the error rate in the STR calling of the software, if we accept that capillary electrophoresis can be regarded as a gold standard. Matched capillary electrophoresis data for the 13 core STRs, generated using Promega Powerplex Fusion 24 assay, has been published (Kim et al., 2019), and our calls for NA12878 were compared against these data (table 5). GangSTR showed identical genotypes when compared with CE data for 11/13 loci, with one locus (D21S11) not called and one (TH01) calling one allele incorrectly. HipSTR showed the same results as GangSTR, except that, in addition, it incorrectly called one allele at D13S317. Both tools were also consistent across 22/23 forensic STRs analysed for mendelian inheritance errors across the 6 trios (Table 6).

Figure 3 - Repeat unit size distribution of Mendelian inconsistencies

a) HipSTR calls that are Mendelian inconsistent, b) all STRs in the reference panel used by both GangSTR and HipSTR and c) GangSTR calls that are Mendelian inconsistent.

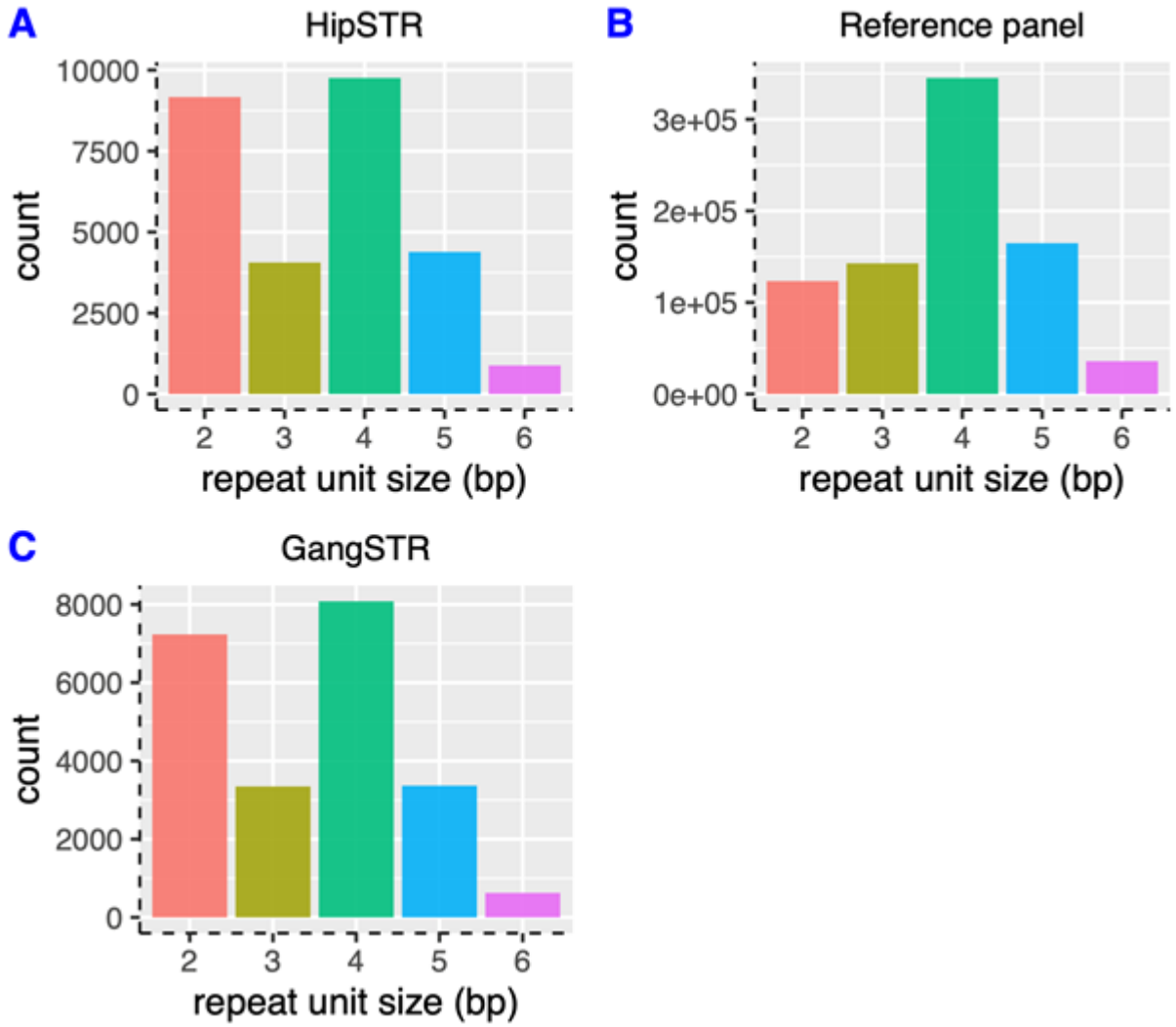


Table 5 - Comparison of STR genotype calls at core forensic loci for NA12878

Gene	Chrom	Start	End	Motif	GangSTR 100x	GangSTR 30x	GIAB HipSTR 100x	HipSTR 30x	CE
CSF1PO	5	150076324	150076375	ATCT	10,11	10,11	10,11	10,11	10,11
D5S818	5	123775556	123775599	ATCT	12,12	12,12	12,12	12,12	12
D7S820	7	84160226	84160277	TATC	8,10	8,10	8,10	8,10	8,10
D13S317	13	82148025	82148068	TATC	11,12	11,12	11.3,12.3	11.3,12.3	11,12
D16S539	16	86352702	86352745	GATA	10,11	10,11	10,11	10,11	10,11
D21S11	21	19181973	19182099	TCTA	F	F	F	F	30
TH01	11	2171088	2171115	AATG	7,8	7,7	7,9.8	7,9.8	7,9.3
TPOX	2	1489653	1489684	AATG	8,8	8,8	8,8	8,8	8
vWA	12	5983977	5984044	AGAT	15,17	15,15	15,17	15,17	15,17
D3S1358	3	45540739	45540802	TCTA	16,17	16,17	16,17	16,17	16,17
D8S1179	8	124894865	124894916	TATC	12,12	12,12	12,12	12,12	12
D18S51	18	63281667	63281738	AGAA	16,17	16,17	16,17	16,17	16,17
FGA	4	154587736	154587823	GGAA	22,24	22,24	22,24	22,24	22,24

N/B: F=failed QC; - =not typed CE=capillary electrophoresis genotype

Table 6 - Mendelian inheritance of HipSTR and GangSTR at forensic STRs

	Trio	Coverage	nOK	nBad	nSkipped	Failed QC
GangSTR	HG004,HG003,HG002	100x	21	0	0	1
	HG004,HG003,HG002	30x	15	1	4	2
	YRI: NA18485, NA18487, NA18489	30x	16	2	3	1
	CEU: NA06984, NA06989, NA12329	30x	17	0	2	3
	CHS: HG00403, HG00404, HG00405	30x	17	1	1	3
	IBS: HG01500, HG01501, HG01502	30x	17	0	2	3
HipSTR	HG004,HG003,HG002	100x	19	1	1	1
	HG004,HG003,HG002	30x	19	1	1	1
	YRI: NA18485, NA18489, NA18487	30x	21	0	0	1
	CEU: NA06984, NA06989, NA12329	30x	21	0	0	1
	CHS: HG00403, HG00404, HG00405	30x	19	0	0	3
	IBS: HG01500, HG01501, HG01502	30x	19	1	1	1

nOK - number of genotypes at which a trio had no missingness and no Mendelian error
nBad: number of genotypes at which a trio had a Mendelian error
nSkipped: number of genotypes at which a trio had at least one individual missing data and therefore could not be considered
Additional forensic STRs analysed: D1S1656,D2S441,D10S1248,PENTA E,D2S1338,PENTA D,D12S391,D19S433,D22S1045.

Detection of STR expansions at known clinical loci

To assess the reliability in calling expanded STRs in a real-world situation, we compared STRetch, EHdn and GangSTR performance in calling expanded STRs in clinical samples with known STR expansions (Table 7). All three methods showed high sensitivity, with EHdn (95% sensitivity) and STretch (97% sensitivity) outperforming GangSTR (73% sensitivity). To assess specificity, 19 samples with known normal allele length at FMR1. EHdn called four of these 19 samples as expanded, while both STRetch and GangSTR called two of the 19 samples as expanded. This is a small dataset to robustly test false positive rate, but suggest that EHdn may have a slightly lower specificity (~80%) than the other two approaches (~90%).

Table 7 - Detection of STR expansions at known clinical loci

Disease	Gene	Total analysed	Identified using EHDN	Identified using STRetch	Identified using GangSTR
Spinal-bulbar muscular atrophy	*AR*	1	0	1	0
Myotonic dystrophy type 1	*DMPK*	13	13	13	10
Fragile X	*FMR1*	26	26	26	22
Friedreich's ataxia	*FXN*	18	18	18	13
Huntington's disease	*HTT*	11	11	11	7
Dentatorubral-pallidoluysian atrophy	*ATN1*	2	2	0	0
Spinocerebellar ataxia type 1	*ATXN1*	3	0	3	2
Spinocerebellar ataxia type 3	*ATXN3*	1	1	1	1
Total		75	71 (95%)	73 (97%)	55 (73%)

Discussion

Genomewide analysis of short tandem repeats using short read sequencing has lagged behind studies of other variations, including single nucleotide variation and, to a certain extent, structural variation. Several software tools have been developed with the aim of accurately calling STR genotypes genomewide, but with different ultimate aims. Some have been developed to detect and genotype large repeat expansions that are outliers from the population distribution of allele lengths for thay STR. Development of these was motivated by well-established repeat expansions causing a variety of Mendelian diseases, and more recent discoveries showing that large repeat expansions can underlie a large amount of complex disease, such as ALS (DeJesus-Hernandez et al., 2011; Renton et al., 2011). Other software tools aim to genotype STRs genomewide irrespective of their alleles, using a catalogue of known STRs.

We selected the most recent software tools most appropriate to our needs in detecting both STR expansions and genomewide STR genotypes in a population-based sample. A previous study describes detection of six disease-causing STR expansions from whole exome sequences using several STR-calling software tools, including GangSTR and STRetch but not EHdn (Rajan-Babu et al., 2021). Expansion Hunter (Dolzhenko et al., 2019), STRetch and exSTRa (Tankard et al., 2018) were used to detect clinically-related STR expansions in those 6 genes with high specificity. Other reports include those focused on forensic STR detection (Valle-Silva et al., 2022), or are focused on genotyping a particular, or small number, of STR loci (Budiš et al., 2019).

Our aim was to assess the feasibility of applying these software tools, both in terms of quality of the resulting data and practicality of using the software, for large cohorts of genomes. We assessed this in a variety of ways. Processor and memory usage, in conjunction with processing time, was measured so that we could assess the feasibility of scaling up analysis to thousands of genomes given current computing resources. For the software using a catalogue to genotype known STRs (GangSTR and HipSTR), we examined the effect of increasing sequencing depth on the proportion of STRs genotyped, and accuracy. We then examined the concordance between the two software tools using STR catalogues, and the relative number of Mendelian inconsistencies in trios to assess STR genotyping accuracy. We found sequencing depth to have no effect on the number or quality of STR genotypes called by HipSTR, but higher sequencing depth had a modest positive effect on GangSTR STR genotype calling. Quality of STR calls for both GangSTR and HipSTR was very similar when given the same STR catalogue, and GangSTR used considerably less memory than HipSTR, but took between 3x and 4x more CPU time. Both GangSTR and HipSTR performed very well, but not perfectly, in genotyping STR loci used for forensics.

The software tools STRetch, EHdn and GangSTR can all call expanded repeats, with only EHdn not requiring a predefined catalogue of STR loci. Because expanded repeats are expected to be rare, and underlie several clinical conditions, we used a previous dataset to test the ability of STRetch, EHdn and GangSTR to genotype known STR expansions. EHdn and STRetch show

higher sensitivity than GangSTR, at least under the conditions tested, with EHdn possibly showing lower levels of specificity compared to GangSTR and STRetch. EHdn used the least resource, analysing a genome in about half an hour using about half a Gb of memory under our conditions.

For case-control studies where thousands of genomes have been sequenced at ~30x coverage for both large expansions and known STRs genomewide, we have decided to use GangSTR and EHdn. EHdn takes fewer computing resources compared to its rivals, and is the only STR genotyper that can genotype expansions without a prior defined catalogue, broadening its scope to identify previously unknown expansions. GangSTR and HipSTR are similar, with the increased computing resources needed by GangSTR offset in our view by its ability to detect larger STR expansions longer than the read length, to support and extend EHdn calls.

We note that, although the clinical repeat expansions shown in table 7 and the forensic STRs in table 6 have been validated by orthogonal data and methods, this is not the case with the other STR genotypes. For STRs that are associated with disease in subsequent analyses, it will be important to validate those particular STRs on a subset of samples using alternative methods, such as capillary electrophoresis. STR genotype calling from high throughput sequencing remains an area of active development, and we hope that further progress can be made in reliably calling STRs from both short read and long read sequencing data.

Our study had some limitations inherent to different tools. First, we did not analyse the 1bp repeat unit motifs. Both GangSTR and EHdn have been optimised to genotype 2-20bp repeat motifs, of which EHdn only genotypes expanded repeats. 1bp repeat motifs present a challenge to many sequencing technologies due to PCR stutter noise and may introduce artefacts in sequence, however, a detailed assessment of accuracy of 1bp repeat motif STR genotyping in PCR-free sequencing is needed. Secondly, apart from EHdn, the rest of the tools require defined STR coordinates built from reference genomes. Therefore STRs not assembled in the reference genome, including the highly complex regions of the genome that are often hard to sequence and assemble, cannot be genotyped. However, with improvements in PCR free protocols and long read sequencing, combined with new genome assemblies, the ability to genotype these regions will improve.

Acknowledgements
JWO is funded by a Wellcome Trust PhD studentship as part of the Wellcome Trust Genetic Epidemiology and Public Health Genomics Doctoral Training Programme. LWV holds a GSK/Asthma+Lung UK Chair in Respiratory Research. The research was partially supported by the National Institute for Health Research (NIHR) Leicester Biomedical Research Centre; the views expressed are those of the author(s) and not necessarily those of the National Health Service (NHS), the NIHR or the Department of Health. This research used the ALICE High Performance Computing Facility at the University of Leicester.

References

Bahlo M, Bennett MF, Degorski P, Tankard RM, Delatycki MB, Lockhart PJ. 2018. Recent advances in the detection of repeat expansions with short-read next-generation sequencing. F1000Research 7:736

Brearley EJ, Singh P, Bhatti JS, Mastana S. 2020. Genetic variation and differentiation among a native British and five migrant South Asian populations of the East Midlands (UK) based on CODIS forensic STR loci. Ann Hum Biol 47:572–583.

Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B. 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet 62:1408–1415.

Budiš J, Kucharík M, Ďuriš F, Gazdarica J, Zrubcová M, Ficek A, Szemes T, Brejová B, Radvanszky J. 2019. Dante: genotyping of known complex and expanded short tandem repeats. Bioinformatics 35:1310–1317.

Burgarella C, Navascués M. 2011. Mutation rate estimates for 110 Y-chromosome STRs combining population and father–son pair data. Eur J Hum Genet 19:70–75.

Butler JM. 2006. Genetics and genomics of core short tandem repeat loci used in human identity testing. J Forensic Sci 51:253–265.

Contente A, Dittmer A, Koch MC, Roth J, Dobbelstein M. 2002. A polymorphic microsatellite that mediates induction of PIG3 by p53. Nat Genet 30:315–320.

Dashnow H, Lek M, Phipson B, Halman A, Sadedin S, Lonsdale A, Davis M, Lamont P, Clayton JS, Laing NG, MacArthur DG, Oshlack A. 2018. STRetch: detecting and discovering pathogenic short tandem repeat expansions. Genome Biol 19:121.

DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, Rutherford NJ, Nicholson AM, Finch NA, Flynn H, Adamson J, Kouri N, Wojtas A, et al. 2011. Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron 72:245–256.

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

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

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

Fotsing SF, Margoliash J, Wang C, Saini S, Yanicky R, Shleizer-Burko S, Goren A, Gymrek M. 2019. The impact of short tandem repeat variation on gene expression. Nat Genet 51:1652–1659.

Gymrek M. 2017. A genomic view of short tandem repeats. Curr Opin Genet Dev 44:9–16.

Gymrek M, Golan D, Rosset S, Erlich Y. 2012. lobSTR: A short tandem repeat profiler for personal genomes. Genome Res 22:1154–1162.

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

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

Hannan AJ. 2020. Expanding genes, repeating themes and therapeutic schemes: The neurobiology of tandem repeat disorders. Neurobiol Dis 144:105053.

Highnam G, Franck C, Martin A, Stephens C, Puthige A, Mittelman D. 2013. Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Res 41:e32.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37:907–915.

Kuilenburg ABP van, Tarailo-Graovac M, Richmond PA, Drögemöller BI, Pouladi MA, Leen R, Brand-Arzamendi K, Dobritzsch D, Dolzhenko E, Eberle MA, Hayward B, Jones MJ, et al. 2019. Glutaminase Deficiency Caused by Short Tandem Repeat Expansion in GLS. N Engl J Med 380:1433–1441.

Lavoie H, Debeane F, Trinh Q-D, Turcotte J-F, Corbeil-Girard L-P, Dicaire M-J, Saint-Denis A, Pagé M, Rouleau GA, Brais B. 2003. Polymorphism, shared functions and convergent evolution of genes with sequences coding for polyalanine domains. Hum Mol Genet 12:2967–2979.

Lieberman AP, Shakkottai VG, Albin RL. 2019. Polyglutamine Repeats in Neurodegenerative Diseases. Annu Rev Pathol 14:1–27.

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

Paulson H. 2018. Chapter 9 - Repeat expansion diseases. In: Geschwind DH, Paulson HL, Klein C, editors. Handbook of Clinical Neurology, Elsevier, p 105–123.

Rajan-Babu I-S, Peng JJ, Chiu R, Birch P, Couse M, Guimond C, Lehman A, Mwenifumbo J, Karnebeek C van, Friedman J, Adam S, Souich CD, et al. 2021. Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions. Genome Med 13:126.

Renton AE, Majounie E, Waite A, Simón-Sánchez J, Rollinson S, Gibbs JR, Schymick JC, Laaksovirta H, Swieten JC van, Myllykangas L, Kalimo H, Paetau A, et al. 2011. A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron 72:257–268.

Saini S, Mitra I, Mousavi N et al. 2018. A reference haplotype panel for genome-wide imputation of short tandem repeats. Nat Commun 9, 4397.

Tankard RM, Bennett MF, Degorski P, Delatycki MB, Lockhart PJ, Bahlo M. 2018. Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data. Am J Hum Genet 103:858–873.

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

Valle-Silva G, Frontanilla TS, Ayala J, Donadi EA, Simões AL, Castelli EC, Mendes-Junior CT. 2022. Analysis and comparison of the STR genotypes called with HipSTR, STRait Razor and toaSTR by using next generation sequencing data in a Brazilian population sample. Forensic Sci Int Genet 58:102676.

Willems T, Gymrek M, Highnam G, 1000 Genomes Project Consortium, Mittelman D, Erlich Y. 2014. The landscape of human STR variation. Genome Res 24:1894–1904.

Willems T, Gymrek M, Poznik GD, Tyler-Smith C, 1000 Genomes Project Chromosome Y Group, Erlich Y. 2016. Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. Am J Hum Genet 98:919–933.

Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. 2017. Genome-wide profiling of heritable and de novo STR variations. Nat Methods 14:590–592.

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, Henaff E, McIntyre ABR, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 3:160025.