scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A high-quality chromosome level genome assembly of the South African indigenous Nguni goat (*Capra hircus*)

Sinebongo Mdyogolo[1]✉, Lucky Tendani Nesengani[2], Thendo Tshilate[2], Rae Smith[1], Tracy Masebe[1] & Mapholi Ntanganedzeni[2]

Nguni goat is an indigenous goat breed of South Africa characterised by small body frame, variable coat colours and exceptional tolerance to harsh environmental conditions. This breed is utilised for meat and milk production thus sustaining the livelihood of communal farmers and is important for cultural purposes. We assembled a high-quality genome using HiFi and Omni-C reads. For HiFi and Omni-C, we used PacBio sequel IIe and the Illumina NovaSeq 6000, respectively. The total HiFi generated was 106 Gb and Omni-C generated 300 million reads. The assembly length was 2.85 Gb, comprising of 114 scaffolds and 127 contigs; with N50/L50 of 97.7 Mb/12 and 81.34 Mb/12, respectively. BUSCO showed 96% completeness with 3% missing genes making it the most complete genome. The genome contained 42.86% repetitive elements, with LINEs being the most abundant, representing 28.82% and accounting for 67% of total repeats. The annotation predicted 22,507 genes with an average length of 33,847.2 bp. The genome assembly of the Nguni goat will advance genomic research and enhance breeding programs to improve both production and conservation.

Background and summary

Relative to the world, South Africa holds a small number of goat population (1.9 million)[1]. The overall distribution of goat population in South Africa per province is dominated by Eastern Cape (39%), Limpopo (17%), KwaZulu Natal and North-West (13%) with the rest of the provinces sharing the remaining 18% of the populations[1]. Within the South African provinces, generally goats are kept for meat, milk, skin and fibre production[2,3]. The dominant breed within the smallholder and communal farmers is the Nguni goat which is indigenous to South Africa and distributed across all agroecological regions within the tropical climate[4,5]. It is characterised by small body frame (i.e., 35 kg–45 kg), glossy short hair coat and regarded as the breed of choice by communal and subsistence farmers due to low management input requirements[6,7]. It is also known for its ability to adapt in harsh environmental conditions and resistance to internal and external parasites[2,8].

Nguni goat carcass characteristics is perceived as inferior to exotic breeds in the commercial market due to its small size (hot carcass weight: 19.9 kg vs 16.7 kg)[6,9,10]. To enhance production traits, the breed is crossbred with exotic breeds therefore compromising its unique genetic composition. Due to its unique genetic composition and observed phenotypic hardiness in comparison to exotic breeds, there are efforts towards preserving and promoting the use of this breed. However, there is a gap in terms of understanding the genetic architecture of this breed. Conservation of Nguni goat for genetic resource is important to boost economic growth, improve production and select for traits of importance thus addressing Sustainable Development Goals 2 (zero hunger), 8 (decent work and economic growth), 12 (responsible consumption and production) and 15 (Life on land)[11,12]. Like many indigenous species, unique genetic resources of Nguni goat have potential to improve communal farming where there is limited resources and infrastructure for improved productivity. Contrary to commercial breeds, Nguni goat is the preferred breed in the communal farming sector due to their low input requirements.

[1]Department of Life and Consumer Sciences, College of Agriculture and Environmental Sciences, UNISA Science Campus, Florida, Johannesburg, South Africa. [2]Department of Agriculture and Animal Health, College of Agriculture and Environmental Sciences, UNISA Science Campus, Florida, Johannesburg, South Africa. ✉e-mail: mdyogs@unisa.ac.za

Fig. 1 An image of the Nguni goat (black coat) at the Mara Research Station, South Africa.

This warrants the need to conduct genomics studies to document important traits (e.g., growth traits) at the molecular level that may be exploited to improve and enhance production.

Numerous studies have been conducted to characterise the breed's phenotypic traits, however, lack the aspects of genetic composition[7,13–18]. These studies have explored the production performance of the Nguni goat ecotype's in communal and commercial farming enterprises; characterised genetic diversity and identified selective sweeps within the populations[13]. In the area of production performance, researchers documented breeding strategies among smallholder and commercial farmers utilising meat producing goat breeds such as indigenous ecotype breeds (e.g., Cape speckled goat, Nguni and composites such as Boer and Savannah goat)[15,19]. Mdladla *et al.*[13] investigated landscape genomics and biological pathways of Boer, Nguni and Savannah breeds and identified several candidate genes (i.e., *PRKG1*, *PLCB1 and JAK2*) associated with adaptation to harsh environmental conditions[13]. The identified genes are associated with protein digestion and absorption, amino acid metabolism and the ability of goats to utilise low roughage. These studies provide important genetic tools that can be used to improve the indigenous goats performance. Having a high-quality reference genome will complement the available genetic tools to establish efficient breeding programs.

In other instances, the availability of genomic data has led to the establishment of reference genomes for most livestock species (i.e. cattle, goat, sheep, chicken); however, indigenous breeds are underrepresented in these publicly available reference genomes. Therefore, the high-quality reference genome of the indigenous Nguni goat generated in this study is of pivotal importance. Availability of high-quality genome assembly will not only be essential for developing genomic research for indigenous goat breeds but will also be a valuable resource for comparative genomics, functional annotation of genes and evolutionary studies.

Material and Methods

Ethics. The sample collection procedure was approved by the University of South Africa ethics committee (reference number: AREC-100818-024). Further approvals were obtained from the Department of Agriculture, Land Reform and Rural Development (DALRRD) under Section 20 of the Animal Diseases Act (Act 35 of 84) (reference number: 12/11/1/1/23 (6508 AC)) and the Limpopo department of Agriculture.

Sample collection and DNA extraction. In this study, a mature, healthy pure female Nguni goat, (Fig. 1) was selected for sampling from Mara Research Station in Limpopo province, South Africa (23° 05′S, 29° 25′E). This was done in line with the ethics requirements / approval from the University of South Africa which stated that we should sample from matured animals from the age of 1.5 to 2 years. Mara is a designated research station for conservation of indigenous livestock species which is connected to the national breeding herd.

Blood (5 ml) was collected using a needle and EDTA vacuum tube from the jugular vein and stored in dry ice. Prior to sample processing, the samples were stored in a −80 °C freezer at Inqaba Biotech. For PacBio HiFi sequencing, high molecular weight (HMW) genomic DNA was extracted using Nanobind protocol[20]. The concentration of the DNA was assessed with a Qubit 2.0 Fluorometer (Thermo Fisher Scientific), and the integrity was confirmed using a Femto Pulse system (Agilent Technologies, Santa Clara, CA).

Library preparation and sequencing. Library preparation for sequencing on the PacBio Sequel IIe platform was carried out using SMRTbell® prep kit 3.0v[21] following the manufacturer's protocol. For Omni-C, a Dovetail Omni-C library (Dovetail Genomics, Cantata bio) was prepared following manufacturer's instructions and subsequently using Illumina NovaSeq 6000 system for sequencing. Sample setup and run design was performed using browser-based SMRTLink software v11.04[20,22]. The HiFi sequencing resulted in 106 Gb data generated at a coverage of 33X. For Omni-C sequencing, a total of 300 million reads were generated, based on the yield of 100 million read pairs per Gb genome. Analysis of the raw reads was documented (Table 1a) to assess average read length of the HiFi and Omni-C sequences and Merqury completeness was also assessed (Table 1b).

Genome assembly. Genome assembly was performed using the Vertebrate Genome Project (VGP) workflows[23] on the galaxy platform. The genome profiling was first conducted with GenomeScope, applying k-mer

Sequencing platform	Data yield (bp)	Mean read length (bp)	Total bases (bp)	Read quality
PacBio HiFi - Read 1	1,921,204	10,800	20,711,353,231	Q36
PacBio HiFi - Read 2	2,667,246	11,991	31,916,263,564	Q34
PacBio HiFi - Read 3	2,186,218	13,844	30,201,754,645	Q34
PacBio HiFi - Read 4	1,531,633	17,846	26,782,536,804	Q32
Illumina NovaSeq 6000 Forward Read	201,1370,19	150	30,255,819,428	Q20
Illumina NovaSeq 6000 Reverse Read	201,1370,19	150	30,253,600,654	Q20

Assembly	unique k-mers	common k-mers	QV
assembly_01	7215	2935552844	69.3167
assembly_02	4777	2794956241	70.8944
Both	11992	5730509085	70.0152

Table 1. (**a**) Summary statistics of the raw sequence reads. (**b**) Merqury assembly statistics.

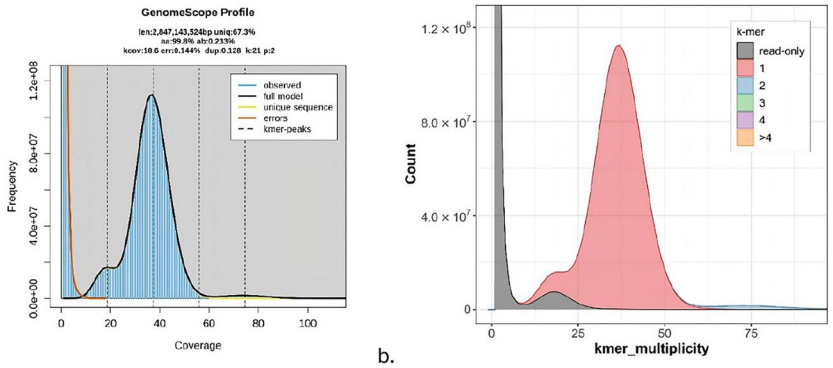


Fig. 2 (**a**) The GenomeScope profile of the Nguni goat displaying predicted genome size based on K-mer analysis, levels of homozygosity (aa), heterozygosity (ab) reads with errors among other information. (**b**) The Merqury ASM plot which shows assembly quality.

Assembly	k-mer set	Solid k-mers in assembly	Solid k-mers in reads	Completeness %	Common k-mers	Unique k-mers	QV
assembly_01	all	1988879398	2077011774	95.7568	2935552844	7215	69.3167
assembly_02	all	1984914029	2077011774	95.5659	2794956241	4777	70.8944
both	all	2069275594	2077011774	99.6275	5730509085	11992	70.0152

Table 2. Merqury statistics completeness.

count (K = 21) to estimate the genome size (2.85 Gb), homozygosity levels of 99.8% and heterozygosity of 0.23% (Fig. 2a)[24]. Assembly was then carried out using Hifiasm[23] with HiC mode, producing two haplotypes (hap1 and hap2) each measuring the genome size of 2.85 Gb. Accuracy of the high-quality genome assembly of the Nguni goat was further assessed using Merqury[25] output (Fig. 2b) and reported 95.76% k-mer completeness and 69–70 quality value (QV) (Table 2). We further assessed the completeness and integrity of the genome assembly by using the Benchmarking Universal Single-Copy Orthologs (BUSCO, version 5.3.2)[26]. BUSCO analysis had a completeness of 96% (Fig. 3). The assembly was scaffolded using YaHs. Following the scaffolding step, the assembly was cleaned of contaminants using Kraken to prepare the genome for manual curation to improve the quality of the assembly.

The Nguni goat genome assembly (**GCA_054126965.1**)[27] statistics was benched marked against five (5) other goat breeds as reported in Table 3. The five breeds included the Telomere-to-Telomere genome of the Inner Mongolian Cashmere (**GCA_040822015.1**)[28], wild goat (*Capra aegagrus*) (**GCA_000978405.1**)[29], Siberian abex (*Capra siberica*) (**GCA_003182615.2**), San Clemente (**GCF_001704415.2**)[30] and Saneen dairy goat (**GCA_015443085.1**)[31].

To further assess the assembly quality, the BUSCO method was used to evaluate completeness of the Nguni Goat genome. Out of the 13 335 BUSCOs searched, 12 797 were complete; 12 564 were complete and single copies, 403 were missing, 135 were fragmented and 233 were complete and duplicated. Overall, the completeness of the Nguni goat genome was 96% (Fig. 3). Among the five goats used for benchmarking, the Nguni goat's BUSCO completeness was high, similar to *Capra Siberica* thus confirm good quality genome assembly.

Alignment of the decontaminated Nguni goat (**GCA_054126965.1**)[27] assembly against the telomere-to-telomere Cashmere genome assembly showed significant collinearity[32] (Fig. 4). What was significant from the alignments was the identification of the sex chromosomes at scaffold 2; the identification of the

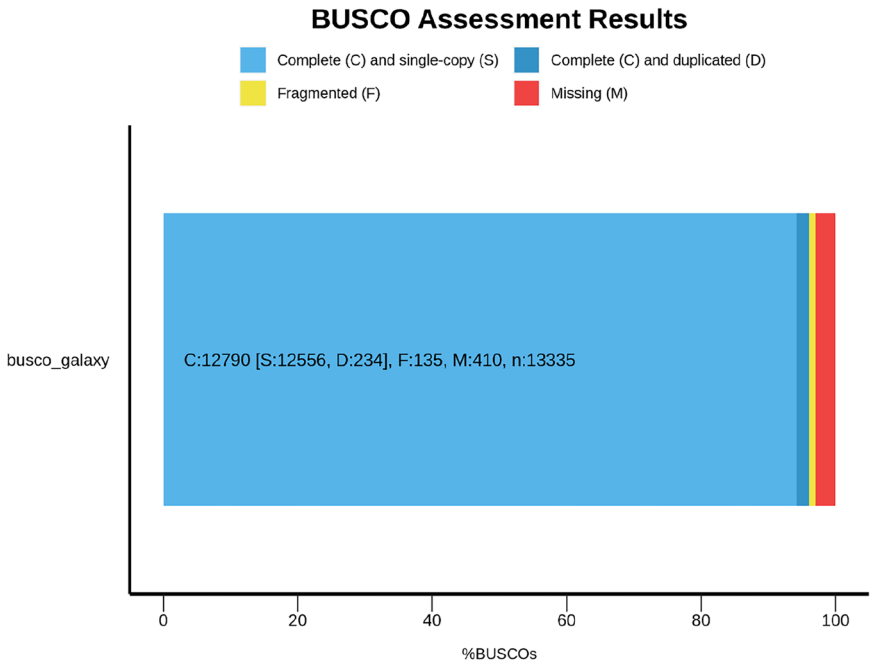


Fig. 3 BUSCO score analysis.

sex chromosomes was based on the reference genome of the Saneen dairy goat and the Cashmere T2T reference genome; these were both recently assembled (i.e., Saneen (**GCA_015443085.1**) in 2020 and Cashmere (**GCA_041053135.1**) in 2024. After decontamination, a Hi-C heat map of the genome assembly was generated in order to view misalignments and mis-joins of hap 1 and hap2 (Fig. 5).

Genome annotation. The Nguni goat genome's repeat elements were found and masked using RepeatMasker as the initial stage of genome annotation[33]. RepeatModeler v0.9, which includes tools like RepeatScout v1.0.6, RECON v1.5.0, and TRF v4.09, was used to undertake de novo repeat modeling to define Nguni-specific repeated elements[34,35]. The Perl script "one_code_to_find_them_all" was used to further annotate the resultant repeats. Repeat analysis revealed that about 1.19 Gb (42.86%) of the genome consists of repetitive elements, a value consistent with other goats. Retroelements are the most abundant among the repeats and account for 37.16% of the genome. Among these, the most abundant are the LINEs (28.82%), followed by SINEs (4.28%) and LTR elements (4.06%) (Fig. 6a). DNA transposons contributed a smaller percentage (1.31%), with contributions from Tc1/mariner and hAT-like elements. 3.24% comprises unclassified repeats and may represent species-specific or highly degenerate sequences. Other repeat classes are small RNAs (3.36%), simple repeats and satellites (1%), and low-complexity sequences (0.15%). Such a repeat composition represents the structural content of the genome and forms the foundation for exploring its evolutionary history and regulatory complexity. Distribution of the repeat elements across the genome is further shown on Fig. 6b.

The repeat-masked assembly was then subjected to genome annotation using TIBERius v1.1.4, a deep learning-based ab initio gene prediction method[36]. Convolutional neural networks, lengthy short-term memory layers, and a differentiable hidden Markov model layer are all used in TIBERius' end-to-end architecture. With an accuracy on par with techniques that depend on outside data, it makes direct predictions about gene architecture from unprocessed genomic sequences. The method generated gene predictions based on inherent structural evidence. This resulted in 22507 predicted protein-coding genes that consisted of exons and introns with average length of 181.78 bp and 4248.25 bp, respectively (Table 4). The BUSCO analysis, based on the cetartiodactyla_odb10 lineage dataset showed that 95% of the BUSCOs were complete, with 94.1% being single-copy and 1.0% duplicated. The annotation process also identified 0.9% fragmented BUSCOs and 4.1% missing BUSCOs, resulting in a total of 13,335 cetartiodactyla BUSCO groups that were searched. Gene distribution across the chromosomes is shown in Fig. 6b. Telomeric regions: The Nguni goat genome displayed a total of seven telomere regions (Fig. 7). The longest telomeric region was found on scaffold 1.

The results from the genome assembly and annotation of the Nguni goat may provide breeders; especially communal and subsistence farmers with crucial information which can be incorporated into their breeding strategies. This present study also paves a way on better understanding the genome architecture of the Nguni goat thus possibly giving researchers an idea of the genome architecture of other indigenous goat breeds.

NCBI submission accession number. JBNQSG000000000

Submitted GenBank Assembly. GCA_054126965.1[27]

Quality metrics	Nguni goat (GCA_054126965.1)	Cashmere goat (GCA_040822015.1)	San Clemente (GCF_001704415.2)	Capra aegagrus (Bezoar/Asian ibex) (GCA_000978405.1)	Saneen Dairy goat (GCA_015443085.1)	Capra Siberica (Siberian ibex) (GCA_003182615.2)
Expected genome size	2.85 Gb	2.8 Gb	2.9 Gb	2.8 Gb	2.7 Gb	2.7 Gb
# scaffolds	114	31	29.906	89.498	1.362	85.609
Scaffold N50	97.7 Mb	95.2 Mb	87.3 Mb	91.3 Mb	102.4 Mb	15.2 Mb
Scaffold L50	12	12	13	13	11	54
# contigs	127	33	30.398	368.693	1.531	113.068
Contig N50	81.34 Mb	95.2 Mb	26.2 Mb	19.3 kb	46.2 Mb	376.6 kb
Contig L50	12	12	32	39.764	20	2.131
GC%	42.8	43	43	42	42.5	42.5
BUSCO	96%	95.9%	93.8%	—	94.3%	96.1%
Genome Coverage	37.80X	50X	50X	83.54X	117X	217X

Table 3. A set of genome assembly statistics between six different breeds of goats.

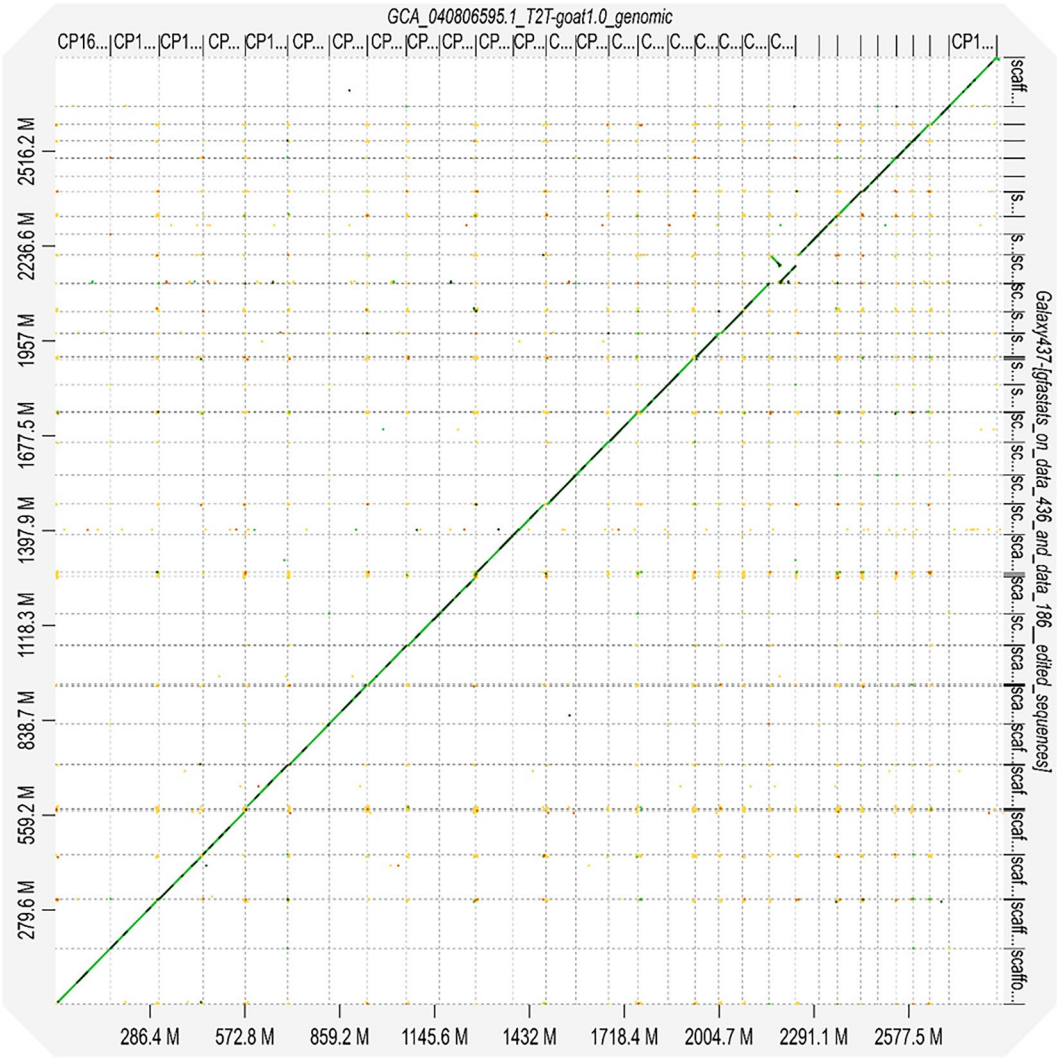


Fig. 4 Alignment of the Nguni goat to the T2T Cashmere goat breed.

NCBI accession for the SRA. PRJNA1227266 (https://www.ncbi.nlm.nih.gov/sra/PRJNA1227266)

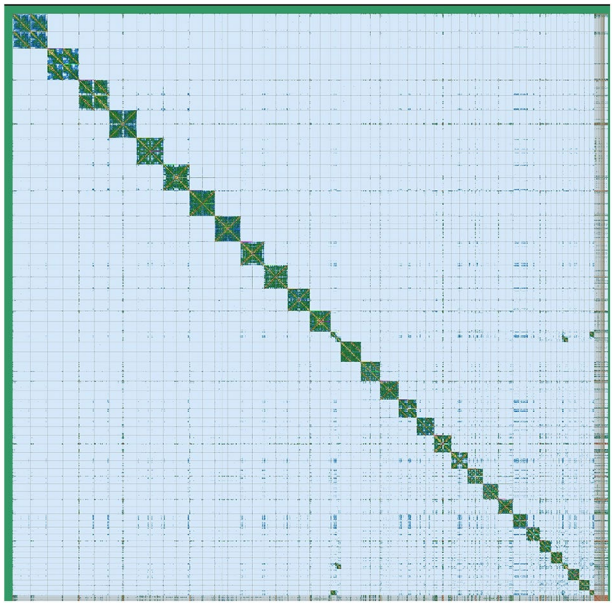
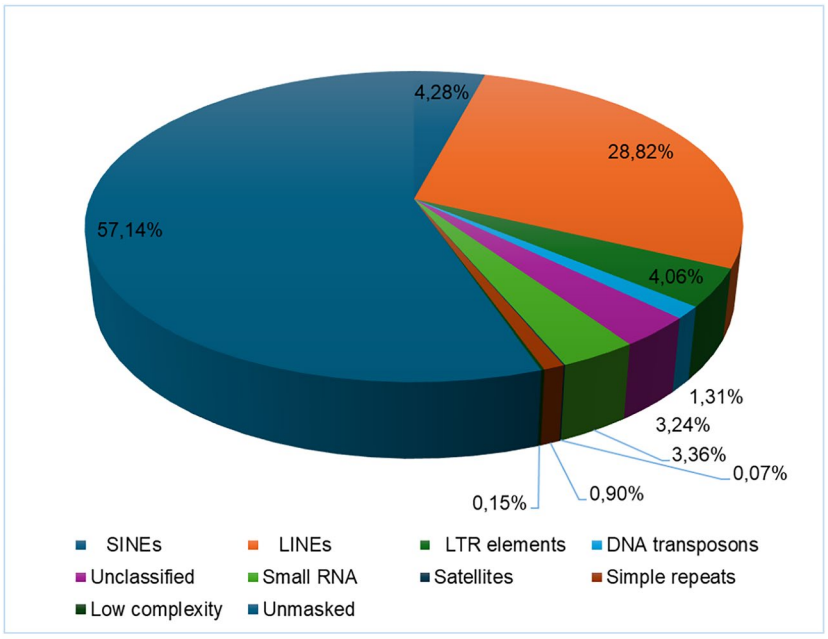


Fig. 5 HiC contact map of the Nguni goat.



a)

Fig. 6 (**a**) Repetitive element distribution in percentage for the Nguni goat genome (**b**) A whole genome visualization integrating chromosome length, GC content (blue), repeat content (red) and gene distribution (green), providing a comparative overview of genomic architecture in the Nguni goat.

Number of genes	22507
Mean gene length (bp)	33847.2
Number of exons per gene	8.59924
Mean exon length (bp)	181.78
Mean intron length (bp)	4248.25
Total intron length	726.617
Average transcript length (bp)	33847.2

Table 4. Genome annotation of the Nguni goat.

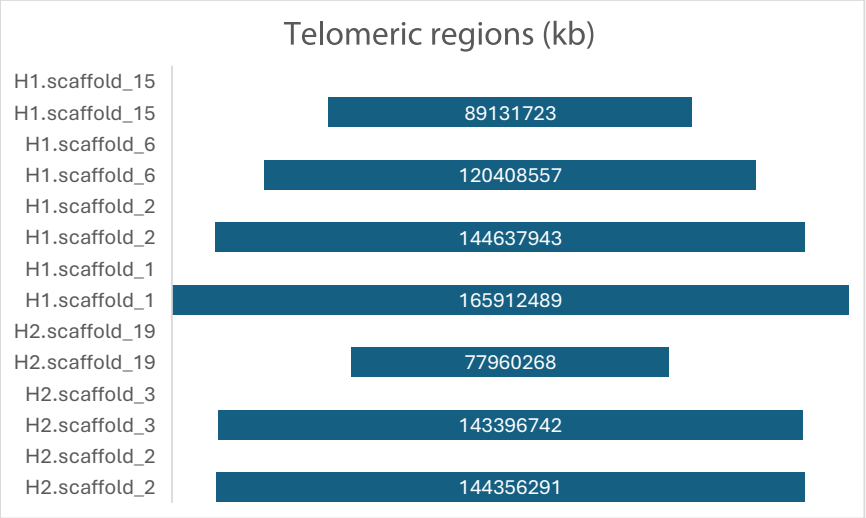


Fig. 7 Total telomeric regions on the Nguni goat after scaffolding with Omni-C.

Data Records

The final genome assembly and raw HiFi reads for the Nguni Goat have been submitted to the NCBI. The submission number for the genome assembly is SUB15281210 and the accession for the completed genome assembly is JBNQSG000000000 and the GenBank Assembly code is GCA_054126965.1[27]. The SRA records are available via SRP576325[37]

Technical Validation

The high-quality reference genome assembly of Nguni Goat was supported by VGP pipeline which is designed to reduce human error by employing regularly updated workflows that produce near error-free high-quality reference genome. The pipeline is comprised of different workflows that make use of methods such as Genome profiling, BUSCO analysis, Mercury and GFA statistics to assess the integrity of the genome.

Data availability

All data are available under the Bioproject PRJNA1227266 or see the specific accession numbers and citations in the Data Record section.

Code availability

All the analyses done in the current study were processed by employing the VGP pipeline, pretext view for editing the omni-C contact map and RepeatMasker, MaskModeler and Tiberius for annotation. All the commands and pipelines were executed following the manual and protocols of the corresponding bioinformatics software. In all the workflows, unless mentioned and where necessary, we used the default parameters.

Received: 4 June 2025; Accepted: 26 January 2026;
Published online: 31 January 2026

References

1. *AGRICULTURAL SECTOR 2023 Agriculture, Land Reform & Rural Development Agriculture, Land Reform and Rural Development REPUBLIC OF SOUTH AFRICA*.
2. Dziba, L. E., Scogings, P. F., Gordon, I. J. & Raats, J. G. Effects of season and breed on browse species intake rates and diet selection by goats in the False Thornveld of the Eastern Cape, South Africa. *Small Ruminant Research* **47**, 17–30 (2003).
3. Slippers, S. C., Letty, B. A. & De Villiers, J. F. *Prediction of the Body Weight of Nguni Goats*. *South African Journal of Animal Science* vol. 30 http://www.sasas.co.za/Sajas.html (2000).
4. Mathapo, M. C., Mugwabana, J. T. & Tyasi, T. L. Phenotypic characterization of Nguni goats in four agro-ecological zones of Limpopo province, South Africa. *PLoS One* **19** (2024).
5. Campbell, Q. P. *The Origin and Description of Southern Africa's Indigenous Goats*. vol. 4 http://www.sasas.co.za/Popular/Popular.html18 (2003).
6. Rashijane, L. T., Mokoena, K. & Tyasi, T. L. Using Multivariate Adaptive Regression Splines to Estimate the Body Weight of Savanna Goats. *Animals* **13** (2023).
7. Ndou, S. P., Muchenje, V. & Chimonyo, M. Behavioural responses of four goat genotypes to successive handling at the farm. *Afr. J. Biotechnol.* **9**, 8118–8124 (2010).
8. Monau, P., Raphaka, K., Zvinorova-Chimboza, P., Gondwe, T. Sustainable Utilization of Indigenous Goats in Southern Africa. Diversity, **12**, 20. https://doi.org/10.3390/d12010020 (2020).
9. Pieters_South (2009).
10. Monau, P. I., Visser, C., Nsoso, S. J. & Van Marle-Köster, E. Phenotypic and genetic characterization of indigenous Tswana goats. *S. Afr. J. Anim. Sci.* **48**, 925 (2019).
11. *The 2030 Agenda for Sustainable Development's 17 Sustainable Development Goals (SDGs)*.
12. Visser, C. & Snyman, M. A. Incorporating new technologies in breeding plans for South African goats in harsh environments. *Animal Frontiers* **13**, 53–59 (2023).

13. Mdladla, K., Dzomba, E. F. & Muchadeyi, F. C. Landscape genomics and pathway analysis to understand genetic adaptation of South African indigenous goat populations. *Heredity (Edinb)* **120**, 369–378 (2018).
14. Chokoe, T. C. *et al*. Genetic diversity of south african indigenous goat population from four provinces using genome-wide SNP data. *Sustainability (Switzerland)* **12**, 1–16 (2020).
15. Thobile, F. M. *et al*. Goat farmers production objectives and trait preferences in the North West province of South Africa: An approach to identify selection criteria for community-based breeding program. *International Journal of Livestock Production* **12**, 64–75 (2021).
16. Tyasi, T. L., Ng'ambi, J. & Mogashoa, S. Breeding practices and trait preferences of goat keepers at Lepelle-Nkumpi Local Municipality, South Africa: implication for the design of breeding programmes. *Trop. Anim. Health Prod*. **54** (2022).
17. Malan, S. W. *The Improved Boer Goat*.
18. Panayotov, D. *Heart Girth (HG), Cannon Circumference (CC), Head Length (HDL), Head Width (HDL), Horns Length (HoL). The BW of Does Was 44. Bulgarian Journal of Agricultural Science* vol. 28 https://www.bo- (2022).
19. Visser, C. A review on goats in southern Africa: An untapped genetic resource. Small Ruminant Research 176, 11–16. https://doi.org/10.1016/j.smallrumres.2019.05.009 (2019).
20. Mitchell, M. W. *et al*. *High Molecular Weight DNA Extraction and Long-Read next-Generation Sequencing of Human Genomic Reference Standards*.
21. Ekholm, J. *et al*. *Single Tube Addition-Only Reactions to Generate SMRTbell Libraries (Approx. 4 Hrs*) QC and GDNA Shearing Anneal Primer and Bind Polymerase for Sequencing ABSTRACT LARGE-INSERT GDNA QC AND SEQUENCING RESULTS Multiplexed Microbial Genome Assembly Results Streamlined SMRTbell ® Library Generation Using Addition-Only, Single Tube Strategy for All Library Types Reduces Time to Results. EXPRESS 2.0 WORKFLOWS CONCLUSION MULTIPLEXED MICROBIAL GDNA QC AND SEQUENCING RESULTS* (2019).
22. PacBio. *SMRT ® Link Software Installation Guide (V11.0)*. http://www.pacificbiosciences.com/licenses.html. (2022).
23. Larivière, D., Abueg, L., Brajuka, N. *et al*. Scalable, accessible and reproducible reference genome assembly and evaluation in Galaxy. Nat Biotechnol 42, 367–370. https://doi.org/10.1038/s41587-023-02100-3 (2024).
24. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun*. **11** (2020).
25. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: Reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. **21** (2020).
26. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr. Protoc*. **1** (2021).
27. Mdyogolo, S. *et al*. Genbank https://identifiers.org/ncbi/insdc.gca:GCA_054126965.1 (2025).
28. Wang, Z. *et al*. Chromosome-level genome assembly of the cashmere goat. *Sci. Data* **11**, 1107 (2024).
29. Dong, Y. *et al*. Reference genome of wild goat (capra aegagrus) and sequencing of goat breeds provide insight into genic basis of goat domestication. *BMC Genomics* **16** (2015).
30. *USDA ARS Capra Hircus Breed:San Clemente Genome Sequencing and Assembly. National Center for Biotechnology Information. Dataset*. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_001704415.2/ (2016).
31. Li, R. *et al*. A near complete genome for goat genetic and genomic research. *Genetics Selection Evolution* **53** (2021).
32. Cabanettes, F. & Klopp, C. D-GENIES: Dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **2018** (2018).
33. Tarailo-Graovac, M. & N. Chen. "Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences." Current Protocols in Bioinformatics 25, no. 1: 4.10.1–4.10.14. https://doi.org/10.1002/0471250953.bi0410s25 (2009).
34. J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, & A. F. Smit, RepeatModeler2 for automated genomic discovery of transposable element families, *Proc. Natl. Acad. Sci. USA* 117 (17) 9451-9457, https://doi.org/10.1073/pnas.1921046117 (2020).
35. Manchanda, N. *et al*. GenomeQC: A quality assessment tool for genome assemblies and gene structure annotations. *BMC Genomics* **21** (2020).
36. Gabriel, L., Becker, F., Hoff, K. J. & Stanke, M. Tiberius: end-to-end deep learning with an HMM for gene prediction. *Bioinformatics* **40** (2024).
37. Mdyogolo, S. *et al*. *NCBI Sequence Read Archive*, https://identifiers.org/ncbi/insdc.sra:SRP576325 (2025).

Acknowledgements

This project is supported by University of South Africa Africa-nuanced sustainable development goals research support programme. The samples for this project were provided by Mara Research Station, Limpopo province of South Africa and the technical support was provided by the Africa BioGenome Project. A special thanks to Mapholi labs for providing the resources and managing the project.

Author contributions

Conceptualization: S.M., L.T.N., N.M.; Data Curation: S.M., T.T., L.T.N.; Formal Analysis: S.M., L.T.N., T.T.; Funding Acquisition: N.M., T.M.; Investigation: S.M., T.T., L.T.N.; Methodology: S.M., L.T.N., T.T., R.M.S.; Project Administration: N.M., T.M.; Resources: L.T.N., S.M., Software: S.M.; Validation: S.M., L.T.N., T.T., R.M.S.; Visualisation: S.M.; Writing of the original draft: S.M.; Reviewing and editing of the manuscript: S.M., L.T.N., T.T., R.M.S., N.M., T.M.

Competing interests

The authors declare no competing interests..

Additional information

Correspondence and requests for materials should be addressed to S.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.