University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Halman, A;Oshlack, A

Title:
Accuracy of short tandem repeats genotyping tools in whole exome sequencing data

Date:
2020-01-01

Citation:
Halman, A. & Oshlack, A. (2020). Accuracy of short tandem repeats genotyping tools in whole exome sequencing data. F1000research, 9, pp.200-. https://doi.org/10.12688/f1000research.22639.1.

Persistent Link:
https://hdl.handle.net/11343/274410

License:
CC BY

Check for updates

RESEARCH ARTICLE

Accuracy of short tandem repeats genotyping tools in whole exome sequencing data [version 1; peer review: 2 approved, 1 approved with reservations]

Andreas Halman [1-4], Alicia Oshlack [1,2,5]

[1]Murdoch Children's Research Institute, Royal Children's Hospital, Parkville, VIC, 3052, Australia
[2]Peter MacCallum Cancer Centre, 305 Grattan St, Melbourne, VIC, 3000, Australia
[3]Florey Institute of Neuroscience and Mental Health, University of Melbourne, Parkville, VIC, 3052, Australia
[4]School of Natural Sciences and Health, Tallinn University, Tallinn, 10120, Estonia
[5]School of BioSciences, University of Melbourne, Parkville, VIC, 3052, Australia

v1 **First published:** 23 Mar 2020, **9**:200
https://doi.org/10.12688/f1000research.22639.1
Latest published: 23 Mar 2020, **9**:200
https://doi.org/10.12688/f1000research.22639.1

Abstract

Background: Short tandem repeats are an important source of genetic variation. They are highly mutable and repeat expansions are associated dozens of human disorders, such as Huntington's disease and spinocerebellar ataxias. Technical advantages in sequencing technology have made it possible to analyse these repeats at large scale; however, accurate genotyping is still a challenging task. We compared four different short tandem repeats genotyping tools on whole exome sequencing data to determine their genotyping performance and limits, which will aid other researchers in choosing a suitable tool and parameters for analysis.
Methods: The analysis was performed on the Simons Simplex Collection dataset, where we used a novel method of evaluation with accuracy determined by the rate of homozygous calls on the X chromosome of male samples. In total we analysed 433 samples and around a million genotypes for evaluating tools on whole exome sequencing data.
Results: We determined a relatively good performance of all tools when genotyping repeats of 3-6 bp in length, which could be improved with coverage and quality score filtering. However, genotyping homopolymers was challenging for all tools and a high error rate was present across different thresholds of coverage and quality scores. Interestingly, dinucleotide repeats displayed a high error rate as well, which was found to be mainly caused by the AC/TG repeats. Overall, LobSTR was able to make the most calls and was also the fastest tool, while RepeatSeq and HipSTR exhibited the lowest heterozygous error rate at low coverage.
Conclusions: All tools have different strengths and weaknesses and the choice may depend on the application. In this analysis we demonstrated the effect of using different filtering parameters and offered recommendations based on the trade-off between the best

Open Peer Review

Reviewer Status ✓ ? ✓

	Invited Reviewers		
	1	**2**	**3**
version 1 23 Mar 2020	✓ report	? report	✓ report

1. **Elizabeth A. Worthey**, University of Alabama at Birmingham, Birmingham, USA

2. **Jan Radvanszky**, Slovak Academy of Sciences, Bratislava, Slovakia
Comenius University in Bratislava, Bratislava, Slovakia
Geneton Ltd., Bratislava, Slovakia
Jaroslav Budiš, Comenius University in Bratislava, Bratislava, Slovakia
Geneton Ltd., Bratislava, Slovakia
Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia

3. **Maria Anisimova**, Zurich University of Applied Sciences (ZHAW), Waedenswil, Switzerland
Swiss Institute of Bioinformatics, Lausanne,

accuracy of genotyping and the highest number of calls.

Keywords
short tandem repeats, microsatellites, gangstr, lobstr, hipstr, repeatseq

Switzerland

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Alicia Oshlack (alicia.oshlack@petermac.org)

Author roles: Halman A: Conceptualization, Formal Analysis, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Oshlack A**: Conceptualization, Funding Acquisition, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: A.O. is supported by an NHMRC fellowship [GNT1126157].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2020 Halman A and Oshlack A. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Halman A and Oshlack A. **Accuracy of short tandem repeats genotyping tools in whole exome sequencing data [version 1; peer review: 2 approved, 1 approved with reservations]** F1000Research 2020, **9**:200 https://doi.org/10.12688/f1000research.22639.1

First published: 23 Mar 2020, **9**:200 https://doi.org/10.12688/f1000research.22639.1

Introduction

Overview of short tandem repeats (STRs) and methods of analysis

STRs, also known as microsatellites, consist of repeated units of 1 to 6 base pairs (bp) in length and cover about 3% of the human genome (Gymrek, 2017). STRs are highly mutable and often vary in their number of repeat units across the population. They can be found in various regions of the genome, including in or near protein coding regions and introns (Hannan, 2018). Expanded variants contribute to several dozen human disorders, including Huntington's disease, fragile X syndrome, spinocerebellar ataxias and other diseases. In addition, variation in STR length has been shown to associate with quantitative traits such as gene expression (Gymrek, 2017). The standard method to genotype the length of STRs is to perform polymerase chain reaction (PCR) amplification on the region of interest and gel electrophoresis (Tang & Nzabarushimana, 2017). Sanger sequencing has high accuracy, but low throughput, limiting analysis to a few genes at a time (Caspar *et al.*, 2018).

Recent technology advances in high-throughput sequencing (HTS) have revolutionised the genomics field and brought us the opportunity to detect sequence variants at a scale that was impossible before (Caspar *et al.*, 2018). HTS of the whole genomes provides the potential to profile over a million STRs in the human genome. Recent advances in bioinformatics have brought us tools to analyse STRs from sequencing data (HTS), but genotyping still remains challenging for many reasons, including issues with extreme GC content, short read lengths that do not span over the entire repeat, and issues with alignment due to variation in STRs appearing as large insertions or deletions relative to the reference. In addition, using PCR amplification during library preparation will often cause stutter noise and produce artificial variability in the sequence (Caspar *et al.*, 2018; Gymrek, 2017). Stutter noise is a result of *in vitro* slippage of DNA polymerase during PCR cycles that leads to erroneous reads of incorrect repeat length (Willems *et al.*, 2014), which contributes to challenges in genotyping.

Illumina has developed a method for amplification-free (PCR−) library preparation (Kozarewa *et al.*, 2009), which theoretically eliminates the STR stutter error during PCR amplification in sample preparation (PCR+) and therefore improves the accuracy of STR genotyping. The developers of STR-FM evaluated the new protocol by running their tool in both PCR− and PCR+ samples and found that the PCR− protocol compared to PCR+ has up to nine-fold fewer errors (Fungtammasan *et al.*, 2015). However, huge amounts of sequencing data have already been generated by using the PCR+ protocol, where some data will not be resequenced due to time and/or cost (Fungtammasan *et al.*, 2015). In addition, despite the advantages of whole genome sequencing (WGS), whole exome sequencing (WES) is still widely used in human genetics due to its lower cost and higher coverage and WES is a PCR+ process (Björn *et al.*, 2018). Therefore, tools that can accurately genotype STRs not only from PCR− but also from PCR+ data are essential.

While there are a number of computational tools that have been developed to genotype STR alleles in HTS data, there have been few independent comparisons of their performance. Evaluation of methods for genotyping STRs is difficult. The gold standard measurement of STRs is by capillary electrophoresis (Willems *et al.*, 2014), but these methods have low throughput. Further evaluations have used Mendelian inheritance as a measure of accuracy (Gymrek *et al.*, 2012; Highnam *et al.*, 2013; Mousavi *et al.*, 2019). Other studies have used simulated data for the evaluation of genotyping accuracy (Fungtammasan *et al.*, 2015; Highnam *et al.*, 2013). While simulation can generate many loci with known alleles, it is difficult to simulate the true complexity of real data.

Here we propose to compare and evaluate STR genotyping methods on exome data using a different but complementary approach. We used the natural hemizygous state of the X chromosomes in males to look for incorrect calls revealed by a heterozygous call. With repeats on the X chromosome in males there is only one allele, so we expect all calls to be homozygous. While this approach does not evaluate the accuracy of the allele length, it has advantages in that (a) the data sets are large so we can test thousands of calls, and (b) the data comes from real patients with all the noise and biases found in real data.

In our study, we compared LobSTR (Gymrek *et al.*, 2012), RepeatSeq (Highnam *et al.*, 2013), HipSTR (Willems *et al.*, 2017) and a recently published tool GangSTR (Mousavi *et al.*, 2019). In addition, we included a common variant calling tool GATK HaplotypeCaller (McKenna *et al.*, 2010) as a comparison of genotyping accuracy.

There are a number of tools that have been developed for STR analysis and which were excluded from this analysis. For example, popSTR (Kristmundsdóttir *et al.*, 2017) is a population based STR genotyper and is optimised for whole genome sequencing (WGS) data. STRviper is another method for genotyping STRs that is able to pick up repeats longer than the read length; however, it has no built-in stutter model and it is not suitable for

diploid dataset as it assumes only one allele (Cao *et al.*, 2014). Galaxy environment has an STR analysis tool called STR-FM which we were unable to run (Fungtammasan *et al.*, 2015). Dante (Budiš *et al.*, 2019) and STRScan (Tang & Nzabarushimana, 2017) are designed for targeted searches and require a user-defined list of STR loci.

Tools such as Expansion Hunter (Dolzhenko *et al.*, 2017; Dolzhenko *et al.*, 2019), TREDPARSE (Tang *et al.*, 2017), STRetch (Dashnow *et al.*, 2018) and exSTRa (Tankard *et al.*, 2018) were excluded from our analysis as well because they are classified as tools specifically looking for expansions that might be disease causing and are often longer than the physical read length or expansion relative to a control set.

Our analysis focuses on comparing the performance of STR genotyping tools on the X chromosome of more than 400 males. Using this data set, we investigate the overall ability of tools to call genotypes, the accuracy as a function of coverage and repeat unit and also investigate quality scores of the tools. We find most tools are able to call a majority of homozygous alleles and different tools have different advantages in terms of repeat unit and coverage.

Computational tools to genotype STRs from HTS data

First, we will give a short overview of STR genotyping tools included in our analysis and their reported accuracy. The tools evaluated in our analysis are summarised in Table 1. All of the tools require a set of defined STR loci. Tandem Repeats Finder (TRF) is a tool that can be used to detect STRs that have two or more copies of the same repeat unit in a row in the reference genome (Benson, 1999). In addition, it can detect repeats for which the repeat unit size is up to several hundred of bp long. Running TRF generates a report that includes all the loci detected in the genome, with genomic start and end location of the STR, repeat unit and its size, number of copies aligned with the consensus pattern and other relevant information. For this study, we limited the loci defined by TRF to repeat units up to 6 bp (see *Methods*).

LobSTR. LobSTR was one of the first successful STR genotyping tools for HTS data. It initially used its own inbuilt aligner but can also use data aligned with BWA-MEM (Li, 2013). LobSTR identifies reads that completely contain the STR and which also have flanking sequence with no repetitive sequence when aligned to a reference genome. As mentioned, PCR amplification during library preparation can create stutter noise at an STR locus, and LobSTR tackles this issue with an included stutter model that aims to detect and account for noise to improve genotyping accuracy. The stutter noise model used can be custom generated from the data or the standardised one supplied by the tool developers. As a result, LobSTR determines and reports the maximum likelihood estimates of the genotype in each locus (Gymrek *et al.*, 2012).

LobSTR was validated using concordance of biological replicates (blood and saliva samples) from the same subject to measure the precision of the tool. At 21x coverage, the discordance rate for genotype was 3% and for allelotype was 2%. While for lower 5x coverage, the discordance rate for genotype was 11% and for allelotype was 5%. STR length differences were analysed in discordant calls that were heterozygous in both blood and saliva samples and found that at coverage 5x or higher, 90% of the errors were one repeat unit difference and 99% of errors were in 2 bp repeat unit size (Gymrek *et al.*, 2012). However, it is important to note that LobSTR validated 2–6 bp repeat unit

Table 1. Feature comparison of short tandem repeats (STRs) specific genotyping tools used in our analysis.

	RepeatSeq	LobSTR	HipSTR	GangSTR
Latest version of the tool used in this study	0.8.2 (2014)	4.0.6 (2016)	0.6.2 (2018)	2.4 (2019)
Built-in stutter noise model	✓	✓	✓	✓
Ability to detect STRs that are longer than the read length	✗	✗	✗	✓
Input file types	BAM	BAM FASTA FASTQ	BAM CRAM	BAM CRAM
Sequencing read types	Single- and paired-end reads	Single- and paired-end reads	Single- and paired-end reads	Paired-end reads

size STRs and did not validate homopolymers (Gymrek *et al.*, 2012), which are common in the genome and a known source of genetic variation (Highnam *et al.*, 2013).

RepeatSeq. RepeatSeq (Highnam *et al.*, 2013) uses data aligned by an external tool, such as BWA or Bowtie. It uses Bayesian model selection to determine the most probable genotype and requires all reads to fully contain STRs and at least two reads at a locus to make a call. The RepeatSeq noise model is based on genomes derived from over 100 inbred isolates of fly.

RepeatSeq's accuracy was evaluated by analysing a trio WGS data to test consistency with Mendelian inheritance. The authors reported that on minimum coverage of two, 92.1% of repeat calls were consistent with the Mendelian inheritance, while with a minimum coverage of nine it was 95.3% and on minimum coverage of 17 it was 98.0% (Highnam *et al.*, 2013).

GangSTR. One of the major drawbacks of the first series of STR profiling methods was that they were limited to genotyping repeats within the read length in HTS data. GangSTR (Mousavi *et al.*, 2019) is a more recent method that incorporates additional information besides repeat-enclosing reads to estimate the length of repeats. This includes available information such as fragment length, coverage and information about partially enclosing reads where only one end contains flanking sequence. More specifically, reads are divided into four classes: 1) enclosing read pairs that have at least one read that includes the whole STR and a flanking region in both ends; 2) spanning read pairs that have a mate pair where one read is aligned to one side of the STR and the second read of the pair on the other side; 3) flanking read pairs that include a read which partially extends into the STR region; and 4) fully repetitive read pairs that have one or two reads which are entirely made of STR (Mousavi *et al.*, 2019). These four classes of reads are used to not only genotype repeats less than the read length but can also be used to genotype longer alleles such as repeat expansions.

The GangSTR method was evaluated by first simulating paired-end 150 bp reads (40x coverage) for 14 repeat expansions involved in STR disorders. Tool accuracy was measured by comparing true and observed alleles and also compared to TREDPARSE and ExpansionHunter. In this evaluation, GangSTR showed a lower root mean square error (RMSE) rate between true and observed allele lengths for all tested repeats. The authors demonstrated that GangSTR had an advantage over ExpansionHunter and TREADPARSE, especially in alleles that were close to the read length or longer. In addition, GangSTR and ExpansionHunter improved significantly with higher coverage and longer read length.

GangSTR genotyping for disease causing alleles was also tested on validated 14 Huntington's Disease and 25 Fragile X Syndrome real PCR-free WGS data and they reported an RMSE (7.9 and 29.3, respectively) that was lower than for TREDPARSE (8.3 and 34.8, respectively) and ExpansionHunter (10.1 and 27.3, respectively). In evaluations of genotyping a WGS trio, GangSTR was found to have similar performance to HipSTR for shorter alleles (Mousavi *et al.*, 2019).

HipSTR. HipSTR (Willems *et al.*, 2017) is a haplotype-based method for genotyping, haplotyping and phasing STRs. While other STR tools are made for finding true length of repeats independently along the genome, HipSTR takes into account the whole repeat structure on the allele, which may also have missing data. HipSTR accuracy was tested by comparing calls from 118 PCR− WGS samples to capillary electrophoresis data, reporting about 98.8% consistency between the two datasets (Willems *et al.*, 2017).

GATK HaplotypeCaller. GATK HaplotypeCaller (GATK-HC) (McKenna *et al.*, 2010) can also be used for finding SNPs and indels in repeat regions, but it is not specifically made for STR analysis. It has been widely documented that indel calling is not as accurate as SNP calling and indel callers are not ideal for identifying STR mutation due to the lack of reporting repeat genotypes. Instead, indel callers report insertions or deletions of bases relative to the reference, which may or may not be a multiple of the repeat unit, as well as including SNP differences. Dedicated STR callers, however, use information about the repeat unit, composition and repeat length in order to make more accurate genotype calls (Highnam *et al.*, 2013).

Results

In order to evaluate the accuracy and performance of STR genotyping methods, we used a novel evaluation approach applied to exome sequencing data of more than 430 individuals. Several previous comparisons determined accuracy by comparing the estimated lengths of repeats to a known truth, determined from either simulations or alternative assays such as PCR. Here, we took only male individuals and looked at the heterozygosity of the calls

only on the X chromosome. As there is only one X chromosome in males, a method that reported only homozygous calls was defined to be more accurate than those that reported heterozygous calls.

Dataset

We began with a dataset of 472 males from the Simons Simplex Collection. We had to remove 39 samples for a variety of reasons: six samples were not sequenced with a paired end approach, three samples had no coverage on the Y chromosome so were assumed to be females mislabelled as males, 28 samples produced an error in GangSTR and two samples could not be aligned to the reference genome by BWA due to a software error. The remaining 433 samples (Supplementary Table 1, see *Underlying data*) (Halman, 2020a) were analysed with LobSTR (Gymrek *et al.*, 2012), RepeatSeq (Highnam *et al.*, 2013), HipSTR (Willems *et al.*, 2017) and GangSTR (Mousavi *et al.*, 2019) (see *Methods*). In addition, variant calling was performed using the GATK best practices pipeline.

In brief, the FASTQ files were mapped using BWA-MEM (Li, 2013) and the same BAM files were used as the starting point for running each STR calling method. Each method requires a set of intervals that define repeats to be genotyped. To generate this, we used Tandem Repeats Finder (Benson, 1999) to locate tandem repeats in the hg19 reference genome and detected 224774 STR loci in the X chromosome. Because we are using exome data, we only analysed calls in the 6860 capture regions on the X chromosome. In total, we found 2322 STR loci overlapping the capture regions (Figure 1A), where almost 60% of loci consist of 6 bp repeat units (Figure 1B). In our full data set, across the 433 individuals, we have over a million STR loci for analysis.

First, we looked the ability of a method to make a call at any given locus in the capture regions. By looking at the total number of calls on the X chromosome for each method, we found that LobSTR reported the highest number of loci (Figure 2A). However, the number reported for each individual was variable. Figure 2B shows the distribution of the number of reported loci per individual, with the highest median number of calls by LobSTR (2015), outperforming GangSTR (1967), RepeatSeq (1847) and HipSTR (1834). GATK-HC reported a median of 11 loci per individual but, rather than genotyping all loci, GATK only makes a report when it deems there is a difference from the reference genome. Out of the 2322 loci we investigated, there were 23 loci for which the reference STR length was longer than the read length, but in closer examination none of the tools reported alleles in these loci that were over the read length. However, when looking only at the allele lengths then GangSTR reported 22 unique loci over all samples where at least one of the alleles were longer than the read length, whereas LobSTR reported five and both HipSTR and RepeatSeq four loci.

GATK only makes calls at positions where there is evidence that the allele is different from the reference. In this dataset, GATK made calls in a total of 346 (14.9%) different loci and 21.4% of these were heterozygous, giving a minimum overall heterozygous rate of 3.1%, assuming all uncalled positions are homozygous. No call either means the allele is a homozygous reference or there is not enough data to make a call. This is one reason why specialised STR callers are better suited for genotyping STR loci.

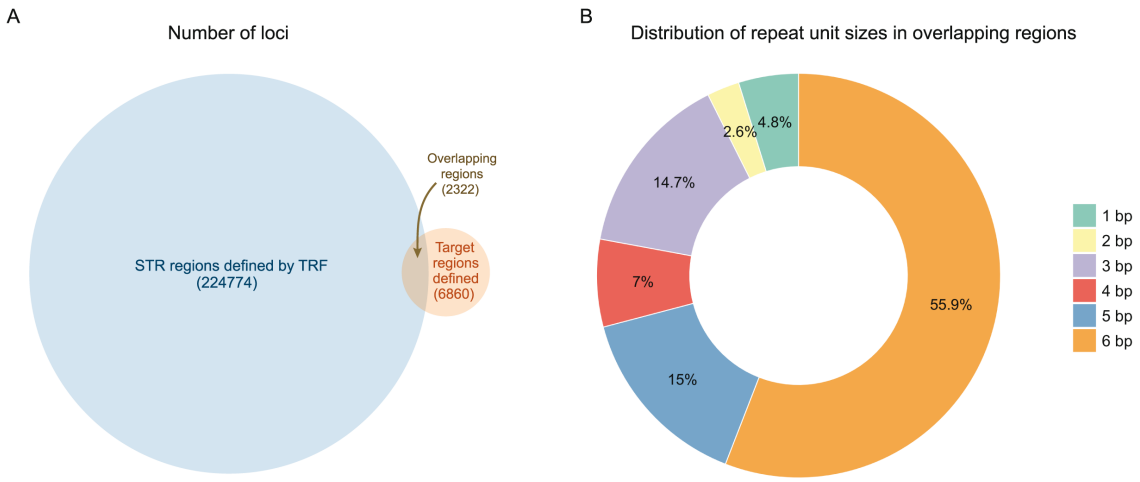


Figure 1. STR loci and repeat unit distribution in the X chromosome. (**A**) Number of STR loci defined by TRF and number of regions in the capture regions of the X chromosome. Overlapping regions include all STR loci that are completely or partially overlapping the target region. Total number of STRs found in target regions is 2322. (**B**) Distribution of all repeat unit sizes in overlapping regions (2322). STR, short tandem repeat; TRF, Tandem Repeats Finder.

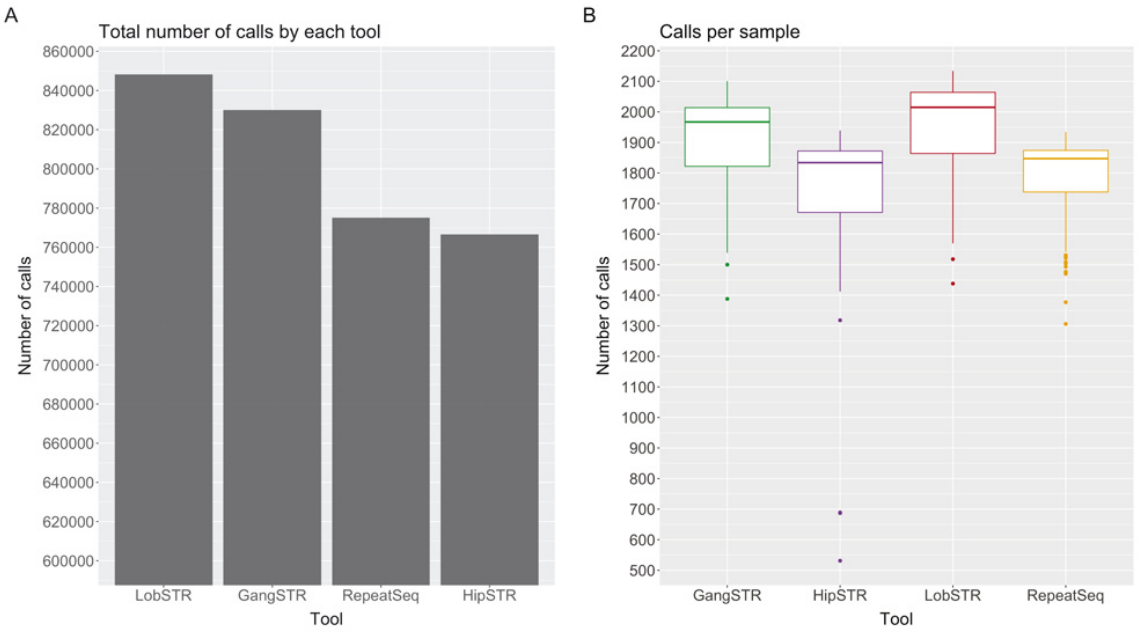


Figure 2. Comparison of different short tandem repeat genotyping tools. (**A**) Total number of calls on the X chromosome over all 433 samples made by each tool. (**B**) Number of calls made per sample by each tool out of a possible 2322.

To determine the genotyping accuracy of the four specialised STR callers, we first looked at the overall percentage of heterozygous calls to estimate the error rate for each method. Overall, RepeatSeq had the lowest median error rate, with 8675 (1.09%) of its calls being heterozygous. Next was HipSTR with 19459 (2.23%), LobSTR with 27410 (2.96%) and GangSTR with 33204 (3.29%). Again, error rates were variable across individual samples, ranging between 0% and 47.3%. RepeatSeq, HipSTR and LobSTR are generally consistent, with three sample outliers with respect to error rate, while GangSTR has higher variability in error rates across samples (Supplementary Figure 2, *Extended data*) (Halman, 2020a). Interestingly, for these three samples, the heterozygous percentage increased for LobSTR after the strict filtering. All tools except RepeatSeq recommend filtering the outputs based on quality metrics of the calls (see *Methods* for details on filtering parameters that were used). Once the recommended filters are applied, we found that the performance of GangSTR improved by 2.76% to 0.53%, HipSTR by 2.14% to 0.09% and LobSTR by 1.32% to 1.64% (Figure 3). However, the median number of calls per sample dropped for LobSTR by 201, HipSTR by 1462 loci and GangSTR by 1512.

The recommended filters for each tool were different (for instance, minimum coverage of 100 for HipSTR and 50 for GangSTR) and therefore we next decided to analyse the effect of these filtering parameters separately.

Effects of repeat unit and coverage on accuracy

We investigated the number of heterozygous calls as a function of the repeat unit length, ranging from 1 to 6 bp. We found that all tools exhibit high error rates for 1 bp repeats, which is not surprising as it is difficult to genotype homopolymers due to higher rates of polymerase slippage events. More surprisingly, 2 bp repeats were poorly genotyped by HipSTR, LobSTR and GangSTR and the best results were obtained with RepeatSeq. All other repeat unit lengths produced much more accurate genotypes.

To investigate the effect of coverage and quality scores on results, we applied call-level filters to our data according to developers' recommendations to get two different datasets: low filtering, where we included all suggested filters except coverage and quality scores, and strict filters, where we also included filters for coverage and quality scores. Then, we looked at the effects of coverage and plotted the error rate as well as percentage of the remaining number of calls as a function of the minimum number of reads supporting the call. We expected the error rate to drop as coverage increased and this was the case for 3, 4, 5 and 6 bp repeat units. However, mono- and dinucleotide repeats did not follow a consistent pattern and the pattern was different between tools. While dinucleotide repeats showed a trend towards a lower error rate with increasing coverage for RepeatSeq and LobSTR, the trend was reversed for GangSTR (Figure 4).

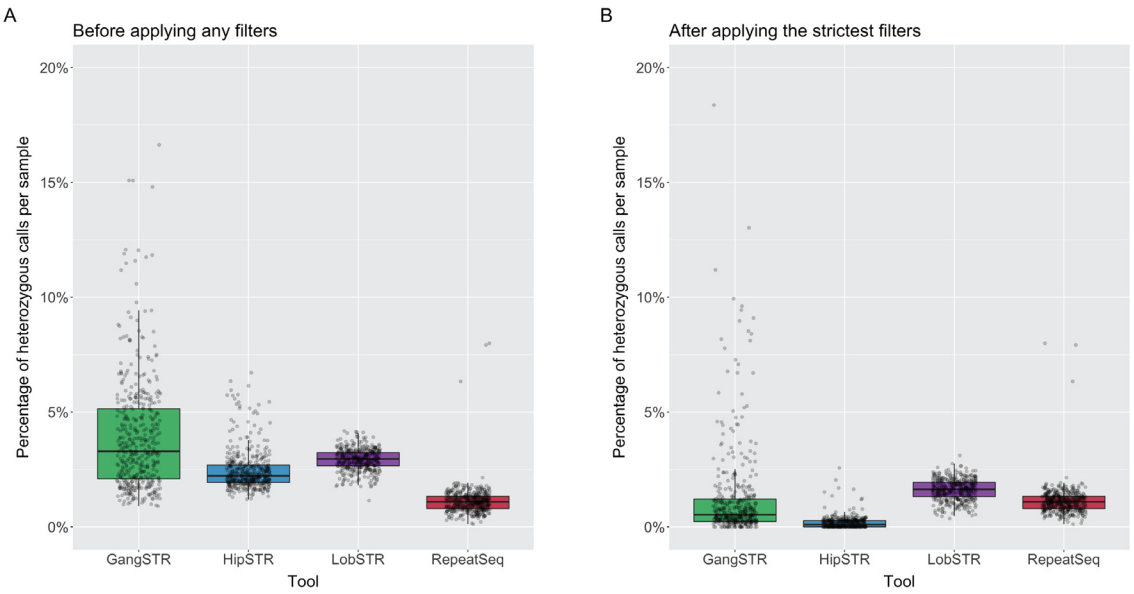


Figure 3. Comparison of different short tandem repeat genotyping tools on making heterozygous calls with and without filters. (**A**) Percentage of heterozygous calls over all samples (each dot is a sample) - no filters applied. (**B**) Percentage of heterozygous calls over all samples after applied the strictest recommended filters for the tools. Since no filters were recommended for RepeatSeq then it has the same values on both plots.

Two call-level filters are recommended for GangSTR: level 1 filters, which require filtering out calls that have less than 20 reads, and level 2 filters, which require at least 50 reads to support a call. Filtering out all calls with coverage of less than 20 reads brings the heterozygous error rate for the 3–6 bp repeats to between 1.6 and 3.1% at this minimum coverage. Filtering out calls with less than 50 reads improves the error rate even further to 0.83–1.97%. However, by filtering out calls with coverage of less than 20, we lose on average 39.4% of 2–6 bp repeats data, with a median of 1120 loci reported per sample. By filtering out calls with coverage less than 50, we lose on average 71.7% of 2–6 bp repeats data and have a median of 464 loci reported. Interestingly, we see an increase in error rate as a function of coverage when genotyping 2 bp STRs (Figure 4A). Unlike the other tools, where heterozygous percentage decreases when coverage increases and remains relatively steady, GangSTR is not so consistent and fluctuates around 1%.

HipSTR excludes calls that have coverage less than 100 by default unless specified otherwise. This coverage gives high accuracy but also filters out 75.9% of data. We investigated reducing the minimum coverage and found HipSTR has excellent accuracy even with minimum coverage of 25: 0.02–0.04% heterozygous rate for 3, 5 and 6 bp and 0.16% for 4 bp repeats (Figure 4B). In addition, a minimum coverage of 50 improves results even more and the maximum error rate for these repeats is 0.01% at this minimum coverage (median of 899 calls per sample). HipSTR also struggles to genotype 2 bp repeats, having a heterozygous error rate around 5.2% for coverage of both 25 and 50. Therefore, by decreasing the filtering parameters for coverage to only exclude the calls less than 25x coverage, we retain on average 76.6% of data for 2–6 bp repeat units, with a moderate call rate (median of 1369 calls per sample).

LobSTR recommends filtering out calls with coverage less than five, which gives us a heterozygous percentage for 3–6 bp repeat units at this minimum coverage of 0.44–0.90% per repeat unit length. This filters out on average 10.0% of calls for 2–6 bp repeats, giving a median of 1840 reported alleles per sample. The accuracy for these repeats increases as a function of coverage and considering the amount of calls that are filtered out, a minimum coverage of 5–10 might be the best compromise. By filtering out calls with coverage less than 10, the heterozygous error rate at this minimum coverage is 0.55% for 4 bp and 5 bp repeats, 0.48% for 3 bp repeats and 0.36% for 6 bp repeats, while still retaining 74.5% of calls on average for 2–6 bp repeats (median of 1555 calls per sample). Increasing the minimum coverage to 20 reads would improve our results by a further 0.2% but also filter out significantly more calls. For dinucleotide repeats, accuracy seems to get better with higher coverage and reaches 1% error rate when the minimum coverage is 33; however, the error rate increases again when coverage increases over 42 (Figure 4C).

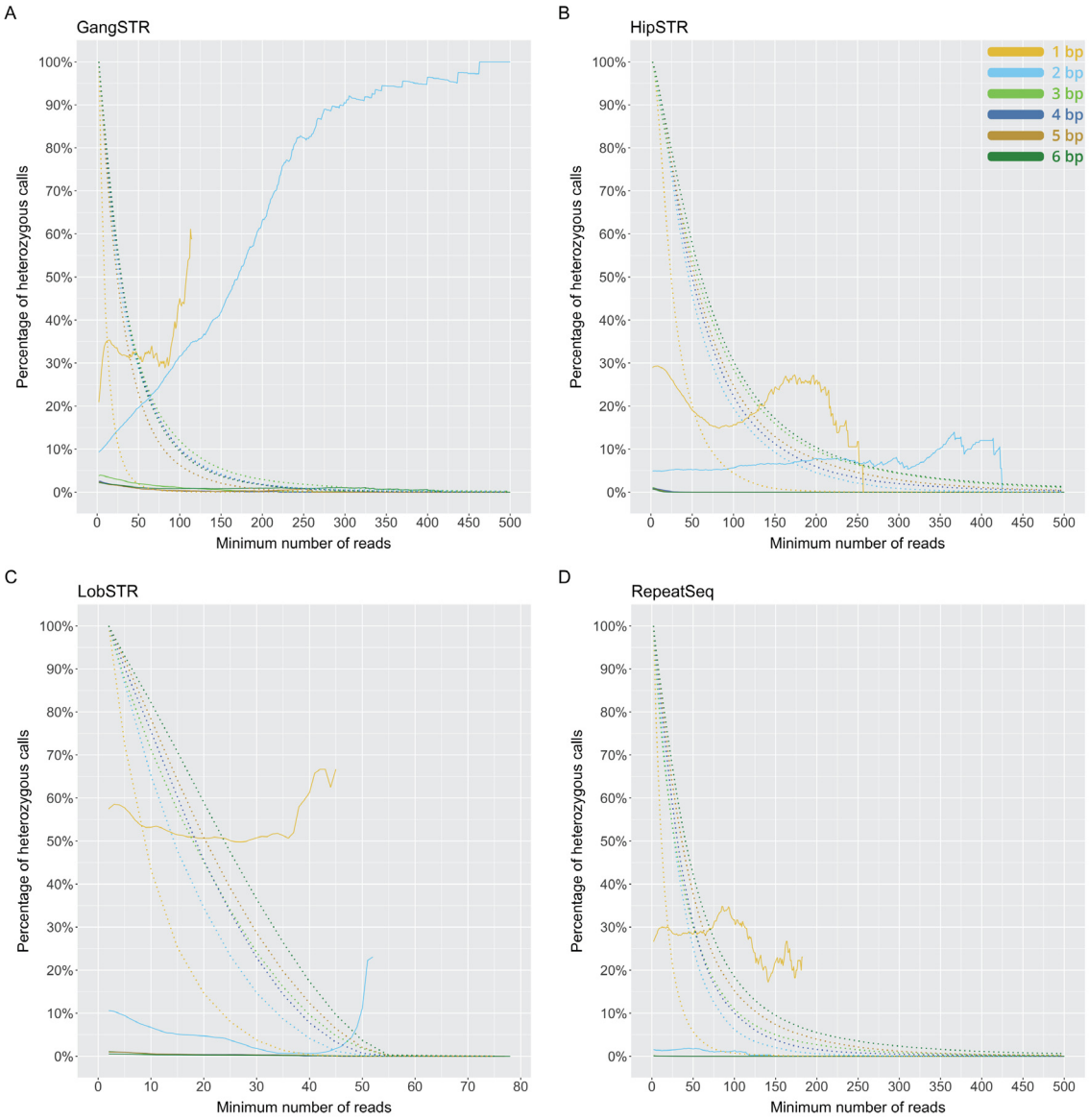


Figure 4. Percentage of heterozygous calls per minimum number of reads for each short tandem repeat genotyping tool. (**A**) GangSTR, (**B**) HipSTR, (**C**) LobSTR, (**D**) RepeatSeq. Solid line shows the percentage of heterozygous calls as a function of minimum number of reads. Dotted line represents the percentage of remaining calls as a function of minimum number of reads. Y-axis limited to 500 reads. Heterozygous calls are represented in percentages, but the total number of calls is different for each tool, where 100% is 796775 calls for GangSTR, 757432 calls for HipSTR, 848252 calls for LobSTR and 775030 calls for RepeatSeq.

Authors of RepeatSeq do not recommend any additional filtering and we found high accuracy even at low coverage. Filtering out all calls less than coverage of five will result in an error rate for 3–6 bp repeat units of 0.04–0.08% and only an average 7.1% of 2–6 bp repeat calls filtered out, leaving an average of 1728 reported calls per sample. Among all tools, RepeatSeq shows the best accuracy for dinucleotides, having an error rate no more than 1.81% (Figure 4D).

As mentioned, we saw unusually high error rates for dinucleotide repeats in nearly all tools, so we examined these repeats in more detail. Curiously, we found that for GangSTR and LobSTR, the unusually high error rate of 2 bp repeats were due to AC/TG repeats, while other repeat units do not exhibit these characteristics (Figure 5). The same pattern was observed for RepeatSeq but at lower error rates. This information is not easily available for HipSTR as it does not report the repeat unit.

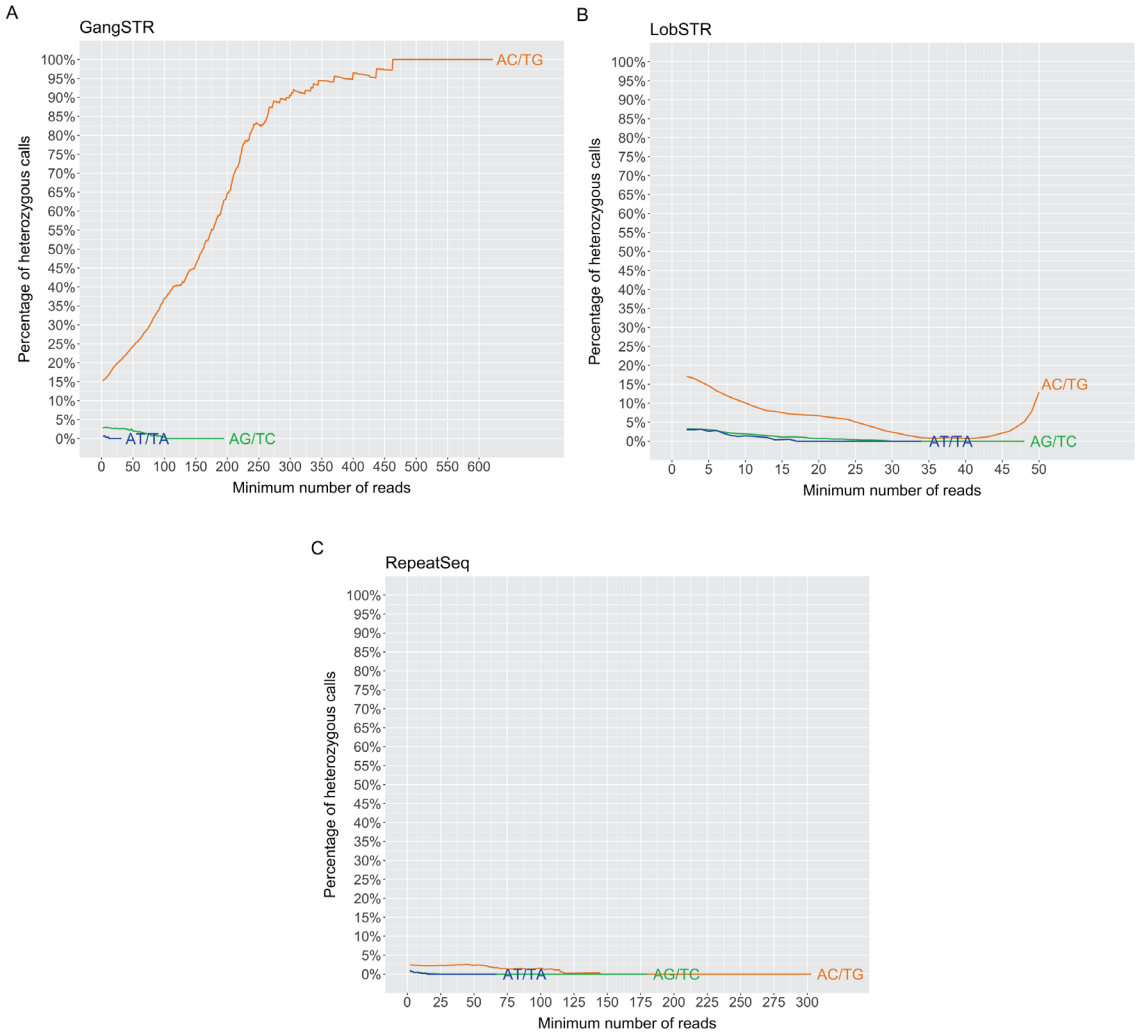


Figure 5. Percentage of heterozygous calls per minimum number of reads for dinucleotide repeats for (**A**) GangSTR, (**B**) LobSTR and (**C**) RepeatSeq, divided by the three possible repeat unit sequences.

Effects of repeat unit and quality scores on accuracy

Besides coverage, the second parameter which is commonly used for filtering in all tools is the call quality or quality score produced by each algorithm. We next investigated the effects of quality scores on accuracy by relaxing the coverage filtering and looking at the quality scores in different bins across the score range (Figure 6). GangSTR's level 2 filters recommend filtering out calls that have quality scores below 0.9. We see in Figure 6A that the heterozygous error rate fluctuates at lower scores and starts to decrease from quality scores above 0.6 for all repeats except for homopolymers. We reach a 1% heterozygous error rate when excluding calls that have quality scores below 0.81 for 3 bp repeat units, 0.66 for 4 bp units, 0.77 for 5 bp units and 0.93 for 6 bp repeat units. We see the lowest heterozygous error rate at the quality score of 1.0, but we also determined a sharp drop in the number of reported genotypes after excluding calls with quality scores below 1.0, and we also lose 61.7% of 3–6 bp repeats data on average. The recommended 0.9 seems a reasonable suggestion for balancing the accuracy with how much data we will have left after the filtering. We can also see that the accuracy of 1 bp repeat calls improves with the highest quality score (1.0) and a stronger filter for this repeat unit may be appropriate.

HipSTR, similarly to GangSTR, recommends filtering out calls with a quality score less than 0.9. In Figure 6B, we can see that on average the calls with a score below 1.0 have a high heterozygous error rate, and keeping only the ones with the highest quality score will improve overall accuracy. Indeed, 98.5% of calls have a quality score of 1.0 and therefore, we only lose a fraction of data while filtering out calls with a quality score below 1.0, which may be a good trade-off.

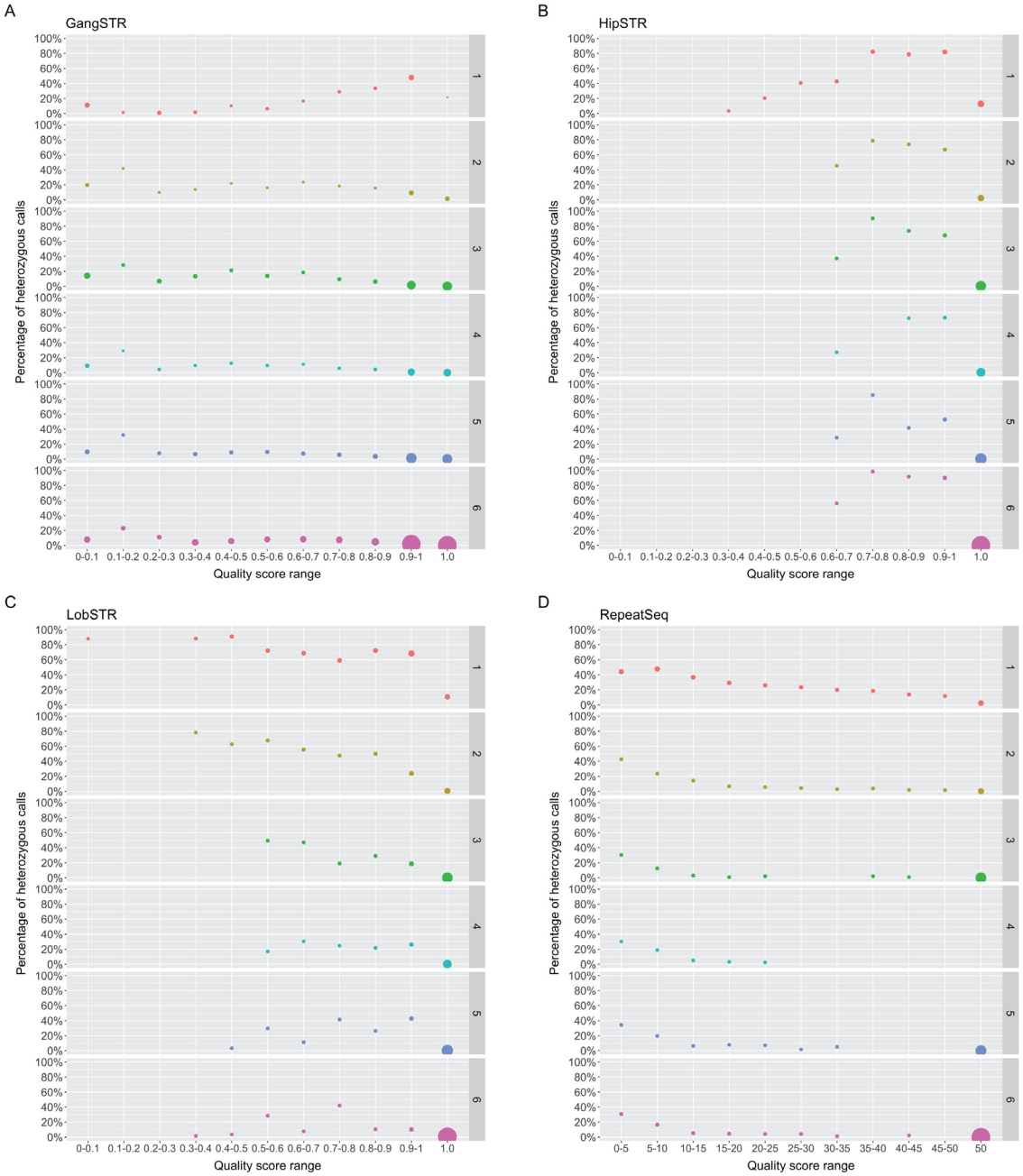


Figure 6. Percentage of heterozygous calls in different bins of quality score separated by each repeat unit length. (**A**) GangSTR, (**B**) HipSTR, (**C**) LobSTR, (**D**) RepeatSeq. Each bin is from marked range (inclusive) to the end value (exclusive), except the last one. Dot size represents the number of calls in that range of the repeat unit length (rows).

LobSTR recommends filtering out calls with a call quality score less than 0.8. As seen in Figure 6C, the heterozygous error rate fluctuates but generally shows higher accuracy with higher quality scores for 1–3 bp repeats, while for 4–6 bp there is no clear trend below a quality score of 0.9. There is no significant improvement in overall accuracy when we remove calls with quality scores less than 0.8, or even 0.9, which might be due to the fact that, similar to HipSTR, the majority of calls (94.3%) have a quality score of 1.0. We do see improvement when only leaving calls with quality of 1.0, particularly in 1–2 bp repeat units, which shows utility in filtering out the least accurate calls and since the majority of data has a quality score of 1.0, this filtering could be a good choice, as also suggested for HipSTR (see Supplementary Figure 3, *Extended data*) (Halman, 2020a).

RepeatSeq does not recommend any filtering and reports quality scores on a Phred scale (Figure 6D). We determined that filtering out calls with Phred quality score of 10 or less improves the accuracy of all repeat units. Accuracy of genotyping mono- and dinucleotide repeats continues to improve as a function of Phred scores, while the best accuracy is observed at the highest quality score. On the flip side, the number of calls also decreases, and at the highest Phred scores, we are left with 31.6% of homopolymers and 87.8% of dinucleotide repeats. However, it only filters out 1.2% of 3–6 bp repeats data. Overall, filtering data based on the quality scores may be reasonable if looking at mono- or dinucleotide repeats and accuracy is an important factor.

Accuracy of GangSTR by looking at only the enclosing class of reads

LobSTR, HipSTR and RepeatSeq use types of reads where the STR region has to be completely in the read. However, GangSTR uses more classes of reads that may give rise to false positives. Therefore, to make a more direct comparison, we decided to look at GangSTR results where we filtered out all other classes of reads besides the enclosing ones, marked here as GangSTR (enc.). Compared to the previous GangSTR results, we now see lower error rates for all repeat units as well as no substantial fluctuation in higher coverage that was apparent previously in Figure 4A. At a coverage of 20, GangSTR (enc.) has a heterozygous error rate of 0.53–0.99% for 3–6 bp repeats, while 58.6% of 2–6 bp repeats data is filtered out, with a median of 699 calls per sample (Figure 7A). When we increase the minimum number of reads to 50, we can see even further improvement. This results in an error rate of 0.01–0.27% for 3–6 bp repeats; however, this also filters out 86.3% of 2–6 bp repeats data (median of 152 calls per sample).

We also looked at the dinucleotides separately (see Supplementary Figure 4, *Extended data*) (Halman, 2020a) and found that it is considerably lower compared to the results of GangSTR. However, we determined that AC/TG repeats still have a higher error rate compared to other repeat units (Figure 4A and Figure 7A). Results of quality scores are quite similar to GangSTR's, where we see improvements at high quality scores (Figure 7B).

Running time of tools

Finally, we benchmarked all tools to determine their average running time. We ran each tool genome wide by using one, two, four and eight cores and determined that only RepeatSeq supports multithreading, which allows the

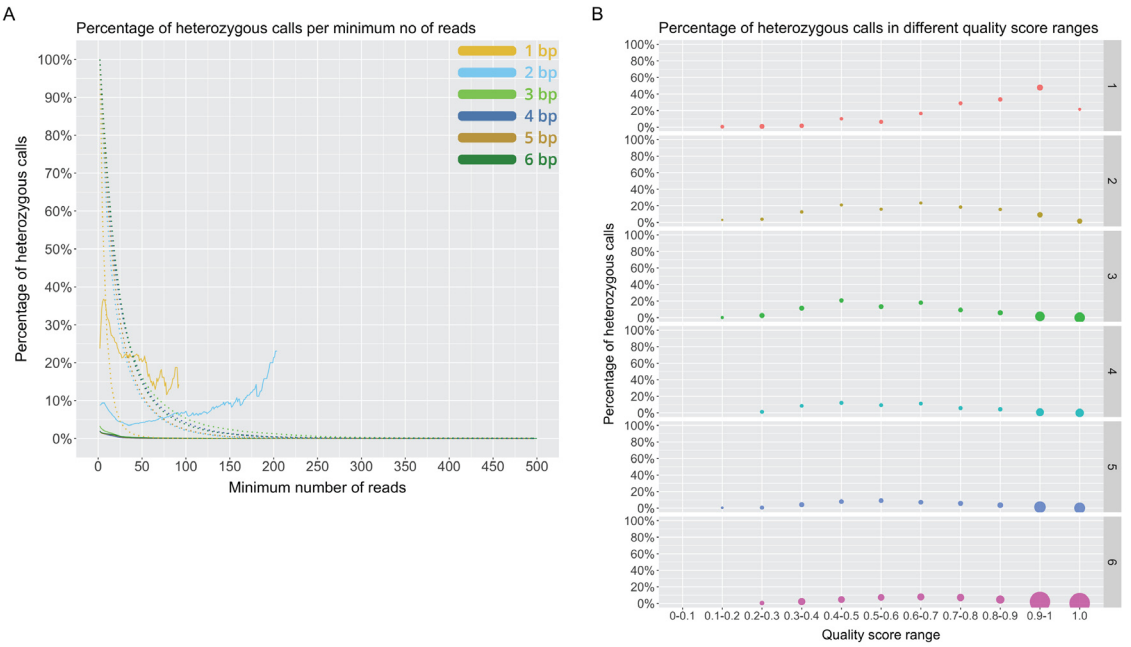


Figure 7. Percentage of (**A**) heterozygous calls per minimum number of reads and (**B**) in different quality score ranges for GangSTR (enc.). Solid line on (**A**) represents the percentage of heterozygous calls per minimum number of reads and dotted line represents the percentage of remaining calls per minimum number of reads. Dot size on (**B**) represents the number of calls in that range of the repeat unit length (rows) and each bin is from marked range (inclusive) to the end value (exclusive), except the last one.

tool to run faster when utilising more processor cores. In particular, we saw that for samples that have coverage on target regions of around 90x, the average running time on one core for LobSTR was 9 min 31 sec, giving it a clear advantage compared with HipSTR at 1 h 22 min, RepeatSeq 2 h 58 min and GangSTR at 3 h 4 min. By utilizing more cores, LobSTR's, GangSTR's and HipSTR's running time remained the same, while RepeatSeq's time improved.

Discussion

We used a novel way of evaluating the error in short tandem repeats genotyping methods where we analysed STR calls on the X chromosome of male samples. Because this is a hemizygous chromosome, we determined a relative error rate as the rate of heterozygous genotypes. We performed this evaluation on a human exome dataset of 433 samples, resulting in the evaluation of more than a million STR loci. Exome sequencing is widely used, but the PCR step in library preparation causes a challenge for STR genotyping tools due to the interference of stutter noise. This is the first independent evaluation of these STR genotyping tools that we are aware of.

Many of the tools do not claim to be able to accurately genotype homopolymers and we found that indeed all tools had difficulty with these repeats, resulting in a high error rate. There was also no clear correlation between minimum coverage and accuracy of genotyping homopolymers, but using the highest quality scores did improve the accuracy. Interestingly, most tools produced high error rates for genotyping dinucleotides as well, which we later found to be mainly caused by AC/TG repeat units. One who analyses dinucleotide repeats with these tools should be aware of the differences in accuracy of genotyping different repeat units and carefully interpret the results of AC/TG repeat units. Repeat units with a length of 3–6 bp were all relatively accurate and similar across tools, with only minor differences. However, genotyping was slightly less accurate for 3 bp length repeats in low coverage and low quality scores, but differences were reduced with proper filtering. We found that LobSTR was able to report the highest number of genotypes at a heterozygous error rate of less than 1%.

There are certain filtering parameters suggested for each tool and we examined the effects of coverage and quality scores across all tools. However, some tools have further parameters that could be explored that were not part of our investigation. In general, we found that higher quality scores increased the accuracy of results at the cost of losing some potentially accurate calls. The relationship with coverage was more complex but some coverage filtering improved results for all tools. Which parameters to use depends on the aim of the analysis. For example, more calls may be desirable to begin a screen, or more accuracy may be desirable if selecting potential disease associated loci. When one does an exploratory analysis to find potential loci of interest that can be followed up with alternative methods, then lowering filtering parameters for coverage and quality scores for certain tools could be a good approach as it leaves us with larger portion of data. We found that even in exome data, we can use these tools to genotype tens of thousands of loci.

Unlike the other tools we used in this analysis, GangSTR utilises four different types of reads, which can help to pick up the locus other tools cannot (such as those longer than the read length). However, these can also produce genotyping errors. In our analysis, we first looked at GangSTR results that included all four classes of reads and then we excluded all calls where only spanning or bounding class reads were present, as suggested by the tool authors. This filtering increased the genotyping accuracy of the tool (we also looked at the results where we skipped this filtering parameter but this did not improve results). Still, compared to other tools, GangSTR showed a higher error rate. Finally, we decided to look only at the enclosing class of reads as the other tools do and determined an error rate around three times lower at 20x and bigger gains at higher coverage. On the other hand, that change will result in losing the ability to genotype alleles longer than the read length, which is GangSTR's important addition. We also found that HipSTR has a very high accuracy for 3–6 bp repeats when coverage is at least 50x. Excellent accuracy was also found for RepeatSeq at very low coverage and this was the most accurate among the tools for genotyping dinucleotides. In addition, RepeatSeq is the only tool that supports multithreading and therefore can run faster by allocating more cores.

Here, we have presented one way of performing an evaluation and this approach does not look at accuracy of the estimated allele length, which is a limitation of the study. In addition, it is difficult to rule out a bias towards tools that default to genotyping an allele as a homozygous reference by the software. Our comparisons were specifically analysing an exome dataset that was PCR amplified, where a tools' noise model may play an important role. Therefore, tools may perform differently when we analyse PCR-free WGS datasets.

In conclusion, all these tools are built to genotype STRs but have different strengths and weaknesses. Based on our analysis there is no clear overall winner. RepeatSeq and HipSTR are the best when considering

genotyping error rate even with low coverage. On the other hand, GangSTR has an advantage because it is the only tool among them that can call alleles longer than the read length but shows a higher error rate, unless looking at only the enclosed class of reads, which in turn would lose the GangSTR's advantage of picking up long genotypes. In addition, GangSTR is the newest tool and so comes with reference files for different reference builds that are periodically updated according to the tool's webpage. The correct choice of a tool and the subsequent filtering depends on the aim of the analysis, and might be influenced by available hardware resources and time limit for running tools.

Methods

Dataset

In order to compare all STR tools, we ran each one of them on the same dataset. We used the data from the publicly available Simons Simplex Collection (SSC) for our analysis.

In total there were 238 families where only males were selected for our analysis to avoid heterozygous sites in the X chromosome, assuming that any multiallelic STR calls should be a result of PCR and/or sequencing errors. Male samples were determined by using metadata of samples (472 samples) and quality controlled by looking at the coverage on X and Y chromosomes. Results of the analysis led to the exclusion of three samples as they had no coverage on the Y chromosome. Out of the remaining 469 samples, we excluded six single-end read sequenced files as well as 28 paired-end read sequenced samples that did not work on GangSTR, and two additional samples that had issues with mapping, which left us with 433 samples in total (Supplementary Table 1, see *Underlying data*) (Halman, 2020a).

Genomic DNA of the final (433) samples used in this analysis was extracted from whole blood, exomes were captured with NimbleGen EZ Exome v2.0 (Roche Nimblegen, Inc., Madison, WI) reagents and sequenced using Illumina (San Diego, CA) GAIIx (N = 271) or HiSeq 2000 (N = 162) at the Yale University School of Medicine.

All computational steps (tools and parameters used) are described in this section.

SRA to FASTQ conversion

All whole-exome sequencing files in Simons Simplex families were downloaded from NCBI Sequence Read Archive (SRA) and converted to FASTQ files by using fastq-dump:

```
fastq-dump \
        --gzip \
        --skip-technical \
        --readids \
        --read-filter pass \
        --split-3 \
        --dumpbase \
        --clip \
        FILE.SRA
```

Defining STR regions

The human reference sequence hg19 (February 2009 assembly) was downloaded from UCSC and the "hg19.fa" file was created by and indexed using Samtools v1.10 (Li *et al.*, 2009):

```
cat *.fa > hg19.fa
samtools faidx hg19.fa
```

Creating FASTA sequence dictionary file for GATK analysis:

```
gatk CreateSequenceDictionary
        -R hg19.fa
        -O hg19.dict
```

Tandem Repeats Finder v4.09 (Benson, 1999) was used to find STRs (1–6 bp repeat unit length) in the hg19 reference genome using the following command and parameters:

```
./trf409.linux64 hg19.fa 2 7 7 80 10 24 6 -h
```

A custom-made Python script named trf2bed.py (Halman, 2020b) was used to extract data from the TRF output file to generate a BED regions file for LobSTR, GangSTR, HipSTR, RepeatSeq and GATK.

```
python3 trf2bed.py \
       --dat hg19.fa.2.7.7.80.10.24.6.dat \
       --bed hg19.fa.2.7.7.80.10.24.6_$TOOL.bed \
       --tool $TOOL
```

FASTQ alignment and calculating BAM coverage

Reads from FASTQ files were aligned to the hg19 reference genome using BWA-MEM v0.7.17 and aligned BAM files were merged and indexed using Samtools v1.10 (Li *et al.*, 2009).

```
bwa mem -M -t 8 -R "@RG\\tID:$id\\tPL:$PLATFORM\\tPU:$BARCODE\\tSM:$SAMPLE\\
tLB:$LIBRARY" hg19.fa $INPUT_FILE1.fastq $INPUT_FILE2.fastq | samtools sort
-@hg19.fa -o $OUTPUT_FILE.bam -

samtools merge $OUTPUT_FILE.merge.bam $INPUT_FILE1.bam $INPUT_FILE2.bam
samtools index $INPUT_FILE.merge.bam
```

To follow the best practices of GATK duplicate reads were removed:

```
gatk MarkDuplicatesSpark \
        INPUT=$INPUT_FILE.merge.bam \
        OUTPUT=$OUTPUT_FILE.bam \
        --remove-sequencing-duplicates \
        --create-output-bam-index
```

Coverage of BAM files on target regions was found with the MosDepth v0.2.4 tool, followed by calculating the median and average coverage:

```
mosdepth \
       -n \
       --fast-mode \
       --by $TARGET_REGIONS.bed \
       $OUPUT_FILE.coverage \
       $INPUT_FILE.merge.bam

gunzip -c $INPUT_FILE.regions.bed.gz | sort -n -k 5 | awk '{a[NR]=\
$5}END{print(NR%2==1)?a[int(NR/2)+1]:(a[NR/2]+a[NR/2+1])/2}' > $OUTPUT_
FILE.avgcov

gunzip -c $INPUT_FILE.regions.bed.gz | sort -n -k 5 | awk '{ sum += \$5;
n++ } END { if (n > 0) print sum / n; }' > $OUTPUT_FILE.medcov
```

Calling STRs and genotyping

GangSTR v2.4 (Mousavi *et al.*, 2019) was executed with the following parameters:

```
GangSTR \
        --bam $INPUT_FILE.merge.bam \
        --ref hg19.fa \
        --regions hg19.fa.2.7.7.80.10.24.6_gangstr.bed \
        --out $OUTPUT.vcf \
        --nonuniform \
        --coverage X*
* where X = mean coverage for the particular sample that was calculated by
MosDepth tool as described previously.
```

Strict filtering was done as recommended by the developer using dumpSTR, which is part of TRTools package:

```
dumpSTR \
        --vcf $INPUT_FILE.vcf \
        --out $OUTPUT_FILE \
        --filter-spanbound-only \
        --filter-badCI \
        --max-call-DP 1000 \
        --min-call-DP 50 \
        --min-call-Q 0.9
```

Since we were looking the relationship between coverage and quality scores, and genotyping accuracy separately, we did additional filtering (partial filtering), where we discarded the filtering on calls with low coverage or low-quality scores:

```
dumpSTR \
        --vcf $INPUT_FILE.vcf \
        --out $OUTPUT_FILE \
        --filter-spanbound-only \
        --filter-badCI \
        --max-call-DP 1000
```

RepeatSeq v0.8.2 (Highnam *et al.*, 2013) was executed with the following parameters:

```
repeatseq \
       $INPUT_FILE.bam \
       hg19.fa \
       hg19.fa.2.7.7.80.10.24.6_repeatseq.bed
```

LobSTR v4.0.6 (Gymrek *et al.*, 2012) was downloaded and a custom lobSTR reference was made using lobstr_index.py and GetSTRInfo.py scripts as follows:

```
python ./lobstr/scripts/lobstr_index.py
       --str hg19.fa.2.7.7.80.10.24.6_lobstr.bed \
       --ref hg19.fa \
       --out ./lobstr/hg19_custom/
python ./lobstr/scripts/GetSTRInfo.py \
       hg19.fa.2.7.7.80.10.24.6_lobstr.bed hg19.fa >       ./lobstr/hg19_
custom/lobstr_hg19_custom_strinfo.tab
```

LobSTR's allelotype was used to call STRs and it was run with default parameters, with and without the the "--no-rmdup" flag:

```
./lobstr/bin/allelotype \
       --command classify \
       --bam $INPUT_FILE.merge.bam \
       --index-prefix ./lobstr/hg19_custom/lobstr_hg19_custom_ref/lobSTR_ \
       --strinfo ./lobstr/hg19_custom/lobstr_hg19_custom_strinfo.tab \
       --noise_model ./lobstr/share/lobSTR/models/illumina_v3.pcrfree \
       --out $OUTPUT_FILE.vcf \
       --no-rmdup
```

Willems and colleagues explored the effects of recommended allelotype options for lobSTR (--filter-mapq0, --filter-clipped, --max-repeats-in-ends and --min-read-end-match), but found the optimal settings for lobSTR does not include these parameters and best results are obtained with default ones, which was also reported for RepeatSeq (Willems *et al.*, 2017), and therefore we decided to run both tools with the default parameters.

We did the strict filtering with the lobSTR's filtering tool, based on the author's recommendations for whole genome data:

```
python ./lobstr/share/lobSTR/scripts/lobSTR_filter_vcf.py \
       --vcf $INPUT_FILE.vcf > $OUTPUT_FILE.vcf \
       --loc-cov 5 \
       --loc-log-score 0.8 \
       --loc-call-rate 0.8 \
       --loc-max-ref-length 80 \
       --call-cov 5 \
       --call-log-score 0.8 \
       --call-dist-end 20
```

And the partial filtering:

```
python ./lobstr/share/lobSTR/scripts/lobSTR_filter_vcf.py \
       --vcf $INPUT_FILE.vcf > $OUTPUT_FILE.vcf \
       --loc-call-rate 0.8 \
       --loc-max-ref-length 80 \
       --call-dist-end 20
```

HipSTR v0.6.2 (Willems *et al.*, 2017) was executed with the following parameters:

```
HipSTR \
       --min-reads 2 \
       --def-stutter-model \
       --fasta hg19.fa \
       --regions hg19_2.7.7.80.10.24.6_hipstr.bed \
       --str-vcf $OUTPUT_FILE.vcf.gz \
       --bams $INPUT_FILE.merge.bam
```

Strict filtering was done according to the developer's recommendations:

```
python ./HipSTR/scripts/filter_vcf.py \
       --vcf $INPUT_FILE.vcf \
       --min-call-qual 0.9 \
       --max-call-flank-indel 0.15 \
       --max-call-stutter 0.15 \
       --min-call-allele-bias -2 \
       --min-call-strand-bias -2 > $OUTPUT_FILE.vcf
```

Since we ran HipSTR with "--min-reads 2" parameter, we additionally filtered out all calls that had less than 100 reads, as this is the default parameter that HipSTR uses.

Partial filtering was done:

```
python ./HipSTR/scripts/filter_vcf.py \
       --vcf $INPUT_FILE.vcf \
       --max-call-flank-indel 0.15 \
       --max-call-stutter 0.15 \
       --min-call-allele-bias -2 \
       --min-call-strand-bias -2 > $OUTPUT_FILE.vcf
```

GATK v4.1.2 was executed with the following standard parameters:

```
gatk HaplotypeCaller \
       --reference hg19.fa \
       --intervals hg19_2.7.7.80.10.24.6_gatk.bed \
       --genotyping-mode DISCOVERY \
       --input $INPUT_FILE.merge.bam \
       --output $OUTPUT_FILE.vcf
```

Data were analysed using GATK best practice guidelines (DePristo *et al.*, 2011) up to variant calling. Variant calling was performed with the HaplotypeCaller in GATK (Poplin *et al.*, 2017).

Data extraction from variant calling files (VCFs) and analysis

A custom-made Python script named extract-data.py (Halman, 2020b) was created to extract data from outputted VCFs of all tools ran:

```
python3 extract-data.py --tool *toolname --vcf ./vcf_folder --chr chrX --
out disk

* Where toolname was either gangstr, lobstr, hipstr, repeatseq or gatk
```

In particularly, only calls in the X chromosome were extracted out. In case of filtering, only calls that passed the filter were extracted out. The output file contained information about all STR loci found in the VCF file, having the following fields: sample name, locus, chromosome, start and end coordinates of the STR region, motif (repeat unit), length of motif, length of the reference, length of alleles, genotype, number of total reads, number of reads supporting the call in each class and the quality score.

The data was then analysed in R by using str-analyse.R together with str-analyse.functions.R script (Halman, 2020b). Bioconductor's GenomicRanges package for R (Lawrence *et al.*, 2013) was used to find and then filter out all calls that fell outside of the target regions. All STR regions that were entirely or partially inside of the target region were included in the analysis, however, all duplicate loci were removed. When calculating heterozygous percentage per minimum number of reads or quality score bins, we only included the

results when there were minimum of ten results (samples) to use and calculated the percentage again for each repeat unit length after each read or after 1/10 quality score bin.

Running time and multithreading

To see the performance of the STR specific tools we selected five WES samples that had a median coverage on target regions closest to 90x (between 88.6x and 91.6x) and calculated the time each tool ran on each sample individually by using either one, two, four or eight processor cores on the same server that has Intel(R) Xeon(R) 2.60 GHz processors and maximum of 16 GB RAM. Each test was repeated three times and the average time was calculated. Timing was performed with the UNIX time command.

An earlier version of this article can be found on bioRxiv (doi: https://doi.org/10.1101/2020.02.03.933002).

Data availability

Underlying data

The Simons Simplex Collection dataset, Accession number SRP010920: https://identifiers.org/insdc.sra:SRP010920

Harvard Dataverse: Supplementary information for the "Accuracy of short tandem repeats genotyping tools in whole exome sequencing data" article. https://doi.org/10.7910/DVN/RWTGWK (Halman, 2020a)

This project contains the following underlying data within the 'Supplementary_information' PDF:

- Supplementary Table 1 (list of all accession numbers used in the analysis)

Extended data

Harvard Dataverse: Supplementary information for the "Accuracy of short tandem repeats genotyping tools in whole exome sequencing data" article. https://doi.org/10.7910/DVN/RWTGWK (Halman, 2020a)

This project contains the following extended data within the 'Supplementary_information' PDF:

- Supplementary Figure 1 (Coverage of X and Y chromosome of samples marked as male in the metadata)
- Supplementary Figure 2 (Percentage of heterozygous calls over all samples with and without filters applied)
- Supplementary Figure 3 (Percentage of heterozygous calls as a function of minimum quality score)
- Supplementary Figure 4 (Percentage of heterozygous calls per minimum number of reads for dinucleotides)

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

Software availability

Source code available from: https://gitlab.com/andreassh/research-str-wes

Archived source code at time of publication: https://doi.org/10.5281/zenodo.3695971 (Halman, 2020b)

License: MIT

Acknowledgments

The authors are thankful to Katrina Bell for their comments and suggestions for the article.

References

Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res.* 1999; **27**(2): 573–80.
PubMed Abstract | Publisher Full Text | Free Full Text

Björn N, Pradhananga S, Sigurgeirsson B, *et al.*: **Comparison of Variant Calls from Whole Genome and Whole Exome Sequencing Data Using Matched Samples.** *Next Gener Seq Appl.* 2018; **5**(1).
Reference Source

Budiš J, Kucharík M, Ďuriš F, *et al.*: **Dante: genotyping of known**

complex and expanded short tandem repeats. *Bioinformatics.* 2019; **35**(8): 1310–17.
PubMed Abstract | Publisher Full Text

Cao MD, Tasker E, Willadsen K, *et al.*: **Inferring short tandem repeat variation from paired-end short reads.** *Nucleic Acids Res.* 2014; **42**(3): e16.
PubMed Abstract | Publisher Full Text | Free Full Text

Caspar SM, Dubacher N, Kopps AM, *et al.*: **Clinical sequencing: From raw data to diagnosis with lifetime value.** *Clin Genet.* 2018; **93**(3): 508–19.
PubMed Abstract | Publisher Full Text

Dashnow H, Lek M, Phipson B, *et al.*: **STRetch: Detecting and discovering pathogenic short tandem repeat expansions.** *Genome Biol.* 2018; **19**(1): 121.
PubMed Abstract | Publisher Full Text | Free Full Text

DePristo MA, Banks E, Poplin R, *et al.*: **A framework for variation discovery and genotyping using next-generation DNA sequencing data.** *Nat Genet.* 2011; **43**(5): 491–98.
PubMed Abstract | Publisher Full Text | Free Full Text

Dolzhenko E, van Vugt JJFA, Shaw RJ, *et al.*: **Detection of long repeat expansions from PCR-free whole-genome sequence data.** *Genome Res.* 2017; **27**(11): 1895–903.
PubMed Abstract | Publisher Full Text | Free Full Text

Dolzhenko E, Deshpande V, Schlesinger F, *et al.*: **ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions.** *Bioinformatics.* edited by Inanc Birol, 2019; **35**(22): 4754–56.
PubMed Abstract | Publisher Full Text | Free Full Text

Fungtammasan A, Ananda G, Hile SE, *et al.*: **Accurate typing of short tandem repeats from genome-wide sequencing data and its applications.** *Genome Res.* 2015; **25**(5): 736–49.
PubMed Abstract | Publisher Full Text | Free Full Text

Gymrek M: **A genomic view of short tandem repeats.** *Curr Opin Genet Dev.* 2017; **44**: 9–16.
PubMed Abstract | Publisher Full Text

Gymrek M, Golan D, Rosset S, *et al.*: **lobSTR: A short tandem repeat profiler for personal genomes.** *Genome Res.* 2012; **22**(6): 1154–62.
PubMed Abstract | Publisher Full Text | Free Full Text

Halman A: **Supplementary information for the "Accuracy of short tandem repets genotyping tools in whole exome sequencing data" article**. [dataset]. *Harvard Dataverse, V1.* 2020a.
http://www.doi.org/10.7910/DVN/RWTGWK

Halman A: **Source code for the "Accuracy of short tandem repeats genotyping tools in whole exome sequencing data" article**. [dataset]. *Zenodo.* 2020b.
http://www.doi.org/10.5281/zenodo.3695971

Hannan AJ: **Tandem repeats mediating genetic plasticity in health and disease.** *Nat Rev Genet.* 2018; **19**(5): 286–98.
PubMed Abstract | Publisher Full Text

Highnam G, Franck C, Martin A, *et al.*: **Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles.** *Nucleic Acids Res.* 2013; **41**(1): e32.
PubMed Abstract | Publisher Full Text | Free Full Text

Kozarewa I, Ning Z, Quail MA, *et al.*: **Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes.** *Nat Methods.* 2009; **6**(4): 291–95.
PubMed Abstract | Publisher Full Text | Free Full Text

Kristmundsdóttir S, Sigurpálsdóttir BD, Kehr B, *et al.*: **popSTR: population-scale detection of STR variants.** *Bioinformatics.* 2017; **33**(24): 4041–48.
PubMed Abstract | Publisher Full Text

Lawrence M, Huber W, Pagès H, *et al.*: **Software for Computing and Annotating Genomic Ranges.** *PLoS Comput Biol.* 2013; **9**(8): e1003118.
PubMed Abstract | Publisher Full Text | Free Full Text

Li H: **Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.** 2013; 1–3.
Reference Source

Li H, Handsaker B, Wysoker A, *et al.*: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics.* 2009; **25**(16): 2078–79.
PubMed Abstract | Publisher Full Text | Free Full Text

McKenna A, Hanna M, Banks E, *et al.*: **The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data.** *Genome Res.* 2010; **20**(9): 1297–303.
PubMed Abstract | Publisher Full Text | Free Full Text

Mousavi N, Shleizer-Burko S, Yanicky R, *et al.*: **Profiling the genome-wide landscape of tandem repeat expansions.** *Nucleic Acids Res.* 2019; **47**(15): e90.
PubMed Abstract | Publisher Full Text | Free Full Text

Poplin R, Ruano-Rubio V, DePristo MA, *et al.*: **Scaling accurate genetic variant discovery to tens of thousands of samples.** *BioRxiv.* 2017; 201178.
Publisher Full Text

Tang H, Kirkness EF, Lippert C, *et al.*: **Profiling of Short-Tandem-Repeat Disease Alleles in 12,632 Human Whole Genomes.** *Am J Hum Genet.* 2017; **101**(5): 700–15.
PubMed Abstract | Publisher Full Text | Free Full Text

Tang H, Nzabarushimana E: **STRScan: targeted profiling of short tandem repeats in whole-genome sequencing data.** *BMC Bioinformatics.* 2017; **18**(Suppl 11): 398.
PubMed Abstract | Publisher Full Text | Free Full Text

Tankard RM, Bennett MF, Degorski P, *et al.*: **Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data.** *Am J Hum Genet.* 2018; **103**(6): 858–73.
PubMed Abstract | Publisher Full Text | Free Full Text

Willems T, Gymrek M, Highnam G, *et al.*: **The landscape of human STR variation.** *Genome Res.* 2014; **24**(11): 1894–904.
PubMed Abstract | Publisher Full Text | Free Full Text

Willems T, Zielinski D, Yuan J, *et al.*: **Genome-wide profiling of heritable and *de novo* STR variations.** *Nat Methods.* 2017; **14**(6): 590–92.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ? ✔

Version 1

Reviewer Report 30 June 2020

https://doi.org/10.5256/f1000research.24995.r63748

© **2020 Anisimova M.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Maria Anisimova**
[1] Institute of Applied Simulation, School of Life Sciences and Facility Management, Zurich University of Applied Sciences (ZHAW), Waedenswil, Switzerland
[2] Swiss Institute of Bioinformatics, Lausanne, Switzerland

In recent years short tandem repeats (STRs) drew increased attention because of their frequent associations with disease, as well as their abundance and a diversity of roles in protein functions. Consequently, many methods were developed for detecting and genotyping tandem repeats. With wide variation in methodology, and its related settings, it is often hard to know which software can provide an optimal performance and be integrated in bioinformatics pipelines. The study by Halman and Oshlack “Accuracy of short tandem repeats genotyping tools in whole exome sequencing data” contributes some insights for scientists who wish to study short tandem repeats (STRs) loci in genomic data. The article is clearly written and should be practically useful for those developing relevant bioinformatics pipelines.

The authors evaluate four STR genotyping tools, they provide a clear description and reasoning for their choices and point out the main differences between the tools. Overall, the results are not surprising and could be expected: the evaluated tools all have different strengths and weaknesses and the choice should be driven by the specific goals of the analysis. Clearly, different filtering settings will produce results of different accuracy, with a trade-off between the genotyping accuracy and the highest number of calls. As expected, homorepeats and dinucleotide repeats are the most problematic and there is currently no tool providing a good solution for such STRs. Apart from the valuable main message, the quantitative results should not be generalized, these are more to get a “feeling” of the expected accuracy than any guarantee. Indeed, it is difficult to directly compare methods that differ in so many ways: conceptually, by assumptions of random noise model, alignment procedures, filtering options and cut-offs, etc. Nevertheless, I applaud the authors for trying to accomplish (at least partially) this difficult task.

At the same time, I would like to raise some points for reflection.
Note, that all evaluated tools required a set of defined STRs loci. These were detected using the TRF tool (Benson 1999)[1] based on self-alignment. In our article Schaper *et al.* (2012)[2], we have

evaluated several tandem repeat (TR) predictors, including TRF. As one can expect (and similar to the findings in the article reviewed above), due to the heterogeneity of the methods in terms of their main concepts and assumptions, we observed a wide spectrum of variable performances over the space of TRs. Specifically, for TRF, we observed that this predictor is very conservative – having a low false positive rate across the TR space (in terms of TR unit numbers and unit length) but also a very low power (high false negative rate). For repeats with very short units, TRF is extremely conservative. Therefore, such bias in STR detection by TRF may have consequences on the observed by the authors distribution of STR unit sizes (e.g., in Figure 1b, the overrepresentation of longer 6-nt units may be explained simply by the TRF-specific bias).

Therefore, I would be curious how the results of Halman and Oshlack would change, both relatively and quantitatively, were they to use a TR annotation tool with contrasting properties, i.e. that is better at detecting STRs specifically. For example, this could be Xtreme (Newman and Cooper, 2007)[3] or a meta TR predictor, each followed by statistical filtering (as we suggest in Schaper *et al* 2012[2] and also in Anisimova *et al.* 2015)[4]. These approaches typically suggest a TR unit distribution dominated by homo- and dinucleotide repeats (eg, see Delucchi *et al*. 2020[5]).

Finally, the authors also mention the current gap in methodology for indel calling. For repeat regions indels become particularly important - as a consequence of TR unit number variation, but also due to regular indel events. In this respect, while methods to tackle this problem are not available, one should be aware of this potential problem for their analyses, interpretation of results and overall conclusions.

References

1. Benson G: Tandem repeats finder: a program to analyze DNA sequences.*Nucleic Acids Res*. 1999; **27** (2): 573-80 PubMed Abstract | Publisher Full Text
2. Schaper E, Kajava AV, Hauser A, Anisimova M: Repeat or not repeat?--Statistical validation of tandem repeat prediction in genomic sequences.*Nucleic Acids Res*. 2012; **40** (20): 10005-17 PubMed Abstract | Publisher Full Text
3. Newman AM, Cooper JB: XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences.*BMC Bioinformatics*. 2007; **8**: 382 PubMed Abstract | Publisher Full Text
4. Anisimova M, Pečerska J, Schaper E: Statistical approaches to detecting and analyzing tandem repeats in genomic sequences.*Front Bioeng Biotechnol*. 2015; **3**: 31 PubMed Abstract | Publisher Full Text
5. Delucchi M, Schaper E, Sachenkova O, Elofsson A, et al.: A New Census of Protein Tandem Repeats and Their Relationship with Intrinsic Disorder.*Genes (Basel)*. 2020; **11** (4). PubMed Abstract | Publisher Full Text

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: computational molecular evolution, bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 17 June 2020

https://doi.org/10.5256/f1000research.24995.r63745

© 2020 Radvanszky J et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Jan Radvanszky
[1] Institute of Clinical and Translational Research, Biomedical Research Center, Slovak Academy of Sciences, Bratislava, Slovakia
[2] Science Park, Comenius University in Bratislava, Bratislava, Slovakia
[3] Geneton Ltd., Bratislava, Slovakia
Jaroslav Budiš
[1] Science Park, Comenius University in Bratislava, Bratislava, Slovakia
[2] Geneton Ltd., Bratislava, Slovakia
[3] Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia

Andreas Halman and Alicia Oshlack prepared and submitted a very interesting manuscript dealing with the results of their comparison of four selected STR genotyping tools (LobSTR, HipSTR, GangSTR and RepeatSeq), which are designed to process data from massively parallel sequencing (here used for exome sequencing data sets). More precisely, they compared five tools in the initial phases of the experiments, however, GATK HaplotypeCaller is not a dedicated STR tool and was disregarded in later phases of the study because of its several limitations. Although the study is well designed, well evaluated and also well written, it has one substantial limitation - the study design itself. The authors chose an elegant and low-cost, but still high-throughput approach that relies on the biological fact that healthy males (at least in humans and other mammals) bear one X chromosome in their genome, so they are hemyzygous for the major part of this chromosome

(when not considering the pseudoautosomal regions). Based on this phenomenon, they "*expect(*ed) *all calls* (coming from the X chromosome) *to be homozygous*" therefore they evaluated the proportion of "*incorrect calls revealed by a heterozygous call*" to determine the "*accuracy of genotyping*". Genotyping accuracy, however, has several aspects and the number of incorrectly called heterozygous positions (i.e. the false heterozygous rate) is only one part of it. The approach chosen by the authors, therefore, does not allow to determine the total genotyping accuracy, only the rate of false heterozygous calls, i.e. the ability of tools to discriminate between heterozygous and homozygous state. We need to mention here, that this limitation is clear from the study and is properly acknowledged by the authors in several places in the manuscript (both in the Introduction and also in the Discussion). We would like to emphasize, therefore, that we do not consider this limitation for an insufficiency of the study design. On the other hand, because of the inadequate wording chosen by the authors (*genotyping accuracy*) we consider the conclusions described by these words only partly supported.

We agree with the authors, however, that they present a "*different but complementary approach*" to evaluate the performance of STR genotyping tools, at least with respect to their ability to make a call, and to correctly assign homozygous calls to homozygous STR positions. We are sure that this approach will become a well established approach not only in future research studies, but also in everyday practice. Following thorough review of the manuscript, we would like to suggest some major and several minor points to be considered by the authors:

Major suggestions:

1. We would like to suggest rewording of "*genotyping accuracy*" (and similar formulations, such as "*accuracy of genotyping*", "*performance of genotyping tools*", etc.) with a more appropriate formulation - throughout the manuscript, everywhere, where they are used with respect to the results of the authors evaluation. Maybe with "*false heterozygous rate*"?

2. Page 10; Part "*Effects of repeat unit and quality scores on accuracy*" - the conclusions, according to which:
    - In case of GangSTR "*We can also see that the accuracy of 1 bp repeat calls improves with the highest quality score (1.0) and a stronger filter for this repeat unit may be appropriate*" , seems not to be correct, since there is a relatively wide range of lower quality scores in which GangSTR leads to lower false heterozygous rates (and also with larger dot sizes) - at least according to Figure 6a
    - For HipSTR and 1 bp repeats there is a similar exception from the general trend that " *keeping only the ones with the highest quality score will improve overall accuracy*"

3. Figure 5 and results for the dinucleotide repeats:
    - Why is the GC/CG dinucleotide missing from the compared possible dinucleotide sequences? If it is by mistake, it would be useful to add this dinucleotide to the comparison. If it has a specific reason, this should be discussed in the text.
    - Similarly to dinucleotides, there were also higher error rates found for homopolymers. It would be interesting to try to evaluate the error rate also for each of the four possible homopolymers and to see whether there are different trends between them (like the marked AC/TG trend with GangSTR and LobSTR).

Minor suggestions:

1. Page 3; First indent of Introduction: with regard to the standard methods to genotype STRs it would be better to change "gel electrophoresis" to "capillary electrophoresis", since gel

electrophoresis most commonly refers to agarose or polyacrylamide gels

2. Page 6; Third and fourth indents of the part called "*Dataset*": We would suggest to merge the two parts dealing with GATK ("*GATK-HC reported a median…*" and "*GATK only makes calls at positions…*") into one, because they are now unnecessarily separated by the part dealing with the "*…23 loci for which the reference STR length was longer…*"

3. Figure 2.: We recommend to:
    - Use the same order of tools in the part A) and B) of the figure
    - Include GATK in the charts, since GATK was also evaluated and compared in this step of the study (even if it resulted in small numbers of calls and its numbers does not necessarily reflect its ability to call genotypes, since it calls only alleles which are different from the reference genome). Although this will most probably lead to "simple rows" in the lower parts of the chart, because of the significantly lower numbers obtained with GATK, we believe that such inclusion will strengthen the visual recognition of the significant differences between the tools
    - We believe also that adding the respective numerical values for each tool (for example up to the top of each bar in the bar chart, and under the median in the box-plot) can further enhance the informational value of the charts
    - Provide also comparison of number of calls after applying the optimal filters (mentioned at the end of the "*Dataset*" section)

4. Page 7; Part "*Effects of repeat unit and coverage on accuracy*":
    - With few exceptions, nearly each experiment of the study is visualized through plots. We think that the number of heterozygous calls as a function of repeat unit length for each compared tool would deserve a plot too, since we believe that readers would appreciate the visualisation of these results
    - We think that the last sentence on the page ("*While dinucleotide repeats showed a trend towards a lower error rate with increasing coverage for RepeatSeq and LobSTR, the trend was reversed for GangSTR*") does not describe accurately the complex trends visualised by the plots on Fig.4 (for LobSTR and HipSTR). LobSTR for example shows a lowering trend only up to a certain point, where it sharply increases. This seems to be, moreover, an opposite trend to that seen for HipSTR, which has a rather increasing error rate (although with little fluctuations) up to a certain point when it suddenly drops
    - Do the authors have an explanation for the sudden change in trends encountered in the LobSTR and HipSTR results for mono- and dinucleotide repeats?

5. Figure 4. (and other line plots): The lines in the plots are difficult to read, especially in a printed form. We suggest reworking of the graphs for better clarity, for example with thicker trend lines, white background instead of grey, etc.

6. Figure 6: Can the authors include an explanation (or at least a hypothetical reason) to the "strange" trend encountered in the HipSTR results?

7. Page 13; The very last sentence of the "*Results*" section: We recommend to specify the respective time also for RepeatSeq, determined after the utilisation of more cores (*while RepeatSeq´s time improved to ….*)

8. Although they comprise only a relatively small part of X chromosome, the pseudoautosomal regions (PAR1 and PAR2) should be mentioned in the study. Together with the information, whether these regions were excluded or included in the target regions for the present study - alternatively, whether any of the called STR loci mapped to these regions (and if yes, then their proportion to the remaining region should be mentioned)

9. Page 13; "*Discussion*": We suggest to move the actual fifth indent ("*Here, we have presented one way of performing…*") after the first one ("*We used a novel way…*"), because together they will form a more consistent "introduction" to Discussion.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Partly

Competing Interests: JR and JB are authors of an STR genotyping tool called Dante which is, however, not commercialized and is not compared in the evaluated manuscript

Reviewer Expertise: JR - human molecular genetics, genomic technologies; JB - bioinformatics, computational biology

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Reviewer Report 26 May 2020

https://doi.org/10.5256/f1000research.24995.r61590

© 2020 Worthey E. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Elizabeth A. Worthey
Center for Computational Genomics and Data Science, Departments of Pediatrics and Pathology, School of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA

This article compares a number of tools released to characterize the presence and nature of repetitive elements from genomic data. The stated goal of the analysis is to generate findings that would help other researchers choose suitable tools and parameters when performing this type of analysis in their own lab. Aspects considered included the ability to call as well as accurately genotype repetitive elements. The analysis is not limited to long disease associated repeats, which has often been the case for obvious reasons.

The authors compared four different short tandem repeats genotyping tools making use of whole exome sequencing data. They sought to determine genotyping performance and limits. They also examined parameters settings to determine if and how to refine the analysis to increase accuracy. The authors made use of a readily accessible dataset for these analyses.

As would be expected, they found that all tools showed reasonable performance when genotyping repeats of 3-6 bp in length. As would be expected, they found that accurate homopolymer genotyping was challenging for all tools, with a high error rate being seen cross the board. They found issues nin mono and dinucleotide repeats as would also be expected/has been shown previously. The authors categorized tools based on a number of factors. For example they found that LobSTR made the most calls (which may or may not be a good thing) and was the fastest tool, while RepeatSeq and HipSTR exhibited the lowest heterozygous error rate at low coverage.
The methods, tools reviewed, and dataset used for testing were clearly presented. The reasoning for selecting those tools was also well laid out. The article cited the current literature in the field.
The study design was appropriate for the purpose of comparing tools to define specific strengths and weaknesses. It might have been nice to have more of an in depth discussion as to how the various tools differed in terms of their algorithm and how this related to the results and conclusions made. The authors stated that a goal of the analysis would be to generate a framework useful for others looking to compare and contrast tools to support selection for research. As such it might have been nice to include some specifics on ease of installation, levels of support etc. for the various tools etc. Maybe they were all very easy and well supported in that regard.

The authors provide specific details of the parameters used to run each tool. The majority appear to be run using defaults; there does not seem to have been a huge amount of parameter refinement . This may be due to limitations in what is possible with these tools. The authors provided links to the datasets used for the comparison. The source data used is not provided per se because it is access controlled. The data is, however, a well known publicly available dataset and it should not be very difficult for others to obtain access. People wishing to recapitulate the study would simple have to apply. The authors provided links to the code they used. All combined; these aspects should make for ease of replication by those interested.

Very minimal statistical analysis and its interpretation was required because of the nature of the analysis presented, but zenodo links to scripts were provided.
The conclusions drawn were adequately supported by the findings that they presented.
Testing of the X chromosome for accuracy in variant calling is not a novel approach. Study of the X to explore the genotype and location of repetitive sequences is also not a novel approach.

Combining the two approaches to use the X chromosome to specifically examine repetitive sequences has not been published.

The study is essentially a comparison of tools. As such, it is well organized and executed and would be of interest for people looking to set up pipelines in their lab. As the authors acknowledge most findings gleaned from their analysis were not surprising; they primarily confirmed known constraints or issues with the tools and their analysis confirmed existing knowledge regarding the nature of the repetitive sequences studied. Selection of the X chromosome was a nice way to undertake analysis that would be likely to be adopted by others. Given this background, this work would seem to have value primarily for education purposes to help researchers select and set up tools. It would be useful for that purpose. It does not present much, if anything, in the way of novel generalizable research knowledge on the repetitive nature of the human genome.

Is the work clearly and accurately presented and does it cite the current literature?
Yes

Is the study design appropriate and is the work technically sound?
Yes

Are sufficient details of methods and analysis provided to allow replication by others?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Yes

Are the conclusions drawn adequately supported by the results?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biology. Genetics.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research