BMC Genomics

RESEARCH

Open Access

Detection of short tandem repeats in the cattle genome: a comparison of bioinformatic tools

Quan Hong Tran[1,2*], Tuan Viet Nguyen[1], Jianghui Wang[1], Iona M. MacLeod[1,2*] and Amanda J. Chamberlain[1,2*]

Abstract

Background Short tandem repeats (STRs) are repetitive DNA sequences with 1–6 nucleotide repeat units, exhibiting high polymorphism due to varying repeat counts. STRs are more variable than SNPs and can cause genetic disorders. With population-scale cattle whole-genome sequencing data available, whole-genome STR identification has attracted new interest, but challenges remain due to the lack of standardized methods, sequencing data limitations, and the diversity of STR-calling tools. This study compared six STR-calling tools: HipSTR, GangSTR, and ExpansionHunter for short-read data, and Straglr, RepeatHMM, and LongTR for Oxford Nanopore (ONT) long-read data—using sequences from five Holstein cattle (two parent–offspring trios with a shared sire). This is the first cattle study to evaluate short- and long-read STR callers using both data types from the same animals.

Results In short-read data, ExpansionHunter identified the highest number of polymorphic STRs (pSTRs) (327,690), followed by HipSTR (205,900) and GangSTR (110,680), with 93,023 loci detected by all three tools. In long-read data, LongTR detected 470,250 pSTRs, RepeatHMM 224,185, and Straglr 90,275, with only 33,253 loci shared among them. Mendelian consistency of STR genotypes in the trio offspring was high (> 0.8) for all short-read tools, with HipSTR and GangSTR highest at 0.98. LongTR was the only long-read tool with high consistency (0.88). Short-read tools also showed higher concordance in STR genotypes among themselves than was observed among long-read tools. However, long-read tools had a clear advantage in detecting large STRs. Relative to computational efficiency, HipSTR and GangSTR (short-reads), and LongTR (long-reads) required less memory and shorter runtimes than the other tools.

Conclusions Tool selection is critical for accurate whole-genome STR identification in cattle. For short-read data, HipSTR showed relatively high Mendelian consistency and concordance compared to the other tools, while ExpansionHunter was able to detect longer STRs but with lower Mendelian consistency. For long-read data, LongTR demonstrated higher consistency and computational efficiency relative to the other tools. Based on these results, HipSTR and LongTR are suggested as preferred options for short-read and ONT long-read datasets, respectively,

*Correspondence:
Quan Hong Tran
quanhong.tran@agriculture.vic.gov.au
Iona M. MacLeod
iona.macleod@agriculture.vic.gov.au
Amanda J. Chamberlain
amanda.chamberlain@agriculture.vic.gov.au

Full list of author information is available at the end of the article

in cattle STR analysis. These recommendations are based on the metrics observed in this study, and confirmatory analyses across additional breeds, larger sample sizes, and validated truth sets are encouraged.

Keywords Short tandem repeat, Cattle, HipSTR, GangSTR, ExpansionHunter, LongTR, Straglr, RepeatHMM

Background

Short tandem repeats (STRs), also known as microsatellites or simple sequence repeats, are repetitive DNA sequences with repeat motifs ranging from 1 to 6 nucleotides. These sequences are widely distributed throughout the organism's genome and demonstrate high polymorphism [1], as shown by differences in the number of repetitions of each motif between individuals. During the late 1990 s and early 2000 s, with the advancements in molecular genetics techniques, STRs became very popular and reliable genetic markers for marker-assisted selection and quantitative trait loci (QTL) discovery studies in cattle [2]. The interest in STRs generally declined in cattle after 2010 due to the development of single nucleotide polymorphism (SNP) arrays. Compared to STRs, SNP arrays offered several advantages, including higher density, higher throughput, and scalability. However, there has been renewed interest in STRs in recent years, as potential causal variants for complex traits. In human studies the expansion of certain STRs has been linked directly to various genetic disorders and complex diseases [3–5]. In addition, while most Genome-Wide Association Studies (GWAS) focus primarily on biallelic SNPs, recent studies have shown that STRs have low linkage disequilibrium (LD) with nearby SNPs, and genome-wide STRs can contribute a new source of genetic information to association studies [6]. The incorporation of genome-wide STRs in the GWAS alongside SNPs and INDELs have been explored in some studies in humans, pigs and cattle [6–12]. Several of these studies reported multiple STRs significantly associated with gene expression and other complex traits, including one in cattle where STRs were more than twofold enriched among expression and splicing QTLs compared to SNPs [7]. Additionally, Margioliash and colleagues estimated that STRs may account for 5%–7% of the causal variants identifiable from GWAS for 44 different human blood phenotypes [10].

Due to the high variability in both length and sequence, accurate genotyping of STRs is challenging. Traditional methods like polymerase chain reaction (PCR) and Southern blot assays are labour-intensive and thus limited to a small set of STRs. Next-generation sequencing technologies provide a powerful approach for identifying STR variants across the entire genome. However, unlike SNP and small INDEL discovery, which have standardized genotype calling formats, STR genotyping is more complicated, because differences in sequencing technologies and bioinformatics tools lead to inconsistencies in identifying and interpreting STR variants and their alleles. The determination of STRs from sequencing data can be classified into two main methods: alignment-based and de novo-based. Alignment-based methods typically involve interrogating reads that have been aligned to a reference sequence and uses a list of known STR positions. This approach is fast and suitable for population-level studies, because it focusses solely on reads spanning known STR loci. De novo-based approaches enable the identification of novel STRs directly from sequencing reads without requiring a reference genome and can detect repeats based on assembly graphs. Several tools have been developed to identify STRs from short-read data using alignment-based approaches, such as lobSTR [13], HipSTR [14], GangSTR [15], ExpansionHunter [16], as well as de novo-based methods, such as REPdenovo [17] RepAHR [18]. Whether alignment-based or de novo-based, STR identification tools based on short-read sequencing are often limited by the length of the read, which is typically between 100 and 150 base pairs (bps). However, some STRs are complex, potentially spanning hundreds to thousands of bases with hundreds of repetitions. Many of these larger STRs have been reported as pathogenic in human [19], for example, the expansion of over 200 CGG repeats in *FMR1* gene has been associated with the fragile X syndrome, while 6 to 55 repeats present as a normal phenotype [20]. It can be challenging to accurately determine the allele length variations with short-read data for these large STRs. Long-read sequencing technologies, like Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), provide the ability to address these challenges. These platforms generate sequence reads that are typically thousands to tens of thousands of bases long, and therefore often cover the entire STR region. Long, contiguous reads improve the accuracy of sequence alignment, allow for more precise determination of repeat numbers, and preserve phase information for determining haplotypes thus improving genotyping and imputation accuracy. Several methods for identifying STRs from long-read data have been developed recently, including both alignment-based and de novo-based approaches. A comprehensive summary of available STR genotyping tools for both short-read and long-read data was provided in the recent review by Tanudisastro et al. [21], which highlights how long reads from Oxford Nanopore and PacBio overcome limitations of short reads by spanning full repeat regions and enabling more accurate characterization of repeat expansions, while also noting challenges such as sequencing

error profiles and the need for specialized algorithms tailored to long-read data.

While several computational tools have been developed for genotyping STRs from short read and long read sequencing data, there have been few independent comparisons of their performance. Several STR genotyping tools were developed and tested on a small set of known STR loci in humans, primarily for distinguishing pathogenic variants from non-pathogenic ones, however their utility for genome-wide STR genotyping in population studies has not been tested [22–24]. Furthermore, all previous comparisons of STR detection tools have focused on human sequence data, for which the reference genome is more complete and well-annotated. Until now, to our knowledge there have been only three published studies genotyping bovine STRs genome-wide. One early study used high-coverage short-read sequencing of five Holstein bulls to generate an STR map using lobSTR [25] and two more recent studies used HipSTR on larger short-read datasets [7, 26]. All of these studies were based on short-read sequencing data and used a single short-read tool, so they did not evaluate the effect of tool selection and were limited in their ability to detect and assess long STRs.

This is the first study in cattle to use matched short-read and long-read whole-genome sequencing data from the same animals, with the purpose of systematically comparing STR genotyping tools for STR discovery and genotyping across sequencing technologies. The findings serve as a reference for selecting appropriate sequencing tools and data for future STR studies in cattle populations with large numbers of individuals. In this study, two parent–offspring trios with sequencing data from both Illumina short-read and ONT long-read platforms were used for benchmarking six STR genotyping tools. HipSTR, GangSTR, and ExpansionHunter were chosen for short-read data, while Straglr, RepeatHMM, and LongTR were chosen for long-read data. These tools were selected based on the following criteria: they exhibit strong performance in previous benchmark studies using human sequence [27–29], have been updated within the past five years, could use alignment files from popular alignment tools (such as bwa [30] for short-read data and minimap2 [31] for long-read data) as input, and could use the same predefined list of STR loci. In multiple human benchmarking studies, comparisons of STR genotyping tools showed that HipSTR, GangSTR, and ExpansionHunter all produced reliable genotypes across a range of repeat unit sizes, with each tool showing different strengths depending on repeat context and filtering thresholds, demonstrating that tool choice and filtering affect heterozygous call rates and error rates in STR genotyping [27–29]. Some other tools were excluded from the list for various reasons, such as not being updated for an extended period (e.g., lobSTR [13]), being primarily intended only for HiFi PacBio data (TRGT [32], Replong [33]), designed for a limited set of STRs rather than whole-genome analysis (NanoRepeat [22]) and requiring uncommon input formats—for example, TRiColor [34] needs haplotype-resolved BAM files, and Tandem-genotypes [35] required alignments produced by LAST [36], which are time-consuming to generate. A summary of the selected STR genotyping tools, including read type, input requirements, output format, genotype quality, and allele length reporting, is provided in Table 1. Given no STR truth set currently exists for cattle, we relied on Mendelian inheritance in the sequenced trios to estimate genotyping accuracy.

Table 1 Summary of STR genotyping tools evaluated in this study with short-read (SR) and long-read sequence (LR). Output shows the results format (Single VCF = per sample, Merge VCF = combined, Table = tabular output); Genotype Quality indicates whether a genotype quality score is reported; Allele Length shows how STR allele sizes are reported (REPCN = repeat copy number, AL = allele length in base pairs, GB = the base pair differences of the genotype from the reference)

Tool	Read Type	Input Files	Output Files	Genotype Quality	Allele length
HipSTR	SR	Bam + bed	Merge VCF	Yes	GB
GangSTR	SR	Bam + bed	Merge VCF	Yes	REPCN
ExpansionHunter	SR	Bam + catalog (.json)	Single VCF	No	REPCN
Straglr	LR	Bam + bed	Single VCF	No	AL
RepeatHMM	LR	Bam + bed	Table	No	REPCN
LongTR	LR	Bam + bed	Merge VCF	Yes	GB

Methods

Dataset and bam preparations

DNA samples from five Holstein dairy cattle, forming 2 parent–offspring trios (with a shared sire), were extracted from either liver tissue (dam and offspring) or semen (sire) using Puregene tissue kit (Qiagen). Tissue samples for the dam and offspring were collected according to the Australian Code of Practice for the Care and Use of Animals for Scientific Purposes [37] as approved by the Agricultural Research and Extension Animal Ethics Committee of the Department of Environment and Primary Industries (application number 2014–23, approved on the 18th December 2014). The sire was owned by a commercial artificial breeding company and the semen samples were collected as a part of routine business operations and donated to Agriculture Victoria for research purposes. For short-read sequencing, the libraries were prepared using TruSeq DNA PCR-Free HT Kit (Illumina)

and were sequenced on the HiSeq3000 genome analyser (Illumina) in a 150-cycle paired end run. For long read sequencing, DNA libraries were prepared using the Genomic DNA Ligation kit (SQK-LSK109, Oxford Nanopore Technology), then each library was loaded onto a R9.4.1 flowcell and sequenced on a PromethelON (Oxford Nanopore Technology). The sequencer was run for 72 h and FAST5 files were saved for re-basecalling. All DNA extraction, library preparation and sequencing steps followed the manufacturer's instructions.

Short read paired-end fastq files were trimmed of adapter sequence (ILLUMINACLIP:fasta:2:30:3:1:true) and poor quality bases (q < 20) and reads with mean qscore less than 20 or length less than 35 bp were filtered out using Trimmomatic v0.38 [38]. Cleaned paired and single reads were then aligned to ARS-UCD1.2 reference genome [39] using bwa-mem [30] with default settings. The resulting bam files were then sorted and indexed using Samtools-1.8 [40]. The mean coverage of the BAM files ranged from 13.8X to 22.2X.

FAST5 files from the PromethelON were basecalled using Guppy v6.4.6 [41] in super high accuracy mode (SUP) with default settings. The output fastq files were trimmed using FiltLong (https://github.com/rrwick/Filtlong) with 2 options: –keep_percent = 90 (keep only the top 90% of reads based on quality) & –mean_q_weight = 10 (gives the mean quality score a weight of 10, prioritizing reads with higher accuracy). Cleaned reads were then aligned to the ARS-UCD 1.2 reference genome [39] by Minimap2 [31] with default parameter settings. Resulting alignments were sorted and indexed by Samtools-1.9 [40]. The average coverage of BAM files for ONT data ranged from 40.3X to 43X.

Reference-guided STR discovery

Because all six benchmark tools required a predefined list of STRs, we constructed a reference STR list according to the HipSTR workflow as follows. Tandem Repeat Finder software [42] was used to scan the cattle reference genome (ARS-UCD1.2) for tandem repeats using default parameters, except that the minimum score was set to 10 (default is 50), which is more suitable for detecting short repeats. The default minimum score of 50 in TRF requires at least 25 matching bases, so it would miss many STRs; for example, a 5-bp motif with only 4 copies (20 bases) or a 2-bp motif with 12 copies (24 bases) would not be detected. The TRF results were converted to BED format, removing low-quality entries and repeat motifs longer than 6 bp, following the instructions provided by the HipSTR authors (https://github.com/HipSTR-Tool/HipSTR-references/tree/master/human). We also filtered out repeats on the sex chromosomes and removed mononucleotide repeats because homopolymer regions show high sequencing error rates in both short-read Illumina [43] and long-read ONT data [44]. The performance of STR-calling tools for these repeats cannot be reliably evaluated with our input data. For population scale whole-genome STR studies, we expect that removing these STRs from the panel will have minimal impact on downstream applications such as imputation or GWAS. After all filters were applied, the remaining list was converted into a BED file with the corresponding format required for each tool.

STR genotyping in trios

For short-read data, STRs were genotyped using HipSTR v4.0.6 [14], GangSTR v2.5 [15] and ExpansionHunter v5.0.0 [16]. HipSTR was run with the BAM files from all five animals together in one run with the –def-stutter-model option, which is designed for STR calling using a reference panel, and with –min-reads = 2, to retain loci with at least two reads present across all individuals. GangSTR was also run with the five samples together in one run, while ExpansionHunter had to be run with each sample separately in Streaming mode, and the variant call format (VCF) outputs of ExpansionHunter were merged using MergeSTR [45]. All other parameters of HipSTR, GangSTR, and ExpansionHunter were set as default.

For long-read data, genotyping was performed using RepeatHMM v2.0.3 [46], Straglr v1.5.0 [47], and LongTR v1.0 [48]. RepeatHMM Scan was run separately for each sample, with the minimum number of support reads set to 4, with other settings at default. The output of RepeatHMM consisted of tables containing the position, STR motif, and number of repeats for each allele. STRs with a repeat count of 0/0 were considered missing genotypes. LongTR was run with all five samples together and produced a merged VCF file. Straglr was run separately for each sample with parameter –min_cluster_size = 1 as recommend from the author. The VCF output files of Straglr were then merged using BCFtools [40] and filtered with VCFtools [49].

For STR filtering in our benchmark, coverage (read depth or supporting reads) and genotype-quality values were the main filters. Among the six tools, only HipSTR, GangSTR, and LongTR provide a genotype-quality score. For HipSTR and GangSTR, we applied a quality cutoff of 0.9, matching the strict filtering recommended by the authors [14, 15]. For LongTR, we used the same quality cutoff of 0.9. Although LongTR does not recommend a fixed threshold, the authors mentioned that quality score is the most reliable indicator of call accuracy, and based on their results, a cutoff near 0.9 retains high-confidence calls [48]. ExpansionHunter, Straglr, and RepeatHMM do not output a genotype-quality score. They rely instead on internal filters based on coverage or supporting reads, which depend strongly on the sequencing depth. To keep filtering consistent, we used only a minimum coverage

requirement across all tools: Read depth (DP) ≥ 4 for HipSTR, GangSTR, LongTR and Straglr, Locus coverage (LC) ≥ 4 for ExpansionHunter and minimum number of support reads = 4 for RepeatHMM. The lack of a genotype quality score in ExpansionHunter, RepeatHMM, and Straglr is an important limitation. Genotypes that were filtered out were converted to missing (./.) in the final VCF files. Only STRs that were polymorphic across the five samples were kept for downstream analysis.

Mendelian consistency

To determine the accuracy of each tool, we used the data from the 2 trios to evaluate consistency with Mendelian inheritance for all polymorphic STRs (pSTRs). Genotypes of pSTRs were extracted from VCF output of HipSTR, GangSTR, ExpansionHunter, LongTR, and Straglr, and Mendelian consistency was then calculated for offspring genotypes after excluding any loci with a missing genotype in at least one member of the trio. In the case of RepeatHMM, because this tool only outputs tables containing the number of repeat motifs instead of genotype, we used the number of repeats as the allele's genotype.

Comparison of concordance in STR allele length

To calculate allele lengths, different approaches were used depending on the output of each tool. For GangSTR, ExpansionHunter, and RepeatHMM, allele length was computed as the number of repeat motif copies (REPCN in the FORMAT field of the VCF file) multiplied by the motif length. For HipSTR and LongTR, allele length was calculated as the length of the reference sequence plus the base pair differences of the genotype from the reference (GB in the FORMAT field of the VCF file). Straglr directly reported allele length (AL in the FORMAT field of the VCF file). GangSTR, ExpansionHunter, and RepeatHMM retained the position and motif of STRs from the BED file in their output. The VCF files from HipSTR and LongTR sometimes reported STR start positions shifted by a few nucleotides relative to the BED file, but the entire STR region in the VCF overlapped the corresponding BED region and the repeat unit length in the VCF matched that in the BED file; therefore, the STR positions in the VCF were adjusted to the start positions in the BED file, ensuring uniform comparisons across tools. For Straglr, only STRs matching both the position and motif length in the BED file were retained for concordance analysis. To calculate the concordance in allele length across each pair of tools, we compared the two alleles reported for the same individual at the same pSTR sites detected by both tools. A single nucleotide difference in allele length was allowed when considering two alleles concordant. The concordance rate was calculated as the number of concordant alleles divided by the total number of alleles evaluated for each pair of tools.

Running time and maximum memory used

To compare the running time and peak memory use of the six tools, we executed each program on our HPC system running CentOS 7 with the Slurm scheduler. Jobs were run on nodes equipped with Intel Xeon Platinum 8168 CPUs. For single-thread tests, we allocated one CPU core and 256 GB of memory; for multithread tests, we allocated eight CPU cores and 256 GB of memory. For short read data, we used a sample with 22 × coverage. For long read data, the BAM file of the same sample was down sampled to 22 × coverage using Samtools v1.17 [30] and was used as input for Straglr and LongTR. HipSTR, GangSTR, LongTR, ExpansionHunter, and Straglr were compared in single-thread mode. ExpansionHunter and Straglr were also evaluated using eight threads. HipSTR, GangSTR, and LongTR do not support multithreading in the versions used in this study (HipSTR v4.0.6, GangSTR v2.5 and ExpansionHunter v5.0.0). Each test was repeated three times, and runtime and peak memory usage were recorded using the time -v command.

Results

STR reference panel

From the reference genome, Tandem Repeat Finder detected 836,981 STRs, covering 21 Mb of the autosomal genome. Overall, the number of STRs per chromosome was positively correlated with chromosome length (Fig. 1A). Among these, dinucleotide STRs were the most abundant, followed by pentanucleotide STRs (Fig. 1B). The AT, GT, and AC motifs were the most common among dinucleotide STRs, while AACTG and AGTTC were the most common among pentanucleotide STRs (Fig. 1C). Of the total, 1,846 STRs were longer than 150 bp. Of the 836,981 reference STRs, 209 STRs located near the ends of chromosomes were excluded from the reference list of ExpansionHunter because they violated the –region-extension-length = 1000 parameter (The distance around on-target and off-target regions, used to search for informative reads for accurate genotyping, was set to 1,000 bp by default.).

Comparison of pSTRs

Among the six tools, HipSTR, GangSTR, ExpansionHunter, LongTR, and Straglr produced VCF files with assigned genotypes, while RepeatHMM provided a table with the copy number for each allele of each STR**.** Therefore, for RepeatHMM, we used the copy number at each allele as the genotype. We focused on STRs that were polymorphic within the five animals (two trios) to demonstrate how well each tool discriminated STR variation. For short-read sequence data, ExpansionHunter identified the largest number of pSTRs across the five individuals, with 327,680 STRs, while GangSTR identified the smallest number (110,680) (Table 1). HipSTR called an

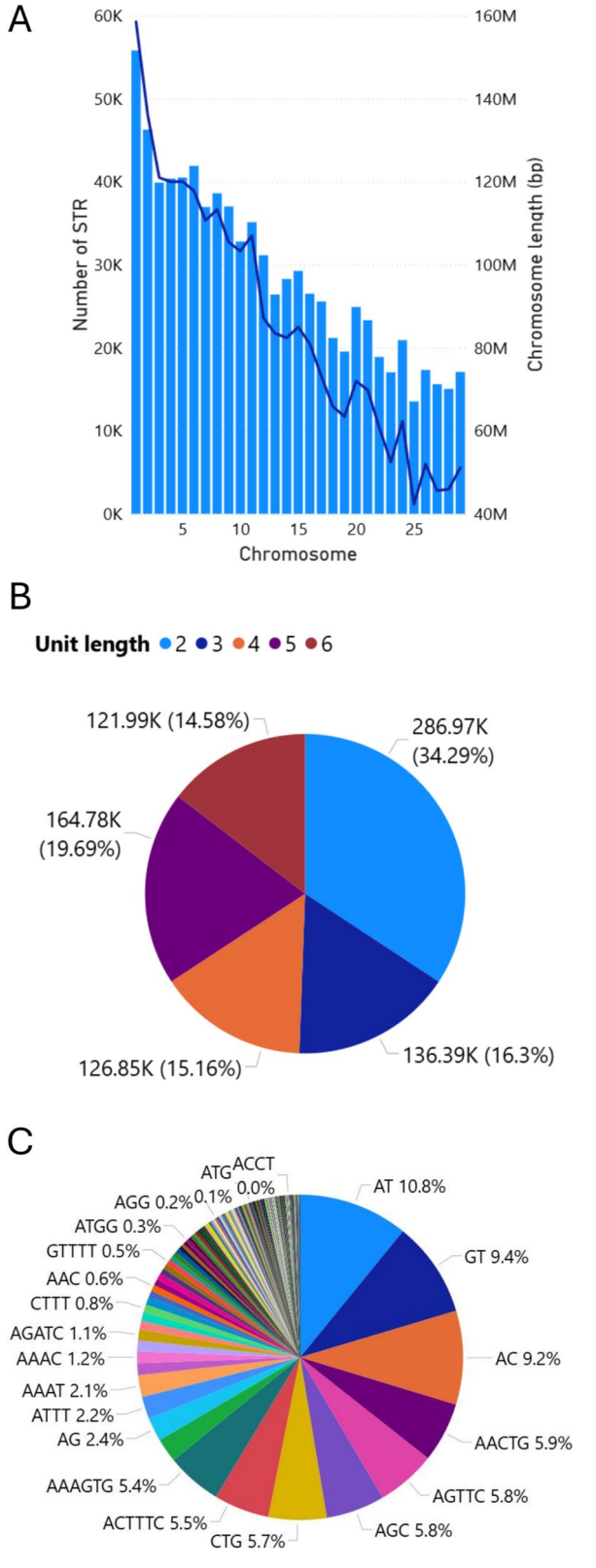

Fig. 1 Distribution of the number of STRs in the reference panel by chromosome (**A**), as well as the proportion of STRs by repeat unit length (2, 3, 4, 5, and 6 nucleotides) (**B**), and by repeat unit sequence (**C**)

intermediate number with 205,900 STRs. For long read sequence data, there was even greater variation in the number of pSTRs among the three tools. LongTR identified the highest number of pSTRs, with 470,250, and Straglr identified the smallest number, with 90,295 STRs (Table 1).

The level of overlap among pSTR discovered by each tool is visualised in Figure S1 (Additional file 1). Only 10,829 of all the polymorphic sites were detected by all six tools. GangSTR and HipSTR showed high agreement with the other tools, including those designed for long-read data, with the number of sites detected unique to GangSTR and HipSTR being only 242 and 2,373 respectively. A total of 40,834 identical pSTRs were detected by GangSTR, HipSTR, RepeatHMM, ExpansionHunter, and LongTR, representing the highest number of STRs detected by any combination of five tools. LongTR, ExpansionHunter, and RepeatHMM, detected more pSTRs than other tools, and also identified the highest number of unique sites. In contrast, Straglr detected the lowest number of pSTRs but also showed the least agreement with the other tools, with 11.6% being unique sites ($N = 10,474$).

Table 2 shows the average heterozygosity rate of STRs based on the total number of pSTRs detected by each tool or for the overlapping STRs identified by all six tools. In short-read sequence data, heterozygosity was relatively uniform and particularly so for the overlapping set of STR (ranging from 62 to 68%). In contrast, there was extreme variability among tools using long read data and this was even more extreme for heterozygosity calculated only among the overlapping STR. LongTR had the highest heterozygosity rate ($88.76 \pm 2.10\%$), while RepeatHMM and Straglr had low to extremely low rates, at $34.96 \pm 1.18\%$ and $2.12 \pm 0.48\%$, respectively.

Given the technological differences between short and long read data, we then focused on results for short-read and long-read tools separately. The number of overlapping sites detected as polymorphic, and the distribution of these sites based on the length of the repeat unit were studied. The detailed numbers for each group are provided in Supplementary Tables S2 and S3 (Additional file 2). For short-read tools only, 93,023 of 382,052 (24.3%) STRs were identified as polymorphic by all three tools: HipSTR, GangSTR, and ExpansionHunter (Fig. 2A). GangSTR, which called the fewest pSTRs, had the highest proportion of overlap; 98.5% (109,071) of its calls were also detected by HipSTR or ExpansionHunter. ExpansionHunter, detected the largest number of pSTRs, including almost 98% of pSTRs detected by GangSTR as well as 70% of pSTRs detected by HipSTR, while also discovering 51% unique pSTRs. HipSTR showed intermediate results in terms of coincidence. For each short-read tool, di-nucleotide repeats were the most common, and

Table 2 Number of pSTRs and the average heterozygosity rate (± standard deviation (SD)) in five individuals for six STR discovery tools as well as the heterozygosity rate (± SD) when considering only STR detected by all six tools

Sequencing platform	Tool	Number of pSTRs	Heterozygosity rate in all sites (mean % ± SD)	Heterozygosity rate in overlapping sites among the six tools (mean % ± SD)
Short-read (Illumina)	HipSTR	205900	53.36 ± 1.36	62.22 ± 2.64
	GangSTR	110680	63.22 ± 3.18	67.97 ± 3.73
	ExpansionHunter	327690	56.29 ± 5.10	67.82 ± 2.71
Long-read (ONT)	LongTR	470250	73.36 ± 1.71	88.76 ± 2.10
	RepeatHMM	224185	28.29 ± 0.95	34.96 ± 1.18
	Straglr	90295	3.88 ± 0.46	2.12 ± 0.48

GangSTR showed the highest proportion of dinucleotide repeats compared to HipSTR and ExpansionHunter (Fig. 2B, horizontal bars). HipSTR showed the highest rate of penta- and hexanucleotide repeats, while ExpansionHunter showed the highest rate of tri- and tetranucleotide repeats. Dinucleotide repeat STRs were also the dominant type among the polymorphic sites detected by all three tools, as well as those detected by only two tools (Fig. 2B, vertical bars). However, in the sets of unique STRs identified by each tool, pentanucleotide repeats were the most frequent in HipSTR, while tri- and tetranucleotide repeats were almost as frequent as dinucleotide repeats in ExpansionHunter (Fig. 2B, vertical bar chart).

Compared to the short-read tools, the three tools used for long-read sequencing data had lower overlap of polymorphic sites, with 33,253 sites called by all three tools out of 535,345 sites (only 6.2%, Fig. 3A). Dinucleotide repeats remained the most common pSTRs detected by each tool (Fig. 3B, horizontal bar chart), as well as the most common among sites detected by all three tools and by only two tools (Fig. 3B, vertical bar chart). When comparing the unique sites identified by each tool, RepeatHMM showed the highest proportion of trinucleotide repeats, while Straglr had the highest proportion of tetranucleotide repeats.

Mendelian consistency of STR genotypes

The accuracy of genotyping was evaluated by Mendelian consistency in the two trio offsprings. For each trio, Mendelian consistency was calculated as the ratio of genotype-consistent STR sites to the total number of segregating STR sites for a given tool, excluding any sites with missing genotypes in the trio. For short-read sequence data, the three tools showed high consistency rates, with GangSTR and HipSTR both at 0.98 (Fig. 4A). ExpansionHunter, which identified the highest number of pSTRs among the three short-read tools, had a lower consistency rate of 0.84. Furthermore, although ExpansionHunter was able to detect STR with alleles > 150 bp these showed an extremely low Mendelian consistency rate (0.12). For long read data, LongTR, which detected the largest number of polymorphic sites, achieved the highest consistency at 0.88 overall and 0.83 for STR with alleles > 150 bp. RepeatHMM and Straglr showed low consistency results, at 0.52 and 0.28 across all STR, respectively. We also examined the effect of repeat unit length on Mendelian consistency for all 6 tools (Fig. 4B). HipSTR, GangSTR, and LongTR showed no clear difference between different repeat unit lengths. However, ExpansionHunter and RepeatHMM showed higher Mendelian consistency rate for trinucleotide repeats and lower Mendelian consistency rate for di-, penta-, and hexanucleotide repeats.

Comparison of concordance in STR allele length

While all tools are designed to genotype STRs, each tool uses a different approach with some differences in outputs that are used to determine genotype. GangSTR, ExpansionHunter, and RepeatHMM estimate the copy number of the repeat unit for each allele. HipSTR and LongTR do not specify the copy number of the repeat unit but instead focus on the number of base pair differences of the alternative allele compared to the reference allele. Additionally, the starting positions of some STRs in the VCF output file were shifted by several nucleotides compared to the original starting positions provided in the BED file of predefined STRs, so these had to be adjusted to the original positions in the BED file to compare the results across tools. All tools, except for Straglr, preserved the repeat unit provided in the BED file. In contrast, Straglr determined the repeat unit sequence using its integrated Tandem Repeat Finder tool and reported the length and copy number of the alternative allele based on the newly identified motif sequence. Among the 90,295 pSTRs identified by Straglr, only 34% of the STR repeat units were the same as those provided in the original BED file, 61% had the same length but the motif starting nucleotide position was shifted by one or more bases (e.g., changed from AC to CA). Both groups were retained for the length-concordance analysis. The remaining 5%, whose repeat units differed from the BED file in both length and sequence, were excluded.

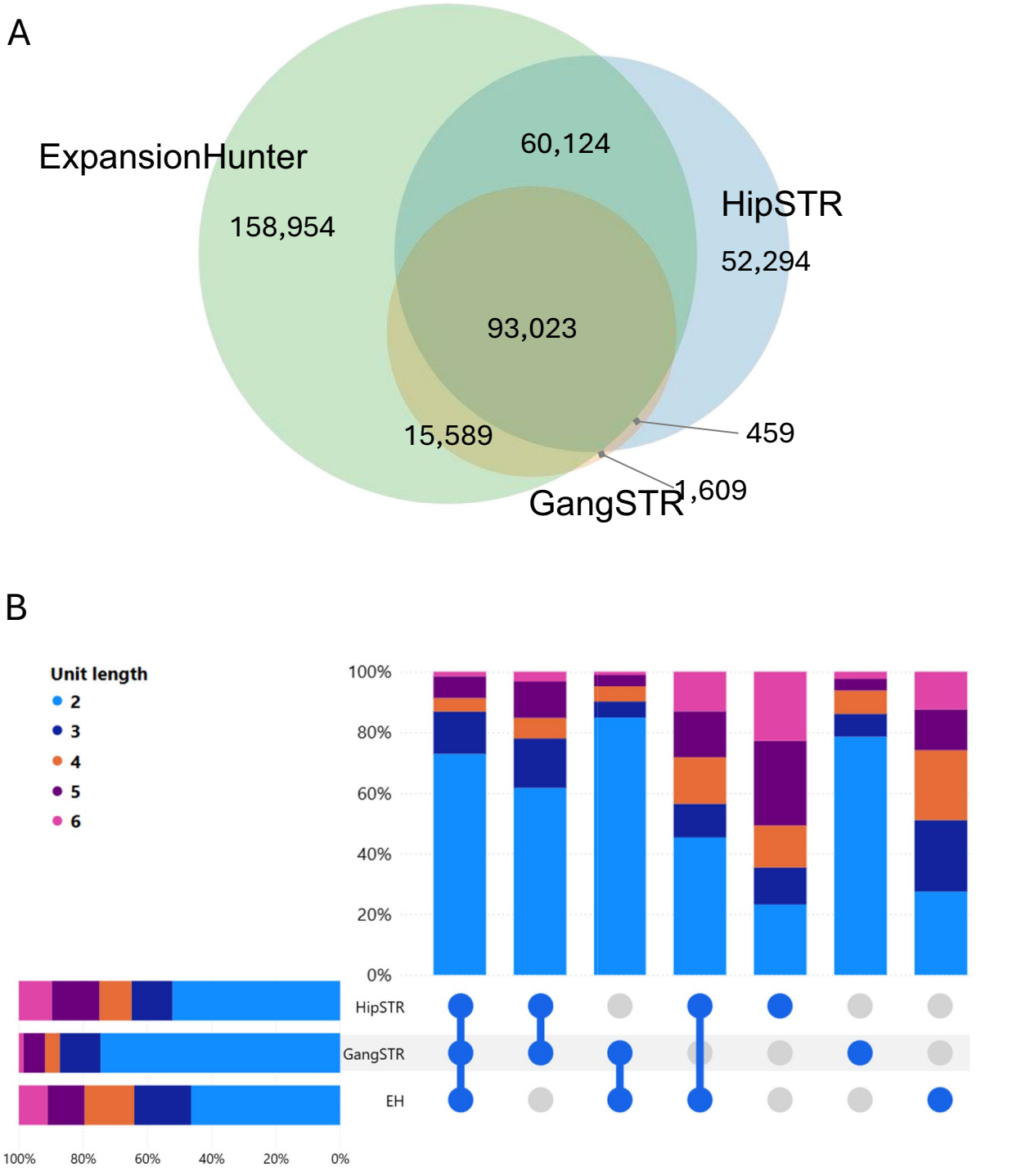

Fig. 2 **A** Overlap of polymorphic sites detected by three short-read tools: HipSTR, GangSTR, and ExpansionHunter [EH]. The size of each area in the Venn diagram corresponds to the number of STR sites, with the numbers indicating the counts of STR sites in each area. **B** Proportion of STR sites based on the length of the repeat unit (2, 3, 4, 5, or 6 nucleotides) detected by each short-read tool (horizontal bar chart), as well as for the intersection groups: sites detected by all three tools, by only two tools, and by only one tool (vertical bar chart)

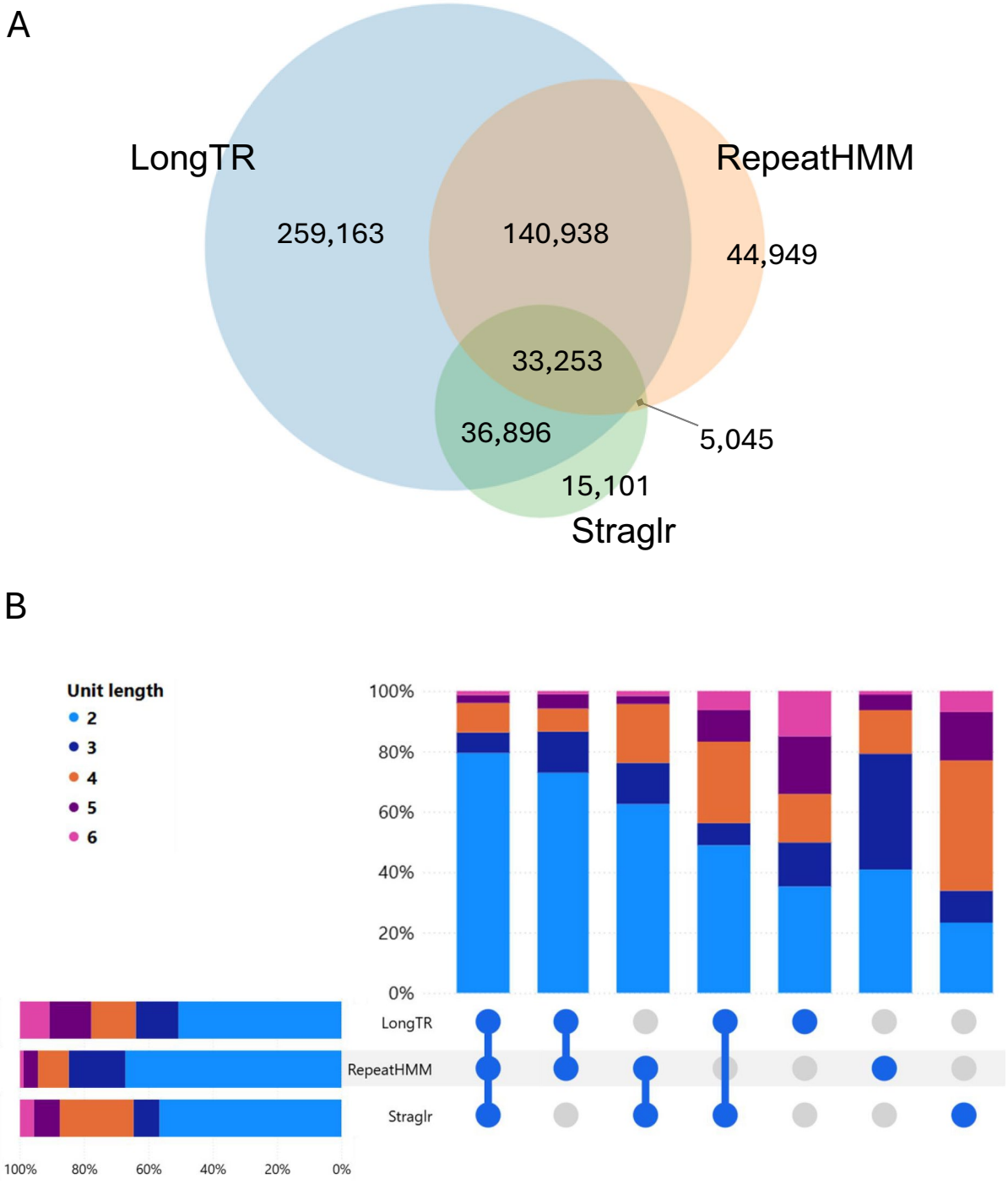

Fig. 3 **A** Overlap of polymorphic sites detected by three long-read tools: LongTR, RepeatHMM, Straglr. The size of each area in the Venn diagram corresponds to the number of STR sites, with the numbers indicating the counts of STR sites in each area. **B** Proportion of STR sites based on the length of the repeat unit (2, 3, 4, 5, or 6 nucleotides) detected by each long-read tool (horizontal bar chart), as well as for the intersection groups: sites detected by all three tools, by only two tools, and by only one tool (vertical bar chart)

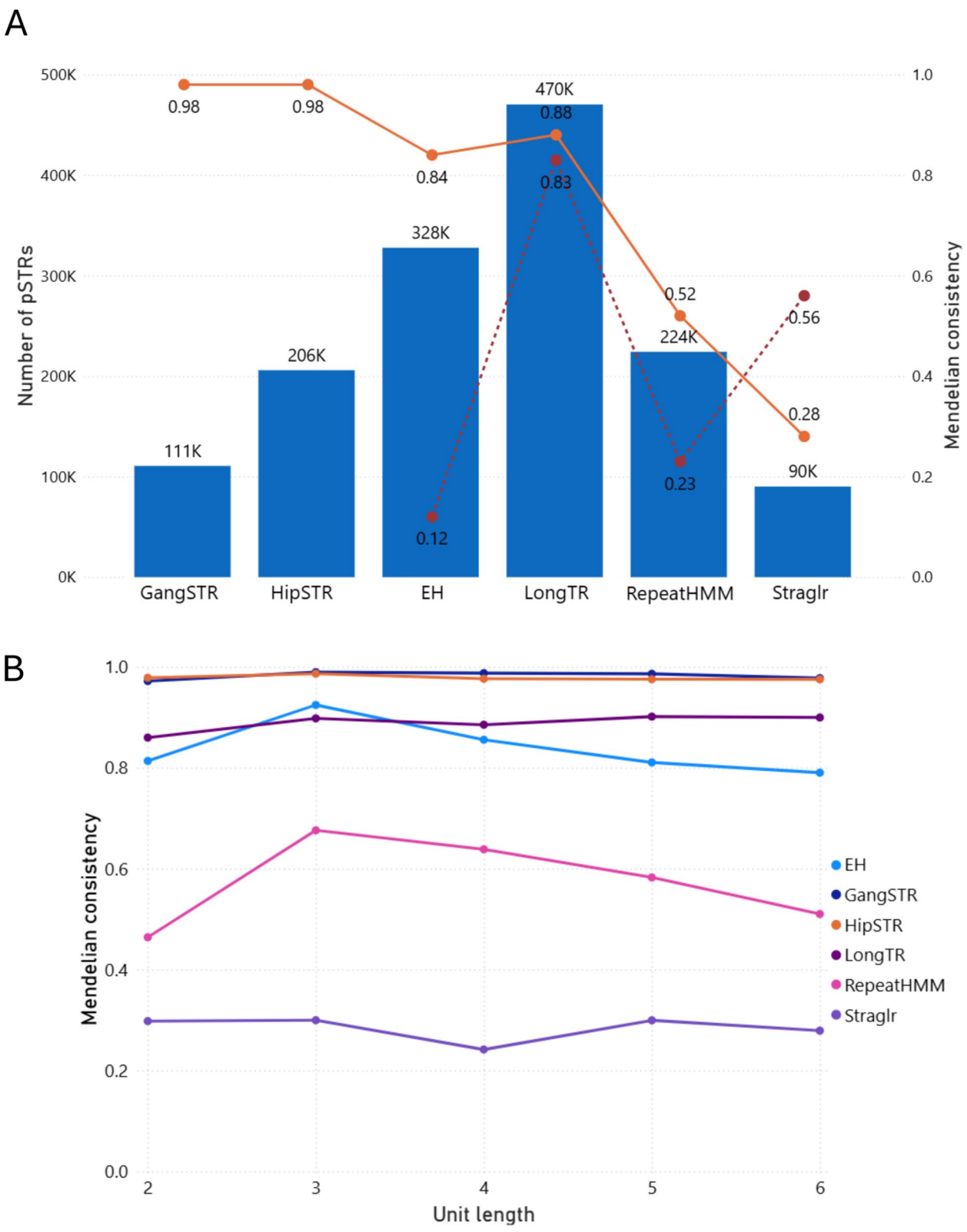

Fig. 4 **A** Comparison of Mendelian consistency across three short-read tools (ExpansionHunter [EH], GangSTR, and HipSTR) and three long-read tools (LongTR, RepeatHMM, and Straglr). The bar chart (left y-axis) shows the number of pSTRs detected by each tool, while the line chart (right y-axis) indicates the average Mendelian consistency across two trios for all pSTR sites. The dashed line represents the average Mendelian consistency for STRs with alleles longer than 150 bp. **B** Comparison of Mendelian consistency by repeat unit length. Each line represents one tool. The y-axis shows the average Mendelian consistency rate across two trios, and the x-axis shows the STR classes based on repeat unit length (2-, 3-, 4-, 5-, and 6-nucleotide repeats)

In addition to comparing genotyping results, the allele lengths at polymorphic sites were evaluated across the tools. Compared to the other five tools, Straglr reported a higher average length of polymorphic alleles (39 bp), while the average lengths for HipSTR, GangSTR, ExpansionHunter, LongTR, and RepeatHMM were similar at 27, 23, 28, 26, and 28 bp, respectively. All pSTRs identified by HipSTR and GangSTR were less than 150 base pairs (bp) in length, while ExpansionHunter, LongTR, RepeatHMM, and Straglr reported pSTR sites exceeding 150 bp in length (1,759, 2,383, 1,731, and 3,763 STR sites with maximum allele lengths over 150 bp, respectively) (Fig. 5A). According to the tool authors, HipSTR can only identify STRs shorter than the length of the read, which in this case is 150 bp in the short-read sequence. Although GangSTR and ExpansionHunter can identify STRs longer than the read length, only ExpansionHunter showed STRs longer than 150 bp among the polymorphic sites identified from short-read data in this study but as detailed above the Mendelian consistency rate for these STRs was very low (Fig. 4A). ExpansionHunter identified STRs with lengths up to 1,000 bp from short-read data, while LongTR, RepeatHMM, and Straglr identified STRs up to 10,000 bp from long-read data.

We compared the concordance of allele lengths across all six tools, examining each pairwise combination (Fig. 5B). The three tools using short-read sequence data showed concordance rates with each other over 60%, with the concordance rate between HipSTR and GangSTR reaching 71%. Tools using long read data showed lower concordance, with the highest rate between LongTR and RepeatHMM at 33%. When comparing across long read and short-read tools, LongTR and HipSTR showed the highest concordance rate of 75%, and in fact LongTR was developed from the same method as HipSTR, with both determining STR alleles based on nucleotide differences from the reference allele. The concordance rates of LongTR with GangSTR and ExpansionHunter were 50% and 40%, respectively. RepeatHMM had concordance rates with GangSTR, HipSTR, and ExpansionHunter of 46%, 40%, and 32%, respectively. Straglr showed the lowest concordance rate for allele length with all the remaining tools.

Comparison of computer power and running time

The running time and peak memory usage of five tools (HipSTR, GangSTR, ExpansionHunter, LongTR, and Straglr) were first compared in single-thread mode, with RepeatHMM excluded from these comparisons. For the remainder of the study, we split the BED file by chromosome to run RepeatHMM in parallel. Its extremely long runtime was a potential drawback when planning to use this tool genome-wide for multiple individuals. For short-read data, GangSTR and HipSTR had the shortest runtimes, each taking approximately 3 h for a 22 × coverage sample. GangSTR used the least memory, 0.027 GB, compared to 0.287 GB for HipSTR. ExpansionHunter required the most resources, using 66.5 GB of RAM and 9.8 h to complete. For long-read data, LongTR was the most efficient, requiring 0.72 GB of RAM and 6.5 h of runtime. Straglr required substantially more memory and time compared to LongTR (Fig. 6A,B).

Next, we compared multithreaded performance for ExpansionHunter and Straglr. ExpansionHunter's runtime decreased from 9.8 h in single-thread mode to 1.4 h when using eight threads. Similarly, Straglr's runtime and memory usage improved with multiple cores, demonstrating the benefit of multithreading for these tools.

Discussion

In this benchmark study, we compared STR genotyping tools at genome-wide scale in five cattle that formed two parent offspring trios. This is the first benchmark study of bovine genome-wide STR genotyping that includes both short-read and long-read sequence data. The six selected STR genotyping tools all use alignment files generated from BWA or Minimap2 as input, which are popular formats in population genomic research. The tools we selected for short-read sequence (HipSTR, GangSTR, and ExpansionHunter) have been quite widely used and there have been some comparison studies among them in humans [28, 29]. In contrast, tools for long read data, including those selected for this study (LongTR, Straglr, and RepeatHMM) are relatively new, and to our knowledge no whole-genome benchmark studies have been conducted to evaluate them, both among themselves and, in comparison to short-read tools. Short-read tools such as HipSTR, GangSTR, and ExpansionHunter report STR genotypes in VCF format, enabling relatively straightforward cross-software comparisons. Meanwhile, long-read STR genotyping tools have only recently been developed, and their outputs require more careful post-processing for benchmarking. Straglr produces VCF files only at the individual level with its own defined STR motifs, while RepeatHMM outputs a table of copy numbers for each individual rather than standard VCF files. Only LongTR, which was developed from the method of HipSTR, provides VCF files with multi-allelic genotypes determined at a population level.

In this study we focused on evaluating each tool's ability to identify multiple alleles in the offspring of two parent–offspring trios, to better understand what approaches might work well at population scale to accurately genotype STRs for imputation and GWAS in cattle. For short-read sequence data, among the three tools, HipSTR showed the best balance between pSTRs discovery rate, Mendelian consistency and computing power requirements for whole genome genotyping.

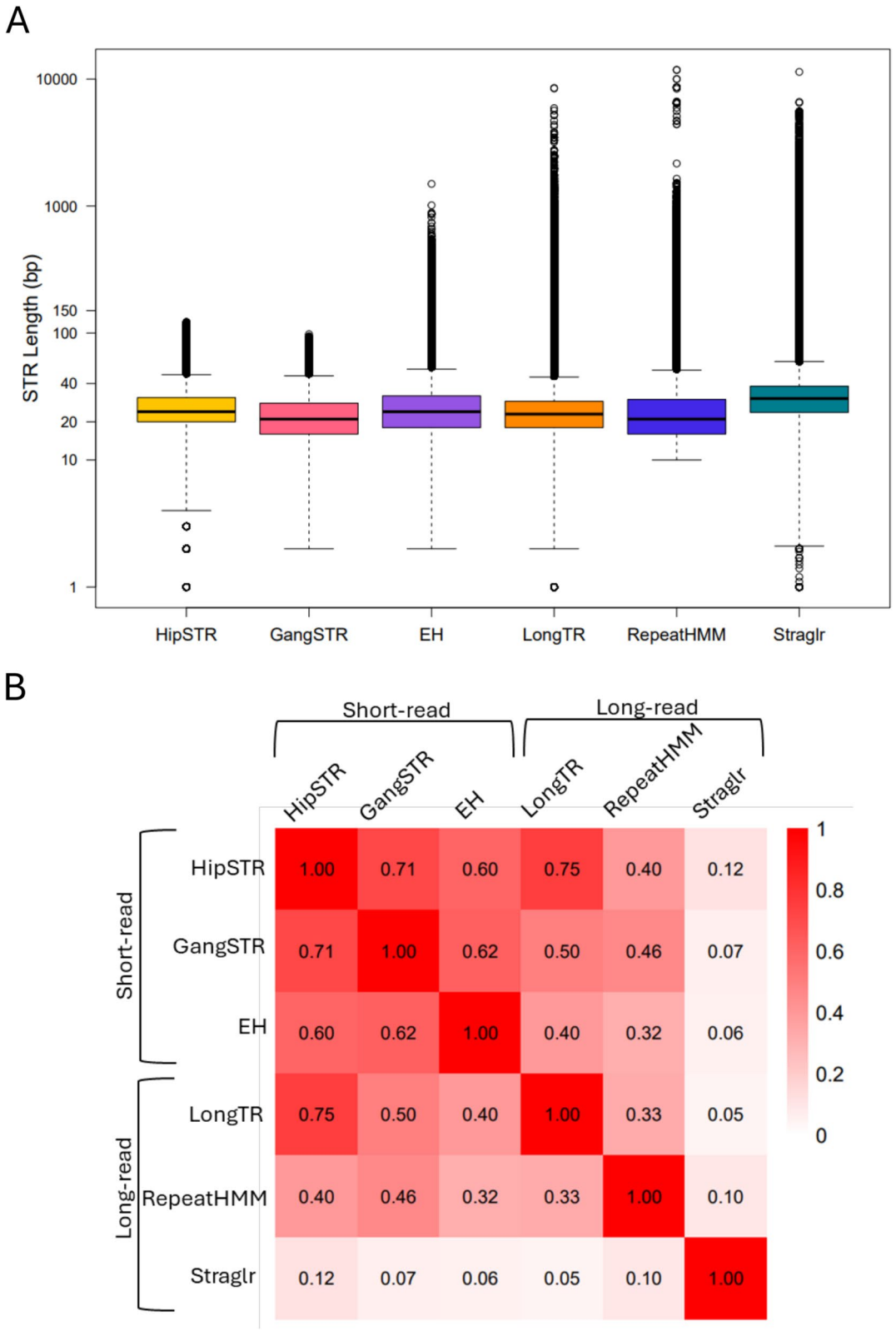

Fig. 5 **A** The distribution of allele lengths at pSTR sites detected by three short-read tools (HipSTR, GangSTR, and ExpansionHunter [EH]) and three long-read tools (LongTR, RepeatHMM, and Straglr). The y-axis represents the allele length of STRs on a log10 scale. **B** Concordance rate of allele length between each pair of tools for their overlapping set of STR. Numbers in the heatmap show the concordance rate for each pair. Darker cells represent higher concordance

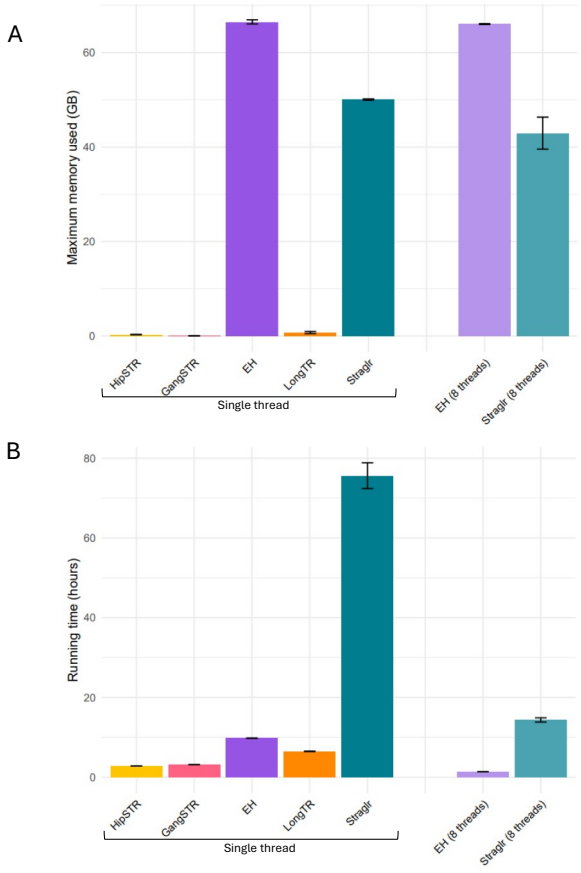

Fig. 6 Comparison of maximum memory usage in gigabytes (GB) (**A**) and running time in hours (**B**) for short-read tools (ExpansionHunter [EH], GangSTR, and HipSTR) and long-read tools (LongTR, Straglr), as well as ExpansionHunter (EH) and Straglr with multithreading (8 threads)

However, HipSTR was limited to detecting repeats that fall entirely within the read length. GangSTR also reported high Mendelian accuracy but appeared to have lower sensitivity to detect pSTRs compared to other tools, consistent with observations in human genomes where GangSTR showed reduced sensitivity for expansions at clinical loci compared to ExpansionHunter [28]. While ExpansionHunter identified the largest number of pSTRs in short-read data, it had lower Mendelian consistency than HipSTR and GangSTR and required more computational resources. This differs from the human benchmarking study where ExpansionHunter showed high accuracy [28], likely due to differences in the reference STRs used. That is, in our analysis, all three tools used the same whole-genome reference STR set, whereas the human benchmark study provided a smaller, well-curated reference set for ExpansionHunter, with a high proportion of trinucleotide repeats, which are patterns commonly reported in disease-related loci [28]. In fact, in our benchmark ExpansionHunter showed variation in genotyping accuracy based on repeat unit size, with trinucleotide repeats having the highest accuracy compared to other types.

With recent advances in long-read technology, datasets for cattle have become increasingly available, allowing the identification of complex genomic variation. While several recent studies have reported genome-wide structural variation in bovine long-read data, none have specifically focused on STR genotyping [50–52]. When comparing three STR calling tools with longread data, the concordance between them was lower than the concordance between tools in the short-read data, despite the long-read data in this study having higher coverage. This may be due to the fact that these tools are still under development, and the lack of a cross-compatible approach to combine and filter outputs from the different tools. Therefore, for long-read datasets, the choice of tool must be considered carefully, as it will significantly impact the results. Among the three tools benchmarked here, LongTR showed the best performance in terms of Mendelian consistency and concordance of pSTR position and length, it also required less computational power compared to RepeatHMM and Straglr. This aligns with human long-read profiling results, where LongTR showed higher genotyping accuracy and efficiency than Straglr for genome-wide TRs, including complex and large alleles [48]. RepeatHMM and Straglr showed poor Mendelian consistency among pSTRs and also tended to be biased toward homozygous genotypes. Because the Mendelian consistency test was conducted only for segregating genotypes within trios, many of the inconsistencies in Straglr and RepeatHMM occurred when all three animals were homozygous but with different numbers of repeats. RepeatHMM and Straglr are primarily copy number-focused and do not report exact STR allele sequences. A recent study that compared Straglr and RepeatHMM with a newly developed tool, HMMSTR, also found that Stragrl and RepeatHMM performed poorly for heterozygous calls and for alleles close in length (less than 100 bp difference) [53]. Another advantage of LongTR is that it was developed from the method of HipSTR [48], so its output format is similar to that of HipSTR, allowing both short-read and long-read STR datasets to be combined in the same study. Among the three long-read tools in this benchmark, LongTR was also the only tool that produced multi-sample output, and this combined with computational efficiency makes it more convenient for population-level STR genotyping. Long-read data, as expected, performed better than short-read data in detecting large STRs. Although the number of large STRs in the cattle genome is much lower than that of short STRs, these large STRs are of particular interest because many large expansion repeats have been reported as disease-associated in humans [20, 54–56].

This study demonstrates that relatively accurate genome-wide genotyping of STRs in both short and long-read sequencing technologies is challenging but possible. Our benchmark results indicate that STR genotyping from short-read data could be used to build a reasonable reference panel for imputation of STR < 150 bp in cattle. Recently, a SNP and STR reference haplotype panel was created in human short-read data of 479 families using HipSTR (and supplemented using a second STR caller, TREDPARSE [12]) and was then used for imputation and GWAS studies [9, 11]. With multiple new tools being developed for STR genotyping in long-read data, careful development of the best practices is required before applying them to large populations and incorporating the results into imputation panels. However, the opportunity to capture longer STR than was possible in short read is worth considering because many pathogenic STRs are found to be very long [20, 54–56] and would be better captured with long-read sequencing. The higher base-calling error rate of platforms such as ONT compared to short-read sequencing presents a challenge in use of long-read data in population studies of STR. This includes elevated error rates in homopolymer regions and mismatch errors in repeat regions, which can affect how STR tools determine the number of repeat units. Also, there is a lack of cross-comparison between long-read tools developed for different long-read sequencing platforms (ONT and PacBio), as some tools were developed only for one type of data—for example, TRGT [32] for PacBio HiFi reads only, while others such as STRkit [57] are not limited to a single platform, but were not yet fully peer reviewed at the time of writing this benchmark. Ideally, for population scale studies the output file format from different tools should be standardised for ease of merging data across different platforms. We expect these challenges will be addressed in the near future, as the base-calling accuracy of long-read technologies continues to improve and long-read data for population studies becomes more accessible. A further limitation of STR benchmarking in cattle is the absence of a validated truth set for accurately evaluating STR genotyping tools. In this study, we used Mendelian consistency at segregating sites as a proxy to estimate genotyping accuracy. In publications by the developers of STR genotyping tools, testing is often performed on a small number of well-characterized STR loci in humans, typically focusing on pathogenic variants. However, these sets are limited and not representative of genome-wide STR diversity. A genome-wide validated STR truth set will be necessary in the future for more accurate evaluation of STR genotyping tools.

This study was conducted in Holstein cattle, so the performance of STR-calling tools in for example *Bos indicus* breeds may not be captured. In addition, one limitation of the tools in this benchmark is their reliance on a list of known STRs. In this study, these STRs were identified using Tandem Repeat Finder based on the reference genome which is from the *Hereford* breed. Building a comprehensive and reliable list of STRs, including those not identifiable from the reference genome, is desirable to enhance population-scale STR genotyping for downstream analyses such as GWAS. Some tools, such as Replong [33], TRiCoLOR [34], ExpansionHunter Denovo [58], can detect novel STRs without relying on a reference. These tools were not included in this benchmark because we focused on reference-based tools, which are generally faster and scale better for population-wide analyses. Further work using these assembly-based tools and/or pan-genome discovery across multiple cattle breeds, is needed to generate a more complete catalogue of STRs that reflects the full extent of STR genetic diversity.

Conclusions

The comparison of six different tools used for genome-wide STR genotyping across two datasets (short-read and long-read data) revealed that short-read tools had higher genotyping accuracy and greater concordance in STR genotyping compared to long-read tools. However, tools used with long-read data performed better at detecting STRs longer than 150 bp. Our results concur with previous benchmark studies that use of a single STR genotyping tool may not be sufficient, and that combining results from multiple tools should be considered. This is especially true for studies requiring high precision on a small list of STRs, such as detecting known pathogenic STRs in clinical testing. For population studies with large cattle datasets, the performance metrics observed in this study favour HipSTR for short-read data and LongTR for long-read data. However, tool performance may vary with sequencing coverage, read quality and validation across additional breeds, larger cohorts, and with orthogonal assays (e.g. PCR or targeted long-read sequencing) is recommended.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12753-4.

Additional file 1: Supplementary Figure S1 – Comparison of polymorphic STR site overlap across six tools. The horizontal bar chart (set size) shows the total number of polymorphic STR sites detected by each tool. The vertical bar chart (intersection size) shows the number of STR sites in each intersection group, including sites detected by all six tools, combinations of five, four, or two tools, and sites uniquely detected by a single tool

Additional file 2: Table S1: Number of loci retained or removed at each TRF filtering step. Table S2: Number of loci in each intersection set of short-read tools, grouped by repeat length. Table S3: Number of loci in each intersection set of long-read tools, grouped by repeat length. Table S4: Average missing genotype rate per tool

Additional file 3: Commands used to run STR tools. A complete list of commands used for all STR genotyping tools included in this study

Acknowledgements
The authors acknowledge funding from the DairyBio project: a joint venture between Agriculture Victoria (Melbourne, Australia), Dairy Australia (Melbourne, Australia) and the Gardiner Foundation (Melbourne, Australia). The research underpinning this publication was undertaken while completing a PhD at La Trobe University, Melbourne, Victoria.

Authors' contributions
Q.H.T. conceived the study, performed the benchmarking analyses and wrote the manuscript. J.W. carried out DNA extraction and long-read sequencing. T.V.N. performed long-read base calling, quality control of raw reads and generated the BAM files. A.J.C. conducted short-read sequencing and generated the BAM files for the short-read data. A.J.C. and I.M.M. conceived and supervised the research, reviewed and edited the manuscript, and contributed equally as senior authors. All authors read and approved the final manuscript.

Funding
DairyBio, a joint venture project between Agriculture Victoria (Melbourne, Australia), Dairy Australia (Melbourne, Australia) and the Gardiner Foundation (Melbourne, Australia). DairyBio also supported the first author's PhD scholarship.

Data availability
The short-read sequencing data are available under BioProject PRJNA431934 with BioSample accessions SAMN19491793, SAMN19491794, SAMN19491795, SAMN19491796 and SAMN19491810. The long-read sequencing data will be made publicly available prior to publication.

Declarations

Ethics approval and consent to participate
DNA samples from the dam and offspring were collected in accordance with the Australian Code of Practice for the Care and Use of Animals for Scientific Purposes, and approved by the Agricultural Research and Extension Animal Ethics Committee of the Department of Environment and Primary Industries (application number 2014–23, approved on 18th December 2014). Semen samples from the sire were collected as part of routine business operations by a commercial artificial breeding company and donated to Agriculture Victoria for research purposes.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Agriculture Victoria, AgriBio, Centre for AgriBioscience, Bundoora, VIC 3083, Australia
[2]School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia

Received: 30 September 2025 / Accepted: 11 March 2026
Published online: 26 March 2026

References

1. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004;5:435–45.
2. Cortes O, Cañon J, Gama LT. Applications of microsatellites and single nucleotide polymorphisms for the genetic characterization of cattle and small ruminants: an overview. Ruminants. 2022;2:456–70.
3. Depienne C, Mandel JL. 30 years of repeat expansion disorders: what have we learned and what are the remaining challenges? Am J Hum Genet. 2021;108:764–85.
4. Gall-Duncan T, Sato N, Yuen RKC, Pearson CE. Advancing genomic technologies and clinical awareness accelerates discovery of disease-associated tandem repeat sequences. Genome Res. 2022;32:1–27.
5. Willems T, Gymrek M, Highnam G, 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. Genome Res. 2014;24:1894–904.
6. Jakubosky D, D'Antonio M, Bonder MJ, Smail C, Donovan MKR, Young Greenwald WW, et al. Properties of structural variants and short tandem repeats associated with gene expression and complex traits. Nat Commun. 2020;11:2927.
7. Bhati M, Mapel XM, Lloret-Villas A, Pausch H. Structural variants and short tandem repeats impact gene expression and splicing in bovine testis tissue. Genetics. 2023;225:iyad161.
8. Blaj I, Tetens J, Bennewitz J, Thaller G, Falker-Gieske C. Structural variants and tandem repeats in the founder individuals of four F2 pig crosses and implications to F2 GWAS results. BMC Genomics. 2022;23:631.
9. Bustos BI, Billingsley K, Blauwendraat C, Gibbs JR, Gan-Or Z, Krainc D, et al. Genome-wide contribution of common short-tandem repeats to Parkinson's disease genetic risk. Brain. 2022;146:65–74.
10. Margoliash J, Fuchs S, Li Y, Zhang X, Massarat A, Goren A, et al. Polymorphic short tandem repeats make widespread contributions to blood and serum traits. Cell Genom. 2023;3:100458.
11. Ohlei O, Paul K, Nielsen SS, Gmelin D, Dobricic V, Altmann V, et al. Genome-wide meta-analysis of short-tandem repeats for Parkinson's disease risk using genotype imputation. Brain Commun. 2024;6:fcae146.
12. Saini S, Mitra I, Mousavi N, Fotsing SF, Gymrek M. A reference haplotype panel for genome-wide imputation of short tandem repeats. Nat Commun. 2018;9:4397.
13. Gymrek M, Golan D, Rosset S, Erlich Y. lobSTR: a short tandem repeat profiler for personal genomes. Genome Res. 2012;22:1154–62.
14. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. Nat Methods. 2017;14:590–2.
15. Mousavi N, Shleizer-Burko S, Yanicky R, Gymrek M. Profiling the genome-wide landscape of tandem repeat expansions. Nucleic Acids Res. 2019;47:e90.
16. Dolzhenko E, Deshpande V, Schlesinger F, Krusche P, Petrovski R, Chen S, et al. Expansionhunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. Bioinformatics. 2019;35:4754–6.
17. Chu C, Nielsen R, Wu Y. REPdenovo: inferring de novo repeat motifs from short sequence reads. PLoS ONE. 2016;11:e0150719.
18. Liao X, Gao X, Zhang X, Wu F-X, Wang J. RepAHR: an improved approach for de novo repeat identification by assembly of the high-frequency reads. BMC Bioinformatics. 2020;21:463.
19. Chintalaphani SR, Pineda SS, Deveson IW, Kumar KR. An update on the neurological short tandem repeat expansion disorders and the emergence of long-read sequencing diagnostics. Acta Neuropathol Commun. 2021;9:98.
20. Willemsen R, Levenga J, Oostra BA. CGG repeat in the FMR1 gene: size matters. Clin Genet. 2011;80:214–25.
21. Tanudisastro HA, Deveson IW, Dashnow H, MacArthur DG. Sequencing and characterizing short tandem repeats in the human genome. Nat Rev Genet. 2024;25:460–75.
22. Fang L, Monteys AM, Dürr A, Keiser M, Cheng C, Harapanahalli A, et al. Haplotyping SNPs for allele-specific gene editing of the expanded huntingtin allele using long-read sequencing. Hum Genet Genomics Adv. 2023;4:100146.
23. Giesselmann P, Brändl B, Raimondeau E, Bowen R, Rohrandt C, Tandon R, et al. Analysis of short tandem repeat expansions and their methylation state with nanopore sequencing. Nat Biotechnol. 2019;37:1478–81.
24. Lang J, Xu Z, Wang Y, Sun J, Yang Z. NanoSTR: a method for detection of target short tandem repeats based on nanopore sequencing data. Front Mol Biosci. 2023;10:1093519.
25. Xu L, Haasl RJ, Sun J, Zhou Y, Bickhart DM, Li J, et al. Systematic profiling of short tandem repeats in the cattle genome. Genome Biol Evol. 2016;9:20–31.
26. Wang P, Sheng X, Xia X, Wang F, Li R, Ahmed Z, et al. The genomic landscape of short tandem repeats in cattle. Anim Genet. 2025;56:e13498.
27. Halman A, Oshlack A. Accuracy of short tandem repeats genotyping tools in whole exome sequencing data. F1000Res. 2020;9:200.
28. Oketch JW, Wain LV, Hollox EJ. 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. PLoS ONE. 2024;19:e0300545.
29. Rajan-Babu I-S, Peng JJ, Chiu R, Birch P, Couse M, Guimond C, et al. Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions. Genome Med. 2021;13:126.
30. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Bioinformatics. 2009;14:1754–60.

31. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
32. Dolzhenko E, English A, Dashnow H, De Sena Brandine G, Mokveld T, Rowell WJ, et al. Characterization and visualization of tandem repeats at genome scale. Nat Biotechnol. 2024;42:1606–14.
33. Guo R, Li Y-R, He S, Ou-Yang L, Sun Y, Zhu Z. RepLong: de novo repeat identification using long read sequencing data. Bioinformatics. 2018;34:1099–107.
34. Bolognini D, Magi A, Benes V, Korbel JO, Rausch T. TRiCoLOR: tandem repeat profiling using whole-genome long-read sequencing data. Gigascience. 2020;9:giaa101.
35. Mitsuhashi S, Frith MC, Mizuguchi T, Miyatake S, Toyota T, Adachi H, et al. Tandem-genotypes: robust detection of tandem repeat expansions from long DNA reads. Genome Biol. 2019;20:58.
36. Frith MC, Khan S. A survey of localized sequence rearrangements in human DNA. Nucleic Acids Res. 2018;46:1661–73.
37. Australian code for the care and use of animals for scientific purposes | NHMRC. https://www.nhmrc.gov.au/about-us/publications/australian-code-care-and-use-animals-scientific-purposes. Accessed 9 Sep 2025.
38. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
39. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience. 2020;9:giaa021.
40. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10:giab008.
41. Guppy protocol. Oxford Nanopore Technologies. 2018. https://nanoporetech.com/document/Guppy-protocol. Accessed 8 Jul 2025.
42. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.
43. Fungtammasan A, Ananda G, Hile SE, Su MS-W, Sun C, Harris R, et al. Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res. 2015;25:736–49.
44. Delahaye C, Nicolas J. Sequencing DNA with nanopores: troubles and biases. PLoS ONE. 2021;16:e0257521.
45. Mousavi N, Margoliash J, Pusarla N, Saini S, Yanicky R, Gymrek M. TRTools: a toolkit for genome-wide analysis of tandem repeats. Bioinformatics. 2021;37:731.
46. Liu Q, Zhang P, Wang D, Gu W, Wang K. Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing. Genome Med. 2017;9:65.
47. Chiu R, Rajan-Babu I-S, Friedman JM, Birol I. Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences. Genome Biol. 2021;22:224.
48. Ziaei Jam H, Zook JM, Javadzadeh S, Park J, Sehgal A, Gymrek M. LongTR: genome-wide profiling of genetic variation at tandem repeats from long reads. Genome Biol. 2024;25:176.
49. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
50. Nguyen TV, Vander Jagt CJ, Wang J, Daetwyler HD, Xiang R, Goddard ME, et al. In it for the long run: perspectives on exploiting long-read sequencing in livestock for population scale studies of structural variants. Genet Sel Evol. 2023;55:9.
51. Gao Y, Yang L, Kuhn K, Li W, Zanton G, Bowman M, et al. Long read and preliminary pangenome analyses reveal breed-specific structural variations and novel sequences in Holstein and Jersey cattle. J Adv Res. 2026;79:137–50.
52. Leonard AS, Crysnanto D, Fang Z-H, Heaton MP, Vander Ley BL, Herrera C, et al. Structural variant-based pangenome construction has low sensitivity to variability of haplotype-resolved bovine assemblies. Nat Commun. 2022;13:3012.
53. Van Deynze K, Mumm C, Maltby CJ, Switzenberg JA, Todd PK, Boyle AP. Enhanced detection and genotyping of disease-associated tandem repeats using HMMSTR and targeted long-read sequencing. Nucleic Acids Res. 2025;53:gkae1202.
54. Fu Y-H, Kuhl DPA, Pizzuti A, Pieretti M, Sutcliffe JS, Richards S, et al. Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell. 1991;67:1047–58.
55. MacDonald ME, Ambrose CM, Duyao MP, Myers RH, Lin C, Srinidhi L, et al. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. Cell. 1993;72:971–83.
56. Verkerk AJ, Pieretti M, Sutcliffe JS, Fu YH, Kuhl DP, Pizzuti A, et al. Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell. 1991;65:905–14.
57. Lougheed DR, Pastinen T, Bourque G. STRkit: precise, read-level genotyping of short tandem repeats using long reads and single-nucleotide variation. 2025. https://doi.org/10.1101/2025.03.25.645269.
58. Dolzhenko E, Bennett MF, Richmond PA, Trost B, Chen S, van Vugt JJFA, et al. Expansionhunter denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. Genome Biol. 2020;21:102.

Publisher's Note