# Genome-wide marker discovery in three South African indigenous cattle breeds (Afrikaner, Drakensberger and Nguni) using whole genome sequencing

**By**

**AVHASHONI AGNES ZWANE**

A thesis submitted in partial fulfilment of the requirements for the degree of

**PHILOSOPHIAE DOCTOR (ANIMAL SCIENCE)**

In the

Faculty of Natural and Agricultural Sciences

Department Animal and Wildlife Sciences

**UNIVERSITY OF PRETORIA**

November 2017

## SUPERVISORY COMMITTEE


E. van Marle-Köster: Department of Animal and Wildlife Sciences
University of Pretoria
P/Bag X20
Hatfield
Pretoria
0028
South Africa

A. Maiwashe: Animal Production Institute
Agricultural Research Council
Private Bag X2,
Irene
Pretoria
0062

J.F. Taylor: Division of Animal Sciences
University of Missouri
920 East Campus Drive
Columbia
Missouri
65211-5300
United States of America

## DECLARATION

I, Avhashoni Agnes Zwane hereby declare that this thesis contains a literature review and original research work undertaken by the undersigned candidate as part of her Ph.D. (Animal Science). All information in this document has been obtained and presented in accordance with academic rules and ethical conduct. All materials and results that are not original to this work have been fully cited and referenced.

Signature: ……

Date: ……...

## ACKNOWLEDGEMENTS

I would like to express my special appreciation and thanks to my supervisors Professor Este van Marle-Koster, Professor Azwihangwisi Maiwashe and Professor Jeremy Taylor for continuous support since the commencement of this research. You have been tremendous mentors for me. Thank you for encouraging my research and for allowing me to grow as a research scientist. I would also like to thank my research team members, Dr Ananyo Choudhury, Dr Mahlako Makgahlela and Associate Professor Robert Schnabel for being there for me when I needed technical and statistical support. Your work has been valuable and contributed to the completion of this work. Thank you for your valuable comments and suggestions.

Thanks to the ARC Biotech Platform for allowing us to explore the newest genomic technology in the country, your efforts to improve South African research are really significant, the South African livestock industry will never be the same. Special thanks to the University of Missouri Animal Genomics team, your training, support and continuous contribution helped me to handle and manage genomic data during data analysis. Your love during my stay in Columbia made me feel at ease. Thank you very much for all your efforts.

A big thanks to the National Research Foundation for allowing me the opportunity to further my studies through the Doctoral Scholarship that sustained me until the completion of this work. I wouldn't be where I am today without your support. I would also like to thank the Red Meat Research and Development Trust, and the Agricultural Research Council for financial support.

A special thanks to my family. Words cannot express how grateful I am to my beloved husband Zwelethu and my kids, Teddy and Thandekile for all of the sacrifices that you've made on my behalf. Your prayers, your support and your tolerance has sustained me thus far. Thank you for being there for me always, and for encouraging me throughout this challenging and life changing experience.

And above all, thanks to the almighty God for giving me the strength, perseverance and sustainability even when I felt like quiting. Your word encouraged me, for I shall be the head and not the tail (Deuteronomy 28:13).

## EXECUTIVE SUMMARY

South Africa (SA) has only recently engaged in programs to establish genomic data for the cattle industry. Due to the important role of indigenous cattle breeds in SA, it is imperative that these breeds be included in the generation of genotypic and sequence data. Genomic data provide opportunity for various genetic investigations including identification of breed-informative markers, selective sweeps and genome-wide association studies (GWAS). In this study sequence data were generated and used in combination with genotypic data to conduct a SNP discovery in the three indigenous SA breeds (Afrikaner, Drakensberger, and Nguni) and study potential selective sweeps. Commercial bovine SNP assays, (BovineSNP50 and GGP-80K) were applied for identifying the breed-informative markers, while an approach of breed pooled samples were used for sequencing. The study was conducted in phases and results were presented in different chapters prepared for submission to specific journals. The thesis is hereby presented with an introduction and literature review followed by Chapter 3 that resulted in first publication in SA Journal of Animal Science. Chapters 4 and 5 have been prepared for submission to the international journals. The referencing style was consequently prepared in a similar manner except for the published article. The thesis was concluded with a critical review and discussion.

## ABSTRACT

Afrikaner, Drakensberger, and Nguni are the South African (SA) landraces that played major roles in the social, cultural and economic history of SA. These breeds are valuable genetic resources for beef production and limited information is available for these breeds at the genome level. The aim of this study was to perform SNP discovery in these three breeds using whole genome sequencing. Ninety cattle representing the three breeds were used to identify more about 17.6 M putative variants including SNPs and Indels. DNA was extracted from blood and hair samples, quantified and prepared at 50ng/µl concentration for sequencing at the Agricultural Research Council Biotechnology Platform using an Illumina HiSeq 2000. The fastq files were used to call the variants using the Genome Analysis Tool Kit. A total of 4,369,879 (16% of the total SNPs) were identified as novel. Annotation of these variants classified them into functional categories. Within the coding regions, 43% of the SNPs were nonsynonymous substitutions that encode for alternate amino acids. Functional enrichment analysis of novel SNPs identified significant number of genes ($p < 0.001$) that were located within 5% of 1,481 100kb windows. Gene ontology terms identified genes such as *MLANA* and *SYT10* that have been associated with coat colour and sense of smell in mouse, respectively, and the *ADAMS3* gene has been associated with fertility in cattle. Furthermore, whole genome screening detected 688 candidate selective sweeps ($ZH_p$ Z-scores $\leq -4$) across all three breeds, of which 223 regions were assigned as being putative selective sweeps ($ZH_p$ scores $\leq -5$). We also identified 96 regions with extremely low $ZH_p$ Z-scores ($\leq -6$) in Afrikaner and Nguni. Several genes such as *KIT* and *MITF* that have been associated with skin pigmentation in cattle, and *CACNA1C*, which has been associated biopolar disorder in human were identified in these regions. Breed-specific SNPs (2,272,667) were identified across the breeds and only 186 of these SNPs were identified as putative breed-specific SNPs. These SNPs were further tested for their ability to assign individuals to a breed and need further validation. This study provides the first analysis of sequence data to discover SNPs in indigenous SA cattle breeds. These results provide insight into the genetic composition of the breeds and offer the potential for further applications in their genetic improvement.

## SCIENTIFIC OUTPUTS

**Publication:**

Zwane, A.A., Maiwashe, A., Choudhury, A., Makgahlela, M.L., Taylor, J.F. and Van Marle-Köster, E. (2016). South African Journal of Animal Science, 46 (3), 1-11.

**Congresses:**

*National*

Zwane, A.A., Maiwashe, A., Makina, S.O., Mapholi, N.O. and Van Marle-Köster, E. Identification of differentiated SNPs for breed assignment in selected South African beef cattle breeds. Proceedings of the 47th South African Society of Animal Science (SASAS) Congress, University of Pretoria, Pretoria, 6-8 July 2014.

Zwane, A.A., Maiwashe, A., Makina, S.O., Mapholi, N.O. and Van Marle-Köster, E. Selection of informative SNPs for breed assignment in South African indigenous and locally-developed beef cattle breeds. Proceedings of the joint South African Genetics Society (SAGS) & South African Society for Bioinformatics SASBI Congress, Kwalata Game Ranch, Hammanskraal, 23-26 September 2014.

Zwane, A.A. Genome-wide selection of informative SNPs in South African indigenous cattle breeds. Proceedings of the World Braford Congress, Pretoria, 15-18 March 2015.

Zwane, A.A., Maiwashe, A., Choudhury, A., Taylor, J.F. and Van Marle-Köster, E. SNP discovery in Afrikaner, Drakensberger and Nguni indigenous breeds of South Africa. Proceedings of the 49th South African Society of Animal Science (SASAS) Conference, Spier Hotel and Congress Centre, Stellenbosch, Western Cape, 3-6 July 2016.

Zwane, A.A., Choudhury, A., Makgahlela, M.L., van Marle-Köster, E., Maiwashe, A. and Taylor, J.F. Identification of selective sweeps in Afrikaner, Drakensberger and Nguni cattle using genome-wide sequence data. Boardwalk Conference Centre, Port Elizabeth, 18-21 September 2017.

*International*

Zwane, A.A., Maiwashe, A., Makina, S.O., Choudhury, A., Mapholi, N.O. and Van Marle-Köster, E. Selection of informative SNPs for breed assignment in South African indigenous and locally-developed beef cattle breeds. Proceedings of the 34th International Society of Animal Genetics (ISAG) Conference, Xi'an, China, 27 July - 1 August 2014.

Zwane, A.A., Choudhury, A., Makgahlela, M.L., van Marle-Köster, E., Maiwashe, A. and Taylor, J.F. SNP discovery in indigenous Afrikaner, Drakensberger and Nguni cattle breeds of South Africa, Dublin, Ireland, 16-21 July 2017.

## TABLE OF CONTENTS

## LIST OF FIGURES

### CHAPTER TWO



### CHAPTER THREE



### CHAPTER FOUR



### CHAPTER FIVE

# LIST OF TABLES

## CHAPTER THREE



## CHAPTER FOUR



## CHAPTER FIVE

# ADDENDA

## LIST OF ABBREVIATIONS

| | |
|---|---|
| A | Adenine |
| ABI | Applied Biosystems |
| AFR | Afrikaner |
| ANG | Angus |
| ARC | Agricultural Research Council |
| AVS | SNP & Variation Suite |
| BRAH | Brahman |
| C | Cytocine |
| CNV | Copy number variation |
| DAFF | Department of Agriculture, Forestry and Fisheries |
| dbSNP | SNP database |
| DNA | Deoxyribonucleic acid |
| DRA | Drakensberger |
| EBVs | Estimated Breeding Values |
| FAO | Food Agricultural Organization |
| FLK | Extended Lewontin and Krakauer statistics |
| $F_{st}$ | Fixation index |
| G | Guanine |
| GATK | Genome Analysis Tool Kit |
| GC | Guanine-cytosine |
| Gb | Gigabase |
| GDP | Gross domestic product |
| GGP-HD | High Density GeneSeek Genomic Profiler |
| GS | Genomic selection |
| GVCF | Genomic variant call format |
| GWAS | Genome-wide association studies |
| He | Heterozygosity |
| HFD | Hereford |
| Ho | Homozygosity |
| HWE | Hardy-Weinberg equilibrium |
| kb | Kilobase |
| KUR | Kuri |
| LD | Linkage disequilibrium |
| MAF | Minor allele frequency |

| | |
|---|---|
| NDAM | N’Ndam |
| NGI | Nguni |
| NGS | Next generation sequencing |
| NGU | Nguni |
| nsSNPs | Nonsynonymous SNPs |
| PCA | Principal component analysis |
| PCR | Polymerase chain reaction |
| qPCR | Real-time polymerase chain reaction |
| QTLs | Quantitative trait loci |
| QTLdb | Quantitative trait loci database |
| RNA | Ribonucleic acid |
| SA | South Africa |
| SD | Standard deviations |
| SGS | Second-generation sequencing |
| SMS | Single-molecule sequencing |
| SNPs | Single nucleotide polymorphisms |
| T | Thymine |
| Ti/Tv | Transition-to-transversion |
| UMD3.1 | *Bos taurus* reference genome |
| VEP | Variant Effect Predictor |
| WGS | Whole-genome sequencing |
| ZHp | Pooled heterozygosity |
| ZMA | Zebu from Madagascar |

# CHAPTER ONE

# Introduction

## 1.1. Motivation for the study

A rapid development of next generation sequencing (NGS) technologies has been witnessed over the past three decades, providing new prospects for the development of genomic tools to enhance genetic progress in livestock production (Anderson & Schrijver, 2010). Large volumes of sequence data can be cost-effectively and accurately generated in a relative short period, to accelerate scientific discoveries in livestock species (Ramos et al., 2011). These technologies have allowed the discovery and high throughput genotyping of thousands of single nucleotide polymorphisms (SNPs) in livestock species including cattle (Nishimura et al., 2013). These tools and developments are believed to play a central role in current and future studies in livestock genetics and improvement (Ramos et al., 2011; Sabir et al., 2014).

The global demand for livestock products, specifically the need for animal protein is escalating due to a growing human population. Future projections suggest a world population of approximately 8 billion by 2025, 11 billion by 2050 and 16 billion world inhabitants by 2100 (Ilea, 2009; FAO, 2011). The current needs already exceed the earth's bio-productive capacity for the total inhabitants. Efforts must be made to increase the productivity of all livestock production systems in order to satisfy the demand for animal products (Webb, 2013; van Marle-Kőster et al., 2015). This should be done through advancing production systems and by employing new technologies.

South Africa (SA) is a diverse country with rich cultural diversity and different types of vegetation, biodiversity, environments and soil types. This country has particular farming regions and various farming practices (Goldblat, 2010). Agricultural activities varies from thorough crop production in winter precepitation, higher summer precipitation zones, to cattle farming in the bushveld and sheep farming in most dry areas (Goldblat, 2010; Shabalala & Combrinck, 2012). Mixed farming systems are employed with livestock being the largest sector (Rust & Rust, 2013; Goldblatt, 2010). The demand for meat and milk generally exceeds production, though there are untapped reserves in the communal farming areas. Beef farming produces about 85% of the consumed meat, while 15% is imported from other African countries and Europe (Webb, 2013). Cattle are found throughout SA regions, but mainly in

Eastern Cape, KwaZulu-Natal, Free State and North West provinces. Herd sizes vary according to production system. In terms of dairy cattle, herd size are estimated from < 50 to > 1 000 while beef cattle herds range from fairly small (< 20 head of cattle) to extensive farms and feedlots (more than 1 000 head). Vryburg, in the North West Province, has some of the largest cattle herds in SA. Weaner's production is a cattle-farming system widely used in SA for feedlot industry, and of all the beef produced, feedlots account for approximately 75%. In SA, the number of cattle was projected at 13.7 million by the end of 2015, comprising of various international beef and dairy cattle breeds including SA indigenous breeds such as the Afrikaner (AFR), Drakensberger (DRA) and Nguni (NGI) (DAFF, 2016). Beef cattle contribute approximately 80% of the total number of cattle in SA, while dairy cattle contribute the remaining 20% (DAFF, 2013). In general, the main agricultural sector contributes about 3% to the country's gross domestic product (GDP), and it represents about 7% of formal employment. The whole agricultural value chain contributes 12 % GDP (DAFF, 2014).

South African indigenous cattle such as AFR, DRA, NGI, Bonsmara and Tuli have played a major role in traditional, social, and commercial history of the country (Scholtz, 2010). These breeds provide valuable farm animal genetic resources for beef production in SA in combination with exotic beef breeds that were introduced in SA many decades ago (Scholtz, 2010). Currently, relatively little information is available on these SA breeds at the genome level, including sequence variation. With recent advances in NGS technologies, it is now possible to sequence these local breeds to identify millions of SNPs, with the potential of identifying genes and mutations that lead to variation in economically important traits (Wiedmann et al., 2008; Buermans & Dunnen, 2014). The use of NGS for variant discovery in these breeds offers opportunities for genome-wide association studies (GWAS) for the discovery of variants underlying adaptive and production traits. The utilization of these variants may provide opportunities for improving breeding objectives to overcome the limitations of traditional breeding programs in SA (Hayes et al., 2015). Genetic improvement has also resulted to the commercialization of number of livestock species in such a way that validation of breeds, in global and domestic markets, has increasingly become essential for the safety and validity of livestock products (Pant et al., 2012). Therefore, the identification and evaluation of breed-specific SNPs is essential for discriminating between cattle breeds, including local beef breeds. The availability of the genome-wide sequence data provides the opportunities to develop the required tools.

In cattle, genetic variants have been intensively discovered and annotated since the completion of the bovine sequencing project in 2009 (Stothard et al., 2011; Mei et al., 2016). SNPs have been the most widely used variants in association studies to identify genes and genomic regions responsible for genetic variation in cattle (Lu et al., 2013; Choi et al., 2015). Currently, SNP assays such as BovineSNP50, High Density GeneSeek Genomic Profiler (GGP-HD) and BovineHD BeadChips are available for genome-wide studies in cattle (Van Tassell et al., 2008; Matukumalli et al., 2009). The design of these assays included mostly common SNPs within European taurine breeds (*Bos taurus*) and the assays contain less informative SNPs for indigenous SA breeds. These assays result in lower minor allele frequency (MAF) and lower levels of linkage disequilibrium (LD) in local SA breeds such as Sanga and indicine compared to the taurine breeds (Edea et al., 2012; Makina et al., 2014). This design bias can have a significant impact on the deployment of these assays for GWAS and the detection of quantitative trait loci (QTLs) and genes associated with economically important traits in local breeds (Albrechtsen et al., 2010). Ascertainment bias in the detection of the origin of SNPs could also generate misleading conclusions in determining the degree of differentiation and similarities between the breeds (McKay et al., 2008). Therefore, it is essential to sequence the whole genomes of indigenous SA cattle to discover new SNPs and search for variants within genes related to traits of interest.

Whole-genome sequencing (WGS) is a powerful approach for mining millions of SNPs leading to the identification of genetic variants present in the populations, and also genes of economic importance in cattle. Studying the nature and extent of genetic variations between individuals provides a basis for understanding the heritability of traits and phenotypes, and offers prospects to study complex issues in molecular ecology, conservation, disease susceptibility, and other related disciplines (Le Roex et al., 2012). The aim of this study was to use an NGS technology to discover new SNPs in South African AFR, DRA and NGI breeds. The availability of full genome sequence data for these breeds would enhance our understanding of breed composition, genomic regions under selection and the selected traits, as well as the level of genetic diversity within and between breeds.

### 1.2. Aim of the study

The aim of this study was to perform genome-wide marker discovery in three South African indigenous cattle breeds, AFR, DRA and NGI using next generation sequencing technology on pooled DNA samples.

### 1.3. Objectives

In order to realize the aim of this study, the following objectives were set:

- To identify breed-informative markers in AFR, DRA and NGI using BovineSNP50 and GGP-80K BeadChip data.

- Sequence pooled DNA samples from AFR, DRA and NGI breeds using next generation sequencing to search for new variants at a genome-wide level.

- To validate newly identified SNPs using Run 5 data from the 1000 Bull Genomes Project and perform functional annotation and enrichment analysis.

- To identify selective sweeps and a panel of SNP markers to discriminate between the three indigenous breeds

### References

Albrechtsen, A., Nielsen, F.C. & Nielsen, R., 2010. Ascertainment biases in SNP chips affect measures of population divergence. Mol. Biol. Evol. 27, 2534-2547.

Anderson, M.W. & Schrijver, I., 2010. Next generation DNA sequencing and the future of genomic medicine. Genes. 1, 38-69.

Buermans, H.P.J. & Den Dunnen, J.T., 2014. Next generation sequencing technology: advances and applications. Biochim Biophys Acta. 1842, 1932-1941.

Choi, J.W., Choi, B.H., Lee, S.H., Lee, S.S., Kim, H.C., Yu, D., Chung, W.H., Lee, K.T., Chai, H.H., Cho, Y.M. & Lim, D., 2015. Whole-genome resequencing analysis of Hanwoo and Yanbian cattle to identify genome-wide SNPs and signatures of selection. Mol. Cells. 38, 466-473.

DAFF., 2013. Trends in the Agricultural Sector. pp 1-73.

DAFF., 2014. Pocket Guide to South Africa 2014/15.

DAFF., 2016. Abstract of agricultural statistics. pp 1-106

Edea, Z., Dadi, H., Kim, S.W., Dessie, T. & Kim, K.S., 2012. Comparison of SNP variation and distribution in indigenous Ethiopian and Korean Cattle (Hanwoo) populations. Genomic Inform. 10, 200-205.

FAO., 2011. Mapping supply and demand for animal-source foods to 2030, by T.P. Robinson & F. Pozzi. Animal Production and Health Working Paper. No. 2. Rome. pp 1-154.

Goldblatt, A., 2010. Agriculture: Facts & Trends. WWF-SA, pp 1-32.

Hayes, M.G., Urbanek, M., Ehrmann, D.A., Armstrong, L.L., Lee, J.Y., Sisk, R., Karaderi, T., Barber, T.M., McCarthy, M.I., Franks, S., Lindgren, C.M., Welt, C.K., Diamanti-Kandarakis, E., Panidis, D., Goodarzi, M.O., Azziz, R., Zhang, Y., James, R.G., Olivier, M., Kissebah, A.H.; Reproductive Medicine Network., Stener-Victorin, E., Legro, R.S. & Dunaif, A., 2015. Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations. Nat. Commun. 6, 7502, 1-13.

Ilea, R.C., 2009. Intensive livestock farming: Global trends, increased environmental concerns, and ethical solutions. J. Agr. Environ. Ethic. 22, 153-167.

Le Roex, N., Noyes, H., Brass, A., Bradley, D.G., Kemp, S.J., Kay, S., Van Helden, P.D. & Hoal, E.G., 2012. Novel SNP discovery in African buffalo, Syncerus caffer, using high-throughput sequencing. PloS One. 7, e48792, 1-6.

Lu, D., Miller, S., Sargolzaei, M., Kelly, M., Vander Voort, G., Caldwell, T., Wang, Z., Plastow, G. & Moore, S., 2013. Genome-wide association analyses for growth and feed efficiency traits in beef cattle. J. Anim. Sci. 91, 3612-3633.

Makina, S.O., Muchadeyi, F.C., van Marle-Köster, E., MacNeil, M.D. & Maiwashe, A., 2014. Genetic diversity and population structure among six cattle breeds in South Africa using a whole genome SNP panel. Front. Genet. 5, 333, 1-7.

Matukumalli, L.K., Lawley, C.T., Schnabel, R.D., Taylor, J.F., Allan, M.F., Heaton, M.P., O'Connell, J., Moore, S.S., Smith, T.P., Sonstegard, T.S. & Van Tassell, C.P., 2009. Development and characterization of a high density SNP genotyping assay for cattle. PloS One. 4, e5350, 1-13.

Mei, C., Wang, H., Zhu, W., Wang, H., Cheng, G., Qu, K., Guang, X., Li, A., Zhao, C., Yang, W., Wang, C., Xin, Y. & Zan, L., 2016. Whole-genome sequencing of the endangered bovine species Gayal (*Bos frontalis*) provides new insights into its genetic features. Sci. Rep. 6, 1978, 1-8.

McKay, S.D., Schnabel, R.D., Murdoch, B.M., Matukumalli, L.K., Aerts, J., Coppieters, W., Crews, D., Neto, E.D., Gill, C.A., Gao, C., Mannen, H., Wang, Z., Van Tassell, C.P., Williams, J.L., Taylor, J.F. & Moore, S.S., 2008. An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. BMC Genet. 9, 37, 1-9.

Nishimura, S., Watanabe, T., Ogino, A., Shimizu, K., Morita, M., Sugimoto, Y. & Takasuga, A., 2013. Application of highly differentiated SNPs between Japanese Black and Holstein to a breed assignment test between Japanese Black and F1 (Japanese Black x Holstein) and Holstein. Anim. Sci. J. 84, 1-7.

Pant, S.D., Schenkel, F.S., Verschoor, C.P. & Karrow, N.A., 2012. Use of breed-specific single nucleotide polymorphisms to discriminate between Holstein and Jersey dairy cattle breeds. Anim. Biotechnol. 23, 1-10.

Ramos, A.M., Megens, H.J., Crooijmans, R.P., Schook, L.B. & Groenen, M.A., 2011. Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing. Anim. Genet. 42, 613-620.

Rust, J.M. & Rust, T., 2013. Climate change and livestock production: A review with emphasis on Africa. SA J. Anim. Sci. 43, 256-267.

Sabir, J., Mutwakil, M., El-Hanafy, A., Al-Hejin, A., Sadek, M.A., Abou-Alsoud, M., Qureshi, M., Saini, K. & Ahmed, M., 2014. Applying molecular tools for improving livestock performance: From DNA markers to next generation sequencing technologies. J. Food, Agric. & Environ. 12, I351-I363.

Scholtz, M.M., 2010. Beef breeding in South Africa (2nd ed.). Asikhulume pixArt, Rooihuiskraal, Pretoria, South Africa.

Shabalala, A.N. & Combrinck, W.L., 2012. Correlation of water quality with farming activities: hydrochemical characteristics of the Bonsma Dam, KwaZulu-Natal. Dams Technol. Paper. 1-4.

Stothard, P., Choi, J.W., Basu, U., Sumner-Thomson, J.M., Meng, Y., Liao, X. & Moore, S.S. (2011). Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. BMC Genomics. 12, 559, 1-14.

Webb, E.C., 2013. The ethics of meat production and quality - a South African perspective. S. Afr. J. Anim. Sci. 43, 1-9.

Wiedmann, R.T., Smith, T.P. & Nonneman, D.J., 2008. SNP discovery in swine by reduced representation and high throughput pyrosequencing. BMC Genet. 9, 81, 1-7.

Van Marle-Kőster, E., Visser, C., Makgahlela, M. & Cloete, S.W., 2015. Genomic technologies for food security: A review of challenges and opportunities in Southern Africa. Food Res. Int. 76, 971-979.

Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C. & Sonstegard, T.S., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Methods. 5, 247-252.

# CHAPTER TWO

## Literature Review

### 2.1. Introduction

Livestock production is one of the largest agricultural sectors around the world (Thornton, 2010). This development is mostly influenced by increasing demand for livestock products due to the significant growth of human population, urbanization and increasing incomes in developing countries (Delgado, 2005; Ilea, 2009; Thornton, 2010). Livestock products contribute about 33% to protein consumption and 17% to calorie consumption worldwide, with great variances between developed, developing and under-developed countries (Thornton, 2010). Globally, the demand for meat and meat products, eggs and milk is likely to rise by 30% in 2020 (Thornton, 2010; Webb, 2013). Global annual meat production is anticipated to twofold from 229 million tons in 1999-2001 to 465 million tons in 2050, with milk yield anticipated to almost twofold from 580 million tons to 1043 million tons. Most of this increase is predicted to occur in low or middle income countries (FAO, 2006). In developing countries, the increasing demand for meat is forcing an extension of intensive agricultural activities into the tropical rainforests, except in some of the African countries that are currently facing droughts and dry periods. Therefore, there are growing concerns, especially in developing countries, about the world's capability to provide meat in a fast growing human population (Thornton, 2010).

Currently, livestock habit nearly a third of the world’s entire land surface, with the majority of the land being permanent pasture, and the remaining arable land providing livestock feed (McMichael et al., 2007). It is estimated that about 35% of greenhouse-gas emissions from agriculture and land use globally emanates from livestock production. About 18% of greenhouse-gas emissions emanates from the deforestation of forage land, soy-feed production, accumulative feed-grains, soil carbon degradation, processing and transporting meat and grains, use of nitrogenous composts, methane gas from animal fertilizer and enteric maturation; and all contribute to global warming (McMichael et al., 2007; Stehfest et al., 2009). Therefore, production systems are under preasure and must be improved and adapted due to their potential contribution to climate change.

In SA, animal production is based on intensive and extensive production systems (Webb, 2013). Cattle production is the most important livestock sub-sector in SA and it contributes about 25-30% to the total agricultural output per annum (Musemwa et al., 2008). In 2010, the Department of Agriculture, Forestry and Fisheries (DAFF) reported that there were approximately 14.1 million cattle in SA, 60% of which were owned by commercial farmers and 40% by emerging and communal farmers (DAFF, 2010). Beef production raised from 512,000 tons in 2000 to over 750,000 tons in 2009, and this indicates a rise in production of about 46.6%. In total, the number of cattle slaughtered yearly raised slightly by about 7% (from 2.7 million to 2.9 million) and beef consumption raised by just over 20% from 671,000 to 815,000 tons per annum during this period. Subsequently, it is evident that SA livestock industry is unable to meet the demand for beef, due to a constant annual shortage of about 10% (DAFF, 2010). Therefore, breeding objectives must be adjusted to accommodate the meat demand of a growing SA population. The agricultural sector should source the scientific knowledge, expertise and technology to respond to these challenges. The aim of this literature review was to explore the challenges facing SA livestock production, with reference to beef production and how this can be addressed given the availability of new next NGS technologies.

### 2.2. Indigenous cattle in Africa

Since the domestication of livestock approximately 10,000 years ago, cattle have played an important role in human cultural and economic activities (Taberlet et al., 2011). Animals were tamed to produce milk, meat and skins, leather and hides, for draft purposes on farms ploughing and as transport for pulling wagons and carts, and also for other socio-economic functions (Musemwa et al., 2008). The establishment of the indigenous cattle breeds of Africa was closely associated with human development and migration (Strydom, 2008). About 180 cattle breeds have been recognised in southern Africa, of which, 150 breeds are indigenous and others are commercial composites or exotic (Rege, 1999; Mwai et al., 2015). However, the genetic differentiation between these cattle breeds, and their ecotypes remain essentially uncharacterized (Rege, 1999; Mwai et al, 2015).

Africa's indigenous cattle incorporate various crosses between Hamitic longhorn cattle (*Bos taurus*), zebu cattle (*Bos indicus*) and shorthorn cattle (Strydom, 2008). As the movement of man proceeded southward through Africa, new cattle breeds (zebu and sanga-types) were developed (Strydom, 2008). The zebu-type cattle include the Boran, Masai and Sokoto breeds,

while sanga-type (also known as *Bos taurus africanus*) includes the Afrikaner, Nguni, Pedi, Mashona and Tuli (Hanotte et al., 2000; Strydom, 2008). Sanga cattle were introduced to SA during the migration of the San and Sudanic Bantu tribes to southern Africa and the arrival of Europeans during the 15$^{th}$ century (Bachmann, 1983). *Bos taurus* are humpless and include two groups of humpless shorthorns and longhorns which are mainly found in Central and West Africa. *Bos indicus* are humped and they are major cattle types in Africa. *Bos indicus* cattle mostly originated from western and eastern parts of Africa, and the commercial taurine breeds, with their crossbreds, are found everywhere across the world. Their substantial body mass and greater production in tsetse-free areas have made these breeds more appealing to the local farmers, which somewhat explains the abundance of these breeds and wide distribution throughout Africa (Mwai et al., 2015). It has been noticed that there are no pure *Bos indicus* on the African continent because all cattle carry taurine mitochondrial DNA (Mwai et al., 2015).

During the migration of black tribes to southern African regions, Sanga cattle accompanied them and adapted to these regions with diverse environments. For example, Nguni cattle adapted in Kwazulu Natal, SA; Pedi cattle in Limpopo, SA; Nkone cattle in Zimbabwe and Zambia; while Tswana and Tuli cattle were adapted in Botswana (Schoeman, 1989). The Nguni breed is widely considered a beef breed with optimal production under harsh African conditions and is regarded as a mainstay of traditional Zulu culture. The breed is variously patterned with multi-coloured hides (Figure 1), is fertile, easy calving, has low maintenance requirements, and

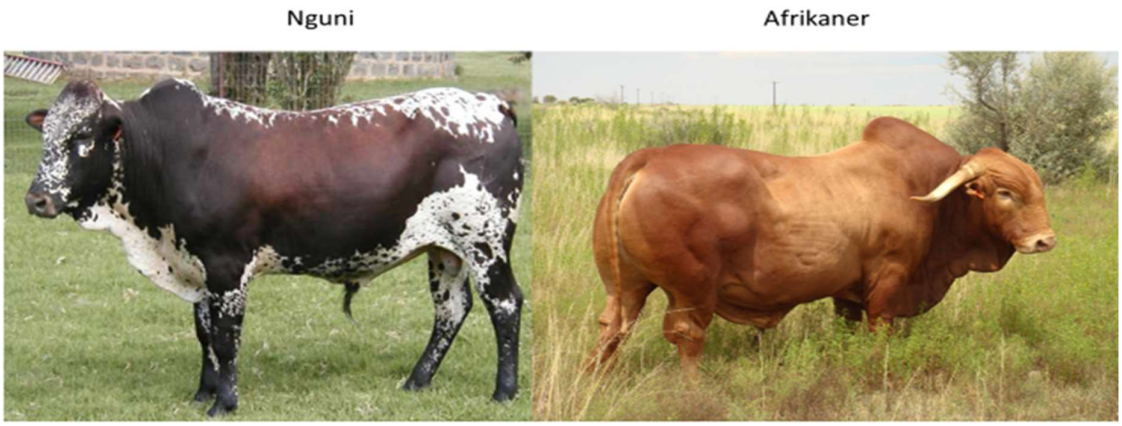


**Figure 1**: Nguni and Afrikaner cattle of South Africa. Pictures taken from zulucattle.com and pinterest.com websites, respectively

has a low susceptibility to parasites and tick-borne diseases (Scholtz, 1988; Mapholi et al., 2014). These characteristics have made Nguni widely used as a dam line and it has been promoted as a maternal breed (Rege, 2001).

Afrikaner cattle are usually deep red in colour with long spreading horns and can be found in various geographical areas in and around southern Africa (Figure 1). The breed was widely used for crossbreeding in extensive cattle production regions due to its adaptive characteristics (Van Marle, 1974) and was used in the development of the Bonsmara breed (a SA composite). The Afrikaner's adaptive traits are complementary to the growth performance, fertility, mothering ability and carcass quality of Hereford and Shorthorn that were used to develop the Bonsmara breed (Bachmann, 1983; Rege, 1999; Strydom, 2008). Tuli is an indigenous breed originated from Zimbabwe, and was derived from Tswana type cattle (Figure 2). Tuli was introduced to SA in the early 1940s, and is adapted to local conditions with a unique ability to utilise poor quality forages (Scholtz, 2010). The Drakensberger is also an indigenous SA breed with a history that is not well documented (Figure 2). The breed is also known for its adaptability, hardiness and resistance to tick-borne diseases and is often used in crossbreeding programs as a dam line (Scholtz, 2010).


**Figure 2**: Drakensberger and Tuli cattle of South Africa showing different phenotypic characteristics. Pictures taken from ultimatebeef.co.za.

Limited studies have been done to understand the genetic variation of indigenous SA breeds at the level of the genome. Recent studies have focussed on the use of microsatellite markers and available bovine SNP assays to determine the extent of genetic diversity among Nguni, Bonsmara, Drakensberger, and Afrikaner cattle (Pienaar, 2014; Makina et al., 2014; Sanarana

et al., 2015). Other studies include the identification of genes for tick resistance and copy number variation (CNV) in Nguni, as well as determining the extent of LD in SA cattle as compared to the European taurine breeds (Wang et al., 2015; Makina et al., 2015; Mapholi et al., 2015). These studies have provided a basis for understanding the genetic diversity and variation among these cattle breeds.

### 2.3. Genetic variation in cattle

Genetic markers have been widely studied to assess the genetic variability among animals, as they provide information from every region of the genome, regardless of the levels of gene expression (Barcaccia et al., 2013). Before SNP data became accessible, pedigree information was used to study the genetic variability and relatedness. However, the incomplete pedigree data limited the extent of these studies and resulted in low prediction accuracies of Estimated Breeding Values (EBVs) (Garcia-Ruiz et al., 2015). Microsatellite markers have been extensively investigated for parentage testing (Koskinen, 2003; Kathiravan et al., 2012) and some studies have explored their practical applications for tracing meat or dairy products at the breed level (Negrini et al., 2008). It is now possible to study the genetic composition of a population or breed using SNP data, without any previous knowledge of ancestry (Sölkner et al., 2010; Frkonja et al., 2012; Garcia-Ruiz et al., 2015).

Information about population structure can be used to study the histories of animals as well as to remove outliers and correct for stratification in GWAS (Negrini et al., 2008; Garcia-Ruiz et al., 2015). Population structure analysis explores whether there is any evidence that the samples are from a homogeneous population or whether they represent a population containing genetically distinct subgroups (Patterson et al., 2006). Breed composition on the other hand, provides information on the extent of crossbreeding as well as the genetic effects of heterosis and its diminution in advanced generations through recombination loss (Frkonja et al., 2012). Therefore, a comprehensive understanding of breed characteristics and genetic composition is required to facilitate their effective management (Sharma et al., 2015).

The extent of genetic diversity among populations has been studied using different analytical methods based on marker information. Principal component analysis (PCA) is done to define the breeds or populations' similarity and can be superimposed with geographic information (Lewis et al., 2011), while admixture analysis is used to define the genetic composition of

individuals within populations (Yonesaka et al., 2016). ). PCA is frequently used to understand the relationships among cattle breeds and it is also used to control for the effects of stratification on false positive discovery in genome-wide association studies of admixed populations (Bovine HapMap Consortium, 2009; Lewis et al., 2011). Admixture analysis is useful in exploring the extent of genetic variation within and between population groups based on genetic markers. It uses the DNA from multiple genetically distinct populations or breeds to study groups at both individual and population levels (Frkonja et al., 2012). An admixed individual' genome with ancestry arising from several distinct progenitor populations, indicate a mixture of chromosomal blocks, each following the statistics of variation in those populations (Sundquist, 2008). By assessing polymorphisms in the admixed individual, the ancestral origins of the individual's haploblocks can be inferred under the assumption that there were *K* ancestral populations (Sundquist, 2008). The size of these haploblocks will differ due to random nature of recombination, but on average, they will be shorter as the number of generations to the original crossbreeding event(s) increases (Garcia-Ruiz et al., 2015).

Persistence of chromosomal phase relationships allows the characterization of the extent of LD at constant genomic distances between populations (De Roos et al., 2008; Garcia-Ruiz et al., 2015). LD characterization is useful to determine if two or more populations can be analysed jointly in genomic studies. This is because markers that are in LD in one population may not be in LD in another population (De Roos et al., 2008). Therefore, in order to make significant statistical population inferences, will rely on the persistence of allele phase relationships among two populations (Lu et al., 2012).

Advances in identifying and genotyping variants now influence a more detailed understanding of the global patterning of genetic variation (Kim et al., 2006). However, for populations that were not represented in the SNP discovery that led to the design of the current assays, it is unclear whether one or a couple of haplotype maps can provide useful information (Kim et al., 2006). Studies proposed that LD characterised in a small subset of populations needs to be expanded to include information about other populations (De Roos et al., 2008). This information is needed to determine the LD patterns and the ancestral origins of genetic variation present in indigenous SA cattle populations, and this will enable mapping of complex traits, including adaptation and disease susceptibility (Campbell & Tishkoff, 2008). Therefore, sequencing of local breeds is essential in order to discover more SNPs that can be included into the existing SNP assays, to increase their density so that they can be compatible for use world-

wide. Sequencing of local populations will not only help identify new SNPs, but will also assist in understanding the composition of the local SA breeds, their genetic variability as well as allow an increase in understanding of the extent of LD among markers assayed in these breeds.

### 2.4. Use of DNA technology in genomic selection

There is currently a potential to use genome-wide markers in animal breeding. Genomic selection provides a more accurate estimation of breeding values earlier in the life of breeding animals, giving more accurate selection and allows lower generation intervals (van der Werf, 2013). In domestic animals, for centuries, artificial selection has been based on phenotypic characteristics of the animals, using selection index theory, and Best Linear Unbiased Prediction (BLUP), which rely on mixed linear models. These methods allowed the use of phenotypes of related individuals to estimate the breeding values for selection (Boichard er al., 2016). The methods were successful in the selection of easy phenotypic traits such as coat colour, body mass, milk yield and other important traits with moderate or high heritability, but were not suitable for complex traits. The mapping of quantitative trait loci (QTLs) using genetic markers have paved a way to marker-assisted selection (MAS), but also with limitatons in identifying complex traits (Schuster, 2011).

South African ndigenous breeds such as Afrikaner, Drakensberger and Nguni have made major contributions to livestock production because of their ability to adapt and produce in different production systems (Abin et al., 2016). These breeds have been participating in animal recording programmes and have an average complete pedigree recording in the first generation varying from 88.5% for the Nguni to 92.5% for the Afrikaner (Abin et al., 2016). The availability of the pedigree records have been essential for genetic evaluation using BLUP model in determining the selection efficiency and actual genetic change (Mostert, 2007; Groeneveld et al., 2009). However, crossbreeding and inbreeding within cattle breeds has been reported to have negative effects on production and fitness traits in beef and dairy cattle (Nazokkarmaher, 2016), and have contributed to loss of diversity in most cattle populations (Pinaar et al., 2014). The use of a small number of selected genotypes increases the chance of having undesirable recessive genes within a population, which may result in inbreeding depression in the near future (Abin et al., 2016). Quantitative breeding methods such as artificial insemination has resulted in more intense selection pressure on a number of traits of economic importance, which could have contributed to an increase in production efficiency.

Therefore, maintaining within-breed genetic diversity is essential for selection (Oltenacu & Broom, 2010).

With the advert of molecular technology, genomic selection can now be accurately estimated using DNA of an individual. Genomic selection has been used in dairy cattle, and has caused a paradigm shift in dairy cattle breeding programs (Harris & Johnson, 2010; Boichard et al., 2016). Genomic selection estimates a prediction equation in a reference population with genotype and phenotype data. This prediction equation can then be used to predict the breeding values in animals without phenotype data, and EBVs can be accurately calculated before sexual maturity. This means that breeders can identify superior animals at earlier age (Schefers et al. 2012; Koopaee and Koshkoiyeh, 2014). The advancement in genotyping high-density SNP chips and the associated reduction in the cost have resulted in large numbers of individuals with genome-wide genotypic data. A large number of candidates can be screened, and selection intensity can be increased. This large-scale screening allows a better use of the available genetic resources (Bassi et al., 2016). The evaluation can be carried out for any trait of interest, including complex traits, such as sex-limited, meat quality, and also disease resistance traits (Boichard et al., 2016). Thus, genomic selection can be widely implemented in farm animals once the accuracy of genomic selection is sufficient (Goddard 2012).

Genomic data also allow improved inbreeding estimates and characterize relationships based on the specific regions of the genome, which can be used to effectively manage areas of low genetic diversity or areas of reduced performance across economically important traits (Biscarini et al., 2015; Howard et al., 2017). The use of these region-specific metrics should allow breeders to efficiently manage the genetic value of the progeny and undesired side effects associated with inbreeding. However, methods to identify regions affected by inbreeding and reltated methods to manage the genome at the herd level, still need to be developed (Howard et al., 2017).

### 2.5. Discovery of SNP markers

Whole genome sequencing has become one of the most important and effective methods for exploring the genetic information present among species (Mei et al., 2016). The discovery of genetic variants identifies markers with power to address various research questions (Imelfort et al., 2009; Le Roex et al., 2012). Sequencing is the process of determining the precise identity

of nucleotides within a deoxyribonucleic acid (DNA) or ribonucleic acid (RNA) molecule (Kumar et al., 2012). Since its inception in 1977, sequencing has enhanced the field of genomics and has increased understanding of the structure and composition of animal genomes (Su et al., 2011, Kumar et al., 2012). Advances in sequencing technologies have led to the generation of large amounts of genomic data in a very short period, and these developments have assembled large catalogs of genetic variation in livestock species (Kumar et al., 2012).

Since the accomplishment of the human genome and the HapMap projects, DNA sequencing and resequencing of members of several livestock species has been accomplished. This includes chicken (Groenen et al., 2000), cow (Bovine HapMap Consortium, 2009), pigs (Ramos et al., 2009), horses (Wade et al., 2009), turkey (Archibald et al., 2010), sheep (Aslam et al., 2012), goat (Tosser-Kloop et al., 2014), as well as other domestic animals such as dogs and cats (Lindblad-Toh et al., 2005; Pontius et al., 2007). Whole genome sequencing strategies for assembling the genomes for most of these animals were directly taken from the human genome sequencing project and combine whole-genome shotgun and BAC-to-BAC sequencing (Fan et al., 2010). However, due to the rapid development of the NGS technologies, these strategies have been modified for the generation of reference genomes for different species (Fan et al., 2010).

The first bovine genome was sequenced by the Baylor College of Medicine using a combination of whole genome shotgun sequencing as well as BAC-to-BAC sequencing (Bovine HapMap Consortium, 2009). Another genome assembly based on the same sequence data was subsequently released by the University of Maryland and this assembly was annotated (Zimin et al., 2009). The two assemblies vary based on the approaches used to assemble the sequence reads, and the availability of these assemblies has provided a valuable resource for genomic studies in beef cattle (Rolf et al., 2010). Through the development of sequencing methods, several breeds have also been sequenced to discover millions of SNPs, e.g. *Bos taurus* breeds such as Hereford, Angus, Hanwoo, Yanbian, and Japanese native cattle, *Bos indicus* breeds such as Brahman, AFR, Gir, Tuli and Nellore, dairy cattle (Holstein, Fleckvieh), and a number of other breeds (Van Tassell et al., 2008; Matukamalli et al., 2009; Stothard et al., 2011; Barris et al., 2012; Canavez et al., 2012; Choi et al., 2013; Choi et al., 2015). These data have provided basis for genetic analysis of complex traits in cattle (Le Roex et al., 2012), as well as providing comprehensive data for relative genomics studies on the function and evolution of important genomic regions and genes (Fan et al., 2010).

Sequencing the genomes of individual animals using NGS methods led to the development of SNP assays that have been widely used worldwide in cattle (Fan et al., 2010). The first high-density and high-throughput genotyping assay (10K) was developed in 2005 (The Bovine HapMap Consortium, 2009), and was commercialised by Affymetrix. However, the SNP' density on this assay was considered inadequate for genomic studies such as genomic selection (GS) and GWAS, and there was a need for a higher density assay. Later, a consortium of animal scientists, using SNP discovery data from Holstein, Angus and other beef cattle breeds, developed the BovineSNP50 assay (Van Tassell et al., 2008) and this SNP assay was commercialised by Illumina early in 2008. The assay provided much higher density (~50,000 SNPs per animal) compared to previous assays (Matukumalli et al., 2009) and has been globally recognised as the standard for population studies, GWAS and GS in cattle.

The Illumina BovineSNP50 assay has proven to be more adequate for different genomic studies, however, higher density assays were developed later to enable building models for GS with utility across the breeds. These assays have played important roles in a broad range of genetic studies, e.g., population studies in cattle (Wilkinson et al., 2011, Edea et al., 2013; Makina et al., 2014, Decker at al., 2014), checking the number of markers needed to form a genomic relationship matrix (Rolf et al., 2010), and resolving the evolutionary relationships among horned ruminants (Decker et al., 2009; MacEachern et al., 2009). Other applications include predictions for GS (Hayes et al., 2009), GWAS (Hayes et al., 2010; Bolormaa et al., 2011), whole-genome LD patterns (Qanbari et al., 2010) and CNV detection (Hou et al., 2011).

### 2.6. DNA sequencing methods

Over the past few years, there has been a major shift from the application of automated Sanger sequencing (first-generation sequencing) for genome analysis. Sanger sequencing is a DNA sequencing method based on the chain terminating of di-deoxynucleotides selectively incorporated by DNA polymerase during *in vitro* DNA replication (Sanger and Coulson, 1975, Jain et al., 2013). This method was the most widely used sequencing method for approximately two decades and has contributed to a number of large scale projects including the completion of the human genome sequence (Chial, 2008; Barba et al., 2014). The limitations of Sanger sequencing showed a need for improved technologies for sequencing large numbers of human and animal genomes (Hert et al., 2008; Barba et al., 2014), and more recently, Sanger sequencing has been replaced by NGS methods (Morozova & Marra, 2008). However, the

Sanger method remains widely used for small-scale projects, validation of NGS results and for locating long contiguous DNA sequence reads (> 500 nucleotides) (Metzker, 2010).

After Sanger sequencing was discovered, Applied Biosystems (ABI) introduced the first automatic sequencing machine, the ABI 370A in 1986, adopting the capillary electrophoresis that made sequencing faster and more precise (Collins et al., 2003; Liu et al., 2012). These successes greatly influenced the development of powerful novel sequencing instruments to increase speed and accuracy, while reducing cost and labour at the same time (Liu et al., 2012). Due to the low throughput and high cost of the first-generation methods, the second-generation sequencing (SGS) was introduced in 2005, i.e., Illumina second-generation sequencing (Schadt et al., 2010; Indap et al., 2013). These tools were highly recognised due to sequencing a large number of DNA molecules in parallel (sequencing hundreds of gigabases in a single run) at a reasonable cost (Schadt et al., 2010; Indap et al., 2013). Subsequently, the single-molecule sequencing (SMS) technologies emerged (third/next-generation sequencing) which offer advantages over the first and SGS technologies (Schadt et al., 2010; Heather & Chain, 2016).

The NGS technologies were designed to achieve higher throughput capacity, faster turnaround time and longer read lengths to enhance *de novo* assembly (Schadt et al., 2010; Pareek et al., 2011; Lee et al., 2013). They also enabled the direct detection of haplotypes and rare variants with small amounts of DNA and at low cost (Schadt et al., 2010). NGS can be used for sequencing whole genomes or can be targeted to specific region of interest, e.g., all coding genes (a whole exome) or individual genes (Behjati & Tarpey, 2013). The methods can capture a broader spectrum of mutations and can interrogate the genome without bias. Different platforms have been used to detect genetic variants using different sequencing technologies (Imelfort et al., 2009; Behjati & Tarpey, 2013), this includes ABI SOLiD, Illumina Genome Analyzer, Illumina HiSeq and Roche 454 platforms. These platforms provide fast and more cost-effective approaches for generating sequence data and discovery of genetic markers (Voelkerding, 2009; Le Roex et al., 2012). Each platform has its own specific chemistry for template preparation, sequencing and data analysis, and therefore, has its own advantages and disadvantages (Le Roex et al., 2012).

## 2.7. Variant detection and the use of SNP assays

One of the objectives of livestock genomic research is to detect the genetic variations responsible for difference in phenotypic traits, particularly economic important traits. Characterization of these genetic variants is important for linking genomic regions or genes to phenotypes (Stothard et al., 2011). A large number of genetic variants, especially SNPs, have been discovered in livestock species and deposited in publicly accessible databases, of which, most of the data came from the 1000 Bull Genomes Project. This facilitated mapping of monogenic and complex traits in cattle (Daetwyler et al., 2014; Iso-Touru et al., 2016). Some of the data comes from the bovine HapMap project (The Bovine HapMap Consortium, 2009), the Bovine Genome Project (Elsik et al., 2009), and large-scale SNP discovery projects (Van Tassell et al., 2008). Nonetheless, there is still much genetic variation that need to be discovered (Stothard et al., 2011).

Comparative studies between the genomes of the domestic animals and human have shown a high level of conservation and orthology for protein coding genes. However, more variations have been observed in non-coding regions, especially the intergenic repetitive regions, known to be one of the major forces driving evolution (Fan et al., 2010). The HapMap studies also discovered abundant genetic variations within and between domestic breeds. Most of these variations were discovered by large-scale genotyping of SNPs and insertions or deletions (Indels) of DNA fragments with different sizes, such as CNVs, which is predicted to partly contribute to the phenotypic variation of domestic animals (Fan et al., 2010).

A large number of genetic variants that were discovered within the genomes of livestock animals were SNPs that required validation due to sequencing errors (Fan et al., 2010; Eynard et al., 2016). Candidate SNPs for assay design were validated and SNPs with a high MAF in the sequenced populations were selected (i.e., Euopean taurine breeds). Consequently, the majority of developed SNP chips targeted SNPs that had approximately uniformly distributed allele frequencies, and from a subset of economically important European taurine populations (Eynard et al., 2016). This led to an ascertainment bias in the initial selection of markers since there was an over-representation of polymorphisms with high MAF and under-representation of polymorphisms with low MAF, which may affect inferences about populations (McTavish & Hillis, 2015). Since the SNPs were selected from certain subpopulations and geographic regions, these factors have influenced variability in ascertainment bias (Heslot et al., 2013).

Thus, sample sizes and the populations in which SNPs were discovered (e.g., European taurine cattle breeds) have affected the characteristics of sampled genetic variants. Therefore, the assay causes the allele frequency distributions to be shifted towards lower frequency alleles, and the LD to be reduced in breeds that are more distantly related to the breeds in which the SNPs were discovered (Porto Neto et al., 2013; Matimba et al., 2009).

The current bovine SNP genotyping assays have successfully been used in a population study in five South African cattle breeds (Makina et al., 2014). Limitations were, however, found for the detection of selection signatures in indigenous SA cattle (Makina et al., 2015) due to SNP ascertainment bias, and small sample sizes. These assays seem to be more appropriate for studies of European taurine breeds, and contain SNPs that are less informative in the local SA breeds. The analysis of the available genotypes for indigenous SA cattle tended to have significantly lower MAF and LD levels as compared to European taurine breeds (Makina et al., 2013) and generally failed to reveal common variation within indigenous SA breeds (Matukumalli et al., 2009). This will also affect the utility of GS in local breeds, detection of QTLs, and identification of genes associated with economically important traits in local breeds (Pool et al., 2010).

### 2.8. Conclusion

The NGS technologies provide a platform for the discovery of substantial numbers of SNPs in indigenous breeds. Sequencing of genomes has the potential to reveal millions of SNPs that may contribute to explaining the genetic composition and relationships among breeds. With the increased efficiency of sequencing using NGS technologies, genes and chromosomal regions that create phenotypes in response to the environmental factors may be identified. The question of how animals respond to different environmental models (e.g., nutrition) at the molecular and cellular levels could also be addressed. It is clear that NGS technologies will assist animal scientists to efficiently raise animals and to improve them for long-term sustainable livestock production, including reducing the susceptibility to infectious diseases.

### References

Abin, S., Theron, H.E. & van Marle-Köster, E., 2016, Population structure and genetic trends for indigenous African beef cattle breeds in South Africa. S. Afr. J. Anim. Sci. 46, 152-156.

Archibald, A.L., Cockett, N.E., Dalrymple, B.P., Faraut, T., Kijas, J.W., Maddox, J.F., McEwan, J.C., Hutton Oddy, V., Raadsma, H.W., Wade, C., Wang, J., Wang, W. & Xun, X., 2010. The sheep genome reference sequence: a work in progress. Anim. Genet. 41, 449-453.

Aslam, M.L., Bastiaansen, J.W., Elferink, M.G., Megens, H.J., Crooijmans, R.P., Blomberg, L.A., Fleischer, R.C., Van Tassell, C.P., Sonstegard, T.S., Schroeder, S.G., Groenen, M.A. & Long, JA., 2012. Whole genome SNP discovery and analysis of genetic diversity in Turkey (*Meleagris gallopavo*). BMC Genomics. 13, 391, 1-14.

Bachmann, M., 1983. Early origins of cattle. Farmer's Weekly, 23, December, pp 18.

Barcaccia, G., Felicetti, M., Galla, G., Capomaccio, S., Cappelli, K., Albertini, E., Buttazzoni, L., Pieramati, C., Silvestrelli, M. & Supplizi, A.V., 2013. Molecular analysis of genetic diversity, population structure and inbreeding level of the Italian Lipizzan horse. Livest. Sci. 151, 124-133.

Barba, M., Miquel, C., Boyer, F., Mercier, C., Rioux, D., Coissac, E. & Taberlet, P., 2014. DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. Mol. Ecol. Resour. 14, 306-323.

Barris, W., Harrison, B.E., McWilliam, S., Bunch, R.J., Goddard, M.E. & Barendse, W., 2012. Next generation sequencing of African and Indicine cattle to identify single nucleotide polymorphisms. Anim. Prod. 52, 133-142.

Bassi, F.M., Bentley, A.R., Charmet, G., Ortiz, R. & Crossa, J., 2016. Breeding schemes for the implementation of genomic selection in wheat (Triticum spp.). Plant Sci. 242, 23-36.

Behjati, S. & Tarpey, P.S., 2013. What is next generation sequencing? Arch. Dis. Child. Educ. Pract. Ed. 98, 236-238.

Biscarini, F., Nicolazzi, E.L., Stella, A., Boettcher, P.J., & Gandini, G., 2015, Challenges and opportunities in genetic improvement of local livestock breeds. Front. Genet. 6, 1-7.

Boichard, D., Ducrocq, V., Croiseau, P. & Fritz, S., 2016. Genomic selection in domestic animals: Principles, applications and perspectives. C. R. Biol. 339, 274-277.

Bolormaa, S., Neto, L. R., Zhang, Y. D., Bunch, R. J., Harrison, B. E., Goddard, M. E. & Barendse, W., 2011. A genome-wide association study of meat and carcass traits in Australian cattle. J. Anim. Sci. 89, 2297-2309.

Bovine Genome Sequencing and Analysis Consortium, Elsik, C,G., Tellam, R.L., Worley, K.C., Gibbs, R.A., Muzny, D.M., Weinstock, G.M., Adelson, D,L., Eichler, E.E., Elnitski, L., Guigó, R. Hamernik, D.L., Kappes, S.M., Lewin, H.A., Lynn, D.J., Nicholas, F.W., Reymond, A., Rijnkels, M., Skow, L.C., Zdobnov, E.M., Schook, L., Womack, J., Alioto, T., Antonarakis, S.E., Astashyn, A., Chapple, C.E., Chen, H.C., Chrast, J., Câmara, F., Ermolaeva, O., Henrichsen, C.N., Hlavina, W., Kapustin, Y., Kiryutin, B., Kitts, P., Kokocinski, F., Landrum, M., Maglott, D., Pruitt, K., Sapojnikov, V., Searle, S.M., Solovyev, V., Souvorov, A., Ucla, C., Wyss, C., Anzola, J.M., Gerlach, D., Elhaik, E., Graur, D., Reese, J.T., Edgar, R.C., McEwan, J.C., Payne, G.M., Raison, J.M., Junier, T., Kriventseva, E.V., Eyras, E., Plass, M., Donthu, R., Larkin, D.M., Reecy, J., Yang, M.Q., Chen, L., Cheng, Z., Chitko-McKown, C.G., Liu, G.E., Matukumalli, L.K., Song, J., Zhu, B., Bradley, D.G., Brinkman, F.S., Lau, L.P., Whiteside, M.D., Walker, A., Wheeler, T.T., Casey, T., German, J.B., Lemay, D.G., Maqbool, N.J., Molenaar, A.J., Seo, S., Stothard, P., Baldwin, C.L., Baxter, R., Brinkmeyer-Langford, C.L., Brown, W.C., Childers, C.P., Connelley, T., Ellis, S.A., Fritz, K., Glass, E.J., Herzig, C.T., Iivanainen, A., Lahmers, K.K., Bennett. A.K., Dickens, C.M., Gilbert, J.G., Hagen, D.E., Salih, H., Aerts, J., Caetano, A.R., Dalrymple, B., Garcia, J.F., Gill, C.A., Hiendleder, S.G., Memili, E., Spurlock, D., Williams, J.L., Alexander, L., Brownstein, M.J, Guan, L., Holt, R.A., Jones, S.J., Marra, M.A., Moore, R., Moore, S.S., Roberts, A., Taniguchi, M., Waterman, R.C., Chacko, J., Chandrabose, M.M., Cree, A., Dao, M.D., Dinh, H.H., Gabisi, R.A., Hines, S., Hume, J., Jhangiani, S.N., Joshi, V., Kovar, C.L., Lewis, L.R., Liu, Y.S., Lopez, J., Morgan, M.B., Nguyen, N.B., Okwuonu, G.O., Ruiz, S.J., Santibanez, J., Wright, R.A., Buhay, C., Ding, Y., Dugan-Rocha, S., Herdandez, J., Holder, M., Sabo, A., Egan, A., Goodell, J., Wilczek-Boney, K., Fowler, G.R., Hitchens, M.E., Lozado, R.J., Moen, C., Steffen, D., Warren, J.T., Zhang, J., Chiu, R., Schein, J.E., Durbin, K.J., Havlak, P., Jiang, H., Liu, Y., Qin, X., Ren, Y., Shen, Y., Song, H., Bell, S.N., Davis, C., Johnson, A.J., Lee, S., Nazareth, L.V., Patel, B.M., Pu, L.L., Vattathil, S., Williams, R.L. Jr., Curry, S., Hamilton, C., Sodergren, E., Wheeler, D.A., Barris, W., Bennett, G.L., Eggen, A., Green, R.D., Harhay, G.P., Hobbs, M., Jann, O., Keele, J.W., Kent, M.P., Lien, S., McKay, S.D., McWilliam, S., Ratnakumar, A., Schnabel, R.D., Smith, T., Snelling, W.M., Sonstegard, T.S., Stone, R.T., Sugimoto, Y., Takasuga, A., Taylor, J.F., Van Tassell, C.P., Macneil, M.D., Abatepaulo, A.R., Abbey, C.A., Ahola, V., Almeida, I.G., Amadio, A.F., Anatriello, E., Bahadue, S.M., Biase, F.H., Boldt, C.R., Carroll, J.A., Carvalho, W.A., Cervelatti, E.P., Chacko, E., Chapin, J.E., Cheng, Y., Choi, J., Colley, A.J., de Campos, T.A., De Donato, M., Santos, I.K., de Oliveira, C.J., Deobald, H., Devinoy, E., Donohue, K.E., Dovc, P., Eberlein, A., Fitzsimmons, C.J., Franzin, A.M., Garcia, G.R.,

Genini, S., Gladney, C.J., Grant, J.R., Greaser, M.L., Green, J.A., Hadsell, D.L., Hakimov, H.A., Halgren, R., Harrow, J.L., Hart, E.A., Hastings, N., Hernandez, M., Hu, Z.L., Ingham, A., Iso-Touru, T., Jamis, C., Jensen, K., Kapetis, D., Kerr, T., Khalil, S.S., Khatib, H., Kolbehdari, D., Kumar, C.G., Kumar, D., Leach, R., Lee, J.C., Li, C., Logan, K.M., Malinverni, R., Marques, E., Martin, W.F., Martins, N.F., Maruyama, S.R., Mazza, R., McLean, K.L., Medrano, J.F., Moreno, B.T., Moré, D.D., Muntean, C.T., Nandakumar, H.P., Nogueira, M.F., Olsaker, I., Pant, S.D., Panzitta, F., Pastor, R.C., Poli, M.A., Poslusny, N., Rachagani, S., Ranganathan, S., Razpet, A., Riggs, P.K., Rincon, G., Rodriguez-Osorio, N., Rodriguez-Zas, S.L., Romero, N.E., Rosenwald, A., Sando, L., Schmutz, S.M., Shen, L., Sherman, L., Southey, B.R., Lutzow, Y.S., Sweedler, J.V., Tammen, I., Telugu, B.P., Urbanski, J.M., Utsunomiya, Y.T., Verschoor, C.P., Waardenberg, A.J., Wang, Z., Ward, R., Weikard, R., Welsh, T.H. Jr., White, S.N., Wilming, L.G., Wunderlich, K.R., Yang, J. & Zhao, F,Q., 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. 324, 522-528.

Bovine HapMap Consortium., 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science. 324, 528-532.

Campbell, M.C. & Tishkoff, S.A., 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Ann. Rev. Genom. Hum. Genet. 9, 403-433.

Canavez, F.C., Luche, D.D., Stothard, P., Leite, K.R., Sousa-Canavez, J.M., Plastow, G., Meidanis, J., Souza, M.A., Feijao, P., Moore, S.S. & Camara-Lopes, L.H., 2012. Genome sequence and assembly of *Bos indicus*. J. Hered. 103, 342-348.

Chial, H., 2008. DNA sequencing technologies key to the Human Genome Project. Nat. Educ. 1, 219,

Choi, J.W., Choi, B.H., Lee, S.H., Lee, S.S., Kim, H.C., Yu, D., Chung, W.H., Lee, K.T., Chai, H.H., Cho, Y.M. & Lim, D., 2015. Whole-genome resequencing analysis of Hanwoo and Yanbian cattle to identify genome-wide SNPs and signatures of selection. Mol. Cells. 38, 466-473.

Choi, J.W., Liao, X., Park, S., Jeon, H.J., Chung, W.H., Stothard, P., Park, Y.S., Lee, J.K., Lee, K.T., Kim, S.H., Oh, J.D., Kim, N., Kim, T.H., Lee, H.K. & Lee S.J., 2013. Massively parallel

sequencing of Chikso (Korean brindle cattle) to discover genome-wide SNPs and Indels. Mol. Cells. 36, 203-211.

Collins, F.S., Morgan, M. & Patrinos, A., 2003. The Human Genome Project: lessons from large-scale biology. Science. 300, 286-290.

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brøndum, R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C., Esquerre, D., Bouchez, O., Rossignol, M.N., Klopp, C., Rocha, D., Fritz, S., Eggen, A., Bowman, P.J., Coote, D., Chamberlain, A.J., Anderson, C., VanTassell, C.P., Hulsegge, I., Goddard, M.E., Guldbrandtsen, B., Lund, M.S., Veerkamp, R.F., Boichard, D.A., Fries, R. & Hayes, B.J., 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46, 858-865.

DAFF., 2010. News Letter: National livestock numbers. February 2010. Directorate: Agricultural Statistics, Department of Agriculture, Forestry and Fisheries, Pretoria.

De Roos, A.P.W., Hayes, B.J., Spelman, R.J. & Goddard, M.E., 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics. 179, 1503-1512.

Decker, J.E., McKay, S.D., Rolf, M.M., Kim, J., Alcala, A.M., Sonstegard, T.S., Hanotte, O., Gotherstrom, A., Seabury, C.M., Praharani, L. Babar, M.E., de Almeida Regitano, L.C., Yildiz, M.A.,. Heaton, M.P., Liu, W., Lei, C., Reecy, J.M., Saif-Ur-Rehman, M., Schnabel, R.D. & Taylor, J.F., 2014. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. PLoS Genet. 10, e1004254, 1-14.

Decker, J.E., Pires, J.C., Conant, G.C., McKay, S.D., Heaton, M.P., Chen, K., Cooper, A., Vilkki, J., Seabury, C.M., Caetano, A.R., Johnson, G.S., Brenneman, R.A., Hanotte, O., Eggert, L.S., Wiener, P., Kim, J., Kim, K.S., Sonstegard, T.S., Van Tassell, C.P.,. Neibergs, H.L., McEwan, J.C., Brauning, R., Coutinho, L.L., Babar, M.E., Wilson, G.A., McClure, M.C., Rolf, M.M., Kim, J., Schnabel, R.D. & Taylor, J.F., 2009. Resolving the evolution of extant and extinct ruminants with high-throughput phylogenomics. Proc. Natl. Acad. Sci. 106, 18644-18649.

Delgado, C.L., 2005. Rising demand for meat and milk in developing countries: implications for grasslands-based livestock production. IN McGilloway, D.A. (ed.). Grassland: a global

resource. Proceedings of the twentieth International Grassland Congress, Dublin, Ireland, 26-30 June 2005, pp 29-39.

Edea, Z., Dadi, H., Kim, S.W., Dessie, T., Lee, T., Kim, H., Kim, J.J. & Kim, K.S., 2013. Genetic diversity, population structure and relationships in indigenous cattle populations of Ethiopia and Korean Hanwoo breeds using SNP markers. Front. Genet. 4, 1-9.

Eynard, S.E., Windig, J.J., Leroy, G., Van Binsbergen, R. & Calus, M.P., 2015. The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genet. 16, 1-12.

Fan, B., Du, Z.Q., Gorbach, D.M. & Rothschild, M.F., 2010. Development and application of high-density SNP arrays in genomic studies of domestic animals. Asian-Australasian J. Anim. Sci. 23, 833-847.

FAO., 2006. Livestock a major threat to environment. pp 1-2.

Frkonja A., Gredler B., Schnyder U., Curik I. & Sölkner J., 2012. Prediction of breed composition in an admixed cattle population. Anim. Genet. 43, 696-703.

Garcia-Ruiz, A., Ruiz-Lopez, F.D.J., Van Tassell, C.P., Montaldo, H.H. & Huson, H.J., 2015. Genetic differentiation of Mexican Holstein cattle and its relationship with Canadian and US Holsteins. Front. Genet. 6, 7, 1-7.

Goddard, M.E., 2012. Uses of genomics in livestock agriculture. Anim. Prod. Sci. 52, 73-77.

Groenen, M.A.M., Cheng, H.H., Bumstead, N., Benkel, B.F., Briles, W.E., Burke, T., Burt, D.W., Crittenden, L.B., Dodgson, G., Jossi Hillel, J., Lamont, S., Ponce de Leon, A., Soller, M., Takahashi, H. & Vignal, A., 2000. A consensus linkage map of the chicken genome. Genome Res. 10, 137-147.

Groeneveld, E., Van der Westhuizen, B., Maiwashe, A., Voordewind, F. & Ferraz, J.B.S., 2009. POPREP: A generic report for population management. Genet. Mol. Res. 8, 1158-1178.

Hanotte, O., Tawah, C.L., Bradley, D.G., Okomo, M., Verjee, Y., Ochieng, J. & Rege, J.E.O., 2000. Geographic distribution and frequency of a taurine *Bos taurus* and an indicine *Bos indicus* Y specific allele amongst sub-Saharan African cattle breeds. Mol. Ecol. 9, 387-396.

Harris, B.L. & Johnson, D.L., 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. J. Dairy Sci. 93, 1243-1252.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. & Goddard, M.E., 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92, 433-443.

Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J. & Goddard, M.E., 2010. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. PLoS Genet. 6, e1001139, 1-11.

Heather, J.M. & Chain, B., 2016. The sequence of sequencers: The history of sequencing DNA. Genomics, 107, 1-8.

Hert, D.G., Fredlake, C.P. & Barron, A.E., 2008. Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods. Electrophoresis. 29, 4618-4626.

Heslot, N., Rutkoski, J., Poland, J., Jannink, J.L. & Sorrells, M.E., 2013. Impact of marker ascertainment bias on genomic selection accuracy and estimates of genetic diversity. PLoS One. 8, e74612, 1-8.

Hou, Y., Liu, G.E., Bickhart, D.M., Cardone, M.F., Wang, K., Kim, E.S., Matukumalli L.K., Ventura M., Song J., VanRaden P.M. & Van Tassell, C. P., 2011. Genomic characteristics of cattle copy number variations. BMC Genomics. 12, 127, 1-11.

Howard, J.T., Pryce, J.E., Baes, C. & Maltecca, C. 2017. Invited review: Inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genomic variability. J. Dairy Sci. 100, 6009-6024.

Ilea, R.C., 2009. Intensive livestock farming: Global trends, increased environmental concerns, and ethical solutions. J. Agr. Environ. Ethic. 22, 153-167.

Imelfort, M., Duran, C., Batley, J. & Edwards, D., 2009. Discovering genetic polymorphisms in next-generation sequencing data. Plant Biotech. J. 7, 312-317.

Indap, A.R., Cole, R., Runge, C.L., Marth, G.T. & Olivier, M., 2013. Variant discovery in targeted resequencing using whole genome amplified DNA. BMC Genomics. 14, 468, 1-13.

Iso-Touru, T., Sahana, G., Guldbrandtsen, B., Lund, M.S. & Vilkki, J., 2016. Genome-wide association analysis of milk yield traits in Nordic Red Cattle using imputed whole genome sequence variants. BMC Genet. 17, 55, 1-12.

Jain, S., Chaudhary, H. & Bhatnagar, V., 2013. An information security-based literature survey and classification framework of data storage in DNA. IJNVO. 13, 176-201.

Kathiravan, P., Kataria, R.S. and Mishra, B.P., 2012. Power of exclusion of 19 microsatellite markers for parentage testing in river buffalo (*Bubalus bubalis*). Mol. Biol. Rep. 39, 8217-8223.

Kim, K.J., Lee, H., Park, M., Cha, S., Kim, K., Kim, H., Kimm, K., Oh, B. & Lee, J., 2006. SNP identification, linkage disequilibrium, and haplotype analysis for a 200-kb genomic region in a Korean population. Genomics. 88, 535-540.

Koopaee, H.K. & Koshkoiyeh, A.E., 2014. SNPs genotyping technologies and their applications in farm animals breeding programs: Review. Braz. Arch. Biol. Technol. 57, 87-95.

Koskinen M.T., 2003. Individual assignment using microsatellite DNA reveals unambiguous breed identification in the domestic dog. Anim. Genet. 34, 297-301.

Kumar, S., Banks, T.W. & Cloutier, S., 2012. SNP discovery through next-generation sequencing and its applications. Int. J. Plant Genomics. 2012, 1-16.

Le Roex, N., Noyes, H., Brass, A., Bradley, D.G., Kemp, S.J., Kay, S., Van Helden, P.D. & Hoal, E.G., 2012. Novel SNP discovery in African buffalo, *Syncerus caffer*, using high-throughput sequencing. PloS One. 7, e48792, 1-6.

Lewis, J., Abas, Z., Dadousis, C., Lykidis, D., Paschou, P. & Drineas, P., 2011. Tracing cattle breeds with principal components analysis ancestry informative SNPs. PLoS One. 6, e18007, 1-8.

Lindblad-Toh, K., Wade, C.M., Mikkelsen, T.S., Karlsson, E.K., Jaffe, D.B., Kamal, M., Clamp, M., Chang, J.L., Kulbokas, E.J., Zody, M.C., Mauceli, E., Xie, X., Breen, M., Wayne, R.K., Ostrander, E.A., Ponting, C.P., Galibert, F., Smith, D.R., deJong, P.J., Kirkness, E., Alvarez, P., Biagi, P., Brockman, W., Butler, J., Chin, C., Cook, A., Cuff, J., Daly, M.J., DeCaprio, D., Gnerre, S., Grabherr, M., Kellis, M., Kleber, M., Bardeleben, C., Goodstadt, L., Heger, A., Hitte, C., Kim, L., Koepfli, K., Parker, H.G., Pollinger, J.P., Searle, S.M.J., Sutter, N.B., Thomas, R., Webber, C., Baldwin, J., Abebe, A., Abouelleil, A., Aftuck, L., Ait-zahra, M., Aldredge, T., Allen, N., An, P., Anderson, S., Antoine, C., Arachchi, H., Aslam, A., Ayotte, L., Bachantsang, P., Barry, A., Bayul, T., Benamara, M., Berlin, A., Bessette, D., Blitshteyn,

B., Bloom, T., Blye, J., Boguslavskiy, L., Bonnet, C., Boukhgalter, B., Brown, A., Cahill, P., Calixte, N., Camarata, J., Cheshatsang, Y., Chu, J., Citroen, M., Collymore, A., Cooke, P., Dawoe, T., Daza, R., Decktor, K., DeGray, S., Dhargay, N., Dooley, K., Dooley, K., Dorje, P., Dorjee, K., Dorris, L., Duffey, N., Dupes, A., Egbiremolen, O., Elong, R., Falk, J., Farina, A., Faro, S., Ferguson, D., Ferreira, P., Fisher, S., FitzGerald, M., Foley, K., Foley, C., Franke, A., Friedrich, D., Gage, D., Garber, M., Gearin, G., Giannoukos, G., Goode, T., Goyette, A., Graham, J., Grandbois, E., Gyaltsen, K., Hafez, N., Hagopian, D., Hagos, B., Hall, J., Healy, C., Hegarty, R., Honan, T., Horn, A., Houde, N., Hughes, L., Hunnicutt, L., Husby, M., Jester, B., Jones, C., Kamat, A., Kanga, B., Kells, C., Khazanovich, D., Chinh Kieu, A., Kisner, P., Kumar, M., Lance, K., Landers, T., Lara, M., Lee, W., Leger, J., Lennon, N., Leuper, L., LeVine, S., Liu, J., Liu, X., Lokyitsang, Y., Lokyitsang, T., Lui, A., Macdonald, J., Major, J., Marabella, R., Maru, K., Matthews, C., McDonough, S., Mehta, T., Meldrim, J., Melnikov, A., Meneus, L., Mihalev, A., Mihova, T., Miller, K., Mittelman, R., Mlenga, V., Mulrain, L., Munson, G., Navidi, A., Naylor, J., Nguyen, T., Nguyen, N., Nguyen, C., Nguyen, T., Nicol, R., Norbu, N., Norbu, C., Novod, N., Nyima, T., Olandt, P., O'Neill, B., O'Neill, K., Osman, S., Oyono, L., Patti, C., Perrin, D., Phunkhang, P., Pierre, F., Priest, M., Rachupka, A., Raghuraman, S., Rameau, R., Ray, V., Raymond, C., Rege, F., Rise, C., Rogers, J., Rogov, P., Sahalie, J., Settipalli, S., Sharpe, T., Shea, T., Sheehan, M., Sherpa, N., Shi, J., Shih, D., Sloan, J., Smith, C., Sparrow, T., Stalker, J., Stange-Thomann, N., Stavropoulos, S., Stone, C., Stone, S., Sykes, S., Tchuinga, P., Tenzing, P., Tesfaye, S., Thoulutsang, D., Thoulutsang, Y., Topham, K., Topping, I., Tsamla, T., Vassiliev, H., Venkataraman, V., Vo, A., Wangchuk, T., Wangdi, T., Weiand, M., Wilkinson, J., Wilson, A., Yadav, S., Yang, S., Yang, X., Young, G., Yu, Q., Zainoun, J., Zembek, L., Zimmer, A. for Broad Sequencing Platform members & Lander, E.S., 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 438, 803-819.

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M., 2012. Comparison of next-generation sequencing systems. J. BioMed & Biotech. 2012, 1-11.

Lu, D., Sargolzaei, M., Kelly, M., Li, C., Vander Voort, G., Wang, Z., Plastow, G., Moore, S., Miller, S. 2012. Linkage disequilibrium in Angus, Charolais, and Crossbred beef cattle. Front. Genet. 3, 152, 1-10.

MacEachern, S., McEwan, J. & Goddard, M., 2009. Phylogenetic reconstruction and the identification of ancient polymorphism in the Bovini tribe (*Bovidae*, *Bovinae*). BMC Genomics 10, 177, 1-7.

Makina, S.O., Maiwashe, A.N., Van Marle-Köster, E. & Muchadeyi, F.C., 2013. Estimating the extent of linkage disequilibrium in four SA cattle breeds. Proc. 46th Cong. S. Afr. Soc. Anim. Sci., University of the Free State, Bloemfontein. p. 99.

Makina, S.O., Muchadeyi, F.C., Marle-Köster, E., Taylor, J.F., Makgahlela, M.L. & Maiwashe, A., 2015. Genome-wide scan for selection signatures in six cattle breeds in South Africa. Genet. Sel. Evol. 47, 92, 1-14.

Makina, S.O., Muchadeyi, F.C., van Marle-Köster, E., MacNeil, M.D. & Maiwashe, A., 2014. Genetic diversity and population structure among six cattle breeds in South Africa using a whole genome SNP panel. Front. Genet. 5, 333, 1-7.

Mapholi, N.O., 2015. Exploring genetic architecture of tick resistance in South African Nguni cattle. Doctoral dissertation, Stellenbosch University, Stellenbosch, South Africa. pp 1-120.

Mapholi, N.O., Marufu, M.C., Maiwashe, A., Banga, C.B., Muchenje, V., MacNeil, M.D., Chimonyo, M. & Dzama, K., 2014. Towards a genomics approach to tick (*Acari: Ixodidae*) control in cattle: A review. Ticks Tick Borne Dis. 5, 475-483.

Matimba, A., Del-Favero, J., Van Broeckhoven, C. & Masimirembwa, C., 2009. Novel variants of major drug-metabolising enzyme genes in diverse African populations and their predicted functional effects. Hum. Genomics. 3, 169-190.

Matukumalli, L.K., Lawley, C.T., Schnabel, R.D., Taylor, J.F., Allan, M.F., Heaton, M.P., O'Connell, J., Moore, S.S., Smith, T.P., Sonstegard, T.S. & Van Tassell, C.P., 2009. Development and characterization of a high density SNP genotyping assay for cattle. PloS One. 4, e5350, 1-13.

McMichael, A.J., Powles, J.W., Butler, C.D. & Uauy, R., 2007. Food, livestock production, energy, climate change, and health. The Lancet. 370, 1253-1263.

McTavish, E.J. & Hillis, D.M., 2015. How do SNP ascertainment schemes and population demographics affect inferences about population history? BMC Genomics. 16, 266.

Mei, C., Wang, H., Zhu, W., Wang, H., Cheng, G., Qu, K., Guang, X., Li, A., Zhao, C., Yang, W. & Wang, C., 2016. Whole-genome sequencing of the endangered bovine species Gayal (*Bos frontalis*) provides new insights into its genetic features. Sci. Rep. 6, 19787, 1-6.

Metzker, M.L., 2010. Sequencing technologies-the next generation. Nature. Rev. Genet. 11, 31-46.

Morozova, O. & Marra, M.A., 2008. Applications of next-generation sequencing technologies in functional genomics. Genomics. 92, 255-264.

Mostert, B.E., 2007. PhD Thesis: The suitability of test-day models for genetic evaluation of dairy cattle in South Africa. University of Pretoria, South Africa. pp 1-86.

Musemwa, L., Mushunje, A., Chimonyo, M., Fraser, G., Mapiye, C. & Muchenje, V., 2008. Nguni cattle marketing constraints and opportunities in the communal areas of South Africa: Review. Afr. J. Agric. Res. 3, 239-245.

Mwai, O., Hanotte, O., Kwon, Y.J. & Cho, S., 2015. African indigenous cattle: unique genetic resources in a rapidly changing world. Asian Australas. J. Anim. Sci. 28, 911-921.

Negrini, R., Nicoloso, L., Crepaldi, P., Milanesi, E., Marino, R., Perini, D., Pariset, L., Dunner, S., Leveziel, H., Williams, J.L. & Marsan, P.A., 2008. Traceability of four European protected geographic indication (PGI) beef products using single nucleotide polymorphisms (SNP) and Bayesian statistics. Meat Sci. 80, 1212-1217.

Nozokkarmaher, M., 2016. MSc Thesis: The effect of inbreeding on Holstein-Friesian breed. Utah State University, USA. pp 1-39.

Oltenacu, P.A. & Broom, D.M., 2010. The impact of genetic selection for increased milk yield on the welfare of dairy cows. Anim. Welf. 19, 39-49.

Pareek, C.S., Smoczynski, R. & Tretyn, A., 2011. Sequencing technologies and genome sequencing. J. Appl. Genet, 52, 413-435.

Patterson, N., Price, A.L. & Reich, D., 2006. Population structure and eigenanalysis. PLoS Genet. 2, e190, 1-20.

Pienaar, L., 2014. MSc Thesis: Genetic diversity in the Afrikaner cattle breed. University of Freestate, Bloemfontein, South Africa. pp 1-107.

Pontius, J.U., Mullikin, J.C., Smith, D.R., Team, A.S., Lindblad-Toh, K., Gnerre, S., Clamp, M., Chang, J., Stephens, R., Neelam, B., Volfovsky, N., Schäffer, A.A., Agarwala, R., Narfström, K., Murphy, W.J., Giger, U., Roca, A.L., Antunes, A., Menotti-Raymond, M., Yuhki, N., Pecon-Slattery, J., Johnson, W.E., Bourque, G., Tesler, G., NISC Comparative Sequencing Program. & O'Brien, S.J., 2007. Initial sequence and comparative analysis of the cat genome. Genome Res. 17, 1675-1689.

Pool, J.E., Hellmann, I., Jensen, J.D. & Nielsen, R., 2010. Population genetic inference from genomic sequence variation. Genome Res. 20, 291-300.

Porto-Neto, L.R., Sonstegard, T.S., Liu, G.E., Bickhart, D.M., Da Silva, M.V., Machado, M.A., Utsunomiya, Y.T., Garcia, J.F., Gondro, C. & Van Tassell, C.P., 2013. Genomic divergence of zebu and taurine cattle identified through high-density SNP genotyping. BMC Genomics 14, 876, 1-12.

Qanbari, S., Pimentel, E.C.G., Tetens, J., Thaller, G., Lichtner, P., Sharifi, A.R. & Simianer, H., 2010. The pattern of linkage disequilibrium in German Holstein cattle. Anim. Genet. 41, 346-356.

Ramos, A.M., Crooijmans, R.P., Affara, N.A., Amaral, A.J., Archibald, A.L., Beever, J.E., Bendixen, C., Churcher, C., Clark, R., Dehais, P. & Hansen, M.S., 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PloS One 4, e6524, 1-13.

Rege, J.E.O., 1999. The state of African cattle genetic resources I. Classification framework and identification of threatened and extinct breeds. Anim. Genet. Resour. 25, 1-26.

Rege, J.E.O., Kahi, A.K., Okomo-Adhiambo, M., Mwacharo, J. & Hanotte, O., 2001. Zebu cattle of Kenya: Uses, performance, farmer preferences, measures of genetic diversity and options for improved use. ILRI (International Livestock Research Institute), Nairobi, Kenya. pp 103-104.

Rolf, M.M., McKay, S.D., McClure, M.C., Decker, J.E., Taxis, T.M., Chapple, R.H., Vasco, D.A., Gregg, S.J., Kim, J.W., Schnabel, R.D. & Taylor, J.F., 2010. How the next generation of genetic technologies will impact beef cattle selection. In Proceedings of the Beef Improvement Federations 42nd Annual Research Symposium and Annual Meeting, Columbia, MO, USA, pp 46-56.

Sanarana, Y., Visser, C., Bosman, L., Nephawe, K., Maiwashe, A. & van Marle-Köster, E., 2015. Genetic diversity in South African Nguni cattle ecotypes based on microsatellite markers. Trop. Anim. Health. Prod. 48, 379-385.

Sanger, F. & Coulson, A.R., 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 94, 441-448.

Schadt, E.E., Turner, S. & Kasarskis, A., 2010. A window into third-generation sequencing. Hum. Mol. Gen. 19, R227-R240.

Schefers, J.M. & Eeigel, K.A., 2012. Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. Anim. Front. 1, 4-9.

Schoeman, S.J.,1989. Recent research into the production potential of indigenous cattle with special reference to the Sanga. S. Afr. J. Anim. Sci. 19, 55-61.

Scholtz, M.M., 1988. Selection possibilities of hardy beef breeds in Africa: The Nguni example. In 3. Congres Mondial de Reproduction et Selection des Ovins et Bovins a Viande, (Paris France), 19-23 Jun 1988. INRA.

Scholtz, M.M., 2010. Beef breeding in South Africa (2nd ed.). Asikhulume pixArt, Rooihuiskraal, Pretoria, South Africa.

Scholtz, M.M., McManus, C., Okeyo, A.M. & Theunissen, A., 2011. Opportunities for beef production in developing countries of the southern hemisphere. Livest. Sci. 142, 195-202.

Schuster, I., 2011. Marker-assisted selection for quantitative traits. Crop Breed. Appl. Biotechnol. 11, 50-55.

Sharma, R., Kishore, A., Mukesh, M., Ahlawat, S., Maitra, A., Pandey, A.K. & Tantia, M.S., 2015. Genetic diversity and relationship of Indian cattle inferred from microsatellite and mitochondrial DNA markers. BMC Genet. 16, 73, 1-12.

Sölkner, J., Frkonja, A., Raadsma, H.W., Jonas, E., Thaller, G., Gootwine, E., Seroussi, E., Fuerst, C., Egger-Danner, C. & Gredler, B., 2010. Estimation of individual levels of admixture in crossbred populations from SNP chip data: examples with sheep and cattle populations. Interbull Bull. 42, 62-66.

Stehfest, E., Bouwman, L., Van Vuuren, D.P., Den Elzen, M.G., Eickhout, B. & Kabat, P., 2009. Climate benefits of changing diet. Climatic Change. 95, 83-102.

Stothard, P., Choi, J.W., Basu, U., Sumner-Thomson, J.M., Meng, Y., Liao, X. & Moore, S.S., 2011. Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. BMC Genomics 12, 559, 1-14.

Strydom, P.E. 2008. Do indigenous Southern African cattle breeds have the right genetics for commercial production of quality meat? Meat Sci. 80, 86-93.

Su, Z., Ning, B., Fang, H., Hong, H., Perkins, R., Tong, W. & Shi, L., 2011. Next-generation sequencing and its applications in molecular diagnostics. Expert Rev. Mol. Diagn. 11, 333-343.

Sundquist, A., Fratkin, E., Do, C.B. & Batzoglou, S., 2008. Effect of genetic divergence in identifying ancestral origin using HAPAA. Genom. Res. 18, 676-682.

Taberlet, P., Coissac, E., Pansu, J. & Pompanon, F., 2011. Conservation genetics of cattle, sheep, and goats. C. R. Biol. 334, 247-254.

Thornton, P.K., 2010. Livestock production: recent trends, future prospects. Philos. Trans. R. Soc B Biol. Sci. 365, 2853-2867.

Tosser-Klopp, G., Bardou, P., Bouchez, O., Cabau, C., Crooijmans, R., Dong, Y., Donnadieu-Tonon, C., Eggen, A., Heuven, H.C., Jamli, S., Jiken, A.J., Klopp, C., Lawley, C.T. McEwan, J., Martin, P., Moreno, C.R., Mulsant, P., Nabihoudine, I., Pailhoux, E., Palhière, I., Rupp, R., Sarry, J.,. Sayre, B.L., Tircazes, A., Wang, J., Wang, W., Zhang, W. & the International Goat Genome Consortium., 2014. Design and characterization of a 52K SNP chip for goats. PLoS One 9, e86227, 1-8.

Van der Werf, J., 2013. Genomic selection in animal breeding programs. Methods Mil. Biol. 1019, 543-561.

Van Marle, J., 1974. The breeding of beef cattle in South Africa: Past, present and future. S. Afr. J. Anim. Sci. 4, 297-304.

Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C. & Sonstegard, T.S., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Meth. 5, 247-252.

Voelkerding, K.V., Dames, S.A. & Durtschi, J.D., 2009. Next-generation sequencing: from basic research to diagnostics. Clin. Chem. 55, 641-658.

Wade, C.M., Giulotto, E., Sigurdsson, S., Zoli, M., Gnerre, S., Imsland, F., Lear, T.L., Adelson, D.L., Bailey, E., Bellone, R.R., Blöcker, H., Distl, O., Edgar, R.C., Garber, M., Leeb, T., Mauceli, E., MacLeod, J.N., Penedo, M.C., Raison, J.M., Sharpe, T., Vogel, J., Andersson, L., Antczak, D.F., Biagi, T., Binns, M.M., Chowdhary, B.P., Coleman, S.J., Della, V.G., Fryc, S., Guérin, G., Hasegawa, T., Hill, E.W., Jurka, J., Kiialainen, A., Lindgren, G., Liu, J., Magnani, E., Mickelson, J.R., Murray, J., Nergadze, S.G., Onofrio, R., Pedroni, S., Piras, M.F., Raudsepp, T., Rocchi, M., Røed, K.H., Ryder, O.A., Searle, S., Skow, L., Swinburne, J.E., Syvänen, A.C., Tozaki, T., Valberg, S.J., Vaudin, M., White, J.R., Zody, M.C., Broad Institute Genome Sequencing Platform, Broad Institute Whole Genome Assembly Team, Lander, E.S. & Lindblad-Toh. K., 2009. Genome sequence, comparative analysis, and population genetics of the domestic horse. Science. 326, 865-867.

Wang, M.D., Dzama, K., Hefer, C.A. & Muchadeyi, F.C., 2015. Genomic population structure and prevalence of copy number variations in South African Nguni cattle. BMC genomics. 16, 894, 1-16.

Webb, E.C., 2013. The ethics of meat production and quality - a South African perspective. S. Afr. J. Anim. Sci. 43, S2-S11.

Wilkinson, S., Wiener, P., Archibald, A.L., Law, A., Schnabel, R.D., McKay, S.D., Taylor, J.F. & Ogden, R., 2011. Evaluation of approaches for identifying population informative markers from high density SNP Chips. BMC Genet. 12, 45, 1-15.

Yonesaka, R., Sasazaki, S., Yasue, H., Niwata, S., Inayoshi, Y., Mukai, F. & Mannen, H., 2016. Genetic structure and relationships of 16 Asian and European cattle populations using DigiTag2 assay. Anim. Sci. J. 87, 190-196.

Zimin, A.V., Delcher, A.L., Florea, L., Kelley, D.R., Schatz, M.C., Puiu, D., Hanrahan, F., Pertea, G., Van Tassell, C.P., Sonstegard, T.S., Marçais, G., Roberts, M., Subramanian, P., Yorke, J.A., Salzberg, S.L., 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol. 10, R42, 1-10.

# CHAPTER THREE

## Genome-wide identification of breed-informative single-nucleotide polymorphisms in three South African indigenous cattle breeds

**A.A. Zwane[1,2#], A. Maiwashe[1,5], M.L. Makgahlela[1], A. Choudhury[3], J.F. Taylor[4] & E. Van Marle-Köster[2]**

[1]Department of Animal Breeding and Genetics, ARC-API, P/Bag X2, Irene, 0062,

[2]Department of Animal and Wildlife Sciences, University of Pretoria, P/Bag X20, Hatfield, Pretoria, 0028,

[3]Sydney Brenner Institute of Molecular Bioscience, University of the Witwatersrand, P/Bag 3, Wits, Gauteng, 2050,

[4]Division of Animal Sciences, University of Missouri, 920 East Campus Drive, Columbia, MO 65211-5300, USA,

[5]Department of Animal, Wildlife and Grassland Sciences, University of the Free State, Bloemfontein 9300, South Africa

**Published in the South African Journal of Animal Science 2016, 46 (No. 3)**

# Genome-wide identification of breed-informative single-nucleotide polymorphisms in three South African indigenous cattle breeds

**A.A. Zwane[1,2#], A. Maiwashe[1,5], M.L. Makgahlela[1], A. Choudhury[3], J.F. Taylor[4] & E. Van Marle-Köster[2]**

[1]Department of Animal Breeding and Genetics, ARC-API, P/Bag X2, Irene, 0062, [2]Department of Animal and Wildlife Sciences, University of Pretoria, P/Bag X20, Hatfield, Pretoria, 0028, [3]Sydney Brenner Institute of Molecular Bioscience, University of the Witwatersrand, P/Bag 3, Wits, Gauteng, 2050, [4]Division of Animal Sciences, University of Missouri, 920 East Campus Drive, Columbia, MO 65211-5300, USA, [5]Department of Animal, Wildlife and Grassland Sciences, University of the Free State, Bloemfontein 9300, South Africa

*(Received 1 June 2016; Accepted 31 July 2016; First published online September 2016)*



## Abstract

Access to genotyping assays enables the identification of informative markers that discriminate between cattle breeds. Identification of these markers can assist in breed assignment, improvement and conservation. The objective of this study was to identify breed informative markers to discriminate between three South African indigenous cattle breeds. Data from BovineSNP50 and GeneSeek Genomic Profiler (GGP-80K) assays were generated for Afrikaner, Drakensberger and Nguni, and were analysed for their genetic differentiation. Hereford and Angus were included as outgroups. Breeds were differentiated using principal component analysis (PCA). Single-nucleotide polymorphisms (SNPs) within the breeds were determined when minor allele frequency (MAF) was ≥ 0.05. Breed-specific SNPs were identified using Reynolds $F_{st}$ and extended Lewontin and Krakauer's (FLK) statistics. These SNPs were validated using three African breeds, namely N'Dama, Kuri and Zebu from Madagascar. PCA discriminated among the breeds. A larger number of polymorphic SNPs was detected in Drakensberger (73%) than in Afrikaner (56%) and Nguni (65%). No substantial numbers of informative SNPs ($F_{st}$ ≥ 0.6) were identified among indigenous breeds. Eleven SNPs were validated as discriminating the indigenous breeds from other African breeds. This is because the SNPs on BovineSNP50 and GGP-80K assays were ascertained as being common in European taurine breeds. Lower MAF and SNP informativeness observed in this study limits the application of these assays in breed assignment, and could have other implications for genome-wide studies in South African indigenous breeds. Sequencing should therefore be considered to discover new SNPs that are common among indigenous South African breeds and also SNPs that discriminate among these indigenous breeds.

**Keywords:** Beef cattle, genetic differentiation, minor allele frequency, polymorphisms
[#]Corresponding author: zwanea@arc.agric.za

## Introduction

In southern Africa, livestock has always played a vital role in the agricultural economies of countries such as South Africa, Namibia and Botswana, where commercial livestock enterprises, smallholder and communal farming contribute to food production, social needs and the general wellbeing of rural households (Bettencourt *et al*., 2013). The demand for animal protein is increasing. Meeting this demand will require more efficient production to ensure long-term sustainability of production and environmental conservation (Otten & Van den Weghe, 2011). Developing countries are often richly endowed with indigenous livestock resources that are well adapted to environmental challenges, but lack productivity (Mwai *et al.,* 2015) in milk and meat production compared with imported commercial breeds (Renaudeau *et al.,* 2012).

South African indigenous cattle have unique morphological features that distinguish them from other cattle breeds (Makina *et al*., 2014). These breeds include Nguni, Afrikaner and Drakensberger,

known as Sanga cattle, which belong to the subspecies *Bos taurus africanus*. Sanga cattle breeds originated from eastern and northern Africa, the home of Sanga and Zenga cattle. Sanga are possibly crossbreds between the indigenous humpless cattle *(Bos taurus)* and Zebu *(Bos indicus),* whereas Zenga are crossbreds between Zebu and Sanga (Rege, 1999). These breeds inhabit eastern and southern Africa and are known to be well adapted to harsh environmental conditions (Okello & Sabiiti, 2006). Nguni cattle are recognized for their ability to survive when exposed to high temperatures and low-quality grass and for their resistance to parasites and tick-borne diseases (Scholtz, 1988; Mapholi *et al*., 2014). The Afrikaner is a hardy beef cattle breed, known for its adaptation to harsh conditions. It was used in the development of the Bonsmara, a South African composite breed (Van Marle, 1974; Mason, 1996; Strydom, 2008). The Drakensberger is known for its adaptability, especially to Sourveld regions and is regarded as one of the local indigenous breeds of South Africa. Its origin and history have not been well documented (Scholtz, 2010). Other commercial breeds such as Angus and Hereford have been continuously selected for production traits including beef yield (Scholtz, 2010; Kugonza *et al*., 2011). Hereford and Angus are common European taurine breeds that were introduced to South Africa in 1892 and 1895, respectively (Hanotte *et al*., 2002; Scholtz, 2010). These breeds are well known internationally, and possess good mothering ability, good growth rate, early marketability, grazing performance and good temperament (Vasconcellos *et al*., 2003). The Hereford has contributed to the development of the South African beef industry through its role in the development of the Bonsmara breed (Scholtz, 2010).

South African beef cattle are genetically diverse. However, certain populations have been identified as critically endangered, namely Pedi and Shangaan cattle (both South African Sanga) (Rege, 1999; Mwai *et al*., 2015). Indigenous breeds are often subjected to indiscriminate crossbreeding with exotic breeds to improve production, especially in rural areas. However, this practice leads to the loss of genetic diversity. The future utilization of indigenous genetic resources depends on their conservation, promotion and improvement (Frese *et al*., 2014). There is an urgent need to characterize South African indigenous cattle populations using genomic information to enhance their productivity and to inform on their utilization in breeding programmes (Hanotte *et al*., 2010; Mwai *et al*., 2015). However, there is limited knowledge of their genetic composition (e.g., Makina *et al*., 2014).

Knowledge of breed composition may enable better understanding of the basis of adaptive traits of cattle in their own production environments, which is critical for genome-wide association studies (Kuehn *et al*., 2011) and for the assigning individuals to their population of origin (Sanz *et al*., 2014). Furthermore, understanding the breed composition of these cattle populations could be useful in predicting heterosis (Kuehn *et al*., 2011), and assisting with the proper management of genetic resources for long-term sustainability (Gorbach *et al*., 2010).

Several studies have shown the utility of SNP markers for breed differentiation and individual assignment (Yoon *et al*., 2008; Negrini *et al*., 2009; Pariset *et al*., 2010; Kuehn *et al*., 2011; Lewis *et al*., 2011; Wilkinson *et al*., 2011; Dimauro *et al*., 2013; Hulsegge *et al*., 2013). Individual assignment uses genetic information to allocate an individual to a population and to determine the origin of unknown individuals (Negrini *et al*., 2008). Methods of selecting informative markers to discriminate among breeds and assign individuals to their population of origin have been described (Negrini *et al*., 2009; Ramos *et al*., 2011; Wilkinson *et al*., 2011; Opara *et al*., 2012). A relatively small number of SNPs can be used to elucidate the genetic structure among breeds (Wilkinson *et al*., 2011), and only a small set of informative SNPs, if chosen appropriately, may be needed for accurate breed assignment (Mackay *et al*., 2008; Hulsegge *et al*., 2013; Martinez-Camblor *et al*., 2014). High-density SNP assays, such as the BovineSNP50 and bovine high-density (BovineHD) are now available with large numbers of SNPs from which the most informative SNPs can be selected for breed assignment (Matukumalli *et al*., 2009). The objective of this study was to identify breed informative SNPs for differentiating among the three South African cattle breeds using genotype data generated with the BovineSNP50 and GGP-80K assays.

## Materials and Methods

Genotype data generated from previous projects were available for this study (Makina *et al*., 2014). Data from five breeds were studied, including three indigenous South African breeds (20 Afrikaner, 48 Drakensberger and 47 Nguni) and two exotic British breeds (31 Angus and 33 Hereford), genotyped with the Illumina BovineSNP50 (Illumina, San Diego, Calif, USA) and GGP-80K assay (Neogen, Lincoln, Nebr., USA). Angus and Hereford were considered outgroups in this study because these breeds were included in the development of the BovineSNP50 chip. Afrikaner and Hereford, genotyped with the GGP-80K assay, were provided by the Department of Animal & Wildlife Sciences

at the University of Pretoria, while Angus, Drakensberger and Nguni, genotyped with the BovineSNP50 assay, were provided by the Agricultural Research Council (ARC).

BovineSNP50 BeadChip genotypes for 54 609 SNPs and GGP-80K assay genotypes for 88 683 SNPs were available, respectively. These chips contain highly informative SNPs that are evenly distributed throughout the autosomal genome of the major European cattle breeds (Michelizzi *et al*., 2011). The GGP-80K assay consists of SNPs common to taurine cattle that were derived from the BovineSNP50 and BovineHD assays, but includes variants derived from these assays that are common in *Bos indicus* (Edea *et al*., 2015). Genotypes from the two genotyping platforms were merged and SNPs that were common to both BeadChips (i.e., 28 261 SNPs) were used for the analysis. The markers were next filtered within breeds to remove those with call rates of less than 98% and samples with more than 10% missing genotypes. After applying these filters, 26 472 SNPs remained for further analysis. To evaluate that the genotypes had been called with the same Illumina format for both assays, sets of Afrikaner ('afr', n = 48), Hereford ('hfd', n = 20) and Angus ('ang', n = 20) animals genotyped with the Illumina BovineSNP50 chip and called with the Illumina A/B format were included in the analysis (Makina *et al*., 2014; Decker *et al*., 2014). Furthermore, three breeds originating from Africa (N'Dama (*Bos taurus taurus*), Kuri (*Bos taurus taurus*) and Zebu from Madagascar (*Bos taurus indicus*) (Decker et al., 2014), also genotyped using BovineSNP50 chip, were used to compare the indigenous South African breeds with these other African breeds. These breeds were among the several African breeds in the dataset, but were selected based on their demographic location since no breeds originate from neighbouring countries to South Africa.

Allele frequency estimates for each SNP marker were used to determine its utility for breed differentiation. To examine the basic indices of genetic variability between the breeds, allele frequency distributions and the proportion of SNPs were estimated within a breed using PLINK version 1.09 (Purcell *et al*., 2007). SNP proportions by MAF ≥ 0.05 were determined for each breed (Edea *et al*., 2012; Grasso *et al*., 2014; Edea *et al*., 2015). Within-population genetic diversity was estimated by calculating observed heterozygosity (Ho), expected heterozygosity (He) and mean inbreeding for each population using GoldenHelix SNP & Variation Suite (SVS) software (GoldenHelix Inc., Bozeman, Mont, USA) (Grasso *et al*., 2014).

For genetic structure analysis with PCA, 26 472 SNPs were further filtered across breeds to remove SNPs with call rates ≤ 98%, MAF ≤ 0.01 or with *P*-value for a chi-square test for Hardy-Weinberg equilibrium (HWE) ≤0.0001 (Lee *et al*., 2013). SNPs were filtered to avoid the effects of ascertainment bias on diversity indexes and genetic distances (Edea *et al*., 2015). The average proportion of alleles shared between animals was calculated using PLINK using the commands - - cluster and - - distance-matrix and the resulting matrix was used to generate the PCA plots (Kijas *et al*., 2012). The genotypes from 'afr', 'hfd' and 'ang' were used to check for the possibility of assay or genotyping call effects between the two datasets (BovineSNP50 and GGP-80K), using PCA analysis (Gurdasani *et al*., 2015).

Breed-specific markers were determined by identifying markers with MAF ≥ 0 (Grasso *et al*., 2014). A SNP was declared to be breed specific when it possessed an allele that was present in only one breed (Ramos *et al*., 2011). To find the SNPs that distinguished between the breeds, pairwise $F_{st}$ (Weir & Cockerham, 1984; Weir, 1996) and an FLK statistic (Lewontin & Krakauer, 1973) were calculated between each pair of breeds (Wilkinson *et al*., 2011; Fariello *et al*., 2013). Pairwise $F_{st}$ was calculated using SVS software and the FLK statistics were calculated with the haplotype-based method in hapFLK (Bonhomme *et al*., 2010). The method uses the genetic distance of Reynolds *et al.* (1983), and builds the population relationship tree using the neighbour-joining algorithm applied to the matrix of Reynolds distances for a specified outgroup (e.g., Angus in this study). SNP pairs with high FLK values ($P$ <0.001) and $F_{st}$ values of 0.60 or greater were selected for each breed comparison (Nishimura *et al*., 2012).

## Results

The average MAF observed for Afrikaner, Drakensberger, Nguni, Hereford and Angus is shown in Table 1. There were no differences between the indigenous and the exotic breeds for the proportion of SNP with MAF ≥ 0.01. A small difference was observed between the Afrikaner and Nguni populations for the proportion of SNP with MAF ≥ 0 and MAF ≥ 0.01. However, for both MAF criteria, the Drakensberger and Angus (a taurine breed) were similar. Less than 1% of the SNPs were fixed within the South African indigenous cattle (data not shown).

As shown in Figure 1, approximately 80% of the markers had MAF ≥ 0.01, while 65% of markers were highly polymorphic with MAF ≥ 0.05 across the South African indigenous breeds. In the Drakensberger, 73% of the markers were highly polymorphic, which was greater than in Angus (68%).

The Afrikaner had the lowest proportion of polymorphic SNPs compared with the other breeds. Angus and Hereford generally had high levels of polymorphism. Table 2 presents the measure of Ho, He and average inbreeding coefficients. There was low observed heterozygosity (Ho) in Hereford compared with the other breeds (Table 2).

A principal component analysis was performed to evaluate the genetic structure and affinities among the five populations included in this study. Figure 2 illustrates the clustering of the five breeds, showing the separation of indigenous South African breeds and exotic breeds (Angus and Hereford). The cluster also shows the relationship between the Drakensberger and exotic breeds. The cluster of the five breeds with 'afr', 'hfd' and 'ang' showed that the observed pattern of clustering separated these populations based on their relatedness rather than on their genotyping platforms (BovineSNP50 versus GGP-80K) or sample batches, as shown in Figure 3. These PCA results, as expected, show that the indigenous populations cluster closer to each other in comparison with the exotic breeds.

***T*able 1** Average minor allele frequency (MAF) and standard deviations (SD) in Afrikaner (AFR), Drakensberger (DRA), Nguni (NGI), Angus (ANG) and Hereford (HFD) cattle breeds

| Population | Breed | No. of animals | Mean ± SD (MAF ≥ 0) | Mean ± SD (MAF ≥ 0.01) | Mean ± SD (MAF ≥ 0.05) |
|---|---|---|---|---|---|
| *Indigenous breeds* | | | | | |
| Afrikaner | AFR | 20 | 0.20 ± 0.159 | 0.24 ± 0.146 | 0.25 ± 0.138 |
| Drakensberger | DRA | 48 | 0.26 ± 0.145 | 0.26 ± 0.142 | 0.28 ± 0.130 |
| Nguni | NGI | 47 | 0.21 ± 0.158 | 0.23 ± 0.151 | 0.27 ± 0.133 |
| *Outgroup taurine breeds* | | | | | |
| Hereford | HFD | 33 | 0.28 ± 0.142 | 0.29 ± 0.136 | 0.30 ± 0.126 |
| Angus | ANG | 31 | 0.26 ± 0.147 | 0.27 ± 0.139 | 0.29 ± 0.126 |

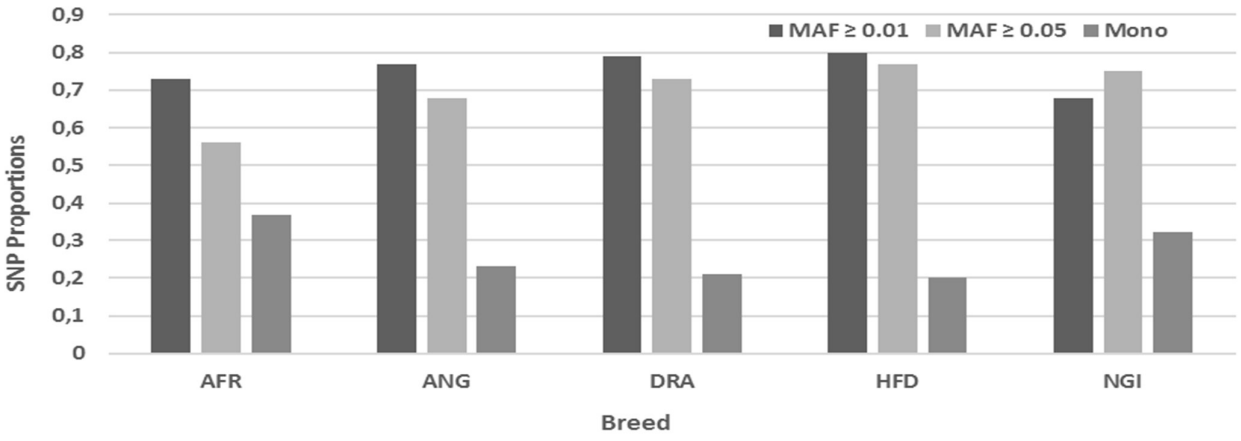


**Figure 1** Single-nucleotide polymorphism and monomorphism as determined by minor allele frequency = 0, MAF ≥ 0.01 and MAF ≥ 0.05 thresholds for each breed (AFR: Afrikaner; DRA: Drakensberger; NGI: Nguni; ANG: Angus; HFD: Hereford).

**Table 2** Indexes of genetic diversity in South African cattle breeds

| Breed | Breed group | Observed heterozygosity (Ho) | Expected heterozygosity (He) | Mean inbreeding coefficient (fi) |
|---|---|---|---|---|
| AFR | Sanga | 0.22 | 0.22 | 0.00 |
| ANG | Taurine | 0.25 | 0.24 | 0.03 |
| DRA | Sanga | 0.26 | 0.25 | 0.01 |
| HFD | Taurine | 0.20 | 0.20 | -0.02 |
| NGI | Sanga | 0.24 | 0.23 | -0.01 |

AFR: Afrikaner; DRA: Drakensberger; NGI: Nguni; ANG: Angus; HFD: Hereford

To examine the influence of indicine introgression in the indigenous South African breeds, PCA of allele sharing was performed between South African and the other African populations (N'Dama (NDAM), Kuri (KUR) and Zebu from Madagascar (ZMA) breeds) (Figure 4). This analysis showed a clear separation between the South African and other African breeds. The clustering still placed Drakensberger on the diagonal axis between European and African taurine breeds. Regardless of the analytical method or subset of breeds analysed, these three groups were consistently observed to be highly differentiated. SNPs with highly differentiated allele frequencies were identified using pairwise Reynolds Fst and hapFLK analyses. From this analysis, 325 informative loci, for example SNPs with Fst ≥ 0.6, were identified between South African breeds, but the hapFLK analysis provided little evidence of the existence of highly breed informative SNPs in these data (P >0.001).

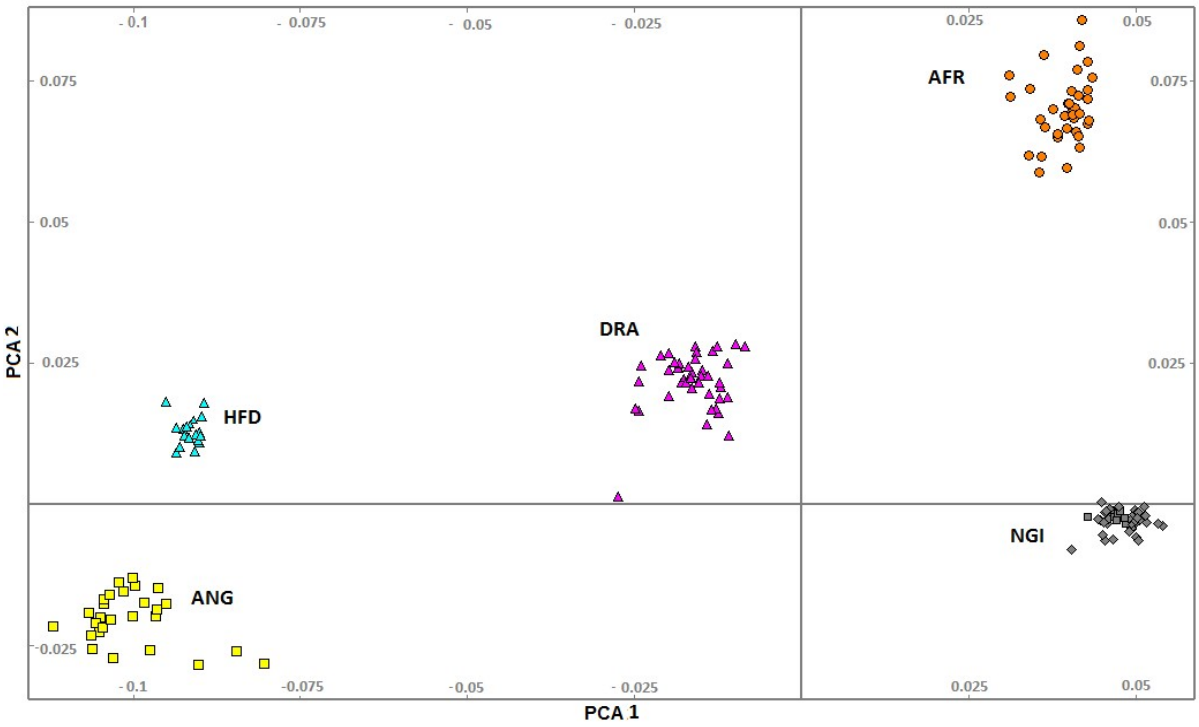


**Figure 2** Principal component analysis for population structure in South African cattle populations showing first two principal components for all the breeds. (AFR: Afrikaner; DRA: Drakensberger; NGI: Nguni; ANG: Angus; HFD: Hereford).

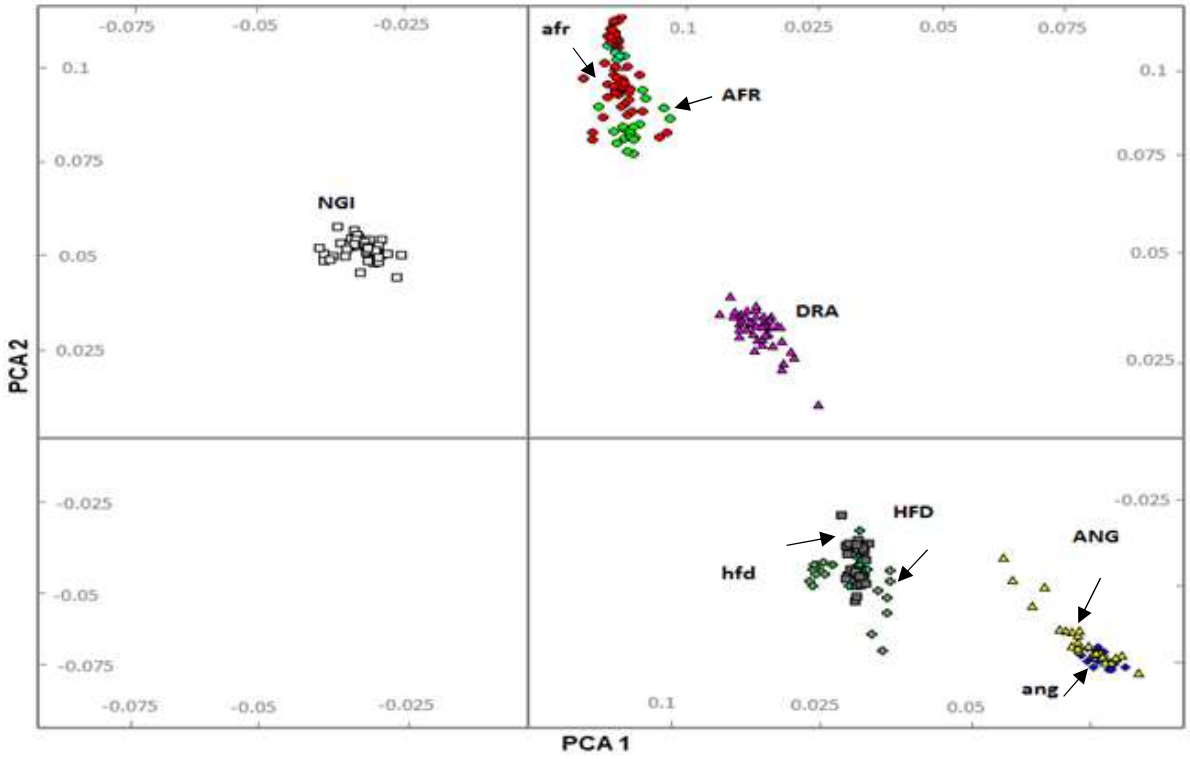


**Figure 3** First two principal components for the eight breeds (five breeds including second Afrikaner (afr), Hereford (hfd) and Angus (ang) group genotyped with BovineSNP50K). (AFR: Afrikaner; DRA: Drakensberger; NGI: Nguni; ANG: Angus; HFD: Hereford).

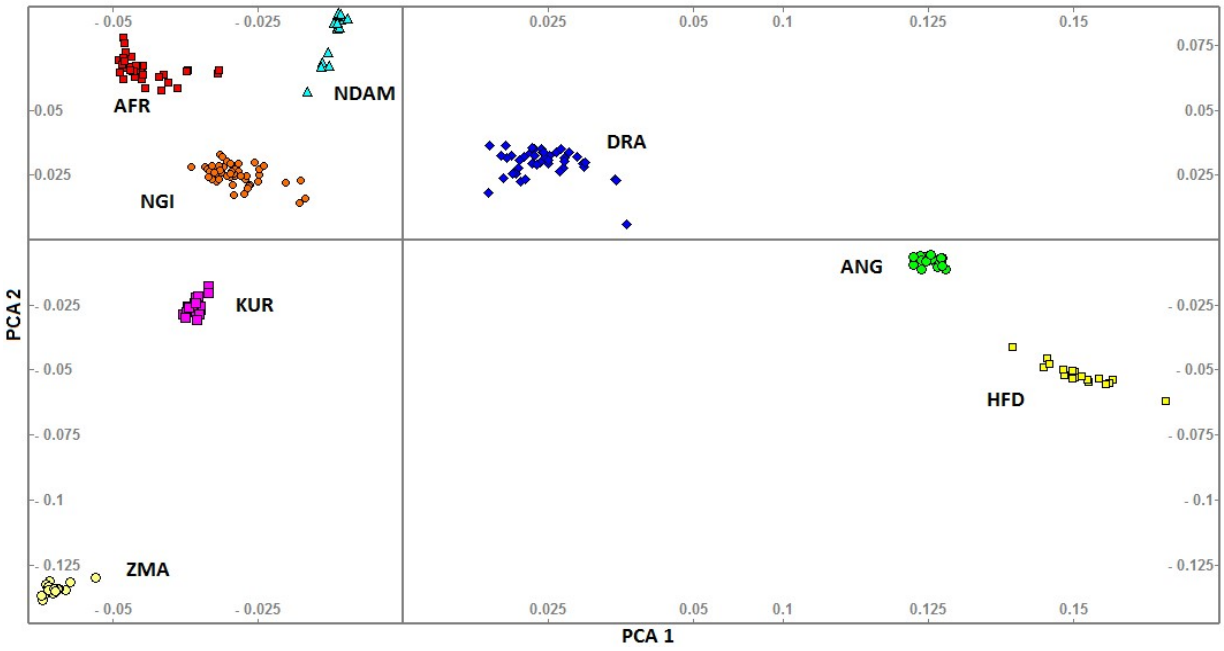


**Figure 4** Clustering of five South African cattle breeds and three African breeds showing separation among the breeds. (AFR: Afrikaner; DRA: Drakensberger; NGI: Nguni; ANG: Angus; HFD: Hereford; NDAM: N'Dama; KUR: Kuri; ZMA: Zebu from Madagascar).

Analyses of South African breeds with African breeds (N'Dama, Kuri and Zebu from Madagascar) revealed a set of informative SNPs that differentiate the South African indigenous and other African (taurine × indicine) breeds. Differentiated regions were identified on nine chromosomes (Table 3), containing SNPs with high FLK values. Furthermore, these SNPs were checked against a SNP allele frequency database at the University of Missouri to determine whether they were fixed only in the African breeds. This revealed that both alleles at each SNP segregated in at least 30 other cattle breeds for all SNPs (data not shown). Consequently, the SNPs producing high FLK values did not possess alleles that were specific to the African breeds, but are sufficiently skewed in frequency to differentiate South African cattle breeds from other African breeds.

**Table 3** Informative single-nucleotide polymorphisms that discriminate between South African and African breeds

| SNP | Chromosome No | Position | MAF |
|---|---|---|---|
| BTB-00187975 | 4 | 57553315 | 0.13 |
| BTB-00432889 | 10 | 62653672 | 0.11 |
| BTB-01642403 | 24 | 6512738 | 0.48 |
| BTB-00363099 | 8 | 84738093 | 0.11 |
| ARS-BFGL-NGS-14285 | 1 | 30024945 | 0.40 |
| BFGL-NGS-109801 | 17 | 55713369 | 0.37 |
| ARS-BFGL-NGS-34121 | 21 | 32113699 | 0.18 |
| BTA-70284-no-rs | 4 | 41895490 | 0.46 |
| Hapmap48127-BTA-93939 | 1 | 142370512 | 0.14 |
| ARS-BFGL-BAC-27254 | 20 | 44861949 | 0.42 |
| BTA-118486-no-rs | 19 | 21716537 | 0.32 |

SNP: Single nucleotide polymorphism; MAF: Minor allele frequency

## Discussion

The analyses performed in this study were conducted to identify breed informative markers for use in discriminating among indigenous South African cattle breeds using the BovineSNP50 and GGP-80K data. A number of studies have shown the usefulness of SNP data for identifying breed informative SNPs for discrimination among breeds (Negrini *et al*., 2009; Wilkinson *et al*., 2011; Nishimura *et al*., 2012; Edea *et al*., 2012). Although the BovineSNP50 and GGP-80K assays were designed to contain variants that were common to taurine breeds, the authors decided to test their usefulness in identifying informative SNPs to discriminate between South African indigenous cattle breeds. The differences between cattle breeds for the mean MAFs ranged from 0.20 to 0.29 with an average of 0.24 (SD = 0.143) and were similar to those observed in previous studies (Chan *et al*., 2008; Edea *et al*., 2012). Studies have shown that MAF limits the utility of markers in association studies owing to the effects of rare alleles that are difficult to estimate (Gurgul *et al*., 2013). Although rare and fixed alleles could be used to explain the distinct loci in a particular population, they may account for the reduced percentage of informative markers within the breeds (Dadi *et al*., 2011). SNPs with low MAF have a frequency imbalance between the two allelic groups, which may reflect functional importance (Cargill *et al*., 1999).

The differences in allele frequencies among the breeds may be caused by genetic drift, selection to adaptation to the local South African environment or ancient divergence among founder populations (MacEachern *et al*., 2009, Dadi *et al*., 2012). The lowest average MAF was observed in Afrikaner (0.20). Drakensberger had a higher average MAF compared with the Afrikaner and Nguni. This agrees with the study by Makina *et al.* (2014), which revealed the closer relationship of Drakensberger to the European taurine breeds. The overall MAF in South African indigenous breeds was lower compared with the European taurine breeds, which may be due to ascertainment bias in the design of the BovineSNP50 assay. McKay *et al.* (2008) also found a lower average minor allele frequency for *Bos indicus* than for *Bos taurus* breeds. The lower MAF in *Bos indicus* could again reflect the lower

representation of indicine populations in the design of these assays such that common loci identified in taurines are generally not the most common in indicines (Chan *et al*., 2008; Edea *et al*., 2012; Espigolan *et al*., 2013). Therefore, it is possible that the SNPs that have been identified as being useful in one population may not necessarily be as useful in another (Allen *et al*., 2010). The differences in observed allele frequencies among breeds show the genetic diversity that exists within and between the breeds (Allen *et al*., 2010).

The proportion of SNP polymorphisms that was common to South African indigenous cattle breeds was generally lower than for the two British breeds, except for Drakensberger. This result was expected because these breeds were included in the design of the bovine SNP assays. The higher level of polymorphisms in Drakensberger is likely related to the admixture that occurred in the development of the breed, which was observed by Makina *et al.* (2014). The higher degree of SNP polymorphisms than the observed monomorphic SNPs in indigenous breeds could have contributed to the inability to find SNPs with alleles that are specific to South African indigenous breeds, and could have other significant impacts on the design and application of marker association studies in South African populations. Studies have indicated that the majority of the SNP markers on the BovineSNP50 BeadChip were discovered in Angus, Holstein and Hereford (Van Tassell *et al*., 2008) and could have influenced the level of SNP informativeness in such a way that the breeds used in the discovery process show higher MAFs and SNP variability, manifested in this study in Angus and Hereford compared with the South African indigenous breeds.

The separation of indigenous South African breeds from Angus and Hereford populations was consistent with the current understanding of ancestry and population structure in South African populations (Makina *et al*., 2014). It was also in agreement with insights from previous studies that included partially overlapping populations (Decker *et al*., 2014). Since subsets of animals from the same breed that were genotyped on assays clustered closely (Figure 3), this indicates that there were no assay or genotyping effects from differences in ascertainment bias, genotyping accuracy, allele calling or other technical variables between the genotyping platforms (Gurdasani *et al*., 2015).

The $F_{st}$ measure for genetic differentiation (Willing *et al*., 2015) and the FLK test, which accounts for unequal population size and the hierarchical structure of relationships among the populations (Bonhomme, 2010), provided little evidence of the existence of breed informative SNPs from these genotyping platforms in indigenous South African breeds. This could be because of lower MAF and high SNP polymorphisms observed between these breeds (Tabangin *et al*., 2009). It is generally considered that uninformative markers (i.e., low MAF loci) add variability and noise to the results and compromise the power of population genetic studies (Liu *et al*., 2005). However, effective exploration of other SNP identification methods, such as genome resequencing, could help to identify the most informative markers, and produce an optimal minimum set of markers that could accurately and efficiently differentiate among populations (Ding *et al*., 2011).

The cluster of South African indigenous breeds and other African breeds sampled from Decker *et al.* (2014) enabled the identification of the regions in the genome that discriminate between the populations. Eleven identified SNPs with higher $F_{st}$ values (>0.6) and higher FLK best described the distinctiveness of the breeds. These SNPs were useful in segregating the South African breeds from the other African breeds. The numbers of SNPs, however, were lower than the 18 informative SNPs found between Japanese Black and Holstein by Nishimura *et al*. (2012) using BovineSNP50 data. This is consistent with the development of the assay in which SNPs with a high MAF across taurine breeds were preferentially selected in the assay design. Consequently, sets of randomly chosen SNP markers in taurine breeds may have sufficient genetic information to produce moderate levels of power to assign individuals to other taurine breeds (Wilkinson *et al*., 2011). Therefore, the allele frequency distribution within other breeds reveals that the 11 identified SNPs do not possess breed-specific alleles. This suggests that the sequencing of indigenous South African breeds should be considered to identify a large number of informative SNPs specific to discriminating among South African breeds.

## Conclusion

The levels of genetic variation for SNPs on the BovineSNP50 and GGP-80K assays identified in this study indicate that these assays have utility for genetic studies in South African populations. The lower average MAF in the indigenous South African breeds reduced the effectiveness of the assays for the selection of breed-informative markers. This may affect their utility in downstream genomic applications. The assays were not adequate for identifying breed informative markers allowing for a small subset of markers to be used to differentiate between the South African indigenous breeds and African breeds. Therefore, identification of SNPs with breed-specific fixation of alternate alleles appears to require the whole genome sequencing of pools of DNA from individuals from the local cattle breeds

to avoid the biases inherent to SNP assays. This would help to overcome the challenge of ascertainment bias, and would improve the MAF distribution of variants available for genotyping South African indigenous breeds.

## Acknowledgements

The authors acknowledge the financial support from the Red Meat Research and Development of South Africa (RMRDSA). The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged by the first author. Opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF.

## Authors' Contributions

AAZ designed the experiment, carried out the analysis and drafted the manuscript. AC, MLM assisted with statistical analysis. AM, EVM and JFT structured scientific content. All authors provided editorial suggestions and revisions, read and approved the final draft.

## Conflict of Interest Declaration

The authors declare that they have no competing interests

### References

Allen, H.L., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S. & Ferreira, T., 2010. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nat. 467, 832-838.

Bettencourt, E.M.V., Tilman, M., Henriques, P.D.D.S., Narciso, V. & Carvalho, M.L.D.S., 2013. The economic and socio-cultural role of livestock in the wellbeing of rural communities of Timor-Leste. CEFAGE-UE Working Paper 2013/01. pp. 1-18.

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S. & SanCristobal, M., 2010. Detecting selection in population trees: the Lewontin and Krakauer test extended. Genet. 186, 241-262.

Bradley, D.G., MacHugh, D.E., Loftus, R.T., Sow, R.S., Hoste, C.H. & Cunningham, E.P., 1994. Zebu-taurine variation in Y chromosomal DNA: a sensitive assay for genetic introgression in West African trypanotolerant cattle populations. Anim. Genet. 25, 7-12.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Lane, C.R., Lim, E.P., Kalyanaraman, N., Nemesh, J. & Ziaugra, L., 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. 22, 231-238.

Chan, E. K. F., Hawken, R. & Reverter, A., 2008. The combined effect of SNP-marker and phenotype attributes in genome-wide association studies. Anim. Genet. 40, 149-156.

Dadi, H., Kim, J.J., Yoon, D. & Kim, K.S., 2011. Evaluation of single nucleotide polymorphisms (SNPs) genotyped by the Illumina Bovine SNP50K in cattle focusing on Hanwoo breed. Asian Australas. J. Anim. Sci. 25, 28-32.

Decker, J.E., McKay, S.D., Rolf, M.M., Kim, J., Alcala, A.M., Sonstegard, T.S., Hanotte, O., Gotherstrom, A., Seabury, C.M., Praharani, L. & Babar, M.E., 2014. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. PLoS Genet. 10 (3), e1004254, 1-14.

Dimauro, C., Cellesi, M., Steri, R., Gaspa, G., Sorbolini, S., Stella, A. & Macciotta, N.P.P., 2013. Use of the canonical discriminant analysis to select SNP markers for bovine breed assignment and traceability purposes. Anim. Genet. 44, 377-382.

Ding, L., Wiener, H., Abebe, T., Altaye, M., Go, R.C., Kercsmar, C., Grabowski, G., Martin, L.J., Hershey, G.K.K., Chakorborty, R. & Baye, T.M., 2011. Comparison of measures of marker informativeness for ancestry and admixture mapping. BMC Genomics. 12 (622), 1-18.

Edea, Z., Dadi, H., Kim, S.W., Dessie, T. & Kim, K.S., 2012. Comparison of SNP variation and distribution in indigenous Ethiopian and Korean Cattle (Hanwoo) populations. Genomics Inform. 10, 200-205.

Edea, Z., Bhuiyan, M.S.A., Dessie, T., Rothschild, M.F., Dadi, H. & Kim, K.S., 2015. Genome-wide genetic diversity, population structure and admixture analysis in African and Asian cattle breeds. Anim. 9 (02), 218-226.

Espigolan, R., Baldi, F., Boligon, A.A., Souza, F.R., Gordo, D.G., Tonussi, R.L., Cardoso, D.F., Oliveira, H.N., Tonhati, H., Sargolzaei, M. & Schenkel, F.S., 2013. Study of whole genome linkage disequilibrium in Nellore cattle. BMC Genomics 14, 1-8.

Fariello, M.I., Boitard, S., Naya, H., SanCristobal, M. & Servin, B., 2013. Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genet. 193, 929-941.

Frese, L., Palme, A. & Kik, C., 2014. On the sustainable use and conservation of plant genetic resources in Europe. Report from Work Package 5, 1-34.

Gorbach, D.M., Makgahlela, M.L., Reecy, J.M., Kemp, S.J., Baltenweck, I., Ouma, R., Mwai, O., Marshall, K., Murdoch, B., Moore, S. & Rothschild, M.F., 2010. Use of SNP genotyping to determine pedigree and breed composition of dairy cattle in Kenya. J. Anim. Breed. Genet. 127, 348-351.

Grasso, A.N., Goldberg, V., Navajas, E.A., Iriarte, W., Gimeno, D., Aguilar, I., Medrano, J.F., Rincón, G. & Ciappesoni, G., 2014. Genomic variation and population structure detected by single nucleotide polymorphism arrays in Corriedale, Merino and Creole sheep. Genet. Mol. Biol. 72, 389-395.
Gurgul, A., Żukowski, K., Pawlina, K., Ząbek, T., Semik, E. & Bugno-Poniewierska, M., 2013. The evaluation of bovine SNP50 BeadChip assay performance in Polish Red cattle breed. Folia Biol. 61, 173-176.
Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A. & Ritchie, G.R., 2015. The African genome variation project shapes medical genetics in Africa. Nat. 517 (7534), 327-332.
Hanotte, O., Bradley, D.G., Ochieng, J.W., Verjee, Y., Hill, E.W. & Rege, J.E.O., 2002. African pastoralism: genetic imprints of origins and migrations. Science 296 (5566), 336-339.
Hanotte, O., Dessie, T. & Kemp, S., 2010. Time to tap Africa's livestock genomes. Sci. (Washington) 328, 1640-1641.
Hayes, B. & Goddard, M., 2010. Genome-wide association and genomic selection in animal breeding. Genome 53 (11), 876-883.
Heaton, M.P., Leymaster, K.A., Kalbfleisch, T.S., Kijas, J.W., Clarke, S.M., McEwan, J., Maddox, J.F., Basnayake, V., Petrik, D.T., Simpson, B. & Smith, T.P., 2014. SNPs for parentage testing and traceability in globally diverse breeds of sheep. PLoS One. 9, e94851, 1-10.
Hulsegge, B., Calus, M.P.L., Windig, J.J., Hoving-Bolink, A.H., Maurice-van Eijndhoven, M.H.T. & Hiemstra, S.J., 2013. Selection of SNP from 50K and 777K arrays to predict breed of origin in cattle. J. Anim. Sci. 91, 5128-5134.
Kijas, J.W., Lenstra, J.A., Hayes, B., Boitard, S., Neto, L.R.P., San Cristobal, M., Servin, B., McCulloch, R., Whan, V., Gietzen, K. & Paiva, S., 2012. Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. PLoS Biol. 10 (2), e1001258, 1-14.
Kuehn, L.A., Keele, J.W., Bennett, G.L., McDaneld, T.G., Smith, T.P.L., Snelling, W.M., Sonstegard, T.S. & Thallman, R.M., 2011. Predicting breed composition using breed frequencies of 50,000 markers from the US Meat Animal Research Center 2,000 Bull Project. J. Anim. Sci. 89, 1742-1750.
Kugonza, D.R., Nabasirye, M., Mpairwe, D., Hanotte, O. & Okeyo, A.M., 2011. Productivity and morphology of Ankole cattle in three livestock production systems in Uganda. Anim. Genet. Res. 48, 13-22.
Lee, S.H., Choi, B.H., Lim, D., Gondro, C., Cho, Y.M., Dang, C.G., Sharma, A., Jang, G.W., Lee, K.T., Yoon, D. & Lee, H.K., 2013. Genome-wide association study identifies major loci for carcass weight on BTA14 in Hanwoo (Korean cattle). PLoS One. 8, e74677, 1-9.
Lewis, J., Abas, Z., Dadousis, C., Lykidis, D., Paschou, P. & Drineas, P., 2011. Tracing cattle breeds with principal components analysis ancestry informative SNP. PLoS One. 6 (4), 1-8.
Lewontin R.C. & Krakauer J., 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genet. 74, 175-195.
Liu, N., Chen, L., Wang, S., Oh, C. & Zhao, H., 2005. Comparison of single-nucleotide polymorphisms and microsatellites in inference of population structure. BMC Genet. 6, S26, 1-5.
MacEachern, S., Hayes, B., McEwan, J. & Goddard, M., 2009. An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (*Bos taurus*) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genomic diversity in domestic cattle. BMC Genomics. 10, 1-19.
Makina, S.O., Muchadeyi, F.C., van Marle-Köster, E., MacNeil, M.D. & Maiwashe, A., 2014. Genetic diversity and population structure among six cattle breeds in South Africa using a whole genome SNP panel. Front Genet. 5, 1-7.
Mapholi, N.O., Marufu, M.C., Maiwashe, A., Banga, C.B., Muchenje, V., MacNeil, M.D., Chimonyo, M. & Dzama, K., 2014. Towards a genomics approach to tick (Acari: Ixodidae) control in cattle: A review. Ticks Tick Borne Dis. 5 (5), 475-483.
Martinez-Camblor, P., Carleos, C., Baro, J.Á. & Canon, J., 2014. Standard statistical tools for the breed allocation problem. J. Appl. Statist. 41 (8), 1848-1856.
Mason, I.L., 1996. A World Dictionary of Livestock Breeds, Types and Varieties (4th ed.), C.A.B. International, Wallingford, Oxofordshire, UK. ISBN 0-85199-102-5
Matukumalli, L.K., Lawley, C.T., Schnabel, R.D., Taylor, J.F., Allan, M.F., Heaton, M.P., O'Connell, J., Moore, S.S., Smith, T.P., Sonstegard, T.S. & Van Tassell, C.P., 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS One. 4, e5350, 1-11.
McKay, S.D., Schnabel, R.D., Murdoch, B.M., Matukumalli, L.K., Aerts, J., Coppieters, W., Crews, D., Neto, D.E., Gill, C.A., Gao, C. & Mannen, H., 2008. An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. BMC Genet. 9, 1-9.
Michelizzi, V.N., Wu, X., Dodson, M.V., Michal, J.J., Zambrano-Varon, J., McLean, D.J. & Jiang, Z., 2011. A global view of 54,001 single nucleotide polymorphisms (SNPs) on the Illumina BovineSNP50 BeadChip and their transferability to Water Buffalo. Int. J. Biol. Sci. 7, 18-27.
Mwai, O., Hanotte, O., Kwon, Y.J. & Cho, S., 2015. African indigenous cattle: unique genetic resources in a rapidly changing world. Asian Australas. J. Anim. Sci. 28, 911-921.
Negrini, R., Nicoloso, L., Crepaldi, P., Milanesi, E., Colli, L., Chegdani, F., Pariset, L., Dunner, S., Leveziel, H., Williams, J.L. & Ajmone Marsan, P., 2009. Assessing SNP markers for assigning individuals to cattle populations. Anim. Genet. 40, 18-26.

Nishimura, S., Watanabe, T., Ogino, A., Shimizu, K., Morita, M., Sugimoto, Y. & Takasuga, A., 2013. Application of highly differentiated SNPs between Japanese Black and Holstein to a breed assignment test between Japanese Black and $F_1$ (Japanese Black x Holstein) and Holstein. Anim. Sci. J. 84, 1-7.

Okello, S. & Sabiiti, EN., 2006. Milk production of indigenous Ankole cattle in Uganda as influenced by seasonal variations in temperature, rainfall and feed quality. Makerere Univ. Res. J. 1, 73-92.

Opara, A., Razpet, A. & Logar, B., 2012. Breed assignment test of Slovenian cattle breeds using microsatellites. Acta. Agric. Slov. 3, 167-170.

Otten, D. & Van den Weghe, H.F., 2011. The Sustainability of Intensive Livestock Areas (ILAS): Network system and conflict potential from the perspective of animal farmers. Int. J. Food Syst. Dyn. 2, 36-51.

Pariset, L., Mariotti, M., Nardone, A., Soysal, M.I., Ozkan, E., Williams, J.L., Dunner, S., Leveziel, H., Maroti-Agots, A., Bodo, I. & Valentini, A., 2010. Relationships between Podolic cattle breeds assessed by single nucleotide polymorphisms (SNPs) genotyping. J. Anim. Breed. Genet. 127, 481-488.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. & Sham, P.C., 2007. PLINK: A toolset for whole-genome association and population-based linkage analysis. Am. J. Hum. Genet. 81, 559-575.

Ramos, A.M., Megens, H.J., Crooijmans, R.P.M.A., Schook, L.B. & Groenen, M.A.M., 2011. Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing. Anim. Genet. 42 (6), 613-620.

Rege, J.E.O. & Tawah, C.L., 1999. The state of African cattle genetic resources II. Geographical distribution, characteristics and uses of present-day breeds and strains. Anim. Genet. Res. Inf. 26, 1-25.

Renaudeau, D., Collin, A., Yahav, S., De Basilio, V., Gourdine, J.L. & Collier, R.J., 2012. Adaptation to hot climate and strategies to alleviate heat stress in livestock production. Anim. 6, 707-728.

Reynolds, J., Weir, B.S. & Cockerham, C.C., 1983. Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genet. 105, 767-779.

Sanz, A., Martin-Burriel, I., Cons, C., Reta, M., Poblador, A., Rodellar, C. & Zaragoza, P., 2014. Genetic diversity, structure and individual assignment of Casta Navarra cattle: a well-differentiated fighting bull population. J. Anim. Breed. Genet. 131 (1), 11-18.

Scholtz, M.M., 1988. Selection possibilities of hardy beef breeds in Africa: The Nguni example. In 3. Congres Mondial de Reproduction et Selection des Ovins et Bovins a Viande, Paris (France), 19-23 Jun 1988. INRA.

Scholtz, M.M., 2010. Beef breeding in South Africa (2nd ed.). Asikhulume pixArt, Rooihuiskraal, Pretoria, South Africa.

Strydom, P.E., 2008. Do indigenous Southern African cattle breeds have the right genetics for commercial production of quality meat? Meat Sci. 80, 86-93.

Tabangin, M.E., Woo, J.G. & Martin, L.J., 2009. The effect of minor allele frequency on the likelihood of obtaining false positives. BMC Proceedings. 3, (7), S41.

Van Marle, J., 1974. The breeding of beef cattle in South Africa: Past, present and future. S. Afr. J. Anim. Sci. 4, 297-304.

Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C. & Sonstegard, T.S., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Meth. 5, 247-252.

Vasconcellos, L.P.D.M.K., Tambasco-Talhari, D., Pereira, A.P., Coutinho, L.L. & Regitano, L.C.D.A., 2003. Genetic characterization of Aberdeen Angus cattle using molecular markers. Genet. Mol. Biol. 26, 133-137.

Yoon, D., Kwon, Y.S., Lee, K.Y., Jung, W.Y., Sasazaki, S., Mannen, H., Jeon, J.T. & Lee, J.H., 2008. Discrimination of Korean cattle (Hanwoo) using DNA markers derived from SNPs in bovine mitochondrial and SRY genes. Asian Aust. J. Anim. Sci. 21 (1), 25-28.

Weir, B. S., 1996. Genetic Data Analysis II: Methods for discrete population genetic data. Sinauer Associates Inc, Sunderland, MA.

Weir, B.S., & Cockerham, C.C., 1984. Estimating F-statistics for the analysis of population structure. Evol. 38, 1358-1370.

Willing, E. M., Dreyer, C., & Van Oosterhout, C., 2012. Estimates of genetic differentiation measured by $F_{st}$ do not necessarily require large sample sizes when using many SNP markers. PLoS One. 7 (8), e42649, 1-7.

Wilkinson, S., Wiener, P., Archibald, A.L., Law, A., Schnabel, R.D., McKay, S.D., Taylor, J.F. & Ogden, R., 2011. Evaluation of approaches for identifying population informative markers from high density SNP chip. BMC Genet. 12, 1-14.

# CHAPTER FOUR

## SNP discovery in indigenous Afrikaner, Drakensberger and Nguni cattle breeds of South Africa

A. A. Zwane[1,2], E. Van Marle-Köster[2], M.L. Makgahlela[1], A. Maiwashe[1,5] and J.F. Taylor[4]

[1]Department of Animal Breeding and Genetics, ARC-API, P/Bag X2, Irene, 0062,
[2]Department of Animal and Wildlife Sciences, University of Pretoria, P/Bag X20, Hatfield, Pretoria, 0028,
[3]Sydney Brenner Institute of Molecular Bioscience, University of the Witwatersrand, 9 Jubilee Road, Parktown, Johannesburg, 2193,
[4]Division of Animal Sciences, University of Missouri, 920 East Campus Drive, Columbia, MO 65211-5300.
[5]Department of Animal, Wildlife and Grassland Sciences, University of the Free State, Bloemfontein 9300, South Africa

**Prepared for Publication**

# SNP discovery in indigenous Afrikaner, Drakensberger and Nguni cattle breeds of South Africa

**A. A. Zwane[1,2], E. Van Marle-Köster[2], M.L. Makgahlela[1], A. Maiwashe[1,5] and J.F. Taylor[4]**

[1]Department of Animal Breeding and Genetics, ARC-API, P/Bag X2, Irene, 0062, [2]Department of Animal and Wildlife Sciences, University of Pretoria, P/Bag X20, Hatfield, Pretoria, 0028, [3]Sydney Brenner Institute of Molecular Bioscience, University of the Witwatersrand, 9 Jubilee Road, Parktown, Johannesburg, 2193, [4]Division of Animal Sciences, University of Missouri, 920 East Campus Drive, Columbia, MO 65211-5300, [5]Department of Animal, Wildlife and Grassland Sciences, University of the Free State, Bloemfontein 9300, South Africa

## Abstract

Single nucleotide polymorphism arrays have created new possibilities for performing genome-wide studies to detect genomic regions harbouring sequence variants that affect complex traits. However, the majority of validated SNPs for which allele frequencies have been estimated are limited to European breeds. The objective of this study was to search for new SNPs in three indigenous SA breeds (Afrikaner, Drakensberger and Nguni) using next generation sequencing. DNA samples from 30 individuals from each of the three breeds were equimolar pooled and sequenced to identify putative SNPs. Approximately 1.8 billion sequence reads were aligned to the UMD3.1 reference genome generating an average depth of 21-fold sequence coverage for each breed. A total of 15.7 million SNPs were identified across the breeds with the highest number of SNPs identified in Nguni. Verification of SNPs against Run 5 data from the 1000 Bull Genomes project suggested that 16% of the SNPs were novel variants. Annotation of the detected variants indicated numerous variants classified within functional genes that may be associated with complex traits in these cattle breeds. Functional enrichment analysis of novel SNPs identified 1,481 genes enriched for novel variants across the breeds. In total, 461, 478 and 542 genomic regions were enriched for novel variants in AFR, DRA and NGI respectively ($p < 0.001$), identified from the top (5%) of genomic windows. These discoveries provide a valuable genomic resource for studying the genetic composition of these breeds.

**Key words**: indigenous breeds, sequencing, mapping, novel variants, annotation

## Introduction

The development of next generation sequencing (NGS) technologies has enabled the rapid and cost-effective generation of sequence data for SNP discovery in cattle (Le Roex et al., 2012). These developments have also enabled the simultaneous estimation of SNP allele frequencies in a diverse range of reference populations (Van Tassell et al., 2008). Low and high-density SNP genotyping assays are available for performing genome-wide analyses in cattle (Matukumalli et al., 2009). However, while the available assays have been shown to be adequate for studies in European taurine breeds, they are less informative when applied to indicine or indigenous SA breeds (Gurgul et al., 2013; Zwane et al., 2016). Studies using the BovineSNP50 assay on indigenous SA breeds have shown substantially lower levels of linkage disequilibrium (LD) and lower minor allele frequencies (MAF) compared to those obtained in European taurine breeds (Edea et al., 2013; Makina et al., 2014). Furthermore, a study by Makina et al. (2015) using the BovineSNP50 assay for the detection of signatures of selection in indigenous SA breeds, also indicated reduced numbers of informative markers. Again, analysis of these markers showed little evidence for the existence of breed-specific markers in indigenous SA cattle breeds (Zwane et al., 2016).

Consequently, there is a reduced utility for the implementation of these assays for genome-wide association studies (GWAS), quantitative trait locus (QTL) detection or for the identification of genes associated with economically important traits in indigenous SA breeds as observed by Albrechtsen et al. (2010). Therefore, sequencing the genomes of indigenous SA cattle could be beneficial in animal production, in understanding the traits of economic importance, animal health and welfare, and in understanding the genetic basis of diseases. Genome sequencing also presents opportunities for increased knowledge of the evolutionary histories of these breeds (Pool & Waddell, 2002)

NGS technologies have identified a large number of SNPs and insertions-deletions (Indels), with many variants remaining to be detected, especially in cattle breeds that are phylogenetically distinct from the extensively studied European breeds (Choi et al., 2013). More than 60,000 putative SNPs were identified from the sequencing of reduced representation DNA libraries generated for 66 cattle from three populations (Van Tassell et al., 2008). More than 2 million novel SNPs were discovered from resequencing of a Fleckvieh bull (Eck et al. 2009). Furthermore, Kawahara-Miki et al. (2011) re-sequenced the genome of a single

Kuchinoshima-Ushi (Japanese native cattle) bull and identified 6.3 million SNPs, of which more than 5.5 million (87%) were novel. Choi et al. (2014) reported a total of 10.4 million SNPs identified in Korean Hanwoo, Jeju Heugu and Holstein cattle, and found 54.12% novel SNPs as well as detected 1,063,267 Indels in these genomes. This indicates that NGS technologies are effective for SNP discovery projects and can also be applied to variant discovery in indigenous South African (SA) cattle.

The use of sequence data for variant discovery and genotyping has the advantage of less SNP ascertainment bias compared to the use of commercially available SNP assays (Nielsen et al., 2011). SNP ascertainment bias influences the extent to which polymorphisms are shared across populations due to the distribution of allele frequencies within studied populations that may result in biases in measures of genetic differentiation, e.g., $F_{st}$ estimates between populations and also affects the weighting of principal components, which in turn, can affect inferences about admixture in populations (McTavish & Hillis, 2015). Consequently, the sequencing of indigenous SA cattle genomes presents the potential to discover new SNPs for inclusion in existing SNP assays or for developing custom-made SNP chips for local SA populations. This information can also improve the accuracy of inferences made in population studies and the genome-wide detection of genes associated with complex traits such as disease resistance (Pool et al., 2010). It also holds potential for the identification of breed informative SNPs for breed assignment in SA populations (Ramos et al., 2011).

To date, limited sequence data have been generated for indigenous SA cattle breeds. Breeds such as Brahman, Afrikaner and Tuli (African indicine), representing Australian populations, have been sequenced and analysed resulting in 3.56 million new SNPs being submitted to dbSNP (Barris et al., 2012). The objective of this study was to search for novel SNPs in three indigenous SA breeds (i.e., Afrikaner (AFR), Drakensberger (DRA) and Nguni (NGI)) by sequencing pooled DNA samples using next generation sequencing.

## Materials and Methods

### Pedigree analyses and sample identification

The available pedigree data for each breed were obtained from the Agricultural Research Council (ARC) Integrated Registration and Genetic Information System (INTERGIS) database. Pedigree analysis of Afrikaner (n=251,964), Drakensberger (n=198,237) and Nguni

(n=241,491) were performed within breed to identify the least related individuals in these populations. Relationship coefficients between individuals were estimated using the method of Meuwissen and Luo (1992) implemented in the PEDIG software (Boichard, 2002), where males born between 2006 and 2012 were considered to be the reference population. In total, 90 least related animals across breeds (i.e., 30 animals per breed) with average relationship coefficients of 0.006, 0.008 and 0.0008 for Afrikaner, Drakensberger and Nguni, were selected across all nine SA provinces for sequencing to span the cattle's genetic diversity, and breeder's consent was obtained from the animal owners.

**Sample collection, library construction and DNA sequencing**

Sampling of blood and hair was performed with the approval of the Animal Ethics Committee of the University of Pretoria (EC: S4285-15), according to guidelines for the proper handling of animals during sample collection. Genomic DNA was extracted from whole blood (200 µl/sample) using the Roche DNA extraction Kit (Roche, Germany) following the standard protocol of the manufacturer. The procedure included a proteinase K digestion followed by column purification for the extraction of high quality DNA. The extraction of DNA from hair roots was performed using an optimized Phenol-Chloroform protocol (Sambrook & Russell, 2006), that included a Proteinase K and Dithiothreitol digestion followed by phenol-chloroform extraction and centrifugal dialysis with Centricon concentrators (Slikas et al., 2000). The quality of the extracted DNA samples was assessed using a Nanodrop UV/Vis Spectrophotometer (Nanodrop ND-1000) and verified using a Qubit® 2.0 Fluorometer (Thermo Scientific). All DNA samples were maintained at a concentration of 50 ng/µl in preparation for NGS sequencing at the ARC Biotechnology Platform.

Equimolar DNA pools were prepared for each breed using 170 ng of DNA per animal, and each DNA pool contained 30 animals per breed. Genomic libraries were prepared with the Paired-end Sequencing Sample Preparation Kit (Illumina, San Diego, CA) using 5 µg of genomic DNA according to the manufacturer's instructions. DNA was fragmented using a Covaris M220 sonicator, end-repaired and A-tailed followed by the ligation of adapters (Nextera Transposase, Illumina) and 12 cycles of polymerase chain reaction (PCR) were performed. The average fragment size for each library was 350 bp. Quantities and the quality of usable material for each of the libraries were estimated by qPCR (KAPA Library Quantification Kit–Illumina Genome Analyzer-SYBR Fast Universal). The automated cBot

Cluster Generation System (Illumina, San Diego, Calif, USA) was used to generate clusters on the flow cell. Each DNA pool was then sequenced (paired-end; read length 125 bp) in a single lane of a flow cell using the Illumina HiSeq 2000 to a target of 30X coverage. The resulting images were analyzed with the HiSeq Pipeline Software v2.0 (Illumina) to generate the raw fastq files (Van Tassell et al., 2008; Ramos et al., 2009; Van et al., 2013; Boutet et al., 2016).

Sequence reads were filtered for base quality using Trimmomatic (Bolger et al., 2014). Reads were trimmed if four consecutive bases had an average Phred-like quality score of less than 20. PCR duplicates were removed using Picard (Li et al., 2009) since these should not be counted as evidence for or against putative variants or for allele frequency estimation (Auwera, 2013). Pairs of DNA sequences for which each read exceeded 35 bp were retained for analysis. Sequence reads were aligned to the *Bos taurus* reference genome (UMD3.1) using the Burrows-Wheeler aligner (BWA), a software package for mapping lowly-divergent sequences against a large reference genome (Li et al., 2009). The alignments were sorted and converted to the BAM format using SAMtools v1.2 (Ramirez-Gonzalez et al., 2012). Data were then formatted for variant calling using Picard tools, by marking duplicate reads (Li et al., 2009) which were ignored by the Genome Analysis Tool Kit (GATK) during variant calling.

**Variant discovery, annotation and functional enrichment analysis**

Variant discovery was performed within breed according to GATK Best Practices using the genomic variant call format (GVCF) workflow (Auwera, 2013). The workflow includes data pre-processing steps and calling variants separately for each population using a command that is specific for paired-end data. The pre-processing steps include realigner target creator to generate intervals for each chromosome for Indel realignment, depth of coverage estimation for each chromosome, base recalibration, analyzing covariates/variables and printing reads. Genotype calling was performed separately for each chromosome to generate GVCF files for variant calling. The workflow included a joint analysis step that empowers variant discovery by providing the ability to leverage population-wide information from a cohort of samples, allowing the detection of variants with greater sensitivity and genotyping samples as accurately as possible (GATK Best Practices; Bareke et al., 2013). Cohorts of variants were generated in VCF files, and the genotypes were called for each breed with a minimum genotype quality of 20, and a read depth of between 1 and 25 (Aslam et al., 2012). To reduce the false discovery rate, hard filtering steps were conducted using the following criteria: Phred scaled

polymorphism probability (QUAL) < 30.0, variant confidence normalized by depth (QD) < 2.0, mapping quality (MQ) < 40.0, strand bias (FS) > 60.0, HaplotypeScore > 13.0, MQRankSum < −12.5, and ReadPosRank-Sum < −8.0 (GATK Best Practices; Choi et al., 2015). All SNPs that passed these criteria were consequently categorized into fixed (homozygous non-reference assembly nucleotide genotypes called in all individuals within the breed) or segregating (variable/heterozygous genotypes identified in the breed) (Aslam et al., 2012).

Minor allele frequencies were estimated for each SNP by directly counting the number of reads representing each allele using PLINK (Purcell et al., 2007; Ramos et al., 2009). Ratios of fixed to segregating SNPs were estimated within each of the populations using PLINK. The transition-to-transversion (Ti/Tv) ratio for each SNP call was calculated for each population as an indicator of potential sequencing errors (Choi et al., 2015) using VCFtools (Danecek et al., 2011). This is the ratio of the number of transitions (interchanges of either purines, A<->G or pyrimidines, C<->T) to the number of transversions (interchanges of purine for pyrimidine bases), for a pair of DNA sequences (Mitchell, 2015).

SNP annotation and the functional consequences of sequence variants were predicted using the Variant Effect Predictor (VEP) tool, Ensembl and dbSNP (Huang et al., 2009; McLaren et al., 2010). For all input variants, VEP provides detailed annotations for transcripts, proteins, and regulatory regions, and also provides phenotype information for known variants (McLaren et al., 2016). The functional effects of each SNP were estimated, and all SNPs were assigned with a diverse range of functional categories based on genomic coordinates, functional class, codon change, gene name, transcript biotype, gene coding, transcript ID, exon rank and corresponding genotype (Choi et al., 2015). Annotation results were downloaded for further downstream analysis. The identified variants were verified by using data from European taurine or indicine breeds that were available from Run5 of the 1000 Bull Genomes Project (July 2015), consisting of 1,682 sequenced animals and 60,223,042 million variants (Daetwyler et al., 2014).

**Identification of novel SNPs**

Identified novel SNPs were further examined to determine their distribution throughout the genome, identify regions enriched for novel SNPs and identify the genes that were associated with novel SNP enriched regions. The files containing novel SNPs identified in each breed

were first compared to identify SNPs that were common between breeds. An in-house developed script was used to parse SNPs that were predicted to be unique to each breed, and these were used for further analysis. A file containing the union set of SNPs identified within breeds (including common SNPs) was used to characterize the percentage of novel SNPs within 100 kb sliding windows throughout the autosomal genome. SNP distributions were then computed using the package "qqman" in the R environment (Turner, 2014). All windows were annotated with the Ensembl Cow database (www.ensembl.org/). If genes were found in the window, the corresponding gene names were provided for each SNP. The comparison of the observed and expected number of SNPs assuming a random SNP distribution throughout the genome was made using the hypergeometric test to generate P-values (Rivals et al., 2007). To correct for multiple testing, a Bonferroni correction was used and the P-value was multiplied by the number of performed scans (Klein et al., 2009).

## Results and Discussion

### Sequencing and Mapping

Sequencing of AFR, DRA and NGI generated approximately 1.8 billion (184 Gb) of high quality paired-end reads using an Illumina HiSeq 2000 sequencer, of which 99 % of the reads were mapped to the bovine reference genome (UMD 3.1). PCR duplicates were removed and reads were realigned around insertion and deletion events resulting in approximately 1.7 billion sequence reads (90.2 %) across the three breeds, with an average coverage of 21.1-fold across the reference genome (Table 1). The sequence depth observed in this study was similar to studies by Eck et al. (2009) and Stothard et al. (2011) but higher than in the study by Choi et al. (2015) with an average coverage of 10.71X for Hanwoo and Yanbian cattle, but lower than the 27X mean coverage obtained by Das et al. (2015) for Danish Holstein dairy cattle.

The Ti/Tv ratio and heterozygous/homozygous variant ratios have commonly been computed in genetic studies as a quality control measure for sequence data. These ratios are helpful for understanding patterns of DNA sequence evolution (Wang et al., 2014). To evaluate the quality

**Table 1:** Sequencing results for indigenous Afrikaner (AFR), Drakensberger (DRA) and Nguni (NGI) cattle breeds.

| Breed | Animals pooled | Raw Reads | Non-duplicated Reads | Properly Paired Reads | Mapped Reads | High Quality Mapped Reads | Average Coverage |
|---|---|---|---|---|---|---|---|
| AFR | 30 | 537,681,018 | 518,717,587 | 500,986,036 | 536,215,468 | 424,043,570 | 21.2X |
| DRA | 30 | 540,797,394 | 498,063,449 | 502,707,076 | 537,486,252 | 385,388,748 | 15.4X |
| NGI | 30 | 682,407,201 | 646,078,421 | 640,580,750 | 680,935,451 | 528,151,411 | 26.6.X |
| **Total** | **90** | **1,760,885,613** | **1,662,859,457** | **1,644,273,862** | **1,754,637,171** | **1,337,583,729** | **21.1X** |

of the detected SNPs, the Ti/Tv ratio was computed and found to be similar for each breed (AFR:2.20, DRA:2.23, NGI:2.22). These results are similar to the studies of Gayal, Red Angus and Japanese Black cattle where the Ti/Tv values were 2.32, 2.17 and 2.18, respectively (Mei et al., 2016). These results suggest that the majority of SNPs identified in this study were accurately identified (Choi et al., 2015). Since the Ti/Tv ratio is a measure of the nature of sequence changes within a population, it accounts for intra-species variation, and therefore, differs from species to species, among populations and individuals of the same species.

**Variant Detection**

A total of 17.6 million variants were identified in the three studied breeds with the greatest number of variants in NGI and AFR and lowest in DRA (Table 2). The detected variants comprised 89 % SNPs and 11 % Indels. DNA sequence variation is primarily comprised of SNPs and Indels. These variants are mostly intergenic, but include mutations in coding or regulatory regions of transcribed sequences, potentially related to phenotypic traits, and include polymorphisms that can be used as markers for genetic association studies and the fine mapping of candidate regions based on linkage disequilibrium (Weckx et al., 2005). Sequencing of individuals can identify millions of SNPs that differ between any two individual genomes (Bischoff et al., 2008). These results also hold the potential for identifying new SNPs that are unique to indigenous SA breeds.

**Table 2: S**ummary of SNPs and Indels identified in Afrikaner (AFR), Drakensberger (DRA) and Nguni (NGI).

| | | SNPs | | Indels | |
|---|---|---|---|---|---|
| **Breed** | No. Variants | No. SNPs | Proportion SNPs | No. Indels | Proportion indels |
| **AFR** | 11,165,172 | 9,950,392 | 0.89 | 1,212,231 | 0.11 |
| **DRA** | 7,049,802 | 6,327,523 | 0.90 | 721,628 | 0.10 |
| **NGI** | 12,514,597 | 11,164,422 | 0.89 | 1,347,215 | 0.11 |
| Total | **17,647,583** | **15,723, 684** | **0.89** | **1,908,137** | **0.11** |

Variants can also be used to identify selective sweep regions that occur during strong selection events and also to identify breed-specific SNPs to differentiate among the breeds. In addition, these variants can be used to study disease susceptibilities, to determine structural effects on protein sequences, and to design association studies aimed at clarifying complex, polygenic phenotypes (Bischoff et al., 2008).

The ratio of homozygous to heterozygous SNPs within each breed was 1:4.3 (1,881,400:8,068,992), 1:4.7 (1,104,006:5,223,517), and 1:7.8 (1,265,926: 9,898,596 ) for the DNA pools for AFR, DRA and NGI, respectively. The heterozygosity in NGI indicates a larger genetic variation in NGI cattle, likely due to a larger effective population size, the presence of different ecotypes, historic admixture and the genetic distance that exists between the three breeds and that of the Hereford reference genome used in this study. The study by Sanarana et al. (2016) reported levels of genetic differentiation between NGI ecotypes based on microsatellite markers, but the levels were relatively low. Both the AFR and DRA populations have been subjected to artificial selection for specific traits of economic importance for many decades (Abin et al., 2016) which has contributed to shaping the patterns of variation in their genomes.

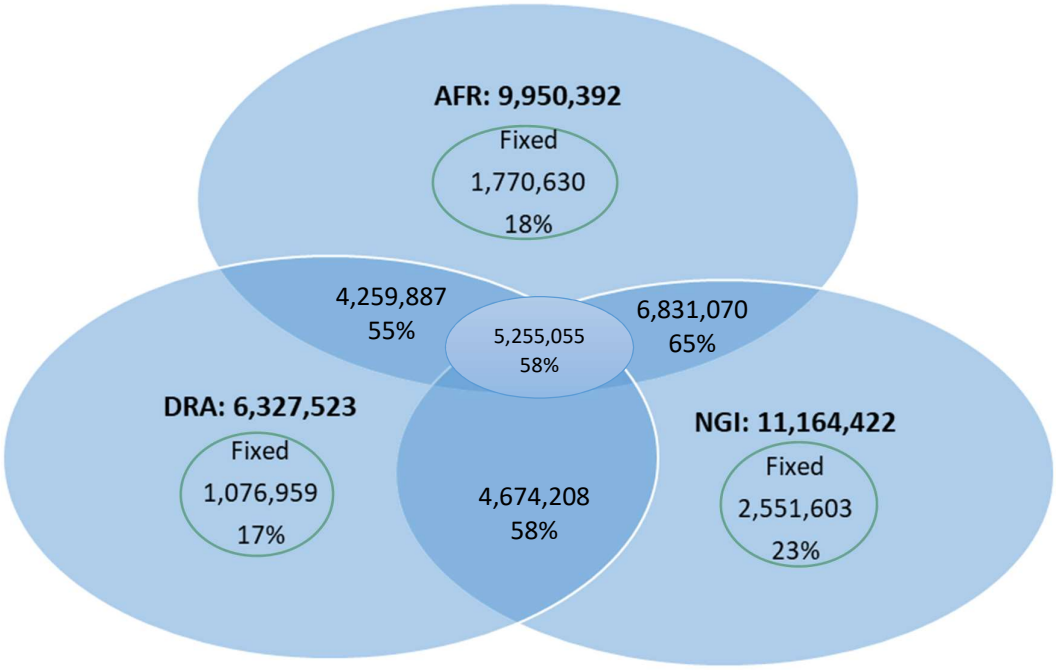


**Figure 1**: The number of SNPs shared and fixed among the three indigenous South African breeds.

From the total number of identified SNPs, on average, 58% of the SNPs were shared among the three cattle populations (Figure 1) with the highest number of SNPs shared between AFR and NGI. The numbers of shared SNPs reflect the potential common ancestries that exist between these cattle population (Dadi et al., 2011; Wang et al., 2016). Due to the history of human migration and trading, it is expected that indigenous breeds will often have multiple genetic signatures of origin and admixture, and this has been confirmed by analyses using available molecular data (Hanotte & Jianlin, 2006; Makina et al., 2014; Decker et al., 2014). These analyses have suggested that several ancestral lineages have contributed to today's genetic pool of livestock (Hanotte et al., 2000; Xuebin, 2004).

**Validation of SNPs using 1000 Bull Genomes Project data**

Run 5 of the 1000 Bull Genomes Project (July 2015) was used to validate SNPs in the three SA breeds that are in common with other cattle breeds worldwide (i.e., *Bos taurus* and African indicine) (Daetwyler et al., 2014). On average, 85 % of all SNPs identified in the three SA indigenous breeds were also shared among the breeds represented in the 1000 Bull Genomes Project data (Table 3). The remaining 16 % of SNPs appear to be unique to SA indigenous breeds, AFR (18 %), DRA (13 %) and NGI (16 %). This proportion was lower than that reported by Choi et al. (2013) where 29.4 % of SNPs were found to be novel in Korean Black Cattle when compared to the dbSNP version 137. This likely reflects the large number of SNPs that have now been discovered in the 1000 Bull Genomes Project (60 million in Run 5).

In the study of Mei et al. (2016), 62.24 % novel SNPs were identified for Gayal cattle, which is much higher than was found in this study reflecting different SNP filtering criteria and the increased divergence of Gayal cattle from the reference genome, Hereford, relative to the SA breeds. The detected SNPs were validated using dbSNP Build 140, which also represents a smaller validation set than was used in this study. The greater number of novel SNPs found in NGI and AFR cattle likely reflects the extent of genetic diversity that exist between these breeds and also their phylogenetic distance from their reference genome. Novel variants characterize the extent of genetic differentiation that exists between individuals and populations (Choudhury et al., 2014). The lower number of novel SNPs found in DRA suggests that the breed might be more closely related to European breeds than the AFR or NGI (Zwane et al., 2016).

**Table 3:** Novel variants identified in the three breeds through comparison to 1000 Bull Genomes Project data.

| | All Variants | | | | SNPs | | | |
|---|---|---|---|---|---|---|---|---|
| **Breed** | **Known** | **Novel** | **Total** | **Proportion Novel variants** | **Known** | **Novel** | **Total** | **Proportion Novel SNPs** |
| AFR | 9,407,874 | 1,757,298 | 11,165,172 | 0.16 | 9,775,327 | 1,751,065 | 9,950,392 | 0.18 |
| DRA | 6,223,599 | 826,203 | 7,049,802 | 0.12 | 5,503,526 | 823,997 | 6,327,523 | 0.13 |
| NGI | 10,723,472 | 1,791,125 | 12,514,597 | 0.14 | 9,369,605 | 1,794,817 | 11,164,422 | 0.16 |
| **Total** | **26,354,945** | **4,374,626** | **30,729,571** | **0.14 (Av)** | **24,648,458** | **4,369,879** | **28,504,873** | **0.16 (Av)** |

The complex origins of cattle are associated with both natural and artificial selection, and gave rise to numerous different breeds displaying a broad spectrum of phenotypes. This happened after the global partitioning of the world-wide cattle genetic diversity into three distinct events, two of which involved domestication, and that resulted in European taurines, West African taurines and Zebu from India spreading all over the world through the migration of different tribes (Gautier et al., 2010; Decker et al., 2014). Figure 2 shows the distribution of variants per chromosome and reveals the extent of variation that exists between the breeds.

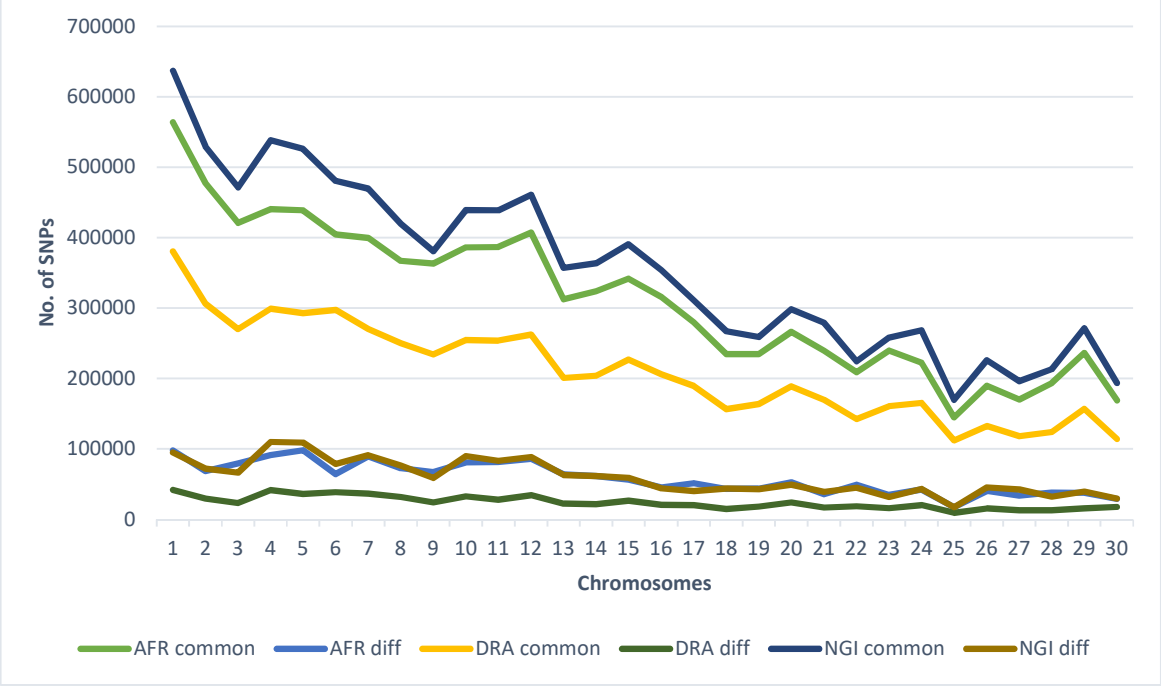


**Figure 2**: Variants shared with breeds represented in the 1000 Bull Genomes project (top three lines) and variants unique to indigenous Afrikaner (AFR), Drakensberger (DRA) and Nguni (NGI) cattle of SA (bottom three lines) by chromosome (X= 30).

**SNP annotation and analysis of functional enrichment**

SNP annotation using VEP Ensembl gene annotation and dbSNP indicated that 62% of the SNPs were located in intergenic regions (AFR:62%, DRA:61%, NGI:62%), 29% were located in genic regions including introns, splice sites, exons and untranslated regions. Fewer SNPs (9%) were located in upstream or downstream regions (transcription start and termination sites) as indicated in Figure 3.

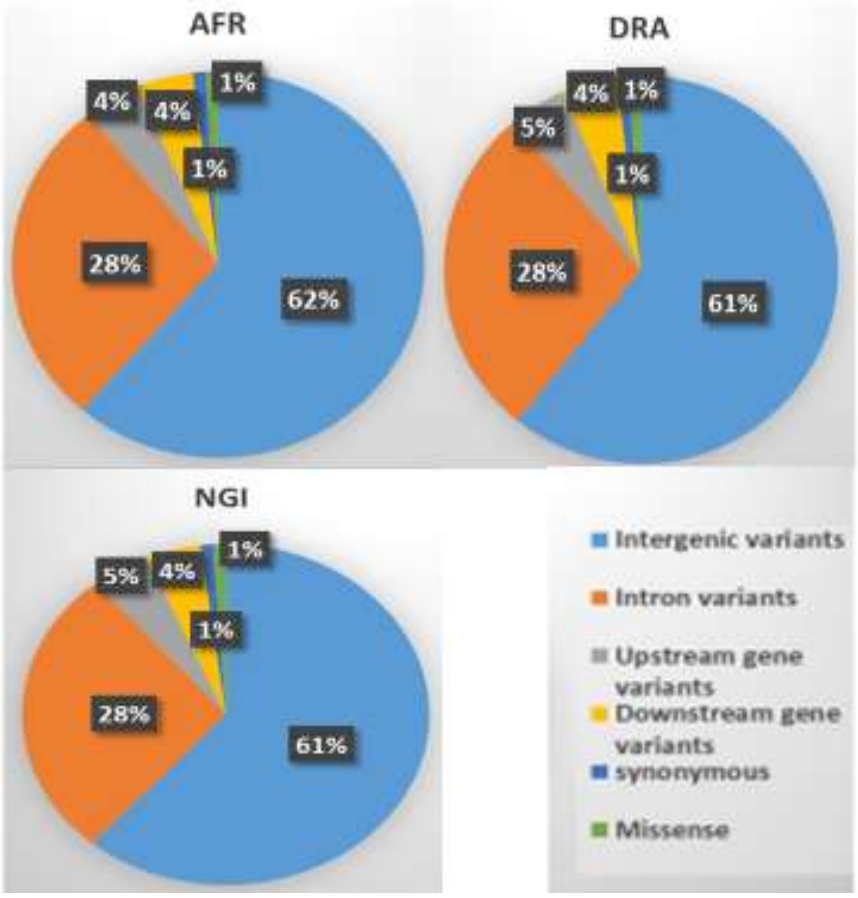


**Figure 3**: Functional classification of variants by breed in Afrikaner (AFR), Drakensberger (DRA) and Nguni (NGI).

Genetic variation in most complex quantitative traits is the result of many mutations of small effects that individually explain only a very small proportion of the genetic variance (Koufariotis et al., 2014). The identification of functional variants such as missense variants, and variants within upstream and downstream genic regions in indigenous SA cattle will enable the testing of these variants for their effects on complex traits (Koufariotis et al., 2014). While the roles of variation in overlapping genes is less clear, studies have suggested that this could be a mechanism allowing the regulation of key genes in eukaryotes (Kim et al., 2009). Further studies of overlapping genes will enable an understanding of the tissue- and developmental-stage regulation of each strand and will provide insight into their mechanisms of evolution (Nakayama et al., 2007). Genetic variants such as insertions, deletions and structural variants can also be tested for association with traits or used in genomic prediction (Koufariotis et al., 2014).

Tables 4 and 5 indicate the distribution of SNPs and Indels detected within each functional class within genic regions. Of the total number of Indels, 61% were located in intergenic regions, 28% in genic regions including introns, exons and splice sites, and 1% were located in up/downstream regions [i.e., 5' and 3'untranslated regions), relative to Figure 3. In AFR there were 433,495 (4.4 %) SNPs located within 5 kb upstream of a transcription start site and 437,355 (4.4%) SNPs within 5 kb downstream of a transcription stop site; 3,974 (0.04%) SNPs were located in a 5' UTR and 18,999 (0.2%) in a 3' UTR. These totals were slightly different in other two breeds, but were slightly lower in NGI.

**Table 4**: Counts of SNPs within each functional class for gene regions.

| **SNP Class** | **Count** | | | | | | **Total** |
|---|---|---|---|---|---|---|---|
| | **AFR** | **%** | **DRA** | **%** | **NGI** | **%** | |
| **Downstream** | 437,355 | 4.4 | 288,515 | 4.6 | 440,357 | 3.9 | 1,166,227 |
| **Stop_lost** | 318 | 0.003 | 200 | 0.003 | 350 | 0.003 | 868 |
| **Stop_gain** | 38 | 0.0004 | 22 | 0.0003 | 15 | 0.0001 | 75 |
| **Splice_site** | 7,650 | 0.08 | 5,305 | 0.008 | 7,553 | 0.07 | 20,508 |
| **Upstream** | 433,495 | 4.4 | 435,935 | 6.9 | 435,955 | 3.9 | 1,305,385 |
| **Intronic** | 2,726,502 | 27.4 | 1,800,155 | 28.4 | 2,731,530 | 24.5 | 7,258,187 |
| **miRNA** | 32,911 | 0.33 | 21,670 | 0.34 | 33,544 | 0.3 | 88,125 |
| **Synonymous_ coding** | 38,537 | 0.4 | 29,836 | 0.47 | 40,694 | 0.36 | 109,067 |
| **Nonsynonymous_ coding** | 31,205 | 0.31 | 22,395 | 0.35 | 31,130 | 0.28 | 84,730 |
| **3'_UTR** | 18,999 | 0.2 | 13,163 | 0.21 | 18,968 | 1.7 | 51,130 |
| **5'_UTR** | 3,974 | 0.04 | 3,055 | 0.05 | 3,805 | 0.034 | 10,834 |
| **Within_non_ coding_gene** | 8,561 | 0.09 | 5,608 | 0.09 | 8,725 | 0.08 | 22,894 |
| **Essential_splice_ site** | 182 | 0.002 | 124 | 0.002 | 192 | 0.002 | 498 |
| **Total** | **3,739,545** | **37.6** | **2,625,859** | **41.5** | **3,752,626** | **33.6** | **10,033,300** |

**Table 5:** Counts of Indels by functional class for gene regions.

| Indel Class | Count | | | | | | Total |
|---|---|---|---|---|---|---|---|
| | AFR | % | DRA | % | NGI | % | |
| **Downstream** | 126,159 | 10.4 | 73,669 | 10.2 | 50,823 | 3.8 | 250,651 |
| **Stop_lost** | 43 | 0.004 | 49 | 0.007 | 26 | 0.002 | 118 |
| **Stop_gain** | 82 | 0.007 | 115 | 0.016 | 34 | 0.003 | 231 |
| **Splice_site** | 2481 | 0.2 | 1667 | 0.23 | 952 | 0.007 | 5,100 |
| **Upstream** | 123,341 | 10.2 | 71,747 | 10.4 | 48,080 | 3.6 | 243,168 |
| **Intronic** | 745,500 | 61.5 | 431,225 | 59.8 | 317,114 | 23.5 | 1,493,839 |
| **miRNA** | 10,296 | 0.85 | 5,644 | 0.8 | 3,816 | 0.28 | 19,756 |
| **Synonymous_coding** | 1,004 | 0.08 | 855 | 0.12 | 449 | 0.33 | 2,308 |
| **Nonsynonymous_coding** | 2,943 | 0.24 | 2,293 | 0.32 | 1,145 | 0.008 | 6,381 |
| **3'_UTR** | 5,574 | 0.46 | 3,165 | 0.44 | 2,166 | 0.16 | 10,905 |
| **5'_UTR** | 842 | 0.07 | 660 | 0.01 | 376 | 0.028 | 1,878 |
| **Within_non_coding_gene** | 2,141 | 0.18 | 1,311 | 0.18 | 545 | 0.04 | 3,997 |
| **Total** | **1,020,406** | **84.1** | **592,400** | **82.1** | **425,526** | **31.6** | **2,038,332** |

A total of 20,508 SNPs across the three breeds were located in splice sites, and 498 SNPs were in splice/donor sites. A total of 109,067 nsSNPs substitutions were observed. These numbers were higher than found by Stothard et al. (2011) in Holstein and Black Angus, and Choi et al. (2013) in Heugu cattle. There were 868 SNPs predicted to cause premature stop codons and 75 to cause gains in coding sequence across the breeds. Variants characterized as nsSNP in coding genes included 84,730 in coding exons and 88,125 in miRNAs across the three breeds. The number of functional genes differs depending on the breeds and the method used for functional annotation (Das et al., 2011). The number of functionally annotated Indels was slightly higher than the number of detected Indel loci, because a SNP or Indel locus may have multiple annotations (Choi et al., 2015). The numbers of SNPs and Indels identified in this study were slightly greater in NGI and AFR than in DRA due to the higher indicine percentage present in their genomes (Makina et al., 2016).

The numbers of nsSNPs segregating in these breeds were greater than for Danish Jutland Cattle which had 34,257 non-synonymous substitutions (34,183 missense and 74 initiator codon variants) were identified. (Das et al., 2011). Non-synonymous SNPs are 'neutral' if the

function of the resulting point-mutated protein is not visible to the mutant, and are 'non-neutral' if the function of the resulting point-mutated protein is visible. Therefore, the ability to identify non-neutral substitutions could help targeting diseases caused by detrimental mutations, and SNPs that increase the fitness of particular phenotypes (Bromberg & Rost, 2007). In human, among all types of variants, nsSNPs are believed to be the major contributors to heritable diseases. They constitute more than half of the disease-causing genetic changes deposited in the Human Gene Mutation Database (HGMD) (Stenson et al., 2009). Therefore, further analysis of these SNPs will assist in determining the genetic changes that contribute to major diseases and phenotypes in cattle.

**Novel SNP enrichment and gene annotation**

In Figure 4, the proportion of novel SNPs occurring in 100 kb windows throughout the genome in each of the three breeds are shown. The figure shows regions that are enriched for novel SNPs throughout the genome and that these regions are breed specific. AFR and NGI possess greater numbers of novel SNP enriched regions, with the greatest differentiation on chromosomes 3 and 22 in AFR and chromosomes 8 and 18 in NGI. The DRA had fewer regions enriched for novel SNPs consistent with the lower overall diversity detected for this breed.

More than 8,237 genes were located within the 1,481 100 kb windows that were enriched for novel SNPs across the breeds. In total, 461, 478 and 542 genomic regions were enriched for novel variants in AFR, DRA and NGI respectively ($p < 0.001$), identified from the top (5%) of windows. These genes were annotated using Ensembl gene annotator (www.ensembl.org) to identify gene ontology terms associated with genes in regions enriched for novel variation in indigenous SA breeds. Most of these genes were protein-coding and regulate biochemical processes, phenotypic characteristics and disease-related phenotypes in human and other model organisms including mouse and zebra fish (Cieslak et al., 2011). These are the genes that might have been subject to natural or artificial selection due to their effects on phenotypic variation (Cieslak et al., 2011).

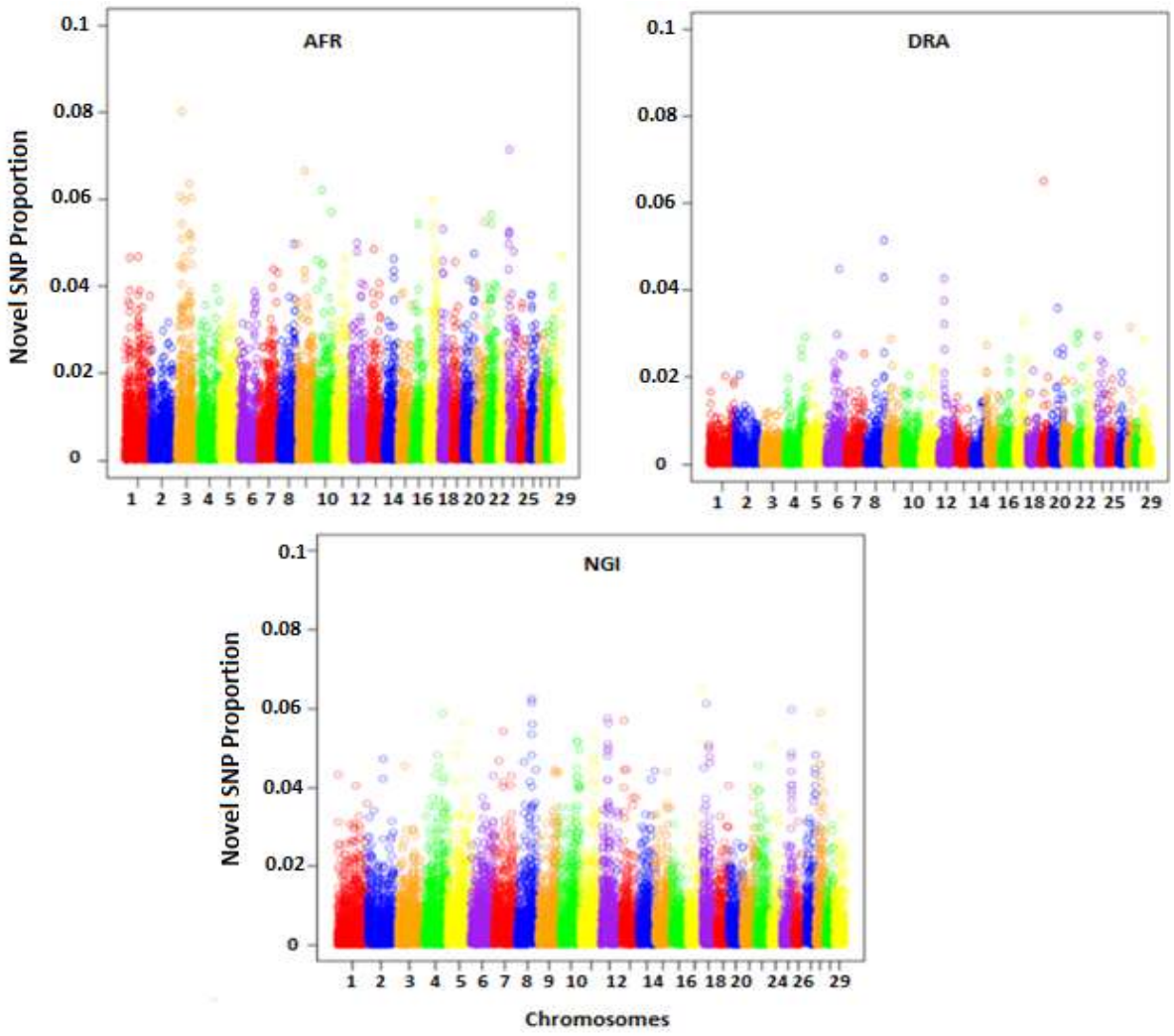


**Figure 4:** Distribution of novel SNP enriched regions across the genome for Afrikaner (AFR), Drakensberger (DRA) and Nguni (NGI) cattle.

The melanin synthesis gene (*MLANA*) on chromosome 8 is located in a region that is enriched for novel SNPs in AFR, is associated with coat colour in mouse (Table 6) and pigmentation in human (Sturm, 2009). In NGI, *SYT10* on chromosome 5 has been associated with sleep disorders and the sense of smell in mouse (Bahbahani et al., 2015) and in cattle has been associated with longevity in Fleckvieh bulls (Meszaros et al., 2014). Since fertility is one of the most important fitness traits contributing to the culling of animals, the longevity of cattle is highly influenced by their reproductive performance (Meszaros et al., 2014). Genes such as *SNTG1* (identified in AFR and NGI) and *ADAMS3* (identified in AFR and DRA) have also

**Table 6:** List of genes within SNP enriched genomic regions in the top 100 kb window.

**Afrikaner**

| Gene | CHR | Function | Species | References |
|---|---|---|---|---|
| *MOV10* | 3 | gene silencing by miRNA | human | Goodier et al., 2012 |
| *MPV17* | 11 | abnormal coat/hair pigmentation, thin skin, decreased body weight, kidney failure, anemia, hypertension, increased heart rate | mouse | Weiher et al., 1990; Viscomi et al., 2009 |
| *UCN* | 11 | increased anxiety, feeding behavior, heart failure, decreased drinking behavior, parkinsonian disorders | mouse, rat | Vetter et al., 2002 |
| *TRIM54* | 11 | premature death, abnormal heart morphology | mouse | Hwang et al., 2010 |
| *DNAJC5G* | 11 | cardiovascular system phenotype, decreased anxiety-related response | mouse | Rovelet-Lecrux et al., 2012 |
| *WNT4* | 2 | serkal syndrome, female sex determination, kidney failure, male sex differentiation | mammals, mouse | Vainio et al., 1999; Brisken et al., 2000 |
| *CDC42* | 2 | negative regulation of gene expression, hair follicle placode formation, spinal cord injuries, bipolar disorder, epilepsy arthritis | mouse, rat | Erschbamer et al., 2005; Park et al., 2009 |
| *MLANA* | 8 | diluted coat color, hair morphology | mouse | Steingrimsson et al., 2006 |
| *KIAA1549* | 4 | decreased total body fat amount, pilocytic astrocytoma (brain tumor) | human, mouse | Hughes, 1998; Antonelli et al., 2015 |
| *HECTD3* | 3 | decreased lean body mass, length, increased total body fat amount | mouse | Zhang et al., 2009 |
| **Drakensberger** | | | | |
| *YTHDC2* | 10 | prostatic neoplasms | rat | Arambula et al., 2016 |
| *DCLRE1B* | 3 | decreased embryo size, neonatal lethality, cell cycle checkpoint | mouse, human | Liu et al., 2009; Dronkert et al., 2000 |
| *AP4B1* | 3 | spastic paraplegia , autosomal recessive | mouse, human | Tuysuz et al., 2014 |
| *PTPN22* | 3 | autoimmune diseases, enlarged spleen, diabetes mellitus, insulin-dependent | human, mouse, rat | Bottini et al., 2006; Michou et al., 2007 |

| | | | | |
|---|---|---|---|---|
| *ZC3HAV1* | 4 | suppression by virus of host molecular function, endosome to lysosome transport | mouse | Lee et al., 2009 |
| *PSMB11* | 10 | increased T-cell proliferation, abnormal self-tolerance | mouse | Anderson & Takahama., 2012 |
| *AJUBA* | 10 | gene silencing by miRNA, wound healing, spreading of epidermal cells, heart contraction, decreased rate, abnormal cell migration | human, zebrafish, mouse | Bergantinos et al., 2010; Wilkinson et al., 2014 |
| *SLC7A8* | 10 | decreased susceptibility to pharmacologically induced seizures | mouse | Dai et al., 2007 |
| *IFT74* | 8 | abnormal lung lobe morphology, notch signaling involved in heart development, cilium assembly | human, mouse | Bhogaraju et al., 2006; Kwong et al., 2007 |
| *SUPT7L* | 11 | abnormal hair texture, decreased body weight, embryonic lethality | mouse | Bardot et al., 2016 |

**Nguni**

| | | | | |
|---|---|---|---|---|
| *RAB33B* | 17 | skeletal system morphogenesis | human | Bonafe et al., 2015 |
| *SYT10* | 5 | shortened circadian period (sleep disorder), sensory perception of smell | mouse | de Anda et al., 2016 |
| *STT3B* | 22 | congenital disorder of glycosylation | human | Scott et al., 2014 |
| *CEACAM16* | 18 | deafness, autosomal dominant 4b | human, mouse | Zheng et al., 2011, Lukashkin et al., 2012 |
| *SRGAP2* | 16 | dendritic spine development | mouse | Charrier et al., 2012 |
| *TMEM98* | 19 | nanophthalmia, hemorrhage | human, mouse | Liao et al., 2016 |
| *CCL17* | 18 | staphylococcal pneumonia, bronchiolitis obliterans | mouse | Montgomery & Daum 2009 |
| *TXN* | 8 | fatty liver, myocarditis, diabetes mellitus | rat | Chung et al., 2011 |
| *COG5* | 4 | congenital disorder | human | Wu et al., 2004 |
| *AIRE* | 1 | reduced fertility, thyroid & eye inflammation | mouse | Schaller et al., 2008 |

been associated with fertility in cattle (Meszaros et al., 2014). Shugoshin 2 (*SGO2*) and protein phosphatase (*SGPP2*) on chromosome 2 have been associated with abnormal spermatid

morphology, abnormal gametogenesis, small testes, male and female infertility in mouse (*SGO2*), while *SGPP2* has been associated with dwarfism in chicken (Taguchi et al., 2016).

In AFR, the region with the greatest enrichment of novel variants was found on chromosome 3 (Figure 5). There were also genes with unknown functions that were located in the novel SNP enriched regions in all of the breeds. It has been suggested that rare or low-frequency variants may explain a substantial proportion of the heritability of many complex diseases, most of which have previously not been fully captured in GWAS studies (Bang et al., 2014). Therefore, the identification of the functions of these uncharacterized genes may increase the utility of these SNPs for selection for productivity, including product quality, increasing animal welfare, disease resistance and reducing environmental impact (Thornton, 2010).

The power to identify variants associated with traits, particularly those of small effect, could be increased if certain regions of the genome were known to be enriched for trait associations (Koufariotis et al. 2014). However, given the typical genetic architecture of complex traits, such regions are likely to be very few. Variants in regions of the genome for which the sequence is strongly conserved across species have been proposed as an important annotation class for prioritization since they are potentially regulatory. The majority of these variants are found in non-coding regions, and it is believed that at least some of these are *cis* regulators for genes (Knight et al., 2011).

**Conclusion**

Identification of novel SNPs including nsSNPs provides the potential for the detection of genes and variants underlying variation in traits of economic importance in these breeds, in particular environmental adaptation. Genes located in genomic regions that are enriched for variation suggests their potential for selection due to effects on phenotypic characteristics. Of the SNPs identified in Afrikaner, Drakensberger and Nguni, 16% were predicted to be unique to these SA indigenous breeds. The results of this study provide a framework for further genetic association and QTL fine-mapping studies in indigenous SA cattle. This work should enable more genetic studies on these breeds, knowing the basis of their unique traits for breed improvement.

## References


Abin, S., Theron, H.E. & Van Marle-Koster, E., 2016. Population structure and genetic trends for indigenous African beef cattle breeds in South Africa: short communication. SA J. Anim. Sci. 46, 152-156.

Albrechtsen, A., Nielsen, F.C. & Nielsen, R., 2010. Ascertainment biases in SNP chips affect measures of population divergence. Mol. Biol. Evol. 27, 2534-2547.

Anderson, G. & Takahama, Y., 2012. Thymic epithelial cells: working class heroes for T cell development and repertoire selection. Trends Immunol. 33, 256-263.

Antonelli, M., Badiali, M., Moi, L., Buttarelli, F.R., Baldi, C., Massimino, M., Sanson, M. & Giangaspero, F., 2015. *KIAA1549*: BRAF fusion gene in pediatric brain tumors of various histogenesis. Pediatr. Blood Cancer. 62, 724-727.

Arambula, S.E., Belcher, S.M., Planchart, A., Turner, S.D. & Patisaul, H.B., 2016. Impact of low dose oral exposure to Bisphenol A (BPA) on the neonatal rat hypothalamic and hippocampal transcriptome: A CLARITY-BPA Consortium Study. Endocrinol. 157, 3856-3872.

Aslam, M.L., Bastiaansen, J.W., Elferink, M.G., Megens, H.J., Crooijmans, R.P., Blomberg, L.A., Fleischer, R.C., Van Tassell, C.P., Sonstegard, T.S., Schroeder, S.G., Groenen, M.A. & Long, J.A., 2012. Whole genome SNP discovery and analysis of genetic diversity in Turkey (*Meleagris gallopavo*). BMC Genomics. 13, 391, 1-14.

Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler D., Gabriel, S. & DePristo, M.A., 2013. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr. Protoc. Bioinf. 43, 11, 1-11.

Bahbahani, H., Clifford, H., Wragg, D., Mbole-Kariuki, M.N., Van Tassell, C., Sonstegard, T., Woolhouse, M. & Hanotte, O., 2015. Signatures of positive selection in East African Shorthorn Zebu: A genome-wide single nucleotide polymorphism analysis. Sci. Rep. 5, 11729, 1-13.

Bang, S.Y., Na, Y.J., Kim, K., Joo, Y.B., Park, Y., Lee, J., Lee, S.Y., Ansari, A.A., Jung, J., Rhee, H., Lee, J.Y., Han, B.G., Ahn, S.M., Won, S., Lee, H.S. & Ba, S.C., 2014. Targeted exon sequencing fails to identify rare coding variants with large effect in rheumatoid arthritis. Arthritis Res. Ther. 16, 447, 1-9.

Bardot, P., Vincent, S.D., Fournier, M., Hubaud, A., Joint, M., Tora, L. & Pourquie, O., 2016. *TAF10* is required for the integrity of *TFIID* and *SAGA* complexes but is initially dispensable for somitogenesis in the mouse embryo. BioRxiv. 071324, 1-42.

Bareke, E., Saillour, V., Spinella, J.F., Vidal, R., Healy, J., Sinnett, D. & Csuros, M., 2013. Joint genotype inference with germline and somatic mutations. BMC Bioinf. 14, S3, 1-11.

Barris, W., Harrison, B.E., McWilliam, S., Bunch, R.J., Goddard, M.E. & Barendse, W., 2012. Next generation sequencing of African and Indicine cattle to identify single nucleotide polymorphisms. Anim. Prod. 52, 133-142.

Bergantinos, C., Corominas, M. & Serras, F., 2010. Cell death-induced regeneration in wing imaginal discs requires JNK signaling. Devel. 137, 1169-1179.

Bhogaraju, S., Cajanek, L., Fort, C., Blisnick, T., Weber, K., Taschner, M., Mizuno, N., Lamla, S., Bastin, P., Nigg, E.A. & Lorentzen, E., 2013. Molecular basis of tubulin transport within the cilium by *IFT74* and *IFT81*. Sci. 341, 1009-1012.

Bischoff, S.R., Tsai, S., Hardison, N.E., York, A.M., Freking, B.A., Nonneman, D., Rohrer, G. & Piedrahita, J.A., 2008. Identification of SNPs and INDELS in swine transcribed sequences using short oligonucleotide microarrays. BMC Genomics. 9, 252, 1-14.

Boichard, D., 2002. PEDIG: a fortran package for pedigree analysis suited for large populations. In Proceedings of the 7th World Congress on Genetics Applied to Livestock Production. 32, 525-528.

Bolger, A.M., Lohse, M. & Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinf. 30, 2114-2120.

Bonafe, L., Cormier-Daire, V., Hall, C., Lachman, R., Mortier, G., Mundlos, S., Nishimura, G., Sangiorgi, L., Savarirayan, R., Sillence, D. & Spranger, J., 2015. Nosology and classification of genetic skeletal disorders: 2015 revision. Am. J. Med. Genet. 167, 2869-2892.

Bottini, N., Vang, T., Cucca, F. & Mustelin, T., 2006. Role of *PTPN22* in type 1 diabetes and other autoimmune diseases. Semin. Immunol. 18, 207-213.

Boutet, G., Carvalho, S.A., Falque, M., Peterlongo, P., Lhuillier, E., Bouchez, O., Lavaud, C., Pilet-Nayel, M.L., Rivière, N. & Baranger, A., 2016. SNP discovery and genetic mapping using Genotyping by Sequencing of whole genome genomic DNA from a pea RIL population. BMC Genomics. 17, 121, 1-14.

Brisken, C., Heineman, A., Chavarria, T., Elenbaas, B., Tan, J., Dey, S.K., McMahon, J.A., McMahon, A.P. & Weinberg, R.A., 2000. Essential function of *Wnt-4* in mammary gland development downstream of progesterone signaling. Genes. Dev. 14, 650-654.

Bromberg, Y. & Rost, B., 2007. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic acids Res. 35, 3823-3835.

Charrier, C., Joshi, K., Coutinho-Budd, J., Kim, J.E., Lambert, N., De Marchena, J., Jin, W.L., Vanderhaeghen, P., Ghosh, A., Sassa, T. & Polleux, F., 2012. Inhibition of *SRGAP2* function by its human-specific paralogs induces neoteny during spine maturation. Cell. 149, 923-935.

Choi, J.W., Choi, B.H., Lee, S.H., Lee, S.S., Kim, H.C., Yu, D., Chung, W.H., Lee, K.T., Chai, H.H., Cho, Y.M. & Lim, D., 2015. Whole-genome resequencing analysis of Hanwoo and Yanbian cattle to identify genome-wide SNPs and signatures of selection. Mol. Cells. 38, 466-473.

Choi, J.W., Liao, X., Park, S., Jeon, H.J., Chung, W.H., Stothard, P., Park, Y.S., Lee, J.K., Lee, K.T., Kim, S.H. & Oh, J.D., 2013. Massively parallel sequencing of Chikso (Korean brindle cattle) to discover genome-wide SNPs and Indels. Mol. Cells. 36, 203-211.

Choi, J.W., Liao, X., Stothard, P., Chung, W.H., Jeon, H.J., Miller, S.P., Choi, S.Y., Lee, J.K., Yang, B., Lee, K.T. & Han, K.J., 2014. Whole-genome analyses of Korean native and Holstein cattle breeds by massively parallel sequencing. PloS One. 9, e101127, 1-13.

Choudhury, A., Hazelhurst, S., Meintjes, A., Achinike-Oduaran, O., Aron, S., Gamieldien, J., Dashti, M.J.S., Mulder, N., Tiffin, N. & Ramsay, M., 2014. Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. BMC Genomics. 15, 437, 1-20.

Chung, J.H., Choi, H.J., Kim, S.Y., Hong, K.S., Min, S.K., Nam, M.H., Kim, C.W., Koh, Y.H. & Seo, J.B., 2011. Proteomic and biochemical analyses reveal the activation of unfolded protein response, *ERK-1/2* and ribosomal protein S6 signaling in experimental autoimmune myocarditis rat model. BMC Genomics. 12, 250, 1-12.

Cieslak, M., Reissmann, M., Hofreiter, M. & Ludwig, A., 2011. Colours of domestication. Biol. Rev. 86, 885-899.

Dadi, H., Kim, J.J., Yoon, D. & Kim, K.S., 2011. Evaluation of single nucleotide polymorphisms (SNPs) genotyped by the Illumina Bovine SNP50K in cattle focusing on Hanwoo breed. Asian Australas. J. Anim. Sci. 25, 28-32.

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., Van Binsbergen, R., Brøndum, R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C. & Esquerre, D., 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat. Genet. 46, 858-865.

Dai, Z., Huang, Y., Sadee, W. & Blower, P., 2007. Chemoinformatics analysis identifies cytotoxic compounds susceptible to chemoresistance mediated by glutathione and cystine/glutamate transport system xc. J. Med. Chem. 50, 1896-1906.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. & McVean, G., 2011. The variant call format and VCFtools. Bioinf. 27, 2156-2158.

Das, A., Panitz, F., Gregersen, V.R., Bendixen, C. & Holm, L.E., 2015. Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. BMC Genomics. 16, 1043, 1-10.

de Anda, F.C., Madabhushi, R., Rei, D., Meng, J., Gräff, J., Durak, O., Meletis, K., Richter, M., Schwanke, B., Mungenast, A. & Tsai, L.H., 2016. Cortical neurons gradually attain a post-mitotic state. Cell Res. 26, 1033-1047.

Decker, J.E., McKay, S.D., Rolf, M.M., Kim, J., Alcala, A.M., Sonstegard, T.S., Hanotte, O., Gotherstrom, A., Seabury, C.M., Praharani, L., Babar, M.E., Correia de Almeida Regitano, L, Yildiz, M.A., Heaton, M.P., Liu, W.S., Lei, C.Z., Reecy, J.M., Saif-Ur-Rehman, M., Schnabel, R.D. & Taylor, J.F., 2014. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. PLoS Genet. 10, e1004254, 1-14.

Dronkert, M.L.G., De Wit, J., Boeve, M., Vasconcelos, M.L., van Steeg, H., Tan, T.L.R., Hoeijmakers, J.H.J. & Kanaar, R., 2000. Disruption of mouse *SNM1* causes increased sensitivity to the DNA interstrand cross-linking agent mitomycin C. Molec. Cell. Biol. 20, 4553-4561.

Eck, S.H., Benet-Pages, A., Flisikowski, K., Meitinger, T., Fries, R. & Strom, T.M., 2009. Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. Genome Biol. 10, R82, 1-8.

Edea, Z., Dadi, H., Kim, S.W., Dessie, T., Lee, T., Kim, H., Kim, J.J. & Kim, K.S., 2013. Genetic diversity, population structure and relationships in indigenous cattle populations of Ethiopia and Korean Hanwoo breeds using SNP markers. Front. Genet. 4, 1-9.

Erschbamer, M.K., Hofstetter, C.P. & Olson, L., 2005. *RhoA*, *RhoB*, *RhoC*, *Rac1*, *Cdc42*, and *Tc10* mRNA levels in spinal cord, sensory ganglia, and corticospinal tract neurons and long-lasting specific changes following spinal cord injury. J. Comp. Neurol. 484, 224-233.

Gautier, M., Laloe, D. & Moazami-Goudarzi, K., 2010. Insights into the genetic history of French cattle from dense SNP data on 47 worldwide breeds. PLoS One. 5, e13038, 1-11.

Goodier, J.L., Cheung, L.E. & Kazazian Jr, H.H., 2012. *MOV10* RNA helicase is a potent inhibitor of retrotransposition in cells. PLoS Genet. 8, e1002941, 1-14.

Gurgul, A., Zukowski, K., Pawlina, K., Zqbek, T., Semik, E. & Bugno-Poniewierska, M., 2013. The evaluation of bovine SNP50 BeadChip assay performance in Polish Red cattle breed. Folia Biol. 61, 173-176.

Hanotte, O. & Jianlin, H., 2006. Genetic characterization of livestock populations and its use in conservation decision-making. The Role of Biotechnology in Exploring and Protecting Agricultural Genetic Resources. FAO, Rome, pp.89-96.

Hanotte, O., Tawah, C.L., Bradley, D.G., Okomo, M., Verjee, Y., Ochieng, J. & Rege, J.E.O., 2000. Geographic distribution and frequency of a taurine *Bos taurus* and an indicine *Bos indicus* Y specific allele amongst sub-Saharan African cattle breeds. Mol. Ecol. 9, 387-396.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37, 1-13.

Hughes, S.G., 1998. Prescribing for the elderly patient: why do we need to exercise caution? Br. J. Clin. Pharmacol. 46, 531-533.

Hwang, C.Y., Holl, J., Rajan, D., Lee, Y., Kim, S., Um, M., Kwon, K.S. & Song, B., 2010. *Hsp70* interacts with the retroviral restriction factor *TRIM5α* and assists the folding of *TRIM5α*. J. Biol. Chem. 285, 7827-7837.

Kawahara-Miki, R., Tsuda, K., Shiwa, Y., Arai-Kichise, Y., Matsumoto, T., Kanesaki, Y., Oda, S.I., Ebihara, S., Yajima, S., Yoshikawa, H. & Kono, T., 2011. Whole-genome resequencing shows numerous genes with nonsynonymous SNPs in the Japanese native cattle Kuchinoshima-Ushi. BMC Genomics. 12, 103, 1-8.

Kim, D.S., Cho, C.Y., Huh, J.W., Kim, H.S. & Cho, H.G., 2009. EVOG: a database for evolutionary analysis of overlapping genes. Nucleic Acids Res. 37, D698-D702.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson,M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J. & Parsey, R.V., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage. 46, 786-802.

Knight, J., Barnes, M.R., Breen, G. & Weale, M.E., 2011. Using functional annotation for the empirical determination of bayes factors for genome-wide association study analysis. PLoS One. 6, e14808-10, 1-8.

Koufariotis, L., Chen, Y.P.P., Bolormaa, S. & Hayes, B.J., 2014. Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle. BMC Genomics. 15, 436, 1-16.

Kwong, L.K., Neumann, M., Sampathu, D.M., Lee, V.M.Y. & Trojanowski, J.Q., 2007. *TDP-43* proteinopathy: the neuropathology underlying major forms of sporadic and familial frontotemporal lobar degeneration and motor neuron disease. Acta Neuropathol. 114, 63-70.

Le Roex, N., Noyes, H., Brass, A., Bradley, D.G., Kemp, S.J., Kay, S., Van Helden, P.D. & Hoal, E.G., 2012. Novel SNP discovery in African buffalo, *Syncerus caffer*, using high-throughput sequencing. PloS One. 7, e48792, 1-6.

Lee, S.M., Gardy, J.L., Cheung, C.Y., Cheung, T.K., Hui, K.P., Ip, N.Y., Guan, Y., Hancock, R.E. & Peiris, J.M., 2009. Systems-level comparison of host-responses elicited by avian *H5N1* and seasonal *H1N1* influenza viruses in primary human macrophages. PloS One. 4, e8072, 1-11.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G. & Durbin R., 2009. The sequence alignment/map (SAM) format and SAMtools. Bioinf. 25, 2078-2079.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinf. 25, 1754-1760.

Liao, X., Lan, C., Liao, D., Tian, J. & Huang, X., 2016. Exploration and detection of potential regulatory variants in refractive error GWAS. Sci. Rep. 6, 33090, 1-9.

Liu, L., Akhter, S., Bae, J.B., Mukhopadhyay, S.S., Richie, C.T., Liu, X. & Legerski, R., 2009. *SNM1B*/Apollo interacts with astrin and is required for the prophase cell cycle checkpoint. Cell Cycle. 8, 628-638.

Lukashkin, A.N., Legan, P.K., Weddell, T.D., Lukashkina, V.A., Goodyear, R.J., Welstead, L.J., Petit, C., Russell, I.J. & Richardson, G.P., 2012. A mouse model for human deafness *DFNB22* reveals that hearing impairment is due to a loss of inner hair cell stimulation. Proc. Natl. Acad. Sci. 109, 19351-19356.

Makina, S.O., Muchadeyi, F.C., Marle-Köster, E., Taylor, J.F., Makgahlela, M.L. & Maiwashe, A., 2015. Genome-wide scan for selection signatures in six cattle breeds in South Africa. Genet. Sel. Evol. 47, 92, 1-14.

Makina, S.O., Muchadeyi, F.C., van Marle-Köster, E., MacNeil, M.D. & Maiwashe, A., 2014. Genetic diversity and population structure among six cattle breeds in South Africa using a whole genome SNP panel. Front. Genet. 5, 333, 1-7.

Makina, S.O., Whitacre, L.K., Decker, J.E., Taylor, J.F., MacNeil, M.D., Scholtz, M.M., Marle-Köster, E., Muchadeyi, F.C., Makgahlela, M.L. & Maiwashe, A., 2016. Insight into the genetic composition of South African Sanga cattle using SNP data from cattle breeds worldwide. Genet. Sel. Evol. 48, 88, 1-7.

Matukumalli, L.K., Lawley, C.T., Schnabel, R.D., Taylor, J.F., Allan, M.F., Heaton, M.P., O'Connell, J., Moore, S.S., Smith, T.P., Sonstegard, T.S. & Van Tassell, C.P., 2009. Development and characterization of a high density SNP genotyping assay for cattle. PloS One. 4, e5350, 1-13.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P. & Cunningham, F., 2016. The Ensembl Variant Effect Predictor. Genome Biol. 17, 122, 1-14.

McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. & Cunningham, F., 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinf. 26, 2069-2070.

McTavish, E.J. & Hillis, D.M., 2015. How do SNP ascertainment schemes and population demographics affect inferences about population history? BMC Genomics. 16, 266, 1-26.

Mei, C., Wang, H., Zhu, W., Wang, H., Cheng, G., Qu, K., Guang, X., Li, A., Zhao, C., Yang, W., Chongzhi, Wang., Yaping, Xin., Linsen & Zan., 2016. Whole-genome sequencing of the endangered bovine species Gayal (*Bos frontalis*) provides new insights into its genetic features. Sci. Rep. 6, 19787, 1-8.

Meszaros, G., Eaglen, S. & Waldmann, P., 2014. A genome wide association study for longevity in cattle. Open J. Genet. 4, 46-55.

Meuwissen, T. H. E., & Luo Z., 1992. Computing inbreeding coefficients in large populations. Genet. Sel. Evol. 24, 305-313.

Michou, L., Lasbleiz, S., Rat, A.C., Migliorini, P., Balsa, A., Westhovens, R., Barrera, P., Alves, H., Pierlot, C., Glikmans, E. & Garnier, S., 2007. Linkage proof for *PTPN22*, a rheumatoid arthritis susceptibility gene and a human autoimmunity gene. Proc. Natl. Acad. Sci. 104, 1649-1654.

Mitchell, K.J., 2015. Doctoral dissertation: Using high-throughput DNA sequencing and molecular phylogenies to investigate the evolution and biogeography of the southern hemisphere. University of Adelaide, Australia. pp 1.

Montgomery, C.P. & Daum, R.S., 2009. Transcription of inflammatory genes in the lung after infection with community-associated methicillin-resistant Staphylococcus aureus: a role for Panton-Valentine leukocidin? Infect. Immun. 77, 2159-2167.

Mullen, M.P., Creevey, C.J., Berry, D.P., McCabe, M.S., Magee, D.A., Howard, D.J., Killeen, A.P., Park, S.D., McGettigan, P.A., Lucy, M.C., Machugh, D.E. & Waters, S.M., 2012. Polymorphism discovery and allele frequency estimation using high-throughput DNA sequencing of target-enriched pooled DNA samples. BMC Genomics. 13, 16, 1-12.

Nakayama, T., Asai, S., Takahashi, Y., Maekawa, O. & Kasama, Y., 2007. Overlapping of genes in the human genome. Int. J. Biomed Sci. 3, 14-19.

Nielsen, R., Paul, J.S., Albrechtsen, A. & Song, Y.S., 2011. Genotype and SNP calling from next-generation sequencing data. Nat. Rev. Genet. 12, 443-451.

Park, S.Y., Lee, J.H., Ha, M., Nam, J.W. & Kim, V.N., 2009. miR-29 miRNAs activate *p53* by targeting *p85α* and *CDC42*. Nat. Struct. Mol. Biol. 16, 23-29.

Pool, J.E., Hellmann, I., Jensen, J.D. & Nielsen, R., 2010. Population genetic inference from genomic sequence variation. Genome Res. 20, 291-300.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. & Sham, P.C., 2007. PLINK: A toolset for whole-genome association and population-based linkage analysis. Am. J. Hum. Genet. 81, 559-575.

Ramirez-Gonzalez, R., Bonnal, R. J., Caccamo, M., & MacLean, D., 2012. Biosamtools: Ruby bindings for Samtools, a library for accessing bam files containing high-throughput sequence alignments. Source Code Biol. Med. 7, 6, 1-6.

Ramos, A.M., Crooijmans, R.P., Affara, N.A., Amaral, A.J., Archibald, A.L., Beever, J.E., Bendixen, C., Churcher, C., Clark, R., Dehais, P. & Hansen, M.S., 2009. Design of a high density SNP genotyping assay in the pig using SNPs identified and characterized by next generation sequencing technology. PloS One. 4, e6524, 1-13.

Ramos, A.M., Megens, H.J., Crooijmans, R.P.M.A., Schook, L.B. & Groenen, M.A.M., 2011. Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing. Anim. Genet. 42, 613-620.

Rivals, I., Personnaz, L., Taing, L. & Potier, M.C., 2007. Enrichment or depletion of a GO category within a class of genes: which test? Bioinf. 23, 401-407.

Rovelet-Lecrux, A., Legallic, S., Wallon, D., Flaman, J.M., Martinaud, O., Bombois, S., Rollin-Sillaire, A., Michon, A., Le Ber, I., Pariente, J., Puel, M., Paquet, C., Croisile, B., Thomas-Antérion, C., Vercelletto, M., Lévy, R., Frébourg, T., Hannequin, D., Campion, D, & Investigators of the GMAJ project., 2012. A genome-wide study reveals rare CNVs exclusive to extreme phenotypes of Alzheimer disease. Eur. J. Hum. Genet. 20, 613-617.

Sambrook, J. & Russell, D.W., 2006. Amplification of cDNA generated by reverse transcription of mRNA. CSH Protoc. 2006, pii: pdb.prot3837, 1-6.

Sanarana, Y., Visser, C., Bosman, L., Nephawe, K., Maiwashe, A. & van Marle-Köster, E., 2016. Genetic diversity in South African Nguni cattle ecotypes based on microsatellite markers. Trop. Anim. Health. Prod. 48, 379-385.

Schaller, C.E., Wang, C.L., Beck-Engeser, G., Goss, L., Scott, H.S., Anderson, M.S. & Wabl, M., 2008. Expression of Aire and the early wave of apoptosis in spermatogenesis. J. Immunol. 180, 1338-1343.

Scott, K., Gadomski, T., Kozicz, T. & Morava, E., 2014. Congenital disorders of glycosylation: new defects and still counting. J. Inherit. Metab. Dis. 37, 609-617.

Slikas, B., Jones, I.B., Derrickson, S.R. & Fleischer, R.C., 2000. Phylogenetic relationships of Micronesian white-eyes based on mitochondrial sequence data. Auk. 117, 355-365.

Steingrimsson, E., Copeland, N.G. & Jenkins, N.A., 2006. Mouse coat color mutations: from fancy mice to functional genomics. Dev. Dynam. 235, 2401-2411.

Stenson, P.D., Ball, E.V., Howells, K., Phillips, A.D., Mort, M. & Cooper, D.N., 2009. The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalised genomics. Human Genomics. 4, 69-72.

Stothard, P., Choi, J.W., Basu, U., Sumner-Thomson, J.M., Meng, Y., Liao, X. & Moore, S.S., 2011. Whole genome resequencing of black Angus and Holstein cattle for SNP and CNV discovery. BMC Genomics 12, 559, 1-14.

Sturm, R.A., 2009. Molecular genetics of human pigmentation diversity. Hum. Mol. Gen. 18, R9-R17.

Taguchi, Y., Allende, M.L., Mizukami, H., Cook, E.K., Gavrilova, O., Tuymetova, G., Clarke, B.A., Chen, W., Olivera, A. & Proia, R.L., 2016. Sphingosine-1-phosphate phosphatase 2 regulates pancreatic islet β-cell endoplasmic reticulum stress and proliferation. J. Biol. Chem. 291, 12029-12038.

Thornton, P.K., 2010. Livestock production: recent trends, future prospects. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 365, 2853-2867.

Turner, S.D., 2014. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. BiorXiv. pp 005165, 1-2.

Tuysuz, B., Bilguvar, K., Koçer, N., Yalçınkaya, C., Çaglayan, O., Gul, E., Şahin, S., Çomu, S. & Gunel, M., 2014. Autosomal recessive spastic tetraplegia caused by *AP4M1* and *AP4B1* gene mutation: expansion of the facial and neuroimaging features. Am. J. Med. Genet. 164, 1677-1685.

Vainio, S., Heikkila, M., Kispert, A., Chin, N. & McMahon, A.P., 1999. Female development in mammals is regulated by Wnt-4 signaling. Nature. 397, 405-409.

Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C. & Sonstegard, T.S., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Meth. 5, 247-252.

Van, K., Kang, Y.J., Han, K.S., Lee, Y.H., Gwag, J.G., Moon, J.K., & Lee, S.H., 2013. Genome-wide SNP discovery in mungbean by Illumina HiSeq. Theor. Appl. Genet. 126, 2017-2027.

Vetter, C.S., Groh, V., thor Straten, P., Spies, T., Brocker, E.B. & Becker, J.C., 2002. Expression of stress-induced *MHC* class I related chain molecules on human melanoma. J. Invest. Dermatol. 118. 600-605.

Viscomi, M.T., Oddi, S., Latini, L., Pasquariello, N., Florenzano, F., Bernardi, G., Molinari, M. & Maccarrone, M., 2009. Selective *CB2* receptor agonism protects central neurons from remote axotomy-induced apoptosis through the *PI3K*/*Akt* pathway. J. Neurosci. 29, 4564-4570.

Wang, J., Li, L. & Zhang, G., 2016. A high-density SNP genetic linkage map and QTL analysis of growth-related traits in a hybrid family of oysters (*Crassostrea gigas*× *Crassostrea angulata*) using genotyping-by-sequencing. G3 (Bethesda). 6, 1417-1426.

Wang, J., Raskin, L., Samuels, D.C., Shyr, Y. & Guo, Y., 2014. Genome measures used for quality control are dependent on gene function and ancestry. Bioinf. 31, 318-323.

Weckx, S., Del-Favero, J., Rademakers, R., Claes, L., Cruts, M., De Jonghe, P., Van Broeckhoven, C. & De Rijk, P., 2005. NovoSNP, a novel computational tool for sequence variation discovery. Genome Res. 15, 436-442.

Weiher, H., Noda, T., Gray, D.A., Sharpe, A.H. & Jaenisch, R., 1990. Transgenic mouse model of kidney disease: insertional inactivation of ubiquitously expressed gene leads to nephrotic syndrome. Cell. 62, 425-434.

Wilkinson, R.N., Jopling, C. & Van Eeden, F.J., 2014. Zebrafish as a model of cardiac disease. Prog. Mol. Biol. Transl. Sci. 124, 65-91.

Wu, X., Steet, R.A., Bohorov, O., Bakker, J., Newell, J., Krieger, M., Spaapen, L., Kornfeld, S. & Freeze, H.H., 2004. Mutation of the *COG* complex subunit gene *COG7* causes a lethal congenital disorder. Nat. Med. 10, 518-523.

Xuebin, Q., 2004. PhD Thesis: Genetic diversity, differentiation and relationship of domestic yak populations: a microsatellite and mitochondrial DNA study. pp 262.

Zhang, L., Kang, L., Bond, W. & Zhang, N., 2009. Interaction between syntaxin 8 and *HECTd3*, a *HECT* domain ligase. Cell Mol. Neurobiol. 29, 115-121.

Zheng, J., Miller, K.K., Yang, T., Hildebrand, M.S., Shearer, A.E., DeLuca, A.P., Scheetz, T.E., Drummond, J., Scherer, S.E., Legan, P.K. & Goodyear, R.J., 2011. Carcinoembryonic antigen-related cell adhesion molecule 16 interacts with α-tectorin and is mutated in autosomal dominant hearing loss (*DFNA4)*. Proc. Natl. Acad. Sci. 108, 4218-4223.

Zwane, A.A., Maiwashe, A., Makgahlela, M.L., Choudhury, A., Taylor, J.F., & van Marle-Köster, E., 2016. Genome-wide identification of breed-informative single-nucleotide polymorphisms in three South African indigenous cattle breeds. SA. J. Anim. Sci. 46, 302-312.

# CHAPTER FIVE

## Identification of selective sweeps and breed-specific SNPs in Afrikaner, Drakensberger and Nguni cattle using genome-wide sequence data

A. A. Zwane[1,2], E. van Marle-Köster[2], A. Maiwashe[1,3] and J.F. Taylor[4]

[1]Department of Animal Breeding and Genetics, ARC-API, P/Bag X2, Irene, 0062,

[2]Department of Animal and Wildlife Sciences, University of Pretoria, P/Bag X20, Hatfield, Pretoria, 0028,

[3]Department of Animal, Wildlife and Grassland Sciences, University of the Free State, Bloemfontein 9300, South Africa,

[4]Division of Animal Sciences, University of Missouri, 920 East Campus Drive, Columbia, MO 65211-5300.

**Prepared for Publication**

# Identification of selective sweeps and breed-specific SNPs in Afrikaner, Drakensberger and Nguni cattle using genome-wide sequence data

**A.A. Zwane[1,2], E. van Marle-Köster[2], A. Maiwashe[1,3] and J.F. Taylor[4]**

[1]Department of Animal Breeding and Genetics, ARC-API, P/Bag X2, Irene, 0062, [2]Department of Animal and Wildlife Sciences, University of Pretoria, P/Bag X20, Hatfield, Pretoria, 0028, , [3]Department of Animal, Wildlife and Grassland Sciences, University of the Free State, Bloemfontein 9300, South Africa, [4]Division of Animal Sciences, University of Missouri, 920 East Campus Drive, Columbia, MO 65211-5300.

## Abstract

The indigenous South African breeds including Afrikaner (AFR), Drakensberger (DRA), and Nguni (NGI) are important genetic resources to world cattle production. The aim of this study was to identify selective sweeps and also to identify breed-specific SNPs for breed distinction among indigenous SA breeds. Whole genome sequencing of pools of DNA from AFR, DRA, and NGI was performed using an Illumina HiSeq 2000 and 17.6 million variants were discovered in the three breeds. A total of 4.3 million novel SNPs, 1,751,065 (AFR), 823,997 (DRA) and 1,794,817 (NGI) were identified when compared to Run 5 of the 1000 Bull Genomes Project. Whole-genome screening was performed to detect selective sweeps throughout the genomes of AFR, DRA, and NGI breeds, and also putative breed-specific SNPs. A total of 96 putative selective sweeps were identified ($ZH_p$ score $\leq$ -5) across the breeds, as well as 186 putative breed-specific SNPs (SNPs that are variable in one SA breed but that are fixed in the two other SA breeds, and that were also found on the BovineSNP50 assay and thus could be used for validation within SA breeds). When tested for breed differentiation, putative breed-specific SNPs showed a 100% accuracy for breed allocation using PCA or GeneClass2. No SNPs were found that were fixed for one allele in one breed and for an alternate allele in the remainder of the breeds. The results of this study indicate that selective sweeps have contributed to the rapid recent phenotypic evolution of cattle in response to strong selection and also provide a suite of SNPs with utility for breed differentiation in these SA breeds.

Key words: indigenous breeds, selective sweeps, breed-specific SNPs, breed differentiation

## Introduction

Identification of recent positive selection signatures in indigenous cattle can provide information on genomic regions that have been subjected to both artificial and natural selection. (Zhao et al., 2015). Artificial selection has resulted in the diversity of cattle breeds that have been tamed for milk and meat production. These selection strategies are likely to have enforced selection pressures on particular regions of the genome that control these production traits, as well as other economic important traits such as disease resistance and adaptation traits (Rubin et al., 2012). Thus, under positive selection pressure, the frequency of favorable alleles in the genome increases, whereas if intensive selection pressure happens over a few generations, it is unlikely that recombination had an impact on haplotype structure, which could result in extended linkage disequilibrium (LD) patterns between the mutation and neighboring loci Sabeti et al., 2002). Analysis of these selection sweeps/signatures can reveal genomic regions of interest for selection and provide insights into the mechanisms of evolution in different cattle breeds such as Afrikaner (AFR), Drakensberger (DRA), and Nguni (NGI), which have never been studied at a sequence level.

The indigenous South African breeds are known for their adaptation to the local environmental conditions with higher tolerance to tick borne diseases than imported exotic taurine breeds (Scholtz et al., 2010; Mapholi et al., 2014). These characteristics have played important roles in the genetic improvement of these breeds. In particular, these breeds have been crossed with local and exotic breeds to produce composites for improved production (Mwai et al., 2015), in order to meet the increasing local demand for meat and meat products. This has also resulted in considerable changes in the morphology and behavior of modern animals, allowing for the formation of more diverse cattle breeds (Flori et al., 2009). Recurrent selection for variants of large effect leads to a loss of variation within the chromosomal regions flanking the selected variants and eventually lead to the complete fixation of a haplotype harbouring the selected variant (Smith & Haigh, 1974). This region of the genome is thererefore refered as the region that subjected to a “selective sweep”. However, such regions may also occur due to random drift (Rubin et al., 2012; Ramey et al., 2013).

Recent advances in genomics studies provide an excellent opportunity for identifying loci subjected to selection and also allow for the validation of new methods to detect selection

signatures (Hayes et al., 2008). Whole-genome sequencing now offers a suitable platform to examine the entire genome for the identification of selective sweeps, copy number of variants (CNVs) and also breed-specific SNPs with which to distinguish between members of different cattle populations (Gorbach et al., 2010). Identification of selective sweeps and breed-specific/informative markers in indigenous SA breeds has been limited using the currently available bovine genotyping assays, which lack extensive numbers of genomic variants discovered and common in these local breeds (Makina et al., 2015; Zwane et al., 2016).

The use of pooled DNA could also have a great value in breed characterisation, due the presence of SNPs common to a particular breed or population; and also an effective method for detecting selective sweeps because heterozygosity can be calculated in sliding windows from sequences drawn from a pool of haplotypes (Rubin et al., 2012; Choi et al., 2015). Markers that are fixed for different alleles in a particular breed are powerful for distinguishing among populations (i.e., markers that are fixed for one allele within a breed and all members of the breed possess the AA genotype, whereas members of the other breeds possess the BB genotype) (Blott et al., 1999; Pant et al., 2012). The markers incorporated on the BovineSNP50 and GGP-80K assays that have been scored in indigenous SA breeds (Afrikaner, Drakensberger and Nguni) have failed to reveal markers with breed-specific alleles in these breeds (Zwane et al., 2016), and were only useful for differentiating the breeds from other African breeds based upon skewed allele frequency differences between the breeds. This reflects the assay design bias that occurred in the development of the assays where common SNPs with high minor allele frequencies (MAF) in European taurine breeds were preferentially selected for incorporation onto the assay. However, the assays are capable of assigning individuals to taurine breeds based upon the differences in allele frequency that occur between these breeds. In this study, next generation sequence (NGS) data will be used to identify breed-specific SNPs for differerentiating among the breeds, and also identify selective sweeps underlying economic important traits among Afrikaner, Drakensberger and Nguni cattle.

## Materials and Methods

### Sampling, DNA isolation and sequencing

A total of 90 samples from three indigenous SA cattle breeds (AFR, DRA, and NGI) collected from nine different provinces of SA were extracted and sequenced in pools of 30 animals representing each breed, using an Illumina HiSeq 2000 (Illumina, San Diego, CA) instrument.

Sampling of blood and hair was performed with the approval of the Animal Ethics Committee of the University of Pretoria (EC: S4285-15). Roche DNA extraction Kit (Roche, Germany) was used to extract genomic DNA from whole blood (200 µl/sample) using the Roche DNA extraction Kit (Roche, Germany) and an optimized Phenol-Chloroform protocol (Sambrook & Russell, 2006) was used to extract DNA from hair roots. Extracted DNA was quantified using a Nanodrop UV/Vis Spectrophotometer (Nanodrop ND-1000) and verified using a Qubit® 2.0 Fluorometer (Thermo Scientific). DNA samples were maintained at 50 ng/µl concentration and samples were sent to Agricultural Research Council (ARC) Biotechnology Platform for whole genome sequencing.

**Sequence data analysis**

The raw Illumina DNA sequence data were trimmed using Trimmomatic (Bolger et al., 2014), aligned to reference genome UMD3.1 using the Burrows-Wheeler aligner (BWA) (Li & Durbin, 2009) and SNPs were called using Genome Analysis Tool Kit (GATK) after sorting the alignments and formatted them for variant calling using Picard tools (Li et al., 2009). An additional sequenced pool of Brahman (BRAH) cattle was used as a reference for testing the breed's relatedness to these SA breeds through principal component analysis (PCA), since BRAH has historically been infused into the indicine beef breeds present in SA. The BRAH sequence data were obtained from the University of Missouri, Animal Genomics Sequence Database, and was sequenced at 10X coverage using the Illumina Platform. BovineSNP50 assay data (AFR (n = 48), DRA (n = 48), and NGI (n = 56)) generated from previous studies (Makina et al. 2014) were used to check for overlaps between the breed specific SNPs identified from sequencing data and the genotypic data. Additional Afrikaner samples (n = 14), Drakensberger (n = 23), genotyped from ARC Biotechnology Platform, and Nguni samples (n = 50) from Mapholi (2015) were used for breed assignment.

**Principal component analysis**

To explore the relatedness among the breeds (AFR, DRA, and NGI), variant allele frequencies called from sequence data were used to cluster the breeds using PCA, using BRAH as an outgroup. The analysis was performed using 15,723,684 SNPs identified from a joint genotyping calling (cohort) of AFR, DRA, NGI, and BRAH. In PCA, the first two principal components account for high variation percentage that exist between populations and these principal components can certainly be used to find clusters (Khodadadi et al., 2011). The PCA

was performed using GENESIS v0.25, a program that is dependent on PLINK 1.9 (Purcell et al., 2007; Purcell & Chang, 2014) and which uses plink2evec to generate an .evec file for the PCA analysis (Buchmann & Hazelhurst, 2015). All non-autosomal SNPs were excluded for this analysis.

**Identification of selective sweeps**

Identification of selective sweeps was performed using the approach of Rubin et al. (2012) that makes provision for the identification of variants from pooled whole genome sequence data. This method determines, for each pool and SNP, the numbers of reads corresponding to the most ($n_{MAJ}$) and least abundant alleles ($n_{MIN}$) and for each window in each breed pool, a pooled heterozygosity score is calculated as:

$$H_p = 2\sum n_{MAJ}\sum n_{MIN} / (\sum n_{MAJ} + \sum n_{MIN})^2,$$

where $\sum n_{MAJ}$ and $\sum n_{MIN}$ are the sums of $n_{MAJ}$ and $n_{MIN}$ for all SNPs in the window. Individual $H_p$ values are then Z-transformed as follows:

$$ZH_p = (H_p - \mu H_p) / \sigma H_p.$$

where $\mu H_p$ and $\sigma H_p$ are the mean and standard deviation for the $H_p$ scores. To detect putative selective sweeps, a whole genome screen was performed to identify genomic regions with an excess of homozygosity (heterozygote deficiency) from the autosomes. All of the SNPs identified in the joint analysis of the three breeds were used to calculate Z-transformations of the pooled heterozygosity ($ZH_p$) in each of the three breeds separately, the numbers of sequence reads containing major and minor alleles were counted. Subsequently, we utilized a 50% overlapping sliding window approach with 150 kb windows, to compute $ZH_p$ in each of the windows, and plot the distribution of SNP counts within these windows. The 150 kb window size was chosen based on studies indicating it to be the most appropriate to detect windows with appropriate length to detect small sweeps (Rubin et al., 2012). Windows with $ZH_p$ Z-scores of $\leq$ -4 were retained as candidate selective sweep regions and regions with $ZH_p$ Z-scores of $\leq$ -5 as putative selective sweeps. In addition, animal QTLdb was used to retrieve quantitative trait loci (QTL) information and visualize the QTL located within the putative selective sweep regions (Hu et al., 2013).

**Identification of breed-specific SNPs and breed allocation**

Novel SNPs (SNPs not found in Run 5 of the 1000 Bull Genomes project) were used to seek breed-specific SNPs. First, novel SNPs were examined to identify if any were fixed for alternate

alleles between the breeds as described by Pant et al. (2012), as these SNPs provide high power for breed allocation. To accomplish this, we required one breed to be fixed for an 'A' allele and the other two breeds to be fixed for an alternate 'B' allele (or *vice versa*) (Blott et al., 1999). A second class of breed-specific SNPs was also determined which included SNPs that were detected as being variable in only one of the three populations (Ramos et al. 2011). These were determined by first removing all of the SNPs that were common between the three breeds. The remaining list of candidate breed-specific SNPs was next compared to the BovineSNP50 manifest to identify candidates included on this assay and for which additional assay data were available for the 152 animals from AFR ( n = 48), DRA (n = 48), and NGI (n = 56) breeds. These are the overlapping SNPs between the two datasets (the chip and the sequence data), referred to as putative breed-specific SNPs by Ramos et al. (2011). Breed assignment was conducted using putative breed-specific SNPs, using a second set of AFR samples (n = 14), DRA (n = 23 and NGI (n = 50), which were genotyped with the BovineSNP50 chip. The second set of 14 AFR, 23 DRA, and 50 NGI BovineSNP50 genotypes were used as the reference populations. Angus was used as an outgroup for PCA analysis.

The assignment test was performed using a PCA and the methods implemented in GeneClass2 (Piry et al., 2004), using 30 randomly selected NGI putative breed-specific SNPs (SNPs that were variable in NGI, but fixed in the other breeds). The assignment method available in GeneClass2 includes the allele frequency based method of Paetkau et al. (1995) and the Bayesian-based methods of Rannala & Mountain (1997). The breed allocation efficiencies and breed misclassification rates were estimated as the proportions of the total number of AFR, DRA, and NGI animals that were correctly or incorrectly classified (Ramos et al., 2011; Pant et al., 2012). Finally, the minor allele frequency (MAF) distribution for the putative breed-specific SNPs was estimated in order to see which SNPs possessed the highest frequencies for breed allocation (Ramos et al., 2011). All SNPs with MAF > 0.2 were regarded as SNPs with higher breed specificity and were included in a FREQ SNP panel. The MAF were calculated using PLINK v 1.9 software (Purcell, 2007).

**Results**

The PCA analysis using the whole genome sequence dataset revealed the genetic distances between the four breeds (AFR, DRA, NGI, and BRAH; Figure 1). All of the breeds clustered distinctly revealing significant genetic differences. The PCA1 clustered the breeds according

to their geographic distribution, suggesting that these breeds originated from different geographic areas; and PCA2 clustered the breeds according to their subgroups, *Bos primigenious indicus* (AFR, and NGI), African taurine (DRA), *Bos indicus* (BRAH). Genetic difference between the three indigenous SA breeds and Brahman was observed; and also the genetic difference between Drakensberger and Brahman. This PCA shows the potential of whole genome sequence data to identify breed-specific SNPs that discriminate between the breeds.

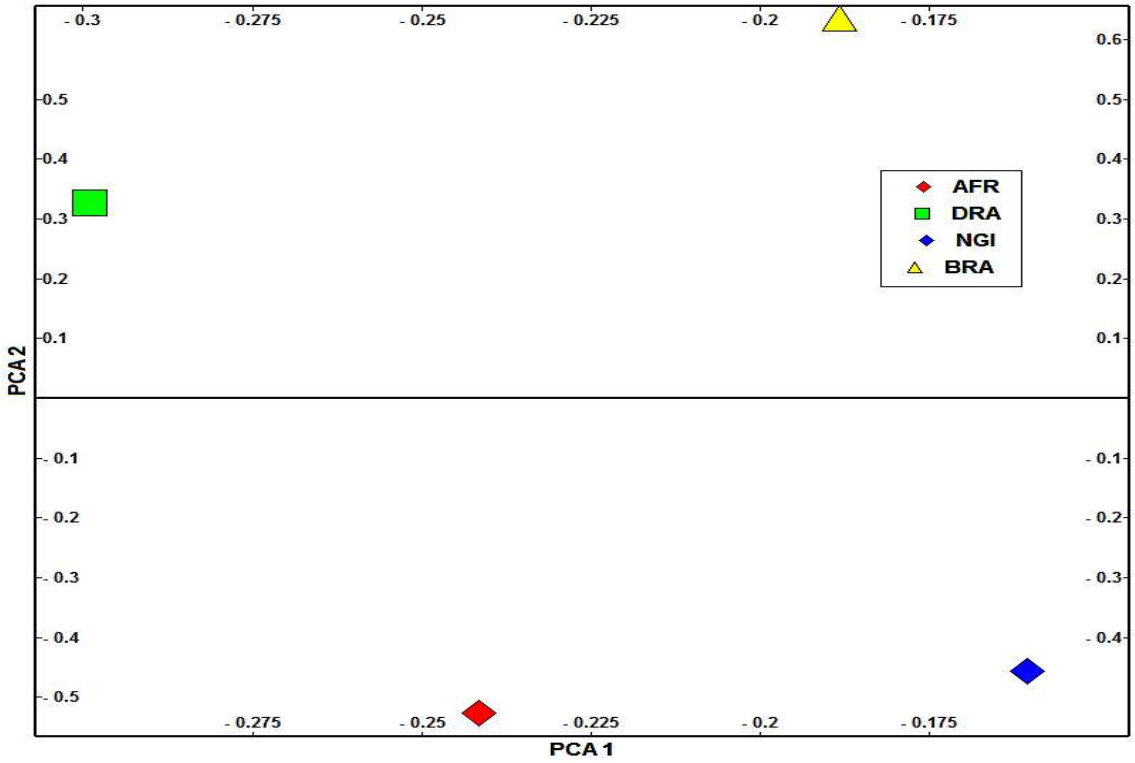


**Figure 1:** PCA1 against PCA2 plot for the three indigenous SA breeds with Brahman as a reference population using whole genome sequence data.

**Identification of selective sweeps**

A total of 33,467 150 kb sliding windows were used to calculate the Z-transformed pooled heterozygosity ($ZH_p$) scores to identify putative selective sweep regions. The $ZH_p$ Z-scores ranged from -10.26 to 2.18, from -5.27 to 1.37, and from -8.27 to 1.94 in AFR, DRA, and NGI, respectively. Thus, there appeared to be regions of excess homozygosity but not excess heterozygosity in the genomes of these animals. Figures 2A, B, and C show the distributions of the $ZH_p$ Z-scores genome-wide for the three indigenous SA breeds. The most noteworthy

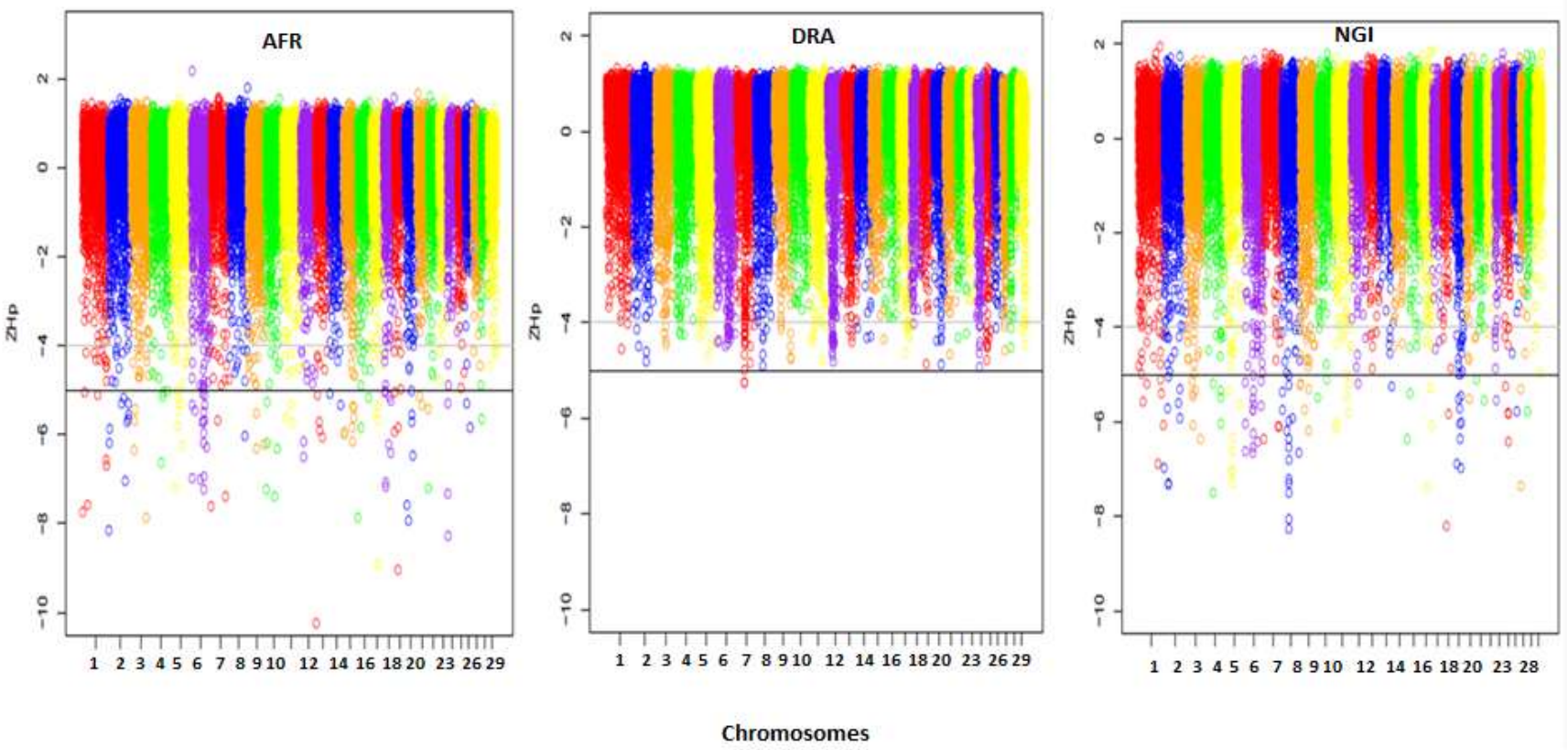


**Figure 2:** Distribution of $ZH_p$ Z-scores across all 29 autosomes for Afrikaner (AFR), Drakensberger (DRA), and Nguni (NGI). The horizontal lines indicate $ZH_p$ Z-score thresholds of -4 and -5 used to define candidate and putative selective sweep regions in this study.

regions of homozygosity were observed in a region spanning < 50 Mb on chromosomes 8 and 19 in NGI, chromosomes 13 and 19 in AFR, and on chromosome 7 in DRA.

The genome-wide screening of the breeds revealed 113 distinct loci with ZHp Z-scores ≤ −5, and 157 loci with ZHp Z-scores ≤ −4 in AFR, 2 and 152, respectively in DRA, and 108 and 156, respectively in NGI (Addendum A). In total, 465 candidate selective sweeps with $ZH_p$ Z-scores ≤ −4 were identified across the genomes of AFR, DRA and NGI and 223 regions were identified as putative selective sweeps ($ZH_p$ Z-scores ≤ −5) (Addendum A). The lowest number of putative selective sweeps was observed in DRA. The regions identified as candidate selective sweeps ($ZH_p$ Z-scores ≤ −4) in DRA were located on 26 different chromosomes. We also identified 93 selective sweep regions with extremely low $ZH_p$ Z-scores ($ZH_p$-scores ≤ −6) as indicated in Table 1. These regions could potentially contribute to the phenotypic differences between AFR, DRA, and NGI. A locus with extremely low $ZH_p$ Z-score of -10.26 was found in AFR on chromosome 13, but no annotated genes were identified in this region. A protein coding gene, family with sequence similarity 101, member B (*FAM101B)* was identified in a sweep region with a $ZH_p$ Z-score of -9.05 in AFR on chromosome 19, and was also found in NGI with a $ZH_p$ Z-score of -8.2 (Table 1). This gene is involved in the regulation of the perinuclear actin network and nuclear shape through interaction with filamins, and plays an essential role in the fornation of cartilaginous skeletal elements in human (http://www.uniprot.org/uniprot/Q8N5W9).

We also detected other genes in selective sweep regions that were common between the three breeds including, *ASIC5*, a gene associated with stress response in chicken (Fallahsharoudi et al., 2016), *DPYS*, a protein coding gene associated with Dihydropyrimidinuria and Dihydropyrimidine Dehydrogenase deficiencies, *DCTN5*, a protein coding gene involved in pathways for transport to the golgi, subsequent modification, and immune system, P*LK1*, essential for successful cell division (van Vugt & Medema, 2005), *ERN2*, which induces translational repression through 28S ribosomal RNA cleavage in response to ER stress, *MCOLN2* and *MLCOLN3,* which exhibit a common 6-membrane-spanning topology, and the *DTMRT3* gene, for which a role has not been well defined.

Among the 23 common genes found between AFR and NGI in selective sweep regions, *GZMK* is associated with heat stress in rat (Zhao et al., 2014), *ESM1* is an immune response gene in cattle (Cai, 2006), and *CNOT6* is associated with ovarian follicle development in cattle (Zielak-

**Table 1:** Putative selective sweep regions with extremely low $ZH_p$ Z-scores ≤ -6 and their associated genes in the two breeds. DRA was not represented in this table due to insufficiently low $ZH_p$ Z-scores.

| CHR | $ZH_p$ | $ZH_p$-Z Score | Associated Genes | Coordinates (bp) | Breed |
|---|---|---|---|---|---|
| 1 | 0.10 | -7.75 | *HIST1H4G* | 1,675,216-1,675,527 | AFR |
| 1 | 0.14 | -6.68 | *NDUFV3* | 144,690,277-144,701,618 | AFR |
| 2 | 0.09 | -8.15 | - | 2,026,957-2,027,063 | AFR |
| 2 | 0.16 | -6.20 | *LIMS2* | 4,780,176-4,819,192 | AFR |
| 3 | 0.16 | -6.35 | *SPRR3* | 17,796,522-17,798,340 | AFR |
| 4 | 0.15 | -6.62 | *KIAA0895* | 61,273,629-61,319,128 | AFR |
| 4 | 0.14 | -6.05 | *FAM71F1* | 93,460,874-93,462,935 | NGI |
| 5 | 0.12 | -7.20 | *PCBP2, PRR13, AMHR2, SP1* | 26,702,879-26,723,947<br>26,719,615-26,719,721<br>26,730,147-26,733,390<br>26,753,574-26,759,487 | AFR |
| 5 | 0.16 | -6.24 | *NUAK1* | 69,816,616-69,892,512 | AFR |
| 6 | 0.13 | -6.98 | - | 6,013,172-6,020,467 | AFR |
| 7 | 0.11, 0.13 | -7.62, -6.37 | *CNOT6* | 493,450-574,753 | AFR, NGI |
| 7 | 0.14 | -6.10 | - | 96,124,692-96,354,407 | NGI |
| 8 | 0.08 | -7.49 | *DMRT1* | 43,916,605-43,972,570 | NGI |
| 8 | 0.11 | -6.80 | *KANK1* | 4,4046,426-44,076,904 | NGI |
| 8 | 0.13 | -6.38 | *DOCK8* | 44,310,613-44,545,537 | NGI |
| 10 | 0.12 | -7.38 | *PYGO1* | 54,865,902-54,887,753 | AFR |
| 10 | 0.12 | -7.24 | *KIAA1191, SIMC1* | 4,950,671-4,964,129<br>4,979,152-5,011,240 | AFR |
| 10 | 0.16 | -6.18 | *PAPD4* | 10,549,964-10,609,975 | AFR |
| 11 | 0.14 | -6.05 | *TTC27* | 15,207,843-15,381,634 | NGI |
| 16 | 1.0, 0.13 | -7.86, -6.37 | - | 659,397-659,500 | AFR, NGI |
| 17 | 0.05, 0.09 | -8.92, -7.42 | *ASIC5* | 44,427,939-44,483,562 | AFR, NGI |
| 17 | 0.14 | -6.06 | *ZNF74, TSSK1B, TSSK2, DGCR14, GSC2* | 74,560,555-74,570,229<br>74,581,054-74,598,153<br>74,607,593-74,609,032<br>74,612,301-74,613,377<br>74,613,480-74,620,081 | NGI |
| 18 | 0.15 | -6.40 | *LSM14A* | 44,800,804-44,854,327 | AFR |
| 19 | 0.05 | -9.05, -8.2 | *FAM101B* | 2,824,635-22,824,710<br>22,824,635-22,824,710 | AFR, NGI |
| 20 | 0.09, 0.11 | -7.94 | *GZMK, ESM1* | 24,097,290-24,107,687<br>24,131,984-24,140,817 | AFR, NGI |
| 20 | 0.11 | -7.57, -6.88 | - | 14,354,229-14,354,294 | AFR |
| 20 | 0.14 | -6.03 | *NPR3* | 40,967,082-41,041,629 | NGI |
| 24 | 0.12 | -7.32 | *ZNF407* | 3,841,029-4,197,665 | AFR |
| 27 | 0.09 | -7.36 | - | 4,996,159-4,999,264 | NGI |

Steciwko et al., 2014). Only three regions were common between AFR and DRA, and four regions were common between DRA and NGI. The rest of the predicted selective sweep

regions were breed-specific. Other putative selective sweeps such as that containing *DMRT1* gene on chromosome 8 in NGI, has been associated with human reproduction and the region detected in AFR, a Histone H4-like protein type G (*HIST1H4G*) gene on chromosome 1, has been associated with nucleosome structure of the chromosomal fibre in eukaryotes (Marzluff et al., 2002), and has also been associated with mastitis resistance in Canadian Holstein cattle (Grossi at al., 2014). These genes lie within regions of the genome that appear to have been under strong selection in many breeds of cattle. The DRA did not have selective sweeps detected with a $ZH_p$ Z-score of ≤ -6.0 Only two regions with a $ZH_p$ Z-score of -5.3 which harboured *PPP2CA,* a protein phosphatase gene which has been associated with fertility in cattle (Walker 2011)*, CDKL3,* a cyclin dependent kinase like 3*,* and *UBE2B*, a protein coding gene which has been associated with male infertility in human (Zhang et al., 2014) were detected in AFR. Most of the genes identified within sweep regions in the three breeds have unknown functions.

There were also a few overlapping common genes that were identified across all the three breeds that could have been associated with breed formation in cattle. The *KIT* and *MITF* genes on chromosomes 6 and 12 respectively, have been associated with pigmentation in cattle, *KDR* on chromosome 6 is a tyrosine kinase receptor, *ERBB4* on chromosome 2 is associated with a signalling pathway involved in the development and progression of melanocytes in human (Choi et al 2010). Other genes include *CACNA1C* on BTA5, *LAMC3* on BTA11, *TAS2R16* on BTA4, *UNC93A* on BTA9, *TNFRSF9* on BTA16, *CAV2* on BTA4 and *DCST1* on BTA3. These genes have previously been identified in selective sweep regions in cattle and have been associated with: 1) major depression, 2) the development of brain cortex and formation of axons, 3) dietary habits, 4) associated with Herpes simplex encephalitis type 1, 5) induced by lymphocyte activation, 6) involved in Cystic Fibrosis, and 7) implicated in Down syndrome, respectively (Qanbari et al., 2014). The keratin genes *KRT24, KRT25, KRT26, KRT27* and *KRT28*; and the heat shock protein gene *HSPB9* found on chromosome 19, which have previously been associated with adaptation to tropical environment in Zebu cattle, were detected in selective sweep regions common to all three breeds. Other associated genes including *ATP2B*, *FMOD*, *WNT5B* and *PRELP* on chromosome 16, have also previously been identified as being under positive selection in cattle, and were located in sweep regions shared across the three breeds.

**Identification of breed-specific SNPs**

Breed-specific SNPs were identified from among the set of novel SNPs identified in this study. From the novel SNPs, no SNPs for which alternate alleles were fixed between the three breeds were identified. Table 2 shows the candidate and putative breed-specific SNPs identified from each breed for which the SNP was variable in one breed and fixed for the same allele in both other breeds and the putative breed-specific SNPs (overlaps) that were identified in the comparison of the sequence and BovineSNP50 data.

**Table 2:** Identification of breed-specific SNPs in all three breeds based on novel SNPs.

| Breed | Novel SNPs | Candidate Breed-specific SNPs[1] | Putative breed-specific SNPs[2] | Proportion |
|---|---|---|---|---|
| AFR | 1,751,065 | 963,522 | 66 | 0.007 |
| DRA | 823,997 | 328,612 | 35 | 0.011 |
| NGI | 1,794,817 | 980,533 | 85 | 0.009 |
| **Total** | **4,369,879** | **2,272,667** | **186** | **0.027** |

[1]SNPs that are variable in the identified breed but are fixed for the same allele in the other two breeds.
[2]Breed-specific SNPs found in the sequence data that were also present on the BovineSNP50 assay.

A total of 186 putative breed-specific SNPs were identified as overlaps between the BovineSNP50 data and the 2,272,667 candidate breed-specific SNPs identified in AFR, DRA, and NGI and could be explored for their utility for breed identification using individual animal genotype data. Higher numbers of overlapping SNPs were detected in NGI (85) than in AFR (66) or DRA (35). However, the overall proportion of identified putative breed-specific SNPs was low (0.27% of total novel SNPs). These SNPs segregate within only one of the three indigenous SA breeds, but are likely to be common in European taurines (Ramey et al., 2015). These results were expected because the three breeds were not included in the design of the BovineSNP50 assay.

**Breed allocation**

The PCA plot based on the 186 chip-based breed-specific SNPs (Figure 3) shows the clustering of the three breeds based on PC1 and PC2. The PCA clearly separated the breeds including the outgroup, and identified outliers.

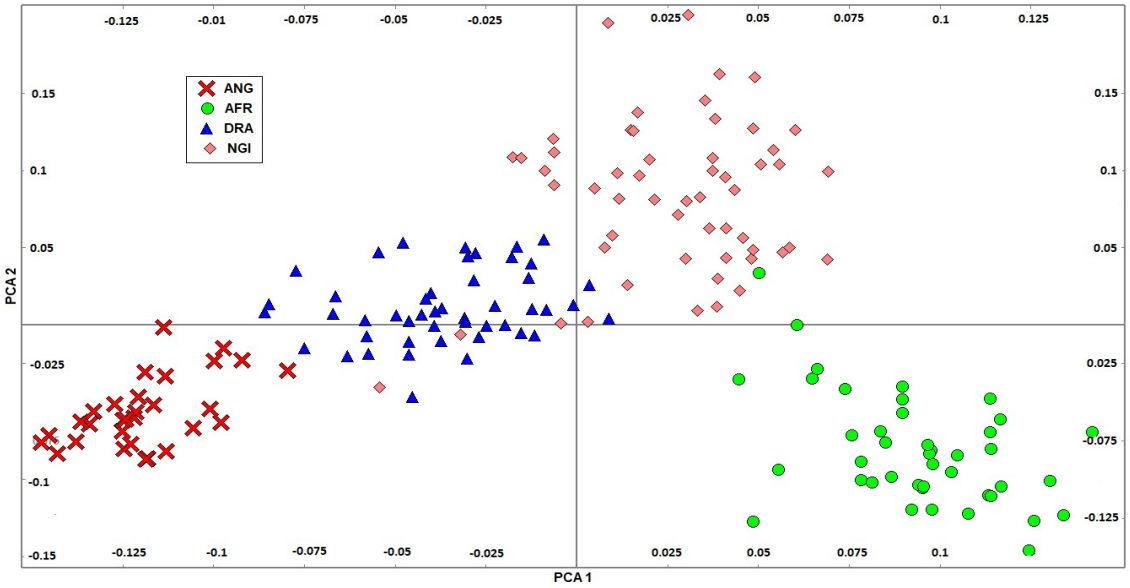


**Figure 3**: Principal component based clustering of genotyped Afrikaner (AFR), Drakensberger (DRA), and Nguni (NGI) using a panel of 186 putative breed-specific SNPs, using Angus (ANG) as an outgroup.

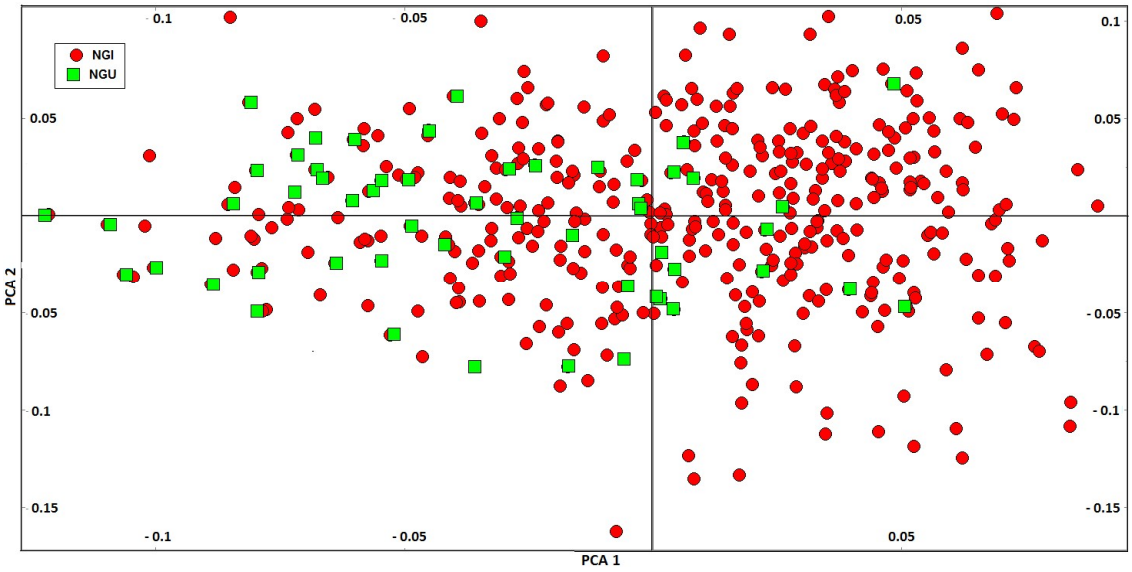


**Figure 4**: The clustering of samples from two Nguni populations (NGI and NGU) using the Nguni putative breed-specific SNPs.

When the putative breed-specific SNPs were used to assign additional 14 AFR, 23 DRA and 50 NGI animals to their breeds, the animals were correctly clustered. Figure 4 indicates the clusetering of additional NGI samples with the reference sample using NGI putative breed-specific SNPs. The breed assignment of the Nguni genotypes (Addenda B & C ), and that of Afrikaner and Drakensberger (data not shown), using GeneClass2 showed 100% breed allocation with the overall probability score equal to 1. There were no animals that were incorrectly assigned. A set of 104 breed-specific SNPs with MAF ≥ 0.2 were identified as the FREQ SNP set across the breeds, and were tested for breed assignment. There was no difference in the allocation efficiencies between the 186 putative breed-specific SNPs and the FREQ SNPs. However, these are the SNPs that were found to be highly discriminative among the three indigenous SA breeds.

**Table 3:** Distribution of minor allele frequencies (MAF) for the putative breed-specific SNPs.

| Genotyped Individuals n = 152 | | | | | |
|---|---|---|---|---|---|
| Breed | No. putative breed-specific SNPs | Minimum MAF | Maximum MAF | Average MAF | No. FREQ SNPs |
| AFR | 66 | 0.03 | 0.5 | 0.31 | 44 |
| DRA | 35 | 0.01 | 0.48 | 0.20 | 17 |
| NGI | 85 | 0.026 | 0.49 | 0.22 | 43 |
| **Total/Avg.** | **186** | **0.022** | **0.49** | **0.24** | **104** |

The MAF distribution for each putative breed-specific SNP is reported in Table 3. The average MAF of the putative breed-specific SNPs across the breeds was 0.24, with the highest frequency of 0.5. The overall maximum frequencies were similar for all of the breeds, with AFR possessing a slightly higher maximum frequency than DRA or NGI.

## Discussion & Conclussion

Improvements in next generation sequencing technologies now allow high volumes of data to be generated at a reasonable price. This study illustrated the usefulness of next generation sequencing data for the identification of selective sweeps and breed-specific SNPs in indigenous SA cattle. The design of this experiment was such that each breed was sequenced

as a pool of 30 animals to maximize the opportunity for the identification of low frequency breed-specific SNPs to discriminate between the breeds, for breed assignment purposes, and also to identify recent positive selection signatures. The results of this study indicated that the sequencing method used in this study was able to identify putative selective sweeps in the Afrikaner and Nguni breeds, represented by genomic regions harbouring SNPs with extremely low $ZH_p$ Z-scores. The method also identified putative breed-specific SNPs in the Afrikaner, Drakensberger, and Nguni breeds, rather than identifying SNPs that were fixed for alternate alleles between breeds. This appears to be due to the evolutionary history of the breeds during their migration into different geographic regions and also histories of crossbreeding and selection. Thus, species may evolve collectively at major loci through the spread of favourable alleles by crossbreeding, while simultaneously differentiating at other loci due to drift and local selection (Morjan & Rieseberg, 2004).

Studies have indicated that regions with extremly lower $ZH_p$ Z-scores indicate putative selective sweeps reflecting significant excesses of homozygosity (Choi et al., 2015). We identified 93 putative selective sweep regions with extremely low $ZH_p$ Z-scores ($\leq$ -6) which represent regions harboring loci subjected to positive selection in the Afrikaner and Nguni breeds as indicated in Table 1. These could have contributed to their adaptation to SA environmental conditions and their distinctive phenotypic characteristics. Genes that were targeted by natural selection during cattle domestication may have been differentially selected between breeds and these sweeps likely occured ~10,000 years or ~2,000 generations ago, allowing sufficient time for new mutations to accumulate in these regions which would likely lead to their not being detected by this methodology. Artificial selection, on the other hand, is the primary cause of the distinct phenotypic traits between cattle breeds and the scan for selection sweeps in genetically distinct populations, is unlikely to be confounded by their similar recent demographic histories (Chen et al., 2010).

Candidate selective sweep regions were also identified in the Drakensberger with higher $ZH_p$ Z-scores. This could indicate the recent events that have occurred in the development of the breed, but appears that few strong selective sweeps have occurred in this breed since regions with extremely low $ZH_p$ Z-scores were not found. Selective sweeps occur due to strong selection events for morphology, physiology and behaviour to human management, or due to strong artificial selection imposed by humans to increase yield, fertility, conformation, and colour patterning. As a result, more than 900 breeds, each with distinct characteristics, have

emerged throughout the world, including the indigenous SA breeds (FAO, 2007). The phenotypes associated with breed development include milk and meat production, fertility, appearance including coat coloration, decreased fearfulness, social motivation, and mild temper (Zeder, 2012). Selection for these phenotypes has left detectable signatures of selection within the genome of modern cattle, some of which appear to have been identified in this study. The common candidate selective sweeps identified between the three breeds, may reflect the similar environmental and demographic forces to which these breeds have been exposed during breed formation.

The modern bovine breeds are grouped into two major types, the taurine and indicine groups. This has led to intra- and inter-group variability in production (milk yield and quality, meat production), morphological (coat colour, presence/absence of horns) and adaptive (disease resistance, heat tolerance) traits (Gouveia et al., 2014). For the identified selective sweep regions, several genes that have previously been associated with phenotypes were identified. These regions harbour genes associated with behavioural characteristics, immune function, reproductive processes, and embryonic development (Ramey et al., 2013). Some of the selective sweep regions identified in this study contain genes with unknown function, which need to have roles established. However, the results of this study provide an insight in genomic variants that underlie complex traits in indigenous SA cattle populations.

This study identified 465 candidate selective sweeps with $ZH_p$ scores $\leq -4$ on 29 chromosomes and 223 regions were identified as putative selective sweeps ($ZH_p$ Z-scores $\leq -5$) on 17 chromosomes. Using BovineSNP50 data, Ramey et al. (2013) identified 28 genomic regions on 15 chromosomes as putatively harbouring selective sweeps in 14 breeds. They also identified 85 putative selective sweep regions from 200 – 846 kb in size using the very high density AFFXB1P assay. Only 11 regions were validated as putative selective sweeps using both assays and no selective sweeps overlapped between the taurine and indicine breeds. For several of the detected sweep regions, Ramey et al. (2013) were able to identify the phenotypes and genes that were likely subjected to selection. However, for many of these regions, the selected genes and phenotypes were unclear. But when using next generation sequencing, Qanbari et al. (2014) identified 146 regions of positive selection in non-overlapping 40 kb windows across the genome. They were able to localise regions/genes harbouring phenotypic characteristics such as patterned pigmentation, brain development and neurobehavioral functioning, sensory perception, immune system, genetic disorders, and blood coagulation.

This shows that the amount of data used and the analytical method employed both impact the identification of the number of regions of positive selection. In this study, the number of identified selective sweeps was even higher.

Using PCA, genetic distance between Afrikaner, Brahman, Drakensberger, and Nguni was obseved. There was a distinction between the three indigenous SA breeds and Brahman, and also between the Drakensberger and Brahman. The genetic distance between Drakensberger and Brahman could suggests that the Drakensberger is more closely related to B. taurus than to B. indicus as indicated in previous studies (Makina et al., 2014; Zwane et al., 2016). The study by Makina et al. (2016) also suggested that Drakensberger is an admixture of European taurine, African Taurine and indicine with a greater percentage of European taurine than African taurine or indicine. In the study by Makina et al. (2016), clustering of Afrikaner and Nguni with world-wide breeds indicated that these breeds are more African taurine than indicine, and this was also observed in this study. The PCA analysis also showed that the breeds used in this study originated from different ancestry lineage and distributed in different geographic locations. Sanga cattle were introduced to SA during migration of African tribes to southern Africa and the arrival of Europeans during the 15th century (Bachmann, 1983). Bos taurus cattle are distributed all over the world and in Africa, they were primarily found in West and Central Africa. Bos indicus represent the majority of cattle types found in Africa. These breeds are mostly found in the western and eastern parts of Africa (Mwai et al., 2015).

From the total number of candidate breed-specific SNPs identified in this study, only 0.03% could be validated for their utility for breed classification via their presence on the BovineSNP50 BeadChip (Ramos et al., 2011). The SNPs on the BovineSNP50 BeadChip were selected and optimized based on their having intermediate allele frequencies across several European taurine breeds with no regard to their allele frequencies in indigenous SA breeds (Ramos et al., 2011). Clustering of the animals by PCA using the putative breed-specific SNPs allowed the identification of outliers and indicated that the identified 186 overlapping SNPs were not completely adequate to differentiate between all animals in these three breeds. Identification of outliers among the samples could be due to admixture in some animals, or due to mislabelling of some samples prior to genotyping. Crossbreeding of different breeds to enhance production has led to Nguni-type and Afrikaner-type breeds that have never been characterised genetically (Sanarana, 2015). The fact that we have different Nguni ecotypes distributed in different geographical areas of SA could have led to the clustering of these

animals into separate clusters. However, the putative breed-specific panel identified in this study demonstrated potential for breed differentiation in these breeds.

The assignment test using a panel of 30 randomly selected putative breed-specific markers assigned the animals to their respective breed with high assignment probabilities. There was no difference between the clusters using the FREQ SNP panel and the putative breed-specific set. This shows that these SNP sets can be used for breed allocation. Studies have indicated that a set of only five SNPs could be sufficient for correctly assigning more than 95% of the pure animals belonging to a particular breed with less than a 5% misclassification rate. Ultimately, the number of SNPs required to allocate animals to a particular breed depends on the number of breeds to be allocated and the extent of crossbreeding among the animals to be allocated. If there are more breeds than the number of breed-specific SNPs used to allocate breeds, it is possible that the other breeds will not be represented in the breed-specific SNP set used, and those breeds could be misclassified into other breeds (Pant et al., 2012). In this study, the SNP set were able to allocate individuals to their respective breeds. Validation steps are still necessary, to determine real and false positive SNPs, and also to use validated SNPs to allocate crossbreds and individuals of unknown breed origin.

This study provides a broader insight into the events that happened during recent selection events and artificial selection processes that have shaped the livestock genome in SA indigenous cattle breeds. The ability to detect selective sweep regions provide useful genomic information for these breeds whereas functional analysis of these regions revealed the presence of genes of biological and economic importance. Candidate and putative selective sweep regions will be useful in identifying regions associated with important traits subjected to strong selection, while the panel of breed-specific SNPs will be used for breed assignment in SA indigenous breeds. However, further analysis to validate the breed specificity of these SNPs is needed, especially in crossbreds.

**References**

Bachmann, M., 1983. Early origins of cattle. Farmer's Weekly, 23, December, pp 18.

Bernatchez, L. & Duchesne, P., 2000. Individual-based genotype analysis in studies of parentage and population assignment: how many loci, how many alleles? Can. J. Fish. Aquat. Sci. 57, 1-12.

Blott, S.C., Williams, J.L. & Haley, C.S., 1999. Discriminating among cattle breeds using genetic markers. Heredity. 82, 613-619.

Bolger, A.M., Lohse, M. & Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinf. 30, 2114-2120.

Buchmann, R. & Hazelhurst, S., 2014. Genesis manual. University of the Witwatersrand, Johannesburg. (available at: http://www. bioinf. wits. ac. za/software/genesis/Genesis. pdf.).

Cai, Y., 2006. Doctoral dissertation: Identification of immune responsive genes in bovine airway using suppression subtractive hybridization. Oklahoma State University, USA. pp 1-58.

Chen, H., Patterson, N. & Reich, D., 2010. Population differentiation as a test for selective sweeps. Genome Res. 20, 393-402.

Choi, J., Young, J.A. & Callaway, E.M., 2010. Selective viral vector transduction of ErbB4 expressing cortical interneurons in vivo with a viral receptor–ligand bridge protein. Proc. Nat. Acad. Sci. 107, 16703-16708.

Choi, J.W., Choi, B.H., Lee, S.H., Lee, S.S., Kim, H.C., Yu, D., Chung, W.H., Lee, K.T., Chai, H.H., Cho, Y.M. & Lim, D., 2015. Whole-genome resequencing analysis of Hanwoo and Yanbian cattle to identify genome-wide SNPs and signatures of selection. Mol. Cells. 38, 466-473.

Fallahsharoudi A., de Kock N., Johnsson M., Ubhayasekera, S.J., Bergquist, J., Wright, D. & Jensen, P., 2005. Domestication effects on stress induced steroid secretion and adrenal gene expression in Chickens. Sci. Rep. 5, 15345, 1-10.

FAO, 2007. The state of the world's animal genetics resources for food and agriculture. ftp://ftp.fao.org/docrep/fao/010/a1250e/a1250e02.pdf.

Gorbach, D.M., Makgahlela, M.L., Reecy, J.M., Kemp, S.J., Baltenweck, I., Ouma, R., Mwai, O, Marshall, K., Murdoch, B., Moore, S. & Rothschild, M.F., 2010. Use of SNP genotyping to determine pedigree and breed composition of dairy cattle in Kenya. J. Anim. Breed. Genet. 127, 348-351.

Gouveia, J.J.D.S., Silva, M.V.G.B.D., Paiva, S.R. & Oliveira, S.M.P.D., 2014. Identification of selection signatures in livestock species. Genet.Mol. Biol. 37, 330-342.

Grossi, A.B., Agerholm, J.S., Christensen, K., Jensen, H.E., Leifsson, P.S., Bendixen, C., Karlskov-Mortensen, P. & Fredholm, M., 2014. A hereditary disposition for bovine peripheral nerve sheath tumors in Danish Holstein cattle. Acta Vet. Scand. 56, 85, 1-5.

Hu, Z.L., Park, C.A., Wu, X.L. & Reecy, J.M., 2013. Animal QTLdb: an improved database tool for livestock animal QTL/association data dissemination in the post-genome era. Nucleic Acids Res. 41, D871-D879.

Khodadadi, M., Mohammad, F. & Miransari, M., 2011. Genetic diversity of wheat (*Triticum aestivum L.*) genotypes based on cluster and principal component analyses for breeding strategies. Aust. J. Crop Sci. 5, 17-24

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinf. 25, 1754-1760.

Makina, S.O., Muchadeyi, F.C., Marle-Köster, E., Taylor, J.F., Makgahlela, M.L. & Maiwashe, A., 2015. Genome-wide scan for selection signatures in six cattle breeds in South Africa. Genet. Sel. Evol. 47, 92, 1-14.

Makina, S.O., Muchadeyi, F.C., van Marle-Köster, E., MacNeil, M.D. & Maiwashe, A., 2014. Genetic diversity and population structure among six cattle breeds in South Africa using a whole genome SNP panel. Front. Genet. 5, 333, 1-7.

Makina, S.O., Whitacre, L.K., Decker, J.E., Taylor, J.F., MacNeil, M.D., Scholtz, M.M., van Marle-Köster, E., Muchadeyi, F.C., Makgahlela, M.L. & Maiwashe, A., 2016. Insight into the genetic composition of South African Sanga cattle using SNP data from cattle breeds worldwide. Genet. Sel. Evol. 48, 88, 1-7.

Mapholi, N.O., 2015. Exploring genetic architecture of tick resistance in South African Nguni cattle. Doctoral dissertation, Stellenbosch University, Stellenbosch, South Africa. pp 1-120.

Mapholi, N.O., Marufu, M.C., Maiwashe, A., Banga, C.B., Muchenje, V., MacNeil, M.D., Chimonyo, M. & Dzama, K., 2014. Towards a genomics approach to tick (*Acari: Ixodidae*) control in cattle: A review. Ticks Tick Borne Dis. 5, 475-483.

Marzluff, W.F., Gongidi, P., Woods, K.R., Jin, J. & Maltais, L.J., 2002. The human and mouse replication-dependent histone genes. Genomics. 80, 487-498.

Morjan, C.L. & Rieseberg, L.H., 2004. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. Mol. Ecol. 13, 1341-1356.

Mwai, O., Hanotte, O., Kwon, Y.J. & Cho, S., 2015. African indigenous cattle: unique genetic resources in a rapidly changing world. Asian-Australas J. Anim. Sci. 28, 911-921.

Paetkau, D., Slade, R., Burden, M. & Estoup, A., 2004. Genetic assignment methods for the direct, real-time estimation of migration rate: a simulation-based exploration of accuracy and power. Mol. Ecol. 13, 55-65.

Pant, S.D., Schenkel, F.S., Verschoor, C.P. & Karrow, N.A., 2012. Use of breed-specific single nucleotide polymorphisms to discriminate between Holstein and Jersey dairy cattle breeds. Anim. Biotechnol. 23, 1-10.

Pienaar, L., 2014. MSc Thesis: Genetic diversity in the Afrikaner cattle breed. University of Freestate, Bloemfontein, South Africa. pp 1-107.

Piry, S., Alapetite, A., Cornuet, J.M., Paetkau, D., Baudouin, L. & Estoup, A., 2004. GENECLASS2: a software for genetic assignment and first-generation migrant detection. J. Hered. 95, 536-539.

Purcell, S. & Chang, C., 2014. PLINK. https://www.cog-genomics.org/plink2

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. & Sham, P.C., 2007. PLINK: A toolset for whole-genome association and population-based linkage analysis. Am. J. Hum. Genet. 81, 559-575.

Qanbari, S., Pausch, H., Jansen, S., Somel, M., Strom, T.M., Fries, R., Nielsen, R. & Simianer, H., 2014. Classic selective sweeps revealed by massive sequencing in cattle. PLoS Genetics. 10, e1004148, 1-13.

Ramey, H.R., Decker, J.E., McKay, S.D., Rolf, M.M., Schnabel, R.D. & Taylor, J.F., 2013. Detection of selective sweeps in cattle using genome-wide SNP data. BMC Genomics. 14, 382, 1-18.

Ramos, A.M., Megens, H.J., Crooijmans, R.P.M.A., Schook, L.B. & Groenen, M.A.M., 2011. Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing. Anim. Genet. 42, 613-620.

Rannala, B. & Mountain, J.L., 1997. Detecting immigration by using multilocus genotypes. Proc. Natl. Acad. Sci. 94, 9197-9201.

Rubin, C.J., Megens, H.J., Barrio, A.M., Maqbool, K., Sayyab, S., Schwochow, D., Wang, C., Carlborg, Ö, Jern, P., Jørgensen, C.B., Archibald, A.L., Fredholm, M., Groenen, M.A. & Andersson, L., 2012. Strong signatures of selection in the domestic pig genome. Proc. Natl. Acad. Sci. 109, 19529-19536.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z.P., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., Ackerman, H.C., Campbell, S.J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R. & Lander, E.S., 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419, 832-837.

Sambrook, J. & Russell, D.W., 2006. Amplification of cDNA generated by reverse transcription of mRNA. CSH Protocol 2006, pii: pdb.prot3837, 1-6.

Sanarana, Y.P., 2015. MSc Thesis: Genetic characterization of South African Nguni cattle ecotypes using microsatellite markers. University of Pretoria, South Africa, pp 1-85.

Scholtz, M., Bosman, D. J., Erasmus, G. J. & Maiwashe, A., 2010. Selection as the base of improvement in beef cattle. Beef Breeding in South Africa, 2nd Ed, pp 2-10.

Smith, J.M. & Haigh, J., 1974. The hitch-hiking effect of a favourable gene. Genet. Res. 23, 23-35.

Strydom, P.E., 2008. Do indigenous Southern African cattle breeds have the right genetics for commercial production of quality meat? Meat Sci. 80, 86-93.

Strydom, P.E., Frylinck, L. & Smith, M.F., 2011. Variation in meat quality characteristics between Sanga (*Bos taurus africanus*) and Sanga-derived cattle breeds and between Sanga and Brahman (*Bos indicus*). Animal. 5, 483-491.

van Vugt, M.A. & Medema, R.H., 2005. Getting in and out of mitosis with Polo-like kinase-1. Oncogene. 24, 2844-2859.

Walker, C.G., Littlejohn, M.D., Mitchell, M.D., Roche, J.R. & Meier, S., 2011. Endometrial gene expression during early pregnancy differs between fertile and sub-fertile dairy cow strains. Physiol. Genomics DOI, 10.1152/physiolgenomics.00254.2010, 47-58.

Zeder, M.A., 2012. The domestication of animals. J. Anthropol. Res. 68, 161-190.

Zhang, Q., Mou, L.S., Gui, Y.T. & Cai, Z.M., 2014. UBE2B gene and male infertility: an update. Zhonghua Nan. Ke. Xue. 20, 367-371.

Zhao, H., Wang, H., Jia, D., Yan, T. & Liu, F., 2014. Cluster analysis of differentially expressed heat stress genes in rat *Jejunal Mucosal*. Agric. Sci. Tech. 15, 1082-1085.

Zhao, F., McParland, S., Kearney, F., Du, L. & Donagh, P.B., 2015. Detection of selection signatures in dairy and beef cattle using high-density genomic information. Genet. Sel. Evol. 47, 92, 1-12.

Zielak-Steciwko, A.E., Browne, J.A., McGettigan, P.A., Gajewska, M., Dzieciol, M., Szulc, T. & Evans, A.C., 2014. Expression of microRNAs and their target genes and pathways associated with ovarian follicle development in cattle. Physiol. Genomics. pp.physiolgenomics-00036, 735-745.

Zwane, A.A., Maiwashe, A., Makgahlela, M. L., Choudhury, A., Taylor, J.F. & van Marle-Köster, E., 2016. Genome-wide identification of breed-informative single-nucleotide polymorphisms in three South African indigenous cattle breeds. SA J of Anim. Sci. 46, 302-312.

# CHAPTER SIX

## Critical Discussion

Next generation sequencing (NGS) methods in South Africa (SA) are still new as compared to other countries, especially developed countries. Introduction of these technologies has provided a platform for generating large amounts of genome sequence data, with broad applications in cattle and other livestock species. Even though several studies have reported reduced costs of sequencing in most countries (Caulfield et al., 2013; Muir et al., 2016), running costs and maintenance of equipment are still high in SA due to importation, coupled with the economic instability within the country. However, access to NGS technologies now guarantees improvements in the quality of livestock research and will facilitate the generation of genome-wide sequence data in a more efficient way.

In SA, studies of indigenous breeds including Afrikaner, Drakensberger and Nguni have been limited to the use of microsatellites and random amplified polymorphic DNA markers for determining the genetic diversity among these breeds (Pienaar, 2014; Sanarana, 2015). These studies have formed the basis for the characterisation of cattle breeds to understand the genetic diversity that exists within and between SA cattle. However, due to the reduced informativeness of these markers, more information was needed to characterise these breeds on a genome-wide level to increase the efficient of use of genomic data (Reyes-Valdes, 2013).

The development of SNP technology in livestock species has opened opportunities to use these markers to identify QTLs and genes underlying traits of economic importance (de Oliveira et al., 2014; Ogorevc et al., 2009; Zhang et al., 2013; Sajjanar et al., 2015). Several commercialised SNP assays with different densities are currently available for studies in cattle, including BovineSNP50, GGP-80K, GGP-150K and BovineHD. It is only recently that these assays have been used to study SA cattle breeds, especially the BovineSNP50 assay. However, these assays were designed using common SNPs primarily identified in European cattle breeds. Lower $F_{st}$ and minor allele frequency (MAF) for these SNPs were observed within and between the SA breeds relative to European breeds. With other limitations of the BovineSNP50 assays observed by Makina et al. (2014; 2015) in similar breeds, it has been shown that these assays were at most only adequate for genetic diversity studies among SA breeds but were limited in utility for other downstream genomic applications.

The main objectives of this study were to: 1) identify breed-informative markers in Afrikaner, Drakensberger and Nguni using BovineSNP50 and GGP-80K BeadChip data, 2) sequence pooled DNA samples from Afrikaner, Drakensberger and Nguni breeds using next generation sequencing to search for new genetic variants at a genome-wide level., 3) validate newly identified SNPs using Run 5 of 1000 Bull Genomes Project data and perform functional annotation and enrichment anlysis, and 4) identify selective sweeps and a panel of SNP markers to discriminate between the three indigenous breeds. The choice of the breeds for use in this study was based on their being the only three breeds that are regarded as landraces in SA in view of their historical significance (Scholtz et al., 2010). These breeds are widely used in SA for commercial beef production. Even though these breeds are still secure in terms of their numbers, their production performance is low compared to the European breeds used in SA, including the Drakensberger which has been observed to be closely related to these European breeds (Makina et al., 2014; Zwane et al., 2016).

The NGS platform established by the Biotechnology Platform of the Agricultural Research Council was the core facility for NGS sequencing used in this study. A minimum of 30 samples from each breed were collected across different regions of SA and sequenced as a pool. The samples were chosen to capture the maximum possible genetic diversity present in these breeds, covering all nine provinces of SA by including the most influential sires recorded in the ARC-Intergis database for Nguni, Afrikaner and Drakensberger cattle.

The strategy of pooling DNA samples prior to sequencing was chosen due to the cost implications associated with NGS sequencing, but nonetheless, also considering that this method has been used in SNP discovery for most species, including cattle (Van Tassell et al., 2008; Ingman & Gyllensten, 2009; Out et al., 2009; Mullen et al., 2012; Vandepitte et al., 2013; Fracassetti et al., 2015). Pooling individual DNA samples has been regarded as effective both for SNP discovery and for the estimation of allele frequencies (for population genomic analyses), and as a result, can be more cost effective due to a reduced sequencing effort required to obtain similar precisions of allele frequency estimates (Cutler & Jensen, 2010). Research has indicated that the results obtained from this sequencing strategy may depend on the model and the statistical analyses applied to examine the data (Futschik & Schlötterer, 2010). There are however, limitations associated with the pooling strategy, the inability to detect variant carriers, which is of high importance for disease-association studies of rare variants, or individual variation, of which the novel SNPs or selective sweeps identified may not be

polymorphic to all the individuals represented in the breed. Studies has also indicated that pooling does not provide variant allele frequency (VAF) estimation, which is commonly used in testing associations for case-control studies (Wang et al., 2013).

For the purpose of this study, pooling of DNA was shown to be an efficient strategy to discover approximately 17.6 million variants across the breeds including Indels, with 16% of SNPs validated as being novel SNPs in the indigenous SA breeds (Afrikaner, Drakensberger and Nguni) using the Run 5 of 1000 Genomes Project data. The proportions of novel SNPs identified in these breeds suggest that a large number of DNA variants remain to be identified in cattle. The ratio of homozygous to heterozygous loci indicated low numbers of loci with homozygous alleles which indicates that these breeds were not greatly affected by inbreeding. Therefore, considerable variation still exists between the breeds. The high number of shared SNPs between these breeds may reflect common ancestries prior to breed development. This study marks the first discovery of SNPs in these breeds and demonstrated the potential of these SNPs for determining breed composition and also for identifying trait-associated genes for breed improvement. The strategy used for SNP discovery in these breeds was successful, however, a higher genome coverage of 30X would have been preferred as only 21X coverage were realised. The number of samples used in the DNA pool could also be increased to span a greater range of breed diversity and influence the precision for allele frequency estimation.

Following genome sequencing, the next critical step was gene annotation which included marking the genomic position and structure of the genes, naming genes (gene ontology) and functional annotation, i.e., identifying their biological function (Beiki et al., 2016). In this study, functional annotation was used to classify the variants according to their functional classes and their gene ontology terms were also determined. Initial annotation of the bovine genome identified more than 22,000 genes, with 14,345 orthologs shared among seven mammalian species (Elsik et al., 2009). However, despite these efforts, the function of most genes are only partly understood. Variants were found in coding regions, non-coding regions, splice sites and in regulatory regions. The number of synonymous SNPs identified in this study was almost similar to the number of non-synonymous SNPs. Nevertheless, some advantageous quantative trait loci (QTL) alleles for economically important traits never reach complete fixation in populations because they are pleiotropic and balancing selection allows them to achieve only intermediate allele frequencies (Takasuga et al., 2015). However, these kinds of events need to be further investigated. Moreover, the presence of non-synonymous SNPs,

together with SNPs in regulatory regions, is believed to have the highest impact on phenotypes (Ramensky et al., 2002).

Most of the genes identified in this study were protein-coding that regulate biochemical processes, phenotypic characteristics and disease-related phenotypes. Candidate genes with known and unknown functions were identified in the regions enriched for novel SNPs and were distributed across the genomes of Afrikaner, Drakensberger, and Nguni based on a 100 kb sliding window. More than 400 genomic regions that were found to be enriched for novel variants were identified in Nguni. These genes may be involved in adaptation to diverse geographic environments and selection may have contributed to the shared and population-specific phenotypes in cattle populations (Scholtz et al., 2005; Abin, 2014). This information, provides a rich resource for researchers to engage in defining the real biological functions and phenotypes governed by these genes.

Selection changes the frequency of variants and their neighbouring polymorphic sites, sweeping the genome and leaving patterns that become identifiable in a population as selection signatures (Utsunomiya et al., 2013). Identification of these patterns of selection in beef cattle can assist in detecting chromosomal regions that underwent not only natural but also anthropogenic selection that may be associated with traits of economic or biological interest. Phenotypic selection has created a wide diversity of breeds that appear differently, that have adapted to different climatic conditions and are used for different purposes. The history of African cattle remains complex and is still under investigation (Decker at al., 2014; Mwai et al., 2015; Makina et al., 2016). However, it is believed that African cattle populations evolved to be adapted to various local environmental conditions including disease and parasite resistance (Mwai et al., 2015). Today, there is a variety of genetically diverse cattle populations across the country, from the purest *Bos taurus* to the nearly pure *Bos indicus* exhibiting production potential (Landry, 2015). The data obtained in this study reveals genetic characteristics that are unique to indigenous SA cattle, some of which may be used to define their adaptability and disease resistance as compared to other breeds introduced to SA.

A number of patterns of strong selection in these breeds on a genome-wide level were identified, and also regions that have been exposed to strong positive selection. These regions had low $ZH_p$ Z-scores arising from a 150 kb window scan genome-wide. A total of 465 candidate selective sweeps and 223 putative selective sweeps were identified in the three breeds. Only 93 putative selective sweeps with extremely low $ZH_p$ Z-scores ($ZH_p$ Z-scores $\leq$ -

6.0) were identified across the genomes of Afrikaner and Nguni. This information will help in the discovery of disease resistance alleles and for the inference of the events that moulded the genetic structure of these population. These imprints of historic selection/adaptation episodes left in cattle genomes allow one to interpret modern and ancestral gene origins and modifications (Qanbari et al., 2014). Searching for genomic regions of reduced variability as signatures of strong positive selection can also help in identifying causal mutations controlling selected phenotypes (Voight et al., 2006). Selective sweeps and their associated genes provide an insight into the genomic footprints left by natural and artificial selection in indigenous SA breeds. While identifying a selective sweep in the same region in different breeds provides support that a particular genomic region has undergone selection for a given trait, many selection signatures appear to be breed-specific (Gutierrez-Gil et al., 2015).

Candidate and putative breed-specific SNPs were also identified where these forms a basis for the identification of a breed-specific SNP panel for individual identification/traceability in SA cattle populations and needs further validation. However, the small set of identified chip-based SNPs can be used to identify pure indigenous animals, and will also help in the development of a breed-specific SNP panel for individual animal identification and traceability in SA cattle. The inability to identify breed-specific SNPs using BovineSNP50 and the GGP-80K Beadchip earlier in this study could be related to the limited number of SNPs used, of which only common SNPs between the two beadchips were used, and also that the design of the Beadchips didn't consider the inclusion of breed-specific SNPs, rather common SNPs between most European cattle breeds.

One of the objectives was to identify novel SNPs that are unique to the three indigenous SA breeds, Afrikaner, Drakensberger, and Nguni. Since SNPs occur at much higher density than other markers, they are useful in distinguishing closely related individuals. Panels of SNPs are often developed by comparing large quantities of DNA sequence data across multiple individuals to identify polymorphic sites (Ramos et al., 2012). When these markers are studied in different cattle populations, they properly characterise the robustness of associations of polymorphisms in candidate genes with economically important traits (Kumar et al., 2012). Results from this study provide an insight into the amount of genetic variation segregating between breeds and creates a basis for genome-wide association studies (GWAS), to elucidate the molecular mechanisms underlying disease resistance (Jiang et al., 2010). It is assumed that beneficial genetic variants that might have been lost as a result of selection in modern breeds are still segregating in the purebred populations of old breeds. Therefore, identification of

genetic variants in these breeds could provide a resource for restoring favourable alleles underlying economically important traits and correcting inherited genetic defects in cattle. Therefore, it is also essential to sequence members of the ancestral populations of these breeds to determine how the environmental or demographic factors have impacted the formation of modern cattle breeds, and how genetic drift has shaped current cattle populations.

SNPs identified in this study provide a resource for SA local beef cattle populations and the newly identified SNPs could be utilized for the development of a custom SNP chip for local use, or could be incorporated into redesigns of the existing bovine SNP chips to avoid the bias inherent in the content of current bovine SNP arrays. This will increase the efficiency of using the current bovine SNP assays and also allow the customised SNP assays for use in genomic selection and GWAS studies, for enhanced beef production in SA. Genomic regions harbouring selective sweeps provide insight into the selection processes involved in the development of the breeds that differentiates them from the other African and world breeds. Overlapping sweep regions were identified in this study that require further investigation with regard to adaptation to SA environments.

**Conclusion**

Next generation sequencing technologies have made SNP discovery affordable even in complex genomes and the technologies have improved tremendously in efficiency and cost in recent years. The SNPs and Indels identified in this study will serve as useful genetic tools, and as candidates in searches for phenotype-altering DNA differences. Novel SNPs identified in this study will provide an insight into the genomic regions that are unique to each breed. These data will also contribute to the development of a customised SNP chip for indigenous SA breeds. Candidate genes, selective sweeps and breed-specific SNPs identified in this study may assist in defining the uniqueness of these breeds. Continued characterization of genetic variation, particularly in breeds that have not been thoroughly examined, will be an important step towards decoding the molecular mechanisms underlying trait variation. More work is needed to characterise the selected genomic regions and genes identified in this study.

## Recommendations

Genetic variation in the cattle genome includes SNPs, tandem repeats, Indels, and CNVs, which may occur within regions that manifest loss of heterozygosity resulting in excessive homozygosity. The distance between these genetic variations ranges from single nucleotides to kilobases. The baseline data generated in this study could be used for further disease related studies and other genomic applications in these cattle breeds. The annotation of the novel SNPs and selective sweeps provided descriptions of protein function, gene names and identifiers, gene ontology information, and known phenotypes in cattle, humans and other model organisms such as mice. This information, in conjunction with QTL mapping or genome-wide association results (which were not part of this study), could be useful for future work aimed at a better understanding of the genetic mechanisms underlying phenotypic differences in cattle. On the other hand, further studies are needed to characterise the number of genes identified that harbour novel SNPs and that occur in selective sweep regions. This will provide insight into the traits of economic importance underlying the variations found in these studied cattle breeds. Validation of the candidate and putative breed-specific SNPs identified from Afrikaner, Drakensberger, and Nguni is necessary to develop a validated set of breed-specific SNPs for breed assignment. Allocation of unknown individuals from a larger reference population is needed, and if possible, also to identify more overlapping SNPs from the BovineHD assay. Furthermore, efforts should be made to build a customised assay for local use. It is recommended that similar studies be conducted in SA cattle breeds in order to provide genomic reference data for genetic analysis and to study their genomic characteristics for enhanced productivity.

## References

Abin, S.A., 2014. Doctoral dissertation: Animal recording as a tool for improved genetic management in African beef cattle breeds, University of Pretoria, South Africa, pp 1-122.

Beiki, H., Nejati-Javaremi, A., Pakdel, A., Masoudi-Nejad, A., Hu, Z.L. & Reecy, J.M., 2016. Large-scale gene co-expression network as a source of functional annotation for cattle genes. BMC Genomics. 17, 846, 1-13.

Bovine Genome Sequencing and Analysis Consortium, Elsik, C,G., Tellam, R.L., Worley, K.C., Gibbs, R.A., Muzny, D.M., Weinstock, G.M., Adelson, D,L., Eichler, E.E., Elnitski, L., Guigó, R., Hamernik, D.L., Kappes, S.M., Lewin, H.A., Lynn, D.J., Nicholas, F.W., Reymond,

A., Rijnkels, M., Skow, L.C., Zdobnov, E.M., Schook, L., Womack, J., Alioto, T., Antonarakis, S.E., Astashyn, A., Chapple, C.E., Chen, H.C., Chrast, J., Câmara, F., Ermolaeva, O., Henrichsen, C.N., Hlavina, W., Kapustin, Y., Kiryutin, B., Kitts, P., Kokocinski, F., Landrum, M., Maglott, D., Pruitt, K., Sapojnikov, V., Searle, S.M., Solovyev, V., Souvorov, A., Ucla, C., Wyss, C., Anzola, J.M., Gerlach, D., Elhaik, E., Graur, D., Reese, J.T., Edgar, R.C., McEwan, J.C., Payne, G.M., Raison, J.M., Junier, T., Kriventseva, E.V., Eyras, E., Plass, M., Donthu, R., Larkin, D.M., Reecy, J., Yang, M.Q., Chen, L., Cheng, Z., Chitko-McKown, C.G., Liu, G.E., Matukumalli, L.K., Song, J., Zhu, B., Bradley, D.G., Brinkman, F.S., Lau, L.P., Whiteside, M.D., Walker, A., Wheeler, T.T., Casey, T., German, J.B., Lemay, D.G., Maqbool, N.J., Molenaar, A.J., Seo, S., Stothard, P., Baldwin, C.L., Baxter, R., Brinkmeyer-Langford, C.L., Brown, W.C., Childers, C.P., Connelley, T., Ellis, S.A., Fritz, K., Glass, E.J., Herzig, C.T., Iivanainen, A., Lahmers, K.K., Bennett. A.K., Dickens, C.M., Gilbert, J.G., Hagen, D.E., Salih, H., Aerts, J., Caetano, A.R., Dalrymple, B., Garcia, J.F., Gill, C.A., Hiendleder, S.G., Memili, E., Spurlock, D., Williams, J.L., Alexander, L., Brownstein, M.J, Guan, L., Holt, R.A., Jones, S.J., Marra, M.A., Moore, R., Moore, S.S., Roberts, A., Taniguchi, M., Waterman, R.C., Chacko, J., Chandrabose, M.M., Cree, A., Dao, M.D., Dinh, H.H., Gabisi, R.A., Hines, S., Hume, J., Jhangiani, S.N., Joshi, V., Kovar, C.L., Lewis, L.R., Liu, Y.S., Lopez, J., Morgan, M.B., Nguyen, N.B., Okwuonu, G.O., Ruiz, S.J., Santibanez, J., Wright, R.A., Buhay, C., Ding, Y., Dugan-Rocha, S., Herdandez, J., Holder, M., Sabo, A., Egan, A., Goodell, J., Wilczek-Boney, K., Fowler, G.R., Hitchens, M.E., Lozado, R.J., Moen, C., Steffen, D., Warren, J.T., Zhang, J., Chiu, R., Schein, J.E., Durbin, K.J., Havlak, P., Jiang, H., Liu, Y., Qin, X., Ren, Y., Shen, Y., Song, H., Bell, S.N., Davis, C., Johnson, A.J., Lee, S., Nazareth, L.V., Patel, B.M., Pu, L.L., Vattathil, S., Williams, R.L. Jr., Curry, S., Hamilton, C., Sodergren, E., Wheeler, D.A., Barris, W., Bennett, G.L., Eggen, A., Green, R.D., Harhay, G.P., Hobbs, M., Jann, O., Keele, J.W., Kent, M.P., Lien, S., McKay, S.D., McWilliam, S., Ratnakumar, A., Schnabel, R.D., Smith, T., Snelling, W.M., Sonstegard, T.S., Stone, R.T., Sugimoto, Y., Takasuga, A., Taylor, J.F., Van Tassell, C.P., Macneil, M.D., Abatepaulo, A.R., Abbey, C.A., Ahola, V., Almeida, I.G., Amadio, A.F., Anatriello, E., Bahadue, S.M., Biase, F.H., Boldt, C.R., Carroll, J.A., Carvalho, W.A., Cervelatti, E.P., Chacko, E., Chapin, J.E., Cheng, Y., Choi, J., Colley, A.J., de Campos, T.A., De Donato, M., Santos, I.K., de Oliveira, C.J., Deobald, H., Devinoy, E., Donohue, K.E., Dovc, P., Eberlein, A., Fitzsimmons, C.J., Franzin, A.M., Garcia, G.R., Genini, S., Gladney, C.J., Grant, J.R., Greaser, M.L., Green, J.A., Hadsell, D.L., Hakimov, H.A., Halgren, R., Harrow, J.L., Hart, E.A., Hastings, N., Hernandez, M., Hu, Z.L., Ingham, A., Iso-Touru, T., Jamis, C., Jensen, K., Kapetis, D., Kerr, T., Khalil, S.S., Khatib, H.,

Kolbehdari, D., Kumar, C.G., Kumar, D., Leach, R., Lee, J.C., Li, C., Logan, K.M., Malinverni, R., Marques, E., Martin, W.F., Martins, N.F., Maruyama, S.R., Mazza, R., McLean, K.L., Medrano, J.F., Moreno, B.T., Moré, D.D., Muntean, C.T., Nandakumar, H.P., Nogueira, M.F., Olsaker, I., Pant, S.D., Panzitta, F., Pastor, R.C., Poli, M.A., Poslusny, N., Rachagani, S., Ranganathan, S., Razpet, A., Riggs, P.K., Rincon, G., Rodriguez-Osorio, N., Rodriguez-Zas, S.L., Romero, N.E., Rosenwald, A., Sando, L., Schmutz, S.M., Shen, L., Sherman, L., Southey, B.R., Lutzow, Y.S., Sweedler, J.V., Tammen, I., Telugu, B.P., Urbanski, J.M., Utsunomiya, Y.T., Verschoor, C.P., Waardenberg, A.J., Wang, Z., Ward, R., Weikard, R., Welsh, T.H. Jr., White, S.N., Wilming, L.G., Wunderlich, K.R., Yang, J. & Zhao, F,Q., 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. Science. 324, 522-528.

Caulfield, T., Evans, J., McGuire, A., McCabe, C., Bubela, T., Cook-Deegan, R., Fishman, J., Hogarth, S., Miller, F.A., Ravitsky, V., Biesecker, B., Borry, P., Cho, M.K., Carroll, J.C., Etchegary, H., Joly, Y., Kato, K., Lee, S.S., Rothenberg, K., Sankar, P., Szego, M.J., Ossorio, P., Pullman, D., Rousseau, F., Ungar, W.J. & Wilson, B., 2013. Reflections on the cost of" low-cost" whole genome sequencing: framing the health policy debate. PLoS Biol. 11, e1001699, 1-6.

Cutler, D.J. & Jensen, J.D., 2010. To pool, or not to pool? Genet. 186, 41-43.

de Oliveira, P.S., Cesar, A.S., do Nascimento, M.L., Chaves, A.S., Tizioto, P.C., Tullio, R.R., Lanna, D.P., Rosa, A.N., Sonstegard, T.S., Mourao, G.B., Reecy, J.M., Garrick, D.J., Mudadu, M.A., Coutinho, L.L., Regitano, L.C., 2014. Identification of genomic regions associated with feed efficiency in Nelore cattle. BMC Genet. 15, 100, 1-10.

Decker, J.E., Pires, J.C., Conant, G.C., McKay, S.D., Heaton, M.P., Chen, K., Cooper, A., Vilkki, J., Seabury, C.M., Caetano, A.R., Johnson, G.S., Brenneman, R.A., Hanotte, O., Eggert, L.S., Wiener, P., Kim, J., Kim, K.S., Sonstegard, T.S., Van Tassell, C.P.,. Neibergs, H.L., McEwan, J.C., Brauning, R., Coutinho, L.L., Babar, M.E., Wilson, G.A., McClure, M.C., Rolf, M.M., Kim, J., Schnabel, R.D. & Taylor, J.F., 2014. Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle. PLoS Genet. 10, e1004254, 1-14.

Fracassetti, M., Griffin, P.C. & Willi, Y., 2015. Validation of pooled whole-genome re-sequencing in Arabidopsis lyrata. PloS One. 10, e0140462, 1-15.

Futschik, A., & Schlötterer, C., 2010. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics. 186, 207-218.

Gutierrez-Gil, B., Arranz, J.J. & Wiener, P., 2015. An interpretive review of selective sweep studies in *Bos taurus* cattle populations: identification of unique and shared selection signals across breeds. Front. Genet. 6, 167, 1-20.

Ingman, M. & Gyllensten, U., 2009. SNP frequency estimation using massively parallel sequencing of pooled DNA. Eur. J. Hum. Genet. 17, 383-386.

Jiang L, Liu J, Sun D, Ma P, Ding X., Yu, Y. & Zhang, Q., 2010. Genome wide association studies for milk production traits in Chinese Holstein population. PLoS One 5, e13661, 1-12.

Kumar, S., Banks, T.W. & Cloutier, S., 2012. SNP discovery through next-generation sequencing and its applications. Int. J. Plant Genomics. 2012, 831460, 1-16.

Landry, H., 2015. Challenging evolution: How GMOs can influence genetic diversity. SITN Science In The News. pp 1-11.

Makina, S.O., Muchadeyi, F.C., van Marle-Köster, E., MacNeil, M.D. & Maiwashe, A., 2014. Genetic diversity and population structure among six cattle breeds in South Africa using a whole genome SNP panel. Front. Genet. 5, 333, 1-7.

Makina, S.O., Muchadeyi, F.C., van Marle-Köster, E., Taylor, J.F., Makgahlela, M.L. & Maiwashe, A., 2015. Genome-wide scan for selection signatures in six cattle breeds in South Africa. Genet. Sel. Evol. 47, 92, 1-14.

Makina, S.O., Whitacre, L.K., Decker, J.E., Taylor, J.F., MacNeil, M.D., Scholtz, M.M., van Marle-Köster, E., Muchadeyi, F.C., Makgahlela, M.L., Maiwashe, A., 2016. Insight into the genetic composition of South African Sanga cattle using SNP data from cattle breeds worldwide. Genet. Se.l Evol. 48, 88, 1-7.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D.J., Salichos, L., Zhang, J., Weinstock, G.M., Isaacs, F., Rozowsky, J. & Gerstein, M., 2016. The real cost of sequencing: scaling computation to keep pace with data generation. Genome Biol. 17, 78, 1-9.

Mullen, M.P., Creevey, C.J., Berry, D.P., McCabe, M.S., Magee, D.A., Howard, D.J., Killeen, A.P., Park, S.D., McGettigan, P.A., Lucy, M.C., Machugh, D.E. & Waters, S.M., 2012. Polymorphism discovery and allele frequency estimation using high-throughput DNA sequencing of target-enriched pooled DNA samples. BMC Genomics. 13, 16, 1-12.

Mwai, O., Hanotte, O., Kwon, Y.J. & Cho, S., 2015. African indigenous cattle: unique genetic resources in a rapidly changing world. Asian Australas. J. Anim. Sci. 28, 911-921.

Ogorevc, J., Kunej, T., Razpet, A. & Dovc, P., 2009. Database of cattle candidate genes and genetic markers for milk production and mastitis. Anim. Genet. 40, 832-851.

Out, A.A., van Minderhout, I.J., Goeman, J.J., Ariyurek, Y., Ossowski, S., Schneeberger, K., Weigel, D., van Galen, M., Taschner, P.E., Tops, C.M., Breuning, M.H., van Ommen, G.J., den Dunnen, J.T., Devilee, P. & Hes, F.J., 2009. Deep sequencing to reveal new variants in pooled DNA samples. Hum. Mutat. 30, 1703-1712.

Pienaar, L., 2014. MSc Thesis: Genetic diversity in the Afrikaner cattle breed. University of Freestate, Bloemfontein, South Africa. pp 1-107.

Qanbari, S., Pausch, H., Jansen, S., Somel, M., Strom, T.M., Fries, R., Nielsen, R. & Simianer, H., 2014. Classic selective sweeps revealed by massive sequencing in cattle. PLoS Genet. 10, e1004148, 1-13.

Ramensky, V., Bork, P. & Sunyaev, S., 2002. Human non-synonymous SNPs: server and survey. Nucleic acids Res. 30, 3894-3900.

Reyes-Valdes, M.H., 2013. Informativeness of microsatellite markers. Microsatellites: Methods and Protocols, pp 259-270.

Sajjanar, B., Deb, R., Singh, U., Kumar, S., Brahmane, M., Nirmale, A., Bal, S.K. &. Minhas, P.S., 2015. Identification of SNP in HSP90AB1 and its association with the relative thermotolerance and milk production traits in Indian dairy cattle. Anim. Biotech. 26, 45-50.

Sanarana, Y.P., 2015. MSc Thesis: Genetic characterization of South African Nguni cattle ecotypes using microsatellite markers. University of Pretoria, South Africa, pp 1-85.

Scholtz, M., Bosman, D. J., Erasmus, G. J., & Maiwashe, A., 2010. Selection as the base of improvement in beef cattle. Beef Breeding in South Africa, 2nd Ed, pp 2-10.

Scholtz, M.M., 2005. The role of research and the seed stock industry in the in situ conservation of livestock genetic resources. In 4th All Africa Conference on Animal Agriculture, Arusha, Tanzania, pp 311-316.

Takasuga, A., Sato, K., Nakamura, R., Saito, Y., Sasaki, S., Tsuji, T., Suzuki, A., Kobayashi, H., Matsuhashi, T., Setoguchi, K., Okabe, H., Ootsubo, T., Tabuchi, I., Fujita, T., Watanabe, N., Hirano, T., Nishimura, S., Watanabe, T., Hayakawa, M., Sugimoto, Y. & Kojima, T., 2015. Non-synonymous FGD3 variant as positional candidate for disproportional tall stature accounting for a carcass weight QTL (CW-3) and skeletal dysplasia in Japanese Black cattle. PLoS Genet. 11, e1005433, 1-22.

Utsunomiya, Y.T., O'Brien, A.M.P., Sonstegard, T.S., Van Tassell, C.P., do Carmo, A.S., Meszaros, G., Sölkner, J. & Garcia, J.F., 2013. Detecting loci under recent positive selection in dairy and beef cattle by combining different genome-wide scan methods. PloS One. 8, e64280, 1-11.

Van Tassell, C.P., Smith, T.P., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C. & Sonstegard, T.S., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat. Meth. 5, 247-252.

Vandepitte, K., Honnay, O., Mergeay, J., Breyne, P., Roldán-Ruiz, I. & Meyer, T., 2013. SNP discovery using Paired-End RAD-tag sequencing on pooled genomic DNA of Sisymbrium austriacum (Brassicaceae). Mol. Ecol. Resour. 13, 269-275.

Voight B. F., Kudaravalli S., Wen X., Pritchard J. K., 2006. A map of recent positive selection in the human genome. PLoS Biol. 4, e72, 1-13.

Wang, W., Yin, X., Pyon, Y.S., Hayes, M. & Li, J., 2013. Rare variant discovery and calling by sequencing pooled samples with overlaps. Bionf. 29, 29-38.

Zhang, L., Liu, J., Zhao, F., Ren, H., Xu, L., Lu, J., Zhang, S., Zhang, X., Wei, C., Lu, G., Zheng, Y. & Du, L., 2013. Genome-wide association studies for growth and meat production traits in sheep. PloS One. 8, e66569, 1-12.

Zwane, A.A., Maiwashe, A., Makgahlela, M. L., Choudhury, A., Taylor, J.F. & van Marle-Köster, E., 2016. Genome-wide identification of breed-informative single-nucleotide polymorphisms in three South African indigenous cattle breeds. S. Afr. J. Anim. Sci. 46, 302-312.

# ADDENDA

**Addendum A**: Candidate selective sweep regions with $ZH_p$ Z-scores ≤ -4 and their associated genes in all the breeds.

| CHR | ZHp | ZHp Z-Scores | Loci | Coordinates | Breed |
|---|---|---|---|---|---|
| 1 | 0.22 | -4.81 | *,WDR4* | 144,607,395-144,607,589<br>144,656,767-144,676,874 | AFR |
| 1 | 0.23 | -4.49 | , | 143,368,303-143,368,674<br>143,370,686-143,371,060 | AFR |
| 1 | 0.23 | -4.47 | , | 145,996,975-145,997,666<br>146,010,006-146,010,491 | AFR |
| 1 | 0.24 | -4.35 | *WRB,LCA5L,,* | 140,944,009-140,958,949<br>140,971,623-141,013,360<br>140,992,954-140,993,023<br>140,993,153-140,993,232 | AFR |
| 1 | 0.24 | -4.25 | *GRK7* | 128,144,744-128,185,018 | AFR |
| 1 | 0.24 | -4.16 | *LMLN,,,IQCG* | 70,742,636-70,784,188<br>70,786,714-70,786,824<br>70,788,697-70,792,140<br>70,792,452-70,838,105 | AFR |
| 1 | 0.18 | -5.21 | *KIAA1524,DZIP3* | 53,727,496-53,754,034<br>53,768,326-53,884,123 | NGI |
| 1 | 0.18 | -5.11 | *TMEM108,* | 137,111,101-137,126,880<br>137,147,859-137,148,728 | NGI |
| 1 | 0.21 | -4.51 | | 137,030,485-137,107,221 | NGI |
| 1 | 0.22 | -4.33 | | 53,913,505-53,913,597 | NGI |
| 1 | 0.23 | -4.11 | *NDUFV3* | 144,690,277-144,701,618 | AFR, NGI |
| 2 | 0.18 | -5.89 | *PMS1* | 6,395,005-6,511,175 | AFR |
| 2 | 0.18 | -5.66 | *B3GNT7* | 119,989,628-119,990,869 | AFR |
| 2 | 0.20 | -5.39 | | *119,912,874-119,912,980* | AFR |
| 2 | 0.20 | -5.17 | | 91,475,505-91,548,842 | AFR |
| 2 | 0.22 | -4.80 | | 20,305,577-20,305,683 | AFR |
| 2 | 0.24 | -4.21 | *LYPD6* | 46,723,405-46,759,004 | AFR |
| 2 | 0.24 | -4.18 | *,HECW2* | 85,132,987-85,133,945<br>85,136,811-85,285,250 | AFR |
| 2 | 0.25 | -4.12 | *TNFAIP6* | 44,850,892-44,867,293 | AFR |
| 2 | 0.05 | -4.51 | *GCG* | 34,398,011-34,408,391 | DRA |
| 2 | 0.08 | -4.16 | | 23,443,013-23,608,298 | DRA |
| 2 | 0.16 | -5.61 | | 91,475,505-91,548,842 | NGI |
| 2 | 0.20 | -4.68 | *RIF1* | 44,763,141-44,818,737 | NGI |
| 2 | 0.22 | -4.21 | | 24,807,357-24,860,776 | NGI |
| 2 | 0.22 | -4.19 | | 94,382,387-94,382,773 | NGI |
| 2 | 0.23 | -4.13 | *NCKAP1* | 13,575,102-13,664,827 | NGI |
| 2 | 0.23 | -4.00 | *,TRIP12* | 118,596,201-118,598,002<br>118,622,220-118,729,375 | NGI |
| 3 | 0.18 | -5.70 | *FAM63A,ANXA9,*<br>*CERS2,SETDB1* | 19,808,483-19,817,752<br>19,818,789-19,828,310 | |

| | | | | | |
|---|---|---|---|---|---|
| | | | | 19,835017-19843638<br>19,844,271-19,875,888 | AFR |
| 3 | 0.19 | -5.44 | *HMGCS2,PHGDH* | 23,643,772-23,667,741<br>23,672,339-23,703,529 | AFR |
| 3 | 0.22 | -4.77 | *,EPHX4* | 51,490,101-51,491,654<br>51,494,275-51,531,276 | AFR |
| 3 | 0.22 | -4.70 | *,BRDT* | 51,546,136-51,565,701<br>51574126-51587278 | AFR |
| 3 | 0.22 | -4.66 | *SCMH1* | 105,660,418-105,788,507 | AFR |
| 3 | 0.23<br>0.08 | -4.61<br>-4.10 | *MCOLN3,MCOLN2* | 59,252,316-59,280,289<br>59,303,476-59,351,124 | AFR, DRA |
| 3 | 0.24<br>0.22 | -4.40<br>-4.33 | *SLAMF9,IGSF9<br>,TAGLN2,* | 9,848,246-9,851,355<br>9,859,945-9,877,224<br>9,878,839-9,886,647<br>9,881,758-9,881,830 | AFR, NGI |
| 3 | 0.24 | -4.23 | *TCEANC2,,TMEM59* | 92,701,749-92,745,663<br>92,725,890-92,726,019<br>92,746,059-92,771,330 | AFR |
| 3 | 0.06 | -4.38 | *LPAR3* | 59,403,909-59,459,794 | DRA |
| 3 | 0.08 | -4.28 | *MTF2* | 50,490,450-50,560,463 | DRA |
| 3 | 0.17 | -5.40 | *SLC30A7,* | 42,465,567-42,559,234<br>42,476,551-42,476,656 | NGI |
| 3 | 0.17 | -5.36 | *MCOLN3,MCOLN2* | 59,252,316-59,280,289<br>59,303,476-59,351,124 | NGI |
| 3 | 0.20 | -4.79 | *,DR1* | 50,229,830-50,229,964<br>50,242,939-50,265,059 | NGI |
| 3 | 0.20 | -4.67 | *SPRR3,* | 17,796,522-17,798,340<br>17,814,055-17,814,288 | AFR, NGI |
| 3 | 0.20 | -4.66 | *FCRL6,DUSP23* | 9,991,099-10,000,540<br>10,029,474-10,030,882 | NGI |
| 3 | 0.22 | -4.30 | *TMED5* | 50,448,643-50,467,156 | NGI |
| 3 | 0.22 | -4.17 | *RGS5,RGS4* | 6,228,349-6,426,528<br>6,287,012-6,292,402 | NGI |
| 4 | 0.21 | -5.15 | *HIBADH,* | 68,926,598-69,034,386<br>68,943,157-68,943,227 | AFR |
| 4 | 0.23 | -4.60 | *ZPBP* | 5,727,483-5,820,313 | AFR |
| 4 | 0.09 | -4.06 | | 26,718,560-26,718,680 | DRA |
| 4 | 0.17 | -5.50 | *,IMPDH1* | 93,379,463-93,382,232<br>93,382,312-93,399,661 | NGI |
| 4 | 0.17 | -5.35 | *STRIP2* | 94,153,704-94,200,845 | NGI |
| 4 | 0.22 | -4.24 | *CALU,OPN1SW,CC<br>DC136* | 93,532,937-93,554,971<br>93,557,351-93,560,340<br>93,575,777-93,602,932 | NGI |
| 4 | 0.22 | -4.18 | *,CRHR2* | 66,002,072-66,006,203<br>66,062,161-66,090,181 | NGI |
| 4 | 0.23 | -4.14 | *TSPAN33* | 93,881,396-93,901,223 | NGI |
| 4 | 0.23 | -4.13 | *FLNC,,,,KCP,* | 93,609,408-93,636,607<br>93,619,560-93,619,627<br>93,640,142-93,642,824<br>93,643,427-93,645,493 | |

| | | | | | |
|---|---|---|---|---|---|
| | | | | 93,655,277-93,685,724<br>93,662,296-93,662,746 | NGI |
| 5 | 0.19 | -5.48 | ,, | 44,351,615-44,357,569<br>44,366,491-44,366,579<br>44,390,512-44,390,615 | AFR |
| 5 | 0.20 | -5.17 | ,*YEATS4* | 44,255,549-44,255,662<br>44,318,357-44,342,854 | AFR |
| 5 | 0.21 | -5.01 | *PUS7L* | 36,909,426-36,926,975 | AFR |
| 5 | 0.22 | -4.87 | *TCP11L2* | 70,013,667-70,054,229 | AFR |
| 5 | 0.24 | -4.38 | | 99,192,315-99,193,247 | AFR |
| 5 | 0.25 | -4.089 | , | 99,230,573-99,231,472<br>99,242,115-99,243,044 | AFR |
| 5 | 0.17 | -5.34 | *PCBP2,,PRR13,AMH R2,SP1* | 26,702,879-26,723,947<br>26,719,615-26,719,721<br>26,730,147-26,733,390<br>26,753,574-26,759,487<br>26,769,555-26,804,655 | AFR, NGI |
| 5 | 0.20 | -4.82 | *NELL2* | 35,657,329-36,042,694 | NGI |
| 5 | 0.21 | -4.56 | *IRAK3* | 47,796,256-47,839,675 | NGI |
| 5 | 0.21 | -4.53 | | 47,602,843-47,603,007 | NGI |
| 5 | 0.22 | -4.37 | *HELB,* | 47,713,520-47,751,430<br>47,736,728-47,737,193 | NGI |
| 5 | 0.22 | -4.25 | *WIF1,* | 48,917,722-49,009,466<br>48,952,298-48,952,372 | NGI |
| 6 | 0.18 | -5.72 | *NIPAL1* | 68,364,432-68,391,572 | AFR |
| 6 | 0.20 | -5.34 | | 5,474,607-5,499,008 | AFR |
| 6 | 0.21 | -4.91 | ,*TEC* | 68,405,247-68,449,203<br>68,468,930-68,530,673 | AFR |
| 6 | 0.23 | -4.53 | ,*NOA1* | 73,986,000-73,997,338<br>74,016,234-74,026,536 | AFR |
| 6 | 0.07 | -4.32 | *GABRA2* | 66,519,802-66,605,788 | DRA |
| 6 | 0.07<br>0.20 | -4.24<br>-4.77 | *UVSSA* | 109,351,917-109,372,113 | DRA, NGI |
| 6 | 0.08 | -4.20 | | 85,709,660-85,709,770 | DRA |
| 6 | 0.16 | -5.65 | | 104,991,391-105,015,210 | NGI |
| 6 | 0.18 | -5.17 | | 6,499,944-6,500,030 | NGI |
| 6 | 0.18 | -5.15 | *AFF1* | 103,688,011-103,825,580 | NGI |
| 6 | 0.22 | -4.24 | *BOD1L1* | 113,647,965-113,701,414 | NGI |
| 7 | 0 | -5.267 | *PPP2CA* | 47,425,980-4,745,0747<br>47,450,517-47,450,612 | DRA |
| 7 | 0 | -5.26 | *CDKL3,*<br>*UBE2B* | 47,493,409-47,524,776<br>47,528,581-47,541,198 | DRA |
| 7 | 0.04 | -4.67 | *TCF7,* | 47,352,223-47,385,009<br>47,389,357-47,401,412 | DRA |
| 7 | 0.08 | -4.23 | *DDX46,C5orf24* | 47,852,147-47,903,296<br>47,910,525-47,917,813 | DRA |
| 7 | 0.22 | -4.73 | *LYL1,*<br>*NFIX* | 13,585,511-13,587,648<br>13,596,367-13,658,112 | AFR |
| 7 | 0.24 | -4.31 | *STK32A* | 60,450,229-60,579,589 | AFR |

| | | | | | |
|---|---|---|---|---|---|
| 7 | 0.24 | -4.22 | *,COL23A1,* | 40,670,261-40,670,368<br>40,675,390-40,700,658<br>40,722,342-40,722,448 | AFR |
| 7 | 0.24 | -4.17 | *CAPS,*<br>*VMAC,* | 19,691,006-19,692,540<br>19,694,193-19,697,346<br>19,698,987-19,704,035 | AFR |
| 7 | 0.25 | -4.08 | | 136,889-136,976 | AFR |
| 7 | 0.05 | -4.51 | | 81,396,101-81,396,804 | DRA |
| 7 | 0.064 | -4.39 | , | 49,059,242-49,095,515<br>49,108,507-49,108,599 | DRA |
| 7 | 0.07 | -4.29 | *SMAD5* | 49,155,483-49,217,780 | DRA |
| 7 | 0.07 | -4.26 | *TIFAB* | 48,524,556-48,525,038 | DRA |
| 7 | 0.07 | -4.23 | *CXXC5* | 52,505,097-52,513,059 | DRA |
| 7 | 0.09 | -4.07 | *,TXNDC15* | 47,929,942-47,930,060<br>47,943,546-47,959,145 | DRA |
| 7 | 0.09 | -4.05 | | 48,858,234-48,904,411 | DRA |
| 7 | 0.22 | -4.33 | *RIOK2* | 99,013,511-99,035,256 | NGI |
| 8 | 0.22 | -4.79 | *RNF20* | 92,911,255-92,935,750 | AFR |
| 8 | 0.23 | -4.61 | | 33,178,579-33,178,676 | AFR |
| 8 | 0.23<br>0.08 | -4.50<br>-4.10 | *,DMRT3* | 43,826,048-43,826,141<br>43,855,333-43,867,721 | AFR, DRA |
| 8 | 0.25 | -4.09 | *HABP4* | 84,579,583-84,615,685 | AFR |
| 8 | 0.03 | -4.91 | *KANK1* | 44,046,426-44,076,904 | DRA, NGI |
| 8 | 0.08 | -4.14 | | 66.517.407-6.6517.687 | DRA |
| 8 | 0.15 | -5.93 | | 66,517,407-66,517,687 | NGI |
| 8 | 0.18 | -5.27 | *,DMRT3* | 43,826,048-43,826,141<br>43,855,333-43,867,721 | NGI |
| 8 | 0.19 | -4.97 | | 30,260,472-30,260,561 | NGI |
| 8 | 0.21 | -4.50 | *ERCC6L2* | 83,978,960-84,133,609 | NGI |
| 9 | 0.19 | -5.52 | | 49,837,663-49,839,034 | AFR |
| 9 | 0.22 | -4.74 | | 32,980,714-32,980,839 | AFR |
| 9 | 0.24<br>0.20 | -4.28<br>-4.70 | *TRAF3IP2* | 39,309,720-39,341,546 | AFR, NGI |
| 9 | 0.04 | -4.75 | | 98,016,582-98,064,203 | DRA |
| 9 | 0.06 | -4.49 | *SLC22A3* | 97,917,373-98,014,739 | DRA |
| 9 | 0.07 | -4.24 | *SLC22A1* | 97,750,746-97,787,561 | DRA |
| 9 | 0.09 | -4.06 | *,RPF2,* | 39,797,081-39,797,187<br>39,804,099-39,840,649<br>39,819,593-39,820,552 | DRA |
| 9 | 0.16 | -5.77 | , | 27,991,134-27,991,829<br>28,046,280-28,046,356 | NGI |
| 9 | 0.18 | -5.24 | , | 403,685-403,791<br>409,005-409,418 | NGI |
| 9 | 0.19 | -4.83 | | 49,837,663-49,839,034 | NGI |
| 9 | 0.22 | -4.19 | *SOBP* | 42,904,974-43,070,235 | NGI |
| 10 | 0.22 | -4.86 | *,,OR6J1,,,,ABHD4* | 21,975,580-21,976,545<br>21,994,853-21,995,149 | |

| | | | | | |
|---|---|---|---|---|---|
| | | | | 22,007,817-22,009,402<br>22,015,188-22,015,810<br>22,018,182-22,018,744<br>22,038,226-22,038,788<br>22,045,427-22,056,366 | AFR |
| 10 | 0.17 | -5.51 | *KIAA1191,SIMC1* | 4,950,671-4,964,129<br>4,979,152-5,011,240 | AFR, NGI |
| 10 | 0.23 | -4.02 | *,RAB11A,MEGF11* | 12,690,727-12,690,806<br>12,691,046-12,710,296<br>12,721,325-12,807,326 | NGI |
| 11 | 0.08 | -4.12 | *EPCAM,,MSH2* | 29,626,087-29,636,759<br>29,645,318-29,645,452<br>29,648,305-29,730,536 | DRA |
| 11 | 0.23 | -4.51 | *SLC2A8,*<br>*ZNF79* | 98,161,639-98,170,664<br>98,174,566-98,189,943 | AFR |
| 11 | 0.06 | -4.40 | *XDH* | 14,176,298-14,281,717 | DRA |
| 11 | 0.22 | -4.35 | | 14,162,023-14,166,468 | NGI |
| 11 | 0.22 | -4.23 | | 56,863,702-56,863,810 | NGI |
| 12 | 0.04 | -4.69 | *URAD,* | 32,300,110-32,309,281<br>32,317,772-32,322,605 | DRA |
| 12 | 0.06 | -4.43 | *,* | 36,491,414-36,491,611<br>36,493,992-36,520,152 | DRA |
| 12 | 0.07 | -4.35 | *MPHOSPH8,PARP4,* | 36,608,008-36,653,042<br>36,656,631-36,727,198<br>36,660,548-36,660,653 | DRA |
| 12 | 0.22 | -4.78 | | 39,640,396-39,640,466 | AFR |
| 12 | 0.23 | -4.55 | | 74,816,045-74,978,179 | AFR |
| 12 | 0.23 | -4.46 | *,* | 36,491,414-36,491,611<br>36,493,992-36,520,152 | AFR |
| 12 | 0.03 | -4.83 | *PSPC1* | 36,529,631-36,579,763 | DRA |
| 12 | 0.09 | -4.08 | *RXFP2* | 29,234,959-29,280,832 | DRA |
| 12 | 0.21 | -4.59 | | 39,640,396-39,640,466 | NGI |
| 13 | 0.17 | -5.92 | *OTUD1* | 24,655,214-24,656,659 | AFR |
| 13 | 0.21 | -5.12 | | 12,039,484-12,039,612 | AFR |
| 13 | 0.08 | -4.15 | *MKX* | 37,127,926-37,198,158 | DRA |
| 13 | 0.09 | -4.01 | *DDX27* | 78,054,502-78,071,705 | DRA |
| 13 | 0.22 | -4.15 | *SCP2D1* | 39,015,341-39,016,050 | NGI |
| 14 | 0.07<br>0.25 | -4.33<br>-4.13 | *DPYS,* | 62,328,132-62,418,497<br>62,374,866-62,374,961 | AFR, DRA,<br>NGI |
| 14 | 0.21 | -5.09 | *NAPRT,MROH6,,*<br>*ZC3H3* | 2,327,870-2,331,019<br>2,332,751-2,337,785<br>2,341,290-2,346,302<br>2,354,390-2,418,557 | AFR |
| 14 | 0.25 | -4.08 | *OPRK1* | 23,373,836-23,395,443 | AFR |
| 14 | 0.07 | -4.29 | *TAF2* | 83,552,370-83,639,183 | DRA |
| 14 | 0.21 | -4.37 | *DPYS,* | 62,328,132-62,418,497<br>62,374,866-62,374,961 | NGI |

| | | | | | |
|---|---|---|---|---|---|
| 14 | 0.22 | -4.24 | *PABPC1* | 65,816,006-65,833,756 | NGI |
| 14 | 0.23 | -4.05 | *YWHAZ* | 65,584,487-65,617,329 | NGI |
| 15 | 0.17 | -5.94 | | 6,750,617-6,817,541 | AFR |
| 15 | 0.18 | -5.84 | | 54,131,056-54,171,357 | AFR |
| 15 | 0.19 | -5.63 | | 65,457,710-65,497,385 | AFR |
| 15 | 0.23 | -4.61 | *KCNE3* | 54,587,990-54,588,289 | AFR |
| 15 | 0.23 | -4.42 | ,, | 50,562,717-50,563,915<br>50,576,452-50,577,620<br>50,601,202-50,602,143 | AFR |
| 15 | 0.25 | -4.15 | | 61,420,030-61,420,149 | AFR |
| 15 | 0.25 | -4.07 | *OR51M1* | 48,900,173-48,901,135 | AFR |
| 15 | 0.07 | -4.35 | *CREB3L1* | 177,157,186-77,193,643 | DRA |
| 16 | 0.25 | -4.03 | | 55,803,157-55,806,065 | AFR |
| 16 | 0.25 | -4.00 | *ETNK2,REN,KISS1, GOLT1A* | 1,734,650-1,749,646<br>1,752,891-1,762,733<br>1,774,374-1,777,002<br>1,781,213-1,794,807 | AFR |
| 16 | 0.21 | -4.43 | | 19,247,162-19,247,268 | NGI |
| 16 | 0.21 | -4.42 | *RASAL2* | 61,121,078-61,314,123 | NGI |
| 17 | 0.18<br>0.22 | -5.68<br>-4.26 | *FBRSL1* | 45,613,860-45,696,523 | AFR, NGI |
| 17 | 0.24 | -4.40 | *ASPHD2,HPS4, SRRD* | 68,354,300-68,362,110<br>68,368,661-68,396,020<br>68,396,102-68,420,240 | AFR |
| 17 | 0.24 | -4.22 | *HSPB8* | 58,405,437-58,418,688 | AFR |
| 17 | 0.25 | -4.10 | *SLC25A1,HIRA,* | 74,633,475-74,635,952<br>74,663,544-74,697,337<br>74,676,730-74,677,594 | AFR |
| 17 | 0.25 | -4.01 | | 10,817,956-10,818,059 | AFR |
| 17 | 0.09<br>0.18 | -4.03<br>-5.30 | *KCTD10,MYO1H* | 65,950,166-65,979,498<br>65,981,418-66,022,389 | DRA, NGI |
| 17 | 0.21 | -4.40 | *,YPEL1,PPIL2,* | 74,049,772-74,053,881<br>74,069,625-74,073,387<br>74,073,547-74,092,409<br>74,095,456-74,098,680 | NGI |
| 17 | 0.08 | -4.15 | *ASIC5* | 44,427,939-44,483,562 | AFR, DRA, NGI |
| 17 | 0.09 | -4.07 | | 61,342,997-61,343,081 | DRA |
| 18 | 0.21 | -4.46 | *SNX20,NOD2* | 19,157,447-19,166,134<br>19,181,972-19,212,607 | NGI |
| 18 | 0.20<br>0.22 | -5.25<br>-4.30 | *GPI* | 44,979,578-45,007,642 | AFR, NGI |
| 18 | 0.09 | -4.02 | *VPS9D1,ZNF276,FA NCA,SPIRE2* | 14,625,180-14,634,572<br>14,635,302-14,649,693<br>14,649,534-14,686,976<br>14,694,452-14,723,123 | DRA |
| 18 | 0.21 | -5.14 | *CDH3* | 36,095,734-36,140,923 | AFR |
| 18 | 0.21 | -5.05 | *ZDHHC7,KIAA0513* | 11,116,413-11,133,102 | |

| | | | | | |
|---|---|---|---|---|---|
| | | | | 11,172,008-11,188,928 | AFR |
| 18 | 0.21 | -5.02 | *COTL1* | 10,781,076-10,825,733 | AFR |
| 18 | 0.22 | -4.88 | *DPEP1,CHMP1A, CDK10,SPATA2L* | 14,578,306-14,584,239<br>14,592,688-14,599,518<br>14,610,728-14,617,868<br>14,617,872-14,621,516 | AFR |
| 18 | 0.24 | -4.35 | *CMIP* | 8,210,815-8,289,563 | AFR |
| 19 | 0.15 | -5.84 | | 36,032,858-36,035,427 | NGI |
| 19 | 0.19 | -4.87 | *TBCD* | 50,388,667-50,536,976 | NGI |
| 19 | 0.18 | -5.83 | *RPA1* | 23,476,634-23,527,009 | AFR |
| 19 | 0.21 | -4.98 | *SP6,,LRRC46,MRPL 10,OSBPL7* | 39,229,249-39,230,376<br>39,237,358-39,240,647<br>39,241,151-39,245,578<br>39,245,741-39,251,565<br>39,254,140-39,265,047 | AFR |
| 19 | 0.24 | -4.16 | *PRR15L,PNPO,SP2* | 39,171,884-39,177,343<br>39,181,604-39,188,148<br>39,194,570-39,296,546 | AFR |
| 19 | 0.03 | -4.87 | *,FAM101B,,,* | 22,824,635-22,824,710<br>22,835,876-22,841,764<br>22,856,214-22,856,325<br>22,856,388-22,859,271<br>22,872,042-23,011,338 | AFR,DRA, NGI |
| 20 | 0.17 | -5.54 | *PRKAA1,TTC33* | 33,688,291-33,716,677<br>33,722,080-33,747,493 | NGI |
| 20 | 0.19 | -4.99 | *CARD6,RPL37,* | 33,638,644-33,653,345<br>33,667,205-33,669,805<br>33,669,525-33,669,606 | NGI |
| 20 | 0.19 | -4.93 | *OTULIN* | 58,563,065-58,596,022 | NGI |
| 20 | 0.20 | -4.78 | *C7* | 33,549,495-33,606,517 | NGI |
| 20 | 0.20 | -4.67 | *C6* | 33,328,558-33,405,555 | NGI |
| 20 | 0.22 | -4.21 | *FAM105A* | 58,634,394-58,664,249 | NGI |
| 20 | 0.22 | -4.17 | *,* | 24,033,747-24,042,721<br>24,048,537-24,059,238 | NGI |
| 20 | 0.06 | -4.38 | | 22,800,133-22,800,236 | DRA |
| 21 | 0.17 | -5.51 | *SNRPA1* | 29,683,789-29,696,818 | NGI |
| 21 | 0.20 | -4.70 | | 26,874,338-26,962,870 | NGI |
| 21 | 0.19<br>0.22 | -5.45<br>-4.21 | *C14orf2,,TDRD9* | 70,113,045-70,118,400<br>70,114,200-70,114,755<br>70,123,917-70,233,306 | AFR, NGI |
| 21 | 0.20 | -5.34 | *HMG20A* | 33,083,549-33,160,047 | AFR |
| 21 | 0.21 | -5.14 | | 17,516,746-17,516,849 | AFR |
| 21 | 0.04 | -4.66 | | 28,785,430-28,785,527 | DRA |
| 22 | 0.16 | -5.56 | *TNNC1,SEMA3G,PH F7,,DNAH1* | 48,988,984-48,991,875<br>48,999,619-49,008,847<br>49,019,728-49,032,603<br>49,032,830-49,039,996<br>49,042,073-49,118,359 | NGI |
| 22 | 0.18 | -5.11 | *,,,,* | 5,650-9,019 | |

| | | | | | |
|---|---|---|---|---|---|
| | | | | 23,848-26,284<br>33,124-34,397<br>63,640-65,183<br>85,296-127,878 | NGI |
| 22 | 0.23 | -4.00 | *POC1A* | 49,282,960-49,359,377 | NGI |
| 22 | 0.23 | -4.61 | *,* | 20,840,724-20,840,830<br>20,841,408-20,841,516 | AFR |
| 22 | 0.07 | -4.31 | | 5,838,522-5,838,627 | DRA |
| 23 | 0.22<br>0.23 | -4.80<br>-4.11 | *ARMC12,CLPS,*<br>*LHFPL5* | 9,683,947-9,694,015<br>9,723,791-9,726,250<br>9,733,331-9,739,310 | AFR, NGI |
| 24 | 0.23 | -4.09 | *ZNF407* | 3,841,029-4,197,665 | AFR, NGI |
| 25 | 0.15 | -5.84 | *AQP8,ZKSCAN2* | 23,069,274-23,079,432<br>23,091,497-23,104,117 | NGI |
| 25 | 0.15 | -5.82 | *DCTN5,PLK1,ERN2* | 21,541,012-21,574,827<br>21,587,069-21,597,613<br>21,597,704-21,620,834 | NGI |
| 25 | 0.22 | -4.19 | *,C1QTNF8,* | 857,167-858,273<br>865,752-867,111<br>871,131-877,122 | NGI |
| 25 | 0.23 | -4.01 | *IL21R* | 25,233,186-25,268,874 | NGI |
| 25 | 0.03 | 4.78 | *SLC9A3R2,NTHL1,*<br>*TSC2,PKD1,* | 1,575,649-1,589,619<br>1,590,252-1,595,934<br>1,596,730-1,626,967<br>1,627,978-1,666,088<br>1,628,447-1,628,539 | DRA |
| 25 | 0.06 | -4.40 | *MLST8,BRICD5,,,,*<br>*E4F1,DNASE1L2,*<br>*ECI1,,,ABCA3* | 1,738,258-1,742,107<br>1,742,073-1,743,503<br>1,744,661-1,744,720<br>1,744,908-1,744,984<br>17,46,200-1,747,344<br>1,753,636-1,762,994<br>1,763,777-1,766,831<br>1,767,070-1,779,723<br>1,780,540-1,789,539<br>1,792,865-1,792,957<br>1,796,660-1,828,217 | DRA |
| 25 | 0.07 | -4.26 | *,,,,,,,,NOXO1,GFER,,*<br>*ZNF598,NPW* | 1,508,568-1,515,379<br>1,517,626-1,520,287<br>1,520,493-1,522,670<br>1,520,844-1,520,974<br>1,521,490-1,521,623<br>1,522,971-1,523,097<br>1,524,318-1,528,358<br>1,529,747-1,535,805<br>1,536,097-1,538,058<br>1,540,863-1,543,188<br>1,545,800-1,549,999<br>1,552,258-1,562,560<br>1,571,169-1,571,725 | DRA |
| 25 | 0.08 | -4.16 | *DCTN5,PLK1,ERN2* | 21,541,012-21,574,827 | |

| | | | | | |
|---|---|---|---|---|---|
| | | | | 21,587,069-21,597,613<br>21,597,704-21,620,834 | DRA |
| 25 | 0.09 | -4.08 | *CCDC154,,PTX4,,TELO2,IFT140,* | 1,126,769-1,133,966<br>1,135,571-1,156,102<br>1,159,757-1,163,098<br>1,167,351-1,168,312<br>1,172,266-1,181,534<br>1,182,105-1,237,269<br>1,194,760-1,207,350 | DRA |
| 25 | 0.21 | -4.96 | *DCTN5,PLK1,ERN2* | 21,541,012-21,574,827<br>21,587,069-21,597,613<br>21,597,704-21,620,834 | AFR |
| 25 | 0.08 | -4.23 | | 1,838,866-1,945,919 | DRA |
| 25 | 0.09 | -4.04 | *,,TRAF7,CASKIN1* | 1,680,110-1,686,192<br>1,693,737-1,693,819<br>1,702,116-1,712,907<br>1,714,641-1,725,252 | DRA |
| 26 | 0.20 | -5.31 | *BTAF1* | 13,503,454-13,576,250 | AFR |
| 27 | 0.09 | -4.00 | *CYP4V2,KLKB1,F11* | 15,306,993-15,323,391<br>15,326,026-15,346,759<br>15,350,931-15,370,080 | DRA |
| 27 | 0.05 | -4.59 | *SLC25A4* | 14,546,020-14,550,037 | DRA |
| 28 | 0.15 | -5.79 | | 1,894,648-1,894,766 | NGI |
| 28 | 0.22 | -4.89 | | 1,894,648-1,894,766 | AFR |
| 28 | 0.05 | -4.51 | | 5,592,553-5,592,659 | DRA |
| 29 | 0.21 | -4.41 | *HEPHL1* | 653,016-744,427 | NGI |
| 29 | 0.06<br>0.25 | -4.46<br>-4.15 | *MEN1,CDC42BPG,EHD1* | 43,661,747-43,668,039<br>43,678,395-43,696,730<br>43,704,113-43,708,218 | AFR, DRA |

**Addendum B:** Putative breed specific SNPs identified as overlaps between the sequence data and the BovineSNP50 array.

| AFR | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| CHR | SNP | NCHROS | POS | A1 | A2 | CHR | SNP | NCHROS | POS | A1 | A2 |
| 1 | ARS-BFGL-NGS-20624 | 47.3124 | 28675718 | A | T | 8 | Hapmap24202-BTA-150022 | 37.8302 | 21689093 | C | A |
| 1 | Hapmap33671-BTA-153460 | 51.7581 | 31407370 | A | G | 8 | Hapmap33684-BTA-27870 | 54.3155 | 34795275 | G | A |
| 1 | Hapmap32123-BTA-156557 | 71.908 | 49618441 | G | A | 8 | Hapmap31847-BTA-162708 | 60.9688 | 41416460 | A | G |
| 1 | UA-IFASA-5504 | 85.3864 | 60889674 | A | C | 8 | ARS-BFGL-NGS-43453 | 61.1654 | 42696770 | A | G |
| 1 | Hapmap25955-BTA-89120 | 141.104 | 1,22E+08 | G | A | 9 | Hapmap27449-BTA-157147 | 35.1649 | 16685512 | A | G |
| 2 | BTA-112386-no-rs | 8.90708 | 3924868 | A | T | 9 | ARS-BFGL-NGS-42246 | 61.298 | 40164227 | G | A |
| 2 | Hapmap40313-BTA-27938 | 21.4572 | 8430514 | C | A | 9 | Hapmap58666-rs29014693 | 76.1181 | 80015418 | G | A |
| 2 | ARS-BFGL-NGS-38727 | 63.7084 | 38670519 | 0 | G | 9 | ARS-BFGL-NGS-34445 | 105.353 | 1,05E+08 | G | A |
| 2 | Hapmap33788-BTA-154116 | 106.45 | 81796035 | G | A | 10 | Hapmap31371-BTA-114833 | 11.7172 | 3389908 | A | G |
| 2 | ARS-BFGL-NGS-33177 | 109.442 | 85585849 | A | G | 10 | Hapmap24333-BTA-125429 | 62.8554 | 36786669 | A | C |
| 2 | Hapmap31464-BTA-150552 | 128.18 | 1,17E+08 | A | G | 10 | Hapmap32991-BTA-125837 | 105.704 | 85827327 | A | G |
| 4 | Hapmap33098-BTA-72619 | 38.2427 | 22783182 | A | G | 10 | Hapmap25613-BTA-158023 | 122.19 | 1,04E+08 | G | A |
| 4 | Hapmap45129-BTA-72713 | 41.8902 | 24932445 | C | G | 11 | Hapmap25893-BTA-157932 | 48.0872 | 33211386 | A | G |
| 4 | BTA-87380-no-rs | 47.9628 | 28149800 | A | G | 11 | Hapmap29422-BTA-120570 | 76.5233 | 69918099 | A | T |
| 4 | Hapmap33364-BTA-142214 | 73.1916 | 57896067 | A | G | 11 | BTA-107321-no-rs | 87.591 | 80107135 | A | G |
| 5 | Hapmap30820-BTA-142968 | 45.0547 | 25379378 | G | A | 11 | BTA-118661-no-rs | 104.038 | 1,01E+08 | 0 | G |
| 5 | Hapmap32902-BTA-74350 | 96.4119 | 80277740 | A | G | 12 | Hapmap24800-BTA-127498 | 55.5329 | 33150154 | A | G |
| 5 | Hapmap49859-BTA-109537 | 100.223 | 82738732 | A | G | 12 | BTB-00499378 | 72.3689 | 55389189 | A | G |
| 5 | Hapmap28443-BTA-164160 | 102.858 | 87187990 | G | C | 12 | Hapmap23461-BTA-147998 | 74.4374 | 59786584 | A | C |
| 5 | BTA-123417-no-rs | 102.883 | 87225487 | A | G | 12 | BTA-120474-no-rs | 83.1661 | 66690581 | A | G |
| 5 | Hapmap40294-BTA-89295 | 112.14 | 1,03E+08 | C | A | 12 | ARS-BFGL-NGS-8701 | 94.9504 | 79374370 | A | C |
| 5 | Hapmap30029-BTA-153521 | 118.125 | 1,12E+08 | A | G | 13 | Hapmap31075-BTA-128022 | 0 | 7538755 | A | G |
| 7 | BTB-00309317 | 68.9419 | 46453487 | G | A | 13 | Hapmap24199-BTA-147242 | 54.8835 | 51021874 | A | C |
| 7 | Hapmap31774-BTA-145190 | 78.5326 | 62431361 | A | G | 14 | Hapmap29971-BTA-128951 | 0 | 13638715 | A | G |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **AFR** | | | | | | | | | | | |
| 14 | UA-IFASA-9658 | 42.2175 | 38248644 | G | A | 6 | Hapmap25900-BTA-159071 | 83.1532 | 68681708 | G | A |
| 14 | Hapmap32017-BTA-123301 | 44.98 | 45517950 | A | G | 6 | BTB-00707501 | 95.2734 | 90075383 | C | A |
| 15 | BTA-91816-no-rs | 0 | 11395733 | A | C | 7 | Hapmap33901-BES9_Contig395_449 | 36.6738 | 21574886 | G | A |
| 15 | Hapmap24380-BTA-149592 | 0 | 13771288 | C | A | 8 | BTA-111112-no-rs | 61.8475 | 47138456 | A | C |
| 16 | BTB-01292634 | 24.3049 | 9237728 | A | C | 10 | Hapmap24169-BTA-96647 | 122.169 | 1,04E+08 | C | A |
| 16 | ARS-BFGL-NGS-81139 | 41.8661 | 19483245 | A | G | 11 | Hapmap30773-BTA-126669 | 73.679 | 68484211 | A | G |
| 16 | Hapmap25641-BTA-38951 | 57.8307 | 44958340 | A | G | 11 | BTA-118786-no-rs | 52.6489 | 35782717 | A | G |
| 16 | Hapmap33660-BTA-131122 | 84.83 | 67027284 | G | A | 11 | ARS-BFGL-NGS-53123 | 81.7478 | 75662200 | A | G |
| 16 | ARS-BFGL-NGS-92942 | 93.1948 | 75690887 | A | G | 11 | Hapmap24534-BTA-127131 | 103.49 | 1E+08 | A | G |
| 17 | Hapmap27883-BTA-154035 | 70.6794 | 37493844 | G | A | 13 | ARS-BFGL-NGS-27078 | 43.24 | 40628125 | A | G |
| 17 | Hapmap31199-BTA-162028 | 80.4265 | 50083942 | A | G | 13 | Hapmap48292-BTA-32583 | 44.683 | 42448271 | G | A |
| 17 | Hapmap30597-BTA-161427 | 81.0887 | 50509617 | A | G | 14 | Hapmap40718-BTA-34856 | 45.812 | 47770268 | G | A |
| 17 | Hapmap30059-BTA-16379 | 89.6734 | 56100282 | A | G | 15 | Hapmap31804-BTA-153738 | 0 | 1757737 | G | A |
| 22 | BTA-102233-no-rs | 107.508 | 57410486 | G | A | 16 | ARS-BFGL-NGS-27776 | 72.9587 | 55576782 | G | A |
| 23 | ARS-BFGL-NGS-21242 | 0 | 5896623 | G | A | 18 | ARS-BFGL-NGS-109919 | 42.5404 | 17159355 | A | G |
| 23 | ARS-BFGL-NGS-42315 | 0 | 50782334 | C | A | 18 | BTA-43023-no-rs | 59.6321 | 34087311 | A | G |
| 24 | Hapmap33715-BTA-137579 | 0 | 4046179 | G | A | 18 | ARS-BFGL-NGS-112003 | 75.759 | 50233023 | A | C |
| 26 | Hapmap24081-BTA-139000 | 0 | 40056717 | C | A | 18 | ARS-BFGL-NGS-86074 | 78.2733 | 53260898 | G | A |
| **DRA** | | | | | | 20 | Hapmap26766-BTA-161730 | 76.9371 | 48046246 | A | G |
| 1 | Hapmap48975-BTA-99363 | 74.8344 | 54032055 | A | G | 20 | Hapmap26766-BTA-161730 | 76.9371 | 48046246 | A | G |
| 1 | BTB-01568926 | 130.294 | 1,1E+08 | G | A | 20 | Hapmap51737-BTA-50812 | 88.99 | 53757088 | G | A |
| 3 | ARS-BFGL-NGS-13586 | 35.0233 | 15380518 | C | A | 21 | BTA-07854-no-rs | 26.1972 | 26447779 | C | A |
| 3 | BTA-28412-no-rs | 121.005 | 1,15E+08 | A | G | 21 | Hapmap44958-BTA-52773 | 156.079 | 61674924 | G | C |
| 4 | Hapmap27403-BTA-142195 | 68.8517 | 55005864 | A | G | 22 | Hapmap23270-BTA-148536 | 0 | 7798009 | C | A |
| 4 | Hapmap30302-BTA-155091 | 80.38 | 69994454 | A | G | 22 | Hapmap26395-BTA-136544 | 37.4101 | 31960252 | G | A |
| 5 | BTA-73200-no-rs | 14.0696 | 5477307 | A | G | 23 | Hapmap33612-BTA-56564 | 0 | 38469772 | A | G |
| 6 | ARS-BFGL-NGS-113703 | 50.5363 | 27744179 | C | A | 24 | ARS-BFGL-NGS-96359 | 0 | 27850482 | G | A |
| 6 | Hapmap28178-BTA-144142 | 81.1368 | 64963475 | G | A | | | | | | |

| NGI | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | BTA-115317-no-rs | 60.2869 | 38098101 | G | A | 8 | Hapmap33502-BTA-153027 | 101.609 | 97132412 | A | G |
| 1 | ARS-BFGL-NGS-59270 | 79.5396 | 56901848 | A | G | 9 | BTB-01209657 | 53.9449 | 30576191 | A | G |
| 1 | Hapmap27375-BTA-124441 | 119.849 | 98033610 | G | A | 10 | ARS-USMARC-307 | 19.6479 | 6703611 | G | A |
| 1 | BTB-00042676 | 120.644 | 98779294 | A | G | 10 | BTA-60298-no-rs | 21.027 | 7271842 | A | G |
| 1 | BTA-102465-no-rs | 130.9 | 1,11E+08 | 0 | C | 10 | Hapmap23230-BTA-63355 | 60.4487 | 33530540 | A | C |
| 2 | ARS-BFGL-NGS-73961 | 147.642 | 1,34E+08 | A | C | 10 | ARS-BFGL-NGS-43120 | 74.7729 | 57065593 | G | A |
| 3 | BTB-01203471 | 92.1416 | 66734946 | A | G | 11 | BTA-97179-no-rs | 63.97 | 55651215 | C | A |
| 4 | BTB-01378315 | 54.8305 | 36066056 | G | A | 11 | ARS-BFGL-NGS-23392 | 63.97 | 55985011 | 0 | G |
| 5 | Hapmap52967-rs29017027 | 32.0392 | 14044364 | A | G | 11 | Hapmap32619-BTA-154095 | 64.4996 | 58730199 | G | A |
| 5 | Hapmap28755-BTA-148390 | 59.1752 | 37473650 | G | C | 11 | Hapmap32846-BTA-152118 | 65.3401 | 60061414 | G | A |
| 5 | Hapmap33077-BTA-163333 | 64.7653 | 40445940 | G | A | 11 | Hapmap22740-BTA-126683 | 77.5425 | 70487203 | A | C |
| 5 | Hapmap26433-BTA-151604 | 102.683 | 86918783 | G | A | 11 | Hapmap29741-BTA-158857 | 89.3432 | 83072589 | G | A |
| 5 | ARS-BFGL-NGS-10549 | 113.165 | 1,06E+08 | A | G | 11 | ARS-BFGL-NGS-108538 | 107.471 | 1,07E+08 | A | G |
| 6 | BTB-01790614 | 10.1865 | 3726761 | A | G | 12 | BTA-21643-no-rs | 56.0903 | 36566658 | 0 | A |
| 6 | ARS-BFGL-NGS-117236 | 83.8483 | 70065213 | A | C | 12 | BTA-122625-no-rs | 73.598 | 58027379 | A | C |
| 6 | ARS-BFGL-NGS-10480 | 108.267 | 1,06E+08 | A | C | 12 | Hapmap48930-BTA-87766 | 96.528 | 81619169 | A | G |
| 7 | BTA-72600-no-rs | 71.6298 | 49811968 | A | G | 12 | ARS-BFGL-NGS-118242 | 97.4541 | 82631955 | A | G |
| 7 | Hapmap26444-BTA-153939 | 77.8253 | 56655089 | A | G | 13 | Hapmap25801-BTA-128091 | 0 | 15642728 | G | A |
| 7 | BTB-01106344 | 78.96 | 66107951 | A | C | 13 | ARS-BFGL-NGS-79261 | 34.0948 | 34851781 | A | C |
| 7 | Hapmap31776-BTA-145290 | 95.5631 | 82131200 | A | G | 13 | Hapmap30063-BTA-32922 | 56.3715 | 55244135 | A | G |
| 7 | Hapmap27427-BTA-151849 | 98.6582 | 85337935 | A | G | 13 | ARS-BFGL-NGS-107717 | 64.6998 | 71075689 | A | G |
| 7 | ARS-BFGL-NGS-4612 | 108.562 | 98173299 | A | G | 13 | BTA-120560-no-rs | 73.26 | 76994266 | C | A |
| 7 | Hapmap25284-BTA-145409 | 109.861 | 99989176 | A | G | 14 | UA-IFASA-5765 | 0 | 4520969 | C | A |
| 8 | ARS-BFGL-NGS-114722 | 48.1898 | 28464160 | A | G | 14 | ARS-BFGL-NGS-37279 | 12.8659 | 21104637 | G | A |
| 8 | Hapmap25906-BTA-159707 | 55.41 | 37471009 | A | G | 14 | Hapmap26746-BTA-157258 | 45.5579 | 47368746 | A | G |
| 8 | Hapmap31155-BTA-153044 | 60.9902 | 41555974 | A | G | 14 | Hapmap26378-BTA-129537 | 83.0661 | 83582981 | A | G |
| 8 | ARS-BFGL-NGS-43242 | 61.1422 | 42545530 | A | G | 15 | Hapmap30784-BTA-129668 | 21.9219 | 28361551 | A | G |
| 8 | Hapmap30833-BTA-145860 | 70.6658 | 59778455 | A | G | 15 | BTA-18105-no-rs | 66.915 | 64168198 | G | A |

| | | | | | |
|---|---|---|---|---|---|
| 16 | ARS-BFGL-NGS-114924 | 57.919 | 46093561 | G | A |
| 17 | BTA-42150-no-rs | 13.066 | 5191623 | G | A |
| 17 | BTA-40649-no-rs | 51.7905 | 23367681 | A | C |
| 17 | Hapmap27157-BTA-131368 | 68.6482 | 35413944 | A | G |
| 17 | ARS-BFGL-NGS-101808 | 73.5092 | 41089654 | A | G |
| 17 | BTA-102489-no-rs | 111.326 | 69535278 | C | A |
| 18 | ARS-BFGL-NGS-52078 | 60.5624 | 36040190 | A | T |
| 18 | ARS-BFGL-NGS-42678 | 79.6561 | 54112530 | A | T |
| 19 | Hapmap26455-BTA-156640 | 0 | 758595 | A | G |
| 19 | ARS-BFGL-NGS-116964 | 63.2158 | 41759425 | 0 | G |
| 19 | Hapmap39750-BTA-45775 | 75.273 | 52334083 | C | A |
| 20 | Hapmap23357-BTA-134915 | 37.6511 | 13478245 | A | C |
| 21 | BTB-00810924 | 24.9047 | 25007400 | A | G |
| 21 | Hapmap30532-BTA-136054 | 131.508 | 56984645 | A | G |
| 22 | Hapmap23073-BTA-136759 | 102.035 | 56049440 | A | C |
| 22 | BTA-86052-no-rs | 107.032 | 57292276 | A | G |
| 23 | Hapmap51466-BTA-55807 | 0 | 3098756 | A | G |
| 23 | ARS-BFGL-NGS-62736 | 29.9423 | 16123681 | A | G |
| 23 | BTB-00872950 | 0 | 43984482 | A | G |
| 24 | Hapmap39983-BTA-20390 | 0 | 26760851 | A | C |
| 24 | Hapmap22828-BTA-90188 | 0 | 36808804 | C | A |
| 25 | Hapmap36124-SCAFFOLD100400_31720 | 0 | 10686085 | G | A |
| 25 | ARS-BFGL-NGS-38449 | 0 | 32575148 | C | A |
| 26 | BTA-61940-no-rs | 0 | 918433 | A | G |
| 26 | Hapmap26496-BTA-27708 | 0 | 6594101 | A | G |
| 26 | Hapmap24386-BTA-163519 | 0 | 48971298 | C | A |
| 26 | ARS-BFGL-NGS-10954 | 0 | 50082843 | A | G |

**Addendum C:** Computation and the probability score of Nguni breed (NGU) allocation to the breed of origin when NGI was ued as a reference population.

| Breed allocation | | | | | | Probability scores | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Assigned samples** | **Rank 1** | **Score %** | **NGI -log(L)** | **No. of loci** | **Missing loci** | **Assigned sample** | **NGI probability** | **No. of loci** | **Missing loci** |
| /NGU | NGI | 100.000 | 11.228 | 30 | - | /NGU | 0.229 | 30 | - |
| /NGU | NGI | 100.000 | 8.740 | 30 | - | /NGU | 0.813 | 30 | - |
| /NGU | NGI | 100.000 | 10.812 | 30 | - | /NGU | 0.305 | 30 | - |
| /NGU | NGI | 100.000 | 10.218 | 30 | - | /NGU | 0.447 | 30 | - |
| /NGU | NGI | 100.000 | 9.296 | 30 | - | /NGU | 0.681 | 30 | - |
| /NGU | NGI | 100.000 | 8.078 | 30 | - | /NGU | 0.927 | 30 | - |
| /NGU | NGI | 100.000 | 10.398 | 30 | - | /NGU | 0.392 | 30 | - |
| /NGU | NGI | 100.000 | 8.946 | 30 | - | /NGU | 0.762 | 30 | - |
| /NGU | NGI | 100.000 | 10.271 | 30 | - | /NGU | 0.431 | 30 | - |
| /NGU | NGI | 100.000 | 9.554 | 30 | - | /NGU | 0.611 | 30 | - |
| /NGU | NGI | 100.000 | 8.839 | 30 | - | /NGU | 0.789 | 30 | - |
| /NGU | NGI | 100.000 | 10.331 | 30 | - | /NGU | 0.409 | 30 | - |
| /NGU | NGI | 100.000 | 8.901 | 30 | - | /NGU | 0.771 | 30 | - |
| /NGU | NGI | 100.000 | 10.045 | 30 | - | /NGU | 0.492 | 30 | - |
| /NGU | NGI | 100.000 | 8.595 | 30 | - | /NGU | 0.838 | 30 | - |
| /NGU | NGI | 100.000 | 10.051 | 30 | - | /NGU | 0.489 | 30 | - |
| /NGU | NGI | 100.000 | 8.697 | 30 | - | /NGU | 0.823 | 30 | - |
| /NGU | NGI | 100.000 | 10.294 | 30 | - | /NGU | 0.422 | 30 | - |
| /NGU | NGI | 100.000 | 9.925 | 30 | - | /NGU | 0.516 | 30 | - |
| /NGU | NGI | 100.000 | 9.454 | 30 | - | /NGU | 0.636 | 30 | - |
| /NGU | NGI | 100.000 | 11.741 | 30 | - | /NGU | 0.160 | 30 | - |
| /NGU | NGI | 100.000 | 13.582 | 30 | - | /NGU | 0.029 | 30 | - |
| /NGU | NGI | 100.000 | 8.345 | 29 | snp5 | /NGU | 0.805 | 29 | snp5 |
| /NGU | NGI | 100.000 | 8.422 | 30 | - | /NGU | 0.879 | 30 | - |
| /NGU | NGI | 100.000 | 9.986 | 30 | - | /NGU | 0.502 | 30 | - |
| /NGU | NGI | 100.000 | 9.441 | 30 | - | /NGU | 0.639 | 30 | - |
| /NGU | NGI | 100.000 | 10.915 | 30 | - | /NGU | 0.279 | 30 | - |
| /NGU | NGI | 100.000 | 12.990 | 30 | - | /NGU | 0.055 | 30 | - |
| /NGU | NGI | 100.000 | 10.216 | 30 | - | /NGU | 0.447 | 30 | - |
| /NGU | NGI | 100.000 | 8.223 | 30 | - | /NGU | 0.906 | 30 | - |
| /NGU | NGI | 100.000 | 10.804 | 30 | - | /NGU | 0.306 | 30 | - |
| /NGU | NGI | 100.000 | 8.744 | 30 | - | /NGU | 0.813 | 30 | - |
| /NGU | NGI | 100.000 | 7.269 | 30 | - | /NGU | 0.982 | 30 | - |
| /NGU | NGI | 100.000 | 7.899 | 30 | - | /NGU | 0.946 | 30 | - |
| /NGU | NGI | 100.000 | 8.646 | 30 | - | /NGU | 0.832 | 30 | - |
| /NGU | NGI | 100.000 | 14.422 | 30 | - | /NGU | 0.012 | 30 | - |
| /NGU | NGI | 100.000 | 10.979 | 30 | - | /NGU | 0.265 | 30 | - |
| /NGU | NGI | 100.000 | 11.342 | 30 | - | /NGU | 0.212 | 30 | - |
| /NGU | NGI | 100.000 | 10.533 | 30 | - | /NGU | 0.359 | 30 | - |
| /NGU | NGI | 100.000 | 10.549 | 30 | - | /NGU | 0.356 | 30 | - |
| /NGU | NGI | 100.000 | 11.325 | 30 | - | /NGU | 0.214 | 30 | - |
| /NGU | NGI | 100.000 | 9.032 | 30 | - | /NGU | 0.742 | 30 | - |
| /NGU | NGI | 100.000 | 9.353 | 30 | - | /NGU | 0.664 | 30 | - |
| /NGU | NGI | 100.000 | 8.525 | 30 | - | /NGU | 0.859 | 30 | - |
| /NGU | NGI | 100.000 | 7.017 | 30 | - | /NGU | 0.991 | 30 | - |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| /NGU | NGI | 100.000 | 9.483 | 30 | - | /NGU | 0.626 | 30 | - |
| /NGU | NGI | 100.000 | 8.596 | 30 | - | /NGU | 0.837 | 30 | - |
| /NGU | NGI | 100.000 | 11.116 | 30 | - | /NGU | 0.248 | 30 | - |
| /NGU | NGI | 100.000 | 9.678 | 29 | snp3 | /NGU | 0.472 | 29 | snp3 |
| /NGU | NGI | 100.000 | 12.008 | 30 | - | /NGU | 0.121 | 30 | - |
| The allocation of the test breed to the reference population was 100% with only two SNPs that didn't allocate individuals | | | | | | The probability score for all the individials that were allocatedto the reference popolation | | | |