Proc. Assoc. Advmt. Anim. Breed. Genet. 26: 375-378

CHARACTERISATION OF SHORT TANDEM REPEATS (STRs) IN CATTLE USING LONG READ SEQUENCING

Q.H. Tran[1,2], I.M. MacLeod[1,2], T.V. Nguyen[1], J. Wang[1]and A.J. Chamberlain[1,2]

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

SUMMARY

Short tandem repeats (STRs), also known as microsatellites, are highly variable DNA mutations that were widely used as genetic markers in cattle populations. However, previous studies focused only a small set of known STRs, and our understanding of STRs at the whole-genome level and their potential functional impact remains limited. This study is the first to examine STRs across the entire genome in Holstein and Jersey breeds using long-read sequencing. We investigated the number of alleles in multiallelic STRs, their allele lengths, their annotation, and the diversity of STRs between populations. Our study characterised over 352,000 STRs and provided evidence suggesting that there may be strong negative selection pressure acting on STRs that cause frameshift mutations in protein coding regions. This study provides a database of STRs at the whole-genome level that serves as a valuable resource for both functional studies and population structure analysis in cattle.

INTRODUCTION

Short tandem repeats (STRs) also known as microsatellites, are DNA regions characterized by the repetition of 1 to 6 base pair (bp) motifs. STRs are highly variable with multiple allele lengths (i.e. repeat motifs). In humans, STRs have been identified as causal variants for both complex and monogenic traits (Tanudisastro *et al*. 2024). However, unlike SNPs, which have been well characterized at sequence level across many populations, there are limited studies on STRs in cattle at the whole-genome scale. This is likely due to the lack of standardised bioinformatic tools for genotyping STRs, as each tool has its own strengths and limitations. Additionally, available short-read sequence data can be unreliable for detecting STRs longer than the read lengths, because sequence reads cannot cover the entire repeat region (Tanudisastro *et al*. 2024). This study is the first to explore whole-genome STRs in a cattle population using long-read sequencing data. LongTR, a recently developed tool, was chosen for STR genotyping because it demonstrated high performance among STR genotyping tools for long-read data (Ziaei Jam *et al.* 2024). Polymorphic STRs across the entire genome were identified in 108 Holstein and Jersey cattle. This work provides an overall insight of STR diversity and their potential functional impact on the cattle genome.

MATERIALS AND METHODS

Fifty Holstein and 58 Jersey cattle were sequenced using an Oxford Nanopore Technology (ONT) PromethION sequencer with R9.4.1 or R10.4.1 flow cells, achieving an average read coverage of 25X. The FAST5 files were basecalled using either Guppy v6.1.7 or Dorado v0.7.0 in super high accuracy mode. The output sequence reads were filtered using FiltLong (https://github.com/rrwick/Filtlong) and then aligned to the ARS-UCD 2.0 reference genome (Rosen *et al.* 2020) using Minimap2 (Li 2018). Prior to STR genotyping, a reference list of STRs was created from the reference genome using Tandem Repeat Finder (Benson 1999). Only autosomal STR loci with repeat motifs ranging from 2 to 6 bp were selected for the predefined list. This list, along with the 108 aligned sequence files, were input to LongTR with the parameter settings --min-reads 4 (minimum total reads required to genotype a locus) and --max-tr-len 10000 (maximum STR length of the reference allele). The VCF output from LongTR was then further filtered, removing STR loci with a total number of informative reads (DP in INFO fields) that were too low (< 1500) or too high

(> 3000), where the thresholds were chosen based on the observed genome-wide distributions. Additionally, individual genotypes with a quality score (Q) < 0.9, number of valid reads per animal (DP in Format fields) < 8, or an alternate pattern marked as <DEL> (deletion) were set to missing. Finally, only polymorphic loci with a missing genotype rate of less than 20% were retained. To assess the genotyping accuracy, Mendelian consistency was calculated for 2 parent–offspring trios in our population. The filtered STRs were annotated using Variant Effect Predictor (VEP) (McLaren *et al*. 2016) version 113 with the parameters --symbol and --biotype. To evaluate the population structure based on genome-wide STRs, the average pairwise genetic differentiation (weighted Fst) between the two breeds, Holsteins and Jerseys, was calculated for all STR within a window size of 2,000,000 bp and a sliding window of 10,000 bp using VCFtools (Danecek *et al*. 2011).

RESULTS AND DISCUSSION

STR genotyping. A reference list of 836,981 autosomal STR loci from the reference genome was created. After genotyping these STR in 108 animals using LongTR, 352,554 polymorphic STRs were retained after quality filtering. The average Mendelian consistency of the STR genotypes in two parent-offspring trios, was 74% for the raw unfiltered STR and increased to 80% after filtering. The number of STRs per chromosome generally corresponded with the chromosome length, except for chromosomes 19 and 25, which had fewer than expected STRs (Figure 1a). Di- and pentanucleotide repeats were the most common STRs, representing 39.2% and 21.4%, respectively (Figure 1b). Most STRs (84.4%) had between 5 to 15 alleles, with dinucleotide repeats showing the highest variability (Figure 1c). This observation is consistent with other studies in humans and cattle using short-read data and different bioinformatic tools, which may suggest that dinucleotide repeats have a higher mutation rate (Bhati *et al*. 2023; Shi *et al*. 2023).

We calculated the length of the major (i.e. most frequent) allele at each STR locus. The length of the major allele typically ranged from 20 to 40 bp, while dinucleotide STRs showed more variability in major allele length, which aligned with the higher number of alleles (Figure 1d). Overall, only 0.5% of major alleles were longer than 150 bp (the common length of reads in short-read sequencing), while the maximum lengths of major STRs for di-, tri-, tetra-, penta-, and hexanucleotide repeats were 2,175, 3,227, 1,994, 2,682, and 1,937 bp, respectively. Furthermore, 19,115 STR loci had at least one allele longer than 150 bp, accounting for 5.4% of all STR loci and these are likely to be better characterised using long read versus short read sequence. Dinucleotide STRs had the highest proportion of loci with at least one allele longer than 150 bp (10% of total dinucleotide STR loci). For other repeat motifs, the proportion ranged from 2.1% to 3.2%. The maximum allele lengths recorded for di-, tri-, tetra-, penta-, and hexanucleotide repeats were 16,753, 7,569, 8,788, 11,311, and 6,733 bp, respectively. The long repeat expansions with rare frequency are of interest, because they could be potential pathogenic variants, and they are difficult to detect in studies using short-read data.

STR annotation. The VEP annotation found that most STRs were in intergenic (48.9%) and intronic regions (42.8%). Only 0.3% of STRs (1,147 loci) were in coding regions. For non-coding regions, the proportion of STRs by repeat motif corresponded to the overall distribution of STRs in the genome (Figure 1e). However, in coding regions, trinucleotide STRs dominated, comprising 66% of all coding-region STRs, followed by hexanucleotide STRs at 17% (Figure 1e). Conversely, there were a considerably lower proportions of di-, tetra-, and pentanucleotide STRs in coding regions, suggesting that these may be under strong negative selection because they cause frameshifts in the DNA codons with a much higher risk of detrimental impact on protein function compared to tri- or hexanucleotide STRs.

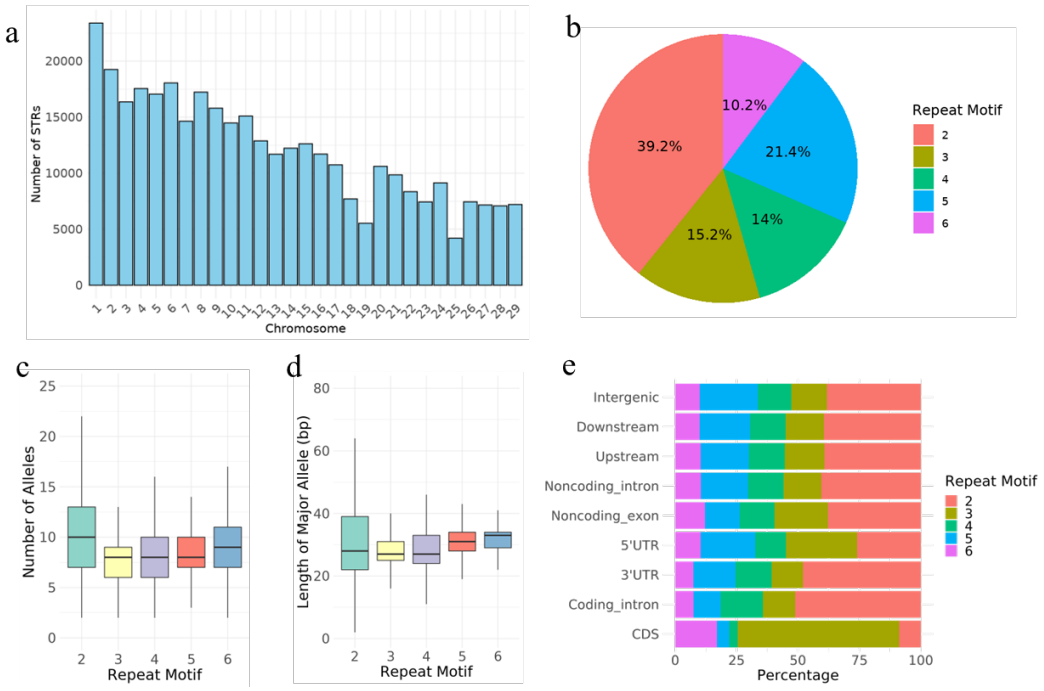

Figure 1. Polymorphic STR characteristics. a. Number of STR by chromosomes. b. Proportion of STR by repeat motif size. c Distribution of number of alleles per STR by repeat motif size. d Distribution of length of major allele by repeat motif size. e Percentage of STRs of repeat motif size within regions of different function

Population structure. In the past STRs have traditionally been used to investigate population differences in cattle. However, previous studies were limited to a small set of known STRs. In this study, we examined the use of genome-wide STRs to assess population structure. We calculated pairwise estimates of genomic differentiation (Fst) between the Holstein and Jersey populations using STRs across chromosome regions. We identified two regions with the most highly differentiated STRs (average Fst > 0.15) (Figure 2). The first region was from 40 to 43 Mb on chromosome 7, and the second was from 26 to 28.7 Mb on chromosome 19. For example, in Jerseys, the major allele (82% of all alleles) for the STR at position 41,775,200 on chromosome 7 had 7 repeats of the AGC motif, while the major allele (68% of alleles) for Holsteins at this STR had 4 repeats of AGC. Interestingly, the region from 40 to 43 Mb on chromosome 7 was reported as part of a run of homozygosity (ROH) island across United States, New Zealand and Australian Jersey populations based on SNP data indicating potential selection pressure across this region (Howard *et al.* 2015). While regions with high Fst do not necessarily implicate STR as causal variants, it is of interest to further investigate STR functional impact and associations with key dairy traits.

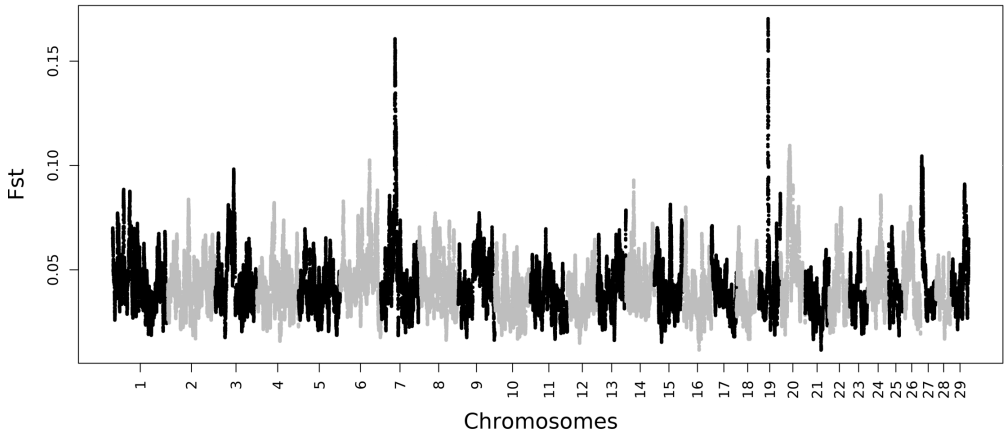

Figure 2. Genome-wide Manhattan plot of average Fst between Holstein and Jersey populations

CONCLUSIONS

This study is the first to examine the diversity of STRs in a cattle population at the whole-genome scale using long-read sequence data. While the study found that challenges still exist in accurately genotyping STRs from whole-genome sequence data, we demonstrated that large numbers of STRs were accurately genotyped and many were highly polymorphic. The considerably higher number of tri- and hexanucleotide STRs and lower number of di-, tetra-, and pentanucleotide STRs in coding regions reflected the potential functional impact of STRs and suggests selection pressure against frameshift causing mutations. The high variability of STRs also made them a powerful approach for identifying genomic regions that may be under different selective forces in the two different breeds.

ACKNOWLEDGEMENTS

The authors acknowledge financial support from the DairyBio program, which is jointly funded by Agriculture Victoria, Dairy Australia and The Gardiner Foundation.

REFERENCES

Benson G. (1999) *Nucleic Acids Res*. **27**: 573.

Bhati M., Mapel X.M., Lloret-Villas A. and Pausch H. (2023) *Genetics* **225**: iyad161.

Danecek P., Auton A., Abecasis G., *et al*. (2011) *Bioinformatics* **27**: 2156.

Howard J.T., Maltecca C., Haile-Mariam M., Hayes B.J. and Pryce J.E. (2015) *BMC Genomics* **16**: 187.

Li H. (2018) *Bioinformatics* **34**: 3094.

McLaren W., Gil L., Hunt S.E., Riat H.S., Ritchie G.R.S., Thormann A., Flicek P. and Cunningham F. (2016) *Genome Biol*. **17**: 122.

Rosen, B.D., Bickhart, D.M., Schnabel, R.D., *et al*. (2020) *GigaScience* **9**: 1.

Shi Y., Niu Y., Zhang P., *et al*. (2023) *Nat. Commun*. **14**: 2092.

Tanudisastro H.A., Deveson I.W., Dashnow H. and MacArthur D.G. (2024) *Nat. Rev. Genet*. **25**: 460.