Mapholi, N., Tshilate, T., Smith, R., Molotsi, A., Hlongwane, N., Nesengani, L. (2026) High-quality genome assemblies of three Nguni cattle ecotypes using PacBio HiFi and Omni-C sequencing, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 23909: https://doi.org/10.31274/wcgalp.23909 ©2026 Mapholi, et al. CC BY-NC-ND

High-quality genome assemblies of three Nguni cattle ecotypes using PacBio HiFi and Omni-C sequencing

N. Mapholi*, T. Tshilate, R. Smith, A. Molotsi, N. Hlongwane, and L.T. Nesengani

*University of South Africa, Johannesburg, South Africa; *Corresponding author: maphon@unisa.ac.za*

Introduction

Nguni cattle are indigenous to Southern Africa and are well known for their adaptability, fertility, disease tolerance and do well under low-input management system (Bester *et al*., 2003). They are a breed of choice among smallholder farmers (Tada *et al*., 2013). Nguni cattle were introduced to the Southern African region by migrating Bantu-speaking people (Faku & Hebinck, 2013). Over centuries, the breed adapted to the different agro-ecological regions representing a diverse adaptative genetic variation which resulted in distinct ecotypes (Sanarana *et al*., 2016). Despite their importance for food security, Nguni cattle remain underrepresented in high-resolution genomic resources such as high-quality reference genomes (Bruford *et al*., 2015). High quality reference genomes are essential for advancing genomic selection for sustainable improvement of livestock and conservation purposes. The current cattle reference genome may not fully represent the genome of the Nguni cattle ecotypes due to their distinct genetic diversity because it was generated from a different breed. The use of the current reference genome inadequately captures structural variation and ecotype-specific diversity, constraining downstream analyses such as variant discovery and functional annotation (Tixier-Boichard *et al*., 2015). In this study used PacBio HiFi long-read sequencing in combination with Omni-C to construct highly contiguous and accurate genome assemblies for three Nguni cattle ecotypes.

Materials & Methods

Sample collection and sequencing. The study was approved by the University of South Africa ethics committee (reference number: AREC-100818-024). Whole blood was collected from Pure Nguni cattle ecotypes (Venda, Shangaan and Pedi) from Mara Research Station, in Limpopo province of South Africa for extraction of high molecular weight DNA using the Nanobind protocol to construct a sequencing library. Library sequencing was performed on the PacBio revio platform using SMRTbell® prep kit 3.0 (Ekholm *et al*., 2019) following the manufacturer's instructions. Dovetail Omni-C library prep was performed from the same sample used for HiFi sequencing following the manufacturer's instructions on NovaSeq 6000 instrument (Illumina).

Genome Assembly. The genome assemblies of the Nguni cattle ecotypes were carried out in the Galaxy platform following the Vertebrate Genome Project (VGP) (Larivière *et al*., 2024) workflows. We first subjected the HiFi reads to GenomeScope (Vurture *et al*., 2017) for genome profiling to estimate the genome size, heterozygosity and homozygosity of the ecotypes. We then employed hifiasm on HIC (Wang, 2022) mode to assembly the genomes. Following the genome assembly with hifiasm, the assemblies were scaffolded with Omni-C reads using YaHS (Zhou *et al*., 2023) then ran the decontamination using dual decontamination preparation workflow which makes use of Kraken2 (Wood & Salzberg, 2014). The dual decontamination workflow generates scaffolds versus Omni-C reads contact map with both

haplotypes for curation. Pretextview was used to manually curate the scaffolds and by orienting and correcting mis-assemblies using the genomic proximity signal of the Omni-C reads alignment against scaffolds. The genome completeness was further assessed using Benchmarking Universal single-copy Orthologs (BUSCO) with a dataset lineage of cetartiodactyla_odb10.

Genome annotation. The primary phase in genome annotation entails the identification and masking of repeat elements in the Nguni cattle ecotypes of genomes. This process utilizes RepeatMasker (Tarailo-Graovac & Chen, 2009), a software tool that facilitates the identification and masking of repetitive elements within genomic sequences. Nguni ecotype-specific repeats were characterized using RepeatModeler v0.9 (Tarailo-Graovac & Chen, 2009), a software that incorporates tools such as RepeatScout v1.0.6, RECON v1.5.0, and TRF v4.09 (Bailly-Bechet *et al*., 2014). The identified repeat elements were then subjected to annotation using the "one_code_to_find_them_all" Perl script. Protein coding genes for repeat masked assemblies were predicted using TIBERius v1.1.4 (Gabriel *et al*., 2024).

Results & Discussion

The genome sizes are within the expected genome size of the bovine (Figure 1). These are comparable to the reference genome of the Hereford cattle (University of California, 2020) which has a genome size of 2.8 Gb, while the Wagyu has a genome size of 3.1 Gb.

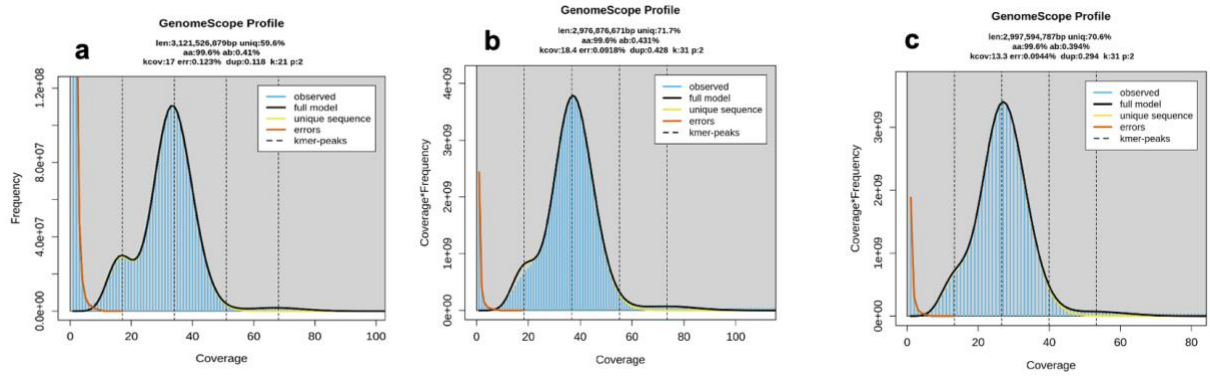

Figure 1. The linear plot generated by genome scope showing the observed K-mer profile for South African indigenous cattle. **a,** Nguni Pedi **b,** Nguni Venda **c,** Nguni Shangaan

The assembly stats of the current genomes are comparable to the reference genome in quality; the Bovine reference of Hereford cattle has scaffold N50 of 103.3 Mb and contig N50 of 26.4. These are comparable to the genomes of the Hereford cattle (University of California, 2020) which has BUSCO completeness of 98.8%. To further explore the genome structural differences between the Nguni ecotypes, we employed the structural variant analysis which uncovered breed specific diversity patterns, where the UpSet analysis in Figure 3 revealed that most features were unique to a breed with Hereford cattle contributing the largest number of specific elements (17,676). Nguni cattle ecotypes also showed substantial numbers of unique features, including Venda (3,217), Shangaan (2,868), and Pedi (1,025). Interestingly, shared features among the breeds were comparatively few, and intersections involving all the breeds were rare, indicating limited overlap across breeds.

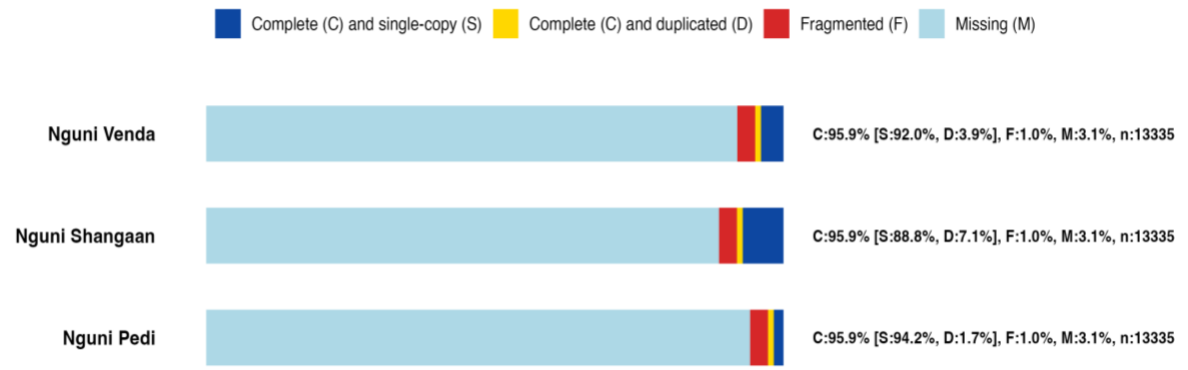

Figure 2. The quality assessment of the genome assemblies for Nguni cattle ecotypes; Nguni Venda, Nguni Shangaan and Nguni Pedi as indicated by BUSCO completeness.

Table 1. Assembly stats for Nguni Cattle ecotypes (Pedi, Venda and Shangaan) as compared to the assembly stats for Genome references of Hereford (University of California, 2020)

Metrics	PEDI	VENDA	SHANGAAN	HEREFORD
Genome Size	3.12 Gb	2.98 Gb	2.99 Gb	2.8 Gb
# Scaffolds	143	1 321	2 224	1 958
Scaffold N50	107 Mb	63 Mb	61 Mb	103 Mb
% Gaps	0.000%	0.000%	0.000%	0.000%
Contig N50	75 Mb	63 Mb	61 Mb	26.4 Mb
GC content	43.45%	45.02%	45.10%	46.50%
***K*-mer complete**	99.5%	99.6%	99.5%	X
Qv score	71.3	66.8	64.3	X

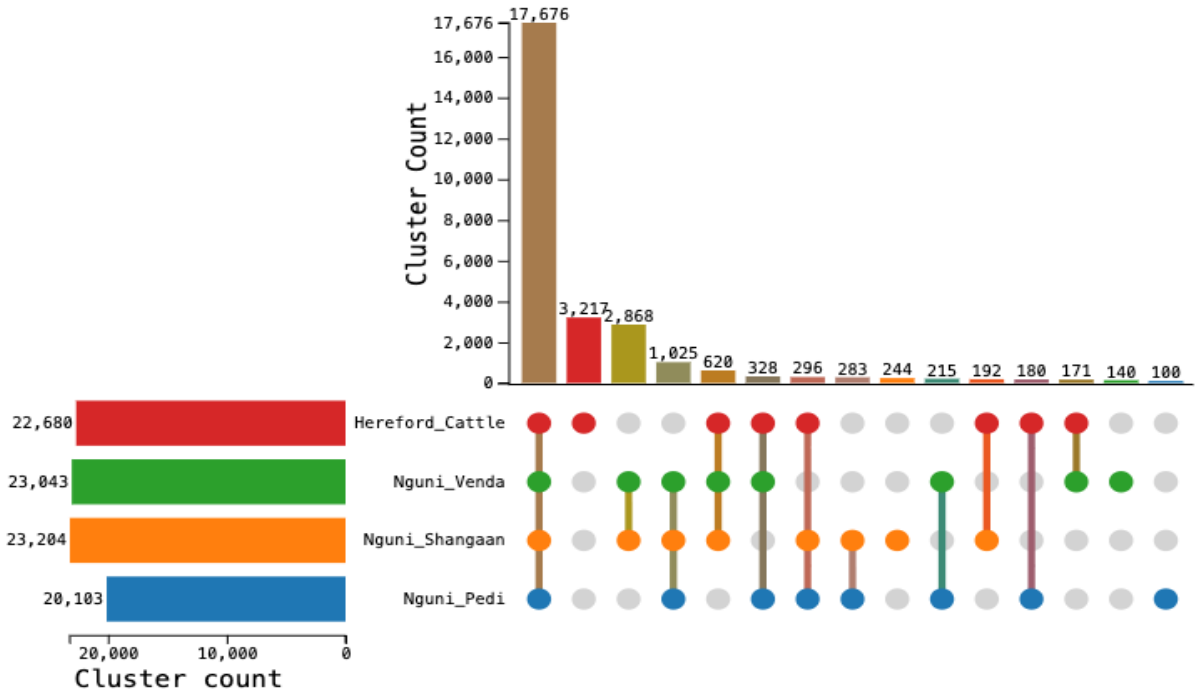

Figure 3. UpSet plot illustrating unique and shared features among the Nguni cattle ecotypes and the bovine reference, Hereford cattle (University of California, 2020). Set sizes are shown on the left, and intersection sizes are shown by vertical bars above the matrix.

Conclusions

The current assemblies exhibit high contiguity and completeness, enabling accurate identification of structural variants and genome-wide restriction sites. Comparative analyses among the three ecotypes revealed genomic differences that may underlie ecotype-specific adaptations. These assemblies provide a valuable genomic resource for conservation, the development of breeding programs, and the study of adaptive traits in African cattle.

References

Bailly-Bechet, M., Haudry, A., and Lerat, E. (2014). Mobile DNA, *5*, 1–15.

Bester, J., Matjuda, L. E., Rust, J. M., *et al*. (2003). A.(Eds), 45–68.

Bruford, M. W., Ginja, C., Hoffmann, I., *et al*. (2015). Frontiers in Genetics, *6*, 314.

Ekholm, J., Gu, J., Burke, J., *et al*. (2019). Journal of Biomolecular Techniques: JBT, 30(Suppl), S38–S38.

Faku, N., and Hebinck, P. (2013). p281–294.

Gabriel, L., Becker, F., Hoff, K. J., *et al*. (2024). Bioinformatics, 40(12), btae685. https://doi.org/10.1093/bioinformatics/btae685

Larivière, D., Abueg, L., Brajuka, N., *et al*. (2024). Nature Biotechnology, *42*(3), 367–370. https://doi.org/10.1038/s41587-023-02100-3

Sanarana, Y., Visser, C., Bosman, L., *et al*. (2016). Tropical Animal Health and Production, 48(2), 379–385.

Tada, O., Muchenje, V., and Dzama, K. (2013). Springerplus, *2*(1), 195.

Tarailo-Graovac, M., and Chen, N. (2009). Current Protocols in Bioinformatics, *25*(1), 4–10.

Tixier-Boichard, M., Verrier, E., Rognon, X., *et al*. (2015). Frontiers in genetics (Vol. 6, p. 153). Frontiers Media SA.

University of California, D. (2020). Genome assembly *ARS-UCD2.0*. NCBI Assembly.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., *et al*. (2017). Bioinformatics, *33*(14), 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Wang, Y. (2022). Proceedings of the 2022 5th International Conf. on Mathematics and Statistics, 100–104.

Wood, D. E., and Salzberg, S. L. (2014). Genome Biology, *15*, R46. https://doi.org/10.1186/gb-2014-15-3-r46

Zhou, C., McCarthy, S. A., and Durbin, R. (2023). Bioinformatics, *39*(1), btac808.