Genotype imputation as a genomic strategy for the SA Drakensberger beef breed

by

SIMON FREDERICK LASHMAR

Submitted in fulfillment of the requirements for the degree

PHILOSOPHIAE DOCTOR (ANIMAL SCIENCE)

In the Faculty of Natural & Agricultural Sciences
University of Pretoria
Pretoria

February 2020

SUPERVISORY COMMITTEE

Dr Carina Visser

Department of Animal and Wildlife Science
University of Pretoria
Private Bag X20
Hatfield
0028
South Africa

Dr Farai C. Muchadeyi

Biotechnology Platform
Agricultural Research Council
Private Bag X5
Onderstepoort
0110
South Africa

Dr Donagh P. Berry

Animal and Grassland Research and Innovation Centre
Teagasc
Moorepark
Fermoy
Ireland

DECLARATION

I, Simon Frederick Lashmar, declare that the thesis/dissertation, which I hereby submit for the degree PhD Animal Science at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature: …………………………………

Date: ....

ACKNOWLEDGEMENTS

I would firstly like to acknowledge and thank my PhD supervisory committee for their guidance during the planning, execution and reviewing of this thesis. Their expertise and inputs were invaluable in enabling me to complete this degree. Dr Carina Visser (Department of Animal & Wildlife Science, University of Pretoria), a special thanks to you for believing in me throughout the entire process (you have been my mentor since my MSc Animal Breeding & Genetics) - your guidance and motivation drove me to complete this thesis. I know that we will work together for many more years in the future. Dr Farai Muchadeyi, thank you for allowing me the opportunity to be a part of the Agricultural Research Council's Professional Development Program and to join your group - your confidence in my abilities made me believe more in myself as well. Dr Donagh Berry, thank you that you allowed me the opportunity to visit TEAGASC in Ireland (2016) and for mentoring me since then - your expertise in the field of cattle genomics was invaluable in refining this thesis.

I would like to thank all other researchers involved in the work presented in this thesis. Prof van Marle-Köster, thank you for providing additional data that added value to Chapter 3 (when there were still few genotyped animals available) and for sharing your incredible knowledge about our field. Dr Rian Pierneef, your assistance in formulating the bioinformatics scripts to execute the data analyses for Chapter 4 was crucial - I really cannot thank you enough. Dr Bobbie van der Westhuizen, thank you for your assistance with the analyses of Chapter 5 and for allowing me to utilize SA Stud Book's software and facilities - your assistance and advice was much appreciated.

I would like to acknowledge the funding bodies involved: the Beef Genomics Program (BGP) and the Red Meat Research and Development (RMRD). Without the funding provided by both the BGP and RMRD, this project would not have been possible. The funding provided was key in the accumulation of a sufficient number of genotyped animals to allow imputation and genomic selection analyses. I would further like to acknowledge the Agricultural Research Council (ARC) for providing me with a bursary that sustained me financially while this project was carried out to completion. The financial assistance of the National Research Foundation (NRF) towards this research is hereby also acknowledged. Opinions expressed and conclusions arrived at, are mine, the student and principle author of this thesis, and are not necessarily to be attributed to the NRF.

I would like to especially mention my family in Cape Town and Pretoria. Thank you to my family in Cape Town: my parents (Monica and Steven Lashmar), both sets of grandparents, my siblings (Jason and Emma Lashmar) and all other relatives - your support has been amazing. Special thanks to my parents for their unconditional love as well as both emotional and financial support throughout my studies (the entire 10 years - from undergrad to the completion of this PhD). Thank you to my family in Pretoria: the Vermaak family. I would like to especially mention Chris - you played a big role in my life during the completion of this degree. Your support has carried me through and I cannot thank you enough for the sacrifices that you've made as well. Your parents (Elize and Christo Vermaak) and siblings have been my "family" and support system while I was so far away from home.

To my friends and fellow students at both the University of Pretoria and the ARC's Biotechnology Platform (BTP): I cannot really put into words what your support has meant to me. I am eternally grateful to be able to call you my friends and colleagues. I would like to especially mention the following people: Maruschke Andrews, Rulien Grobler, Amelia du Preez, Jani de Vos, Yandisiwe Sanarana, Stephanie Cornellissen, Minique de Castro (Ouma), Annie Gous, Erika Viljoen, Nompilo Hlongwane, Khanyi Hadebe and Mary Ranketse (my fellow student representative at the BTP).

Last, but not least, thank you God. Thank you for blessing me with the physical and mental wellbeing to be able to finish my degree. There were moments where I felt like giving up, but You never allowed that to happen - You were always there, every step of the way, looking after me.

"*Accentuate the positive. Eliminate the negative*."

EXECUTIVE SUMMARY

Indigenous Sanga cattle (*Bos taurus africanus*) breeds such as the South African (SA) Drakensberger are economically important to the commercial sector of the local beef-producing industry. The adaptive qualities of this breed make it a competitive performer across the variety of beef-producing environments of SA and hence there has been interest in applying genomic technologies to its breed improvement strategies. There has been a recent surge in genotyping of local breeds, including the SA Drakensberger, which was facilitated by the establishment of a national Beef Genomics Program (BGP). This has generated sufficient genomic data to explore the utility of various genomic methodologies in local breeds, including genomic selection (GS). The implementation of GS will require routine genotyping and hence frequent updating of genomic breeding value (GEBV) prediction equations, in terms of including more animals, to ensure sustainability. Maintaining a genome-enhanced breeding program can become financially unfeasible, especially if the uptake by farmers is low. As a strategy to reduce the cost of high-density genotyping, efficient imputation from a more cost-effective low-density single nucleotide polymorphism (SNP) panel to higher density can be integrated into routine GS pipelines. The aim of this study was to validate the utility of imputation as a genomic strategy for a local breed such as the SA Drakensberger. A commercial genotyping panel consisting of 139 480 SNPs was used to perform analyses that could be organized into three phases: pre-imputation, imputation and post-imputation. These three phases are presented in three experimental chapters (Chapter 3, 4 and 5) in this thesis of which Chapter 3 is published. The animals sampled were pre-selected according to the main research aim of the BGP, which was the eventual ability to impliment GS, and included high-impact animals that had sufficiently accurate estimates of conventional breeding values (EBVs). A subset of these animals were used to carry out the diversity study (Chapter 3), whereafter offspring and/or ancestors of these animals were sampled independently of BGP funding to ensure improved relatedness for imputation (Chapter 4) and GEBV estimation (Chapter 5). An introductory chapter (Chapter 1) as well as a published review chapter (Chapter 2) precedes the experimental chapters. The thesis is concluded with a critical review chapter (Chapter 6), which includes a general discussion, recommendations and a conclusion. The referencing style is consistent throughout except for published sections, in which case the style followed was as required by the relevant journals.

ABSTRACT

Indigenous breeds such as the South African (SA) Drakensberger are economically important genetic resources in local beef production because of their adaptive traits and ability to perform competitively at a commercial level. Genomic selection (GS) is a promising technology to accelerate genetic progress in traits relevant to commercial beef production. A major obstacle in applying this methodology has been the cost of genotyping at high densities of single nucleotide polymorphisms (SNPs). Cost reduction can be achieved by exploiting genotype imputation in GS workflows by means of genotyping at lower densities and imputing upwards. The overarching aim of this study was to conduct an investigation into the practicality of applying imputation in such a workflow utilizing genotypic data for 1 135 SA Drakensberger animals genotyped for 139 480 SNPs. As a pre-imputation step, the objective was firstly to elucidate inter- and intra-chromosomal patterns in genomic characteristics that may contribute to variability in achievable imputation accuracy across the genome. Inter-chromosomal differences in the proportion of low minor allele frequency (MAF) SNPs estimated varied from 6.6% for *Bos Taurus* autosome (BTA) 23 to 16.0% for BTA14. Pairwise linkage disequilibrium (LD), between adjacent SNPs, ranged from r^2=0.11 (BTA28) to 0.17 (BTA14). The largest run of homozygosity (ROH), located on BTA13, was 225.82 kilobases (kb) in length and was identified in 23% of the animals sampled. The ROH-based inbreeding coefficients (F_{ROH}) estimated (e.g. $F_{\mathrm{ROH>1Mb}}$=0.07, where $F_{\mathrm{ROH>1Mb}}$ denotes F_{ROH} calculated for all ROH longer than 1 megabase pair), indicated sufficient within-breed relatedness to achieve accurate imputation. During the imputation step, imputation accuracy from several custom-derived lower density panels varying in SNP density and the SNP selection strategy were compared. Imputation accuracy increased as SNP density increased; a genotyping panel consisting of 10 000 SNPs, selected based on a combination of their MAF and LD with neighbouring SNPs, could be used to achieve <3% imputation error on average. At this density of SNPs, a mean correlation coefficient (±standard deviation) between true- and imputed SNPs of 0.972±0.024 was achieved in a set of validation animals (n=235). Low MAF SNPs were imputed with lesser accuracy; a difference of 0.071 units was observed between the mean accuracy of imputed SNP categorized into low- (0.01<MAF≤0.1) versus high MAF (0.4<MAF<0.5) classes. Post-imputation, the utility of imputed genotypes in genomic breeding value (GEBV) estimation was evaluated by comparing prediction accuracies achieved from the use of true versus imputed SNPs in generating the H-inverse matrix applied in single-step GS. Breeding values were estimated for two growth traits, considering direct and maternal components. Prediction accuracies were improved by using genomic information in addition to traditional pedigree information; the largest improvement (0.026 units increase in accuracy) was observed for maternal birth weight. Marginal differences were observed between GEBV accuracies produced from true ($\mathrm{GEBV_{TRUE}}$) versus imputed genotypes ($\mathrm{GEBV_{IMPUTED}}$); for example the mean±standard deviation in $\mathrm{GEBV_{TRUE}}$=0.774±0.056 versus $\mathrm{GEBV_{IMPUTED}}$=0.773±0.055 accuracy was observed for direct birth weight, suggesting that imputation errors had an almost negligible influence. Results presented in this

thesis demonstrated the usefuleness of imputation as a viable genomic strategy towards low-cost implementation of genomically enhanced prediction of EBVs for a breed such as the SA Drakensberger.

THESIS OUTPUTS

Publications:

Peer-review journals

- **Lashmar, S.F.**, Visser, C., van Marle-Köster, E. & Muchadeyi, F.C., 2018. Genomic diversity and autozygosity within the SA Drakensberger beef breed. Livestock Science 212, 111-119.

- **Lashmar, S.F.**, Muchadeyi, F.C. & Visser, C., 2019. Genotype imputation as a cost saving genomic strategy for South African Sanga cattle: A review. South African Journal of Animal Science 49, 262-280.

Peer-review conference paper

- **Lashmar, S.F.**, Visser, C. & Muchadeyi, F.C., 2018. Factors influencing imputation accuracy for the South African Drakensberger beef cattle breed. In: Proceedings of the 11th World Congress on Genetics Applied to Livestock Production. Auckland, New Zealand, 11-16 February 2018. Paper 11.472.

  * This work was also presented as a 15-minute theatre contribution at the 11th WCGALP conference on the 16th of February during the "Genetic Gain: Strategies for local breeds" session

Popular science

- **Lashmar, S.F.,** 2017. Using imputation to save on genotyping costs for indigenous cattle. SA Drakensberger Cattle Breeders' Society 2017 Newsletter, pp28-29. Available online: https://www.drakensbergers.co.za.

Congresses:

National

- **Lashmar, S.F.**, Muchadeyi, F.C., van Marle-Köster, E. & Visser, C., 2016. Linkage disequilibrium and effective population size in South African Drakensberger cattle. In: Proceedings of the 49th South African Society for Animal Science congress. Stellenbosch, Western Cape, 3-6 July 2016.

- **Lashmar, S.F.**, Visser, C. & Muchadeyi, F.C., 2017. Genotype imputation as a genomic strategy for the South African Drakensberger beef breed. In: Proceedings of the 50th South African Society for Animal Science congress. Port Elizabeth, Eastern Cape, 18-21 September 2017.

- **Lashmar, S.F.,** 2017. Genotype imputation as a genomic strategy for the SA Drakensberger beef breed. In: Proceedings of the 4th Professional Development Programme conference of the Agricultural Research Council.

  * This work was presented as a 15-minute oral presentation

  ** Mr Lashmar was awarded 2nd prize for "Best PhD Scientific Paper"

International

- **Lashmar, S.F.**, Visser, C. & Muchadeyi, F.C., 2017. Genomic diversity and autozygosity within the SA Drakensberger beef cattle breed. In: Proceedings of the 35th International Conference on Animal Genetics. Salt Lake City, Utah, USA, 23-27 July 2017.

- **Lashmar, S.F**., Berry, D.P., Pierneef, R., Muchadeyi, F.C. & Visser, C., 2019. Single nucleotide polymorphism selection methods to optimize imputation accuracy for South African Drakensberger beef cattle. In: Proceedings of the 37th International Conference on Animal Genetics. Lleida, Spain, 7-12 July 2019.

Reports to industry:

- Beef Genomics Project (BGP) workshop. CSIR International Convention Centre, 17 October 2017.
- SA Drakensberger Breeders's Society annual general meeting. AfriDome, Parys, Free State, 20 June 2018.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

LIST OF FIGURES

CHAPTER 3

CHAPTER 4

CHAPTER 5

LIST OF ADDENDA

LIST OF ABBREVIATION

2.5K	2 500 single nucleotide polymorphism panel
5K	5 000 single nucleotide polymorphism panel
10K	10 000 single nucleotide polymorphism panel
20K	20 000 single nucleotide polymorphism panel
50K	50 000 single nucleotide polymorphism panel
ACR	Allele concordance rate
$\textbf{ACR}_{\textbf{ANIM}}$	Allele concordance rate per animal
$\textbf{ACR}_{\textbf{SNP}}$	Allele concordance rate per single nucleotide polymorphism
AER	Allele error rate
AI	Artificial Insemination
ARC	Agricultural Research Council
BGP	Beef Genomics Program
BLUP	Best linear unbiased prediction
BTA	Bos Taurus Autosome
cM	CentiMorgan
CNV	Copy number variation
COR	Correlation
$\textbf{COR}_{\textbf{ANIM}}$	Correlation per animal
$\textbf{COR}_{\textbf{SNP}}$	Correlation per single nucleotide polymorphism
cROH	Consensus runs of homozygosity
DAFF	Department of Agriculture, Forestry and Fisheries
DNA	Deoxyribonucleic acid
DGV	Direct genomic value
DISTMAF	Equidistant and minor allele frequency selection method
EBV	Estimated breeding value
EV	Eigenvector
$\boldsymbol{F}_{\textbf{PED}}$	Pedigree-based inbreeding coefficient
$\boldsymbol{F}_{\textbf{ROH}}$	Runs of homozygosity-based inbreeding coefficient
$\boldsymbol{F}_{\textbf{SNP}}$	Single nucleotide polymorphism-based inbreeding coefficient
GBS	Genotype-by-sequencing
GCR	Genotype concordance rate
$\textbf{GCR}_{\textbf{ANIM}}$	Genotype concordance rate per animal
$\textbf{GCR}_{\textbf{SNP}}$	Genotype concordance rate per single nucleotide polymorphism
GCTA	Genome-wide Complex Trait Analysis
GEBV	Genomic estimated breeding value
$\textbf{GEBV}_{\textbf{TRUE}}$	GEBV using true genotypes

GEBV$_{\text{IMPUTED}}$	GEBV using imputed genotypes
GER	Genotype error rate
GGP	GeneSeek® Genomic Profiler
GS	Genomic selection
GWAS	Genome-wide association study
HD	High density
HMM	Hidden Marcov model
ICBF	Irish Cattle Breeding Federation
IQS	Imputation quality score
Kb	Kilobase pair(s)
LD	Linkage disequilibrium
MAF	Minor allele frequency
MAFLD	Minor allele frequency and linkage disequilibrium selection method
Mb	Mega base pair(s)
MID	Mid-point selection method
NCBI	National Centre for Biotechnology Information
N$_e$	Effective population size
NGS	Next-generation sequencing
NRF	National Research Foundation
PAM	Partitioning-around-medoids
PCA	Principal component analysis
QC	Quality control
QTL	Quantitative trait loci
RAN	Random selection method
RMRD-SA	Red Meat Research and Development South Africa
RMSE	Root mean square error
ROH	Run(s) of homozygosity
SA	South Africa (n)
SNP	Single nucleotide polymorphism
ssGBLUP	Single-step genomic best linear unbiased prediction
TIA	Technology Innovation Agency
UN	United Nations
UP	University of Pretoria

CHAPTER ONE

INTRODUCTION

1.1 Motivation

During the period from 2008 to 2018, the South African (SA) human population grew from 49.6 million to 57.4 million people, with forecasts indicating an exponential growth to 72.8 million people by 2050 (United Nations, 2017). Securing adequate food resources to supply this increase in nutritional demand in an efficient and sustainable manner will be a main focus in animal agriculture. The SA population consumes approximately 38.7kg, 11.9kg, 4.1kg and 3.6kg of chicken, beef and veal, pork, and mutton and lamb per capita, respectively (OECD, 2019). Beef is the second largest animal protein commodity in SA and the industry has a gross value of R33 billion (USDA, 2018). Although it is comparatively small in the global perspective, constituting a mere 1.4% (1.010 million tonnes) of global beef production, beef products in SA contribute nearly a quarter of the African continent's beef production (USDA, 2018). The SA beef industry is dualistic in nature, consisting of both developed and developing sectors with relation to production capacity. More than 75% of beef produced originates from the developed sector and the percentage consumer purchases from this sector is almost double that from the developing sector (33% versus 17%; USDA, 2018). With a key focus on the improvement of indigenous cattle for large-scale beef production, the emphasis in this thesis will be on these cattle within the context of the commercial or developed sector.

Approximately 82.3% of SA land can be classified as farming land, which comprises an estimated 13.7% of land that is potentially arable and 68.6% of land that is suitable for grazing (DAFF, 2018). Ample land is therefore available for utilization in extensive production systems characteristic of farming practices used in rearing meat-producing ruminants such as beef cattle (van Marle-Köster & Visser, 2018). Ruminants raised under these systems are more vulnerable to the impact of external stressors unique to their production environments, including adverse weather conditions, suboptimal grazing quality and disease-causing parasites (Scholtz *et al*., 2014; Nyamushamba *et al*., 2017). In the 2015/2016 calendar year, for example, SA was confronted with extreme drought that forced farmers to slaughter cattle because of grazing shortage and spikes in feed cost (USDA, 2018). This was followed by the commencement of a herd-rebuilding phase in 2017 whereby fewer cattle were being slaughtered and the price of beef became less affordable in comparison with other animal protein commodities (USDA, 2018). Even though the beef industry is recovering, this period of drought and the consequences thereof, provided evidence of the volatile nature of SA farming environments.

The productivity and efficiency of commercial beef-producing operations will therefore increasingly rely on cattle breeds that are adapted to current farming environments and that will be able to withstand uncertain future environments. The Sanga subspecies of cattle (*Bos taurus africanus*), and the composite breeds carrying its bloodlines, are native to SA. The recognized Sanga breeds that inhabit SA are the Afrikaner, Bonsmara, Drakensberger, Nguni and Tuli breeds. These breeds display a unique potential to reproduce and produce well under harsh local conditions. Evidence of this potential was through competitive growth performance in both low-input conditions (e.g. Collins-Lusweti, 2000) and commercial feedlot conditions (e.g. Strydom, 2008) in SA, and has largely been attributed to their unique adaptive qualities. These breeds have an innate resistance to tick-borne diseases present in local production environments and the presence of ticks have a minor impact on their growth performance and even carcass characteristics (Schoeman, 1989; Rechav & Kostrzewski, 1991; Muchenje *et al*., 2008). Their adaptive abilities also include their tolerance to heat, drought, endemic diseases and various other environmental stressors (Gororo *et al*., 2018). These attributes make them fit for inclusion in pure and crossbred breeding programmes, which will assist in dealing with the projected changes in livestock production dynamics (Gororo *et al*., 2018), and important genetic resources as climate change becomes a major challenge to livestock productivity.

The SA Drakensberger is a medium-framed breed with a sleek, black coat. This breed was the first to receive estimated breeding values (EBVs) using best linear unbiased prediction (BLUP; Henderson, 1984) methodology in the mid-1980s (Niemand, 2013) due, in part, to an extensive history of compulsory performance recording (SA Drakensberger Breeders' Society, 2017). At present, 100% of breeders participate in SA Stud Book's *Logix Beef* performance-recording scheme (SA Stud Book, 2016). For this breed, and most other SA breeds, routine measurements have, however, been more diligent for growth-related traits; fertility, health, meat quality and other expensive-to-measure traits have largely being neglected until recently (van Marle-Köster & Visser, 2018). The availability of selection tools such as EBVs has resulted in positive genetic gain being made for such traits related to growth performance. In relation to exotic breeds, the SA Drakensberger has shown competitive growth performance, intermediate between *Bos indicus* and *Bos Taurus* breeds (e.g. average daily gain: Brahman=1345g, Drakensberger=1550g, SA Angus=1805g; Bosman, 2002). Given the sufficiency in the phenotypic records available for growth traits, it would therefore be more appropriate to initially focus genomic selection strategies on these traits. To justify the substitution of traditional EBVs with genome-enhanced breeding values (GEBVs), improvements in accuracy will have to be demonstrated. Facilitating improved accuracies will, however, rely on the availability of sufficient single nucleotide polymorphism (SNP) genotype data to compliment existing phenotypic records.

Establishing a training population that is at least a 1000 genotyped animals, which has been accepted as a general rule-of-thumb to achieve acceptable GEBV accuracies (Meuwissen *et al*., 2001), is still an expensive endeavour in most developing countries due to the high cost of SNP genotyping (Mrode *et al*., 2018). It should further be noted that the ballpark figure of a 1000 animals has been suggested for simulated or true dairy populations that are characterized by small effective poulations sizes and strong linkage disequilibrium (LD); this figure might therefore be higher for local breeds that are genomically more diverse. Genotyping costs therefore present an obstacle in implementing routine genomic evaluations for many breeds and presently GS is only undertaken in a few breeds in developing countries, for example the Nellore (*Bos indicus*) of Brazil (Carvalheiro, 2014). A national

Beef Genomics Program (BGP) was recently established in SA with both research and commercial aims (www.livestockgenomics.co.za) relevant to indigenous breeds and production systems of the local industry. Since the initiation of the BGP, SNP genotypes have been generated for approximately 7 000 animals across 16 participating breeds (van Marle-Köster & Visser, 2018). Over the three-year funding period, a cohort of approximately 800 SA Drakensberger animals were genotyped using a genotyping panel consisting of 139 480 SNPs. With additional funding from organizations such as Red Meat Research and Development SA (RMRD-SA; www.rmrdsa.co.za), approximately 1 200 SA Drakensberger animals have been collectively genotyped to date. For the first time in SA, the number of high-density genotyped animals has allowed the exploitation of GS methodology for indigenous cattle breeds such as the SA Drakensberger.

Genetic progress resulting from the implementation of GS internationally has been contingent on routine predictions made annually i.e. prediction equations have been regularly updated (Mrode *et al*., 2018). Post-BGP, sustaining genomic progress from GS will depend on the participation and uptake by breeders, which will be an unrealistic expectation if the benefit of this methodology, in relation to its cost, has not been properly demonstrated (Berry *et al*., 2016). The procurement of genotypes at a reduced cost can be achieved through genotyping on lower density SNP panels and imputing to higher density (Berry *et al*., 2013). These lower density SNP panels can then be employed to generate GEBVs from a combination of actual and inferred genotypes (e.g. Raoul *et al*., 2017). Imputation is a statistical methodology that relies on the genomic segments shared within a breed, or a group of genetically similar breeds, to predict genotypic information for SNPs that were not physically genotyped (Marchini *et al*., 2007). Higher density genotypes can therefore be inferred for selection candidates that were only genotyped for lower densities by using a haplotype library constructed from breed-representative animals with dense marker genotypes available (Marchini *et al*., 2007).

The utility of imputed genotypes in genomic applications is compromised if the imputation methodology is not carried out correctly. Incorrectly imputed genotypes may lead to false positive associations in genome-wide association studies (GWAS) and the risk is especially concerning for rare SNPs (Marchini & Howie, 2010). Large numbers of incorrectly imputed genotypes may impact the accuracy of predicted GEBVs but previous research has documented this influence to be mostly negligible even for breeds with admixed genomes (e.g. Aliloo *et al*., 2018). Imputation accuracy is directly proportional to the breed-specific genomic characteristics of a SNP itself or its relationship with neighbouring SNPs. This may present a challenge for admixed breeds because the entire set of genotyped SNPs might not be informative in these breeds. Kim *et al*. (2017) previously indicated greater genomic diversity for indigenous breeds in comparison with internationally commercial breeds; this tendency has also been observed for SA breeds (e.g. Makina *et al*., 2016; Zwane *et al*., 2016; Pierce *et al*., 2018). To fully reap the benefits of imputation, a specific set of SNPs will need to be identified in terms of density and characteristics, which will be useful in applying GS without compromising either genotype imputation accuracy or GEBV prediction accuracy. Validation of a GS pipeline that includes imputation in its workflow will enable accelerated breed improvement of the SA Drakensberger, which will benefit the beef cattle gene pool of the SA beef industry in the future.

1.2 Aim of the study

The aim of this thesis was to validate genotype imputation as a method of inferring high density, *in silico*, genotypic data to enable genome-based breed improvement of an economically important indigenous breed, the SA Drakensberger.

The aim of this study was accomplished by executing the following objectives:

- To quantify inter-chromosomal genomic diversity and autozygosity that may influence the accuracy of genotype imputation as a genomic strategy for the SA Drakensberger breed.

- To quantify imputation accuracy from several custom-derived low-density panels, varying in 1) SNP density and 2) design, i.e. inclusion criteria for SNPs; to higher density (GGP® 150K) for the SA Drakensberger breed.

- To determine the value of using imputed genotypes in the prediction of genome-enabled breeding values (GEBVs) for birth- and weaning weight traits, i.e. both direct and maternal components, of the SA Drakensberger breed using a single-step genomic evaluation.

References

Aliloo, H., Mrode, R., Okeyo, A.M., Ni, G., Goddard, M.E. & Gibson, J.P., 2018. The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa. J. Dairy Sci. 101(10), 9108-9127.

Berry, D.P., McClure, M.C. & Mullen, M.P., 2013. Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. J. Anim. Breed. Genet. 131(3), 165-172.

Berry, D.P., Garcia, J.F. & Garrick, D.J., 2016. Development and implementation of genomic predictions in beef cattle. Anim. Front. 6(1), 32-38.

Bosman, D.J., 2002. Cattle breeds and types for the feedlot. Chapter 6 in: Feedlot management. Ed. Leeuw, K-J. Agricultural Research Council Animal Production Institute, Irene. pp. 84-90.

Carvalheiro, R., 2014. Genomic selection in Nelore cattle in Brazil. In: Proceedings of the 10th World Congress on Genetics Applied to Livestock Production. Vancouver, Canada, 17–22 Aug 2014.

Collins-Lusweti, E., 2000. Performance of Nguni, Afrikander and Bonsmara cattle under drought conditions in the North West Province of southern Africa. S. Afr. J. Anim. Sci. 30, 33-33.

DAFF, 2018. A profile of the South African beef market value chain. Available online at URL: https://www.nda.agric.za/doaDev/sideMenu/Marketing/Annual%20Publications/Commodity%20Profiles/Beef%20Market%20Value%20Vhain%20Profile%202018.pdf. pp 3-53.

Gororo, E., Makuza, S.M., Chatiza, F.P., Chidzwondo, F. & Sanyika, T.W., 2018. Genetic diversity in Zimbabwean Sanga cattle breeds using microsatellite markers. S. Afr. J. Anim. Sci. 48(1), 128-141.

Henderson, C.R., 1984. Applications of linear models in animal breeding. Volume 462. Guelph: University of Guelph.

Kim, J., Hanotte, O., Mwai, O.A., Dessie, T., Bashir, S., Diallo, B., Agaba, M., Kim, K., Kwak, W., Sung, S. & Seo, M., 2017. The genome landscape of indigenous African cattle. Genome Biol. 18(1), 34.

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(1), 88.

Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P., 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39, 906-913.

Marchini, J. & Howie, B., 2010. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11, 499-511.

Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genet. 157, 1819-1829.

Mrode, R., Ojango, J.M., Okeyo, A.M., & Mwacharo, J.M., 2018. Genomic selection and use of molecular tools in breeding programs for indigenous and crossbred cattle in developing countries: current status and future prospects. Front. Genet. 9, 694.

Muchenje, V., Dzama, K., Chimonyo, M., Raats, J.G. & Strydom, P.E., 2008. Tick susceptibility and its effects on growth performance and carcass characteristics of Nguni, Bonsmara and Angus steers raised on natural pasture. Anim. 2(2), 298-304.

Niemand, M., 2013. Feedlot performance of the Drakensberger in comparison with other cattle breeds: A Meta-analysis. PhD dissertation, University of Pretoria.

Nyamushamba, G.B., Mapiye, C., Tada, O., Halimani, T.E. & Muchenje, V., 2017. Conservation of indigenous cattle genetic resources in Southern Africa's smallholder areas: turning threats into opportunities - A review. Asian-Australasian J. Anim. Sci. 30(5), 603.

OECD, 2019. Meat consumption (indicator). DOI: 10.1787/fa290fd0-en (Accessed on 20 August 2018).

Pierce, M.D., Dzama, K. & Muchadeyi, F.C., 2018. Genetic diversity of seven cattle breeds inferred using copy number variations. Front. Genet. 9, 163.

Raoul, J., Swan, A.A. & Elsen, J.M., 2017. Using a very low-density SNP panel for genomic selection in a breeding program for sheep. Genet. Sel. Evol. 49(1), 76.

Rechav, Y. & Kostrzewski, M.W., 1991. The relative resistance of six cattle breeds to the tick Boophilus decoloratus in South Africa. Onderstepoort J. Vet. Res. 58, 181-186.

SA Drakensberger Breeders' Society, 2017. Drakensberger Handbook. 3rd Edition. Available online at URL: https://www.drakensbergers.co.za/files/DRAKENSBERGER%20HANDBOOK%203rd%20edition%20(April%202017)%20-%20small.pdf.

SA Stud Book, 2016. SA Stud Book annual report. Available online at: http://www.sastudbook.co.za/images/photos/Annual_Report_2016_a.pdf.

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., Maiwashe, A., Neser, F.W.C., Theunissen, A., Olivier, W.J., Mokolobate, M.C. & Hendriks, J., 2014. Livestock breeding for sustainability to mitigate global warming, with the emphasis on developing countries. S. Afr. J. Anim. Sci. 43, 269-281.

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.

United Nations, 2017. World population prospects: The 2017 revision. Volume I: Comprehensive tables (ST/ESA/SER.A/399). Department of Economic and Social Affairs, Population Division. Available online at URL: https://population.un.org/wpp/Publications/Files/WPP2017_Volume-I_Comprehensive-Tables.pdf.

USDA, 2018. GAIN report: South African beef imports expected to remain flat in 2018. Available online at URL: https://gain.fas.usda.gov/Recent%20GAIN%20Publications/South%20African%20Beef%20Imports%20Expected%20to%20Remain%20Flat%20in%202018_Pretoria_South%20Africa%20%20Republic%20of_4-5-2018.pdf.

Van Marle-Köster, E. & Visser, C., 2018. Genetic improvement in South African livestock: Can genomics bridge the gap between the developed and developing sectors? Front. Genet. 9, 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(3), 302-312.

CHAPTER TWO

LITERATURE REVIEW

Genotype imputation as a cost-saving genomic strategy for South African Sanga cattle: A review

S.F. Lashmar[1,2#], F.C. Muchadeyi[2] & C. Visser[1]

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

[2] Biotechnology Platform, Onderstepoort Veterinary Institute, Agricultural Research Council, P/Bag X05, Onderstepoort, Pretoria, 0110

Published in the South African Journal of Animal Science 2019, 49(2), 262-280.

Genotype imputation as a cost-saving genomic strategy for South African Sanga cattle: A review

S.F. Lashmar[1,2#], F.C. Muchadeyi[2] & C. Visser[1]

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

[2] Biotechnology Platform, Onderstepoort Veterinary Institute, Agricultural Research Council, P/Bag X05, Onderstepoort, Pretoria, 0110,

(Received 21 May 2018; Accepted 17 February 2019; First published online 11 April 2019)

Abstract

The South African beef cattle population is heterogeneous and consists of a variety of breeds, production systems and breeding goals. Indigenous cattle breeds are uniquely adapted to their native surroundings, necessitating conservation of these breeds as usable genetic resources to sustain efficient production of beef. Current projections indicate positive growth in human population size, with parallel growth in nutritional demand, in the midst of intensifying environmental conditions. Sanga cattle, therefore, are invaluable assets to the South African beef industry. Modern genomic methodologies allow for an extensive insight into the genome architecture of local breeds. The evolution of these methodologies has also provided opportunities to incorporate deoxyribonucleic acid (DNA) information into breed improvement programmes in the form of genomic selection (GS). Certain challenges, such as the high cost of generating adequate numbers of dense genotypic profiles and the introduction of ascertainment bias when non-commercial breeds are genotyped with commercial single nucleotide polymorphism (SNP) panels, have caused a lag in progress on the genomics front in South Africa. Genotype imputation is a statistical method that infers unavailable or missing genotypic data based on shared haplotypes within a population using a population or breed representative reference sample. Genotypes are generated *in silico,* providing an animal with genotypic information for SNP markers that were not genotyped, based on predictive model-based algorithms. The validation of this method for indigenous breeds will enable the development of cost-effective low-density bead chips, allowing more animals to be genotyped, and imputation to high-density information. The improvement in SNP densities, at lower cost, will allow enhanced power in genome-wide association studies (GWAS) and genomic estimated breeding value (GEBV)-based selection for these breeds. To fully reap the benefits of this methodology, however, will require the setting up of accurate and reliable frameworks that are optimized for its application in Sanga breeds. This review paper aims, first, to identify the challenges that have been impeding genomic applications for Sanga cattle and second, to outline the advantages that a method such as genotype imputation might provide.

Keywords: developing countries, indigenous breeds, genomics, breed improvement

# Corresponding author: s.lashmar@gmail.com

Introduction

South Africa accommodates a human population of approximately 57.7 million people (Statistics South Africa, 2018). According to UN projections, this number will increase to 72.8 million people by 2050 (United Nations, 2017). Nutritional demand, specifically the demand for animal protein, is expected to grow in parallel with population size and the responsibility of meeting this demand will weigh heavily on livestock production systems. Global warming will make bearing this responsibility a challenging task, with extreme environmental changes expected for developing countries south of the equator (Scholtz *et al*., 2013). Widespread regions of South Africa are experiencing a state of drought, which is the result of the strong El Niño event that occurred in the 2015/2016 year (South African Weather Service, 2016). The shortage of water in the country has raised concerns, among other things, about the amount of water it takes to finish cattle off in feedlots (Meissner *et al*., 2013). Approximately 68.6% of South African land is available for grazing, which is an ideal situation for extensive livestock production systems that rely on natural veld as feed source (Directorate: Knowledge and Information Management, 2017). Global warming, however, will be responsible for fluctuations in the nutritional value of the natural veld (Scholtz *et al*., 2014). Moreover, South Africa is geographically diverse and hosts a wide range of climatic zones, vegetation and soil types as well as a series of tick-borne and other endemic cattle diseases, each unique to its nine terrestrial biomes.

The South African beef population includes the Sanga cattle subspecies (*Bos taurus africanus*), which is indigenous. The breeds that belong to this subspecies include the Afrikaner, Drakensberger, Nguni and Tuli and the experimentally developed composite Bonsmara (Scholtz, 2010). These cattle, which are phenotypically distinguishable by their cervico-thoracic humps, resulted from historic crossbreeding between taurine and indicine cattle subspecies in eastern Africa (Payne & Hodges, 1997; Felius *et al*., 2014). From there, they were brought to southern Africa by migrating tribes (Schoeman, 1989), reaching South Africa by 250–500 AD (Payne & Hodges, 1997; Felius *et al*., 2014). Prolonged exposure to and endurance of the natural elements specific to South Africa have shaped the genomes of Sanga breeds to become habituated to the country's various extensive farming environments. Many trial-based studies have confirmed the ability of these breeds to adapt to, survive and reproduce in the varying beef-producing regions of South Africa. In the past decades several South African animal scientists have reported on and reviewed so-called proxy-indicator traits of adaptability for Sanga breeds. These studies have highlighted the potential of Sanga breeds to be highly fertile (e.g. Maule, 1973), calve easily and frequently (e.g. Scholtz, 1988; Schoeman, 1989; Collins-Lusweti, 2000), resist ticks and tick-borne diseases (e.g. Bonsma, 1980; Schoeman, 1989;

Rechav & Kostrzewski, 1991), and be well suited to extensive finishing systems of South Africa (e.g. Du Plessis *et al*., 2006). The genetic architecture of SA Sanga breeds and the molecular mechanisms underlying their adaptation abilities, however, have been only partially investigated.

The number of genomics studies on beef cattle has grown steadily in South Africa over the past five years. The initial validation of the utility of the Illumina® bovine SNP50 bead chip for indigenous South African cattle by Qwabe *et al*. (2013) provided the first insight into genomics of Sanga cattle. Validation was important because no Sanga breeds were included in the SNP discovery process of this bead chip (Matukumalli *et al*., 2009). As was expected, lower minor allele frequency (MAF) was observed and consequently a lower number of informative SNPs for indigenous, non-discovery breeds and crossbreeds. Zwane *et al*. (2016) and Lashmar *et al*. (2018) confirmed the tendency towards low MAF in Sanga breeds using GeneSeek®'s GGP 80K and GGP 150K bead chips, respectively. In a series of successive studies, the authors who did the original validation investigated the genetic diversity (Makina *et al*., 2014; Makina *et al*., 2016), linkage disequilibrium (LD) and effective population size (Ne) (Makina *et al*., 2015a) as well as selection signatures (Makina *et al*., 2015b) of indigenous breeds. The SNP50 bead chip was also utilized to identify copy number variations (CNVs) in Nguni cattle (Wang *et al*., 2015). Genes associated with biological processes such as immune and abiotic stress were among the genes identified within these CNV regions (Wang *et al*., 2015). These results were in partial agreement with the study by Makina *et al*. (2015b), in which it was revealed that a heat shock protein gene (*HSPB9*) and immune response genes were under selection in indigenous breeds. Mapholi *et al*. (2016) used SNP50 bead chip data to perform a GWAS to investigate tick resistance in a sample of close to 600 Nguni cattle. These authors identified several regions, in concordance with previous research, which harbour quantitative trait loci (QTL) that are linked to various tick-count traits. Only three of these regions, however, reached the threshold for genome-wide association significance, which necessitates further validation.

These studies provided a baseline for further investigation of SNP technology in South African beef cattle, especially for Sanga breeds. The local importance of these breeds, in terms of the role that they will play in enriching the South African beef industry, has attracted attention to the incorporation of genomic information into breed improvement programmes in the form of GS. Genomic selection, however, relies on a good phenotypic and genotypic backbone. Owing to the relative size of the breed (about 120 000 animals) (Van der Westhuizen *et al*., 2014) and the completeness of phenotypic records, the SA Bonsmara was the first South African beef cattle breed for which GEBVs were estimated. This was preceded by a study that elucidated the population genetic structure within a possible reference population for the SA Bonsmara breed (Bosman *et al*., 2017). The Beefmaster, a synthetic breed composed of Brahman, Hereford and Shorthorn genetics (Porter, 1991), followed suit and was the second beef cattle breed for which GEBVs were released (Beefmaster Cattle Breeders Society of South Africa and SA Stud Book, 2017). The lack of accurate pedigree recording still prohibits the application of GS for a large portion of South African beef cattle breeds. Pedigree recording, however, has improved and for the Afrikaner, Drakensberger, Nguni and Tuli breeds the percentage of average pedigree completeness for first-generation animals is about 90% (Abin *et al*., 2016). Although performance testing has been available for commercial beef populations since 1959, and all indigenous breeds are participating actively, participation by some breeders' societies is as low as 32% (Scholtz, 2010; Van Marle-Köster *et al*., 2013).

Accumulation of sufficient genotypic information to initiate GS for breeds such as the SA Bonsmara was partly made possible through the Beef Genomics Project (BGP), which is a collaborative research-focused project that includes researchers from universities and research councils, breeders' societies and other role-players. The BGP is a large-scale project that is funded by the Technology Innovation Agency (TIA; Department of Science and Technology) with the overall aim of instating genomic improvement in the South African beef population (Becker, 2016). This funding initiative allowed the generation of approximately 7 000 genotypes across 16 breeds over a three-year period of routine genotyping (Van Marle-Köster & Visser, 2018). Before the BGP, the scale of genomic studies on beef cattle in South Africa was limited compared with international studies. South African studies have typically included small sample sizes and have focused on genome characterization and population genetic analyses. It is only since the inception of the BGP that improvements in the number of genotyped animals have made applications such as GS a possibility. The main reason for the lag in introducing these genomic technologies earlier for South African beef cattle was related to financial constraints. Even with major reductions in genotyping costs over the past decade, the per-animal price of genotyping is still expensive, especially for researchers in the developing world where adapted indigenous cattle resources, that often have unique and uncharacterized genomes, are located. Many local researchers therefore collaborate with international research groups to fund studies on indigenous breeds. Programmes, such as BGP, therefore assists in mitigating the financial burden on individual research groups in local universities and research organizations to procure dense SNP genotypes towards realizing GS. Routinely generated genotypes could furthermore serve as a data resource for research groups with different research objectives that require large numbers of genotyped animals from certain breeds. These programmes also assist in establishing intra-national collaborations and building local capacity.

Large numbers of genotyped animals are required to set up breed reference populations that are suitable for genomic evaluation. Post BGP, the financial responsibility of genotyping key animals in herds or populations will be that of interested breeders or farmers, and the current high costs of genotyping will probably impede further uptake of genomics if the benefits of this technology, in relation to the cost, cannot be realized. In contrast with the dairy industry (globally and in South Africa), where relatively few large breeds dominate the national herd, the beef industry is diverse and consists of approximately 30 breeds of taurine, indicine and Sanga descent, including taurindicine crosses and composite breeds (Strydom, 2008). These subspecies are diverse, with breeds belonging to each subspecies displaying distinct genetic and phenotypic characteristics. Breeds also differ in their population size and the traits recorded for, as breeding objectives are breeder society specific (Van Marle-Köster *et al*., 2013). Including DNA information in breed improvement is a key objective to achieve genetic gain for local cattle and to exploit the adaptive mechanisms they possess. The effective introduction of GS, however, has phenotypic and genotypic requirements. To fulfil the genotypic requirements would necessitate i) the generation of high-density genotypes to be more affordable; ii) the availability of more genomic profiles per breed (approximately 1000, but depending on the breed structure); iii) the ability to standardize between genotyping platforms; and iv) GS algorithms to be optimized for the genomic structure of local breeds. Genotype imputation is a methodology that uses predictive algorithms to enable the procurement of un-genotyped genomic information, and thereby allows for saving on genotyping costs. It is an ideal candidate methodology to assist in fulfilling the genotypic requirements for GS and, if properly validated and optimized for routine application, can accelerate genomics research in South Africa cost effectively.

This review aims to elucidate genotype imputation as a genomic strategy, focusing on its relevance to indigenous Sanga cattle breeds. The review discusses the methodology behind imputation, the factors that influence imputation accuracy, and its value in future applications such as GWAS and GS.

Genotype imputation

Genotype imputation is a statistical methodology that involves the prediction and simulation of missing SNP genotypes from observed or non-missing genotypes through model-based approaches (Marchini *et al*., 2007). The models are based on population genetic principles, and are used to deduce or extrapolate allelic correlations for sample sets with missing genotypic data using a sample set with dense and ‘complete’ data as a reference (Howie *et al*., 2009). Imputation of missing genotypes generally relies on the fact that extended haplotypes are shared over short distances between animals with a common ancestor (Pei *et al*., 2008). In effect, imputation therefore relies on family linkage, that is, the rules of Mendelian inheritance, or the LD structure within a population. If a reference population is genotyped on a high-density SNP panel and a test population is genotyped for a smaller subset of these SNPs, the assumption is that if they are related in some way, these populations should have similar underlying patterns of LD (Pei *et al*., 2008). The genotypic information of a reference population or sample is used to model patterns in genomic variation (Browning, 2008) and this genomic variation, which is typically shared within population, can therefore be used to infer missing genotypes in a non-reference animal or test sample.

Imputation therefore is population specific, and the accuracy with which genotypes can be imputed depends on the persistence of LD between animals in the reference and test populations. Imputation is hence viable only across breeds or populations if they are genetically similar or related in some way (Berry *et al*., 2014). Synthetic or composite breeds may therefore benefit from imputation if their component breeds are pooled in the reference population (Browning, 2008; Ventura *et al*., 2014). Imputing missing genotypes from a reference population that belongs to an ancestrally different breed from the test population will result in low-quality imputation and will negatively affect the reliability, and hence utility, of imputed genotypes (Browning, 2008).

If high-density genotypes can be imputed reliably from low-density SNP arrays with sufficient accuracy, this would allow for the opportunity to genotype more animals in a more affordable way (García-Ruiz *et al*., 2015). Imputation is therefore an important statistical tool for enriching applications such as GWAS and GS that require higher or more evenly distributed marker densities (Marchini *et al*., 2007). This tool will also enable comparison between SNP bead chip data that i) have been developed by different companies (e.g. AffymetrixTM, Illumina® and Neo-Geneseek®), and ii) that incorporate different SNP densities (e.g. 50K, 150K and 777K) in meta-analytic approaches (Ellinghaus *et al*., 2007). This can be done by identifying a set of SNPs that are common to two platforms and then imputing to a standard density. For example, if low- (e.g. 7K), medium- (e.g. 50K) and high-density (e.g. 150K) genotypic information is available for a breed, the low- and medium-density information can be imputed to the high density and this high-density information can be merged and collectively used for downstream analyses. This standardization might be useful when one considers the fast rate at which new and improved genotyping platforms are becoming available (Nicolazzi *et al*., 2015).

Imputation algorithms can generally be categorized as population-based and pedigree-based methods. Population-based algorithms assume large numbers of animals with unknown pedigree data and therefore rely on population-wide LD between markers (Weigel *et al*., 2010). These algorithms use a probabilistic approach to perform imputation and rely on shorter haplotypes that are typically less than 1 centiMorgan (cM) in length (Antolín *et al*., 2017). Probabilistic methodologies are ideal for natural populations and are more viable for high-density bead chips (Weigel *et al*., 2010; Mulder *et al*., 2012) because using methods that rely solely on LD information would be affected negatively when SNP densities are sparse and LD is low (Wang *et al*., 2016).

Pedigree-based algorithms use heuristic methods that assume the existence of family structure and rely on long haplotypes, typically larger than 10 cM in length, which are shared between closely related animals (Antolín *et al*., 2017). Imputation therefore cannot be reliably performed with these methods if accurate pedigree information is lacking. These methods are more suited to case-control studies or studies in which family trios – both parents and offspring – have been sampled. A number of imputation software programs exist, which differ in their approach to employing the methods discussed. The most popular imputation software programs in animal breeding and genetics research are listed in Table 2.1, which was adapted from Calus *et al*. (2014) and Antolín *et al*. (2017).

Table 2.1 List of commonly used imputation software programs in animal genetics research.

Software	Method	Reference
AlphaImpute	Heuristic	Hickey *et al*. (2011)
Beagle	Probabilistic	Browning & Browning (2007)
CHROMIBD	Heuristic	Druet & Farnir (2011)
DAGPHASE	Heuristic	Druet & Georges (2010)
*fast*PHASE	Probabilistic	Scheet & Stephens (2006)
FImpute	Heuristic	Sargolzaei *et al*. (2014)
Findhap.f90	Heuristic	Van Raden & Sun (2014)
IMPUTE1	Probabilistic	Marchini *et al*. (2007)
IMPUTE2	Probabilistic	Howie *et al*. (2009)
MaCH	Probabilistic	Li *et al.* (2010)
minimac	Probabilistic	Howie *et al*. (2012)
minimac2	Probabilistic	Fuchsberger *et al*. (2014)
PedImpute	Heuristic	Nicolazzi *et al. (*2013*)*
PLINK	Probabilistic, sporadic	Purcell *et al*. (2007)

To summarize, probabilistic methods that rely on LD information generally use hidden Markov models (HMM) to perform imputation. HMMs are probabilistic models that allow assumptions and inferences to be made about hidden variables, each with a finite set of possible 'states', based on observable outputs (Rabiner, 1989). These models estimate the probability that a certain state could be responsible for producing a certain observable output at a given time (Rabiner, 1989). These models therefore use the underlying relationship, or correlation, between observed and unobserved SNPs to infer the most probable genotype for the unobserved SNPs. HMM methodology can be computationally demanding because it requires that genotypes are phased and estimate recombination rates. However, it can be supplemented with heuristic methodology that exploits more accurate phasing information owing to the incorporation of family linkage (Antolín *et al*., 2017). The addition of pedigree information, albeit not a necessity, will therefore boost imputation accuracy. In the same way, HMM methodology can add value to heuristic methods if pedigree information for only one parent is available.

The algorithms and exact methodology incorporated in each of the software programs (Table 2.1) are discussed in detail in each of the scientific papers and have been reviewed by for example Antolín *et al*. (2017) and Wang *et al*. (2016). Various statistical models have been proposed including HMM, haplotype clustering algorithms, linear regression models and expectation-maximization algorithms (Pei *et al*., 2008; Howie *et al*., 2009; Wang *et al*., 2016). Statistical methods differ in their approach to capturing haplotypes that are shared in a population (Pei *et al*., 2008). The choice of software therefore influences imputation accuracy. Several factors that are within and beyond the researcher's control might affect the quality of imputed genotypes.

Factors that affect imputation accuracy

The accuracy with which SNP genotypes can be imputed will determine the utility and reliability of a given imputation method for a given population. Parameters to quantify imputation accuracy can generally be subdivided into two groups, namely those that determine i) the proportion of alleles or genotypes that were correctly (e.g., Weigel *et al*., 2010) or incorrectly imputed (e.g. Druet *et al*., 2010; Zhang & Druet, 2010), and ii) the correlation or squared-correlation, usually the Pearson method of correlation, between true and imputed genotypes (e.g. Huang *et al*., 2009; Mulder *et al*., 2012; Ma *et al*., 2012). An alternative parameter, the imputation quality score (IQS), which determines imputation accuracy on the basis of statistics of agreements and adjusts for chance concordance has also been proposed (Lin *et al*., 2010). Methods that determine the concordance rate of imputed alleles (that is, proportion correctly imputed) will be more informative for imputing genotypes for GS, since GS algorithms assume additive allele effects (Berry *et al*., 2014). These methods, however, tend to inflate imputation accuracies because of sensitivity to low-frequency alleles (MAF<5%) (Ramnarine *et al*., 2015). Imputation of rare variants might therefore be better assessed using for example correlation-based accuracy measures that are less sensitive to MAF. All of these parameters are dependent on the availability of true and imputed genotypes, that is, scenarios in which a proportion of the true genotypes are available, but masked. A third category of accuracy quantification involves statistics that are built into programs such as BEAGLE (Browning & Browning, 2007), which do not require the availability of 'true' genotypic data, but rather estimate accuracies based on the likelihood or expectation of genotypes and allele dosage (Ramnarine *et al*., 2015).

Imputation accuracy depends on many factors including the imputation method, the MAF of the SNP to be imputed, LD between SNPs, the chromosomal position of the SNP (that is, whether it is located in the centre of the chromosome or on the chromosomal extremes), the quality of the SNP map, the discrepancy in SNP densities between high- and low-density SNP panels, and the size and composition of the reference population (Schrooten *et al.*, 2014). Setting up a framework for incorporating imputation, as a routine genomic procedure, would require the optimization of imputation methodology for specific breeds or breed groups such as Sanga cattle. This might necessitate adjustments for example for national breed size (that is, numerically small versus numerically large) and variations that might occur across the genome, which might be the case for admixed populations. In addition, the algorithms that are built into certain imputation software might not be suited to the genomic architecture of certain breeds or populations. The main factors that need consideration are detailed below and are focused on the Sanga context.

Reference population size and degree of relatedness to the test population

The main consideration in the experimental design of an imputation study is determining an appropriate set of reference haplotypes or animals to achieve accurate predictions (Li *et al*., 2009). The most common method for assessing the influence of reference population size on the accuracy of genotype imputation is by masking varying subsets of reference animals and comparing the accuracies recovered for increasing reference population sizes (Li *et al*., 2009). Accuracy measures generally improve with larger reference population sizes, and this trend has been well documented (e.g. Hayes *et al*., 2012; Pausch *et al*., 2013; Piccoli *et al*., 2014; Ogawa *et al*., 2016). The improvement in accuracy, however, is not so pronounced for high-density bead chips compared with lower density ones (Ogawa *et al*., 2016). That is, the effect of reference population size is reduced if fewer genotypes are to be imputed. The improvement in imputation accuracies with increasing reference sample could be because more animals are used to construct haplotype libraries from which to impute. The inclusion of more reference animals increases the probability of including more breed or population representative haplotypes. The rule of thumb has generally been that a reference sample size of about 1000 animals will be sufficient for imputation. However, the ideal population size for a breed will depend on breed dynamics.

The breed dynamic of the South African beef population has seen significant changes in the past. In the 1970s, when the national herd was approximately 6 million head smaller than the approximately 13.6 million animals recorded today (Meissner *et al*., 2013), the indigenous Afrikaner dominated national herd numbers. Purebred Afrikaner and Afrikaner crosses represented approximately 70% of all the cattle slaughtered (Van Marle, 1974). Today, the experimentally developed SA Bonsmara composite is the most abundant breed in South Africa (about 120 000 registered animals) (SA Stud Book, 2016). The Nguni is the most abundant Sanga breed (about 38 000 registered animals) (SA Stud Book, 2016). Registered animals that belong to Sanga breeds such as the Afrikaner (about 7300 registered animals) (SA Stud Book, 2016), Drakensberger (about 12 800 registered animals) (SA Stud Book, 2016) and Tuli (about 9500 registered animals) (SA Stud Book, 2016) are far outnumbered by composite breeds such as the SA Bonsmara and SA Beefmaster (about 48 000 registered animals) (SA Stud Book, 2016).

The SA Bonsmara was the first South African beef cattle breed to receive genomic evaluations. This is attributable to the relative size and market share of the breed and superior record keeping by the breeders' society, and hence availability of phenotypes. In the BGP, the generation of genotypes per breed was conditional on the size and breeding objectives, the availability of phenotypic data, and the long-term prospects of implementing GS successfully (SA Stud Book, 2016). Approximately 42% of the animals participating in Logix Beef (SA Stud Book's animal recording database) belong to the SA Bonsmara, while only 2.4%, 4.5%, 5.6%, and 3.3% of the animals that participate are Afrikaner, Drakensberger, Nguni and Tuli, respectively (SA Stud Book, 2016). The eventual aim is to implement GS for all these indigenous breeds. Establishing breed-appropriate reference populations, however, requires the accumulation of sufficient genotypic information as well. To date approximately 300, 1

850, 960, 400 and 200 genotypes have been generated for the Afrikaner, SA Bonsmara, Drakensberger, Nguni and Tuli breeds, respectively, during the timeframe of the BGP (personal communication). Genotypic profiles were mostly generated using the GeneSeek Genomic Profiler 150K bovine chip.

The ideal reference population is not influenced solely by the size of the reference population, but by its composition as well. Imputation accuracy improves if there is a level of relatedness between the reference and test samples. Berry & Kearney (2011) for example showed a positive correlation between the average relatedness between reference and test samples and imputation accuracy. This correlation strengthened when the maximum relatedness of test animals were considered instead of the average relatedness (Berry & Kearney, 2011). These authors observed an approximate 6% increase in genotype concordance rate for animals with both – as opposed to no – parents in the reference sample. Even though there are many imputation algorithms that can perform imputation fairly accurately for unrelated samples, the inclusion of parental or familial genotypes in the reference population will assist in boosting accuracy estimates. The reason for higher accuracy is due to closer linkage and therefore sharing of larger genomic segments within families. It would therefore be advisable or beneficial to include parent-offspring pairs or family trios in sampling efforts. This would be easier for dairy cattle, as opposed to beef, when one considers the higher prevalence of reproductive biotechnologies in the dairy industry.

In the SA Bonsmara for example strong genetic linkage between the animals that were genotyped was ensured by encouraging all breeders to submit hair samples of influential herd sires (SA Stud Book, 2017). Sampling was also done to ensure the inclusion of animals with accurate BLUP breeding values (SA Stud Book, 2017). This included female animals with superior breeding value accuracies for traits such as age at first calving, calving interval as well as maternal birth and weaning weights (SA Stud Book, 2017). Animals were selected across a spectrum of good and bad performers for these traits. The same process is utilized to select appropriate animals to genotype for other Sanga breeds. Although the focus of selection was not to specifically sample family trios, selecting genetically influential animals would indirectly assure the genotyping of animals that are directly related. If parent-offspring pairs (sire-offspring or dam-offspring) have been genotyped with BGP, additional funding could possibly be used to complete family trio genotypes by genotyping or imputing the missing parent. Berry *et al*. (2014) tested the imputation of parental genotypes based on their half-sib progeny and concluded that ungenotyped parental genotypes could be inferred accurately if genotypes were available for a sufficient number of offspring.

Genome resources and SNP arrays

There were originally two competing genome assemblies for cattle. The first reference genome, Btau_1.0, was assembled by the Human Genome Sequencing Centre at the Baylor College of Medicine (Elsik *et al*., 2009), whereas the Centre for Bioinformatics and Computational Biology at the University of Maryland initiated the assembly of an alternative genome, namely UMD2 (Zimin *et* *al*., 2009). Efforts to assemble these reference genomes occurred concurrently. The assembled genomes went through several stages of improvement over the years, resulting in the versions (Btau_5.0.1 and UMD3.1.1) that are currently available to researchers through the National Centre for Biotechnology Information (NCBI). At the 11th World Congress for Genetics Applied to Livestock Production the release of a new de novo assembly, ARS-UCD, of the Dominette Hereford genome was announced (Rosen *et al*., 2018). Long-read sequencing was utilized to reach 80X genome coverage with approximately 100- and 200-fold fewer gaps than the Btau_5.0.1 and UMD3.1 assemblies, respectively (Rosen *et al*., 2018).

From the two initial genome assemblies, SNP identification followed. Today various commercial SNP bead chips are available for cattle through three leading companies (AffymetrixTM, Illumina®, Neogen's GeneSeek®) (Nicolazzi *et al*., 2015). The new assembly will aid in i) improving genome continuity, ii) re-mapping reads, and iii) improving marker ordering, which might influence SNP selection for the development of SNP genotyping platforms in the future (Rosen *et al*., 2018). The commercial bead chips that are currently available for cattle are summarized in Table 2.2, adapted from Nicolazzi *et al*., (2015). There are also a growing number of custom-made bead chips that are protected by intellectual property, and are therefore not available for commercial utilization (Nicolazzi *et al*., 2015).

Table 2.2 Summary of available single nucleotide polymorphism bead chips for cattle.

Company	Bead chip	Number of SNPs
Affymetrix®	Axiom® Genome-wide BOS1	648 875
Geneseek®	GeneSeek Dairy Ultra LD v2 GGP-LD	7 049
	- version 1 (GGP9K)	8 610
	- version 2 (GGP20K)	19 721
	- version 3	26 151
	GGP-indicus	35 090
	GGP-HD	76 879
	GGP-150K	139 480
Illumina®	Golden Gate Bovine 3K	2 900
	Bovine LD	
	- version 1	6 909
	- version 1.1	6 912
	- version 2	7 931
	Bovine SNP50	
	- version 1	54 001
	- version 2	54 609
	Bovine HD	777 962

Many of the bead chips that are listed in Table 2.2 have large numbers of SNPs in common. Some are updated versions of previously released bead chips, including additional SNPs that are optimized for specific purposes. Illumina's Bovine LD (6 909 SNPs), which was made available in 2011, for example has mostly replaced the Golden Gate 3K (2 900 SNPs), which was released in 2010 (Wiggans *et al*., 2013). A subtotal of 2 159 SNPs were retained from the Golden Gate (Wiggans *et al*., 2013). Geneseek recently released the bovine Genomic Profiler 150K (GGP 150K) SNP bead chip, which features 139 480 SNPs with an average inter-SNP distance of about 19 kilobase pairs (kb). The GGP 150K bead chip incorporates approximately 74 000, 42 000, 25 000 and 23 000 SNPs that are included on the original GGP HD (80K), Bovine SNP50 (Illumina), GGP LD and Bovine HD (777K, Illumina) bead chips, respectively. One of the most recently released platforms, the GeneSeek GGP Indicus bead chip, features about 35 000 indicine-specific SNPs that were selected from a cohort of breeds, including the Brahman, Nellore, Gyr and Santa Gertrudis and tropical composite breeds (Ferraz *et al*., 2018). The GGP indicus chip was also optimized for high imputation accuracy in indicine breeds, with imputation to the Illumina HD bead chip being up to 97% accurate in these breeds (Ferraz *et al*., 2018). Previous research has shown improved MAF and LD estimates for indigenous Ethiopian cattle (Edea *et al*., 2015) and improved imputation accuracy for indicine Gyr cattle (Boison *et al*., 2015) when indicus-derived SNP panels were used. Since Sanga cattle are taurine-indicine hybrids, it would be recommended to test the utility of this panel for Sanga breeds, in contrast to taurine-derived SNP panels. Lower-density bead chips in general are increasingly being developed with the aim of retaining specific subsets of SNPs, in common with higher density bead chips, to be optimized for low-cost genomic applications (Boichard *et al*., 2012).

The bovine reference genome will undergo many updates in the future as sequencing technologies improve. This is an important consideration because re-mapping of SNPs can cause rearrangements in the haplotypes captured by bead chips, especially if SNPs are re-mapped to different chromosomes (Milanesi *et al*., 2015). Incorrect SNP positioning of more than 5 000 SNPs on the bovine HD bead chip has been suggested (Pausch *et al*., 2013). Pre-imputation, ensuring consistency between SNP positions of low- and high-density panel genotypic data is an important quality check. Software such as the web-based tool SNPchiMp (Nicolazzi *et al*., 2015) provides a platform for standardizing SNP genomic positions by mapping markers to the same reference genome (that is, either to the UMD3.1, Btau_5.0.1 or ARS-UCD genome assemblies). Accurate mapping of SNPs is also important owing to the decrease in imputation accuracy, which has been observed for SNP genotypes on chromosomal extremes, as opposed to SNPs that are located in the centre of the chromosome, in previous studies (e.g. Ventura *et al*., 2016).

Studies that focus on Sanga beef cattle have utilized Illumina SNP50 (e.g. Qwabe *et al*., 2013), GeneSeek GGP 80K (e.g. Zwane *et al*., 2016) and GeneSeek GGP 150K (e.g. Lashmar *et al*., 2018) genotypic data for various genomic applications. Since the initiation of the BGP, data has been routinely generated using the GeneSeek GGP 150K panel, whilst specific research projects have generated low-density (e.g. Illumina 7K) and high-density (e.g. Illumina HD) genotypic data and whole-genome sequencing information, depending on the research interest. The diversity in genomic data that is available for Sanga cattle might therefore require standardization between platforms in future efforts to combine data in meta-analyses and collaborative projects.

The discrepancy in the number of SNPs between panels also influences imputation accuracy, specifically between low- and high-density panels (that is, the number of SNPs to be imputed). Accuracy estimates tend to improve with increasing SNP density of the low-density panel. That is, the fewer the number of SNPs that need to be imputed, the higher the mean imputation accuracy will be. This increase in accuracy can be attributed to the fact that haplotypes can be more accurately resolved with more SNPs present (Tsai *et al*., 2017). Imputing Dutch Holstein genotypes to a custom 60K bead chip, Zhang & Druet (2010) observed decreasing imputation error rate when the SNP density of the low-density panel was improved from 384 SNPs to 6 000 SNPs. These authors then suggested a minimum of 3 000 SNPs to achieve 3% or lower imputation error rate (Zhang & Druet, 2010). Ogawa *et al*. (2016) indicated a similar trend in imputing to 50K genotypes for Japanese Black beef cattle. Imputation accuracy was 2.7% higher when the low-density panel included 10 000 SNPs versus 500. This relationship agrees with imputation experiments on sheep (Hayes *et al*., 2012), salmon (Tsai *et al*., 2017), and maize (Hickey *et al*., 2012) and suggests that a minimum set of SNPs is required to allow optimal imputation.

The minimum number of SNPs that are necessary for accurate imputation must be determined to develop an imputation-driven low-density panel for Sanga cattle. For individual breeds, this will depend on the extent of LD within the population. Breeds that are characterized by lower LD will require higher SNP densities. Developing a low-density panel that is applicable across Sanga breeds will also depend on the persistence of LD across Sanga breeds. If the persistence of LD across breeds is low, a higher number of SNPs would be necessary as a minimum for the low-density panel in imputation.

Pre-imputation processing of genotypes

Pre-imputation procedures such as quality control (QC), DNA strand checking and phasing aid in processing and preparing genotypic data to optimize imputation accuracy. QC is an important first step in any genomics study, and serves to remove uninformative samples and markers in preparation of downstream analyses. Sample-based QC will exclude individual animals with discordant sex information (when pedigree- versus SNP-based gender assignment disagree), that have high percentages of missing genotypes (have a low call rate), display outlying heterozygosity rates and show evidence of non-Mendelian inheritance (Li *et al*., 2009; Anderson *et al*., 2010). Marker-based QC involves excluding SNPs with low genotype call rate, those that deviate significantly from Hardy-Weinberg equilibrium, those with low MAF, and those that have duplicated or unknown genomic positions (Anderson *et al*., 2010; Purfield *et al*., 2016). The stringency of the quality filters applied prior to imputation may influence the accuracy of imputed SNPs. Roshyara *et al*. (2014) proposed that for small to moderate datasets, less stringent or no QC procedures prior to imputation would be best practice owing to the detrimental effect that stringent QC procedures might have on the quality of imputation for such datasets. More stringent QC procedures might also discard SNPs that could have been successfully imputed (Roshyara *et al*., 2014) or rare SNPs that are informative for the expression of traits of interest, such as those pertaining to the adaptive mechanisms of Sanga cattle. Purfield *et al*. (2016) for example investigated the effect of sample call rate as a QC parameter on imputation accuracy, and observed improved genotype and allele concordance rates with increasing animal call rate. Genotype concordance rates improved from 0.41 to 0.95 when animal call rates were <40% versus when call rates were between 95% and 99% (Purfield *et al*., 2016). These authors consequently proposed a cut-off of 85% as the lower limit for exclusion of animals based on call rate. QC procedures should therefore be optimized to retain high-quality data, but should not compromise the representation of SNPs in haplotype libraries used for imputation.

The quality of genotype imputation depends on whether the allele calls that are being generated for the test population are from the same physical DNA strand in relation to the reference genome (Verma *et al*., 2014). Determining the DNA strand orientation is therefore an essential pre-processing step. SNP annotations also differ for datasets, and on the genotype platform and the genotype-calling algorithm (Verma *et al*., 2014). Illumina for example uses a TOP/BOT method, which designates top (TOP) and bottom (BOT) DNA strands based on the SNP and its flanking sequence, and calls alleles based on a generalized 'Allele A' and 'Allele B' nomenclature (Illumina, 2006). Genotypic information can also be provided in a forward/reverse orientation. Inconsistencies between genotypic data between the reference and test populations with regard to strand orientation and allele coding – that is, whether genotypes are coded as A/B or A/C/G/T format – can impede accurate imputation. Certain imputation software programs, such as BEAGLE (Browning & Browning, 2007), IMPUTE2 (Howie *et al*., 2009), and PLINK (Purcell *et al*., 2007), have DNA strand checking utilities to determine strand orientation and to subsequently convert to a flipped strand orientation when necessary. During this procedure, alleles are converted to their complements based on observed alleles and the MAF and LD pattern that is observed for SNPs, and then removed when inconsistencies in these parameters cannot be resolved (Verma *et al*., 2014). The developers of SNPchiMp (Nicolazzi *et al*., 2015) have made bioinformatics tools available in the form of an application called SNPConvert, which can also assist researchers in standardizing allele coding and strand orientations for SNP genotypic data.

Haplotype phasing involves determining from which of the parental chromosomes or haplotypes SNP alleles are derived or on which they are located (Browning & Browning, 2011). SNP genotypic data is generally unphased and for the purpose of imputation it is essential to know the origin and location, i.e. on which DNA strand, of each allele of a bi-allelic SNP (Browning & Browning, 2011). Accounting for unknown phase is, however, computationally intensive and can be time-consuming (Howie *et al*., 2012). Some imputation software such as BEAGLE (Browning & Browning, 2007), however, performs phasing as part of the imputation process. In other cases, third-party software such as SHAPEIT (O'Connell *et al*., 2014) is available for 'pre-phasing' genotypic data in a two-step imputation approach. In the two-step approach, observed genotypes are firstly phased and then the phased genotypic information is used for imputation. Pre-phasing will be useful in speeding up computation time of the overall imputation process but the value thereof will depend on the accuracy with which haplotypes can be estimated (Howie *et al.*, 2012). Nevertheless, haplotype phasing will become increasingly important in future efforts to impute to sequencing data and the methods currently available are comprehensively reviewed in Browning and Browning (2011).

Population-specific parameters: minor allele frequency

Minor allele frequency (MAF) can have significant effects on the reliability of imputation, and a number of authors have investigated the influence of varying levels of MAF on imputation accuracy (e.g. Hozé *et al*., 2013; Schrooten *et al*., 2014; Van Binsbergen *et al*., 2014). The effect of low MAF versus high MAF on imputation accuracy will be determined primarily by how 'accuracy' is defined, that is, whether it is quantified as a proportion of correctly imputed genotypes or as a correlation between observed and imputed genotypes. It has consistently been found that accuracy quantified as a proportion or percentage of correctly imputed genotypes is correlated negatively with increasing MAF, whereas the correlation-based measure shows a positive relationship. Investigating different maize lines, Hickey *et al*. (2012) observed decreasing percentage-based accuracy, as opposed to increasing correlation-based accuracy, for increasing levels of MAF. For proportion-based or percentage-based measures, it is debated that when the frequency of the minor allele is low, there is a greater likelihood that imputation algorithms will predict genotypes as homozygous for the major or common allele (Hickey *et al*., 2012). Conversely, high MAF creates more uncertainty, thereby deeming predictions less reliable and hence less accurate. The correlation-based measure is less dependent on allele frequency and assumes that low-MAF SNPs are not sufficiently segregating within the population, and therefore cannot be easily imputed based on shared haplotypes (Hickey *et al*., 2012). Similar relationships were observed in European cattle, where Ma *et al*. (2012) indicated a higher 'correct rate' for lower MAF versus lower correlations for lower MAF across six widely used imputation software programs. These authors also observed that software that incorporated pedigree information was more sensitive to variation in MAF (Ma *et al*., 2012). In pigs, Badke *et al*. (2013) indicated the same trends with regard to the relationship between accuracy measures and MAF, but found that proportion-based measures also improved with increasing MAF when inter-SNP differences in MAF are adjusted for.

Factors pertaining to the experimental design of imputation studies may influence the effect size of MAF on imputation accuracy. These include the size of the reference population and the method of imputation implemented within software programs. Imputation accuracy can be improved, and error rate lowered (Huang *et al*., 2009), for rare SNP if a larger, more extensive reference population is used (Howie *et al*., 2009). Allele frequencies of rare SNP are typically overestimated when a small reference population is used (Howie *et al*., 2009). Possible adverse effects that low MAF might have on imputation accuracy will therefore be gradually alleviated with increasing animal numbers and improved representation.

MAF is essentially an indication of whether a SNP is segregating within a given population and therefore the composition of the reference population plays an important role as well. Boichard *et al.* (2012) indicated higher MAF, and therefore higher imputation accuracy, for cattle breeds that were used to design the bead chip under study. In humans, Howie *et al*. (2011) observed that although population-specific reference panels tend to outperform HapMap panels for imputation accuracy, reference panels that are 'ancestrally inclusive' and non-specific, may capture poorly represented low-frequency alleles. This would be important when genotypes need to be imputed for composite or crossbreeds of uncertain or unknown genetic composition such as the Drakensberger Sanga breed. Alleles that occur in low frequencies are not necessarily presented in reference haplotypes and therefore certain imputation software may have difficulty deriving the correct allele, which would affect the reliability of accuracy estimates directly (Schrooten *et al*., 2014). Software programs differ in their ability to detect copies of the minor allele. Howie *et al*. (2009) showed that some methods are more prone to erroneous minor allele calls. Certain software programs are better equipped to deal with low-frequency SNPs. The use of the appropriate method to optimize imputation for non-commercial breeds, which might be disadvantaged by ascertainment bias, would be an important consideration.

Lower average MAF has been observed for Sanga versus exotic breeds, verifying the existence of ascertainment bias (Qwabe *et al*., 2013). Furthermore, higher MAF was observed for a Sanga crossbreed (Angus x Nguni) than for a 'pure' Sanga breed (Nguni) (Qwabe *et al*., 2013). The use of a commercial bead chip such as the Illumina bovine SNP50 might therefore be more useful for crossbreeds that carry taurine haplotypes and therefore display higher MAF for the SNPs that were discovered for the chip. In imputation for instance this will be useful only when these haplotypes are represented by an appropriate reference sample that is, either of the component breeds or high-impact animals from the crossbred population. Investigating the impact of MAF on imputation accuracy is important not only because of its direct impact, but also because of the influence it may have on other factors such as LD that determine imputation accuracy. Incorporating imputation as a genomic strategy will therefore require a complete understanding of the complex interplay between various population-specific parameters. Low-MAF SNPs have been observed to underestimate r-squared (r^2) based LD estimates (e.g. Khatkar *et al*., 2008; Qanbari *et al*., 2010), and LD has been shown to increase with increasing MAF for Sanga breeds (e.g. Makina *et al*., 2015a; Lashmar *et al*., 2018). In addition to parameters that characterize individual SNPs, it is important to look at the genomic relationship between SNPs.

Population-specific parameters: linkage disequilibrium and effective population size

Among the factors that have an influence of imputation accuracy, LD is probably the most important and has the potential to be limiting to achievable accuracy. The importance of LD, as a determining factor of imputation accuracy, has previously been shown where the influence of MAF was diminished in regions of high LD (Pei *et al*., 2008). The ability to impute a given genotype is affected directly by the strength of local LD in the genomic region in which that SNP is located (Hickey *et al*., 2012). Consensus has been that stronger LD improves imputation accuracy. Imputation algorithms are able to more accurately identify the haplotypes that are present on each gamete for individuals that are genotyped with a low-density panel, when LD is high (Hickey *et al*., 2012). Low inter-SNP LD is generally characteristic of populations with large effective population sizes (Bovine HapMap Consortium, 2009). This has been shown to impede accurate imputation (Pausch *et al*., 2013). If LD is weak and does not persist over long genomic segments, finding key ancestors that are representative of the breed becomes difficult. Weak persistence of phase and LD across breeds also limits the application of multi-breed imputation. For populations that display weak LD, however, algorithms have been proposed to simulate LD specifically for association studies (Yuan *et al*., 2011). This presents an opportunity to test imputation accuracy on different levels of LD.

Makina *et al*. (2015a) revealed LD of $r^2 \geq 0.2$ to extend to an inter-SNP distance of 100 kb in the Afrikaner breed, while the same level of LD extended only to a distance range of 10-20 kb for other Sanga breeds such as the Drakensberger and Nguni. This corresponds to the relative sizes of these breeds in the national beef herd of South Africa – the Afrikaner is numerically the smallest of the Sanga breeds. Given a standard r^2 value of 0.2, which has been proposed as the ideal r^2 for GS and association studies, the Drakensberger and Nguni breeds would require approximately 150 000 SNPs, as opposed to 30 000 SNPs for the Afrikaner, for within-breed analysis (Makina *et al*., 2015a). The utility of the 150K GGP bovine bead chip therefore needs consideration for the Drakensberger and Nguni breeds.

The Ne of a population gives an indication of the evolution of a breed and can assist in understanding the genetic architecture that underlies traits (Falconer & Mackay, 1996). This parameter essentially gives an indication of the number of animals within a breed that contribute to the genetic makeup of the national herd. The Ne is dependent on the interplay between LD and the recombination distance between SNPs, in which the LD across a greater distance will be indicative of more recent Ne and LD across a shorter distance will indicate Ne in the more distant past or ancestral Ne (Barbato *et al*., 2015). SNPs would be more accurately imputed for breeds with smaller Ne, which display higher within-population LD and therefore share larger haplotypes.

Makina *et al*. (2015a) observed Ne estimates of 41, 87 and 95 for Afrikaner, Drakensberger and Nguni breeds, respectively. These estimates were higher than those observed for exotic breeds such as Angus and Holstein. The discrepancy can be expected, because exotic beef breeds were generally subjected to intense artificial selection much earlier, and more consistently, than local commercial breeds. In comparison with dairy breeds, beef breeds are extensively managed, and breeding practices rely considerably less on reproductive technologies such as AI and MOET, and generally rely on natural mating. According to SA Stud Book's 2016 annual report, 31% of the SA Angus births resulted from AI, while only 0.5%, 8% and 1.6% of Afrikaner, Drakensberger and Nguni calves were born from this technology (SA Stud Book, 2016). The variation in Ne estimates between Sanga breeds can be explained by the higher level of admixture observed within Drakensberger and Nguni breeds in comparison with the Afrikaner breed (Makina *et al*., 2014). In admixed genomes, a higher number of small haplotypes will be shared, as opposed to a smaller number of long genomic segments. The Afrikaner breed has experienced a significant decline in its population size over the past decades. This is postulated to be a result of increased utilization of the breed to develop composites, causing a small number of 'pure' Afrikaner animals to remain (Pienaar *et al*., 2014). It is important to consider Ne within the context of the actual or census population size within the national herd. The Afrikaner breed went from being the most abundant indigenous breed in the 1970s (Pienaar *et al*., 2014) to the numerically smallest Sanga breed, consisting of only 42 herds and approximately 7 300 animals nationwide (SA Stud Book, 2016). The SA Bonsmara for example is currently the most numerous indigenous breed in South Africa with numbers of upwards of 120 000 animals. However, it has an estimated Ne of 77 (Makina *et al*., 2015a). On the contrary, the national Drakensberger herd is approximately a tenth of the size, with an estimated Ne of 87. In the national herd of South Africa, it might therefore in theory be easier to sample animals that contribute to the population-wide genomic profile of the Drakensberger breed, and thereby achieve higher imputation accuracies, if high-impact animals can be identified and records are complete and accurate.

The utility of imputed SNPs in improving genomic applications

Genome-wide association studies

Genome-wide association studies (GWAS) aim to locate QTL or genes responsible for traits of economic interest. These studies use association signals between phenotypic and genotypic (genome-wide SNPs) information to guide researchers towards the location(s) on the genome responsible for expressing traits of interest (Hayes & Goddard, 2010). These candidate regions can then be used as a reference to search for causative SNPs that explain a proportion of the variation that is seen in that trait (Hayes & Goddard, 2010). This methodology may have significant implications for cattle, because it can be used to better understand the genetic mechanisms underlying economically important traits, such as those involved in the adaptability of Sanga cattle. It can also be a diagnostic tool to identify SNPs that are associated with cattle disorders, which can then be used to select against animals that carry deleterious alleles. The utility of SNPs that are identified by GWAS, however, needs to be verified in an independent set of animals to determine their validity and reproducibility. After verification, these SNPs may then be included in commercial or population-specific and production-specific SNP panels. These SNPs, however, may still not be informative for the entire spectrum of cattle breeds if they are monomorphic and both alleles are not segregating within a population.

It has been proposed that most traits follow a trend of 'common disease-common SNP'. However, common SNPs have been shown to limit influence on complex diseases in humans (Pritchard & Cox, 2002). The association signals that are observed for common SNPs may be synthetic, meaning that these signals might not be influenced by common SNPs, but rather by rare SNPs that are in strong LD with common SNPs (Pritchard & Cox, 2002). Rare SNPs could be causative variants, but might not be – and in most cases are not – included on the bead chips that are commonly used to perform GWAS. The association of rare SNPs with common diseases or phenotypes is difficult to capture owing to poor statistical power, and generally requires genotyping of large numbers of animals. The alternative is to capture these SNPs directly by using higher density SNP panels or through whole-genome sequencing efforts. Re-sequencing of whole genomes, which is aimed at discovering novel SNPs, is not always feasible because of the requirement for large datasets of sequenced animals and the relative cost of sequencing per animal (Van Binsbergen *et al*., 2014). Methods such as genotyping by sequencing (GBS), which uses restriction enzymes to target specific segments of the genome, have been proposed to reduce sequencing costs and complexity. Genotyping by sequencing, however, has the limitation of producing high volumes of missing data, owing to the presence of variation in restriction sites, resulting from factors such as genetic divergence and low sequence coverage (Brouard *et al*., 2017). This presents the opportunity for the utilization of methods such as imputation to fill the missing data gaps.

The main purpose of imputation for GWAS is to boost the number of SNPs that can be tested for association and hence improve the power of the study (Marchini & Howie, 2010). Because animals can be genotyped on lower density panels, which can then be imputed to SNP densities that equated to a high-density panel or GBS and sequencing data, more animals can be affordably genotyped and included for analyses. Imputation has proved to increase power by up to 10% for GWAS, with rare SNPs being proposed to gain the most from this method (Marchini *et al*., 2007; Spencer *et al*., 2009; Marchini & Howie, 2010). Furthermore, after signals of association have identified certain genomic regions of interest, imputation can assist in fine mapping these regions, which will improve the chances of identifying the causal SNP or SNPs directly (Marchini & Howie, 2010). In the past, in many cases GWAS results were not replicable because the cost of SNP genotyping limited sample size, which limits comparability, and also owing to the availability of SNP data from different panels. Imputation, however, can be utilized to standardize the number of SNPs from different studies by imputing to a common set of SNPs to allow meta-analysis at each given SNP locus (Marchini & Howie, 2010). The combination of multiple datasets aims to reduce the number of false positive associations (Begum *et al*., 2012). Meta-analysis has been applied successfully, and has resulted in the identification of new loci of interest that had not been identified previously in individual studies (Marchini & Howie, 2010). These studies have been limited for cattle with only few meta-analyses performed for beef (e.g. Minozzi *et al*., 2012; Bolormaa *et al*., 2014). All of these studies, however, have identified novel genomic regions of interest when using merged data sets. The only way to truly capture novel regions of interest or novel SNPs, however, is through genome re-sequencing.

A number of studies have been published that investigated the utility of imputed sequence variants for cattle (e.g. Van Binsbergen *et al*., 2014; Frischknecht *et al*., 2017; Pausch *et al*., 2017; Bernardes *et al*., 2018). Imputation to sequencing data has been simplified by the initiation of the 1 000 Bull

Genomes Project in 2012, which provides a database of whole-genome sequence information that is made available to research groups that are interested in imputation towards GWAS and GS (Daetwyler *et al*., 2014). Sequencing variants have largely been imputed from Illumina's Bovine HD SNP panel. Imputation accuracies are compromised when imputing from lower density SNP panels such as the SNP50 panel, and hence a two-step procedure has been proposed, imputing first from SNP50 to HD, and subsequently from imputed HD to sequencing data (Bernardes *et al*., 2018). Imputation to GBS data, which achieved up to 94% accuracy estimates, has also been tested in for example Canadian dairy cattle (Brouard *et al*., 2017). Re-sequencing data that was generated for Sanga breeds from studies such as Zwane (2017), which aimed to identify novel SNPs, provides a valuable resource for future South African studies that aim to utilize imputed sequence variants in GWAS and GS experiments.

Cost-effective genomic selection

Genomic selection, a concept that was first proposed by Meuwissen *et al*. (2001), is a method that incorporates dense SNP genotypes to estimate GEBVs (Hayes *et al*., 2009). A reference population with known SNP information and adequate performance and pedigree data is used to compile a prediction equation for the estimation of GEBVs in selection candidates (Meuwissen *et al*., 2001). GS essentially captures all locus effects, regardless of size, that contribute to the genetic variation for a trait of interest by summing all estimated effects across the entire genome into a GEBV (Hayes *et al*., 2009). These GEBVs are then used to aid selection decisions. The various models whereby GEBVs are estimated have been discussed in detail in previous literature (e.g. Goddard & Hayes, 2007; Goddard *et al*., 2010; Van Marle-Köster *et al*., 2013).

The four main driving factors that influence GEBV accuracy for a population are i) the population-wide LD; ii) the availability and completeness of genotypic and phenotypic data for the reference population; iii) the heritability of the trait in question; and iv) the distribution of QTL effects (Hayes *et al*., 2009). The latter two factors are subject to the trait being studied. The decay of LD with increasing inter-marker distance has been reported widely, and specifically for non-exotic and admixed populations (e.g. Lashmar *et al*., 2018; Edea *et al*., 2015; Mokry *et al*., 2014). Owing to the costs of acquiring high-density SNP genotypes and whole-genome sequencing data, especially for researchers in developing countries, improving SNP densities through genotyping and sequencing more SNPs is not economically feasible. Genotype imputation, however, will aid in cost-effectively improving SNP densities by genotyping animals with low-density panels and imputing to higher-density information. This would enable more efficient use of funds and genotyping more animals. Cost efficient genotyping will aid in the procurement of the rule of thumb 1 000 animals required for accurate GS (Meuwissen *et al*., 2001) and thereby availing this technology for possibly all Sanga breeds. The availability of sufficient performance and pedigree records, however, might then become the limiting factor. For this reason, the implementation of GS was first focused on Sanga breeds with good histories of animal recording such as the SA Bonsmara.

Developing a working pipeline of cost-effective GS for other Sanga breeds will require a low-density panel, consisting of Sanga-informative SNPs, which is optimized for accurate imputation. SNPs to be included on such a low-density panel will need to be selected from a pool of SNPs that are already included on high-density panels or have been sequenced for Sanga breeds, based on certain marker characteristics. These characteristics include the genome distribution (that is, whether SNPs are evenly spaced across the genome), the MAF (that is, whether SNPs are segregating within the population) and the LD pattern between SNPs. Methods that combine these attributes, such as the Wellman SNP selection method (Wellman *et al*., 2013), and methods that incorporate machine-learning algorithms, such as feature similarity (Phuong *et al*., 2006), have been used to select informative SNPs for Irish cattle (Judge *et al*., 2016). Wu *et al*. (2016) also developed a multi-objective local optimization (MOLO) method for SNP selection, which uses a function that adjusts for gaps in the genomic data and incorporates Shannon entropy and other attributes, such as MAF and distribution, to select optimal SNPs. These methods need to be tested and validated for Sanga cattle to identify the optimal way of selecting Sanga-informative SNPs. Once the optimal SNP selection method and the optimal density (i.e. the minimum number of SNPs necessary from which to impute) have been identified, a low-density panel can be developed. This panel would serve as a backbone for using imputed SNPs in GS, which would allow the estimation of GEBVs at a reduced cost.

The utility of imputed genotypes for GEBV estimation has been studied for beef cattle (e.g. Berry & Kearney, 2011; Cleveland *et al*., 2011; Mulder *et al*., 2012). Mulder *et al*. (2012) confirmed the feasibility of direct genomic value (DGV) estimation using low density bead chips, provided that these bead chips included at least 3 000 SNPs. Cleveland *et al*. (2011) observed some loss in the accuracy of GEBVs using imputed SNPs, but still retrieved accuracy estimates that were higher than those acquired by traditional BLUP. Berry & Kearney (2011) observed a correlation of 97% between DGVs estimated from true genotypes versus imputed ones across a set of 15 functional traits. The correlations, however, depend on the availability of records for a specific trait and hence the reference population size that is used to estimate the SNP effects for that trait (Berry & Kearney, 2011). If more records are available, which is usually the case for easy-to-measure traits (e.g. weaning weight), then more animals can be included in the reference sample and the higher the DGV correlation from true genotypes versus imputed ones. For certain traits (e.g. direct and maternal calving difficulty), low DGV correlations are observed, regardless of relatively large reference population size. This phenomenon may be attributed to large QTLs being responsible for the expression of these traits (Berry & Kearney, 2011). Rutkoski *et al*. (2013) observed that the most accurate method of imputation was not necessarily always responsible for the most accurate GEBV estimation. This was ascribed to non-random imputations errors. These errors can be indicative of possible genetic relationships in the GS model if they are shared between related animals (Weigel *et al*., 2010; Rutkoski *et al*., 2013). Small numbers of wrongly imputed SNPs, however, are expected to have a negligible effect on GEBV accuracy since GS estimates all SNP effects simultaneously, as opposed to the SNP-by-SNPs approach that is followed in GWAS (Badke *et al*., 2013). Nevertheless, imputation as a methodology needs to be optimized to minimize, as far as possible, the effect of imputation variability on GS endeavours, especially for breeds with heterogeneous genomes such as those that belong to the Sanga subspecies.

Conclusion

The promotion of locally adapted Sanga breeds relies on the utilization of genomics in breed characterization and improvement. Imputation provides a cost-saving strategy for applying genomic methodologies such as GWAS and GS that will aid in breed improvement. Implementation of imputation as a routine genomic strategy, however, relies on its accuracy and hence reliability, which is influenced by many variables. The incorporation of imputation would therefore require optimization in Sanga breeds. Once a working pipeline has been set up for utilization, this methodology would hold many advantages for downstream genomic applications that aim to advance indigenous South African beef breeds.

Acknowledgements

The research project that encompasses this review is funded by the Red Meat Research & Development (RMRD-SA) and the Beef Genomics Project (BGP). The financial assistance of the NRF towards this research is also acknowledged. Opinions expressed and conclusions arrived at are those of the authors, and are not necessarily to be attributed to the NRF.

Author’s Contributions

SFL formulated and refined the review article as part of his PhD: Animal Science project. CV and FCM supervised SFL and were responsible for revising, editing and structuring the article.

Conflict of Interest Declaration

None of the authors has a conflict of interest to declare.

References

Abin, S.A., 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.

Anderson, C.A., Pettersson, F.H., Clarke, G.M., Cardon, L.R., Morris, A.P. & Zondervan, K.T., 2010. Data quality control in genetic case-control association studies. Nat. Protoc. 5, 1564-1573.

Antolín, R., Nettelblad, C., Gorjanc, G., Money, D. & Hickey, J.M., 2017. A hybrid method for the imputation of genomic data in livestock populations. Genet. Sel. Evol. 49, 30.

Badke, Y.M., Bates, R.O., Ernst, C.W., Schwab, C., Fix, J., Van Tassell, C.P. & Steibel, J.P., 2013. Methods of tagSNP selection and other variables affecting imputation accuracy in swine. BMC Genet. 14, 8.

Barbato, M., Orozco-terWengel, P., Tapio, M. & Bruford, M.W., 2015. SNeP: A tool to estimate trends in recent effective population size trajectories using genome-wide SNP data. Front. Genet 6, 109.

Becker, J., 2016. Beef genomics programme: sequencing for bovine superiority. Available online at URL: http://www.livestockgenomics.co.za.

Beefmaster Breeders' Society of SA & SA Stud Book, 2017. Joint media release: Genomic breeding values for South African Beefmaster cattle. Available online at: http://www.sastudbook.co.za/images/photos/News-Beefmaster-Genomic-EBVs.pdf

Begum, F., Ghosh, D., Tseng, G.C. & Feingold, E., 2012. Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Res. 40, 3777-3784.

Bernardes, P.A., Al-Mamun, H.A., Suarez, M., Lim, D., Park, B. & Gondro, C., 2018. Imputation accuracy of whole-genome sequence data in Hanwoo cattle. In: Proceedings of the 11th World Congress of Genetics Applied to Livestock Production. Auckland, New Zealand, 6-11 February 2018.

Berry, D.P. & Kearney, J.F., 2011. Imputation of genotypes from low-to high-density genotyping platforms and implications for genomic selection. Animal 5, 1162-1169.

Berry, D.P., McClure, M.C. & Mullen, M.P., 2014. Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. J. Anim. Breed. Genet. 131, 165-172.

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A. & Van Tassell, C.P., 2012. Design of a bovine low-density SNP array optimized for imputation. PloS one 7, e34130.

Boison, S.A., Santos, D.J.A., Utsunomiya, A.H.T., Carvalheiro, R., Neves, H.H.R., Perez O'Brien, A.M., Garcia, J.F., Sölkner, J. & da Silva, M.V.G.B., 2015. Strategies for single nucleotide polymorphism (SNP) genotyping to enhance genotype imputation in Gyr (Bos indicus) dairy cattle: Comparison of commercially available SNP chips. J. Dairy Sci. 98, 4969-4989.

Bolormaa, S., Pryce, J.E., Reverter, A., Zhang, Y., Barendse, W., Kemper, K., Tier, B., Savin, K., Hayes, B.J. & Goddard, M.E., 2014. A multi-trait, meta-analysis for detecting pleiotropic polymorphisms for stature, fatness and reproduction in beef cattle. PLoS Genet. 10, e1004198.

Bonsma, J.C., 1980. Cross-breeding, breed creation and the genesis of the Bonsmara. Livestock production. A Global Approach. Ed. J.C. Bonsma. Tafelberg, Cape Town, South Africa. pp. 126-136.

Bosman, L., Van Marle-Köster, E., Van der Westhuizen, R.R., Visser, C. & Berry, D.P., 2017. Population structure of the South African Bonsmara beef breed using high density single nucleotide polymorphism genotypes. Livest. Sci. 197, 102-105.

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

Brouard, J.S., Boyle, B., Ibeagha-Awemu, E.M. & Bissonnette, N., 2017. Low-depth genotyping-by-sequencing (GBS) in a bovine population: strategies to maximize the selection of high quality genotypes and the accuracy of imputation. BMC Genet. 18, 32.

Browning, S.R., 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. Hum. Genet. 124, 439-450.

Browning, S.R. & Browning, B.L., 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am. J. Hum. Genet. 81, 1084-1097.

Browning, S.R. & Browning, B.L., 2011. Haplotype phasing: Existing methods and new developments. Nat. Rev. Genet. 12, 703-714.

Calus, M.P.L., Bouwman, A.C., Hickey, J.M., Veerkamp, R.F. & Mulder, H.A., 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: A review of livestock applications. Animal 8, 1743-1753.

Cleveland, M.A., Hickey, J.M. & Kinghorn, B.P., 2011. Genotype imputation for the prediction of genomic breeding values in non-genotyped and low-density genotyped individuals. BMC Proc. 5, S6.

Collins-Lusweti, E., 2000. Performance of Nguni, Afrikander and Bonsmara cattle under drought conditions in the North West Province of southern Africa. S. Afr. J. Anim. Sci. 30, 33-33.

Daetwyler, H.D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brøndum, R.F., Liao, X., Djari, A., Rodriguez, S.C., Grohs, C., Esquerré, 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.

Directorate: Knowledge and Information Management, 2017. Abstract of agricultural statistics. Department of Agriculture Forestry and Fisheries, Pretoria, South Africa.

Druet, T. & Farnir, F.P., 2011. Modeling of identity-by-descent processes along a chromosome between haplotypes and their genotyped ancestors. Genet. 188, 409-419.

Druet, T. & Georges, M., 2010. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genet. 184, 789-798.

Druet, T., Schrooten, C. & De Roos, A.P.W., 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J. Dairy Sci. 93, 5443-5454.

Du Plessis, I., Hoffman, L.C. & Calitz, F.J., 2006. Influence of reproduction traits and pre-weaning growth rate on herd efficiency of different beef breed types in an arid sub-tropical environment. S. Afr. J. Anim. Sci. 36, 89-98.

Edea, Z., Dadi, H., Dessie, T., Lee, S-H. & Kim, K-S., 2015. Genome-wide linkage disequilibrium analysis of indigenous cattle breeds of Ethiopia and Korea using different SNP genotyping BeadChips. Genes Genom. 37, 759-765.

Ellinghaus, D., Schreiber, S., Franke, A. & Nothnagel, M., 2007. Current software for genotype imputation. Hum. Genom. 3, 371-380.

Elsik, C.G., Tellam, R.L. & Worley, K.C., 2009. The genome sequence of taurine cattle: A window to ruminant biology and evolution. Science 24, 522-528.

Falconer, D.S. & Mackay, T.F., 1996. Introduction to quantitative genetics. 4th edition. Longman, New York.

Felius, M., Beerling, M.L., Buchanan, D.S., Theunissen, B., Koolmees, P.A. & Lenstra, J.A., 2014. On the history of cattle genetic resources. Diversity 6, 705-750.

Ferraz, J.B.S., Wu, X., Li, H., Xu, J., Ferretti, R., Simpson, B., Walker, J., Silva, L.R., Garcia, J.F., Tait Jr., R.G. & Bauck, S., 2018. Design of a low-density SNP chip for Bos indicus: GGP indicus technical characterization and imputation accuracy to higher density SNP genotypes. In: Proceedings of the 11th World Congress of Genetics Applied to Livestock Production. Auckland, New Zealand, 6-11 February 2018.

Frischknecht, M., Pausch, H., Bapst, B, Signer-Hasler, H., Flury, C., Garrick, D., Stricker, C., Fries, R. & Gredler-Grandl, B., 2017. Highly accurate sequence imputation enables precise QTL mapping in Brown Swiss cattle. BMC Genomics 18, 999.

Fuchsberger, C., Abecasis, G.R. & Hinds, D.A., 2014. minimac2: faster genotype imputation. Bioinf. 31, 782-784.

García-Ruiz, A., Ruiz-Lopez, F.J., Wiggans, G.R., Van Tassell, C.P. & Montaldo, H.H., 2015. Effect of reference population size and available ancestor genotypes on imputation of Mexican Holstein genotypes. J. Dairy Sci. 98, 3478-3484.

Goddard, M.E. & Hayes, B.J., 2007. Genomic selection. J. Anim. Breed. Genet. 124, 323-330.

Goddard, M.E., Hayes, B.J. & Meuwissen, T.H., 2010. Genomic selection in livestock populations. Genet. Res. 92, 413-421.

Hayes, B. & Goddard, M., 2010. Genome-wide association and genomic selection in animal breeding. Genom. 53, 876-883.

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., Bowman, P.J., Daetwyler, H.D., Kijas, J.W. & Van der Werf, J.H.J., 2012. Accuracy of genotype imputation in sheep breeds. Anim. Genet. 43, 72-80.

Hickey, J.M., Kinghorn, B.P., Tier, B., Wilson, J.F., Dunstan, N. & Van der Werf, J.H., 2011. A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet. Sel. Evol. 43, 12.

Hickey, J.M., Crossa, J., Babu, R. & De los Campos, G., 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52, 654-663.

Howie, B.N., Donnelly, P. & Marchini, J., 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5, e1000529.

Howie, B., Marchini, J. & Stephens, M., 2011. Genotype imputation with thousands of genomes. G3: Genes Genom. Genet, 1, 457-470.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R., 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44, 955-959.

Hozé, C., Fouilloux, M.N., Venot, E., Guillaume, F., Dassonneville, R., Fritz, S., Ducrocq, V., Phocas, F., Boichard, D. & Croiseau, P., 2013. High-density marker imputation accuracy in sixteen French cattle breeds. Genet. Sel. Evol. 45, 33.

Huang, L., Wang, C. & Rosenberg, N.A., 2009. The relationship between imputation error and statistical power in genetic association studies in diverse populations. Am. J. Hum. Genet. 85, 692-698.

Illumina, 2006. Technical Note: ‘TOP/BOT’ Strand and ‘A/B’ Allele - A guide to Illumina’s method for determining Strand and Allele for the GoldenGate® and InfiniumTM Assays. Pub. No. 370-2006-018, Available online at URL: https://www.illumina.com/documents/products/technotes/technote_topbot.pdf.

Judge., M.M., Kearney, J.F., McClure, M.C., Sleator, R.D. & Berry, D.P., 2016. Evaluation of developed low-density panels for imputation to higher density in independent dairy and beef cattle populations. J. Anim. Sci. 94, 949-962.

Khatkar M., Nicholas F., Collins A., Zenger K., Cavanagh J., Barris W., Schnabel R., Taylor J. & Raadsma H., 2008. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genomics 9, 187.

Lashmar, S.F., Visser, C., Van Marle-Köster, E. & Muchadeyi, F.C., 2018. Genomic diversity and autozygosity within the SA Drakensberger beef cattle breed. Livest. Sci. 212, 111-119.

Li, Y., Willer, C., Sanna, S. & Abecasis, G., 2009. Genotype imputation. Ann. Rev. Genom. Hum. Genet. 10, 387-406.

Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R., 2010. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet. Epidemiol. 34, 816-834.

Lin, P., Hartz, S.M., Zhang, Z.H., Saconne, S.F., Wang, J., Tischfield, J.A., Edenberg, H.J., Kramer, J.R., Goate, A.M., Bierut, L.J. & Rice, J.P., 2010. A new statistic to evaluare imputation reliability. PloS one 5, e9697.

Ma, P., Brøndum, R.F., Zhang, Q., Lund, M.S. & Su, G., 2012. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish red cattle. J. Dairy Sci. 96, 4666-4677.

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.

Makina, S.O., Taylor, J.F., Van Marle-Köster, E., Muchadeyi, F.C., Makgahlela, M.L., MacNeil, M.D. & Maiwashe, A., 2015a. Extent of linkage disequilibrium and effective population size in four South African Sanga cattle breeds. Front. Genet. 6, 1-12.

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

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.

Mapholi, N.O., Maiwashe, A., Matika, O., Riggio, V., Bishop, S.C., MacNeil, M.D., Banga, C., Taylor, J.F. & Dzama, K., 2016. Genome-wide association study of tick resistance in South African Nguni cattle. Ticks Tick-Borne Dis. 7, 487-497.

Marchini, J. & Howie, B., 2010. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11, 499-511.

Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P., 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39, 906-913.

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, e5350-5063.

Maule, J.P., 1973. The role of the indigenous breeds for beef production in southern Africa. S. Afr. J. Anim. Sci. 3, 111-132.

Meissner, H.H., Scholtz, M.M. & Palmer, A.R., 2013. Sustainability of the South African livestock sector towards 2050 Part 1: Worth and impact of the sector. S. Afr. J. Anim. Sci. 43, 282-297.

Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genet. 157, 1819-1829.

Milanesi, M., Vicario, D., Stella, A., Valentini, A., Ajmone ‐ Marsan, P.A.O.L.O., Biffani, S., Biscarini, F., Jansen, G. & Nicolazzi, E.L., 2015. Imputation accuracy is robust to cattle reference genome updates. Anim. Genet. 46, 69-72.

Minozzi, G., Williams, J.L., Stella, A., Strozzi, F., Luini, M., Settles, M.L., Taylor, J.F., Whitlock, R.H., Zanella, R. & Neibergs, H.L., 2012. Meta-analysis of two genome-wide association studies of bovine paratuberculosis. PLoS one 7, e32578.

Mokry, F.B., Buzanskas, M.E., de Alvarenga Mudadu, M., do Amaral Grossi, D., Higa, R.H., Ventura, R.V., de Lima, A.O., Sargolzaei, M., Meirelles, S.L.C., Schenkel, F.S., da Silva, M.V.G.B., Niciura, S.C.M., de Alencar, M.M., Munari, D.P. & de Almeida Regitano, L.C., 2014. Linkage disequilibrium and haplotype block structure in a composite beef cattle breed. BMC Genomics 15, S6.

Mulder, H.A., Calus, M.P.L., Druet, T. & Schrooten, C., 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95, 876-889.

Nicolazzi, E.L., Biffani, S. & Jansen, G., 2013. Imputing genotypes using PedImpute fast algorithm combining pedigree and population information. J. Dairy Sci. 96, 2649-2653.

Nicolazzi, E.L., Biffani, S., Biscarini, F., Orozco ter Wengel, P., Caprera, A., Nazzicari, N. & Stella, A., 2015. Software solutions for the livestock genomics SNP array revolution. Anim. Genet. 46, 343-353.

O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I. & McQuillan, R., 2014. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet. 10, e1004234.

Ogawa, S., Matsuda, H., Taniguchi, Y., Watanabe, T., Takasuga, A., Sugimoto, Y. & Iwaisaki, H., 2016. Accuracy of imputation of single nucleotide polymorphism marker genotypes from low-density panels in Japanese Black cattle. Anim. Sci. J. 87, 3-12.

Pausch, H., Aigner, B., Emmerling, R., Edel, C., Götz, K.U. & Fries, R., 2013. Imputation of high-density genotypes in the Fleckvieh cattle population. Genet. Sel. Evol. 45, 3.

Pausch, H., MacLeod, I.M., Fries, R., Emmerling, R., Bowman, P.J., Daetwyler, H.D. & Goddard, M.E., 2017. Evaluation of the accuracy of imputed sequence variant genotypes and their utility for causal variant detection in cattle. Genet. Sel. Evol. 49, 24.

Payne, W.J.A. & Hodges, J., 1997. Tropical cattle: origins, breeds and breeding policies. Blackwell Science Ltd., Oxford, UK.

Pei, Y.F., Li, J., Zhang, L., Papasian, C.J. & Deng, H.W., 2008. Analyses and comparison of accuracy of different genotype imputation methods. PloS one 3, e3551.

Phuong, T.M., Lin, Z. & Altman, R.B., 2006. Choosing SNPs using feature selection. J. Bioinform. Comput. Biol. 4, 241-257.

Piccoli, M.L., Braccini, J., Cardoso, F.F., Sargolzaei, M., Larmer, S.G. & Schenkel, F.S., 2014. Accuracy of genome-wide imputation in Braford and Hereford beef cattle. BMC Genet. 15, 157.

Pienaar, L., Grobler, J.P., Neser, F.W.C., Scholtz, M.M., Swart, H., Ehlers, K. & Marx, M., 2014. Genetic diversity in selected stud and commercial herds of the Afrikaner cattle breed. S. Afr. J. Anim. Sci. 44, 80-84.

Porter, V., 1991. Cattle. A Handbook to the Breeds of the World. Christopher Helm, London. pp.145-146.

Pritchard, J.K. & Cox, N.J., 2002. The allelic architecture of human disease genes: common disease–common variant… or not?. Hum. Mol. Genet. 11, 2417-2423.

Purcell, S., Neale, B. & Todd-Brown, K., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559-575.

Purfield, D.C., McClure, M. & Berry, D.P., 2016. Justification for setting the individual animal genotype call rate threshold at eighty-five percent. J. Anim. Sci. 94, 4558-4569.

Qanbari, S., Pimentel, E.C., 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.

Qwabe, S.O., Van Marle-Köster, E., Maiwashe, A. & Muchadeyi, F.C., 2013. Evaluation of the BovineSNP50 genotyping array in four South African cattle populations. S. Afr. J. Anim. Sci. 43, 64-67.

Rabiner, L., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE 77, 257-286.

Ramnarine, S., Zhang, J., Chen, L.S., Culverhouse, R., Duan, W., Hancock, D.B., Hartz, S.M., Johnson, E.O., Olfson, E., Schwantes-An, T.H. & Saccone, N.L., 2015. When does choice of accuracy measure alter imputation accuracy assessments? PloS one 10, e0137601.

Rechav, Y. & Kostrzewski, M.W., 1991. The relative resistance of six cattle breeds to the tick Boophilus decoloratus in South Africa. Onderstepoort J. Vet. Res. 58, 181-186.

Rosen, B.D., Bickhart, D.M., Schnabel, R.D., Koren, S., Elsik, C.G., Zimin, A., Dreischer, C., Schultheiss, S., Hall, R., Schroeder, S.G., Van Tassell, C.P., Smith, T.P.L & Medrano, J.F., 2018. Modernizing the bovine reference genome assembly. In: Proceedings of the 11th World Congress of Genetics Applied to Livestock Production. Auckland, New Zealand, 6-11 February 2018.

Roshyara, N.R., Kirsten, H., Horn, K., Ahnert, P. & Scholz, M., 2014. Impact of pre-imputation SNP-filtering on genotype imputation results. BMC Genet. 15. 88.

Rutkoski, J.E., Poland, J., Jannink, J.-L. & Sorrells, M.E., 2013. Imputation of unordered markers and the impact on genomic selection accuracy. G3 Genes Genomes Genet. 3, 427-439.

SA Stud Book, 2016. SA Stud Book annual report. Available online at: http://www.sastudbook.co.za/images/photos/Annual_Report_2016_a.pdf.

SA Stud Book, 2017. Media Release: Genomic Breeding Values for Bonsmara. Available online at: http://www.sastudbook.co.za/n16/general-news/media-release:-genomic-breeding-values-for-bonsmara.html.

Sargolzaei, M., Chesnais, J.P. & Schenkel, F.S., 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15, 478.

Scheet, P. & Stephens, M., 2006. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. Am. J. Hum. Genet. 78, 629-644.

Scholtz, M.M., 1988. Selection possibilities of hardy beef breeds in Africa: The Nguni example. In: Proceedings of the 3rd World Congress on Sheep and Beef Cattle Breeds. Paris, France, 19-23 June 1988. pp. 303-319.

Scholtz, M.M., 2010. Beef breeding in South Africa. 2nd edition. Pretoria, South Africa.

Scholtz, M.M., McManus, G., Leeuw, K-C., Louvandini, H., Seixas, L., Demelo, C.B., Theunissen, A. & Neser, F.W.C., 2013. The effect of global warming on beef production in developing countries of the southern hemisphere. Nat. Sci. 5, 106-119.

Scholtz, M.M., Maiwashe, A., Neser, F.W.C., Theunissen, A., Olivier, W.J., Mokolobate, M.C. & Hendriks, J., 2014. Livestock breeding for sustainability to mitigate global warming, with the emphasis on developing countries. S. Afr. J. Anim. Sci. 43, 269-281.

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.

Schrooten, C., Dassonneville, R., Ducrocq, V., Brøndum, R.F., Lund, M.S., Chen, J., Liu, Z., González-Recio, O., Pena, J. & Druet, T., 2014. Error rate for imputation from the Illumina BovineSNP50 chip to the Illumina BovineHD chip. Genet. Sel. Evol. 46, 10.

South African Weather Service, 2016. South African Weather Service Annual Report 2016/2017. Available online at: https://nationalgovernment.co.za/entity_annual/1364/2017-south-african-weather-service-annual-report.pdf.

Spencer, C.C., Su, Z., Donnelly, P. & Marchini, J., 2009. Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 5, e1000477.

Statistics South Africa, 2018. Statistical release P0302: Mid-year population estimates 2018. Available online at: https://www.statssa.gov.za/publications/P0302/P03022018.pdf.

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.

Tsai, H.Y., Matika, O., Edwards, S.M., Antolín–Sánchez, R., Hamilton, A., Guy, D.R., Tinch, A.E., Gharbi, K., Stear, M.J., Taggart, J.B. & Bron, J.E., 2017. Genotype imputation to improve the cost-efficiency of genomic selection in farmed Atlantic salmon. G3: Genes Genom. Genet. 7, 1377-1383.

United Nations, 2017. World population prospects: The 2017 revision. Volume I: Comprehensive tables (ST/ESA/SER.A/399). Department of Economic and Social Affairs, Population Division. Available online at: https://population.un.org/wpp/Publications/Files/WPP2017_Volume-I_Comprehensive-Tables.pdf.

Van Binsbergen, R., Bink, M.C., Calus, M.P., Van Eeuwijk, F.A., Hayes, B.J., Hulsegge, I. & Veerkamp, R.F., 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet. Sel. Evol. 46, 41.

Van der Westhuizen, R.R., Van Marle-Köster, E., Theron, H.E. & Van der Westhuizen, J., 2014. Reference population for South African Bonsmara cattle. In: Proceedings of the 10th World Congress of Genetics Applied to Livestock Production. Vancouver, Canada, 17-22 August 2014. Poster presentation 498.

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 Marle-Köster, E. & Visser, C., 2018. Genetic improvement in South African livestock: Can genomics bridge the gap between the developed and developing sectors? Front. Genet. 9, 1-12.

Van Marle-Köster, E., Visser, C. & Berry, D.P., 2013. A review of genomic selection – Implications for the South African beef and dairy cattle industries. S. Afr. J. Anim. Sci. 43, 1-17.

Van Raden, P.M. & Sun, C., 2014. Fast imputation using medium- or low-coverage sequence data. In: Proceedings of the 10th World Congress of Genetics Applied to Livestock Production. Vancouver, Canada, 17-22 August 2014. Comm. 179.

Ventura, R.V., Lu, D., Schenkel, F.S., Wang, Z., Li, C. & Miller, S.P., 2014. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. J. Anim. Sci. 92, 1433-1444.

Ventura, R.V., Miller, S.P., Dodds, K.G., Auvray, B., Lee, M., Bixley, M., Shannon, M.C. & McEwan, J.C., 2016. Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. Genet. Sel. Evol. 48, 71.

Verma, S.S., De Andrade, M., Tromp, G., Kuivaniemi, H., Pugh, E., Namjou-Khales, B., Mukherjee, S., Jarvik, G.P., Kottyan, L.C., Burt, A. & Bradford, Y., 2014. Imputation and quality control steps for combining multiple genome-wide datasets. Front. Genet. 5, 370.

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.

Wang, Y., Lin, G., Li, C. & Stothard, P., 2016. Genotype imputation methods and their effects on genomic predictions in cattle. Springer Sci. Rev. 4, 79-98.

Weigel, K.A., Van Tassell, C.P., O'Connell, J.R., VanRaden, P.M. & Wiggans, G.R., 2010. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J. Dairy Sci. 93, 2229-2238.

Wellmann, R., Preuß, S., Tholen, E., Heinkel, J., Wimmers, K. & Bennewitz, J., 2013. Genomic selection using low density marker panels with application to a sire line in pigs. Genet. Sel. Evol. 45, 28.

Wiggans, G.R., Cooper, T.A., Van Tassell, C. P., Sonstegard, T.S. & Simpson, E.B., 2013. Technical note: Characteristics and use of the Illumina BovineLD and GeneSeek Genomic Profiler low-density bead chips for genomic evaluation. J. Dairy Sci. 96, 1258-1263.

Wu, X.L., Xu, J., Feng, G., Wiggans, G.R., Taylor, J.F., …, Bauck, S., 2016. Optimal design of low-density SNP arrays for genomic prediction: Algorithm and applications. PloS one 11, e0161719.

Yuan, X., Zhang, J. & Wang, Y., 2011. Simulating linkage disequilibrium structures in a human population for SNP association studies. Biochem. Genet. 49, 395-409.

Zhang, Z. & Druet, T., 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. J. Dairy Sci. 93, 5487-5494.

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.

Zwane, A.A., 2017. Genome-wide marker discovery in three South African indigenous cattle breeds (Afrikaner, Drakensberger and Nguni) using whole genome sequencing. PhD thesis, University of Pretoria, November 2017.

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.

CHAPTER THREE

Genomic diversity and autozygosity within the SA Drakensberger beef cattle breed

S.F. Lashmar[1,2#], C. Visser[1], E. van Marle-Köster[1] & F.C. Muchadeyi[2]

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

[2] Biotechnology Platform, Onderstepoort Veterinary Institute, Agricultural Research Council, P/Bag X05, Onderstepoort, Pretoria, 0110

Published in Elsevier's Livestock Science 2018, 212, pp. 111-119. DOI: https://doi.org/10.1016/j.livsci.2018.04.006

Abstract

The SA Drakensberger, as a Sanga beef breed, is a composite of *Bos taurus* and *Bos indicus* subspecies. Variation within admixed genomes will influence downstream applications such as imputation and genomic selection (GS). Being an indigenous breed with unique characteristics, such as the black coat, within-breed selection of the SA Drakensberger has focused on maintaining breed purity, which furthermore predisposes the breed to inbreeding. This study aimed to primarily identify possible patterns of variation in population-specific parameters such as minor allele frequency (MAF) and linkage disequilibrium (LD) that might influence the accuracy of future genomic applications. Second, the study investigated possible patterns of genomic uniformity using runs of homozygosity (ROH) as a measure of inbreeding. Average genome-wide MAF was 0.26 with chromosome-specific MAF ranging from 0.24 (*Bos Taurus* Autosome; BTA14) to 0.28 (BTA21). The proportion of low-MAF (<5%) SNPs supported average estimates, ranging from 6.6% for BTA23 to 16.0% for BTA14. The r^2 measure of LD was 0.14, 0.17 and 0.22, respectively, when SNPs separated by ≤1Mb, ≤0.1Mb and ≤0.05Mb were considered. LD was generally low, ranging from r^2=0.11 (BTA28) to r^2=0.17 (BTA14) for SNPs separated by ≤1Mb and r^2=0.20 extended only up to <30kb. LD was weaker between SNP pairs including low-MAF SNPs. The ROH identified were predominantly shorter in length, with more than 50% (54.5%) of ROH falling within the <4Mb length interval. Consensus ROH segments were identified and the most prevalent of these occurred on BTA14 and was identified in ~23% of the sampled population. All coefficients of inbreeding indicated low levels of inbreeding, which corresponded to 3% (F_{PED}), 1% (F_{SNP}) and 7% ($F_{\text{ROH>1Mb}}$). Correlations of F_{PED} with F_{SNP} and $F_{\text{ROH>1Mb}}$ were moderate equating to values of ~0.63 and ~0.64 (P<0.001), respectively. Such moderate correlations could be attributed to the incompleteness of pedigree records. The direct impact of MAF, LD and relatedness on the accuracy of within-breed genetic improvement strategies and its accompanying methodologies, such as imputation, may influence how different chromosomes are treated or accounted for in future genomic endeavors.

Keyword(s): Cattle, genome variation, runs of homozygosity (ROH), single-nucleotide polymorphism (SNP), within-breed diversity

1. Introduction

The genomics era has brought into existence many tools to aid in the genetic characterization and improvement of livestock species. High-throughput technologies such as next-generation sequencing (NGS) and SNP genotyping provide platforms for the identification and utilization of informative SNP markers in various methodologies such as genome-wide association studies (GWAS) and genomic selection (GS). For beef cattle, efforts to localize genomic regions of interest have been focused on economically important traits pertaining to growth, feed efficiency as well as carcass and meat quality (Sharmaa *et al*., 2015). Genomic evaluations in the form of genomic breeding values (GEBVs), are furthermore already being implemented, or at least researched, for beef cattle in the

Americas, Europe and Australasia (Berry *et al*., 2016). Genetic studies implementing these genomic techniques are, however, scarce in Africa. This is due in part to the complexity of beef industries in most African countries, such as South Africa (van Marle-Köster *et al*., 2013), as well as the myriad of indigenous, non-descript beef breeds of uncertain genomic architecture existing across the continent. A main limitation is, however, the inability to compete financially.

The economic status of a country is a main driver of its inclination to adopt genomic technologies (Dekkers & Hospital, 2002). The relative scale of genomics-based studies, with regards to the number of samples and markers required to draw meaningful conclusions, is a major attributor to the absence of studies of this caliber in the developing world. This highlights the need to identify efficient strategies to alleviate the financial burden involved in genotyping many animals for high densities of SNP markers. These strategies can include the establishment of large-scale collaborations, in the form of consortia, or the implementation of statistical methodologies such as genotype imputation. Imputation is primarily a statistical strategy to fulfill incomplete data or generate higher volumes of data by means of predictions that are based on probabilistic methodology (Marchini *et al*., 2007). This methodology presents an ideal opportunity for application in adapted and economically important local breeds. The accuracy of predictions that can be achieved by this method, however, is influenced by genomic characteristics that are not fully understood for African breeds or breed-groups such as the Sanga sub-species.

Sanga cattle (*Bos taurus africanus*) are indigenous to southern Africa (Schoeman, 1989) and are presumably taurine-indicine hybrids (Grigson, 1991). This was confirmed by initial genome-wide SNP data suggesting that the genome of Sanga cattle has selection signatures of both subspecies (Makina *et al*., 2016). Population genetic analyses indicate that whereas a breed such as the Afrikaner shows strong divergence from its ancestors, Bonsmara and SA Drakensberger genomes are more complex. The latter breeds have an admixture of European- and African taurine as well as indicine footprints (Makina *et al*., 2016). It is common knowledge that the Bonsmara is a composite breed that resulted from experimental crosses (Bonsma, 1980); however, admixture in the SA Drakensberger genome was probably introduced unintentionally (Makina *et al*., 2016).

The modern SA Drakensberger is a medium-framed beef cattle breed with a sleek, black coat (Rege & Tawah, 1999). One of the most important qualities of this breed is its ability to adapt and perfom consistantly under even low quality grazing conditions (Bisschoff & Lotriet, 2013). The genomic makeup of this breed is largely unknown with regards to proportions of the genome descended from taurine- or indicine ancestors and proportions that are entirely unique. Bolormaa *et al*. (2011) showed that the origin of chromosomal segments, whether from taurine or indicine descent, has varying effects on traits of economic importance. Genomic heterogeneity will therefore impact on downstream applications for composite breeds when commercial SNP chips are used, as genomic segments with divergent ancestral origins will have to be treated differently to ensure reliability of these applications.

Identifying inter-chromosomal differences in SNPs, whether markers are segregating within a population or whether there are significant correlations between markers, might aid in how individual chromosomes are dealt with in downstream processes. These aspects would be reflected by differences in population-specific parameters such as MAF and inter-SNP LD. SNPs with low MAF, for example, have been shown to negatively influence imputation accuracy, as alleles occurring in low frequencies are not necessarily presented in the haplotypes identified (Schrooten *et al*., 2014). Ascertainment bias disadvantages indigenous, non-discovery populations and currently available genotyping platforms, such as the Bovine SNP50 bead chip, are expected to include SNPs displaying low MAF in these populations. Cattle populations in Africa furthermore seem to follow a trend of high haplotype diversity and low LD (Mwai *et al*., 2015). This needs to be taken into consideration given a general consensus that stronger LD improves imputation accuracy (e.g. Pei *et al*., 2008; Hickey *et al*., 2012).

Genomic uniformity becomes detectable in the form of long stretches of consecutively homozygous marker genotypes referred to as runs of homozygosity (ROH) (Falconer & Mackay, 1996; Purfield *et al*., 2012). These regions can be the result of consanguineous mating in the distant or recent past and, alternatively, long-term artificial selection. The number of breeding animals within the SA Drakensberger breed is gradually declining (Abin *et al*., 2016), with increasing utilization of sires across herds. Based on available pedigree information, the level of inbreeding has steadily increased since the year 2000 albeit lower than the previously suggested 0.5-1% cut-off per generation (Abin *et al*., 2016). It is therefore expected that their genomes harbour some signatures of inbreeding, which if prevalent will also impact on imputation and other applications. Imputation accuracy will increase with closer within-breed genetic relationships thereby improving the chances of encountering shared haplotypes from common ancestors (Pei *et al*., 2008). Coupled to this and given the prioritization of conserving local animal genetic resources (AnGR), knowledge of inbreeding can further aid in effectively managing breeding practices to prevent inbreeding depression.

This study aimed to investigate diversity within the SA Drakensberger genome. The main objective was to quantify inter-chromosomal variation using genetic parameters such as MAF and inter-SNP LD. Knowledge of these metrics may help underpin appropriate imputation strategies in the future. The study furthermore investigated chromosomes harbouring uniformly homozygous regions and used ROH to determine the level of genomic inbreeding within the SA Drakensberger breed. Inferences were made on the implications of these breed parameters on genetic diversity and possible downstream genomic strategies.

2. Materials and methods

2.1 Animals sampled

A total of 620 SA Drakensberger cattle were sampled, consisting of 184 bulls and 436 cows. Samples formed part of a cohort that was specifically selected to include half-sib families and family trios for imputation. The sample constituted animals ranging in birth date from 1982 to 2016 and originated from approximately 40 breeders. All breeders, and therefore animals, included are registered with South African Stud Book (SA Stud Book). Ethical clearance was obtained from the University of Pretoria's Faculty of Natural and Agricultural Science (EC151106-024) and written consent was given by the Drakensberger Breeders' Society to perform this study. Pedigree data up to five generations deep is currently ~65% and ~30% complete, respectively, for animals born within- and longer than 25 years ago (Abin, 2014).

2.2 Genotyping and quality control

A cohort of 414 SA Drakensberger samples was genotyped at the Agricultural Research Council's Biotechnology Platform (ARC-BTP) as part of the Beef Genomics Project (BGP) currently underway in SA. Hair and semen samples were received from individual breeders. DNA was isolated and SNP genotypes generated with the GeneSeek® Genomic Profiler (GGP) 150K-bead chip according to the standard infinium protocol. Hair samples for the remaining animals (206 animals) were sent to GeneSeek® (Lincoln, USA) where DNA isolation and genotyping were performed according to standard protocols. Samples and markers were subjected to standard quality control (QC) procedures using PLINK (Purcell *et al*., 2007). Individuals with high proportions of uncalled genotypes (call rate<90%) were filtered from analysis. Only autosomal, mapped (UMD 3.1 bovine reference genome) markers were considered. Of these markers, SNPs that displayed low call rates (≤95%) and low MAF (<1%) were removed from further analysis. SNPs with duplicated genomic positions were also excluded. After these QC procedures, 606 animals and 120 218 SNPs remained and were used for downstream analyses. To minimize the effect of shared genetic structure on certain analyses, such as principal component analysis (PCA) and ROH estimation, SNPs were additionally filtered based on high LD, removing SNP in LD exceeding r^2=0.5 with another SNP. The LD-pruned (LDP) data set consisted of 76 581 SNPs.

2.3 Genetic relatedness between animals

Relatedness between sampled animals was assessed with PCA as implemented in GCTA software (Genome-wide Complex Trait Analysis; Yang *et al*., 2011). A genomic relationship matrix (GRM) was constructed based on LDP data and used to estimate eigenvectors and -values.

2.4 Inter-chromosomal variation in SNP parameters

The MAF per SNP was estimated using PLINK software and incorporated all autosomal SNPs with call rates above 95%. No other QC procedures, except for filtering on call rates, were performed before MAF estimation. Summary statistics per chromosome were calculated using R (R Development Core Team, 2013).

The r^2 measure, as proposed by Hill and Robertson (1968), was used to quantify LD and was calculated using the following formula implemented in PLINK software;

$$r^2(p_a, p_b, p_{ab}) = \frac{(p_{ab} - p_a p_b)^2}{p_a(1 - p_a)p_b(1 - p_b)}, \quad [1]$$

where p_{ab} represented the frequency of the haplotype consisting of 2 SNPs; p_a and p_b represented the frequency of allele a at the first locus and allele b at the second locus, respectively. In PLINK software, no restrictions were set for the minimum r^2 (*--ld-window-r2* 0) and inter-SNP distance (*--ld-window-kb* 99999) allowed for LD estimation and this enabled all possible pairwise comparisons. Mean r^2 values were then calculated for SNP pairs separated by ≤0.05Mb, ≤0.1Mb and ≤1Mb. Averages were also calculated for bins (10kb bins, 0-100kb; 100kb bins, 100kb-1Mb; 1Mb bins, 1-4Mb) to observe LD decay. The effect of MAF on LD was furthermore investigated by estimating the difference in the extent of LD decay when MAF-filtering thresholds (<1%, <5%, <10% and <20%) were adjusted. Post-PLINK calculations were performed using custom scripts.

2.5 Runs of homozygosity

Contiguous homozygous segments were called using PLINK's sliding window approach. This analysis was performed on the full data set of 606 animals after sample filtering procedures, considering independent, LD-filtered SNPs. Segments were called as ROH if: 1) it was a minimum of 1Mb in length; 2) it included no more than one heterozygous SNP, but included up to two missing SNPs; 3) had a minimum SNP density of one SNP every 75kb; and 4) the maximum gap between consecutive SNPs was no longer than 1Mb. The thresholds for these parameters were based on PLINK defaults and consensus with previous ROH studies on cattle in order to allow comparison. The minimum number of consecutive homozygous SNPs that constituted a ROH segment was calculated using the following formula as implemented by Purfield *et al.* (2012),

$$l = \frac{log_e \frac{\alpha}{n_s . n_i}}{log_e(1 - \overline{het})} \quad [2]$$

where n_s and n_i were the number of SNPs and individuals, respectively, α represented the proportion of false-positive identifications (set to 0.05) and $\overline{het}$ was the average SNP heterozygosity. Using the formula the minimum number of SNPs constituting a ROH was set to 50.

2.5 Inbreeding coefficients

Three methods were utilized to estimate inbreeding: 1) F_{PED} represented a pedigree-derived estimate, 2) F_{SNP} represented a SNP-by-SNP excess in homozygosity and 3) F_{ROH} represented genome-wide ROH coverage. F_{PED} per animal was estimated by SA Stud Book, which the SA Drakensberger is registered with, and forms part of standard breed evaluations (SA Stud Book). All available pedigree information was utilized. F_{SNP}, and heterozygosity, was calculated on LDP data in PLINK. The ROH-based inbreeding coefficient (F_{ROH}) was estimated per individual as follows,

$$F_{ROH} = \frac{S_{ROH}}{L_{GEN}} \tag{3}$$

where S_{ROH} was the summed length of ROH segments for an individual and L_{GEN} represented the total length of the autosomal genome covered by the SNPs on the specific bead chip. Box plots were generated for each inbreeding coefficient using R. Pearson correlations between coefficients were also calculated using R.

3. Results

3.1 Genomic relationships between individuals within the breed

Within breed genomic relationships were estimated from a set of LDP SNPs and the resulting eigenvectors (EVs) did not indicate separation into different clusters, but rather one cluster with dispersion. Correlations between the a) first- and second as well as b) first- and third PCs are illustrated in Figure 3.1.

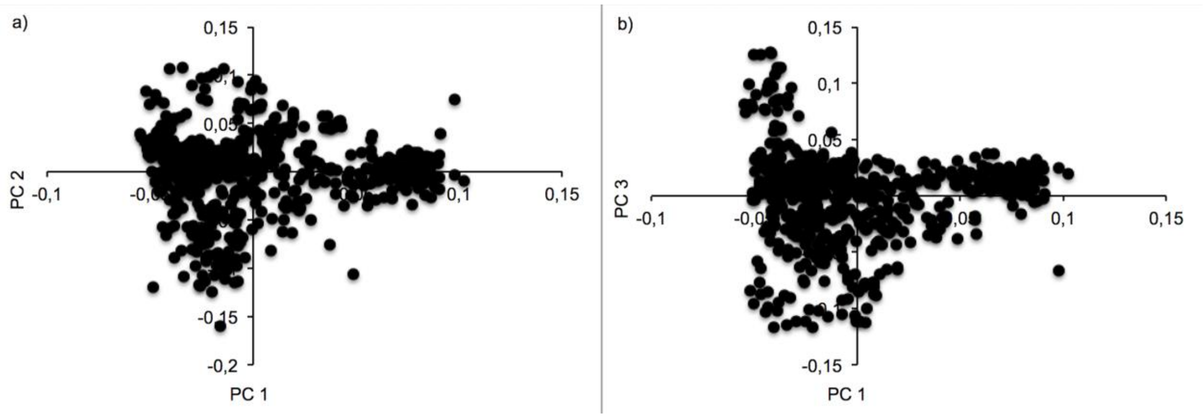


Figure 3.1 Principal component analysis (PCA) of genetic relatedness between SA Drakensberger animals sampled.

PC1 (EV range: -0.055-0.103), PC2 (EV range: -0.160-0.108) and PC3 (EV range: -0.117-0.128) accounted for 46.3%, 28.0% and 25.7% of the variation estimated for the first three principal components. No outliers were identified for EVs estimated for the first PC, however, 59 and 49 outliers were identified based on EVs estimated for PC2 and PC3, respectively.

3.2 Population-specific SNP parameters

The mean MAF across all autosomes was 0.26±0.14 (median: 0.27) with BTA14 and BTA21 having the lowest (0.24) and highest (0.28) mean MAF, respectively. The highest percentage of SNP displaying low MAF (less than 5%) was observed for BTA14 (16.0%), while BTA23 had the smallest percentage of low-MAF SNPs (6.6%). This was consistent with the percentage monomorphic SNPs observed (BTA14=1.3%; BTA23=0.3%). Across all autosomal markers with sufficient call rates (123 505 SNPs) there were only 0.6%, 2.6% and 9.3% SNPs with MAF of 0%, <1% and <5%, respectively. MAF-related trends are illustrated in Figure 3.2.

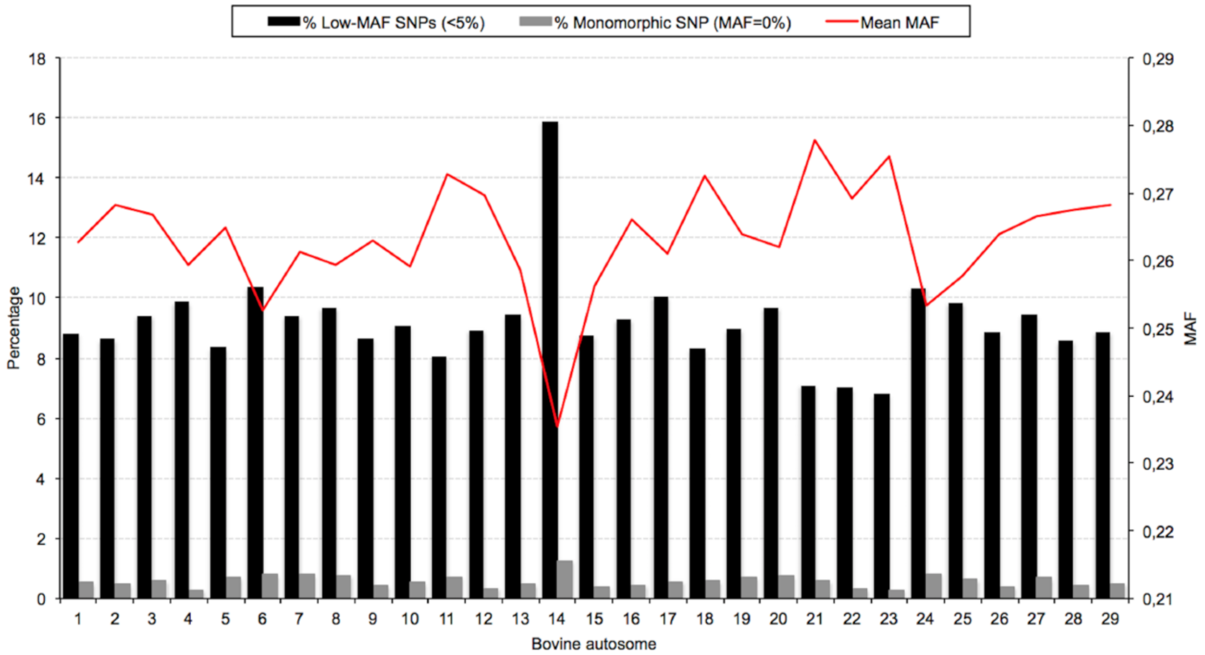


Figure 3.2 Variation in SNP minor allele frequency (MAF) between autosomal chromosomes.

After QC, the mean genome-wide SNP density was 1 SNP/20.9kb and ranged, per chromosome, from 1 SNP/17.9kb (BTA14) to 1 SNP/22.1kb (BTA8). Five autosomes harboured outlying high SNP densities namely BTA6, 7, 14, 20 and 24. Considering SNP pairs separated by ≤1Mb, r^2 ranged from 0.11 (BTA28) to 0.17 (BTA14) with BTA14 identified as displaying outlying high LD. Mean r^2 for shorter inter-SNP distances of up to 100kb and 50kb, respectively, ranged from 0.14 (BTA28) to 0.22 (BTA14) and 0.17 (BTA28) to 0.28 (BTA14). Inter-chromosomal SNP density and LD statistics are depicted in Figure 3.3.

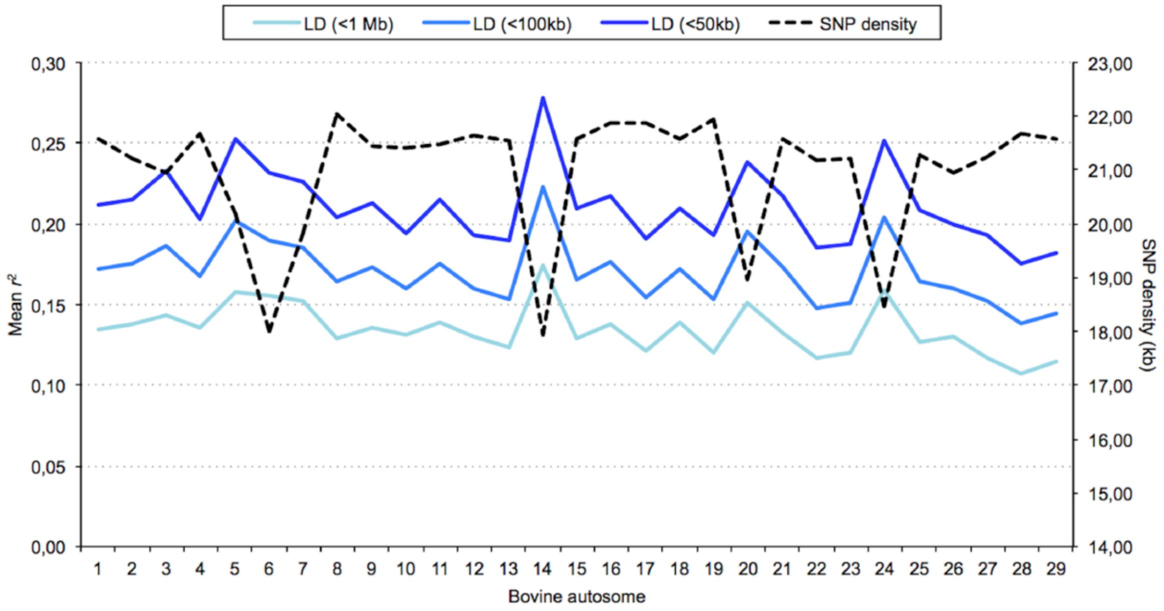


Figure 3.3 Variation in SNP density- and LD between autosomal chromosomes.

On average, the proportion of SNP pairs showing LD of $r^2 \geq 0.2$ increased with 7.1% when only high-MAF SNP (>20%) were included as opposed to lower-MAF SNPs (>1%). Including SNPs with MAF>1%, high LD persisted between SNP pairs <30kb apart. Estimated r^2 of 0.32, 0.24 and 0.21 were observed for SNPs separated by 0-10kb, 10-20kb and 20-30kb, respectively. When only high-MAF (>20%) SNPs were included, LD extended up to approximately double the distance (<60kb). LD decay is illustrated in Figure 3.4.

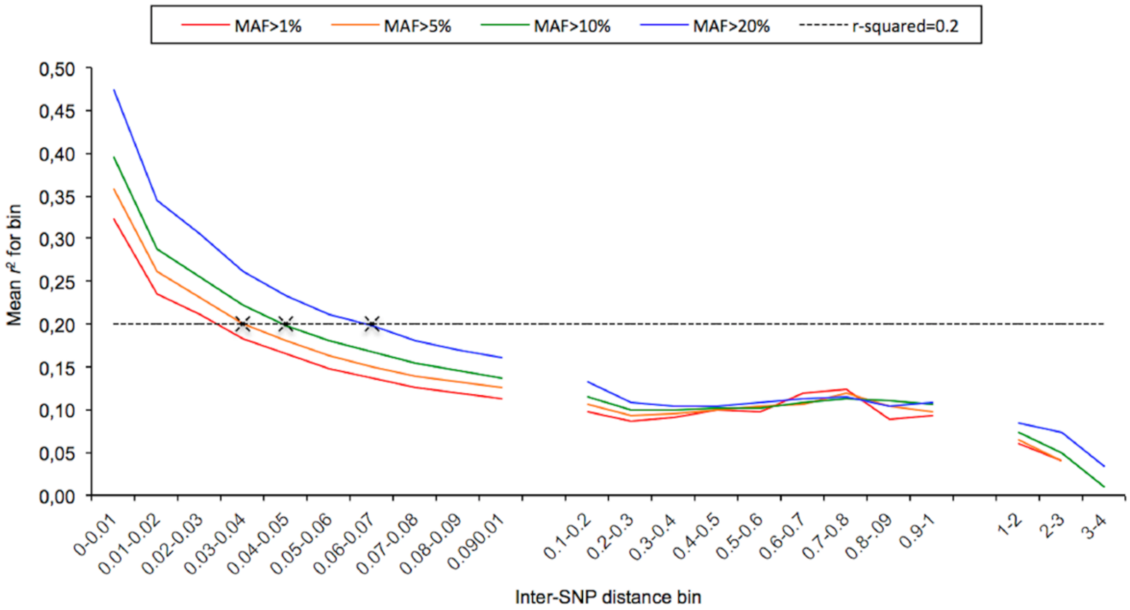


Figure 3.4 The decay of LD with increasing inter-SNP distances.

3.3 ROH analysis

The mean number of ROH (n_{ROH}) per animal was ~33 segments (min=0; max=152). The majority of these segments were between 2-4Mb in length; approximately 18.8%, 35.7%, 25.6%, 14.6% and 5.2% of the segments belonged to the 0-2Mb, 2-4Mb, 4-8Mb, 8-16Mb and >16Mb length categories. Long ROH (>16Mb) were only identified in ~62% of the population and these animals on average had only ~3 of these long segments. On average, the segments identified were composed of 171 SNPs (min=50 SNPs; max=2735 SNPs); the mean distance between homozygous SNPs was 32.78kb (min=11.22kb; max=73kb). Specific clusters of consecutively homozygous SNPs were conserved within varying proportions of the sampled population, and this is illustrated in Figure 3.5.

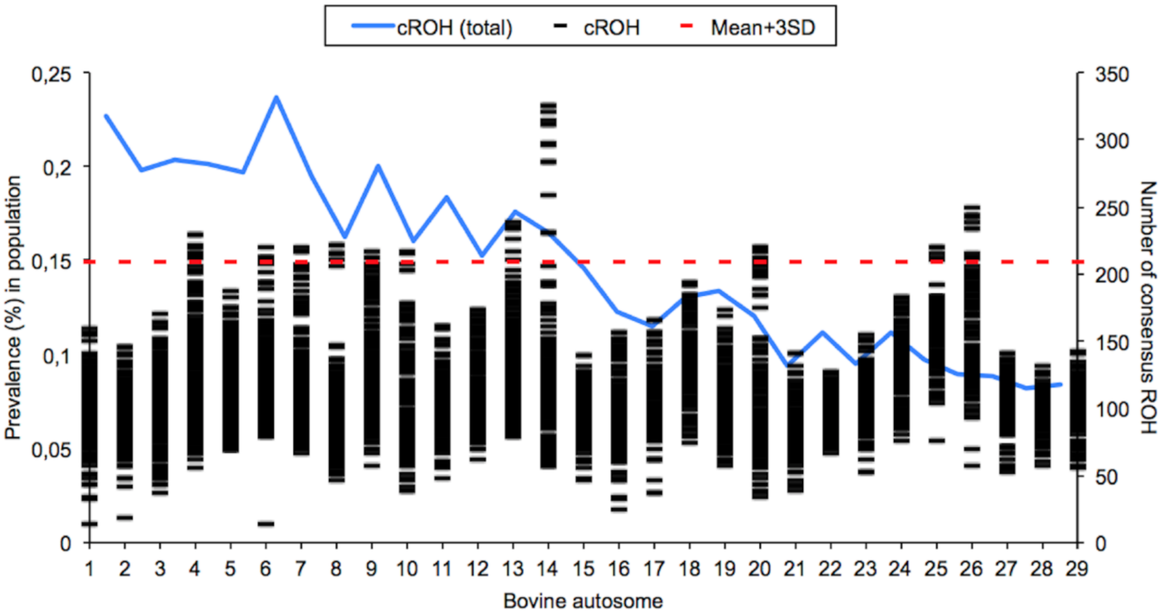


Figure 3.5 Consensus runs of homozygosity (cROH) within the Drakensberger population.

Consensus ROH were on average 86.86kb in length and were composed of ~4 SNPs. BTA6 and BTA28 harbored the highest (331) and lowest (115) number of consensus ROH segments. The largest consensus ROH segment was located on BTA15 and was 1723.83kb in length, consisting of 32 consecutively homozygous SNPs. The most prevalent consensus ROH segment was found in 141 (23.3%) of the sampled animals; this segment was located on BTA14 and constituted 5 SNPs stretching over 225.82kb. Consensus segments occurring in >100 of the sampled animals were also observed on BTA13 and BTA26.

3.4 Inbreeding coefficients

All measures of inbreeding indicated positive inbreeding at a low level. Box plots for each measure are illustrated in Figure 3.6. Only animals with non-zero pedigree-based inbreeding (586 animals) were considered. The mean F_{PED}, F_{SNP} and $F_{ROH>1Mb}$ were calculated as 0.03, 0.01 and 0.07, respectively. Due to the nature of the F_{SNP} measure, the SNP-by-SNP coefficient showed the most variation. F_{SNP} did not change significantly when estimated before- or after LDP and was in agreement with genome-wide heterozygosity estimates that showed a slight loss in genetic diversity (Before LDP: $H_O=0.351<H_E=0.354$; after LDP: $H_O=0.344<H_E=0.347$). Mean F_{ROH} decreased with increasing ROH-length intervals, estimated as 0.07, 0.06 and 0.05 when only ROH>4Mb, >8Mb and >16Mb were included in calculations.

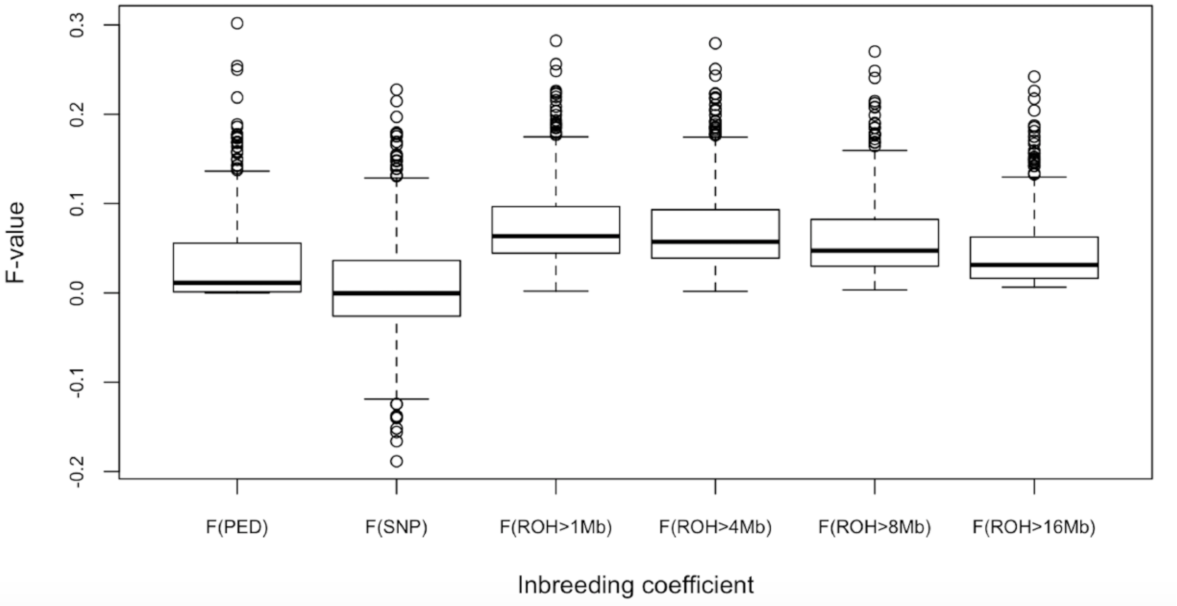


Figure 3.6 Box plots of pedigree-, SNP- and ROH-based inbreeding coefficients.

Pearson correlations between F_{PED} and both F_{SNP} and F_{ROH}, respectively, were moderate. Correlations with F_{ROH} were estimated for different length classes and are indicated in Table 3.1. The highest correlation with F_{PED} was observed when all ROH>8Mb were included in the calculation. Given the fact that ROH were estimated based on SNP data, high correlation between F_{SNP} and F_{ROH} were expected.

Table 3.1 Correlations between pedigree- and molecular-based inbreeding coefficients.

Inbreeding coefficient	**Correlation**
***r*(*F*$_{\mathbf{PED}}$)**	
F_{SNP}	0.633***
$F_{\mathrm{ROH>1Mb}}$	0.642***
$F_{\mathrm{ROH>4Mb}}$	0.639***
$F_{\mathrm{ROH>8Mb}}$	0.655***
$F_{\mathrm{ROH>16Mb}}$	0.523***
***r*(*F*$_{\mathbf{SNP}}$)**	
$F_{\mathrm{ROH>1Mb}}$	0.954***

The F_{SNP} per birth year across the population sampled is indicated in Addendum 1.

4. Discussion

The utility of a specific SNP genotyping platform is influenced by its development and the application of such platforms can adversely impact on population-specific SNP parameters estimated for breeds that were not represented in its design. Breeds with unclear- but presumably diverse ancestry may therefore display variation in these parameters due to the origin of the SNP investigated. The development of the Illumina SNP50 bovine bead chip was based exclusively on the selection of SNPs occurring in taurine beef- and dairy cattle genomes (Matukumalli *et al*., 2009). Conversely, the GGP 80K bovine bead chip was developed to incorporate more SNP of indicine descent (Edea *et al*., 2015). Even though no details regarding the exact composition of the 150K chip have been published, it is believed that SNP selection was also biased towards taurine cattle albeit that indicine cattle were also included (personal communication).

SNP-based genetic studies have observed lower MAF estimates for African breeds when only taurine- as opposed to indicine-derived SNPs were studied. Edea *et al*. (2015), for example, observed significantly lower MAF in Ethiopian cattle using SNP50-genotypes (0.15) as opposed to SNP80-genotypes (0.32). Estimates were similarly low for South African Sanga breeds based on SNP50 genotypes (Nguni: 0.17; Qwabe *et al*., 2013). Given that indigenous Sanga breeds are believed to be hybrids (Grigson, 1991) that harbour signatures of both sub-species, albeit in unknown proportions (Makina *et al*., 2015b), average MAF was expected to be higher when indicine-derived SNP were also included. The average MAF obtained in this study (0.26±0.143) is comparable with estimates obtained by Zwane *et al*. (2016) for the SA Drakensberger breed (0.26 ± 0.145 for MAF≥0%). The latter authors used a set of SNPs that were common between the Illumina SNP50 and GGP 80K chips. Given, firstly, the improvement in SNP density and, secondly, the inclusion of indicine SNPs, ascertainment bias seemed to influence the utility of the SNP150 chip to a lesser extent than the SNP50 chip for a non-discovery, composite breed such as the SA Drakensberger.

Identifying inter-chromosomal variation in MAF is important firstly for its direct impact on downstream analyses and secondly for its influence on local LD. Low-MAF SNPs display lower imputation accuracy as alleles occurring in low frequencies are not represented in reference haplotypes (Schrooten *et al*., 2014). These SNPs furthermore negatively impact on the accuracy of GEBVs (Weng *et al*., 2012). In this study, autosomes with up to ~16% (BTA14) low-MAF SNPs were identified. This will need to be accounted for when imputation, or eventually genomic selection, is implemented especially considering the fact that most imputation methods apply algorithms on a chromosome-by-chromosome basis. For non-discovery breeds it might necessitate the identification of specific markers segregating within these populations by identifying evenly distributed, high-MAF SNPs, and developing breed-specific low-density SNP panels. The relationship between MAF and LD was further investigated and, in agreement with previous research on Sanga- (Makina *et al*., 2015a) and international breeds (eg. Khatkar *et al*., 2008; Qanbari *et al*., 2010), showed low MAF to locally diminish LD. This relationship will be useful for populations, such as the SA Drakensberger, where LD does not persist over long genomic distances (<30kb).

LD is an important population-specific parameter in genomic studies. It serves as a predictor of the density of SNPs required to produce accurate GEBVs and powerful associations in GWAS (Qanbari *et al*., 2010). The strength of local LD furthermore influences the achievable imputation accuracy of specific genomic regions (Hickey *et al*., 2012). Many studies have investigated inter-chromosomal differences in LD in cattle populations (Sargolzaei *et al*., 2008; Bohmanova *et al*., 2010; Qanbari *et al*., 2010; Cañas-Álvarez *et al*., 2014; Edea *et al*., 2015) and the general consensus has been that there is a chromosome-effect influencing LD. In concordance with previous research, the density of SNPs per autosome seemed to be a primary contributor to high local LD. BTA14, which showed outlying high LD, was the most densely covered autosome. High LD on this autosome might therefore be an artifact of close inter-SNP distance and not necessarily a true reflection of strong relationships between SNPs overall. Developing a breed-specific low-density panel that is optimized for imputation would therefore necessitate the selection of high-LD SNP pairs per autosome, while assuring even distribution of SNPs across autosomes. Investigating inter-chromosomal differences in LD is important as high-LD autosomes are expected to produce higher imputation and GEBV accuracies. LD for the sampled population was relatively high when considering only SNP pairs separated by ≤50kb (mean r^2=0.22). At short inter-SNP distances (10-20kb), estimates were comparable with composite Brazilian beef breeds (0.24 versus 0.25; Mokry *et al*., 2014). Furthermore, short distance (~10kb) estimates were on par with what was found for indicine breeds like Brahman (0.25) and Nellore (0.27), but lower than values obtained for taurine breeds (eg. Angus, 0.46; Porto-Neto *et al*., 2014).

LD of r^2=0.2 was found to persist only for short inter-marker distances (<30kb), albeit higher than previously suggested by Makina *et al*. (2015a) for the SA Drakensberger breed (10-20kb). This was expected considering the improvement in SNP density of the SNP150- compared to the SNP50 bead chip used by Makina *et al*. (2015a). In the afore-mentioned study, LD was also estimated in a significantly smaller sample size of SA Drakensberger animals (±40 versus 606). Results were in agreement with Edea *et al*. (2014) who showed LD decay after 20-40kb in indigenous Ethiopian cattle. Rapid LD decay in the SA Drakensberger was furthermore not surprising, as previous studies have suggested shorter persistence of high LD in admixed populations (Toosi *et al*., 2010; Mokry *et al*., 2014). This phenomenon is attributed to more distant common ancestry and therefore the sharing of short haplotype structures within these populations (Mokry *et al*., 2014). Makina *et al*. (2015a) observed the smallest average haplo-block length for the SA Drakensberger breed compared to other Sanga breeds.

High LD reflects genomic regions lacking recombination. This absence of recombination is a key factor in the existence of ROH. The length of a ROH segment is arguably its most important characteristic as this can be used to infer population history (McQuillan *et al*., 2008). Short ROH were significantly more frequent ($ROH_{2\text{-}4Mb}$=35.7%) than longer ROH ($ROH_{>16Mb}$=5.2%) in the SA Drakensberger population sampled. ROH of 16Mb in length has been proposed to represent recent inbreeding of up to ~6 generations ago, whereas ROH of 1Mb corresponds to more ancient inbreeding of up to ~50 generations ago (Ferenčaković, 2015). Given a generation interval (*L*) of ~6 years for the SA Drakensberger breed (Abin, 2014), this represents inbreeding of up to 36 and 300 years ago, respectively. The ROH identified in the SA Drakensberger breed therefore implies that inbreeding is predominantly the result of more ancient consanguinity. Hybridization between taurine and indicine cattle is expected to increase diversity within the genomes and may have lead to the interruption of homozygous stretches in African cattle (Purfield *et al*., 2012). The sharing of homozygous segments between up to 23% of the sampled population may also point towards selection-driven fixation of some segments. It would therefore be beneficial to explore ROH as a trait-association tool, using appropriate phenotypic information, in the future.

Inbreeding has traditionally been quantified by tracing back consanguineous mating through the use of pedigree information. The reliability of this method is, however, dependent on the degree of pedigree recording and hence the completeness of records (McParland *et al*., 2007). Considering that the birth dates of the animals included in this study ranged from 1982 to 2016 and the fact that not all animals had equal depths of pedigree data available, F_{PED} was not the most robust method of inbreeding estimation. This was confirmed by moderate Pearson correlations (~64%) between F_{PED} and F_{ROH}. The relevance of F_{PED} is further brought into question due to the fact that it does not account for recombination within the genome (McQuillan *et al*., 2008; Mastrangelo *et al*., 2016). The insufficiency of in-depth pedigree records therefore deemed F_{ROH} based on larger ROH segments (~66% correlated), which are indicators of more recent inbreeding, a more relevant substitute for F_{PED} than F_{ROH} based on short segments (~64% correlated). Molecular-based measures, such as F_{ROH}, have therefore gained popularity. SNP-based measures such as PLINK's F_{SNP}, however, merely estimates excessive marker homozygosity while the ROH-based measure is representative of the age of inbreeding. Genomic measures of inbreeding were strongly correlated (~95%) and this was in agreement with Mastrangelo *et al*. (2016) who showed correlations ranging from ~83% to ~95% between these measures.

All coefficients indicated low-level inbreeding (1-7%) within the SA Drakensberger breed. This was expected considering a decline in breeding animals observed within the breed (van der Westhuizen & Groeneveld, 2004; Abin, 2014) and the fact that sires are increasingly being used across breeders. PCA results, however, showed that there is still some dispersion, indicating weaker relations, between individuals and herds within the population. Currently within-breed selection focuses on a “standard of excellence”, placing emphasis on the maintenance of breed purity with regards to coat colour, physique and growth (Drakensberger Cattle Breeders’ Society of SA, 2011). The preservation of specific breed characteristics such as the black coat is presumed to have decreased genetic diversity, as was supported by observed- (0.344) versus expected (0.347) heterozygosity, through directional changes in the genotypic frequencies of the loci selected for. Certain historic events in the breed timeline are also believed to have diminished genomic diversity. One such genetic bottleneck was the “Great Trek” in 1834 during which the breed accompanied the migration of Dutch settlers (Drakensberger Cattle Breeders’ Society of SA, 2011).

5. Conclusion

The origin and genetic composition of the SA Drakensberger is uncertain. Previous molecular research has identified *Bos taurus* and *-indicus* signatures within the genome of this breed, albeit in unknown proportions. Considering the fact that modern SNP genotyping platforms incorporate SNPs discovered in predominantly *Bos taurus* breeds, it is uncertain how genomic variation within the admixed genome of this breed will influence downstream applications. Inter-chromosomal differences in MAF and LD conformed to expectations, but suggest that these differences need consideration in future genomic endeavors. The relationship between MAF and LD can be exploited in the selection of informative SNP for the possible development of optimized low-density panels. Inbreeding coefficients indicated low levels of inbreeding, which was expected due to artificial selection practices to maintain breed purity (with regards to characteristics such as the all-black coat colour and adaptability). ROH length characteristics furthermore pointed towards more ancient inbreeding, reflecting known historic bottleneck events. The economic importance of the SA Drakensberger as an adaptable indigenous breed in the SA beef industry has sparked interest in genomics-based breed improvement. Results of this study suggest that genomic applications such as imputation and GS can be further explored if genomic diversity is accounted for. Inferences made on the effect of a heterogeneous genome, such as the SA Drakensberger genome, on downstream applications may apply to other local genetic resources, for example non-descript or composite African breeds, with similarly complex genome architecture.

Statement on conflict of interest

The authors have no conflict of interest to declare.

Acknowledgements

Project funding by the Beef Genomics Program (BGP) and the Red Meat Research and Development Trust of South Africa (RMRDT-SA) is hereby acknowledged. The authors would also like to thank the SA Drakensberger Breeders' Society for providing their consent to proceed with this research project. The financial assistance of the National Research Foundation (NRF) towards this research is hereby also acknowledged. Opinions expressed and conclusions arrived at, are those of the author and are not necessarily to be attributed to the NRF.

References

Abin, S.A., 2014. Animal recording as a tool for improved genetic management in African beef cattle breeds. MSc thesis, University of Pretoria.

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

Berry, D.P., Garcia, J.F., Garrick, D.J., 2016. Development and implementation of genomic predictions in beef cattle. Anim. Front. 6, 32-38.

Bisschoff, C., Lotriet, R., 2013. The Drakensberger as competitive breed of cattle in the South African beef industry. In: The 19th International Farm Management Congress. Warsaw, Poland, 2013. 1-10.

Bohmanova, J., Sargolzaei, M., Schenkel, F.S., 2010. Characteristics of linkage disequilibrium in North American Holsteins. BMC Genomics 11, 421.

Bolormaa, S., Hayes, B.J., Hawken, R.J., Zhang, Y., Reverter, A., Goddard, M.E., 2011. Detection of chromosome segments of zebu and taurine origin and their effect on beef production and growth. J. Anim. Sci. 89, 2050-2060.

Bonsma, J.C., 1980. Cross-breeding, breed creation and the genesis of the Bonsmara. Livestock Production - a Global Approach. Tafelberg, Cape Town, 90–110.

Cañas-Álvarez, J.J., Mouresan, E.F., Varona, L., Díaz, C., Avilés, C., Baró, J.A., Altarriba, J., Casellas, J., Piedrafita, J., 2014. Linkage disequilibrium and persistence of phase in five spanish local beef cattle breeds. In: Proceedings of 10th World Congress of Genetics Applied to Livestock Production. Vancouver, BC, Canada, 2014. Poster presentation, 502.

Dekkers, J.C.M., Hospital, F., 2002. The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. 3, 22-32.

Drakensberger Cattle Breeders' Society of SA, 2011. Drakensberger Handbook, first ed. Volksrust, Mpumalanga, South Africa.

Edea, Z., Dadi, H., Kim, S.W., Park, J.H., Shin, G.H., Dessie, T., Kim, K.S., 2014. Linkage disequilibrium and genomic scan to detect selective loci in cattle populations adapted to different ecological conditions in Ethiopia. J. Anim. Breed. Genet. 131, 358-366.

Edea, Z., Dadi, H., Dessie, T., Lee, S-H., Kim, K-S., 2015. Genome-wide linkage disequilibrium analysis of indigenous cattle breeds of Ethiopia and Korea using different SNP genotyping BeadChips. Genes Genom. 37, 759-765.

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to quantitative genetics, fourth ed. Longman, New York.

Ferenčaković, M., 2015. Molecular dissection of inbreeding depression for semen quality traits in cattle. Doctoral thesis, University of Zagreb.

Grigson, C., 1991. An African origin for African cattle? – some archaeological evidence. Afr. Archaeol. Rev. 9, 119-144.

Hickey, J.M., Crossa, J., Babu, R., De los Campos, G., 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52, 654-663.

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38, 226-231.

Khatkar, M.S., Nicholas, F.W., Collins, A.R., Zenger, K.R., Cavanagh, J.A.L., Barris, W., Schnabel, R.D., Taylor, J.F., Raadsma, H.W., 2008. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genomics 9, 187.

Makina, S.O., Taylor, J.F., van Marle-Köster, E., Muchadeyi, F.C., Makgahlela, M.L., MacNeil, M.D., Maiwashe, A., 2015a. Extent of linkage disequilibrium and effective population size in four South African Sanga cattle breeds. Front. Genet. 6, 1-12.

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

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. Select. Evol. 48, 88.

Marchini, J., Howie, B., Myers, S., McVean, G., Donnelly, P., 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat. Genet. 39, 906-913.

Mastrangelo, S., Tolone, M., Di Gerlando, R., Fontanesi, L., Sardina, M.T., Portolano, B., 2016. Genomic inbreeding estimation in small populations: evaluation of runs of homozygosity in three local dairy cattle breeds. Anim. 10, 746-754.

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.L., Sonstegard, T.S., Van Tassell, C.P., 2009. Development and characterization of a high density SNP genotyping assay for cattle. PloS ONE 4, e5350.

McParland, S., Kearney, J.F., Rath, M., Berry, D.P., 2007. Inbreeding trends and pedigree analysis of Irish dairy and beef cattle populations. J. Anim. Sci. 85, 322-331.

McQuillan, R., Leutenegger, A-L., Abdel-Rahman, R., Franklin, C.S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., MacLeod, A.K., Farrington, S.M., Rudan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S.H., Dunlop, M.G., Wright, A.F., Campbell, H., Wilson, J.F., 2008. Runs of homozygosity in European populations. Am. J. Hum. Genet. 83, 359-372.

Mokry, F.B., Buzanskas, M.E., 2, Mudadu, M.D., Grossi, D.D., Higa, R.H., Ventura, R.V., de Lima, A.O., Sargolzaei, M., Meirelles, S.L.C., Schenkel, F.S., da Silva, M.V.G.B., Niciura, S.C.M., de Alencar, M.M., Munari, D.P., Regitano, L.C.dA., 2014. Linkage disequilibrium and haplotype block structure in a composite beef cattle breed. BMC Genomics 15, S6.

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

Pei, Y-F., Li, J., Zhang, L., Papasian, C.J., Deng, H-W., 2008. Analyses and comparison of accuracy of different genotype imputation methods. PloS ONE 3, e3551.

Porto-Neto, L.R., Kijas, J.W., Reverter, A., 2014. The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. Genet. Select. Evol. 46, 22.

Purcell, S., Neale, B., Todd-Brown, K., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559-575.

Purfield, D.C., Berry, D.P., McParland, S., Bradley, D.G., 2012. Runs of homozygosity and population history in cattle. BMC Genet. 13, 70.

Qanbari, S., Pimentel, E.C., 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.

Qwabe, S.O., van Marle-Köster, E., Maiwashe, A., Muchadeyi, F.C., 2013. Evaluation of the BovineSNP50 genotyping array in four South African cattle populations. S. Afr. J. Anim. Sci. 43, 64-67.

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. Resour. Info. 26, 1-26.

Sargolzaei, M., Schenkel, F.S., Jansen, G.B., Schaeffer, L.R., 2008. Extent of linkage disequilibrium in Holstein cattle in North America. J. Dairy Sci. 91, 2106-2117.

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.

Schrooten, C., Dassonneville, R., Ducrocq, V., Brøndum, R.F., Lund, M.S., Chen, J., Liu, Z., González-Recio, O., Pena, J., Druet, T., 2014. Error rate for imputation from the Illumina BovineSNP50 chip to the Illumina BovineHD chip. Genet. Select. Evol. 46, 10.

Sharmaa, A., Leea, J.S., Dang, C.G., Sudrajad, P., Kim, H.C., Yeon, S.H., Kang, H.S., Lee, S.-H., 2015. Stories and challenges of genome wide association studies in livestock - a review. Asian Austral. J. Anim. Sci. 28, 1371-1379.

Toosi, A., Fernando, R.L., Dekkers, J.C.M., 2010. Genomic selection in admixed and crossbred populations. J. Anim. Sci. 88, 32-46.

Van Marle-Köster, E., Visser, C., Berry, D.P., 2013. A review of genomic selection - Implications for the South African beef and dairy cattle industries. S. Afr. J. Anim. Sci. 43, 1-17.

Van der Westhuizen, R.R., Groeneveld, E., 2004. Population and pedigree analysis of indigenous South African beef breeds, in: Van Der Honing, Y., (Ed.), Book of abstracts of the 55th Annual Meeting of the Europian Association for Animal Production. Bled, Slovenia, pp. 37.

Weng, Z., Zhang, Z., Ding, X., Fu, W., Ma, P., Wang, C., Zhang, Q., 2012. Application of imputation methods to genomic selection in Chinese Holstein cattle. J. Anim. Sci. Biotech. 3, 6.

Yang, J., Hong Lee, S., Goddard, M.E., Visscher, P.M., 2011. GCTA: A Tool for Genome-wide Complex Trait Analysis. Am. J. Hum. Genet. 88, 76-82.

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.

CHAPTER FOUR

Assessing SNP selection methods for the development of a low-density panel optimized for imputation

Extracts of this chapter will be prepared as a manuscript to be submitted for publication.

4. 1 Introduction

Large-scale cattle genotyping is presently undertaken using the commonly termed single nucleotide polymorphism (SNP)-chips with a vast array of different panels available that vary in both genotype density and breed representation (Nicolazzi *et al*., 2015). Most of the genotype panels are constructed by selecting SNPs that are informative in the most populous breeds of either taurine- (e.g. Illumina® Bovine SNP50; Matukumalli *et al*., 2009) or indicine descent (e.g. GeneSeek® GGP^{TM} Indicus; Ferraz *et al*., 2018). The utility of panels favouring either of these subspecies may therefore not be optimal for less common breeds that often have admixed genomes harbouring varying proportions of taurine- and indicine-derived genomic segments. An example of such a breed, within the context of the South African (SA) beef industry, is the Drakensberger.

The SA Drakensberger is an indigenous beef breed with a sleek, black coat that belongs to the Sanga subspecies (*Bos taurus africanus*) of cattle (SA Drakensberger Breeders' Society, 2011). Genome-wide SNP analyses have indicated common ancestry of the breed with both European and African *Bos taurus* as well as *Bos indicus* breeds (Makina *et al*., 2016). In contrast to the composite SA Bonsmara breed that was developed experimentally (Bonsma, 1980), it is believed that the breed development of the modern SA Drakensberger cattle occurred over centuries. An important landmark in the initiation of its developmental process was crossbreeding in the 1700s between taurine cattle imported by European settlers and black-coated, indigenous cattle that were owned by southern African tribes (Niemand, 2013). Following northward migration, and enduring several events such as the Rinderpest disease outbreak in 1896 and the Anglo-Boer War from 1899 to 1902, which led to a severe erosion in cattle numbers, the breed as it is characterized today, was only officially recognized in 1947 when the SA Drakensberger Breeders' Society was formed (SA Drakensberger Breeders' Society, 2011).

In spite of a recent effort to re-sequence the genomes of Sanga breeds (Zwane, 2017), including several Drakensberger animals, a SNP panel applicable within a specific Sanga breed or across Sanga breeds does not exist at present. In the meantime, the success of genome-facilitated breed improvement strategies, such as genomic selection (GS), for breeds such as the SA Drakensberger is contingent on the utility of SNPs included in commercial high-density panels. Despite the ever-reducing cost of genotyping in cattle, the cost of genotyping itself is still the main barrier to adoption. One potential strategy to reduce the cost of genotyping is to reduce the number of SNPs to be genotyped and subsequently impute these SNPs to higher density. The SNPs selected for a reduced density panel must, however, be 1) abundant enough to facilitate acceptable imputation accuracy and 2) informative for the breed(s) in which they will be applied.

When current genotyping panels were applied in the SA Drakensberger, a high proportion of low minor allele frequency (MAF) SNPs as well as weak linkage disequilibrium (LD) between SNPs was observed (Qwabe *et al*., 2013; Zwane *et al*., 2016; Lashmar *et al*., 2018). It is therefore expected that the lower limit of SNPs required on a reduced density panel will be higher than proposed for taurine breeds in which longer haploblocks of these SNPs exist. Careful consideration of the genomic characteristics of SNPs will therefore be necessary so that a reduced panel of SNPs with optimal utility for the SA Drakensberger breed can be selected. Various methods have been proposed to identify, from a larger pool of existing SNPs, the most appropriate low-density SNP sets using certain genomic characteristics as inclusion criteria. These methods have generally considered genomic parameters such as mean MAF (e.g. Corbin *et al*., 2014; Judge *et al*., 2016), and inter-marker relationship estimates, i.e. LD (e.g. Ogawa *et al*., 2016), while maintaining more or less an even dispersion of selected SNPs across the genome. The most efficient selection strategy will enable minimization of the reduced-panel SNP density without compromising imputation accuracy and hence facilitate the use of such a panel in imputation-driven applications for local breeds.

The overarching aim of this chapter was therefore to determine the validity of using imputation from low-density panels, in terms of achievable imputation accuracy, for the indigenous SA Drakensberger breed. The main objectives were achieved by varying 1) the SNP density and 2) the choice of SNPs for various custom-derived low-density panels and imputing to higher density. An additional objective was to determine the impact of relatedness between the validation- and reference populations on the achievable imputation accuracy in the validation animals.

4.2 Materials & methods

Ethical approval to perform this study was granted by the ethics committee of the University of Pretoria's Faculty of Natural and Agricultural Science (ethics number: EC151106-024). Written consent was additionally provided by the SA Drakensberger Breeders' Society and the SA Stud Book Association (SA Stud Book) to utilize SNP genotypic data generated with funding from the Beef Genomics Program (BGP; www.livestockgenomics.co.za).

4.2.1. Genotypic data and quality control

Genotypes generated using the GeneSeek® Genomic Profiler^{TM} uHD SNP panel, consisting of 139 480 SNPs, were available for 214 male and 921 female SA Drakensberger cattle, all of which had a sample call rate exceeding 90%. Single nucleotide polymorphisms were mapped to the UMD3.1 bovine reference genome using SNPchiMp versions 3 and SNPConvert software (Nicolazzi *et al*., 2016). Only SNPs mapped to autosomes (*Bos Taurus* autosome; BTA1 - BTA29) were considered. Only markers with call rates exceeding 95%, MAF exceeding 1% and that did not violate Hardy Weinberg Equilibrium (P<0.01 x 10^{-6}) were retained. A total of 120 608 SNPs (mean call rate=99.4%) remained after edits.

4.2.2. Animal population subsets

The genotyped samples comprised of animals born between the years 1982 and 2017 and originated from 48 breeders nationwide. It should be noted that the discrepancy in birth year range between Chapter 3 and this chapter (1982-2016 versus 1982-2017) existed because the data set used in the former chapter was a subset of the data set used here; more animals, including younger calves, were sampled and genotyped over the entire study period. Available pedigree information for the breed was obtained from SA Stud Book (SA Stud Book) and consisted of 6 074 animals, which comprised the genotyped animals and their ancestors. Parent mismatches within the genotyping data set were identified as pedigree-defined parent-progeny pairs displaying >2% Mendelian inconsistencies. If parentage errors were identified, parents were set to missing in the pedigree file if no alternative matches within the genotype data set were identified, or updated if an alternative match was identified. The data set was then separated into a reference population (n=900) and a validation population (n=235). The validation population consisted of the youngest animals with no more than 3 paternal sibs. The reference population was used to estimate SNP MAF and the extent of LD, both of which were used as criteria for the selection of low-density SNP panels (discussed hereafter), and to model haplotypes used in imputation. The relatedness among individual samples was estimated using a genomic relationship matrix constructed in GCTA software (Genome-wide Complex Trait Analysis; Yang *et al*., 2011). The mean relationship coefficient between each validation animal and its ten closest relatives in the reference data set was estimated.

4.2.3. Single nucleotide polymorphism selection methods

Different strategies were used to develop several custom-derived low-density SNP panels consisting of 2 500, 5 000, 10 000, 20 000 and 50 000 SNPs. These panels may be hereon in referred to as the 2.5K, 5K, 10K, 20K and 50K panels. The number of SNPs selected per autosome was pre-defined and was proportional to the length of each autosome; therefore more SNPs were selected for longer autosomes. The number of SNPs selected per autosome to fulfil each of the different panel densities is outlined in Addendum 2. Five alternative algorithms were used to derive the custom SNP panels and these were implemented as follows:

4.2.3.1. Random selection (RAN)

The pre-defined number of SNPs required per autosome was chosen at random until each of the respective panel densities was reached.

4.2.3.2 Mid-point, equidistant selection (MID)

The length of each autosome, defined as the difference in base pairs between the first and last SNPs per autosome, was divided into equally sized segments and the SNP closest to the physical midpoint of each segment was chosen. The segment size per autosome was calculated as the autosomal length divided by $n - 1$, where n was the pre-defined number of SNPs to be chosen for that specific autosome. Due to uneven distribution of SNPs in specific autosomal regions after quality control procedures, certain segments did not harbour any SNPs and, in these situations, the SNP on the boundary, i.e. closest to the starting position, of the adjacent segment was chosen.

4.2.3.3. Equidistant selection that maximized MAF (DISTMAF)

Markers within segments of equal size were chosen based on an index that maximized MAF whilst attempting to adhere to the ideal inter-SNP spacing per autosome for each panel density. The SNP with the highest MAF was chosen within the first segment per autosome after which SNPs within subsequent segments were chosen based on the highest index score calculated as proposed by Matukumalli *et al*. (2009) and Zhang *et al*. (2010):

$$score_i = MAF_{SNP_i} + [ssize - (ssize - d_{SNP_i,SNP_{i-1}})] \tag*{[2]}$$

where MAF_{SNP_i} represented the MAF of a candidate SNP i, $ssize$ represented the genomic length of each segment within a given autosome and $d_{SNP_i,SNP_{i-1}}$ represented the genomic distance between the base pair position of a candidate SNP_i and SNP_{i-1}, where SNP_{i-1} was the SNP selected in the previous segment. If a segment contained no SNP, a second SNP was chosen in the next segment i.e. the SNP with the second highest index score was also chosen.

4.2.3.4. Segment-based selection combining MAF and LD (MAFLD)

An index score combining MAF and LD information was calculated per SNP and within segments. The MAF and LD per SNP were first standardized, so that the weights on each attribute were equal before summation. The scores were derived as follows:

$$score_{ij} = \frac{MAF_{SNP_i}}{SD_{MAF_{seg_j}}} + \frac{\overline{LD_{SNP_{ij}}}}{SD_{LD_{seg_j}}} \tag*{[3]}$$

where MAF_{SNP_i} represented the MAF of a candidate SNP i in segment j, $SD_{MAF_{seg_j}}$ represented the standard deviation for MAF in segment j, $\overline{LD_{SNP_{ij}}}$ represented the mean LD between a candidate SNP_i and all other SNPs within segment j, and $SD_{LD_{seg_j}}$ represented the standard deviation for all LD interactions within segment j. Within each segment, the SNP with the highest index score was chosen. A second SNP was selected in the segments at both ends of each autosome and the number of segments was therefore equal to the number of SNPs to be chosen minus two. The second SNP was selected based on a score combining MAF and the partial correlation of that SNP with all remaining candidate SNPs in their segment. Adjustments were made to the partial correlation to account for the relationship between each candidate SNP and the SNP already selected in the initial round of selection. This calculation was performed according to Judge *et al*. (2016) as follows:

$$r\left(SNP_i, SNP_j \middle| SNP_{sel}\right) = \frac{\left[r\left(SNP_i, SNP_j\right) - r\left(SNP_i, SNP_{sel}\right).r\left(SNP_j, SNP_{sel}\right)\right]}{\left[1 - r^2\left(SNP_i, SNP_{sel}\right)^{0.5}\right].\left[1 - r^2\left(SNP_j, SNP_{sel}\right)^{0.5}\right]}$$

[4]

where $r\left(SNP_i, SNP_j \middle| SNP_{sel}\right)$ represented the partial correlation between candidate SNP_i and candidate SNP_j corrected for the correlation with the already selected SNP, SNP_{sel}.

4.2.3.5. Partitioning-around-medoids (PAM), equidistant selection

SNPs on each autosome were partitioned into a number of clusters based on their proximity in genomic position using the partitioning-around-medoids (PAM) algorithm implemented in R's "*clara*" package (Kaufman & Rousseeuw, 2009). The number of clusters was equal to the number of pre-defined SNPs to be selected per autosome. The medial SNP within each SNP cluster was chosen. The PAM algorithm was the most computationally demanding, especially for SNP densities exceeding 20 000 SNPs, and because the difference in imputation accuracy was expected to be smallest at the 50K SNP density, this algorithm was only tested up to 20K SNPs for the purpose of this thesis.

4.2.4. Imputation and imputation accuracy

Imputation from each of the low-density panels to the higher density was performed using FImpute version 2.2 software (Sargolzaei *et al.*, 2014) based on both pedigree information and population-wide LD. This software was chosen because its methodology, using a sliding window approach, was deemed most suitable to the breed. Because the sliding window is systematically reduced to account for smaller genomic segments shared, corresponding to more distant relatedness, this software is ideal for utilization in breeds such as the SA Drakensberger with weak LD and high genomic heterogeneity. Imputation with this software is carried out simultaneously on a per chromosome basis. The software's default settings were used with regards to specifications of the sliding window used to capture haplotype similarities (i.e. shrink factor=0.150 and overlap=0.650).

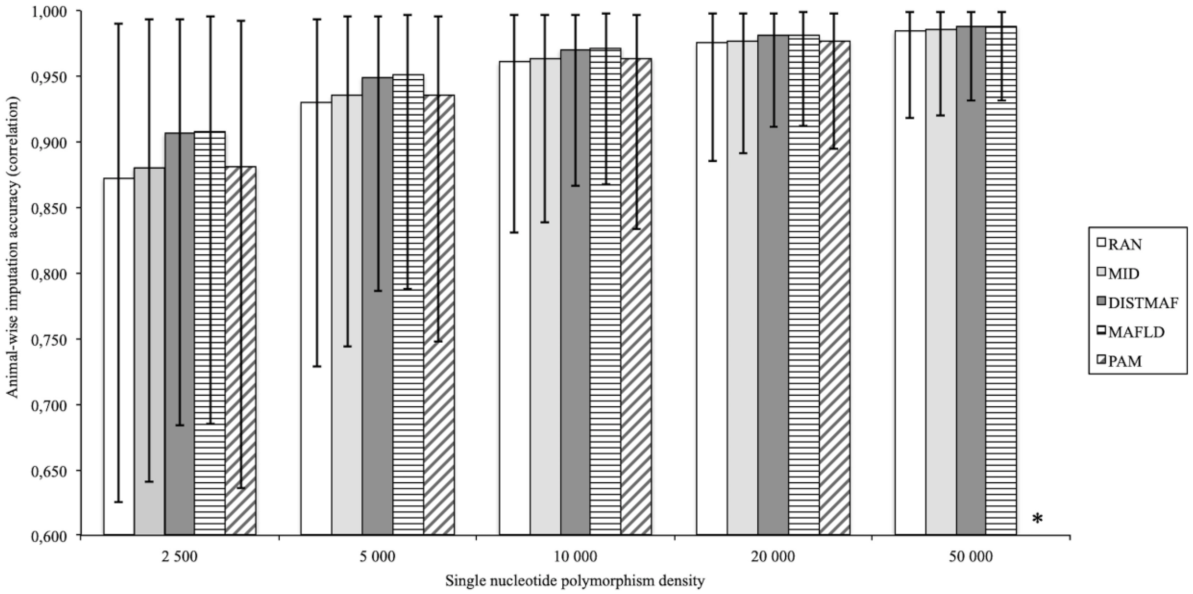
Imputation accuracy was quantified using three parameters namely: 1) genotype concordance rate (GCR), 2) allele concordance rate (ACR) and 3) the Pearson correlation between true- and imputed genotypes (COR). These parameters were averaged per animal (ACR_{ANIM}, GCR_{ANIM} and COR_{ANIM}, respectively) and per SNP (ACR_{SNP}, GCR_{SNP} and COR_{SNP}, respectively). The genotype and allele concordance rates were calculated as the proportion of correctly imputed genotypes and alleles, respectively. For genotype concordance, a score of zero was given to a SNP if it had either one allele (homozygous true versus heterozygous imputed) or both alleles (opposing homozygous for true versus imputed) incorrectly imputed. For allele concordance, a score of 0.5 was given if one allele was correctly imputed i.e. a homozygous true genotype versus a heterozygous imputed genotype. For both these measures, concordance was calculated 1) for only masked genotypes, i.e. that were imputed, and 2) for both masked and unmasked genotypes, i.e. both imputed and actual genotypes. The latter calculation was included to mimic what would be expected to happen in real life (Judge *et al.*, 2016).

4.3. Results

4.3.1 Imputation accuracy per animal

4.3.1.1. Number of SNPs on the lower density panel

The mean imputation accuracy in the validation animals increased with improvements in the number of SNPs included on the low-density panel, irrespective of the strategy used to select the SNPs (Figure 4.1).



[*] PAM was not used to derive a 50 000 SNP genotyping panel because of high computational demand.

Figure 4.1 Mean correlation-based imputation accuracies ($\mathrm{COR}_{\mathrm{ANIM}}$) for different genotyping panel densities derived using five different SNP selection methodologies (RAN: random selection, MID: midpoint selection, DISTMAF: equidistant selection maximizing MAF, MAFLD: combinative selection for MAF and LD, PAM: partitioning-around-medoids selection). Error bars represent minimum and maximum $\mathrm{COR}_{\mathrm{ANIM}}$.

The mean imputation accuracy per animal increased with increasing panel density but did so at a diminishing rate. Animal-wise imputation accuracy, $\mathrm{COR}_{\mathrm{ANIM}}$, ranged (minimum to maximum) from 0.625-0.990, 0.728-0.994, 0.830-0.996, 0.885-0.998 and 0.918-0.999 when 2 500, 5 000, 10 000, 20 000 and 50 000 SNPs were randomly chosen. This was further supported by smaller estimates of standard deviation for $\mathrm{COR}_{\mathrm{ANIM}}$ with increasing panel density; the standard deviation reduced from 0.075 for accuracy per animal to 0.014 as panel density improved from 2 500 (mean $\mathrm{COR}_{\mathrm{ANIM}}$=0.872) to 50 000 SNPs (mean $\mathrm{COR}_{\mathrm{ANIM}}$=0.985). The $\mathrm{COR}_{\mathrm{ANIM}}$ increased by 0.055, 0.043 and 0.043 units for the MID, DISTMAF and MAFLD methods when the number of SNPs doubled from 2 500 SNPs to 5 000 SNPs. The $\mathrm{COR}_{\mathrm{ANIM}}$ increased by only 0.008, 0.007 and 0.007 for MID, DISTMAF and MAFLD when the density increased from 20 000 to 50 000 SNPs. For all panel densities investigated, allele concordance rates were higher than the respective genotype concordance rates with the difference between these measures reducing with increasing SNP density. The difference between $\mathrm{ACR}_{\mathrm{ANIM}}$ and $\mathrm{GCR}_{\mathrm{ANIM}}$ was, for example, 0.051 units for the 2 500 SNP panel versus 0.007 units for the 50 000 SNP panel when the DISTMAF strategy was used.

4.3.1.2. SNP selection method

Across all panel densities evaluated, the poorest imputation accuracy was always achieved when SNPs were randomly selected. Strategies that based the selection of SNPs on scores combining MAF with other attributes (DISTMAF and MAFLD) outperformed the other selection strategies; the MAFLD method resulted in the best estimates of imputation accuracy (Table 4.1) irrespective of panel density.

Table 4.1 Genotype- and allele concordance rates for different low-density SNP panels.

LD panel	Method	Imputation accuracy			
		$\text{GCR}_{\text{MASKED}}$ (SD) [a]	GCR_{ALL} (SD) [b]	$\text{ACR}_{\text{MASKED}}$ (SD) [a]	ACR_{ALL} (SD) [b]
2.5K	RAN	0.857 (0.075)	0.860 (0.074)	0.925 (0.041)	0.926 (0.040)
	MID	0.865 (0.075)	0.868 (0.073)	0.929 (0.041)	0.931 (0.040)
	DISTMAF	0.893 (0.066)	0.896 (0.065)	0.944 (0.036)	0.945 (0.035)
	LDMAF	0.895 (0.066)	0.897 (0.065)	0.945 (0.035)	0.946 (0.035)
	PAM	0.867 (0.074)	0.869 (0.072)	0.930 (0.040)	0.931 (0.040)
5K	RAN	0.918 (0.052)	0.922 (0.050)	0.958 (0.028)	0.959 (0.027)
	MID	0.922 (0.066)	0.925 (0.064)	0.959 (0.042)	0.960 (0.041)
	DISTMAF	0.940 (0.044)	0.943 (0.042)	0.969 (0.023)	0.970 (0.022)
	LDMAF	0.942 (0.042)	0.944 (0.022)	0.970 (0.022)	0.971 (0.021)
	PAM	0.925 (0.050)	0.928 (0.048)	0.961 (0.027)	0.963 (0.026)
10K	RAN	0.954 (0.035)	0.957 (0.032)	0.976 (0.018)	0.978 (0.017)
	MID	0.954 (0.055)	0.957 (0.054)	0.975 (0.037)	0.977 (0.037)
	DISTMAF	0.964 (0.029)	0.967 (0.026)	0.982 (0.015)	0.983 (0.014)
	LDMAF	0.966 (0.028)	0.968 (0.025)	0.982 (0.014)	0.984 (0.013)
	PAM	0.957 (0.034)	0.960 (0.031)	0.978 (0.017)	0.980 (0.016)
20K	RAN	0.970 (0.025)	0.975 (0.021)	0.985 (0.013)	0.987 (0.011)
	MID	0.972 (0.023)	0.977 (0.019)	0.986 (0.012)	0.988 (0.010)
	DISTMAF	0.976 (0.020)	0.980 (0.017)	0.988 (0.010)	0.990 (0.009)
	LDMAF	0.977 (0.020)	0.980 (0.017)	0.988 (0.010)	0.990 (0.009)
	PAM	0.973 (0.023)	0.977 (0.019)	0.986 (0.012)	0.988 (0.010)
50K[c]	RAN	0.981 (0.016)	0.989 (0.010)	0.991 (0.008)	0.994 (0.005)
	MID	0.982 (0.016)	0.990 (0.009)	0.991 (0.008)	0.995 (0.005)
	DISTMAF	0.985 (0.014)	0.991 (0.008)	0.992 (0.007)	0.995 (0.004)
	LDMAF	0.985 (0.014)	0.991 (0.008)	0.992 (0.007)	0.995 (0.004)

[a] GCR_{ALL} and ACR_{ALL} = mean imputation accuracy across the full set of 120 608 SNPs, including both true and imputed SNPs;

[b] $\text{GCR}_{\text{MASKED}}$ and $\text{ACR}_{\text{MASKED}}$ = mean imputation accuracy across masked SNPs only i.e. only the SNPs imputed per density; [c] The PAM method was not tested for the 50 000 SNP panel because of computational demand; SD=standard deviation

Negligible differences were observed between strategies that only considered the genomic location of SNPs (i.e. MID and PAM). The DISTMAF and MAFLD methods were the only selection approaches with mean imputation accuracies exceeding 0.970 at a density of 5 000 SNPs; ACR_{ANIM} ranged from

0.856-0.997 and 0.859-0.997 for DISTMAF and MAFLD at this density of SNPs, respectively. Using the DISTMAF and MAFLD methods, ≥10 000 SNPs and ≥20 000 SNPs was required to yield mean allele concordance rates >0.980 and >0.990, respectively. The improvements in imputation accuracy were marginal when unmasked SNPs were included in the calculation of GCR_{ANIM} and ACR_{ANIM}, increasing with a mean value of 0.004 (standard deviation=0.002) and 0.002 (standard deviation=0.001), respectively, across all densities and selection strategies.

4.3.1.3. Degree of relatedness between validation and reference populations

The genomic relationship of a given animal in the validation population to animals in the reference population directly impacted the imputation accuracy. A scatter plot of each validation animal's imputation accuracy and the mean of its top ten coefficients of relatedness to the reference population is illustrated in Figure 4.2.

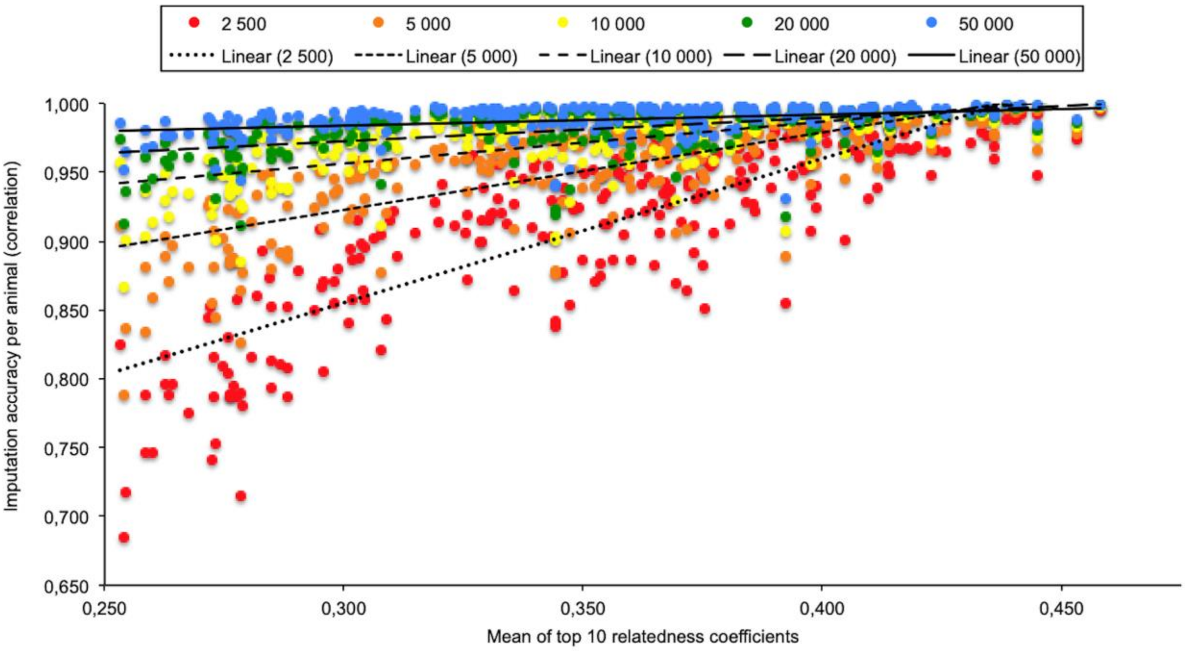


Figure 4.2 The relationship between degree of relatedness to the reference population and imputation accuracy for different SNP density panels derived using the MAFLD selection. Linear (2 500), linear (5 000), linear (10 000), linear (20 000) and linear (50 000) represent the linear regression lines for imputation accuracy on relatedness for each panel density.

The influence of genomic relatedness between reference and validation animals on correlation-based accuracy was more pronounced when fewer SNPs were included on the lower density panel. A minimum (maximum) COR_{ANIM} of 0.825 (0.995), 0.911 (0.997), 0.957 (0.998), 0.974 (0.998) and 0.986 (0.999) was observed for animals that were least and most related to the reference population when 2 500, 5 000, 10 000, 20 000 and 50 000 SNPs were utilized. The regression coefficients, b, in the regression equation, $Y = a + bX$, for the 2 500, 5 000, 10 000, 20 000 and 50 000 SNP panels were 1.047, 0.563, 0.304, 0.171, 0.081, respectively. The corresponding R^2 values for the 2 500, 5 000, 10 000, 20 000 and 50 000 SNP panels were 0.684, 0.557, 0.406, 0.265 and 0.135 (P<0.001); more of the variability in COR_{ANIM} was explained by per animal relatedness to the reference population when the reduced panel density was lower.

4.3.2. Imputation accuracy per SNP

4.3.2.1 Variation between autosomes

Imputation accuracy, described as either mean genotype- (GCR_{SNP}) or allele concordance rates (ACR_{SNP}) differed by autosome (Figure 4.3).

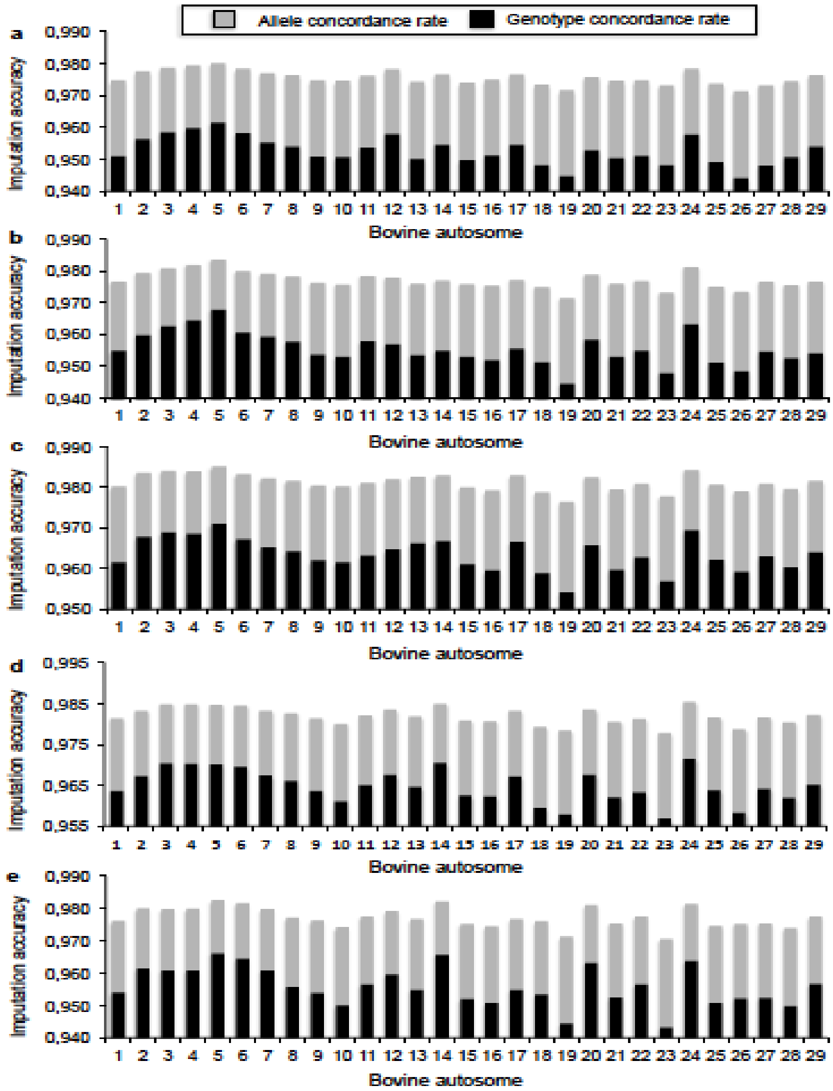


Figure 4.3 Mean concordance-based imputation accuracy measures for 10 000 SNP panels derived using the five different SNP selection methodologies (a: RAN, b: MID, c: DISTMAF, d: MAFLD and, e: PAM).

Using the 10 000 SNP panel as an example, the worst mean±standard deviation allele concordance rate across masked SNPs was for BTA26 (0.971±0.021) when random selection was undertaken. When other SNP selection strategies were employed, imputation accuracy of BTA19 (MID=0.971±0.018; DISTMAF=0.976±0.015) and BTA23 (MAFLD=0.978±0.015; PAM=0.971±0.020) were the worst. The greatest chromosome-wide allele concordance was observed for BTA5 for all selection strategies except the MAFLD strategy; for the MAFLD strategy, BTA24 (0.985±0.12) was the best.

4.3.2.2. Variation within autosomes

Within autosomes, variability in SNP-level imputation accuracy existed by location relative to the centre or the peripheries of the autosome. The distribution of COR$_{SNP}$ across individual autosomes is illustrated in Figure 4.4, comparing 2 500 versus 50 000 SNPs that were randomly selected.

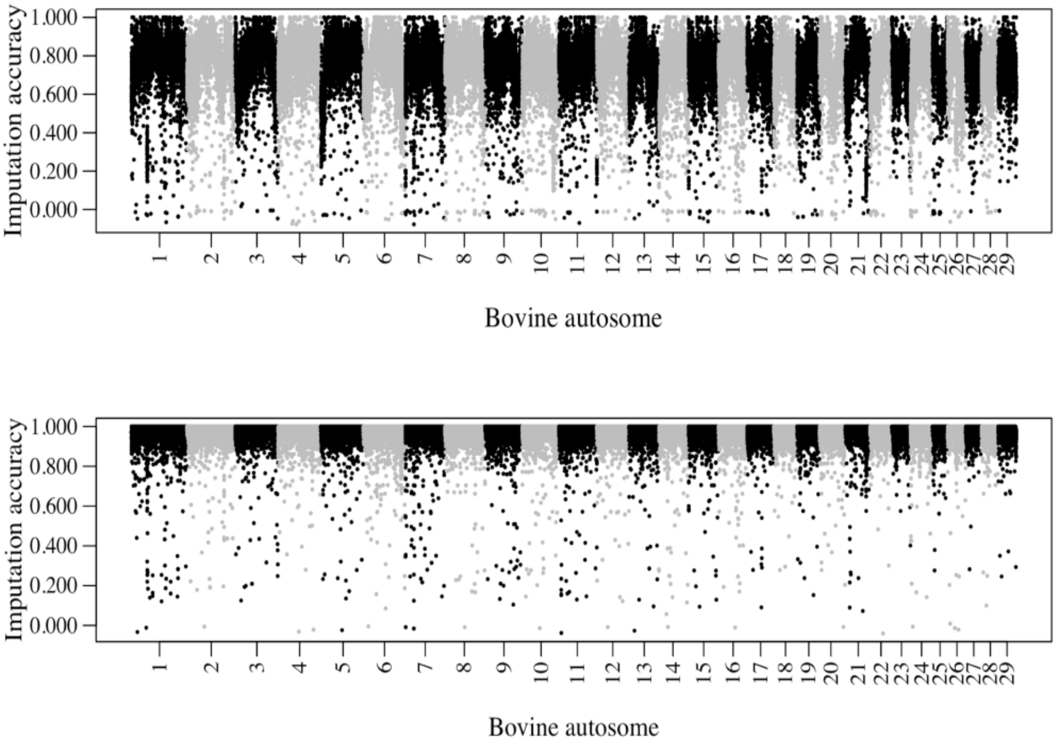


Figure 4.4 SNP-wise imputation accuracy (correlation) per autosome for 2.5K (top) versus 50K (bottom) SNP panels derived by RAN selection methodology.

The pattern of COR$_{SNP}$ distribution on many autosomes indicated a tendency towards worst imputation accuracies on the autosomal extremities and higher accuracies in the centre of the autosome. For the 2 500 and 50 000 SNP panels, mean±standard deviation COR$_{SNP}$ for SNPs located in the central 1Mb (0.5Mb to each side of the physical midpoint of each autosome) of autosomes were 0.781±0.067 and 0.968±0.015 across all autosomes when RAN was used. The corresponding means for the two 1Mb autosomal extremities were 0.682±0.149 and 0.664±0.07, and 0.938±0.130 and 0.968±0.018, for the 2 500 and 50 000 SNP panels. Differences in SNP-level imputation accuracy relative to genome locality were also observed between SNP selection strategies using the same panel density and these differences are illustrated in the Figure 4.5.

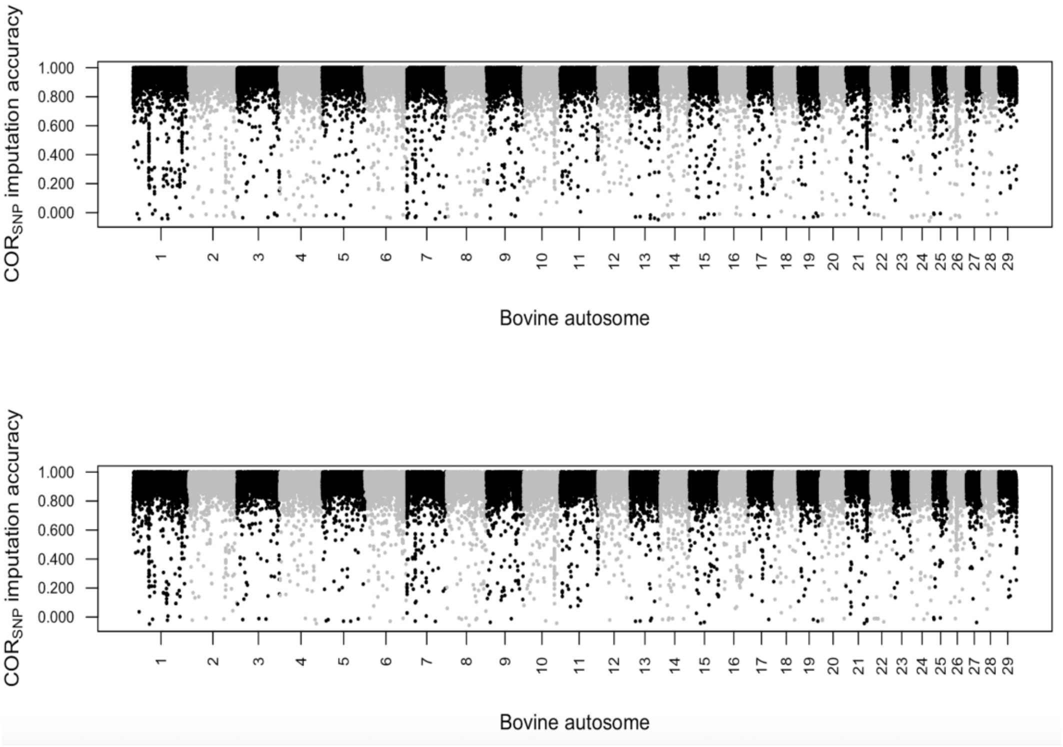


Figure 4.5 SNP-wise imputation accuracy (correlation) per autosome for 10K SNP panels derived by MAFLD (top) and RAN (bottom) selection methodologies.

The mean COR_{SNP} estimated for SNPs on autosomal extremities were lower for RAN than for MAFLD. Using a 10 000 SNP panel, the mean±standard deviation COR_{SNP} for SNPs located within 1Mb of the start, the centre and the end of autosomes were 0.923±0.036, 0.943±0.031 and 0.912±0.030 across all autosomes for MAFLD. For RAN the mean±standard deviation COR_{SNP} for the 1Mb extremities (autosomal start and end) was 0.882±0.122 and 0.892±0.050 across chromosomes, whereas the mean±standard deviation COR_{SNP} for SNPs in the 1Mb autosomal centres was 0.920±0.038. Mean COR_{SNP} for the central and peripheral regions (1Mb) per autosome are depicted in Addendum 3, comparing RAN and MAFLD methods.

4.3.2.3. Variability in imputation accuracy based on SNP MAF

Imputation accuracy differed by SNP MAF. The nature of the relationship between a given SNP's MAF and its imputation accuracy, however, differed depending on the parameter used to quantify accuracy (Figure 4.6). When MAF was binned into ranges of increasing MAF, mean values of concordance measures declined whilst the correlation measure increased with increments of higher MAF ranges.

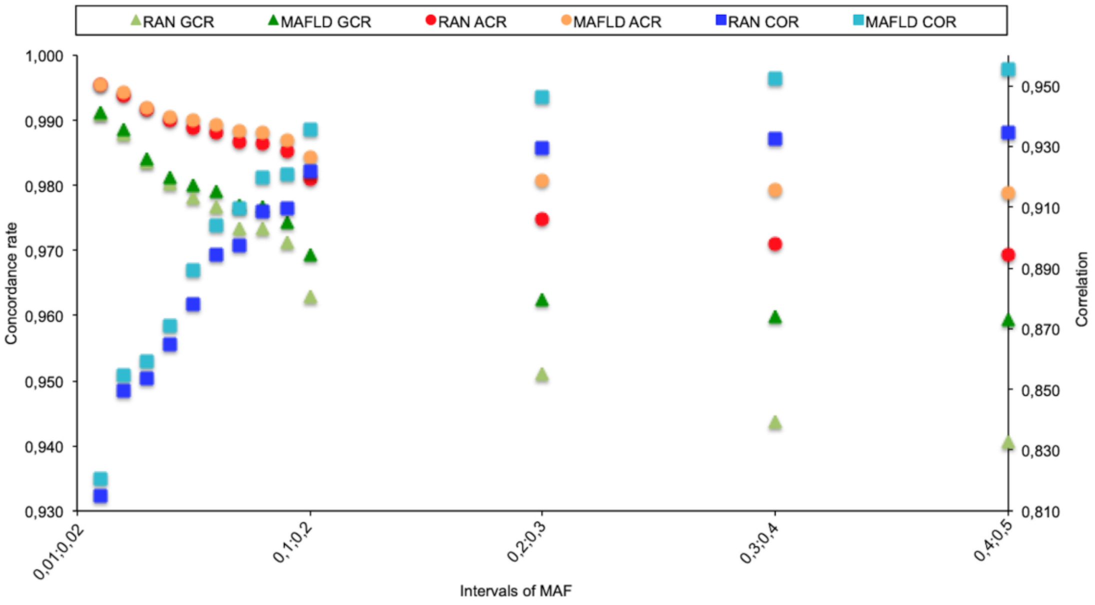


Figure 4.6 Mean genotype and allele concordance rate (primary Y-axis) as well as correlation-based (secondary Y-axis) imputation accuracy for intervals of increasing MAF. Points on the graph <0.1 MAF represent mean imputation accuracy for MAF intervals increasing in increments of 0.01. (RAN: random selection, MAFLD: combinative selection for MAF and LD).

The mean $\mathrm{COR_{SNP}}$ for SNPs classified in the highest MAF bin (0.4<MAF≤0.5) was 0.120 and 0.135 units higher, respectively, than for SNPs classified in the lowest MAF bins (0.01<MAF≤0.02) when the RAN and MAFLD selection strategies were used. For concordance measures, the mean imputation error rate for $\mathrm{GCR_{SNP}}$, defined as one minus mean $\mathrm{GCR_{SNP}}$, was approximately double the error rate for mean $\mathrm{ACR_{SNP}}$, defined as one minus mean $\mathrm{ACR_{SNP}}$, for all MAF bins. The difference between the allele- (AER) and genotype error rates (GER), $\Delta_{AER_{SNP},GER_{SNP}}$, per SNP was calculated as follows according to Ma *et al*. (2012):

$$\Delta_{AER,GER} = 2(1 - \mathrm{ACR}) - (1 - GCR) \qquad [5]$$

The difference increased from zero for the lowest MAF bin (0.01<MAF≤0.02) to 0.002 (RAN) and 0.001 units (MAFLD) for the highest MAF bin (0.4<MAF≤0.5). The discrepancy between both concordance-based measures and the correlation based accuracy measure was more prominent when MAF was low but diminished as MAF increased. Monomorphic SNPs and SNPs with MAF<0.01 were removed during quality control and were therefore not considered.

4.4. Discussion

Since the inception of the BGP in 2015, a large number of high-density SNP profiles have become available for many SA beef cattle (Van Marle-Köster & Visser, 2018). The available genotypic data has facilitated the implementation of genomic selection for certain indigenous breeds such as the SA Drakensberger. Post-BGP, the sustainability of routine genotyping, where the breeder incurs the full cost, will necessitate a transition to a low-density SNP panel that contains fewer SNPs and this should be less costly. The lower associated cost may improve the uptake of such technologies by more beef farmers. The principle motivation of this study was therefore to determine the achievable imputation accuracy for SA Drakensberger cattle from various lower density SNP panels that were constructed using different SNP selection strategies focusing on the key genomic characteristics of the breed. Genomic evaluations are, however, carried out internationally utilizing approximately 50 000 SNPs; for example in Ireland using the Irish Cattle Breeding Federation's (ICBF) 54K custom panel (Mullen *et al*., 2013) and in North America, New Zealand and Australia using the Illumina® 50K panel (Matukumalli *et al*., 2009; Wiggans *et al*., 2017). If genomic evaluation is to be based on medium, or higher, density SNP panels for SA beef cattle; the genotypes on the lower density panel must be imputed to higher density with minimal loss in accuracy. Results from this study may therefore provide guidelines for designing an applicable low-density SNP panel for Drakensberger and may possibly be transferable to other Sanga breeds.

4.4.1. Imputation accuracies per animal

4.4.1.1. SNP density of the reduced panel

The trend of increasing imputation accuracy in the more densely populated SNP panels can be attributed to the fact that haplotypes are more easily and accurately resolved when a greater number of unmasked, neighbouring SNPs are available or, in other words, there are fewer SNPs to impute (Tsai *et al*., 2017). With every incremental increase in panel density, the improvement in accuracy diminished especially when the number of SNPs on the lower of the two panels being compared was already high. This suggests that at higher SNP densities (>10K SNP panel), the density was already adequate to resolve shared haplotypes and to achieve high imputation accuracy. Across all selection methods, the average improvements in $\mathrm{COR}_{\mathrm{ANIM}}$ were 0.05, 0.03 and 0.01 units when panel densities were doubled from 2 500 to 5 000, 5 000 to 10 000 and 10 000 to 20 000 SNPs, respectively. This was in agreement with results reported by Judge *et al*. (2016) that documented improvements in animal-wise correlations of 0.07, 0.02 and 0.01 units when the number of markers were doubled from 1 000 to 2 000, 3 000 to 6 000 and 6 000 to 12 000 SNPs, respectively, in Irish cattle. Carvalheiro *et al*. (2014) documented similarly small gains in correlation-based accuracy (0.01 units) when the density of lower density, custom SNP panels was increased from 11K to 48K and imputed to high-density (777K). Yoshida *et al*. (2018), for example, also documented correlation-based improvements of 0.032 versus 0.003 units when SNP density increased from 500 to 3 000 SNPs versus 3 000 to 6

000 SNPs when imputation to 50 000 SNPs was carried out in Atlantic salmon. The diminishing rates of improvement in the present study, and the reduced associated inter-animal variability in imputation accuracy, for higher density panels suggest that using a panel consisting of more than 20 000 SNPs would have a negligible influence on imputation accuracy and would become less cost efficient. The extent with which the ideal number of low-density SNPs could be further reduced without compromising accuracy estimates, however, was a function of how the SNPs were selected.

4.4.1.2. Degree of relatedness between the validation- and reference populations

Results presented here, and from previous studies (e.g. Ventura *et al*., 2014; García-Ruiz *et al*., 2015), suggest that closer genetic relatedness between reference- and validation populations would enhance imputation accuracy. The diminishing regression coefficients (*d*) for regressions of imputation accuracy on genomic relatedness with increasing panel density suggests that relatedness became a less important factor when SNPs were sufficiently dense; for every unit increase in mean relatedness, there was a smaller unit increase in imputation accuracy for higher SNP densities (e.g. minimum correlation coefficient=0.081 at 50 000 SNPs). Imputation accuracy increasingly became a function of shared population-wide LD between SNPs rather than pedigree relationships; with sufficient SNP density, LD can be better captured and FImpute software (Sargolzaei *et al*., 2014) is better able to use this information instead of relying more on pedigree relationships provided. This software executes imputation under the assumption that all animals are somehow related, with shorter shared haploblocks capturing more distant relationships (Sargolzaei *et al*., 2014). With shorter distances between SNPs, and higher LD (Chapter 3), FImpute is therefore able to capture more distant genomic relatedness, which would be equivalent or superior to an animal's depth of pedigree.

Although results from this study show improved accuracy with closer genetic relatedness between validation and reference animals, it is recommended that future genotyping efforts be focused towards genotyping all major seedstock animals of the SA Drakensberger breed at higher density. To achieve accurate imputation, it is key that the pedigree of the animal imputed is genotyped on higher density (Berry & Kearney, 2011). Because the use of reproductive technologies, such as artificial insemination (AI), is considerably lower in beef cattle than in dairy cattle (Berry *et al*., 2016), it will be more challenging to identify bulls with high genetic impact that are used across herds. Reduction in the cost of genotyping will, however, enable more animals to be genotyped at higher densities, which may facilitate the establishment of larger reference populations and hence more accurate imputation (Berry & Kearney, 2011).

4.4.1.3. Criteria for SNP selection in terms of genomic characteristics

The imputation accuracy achieved per panel was conditional on the characteristics of the selected SNPs. Using the strategy that based SNP selection on MAF and LD, a mean±standard deviation ACR_{ANIM} estimate of 0.970±0.022 could be achieved using only 5 000 SNPs whilst equivalent imputation accuracies were only achieved at a 20 000 SNP panel density when other strategies, e.g. random and mid-point selection, were used.

SNP selection strategies that considered their MAF and an additional attribute, either inter-SNP distance or LD, generally produced better imputation accuracies than strategies that chose SNPs randomly or based solely on genome-wide dispersion. The worst method for SNP selection (RAN) showed significantly more variability in genomic dispersion than the best method (MAFLD) so poorer imputation accuracy was expected for this method. For the 10 000 SNP panels constructed, the mean±standard deviation distance between adjacent SNPs, in kilobase pairs (kb), was 250.30±266.03, 251.12±35.85, 251.00±100.05, 251.10±49.95 and 251.16±101.69 units when the RAN, MID, PAM, DISTMAF and MAFLD methods were employed.

The variability in distance between adjacent SNPs observed for RAN suggested that certain genomic regions were neglected using this method; lower imputation accuracy could be expected considering the weak LD estimated for SA Drakensbergers between adjacent SNPs (Chapter 3). This was corroborated by the fact that the difference in mean imputation accuracy between MAFLD and RAN selection strategies was more pronounced at lower panel densities; mean $\mathrm{COR_{ANIM}}$ and $\mathrm{ACR_{ANIM}}$ estimates were 0.035 and 0.016 units higher at a 2 500 SNP density compared to only 0.003 and 0.001 units higher at a 50 000 SNP density. Similarly, Judge *et al*. (2016) documented 0.031 versus 0.004 units higher accuracy for block selection, which was carried out the same as MAFLD in the present study, as opposed to random selection; these selections methods were also the best and worst methods in their study. A major contributing factor was the enrichment of chromosomal extremities when MAFLD was used; while a second SNP was selected in the first and last chromosomal segments, no SNPs were selected in these regions when RAN was used. Using the MAFLD method, SNPs were furthermore more evenly distributed than RAN because selection was done within genomic segments of even size.

The differences in imputation accuracy between the strategies that maximized MAF (DISTMAF and MAFLD) were negligible despite the fact that the variability in inter-SNP distance was considerably higher for MAFLD; $\mathrm{COR_{ANIM}}$ and $\mathrm{ACR_{ANIM}}$ were both, for example, only ~0.001 units higher for MAFLD than DISTMAF at a 2 500 SNP density. Genomic dispersion of SNPs was therefore not the most important factor impacting imputation accuracy. The mean±standard deviation MAF across the 10 000 SNPs was 0.272±0.140, 0.272±0.138, 0.270±0.140, 0.426±0.065 and 0.420±0.074 for the RAN, MID, PAM, DISTMAF and MAFLD selected panels. The SNP MAF was therefore a key factor in determining imputation accuracy; the two methods with the highest mean and smallest standard deviation MAF (DISTMAF and MAFLD) generated the best and second best imputation accuracies, with minor differences between them (0.001 difference in $\mathrm{COR_{ANIM}}$ using 10 000 SNPs).

In addition to selecting on a score of MAF and LD, the MAFLD strategy indirectly selected for even dispersion across autosomes by means of selecting within genomic segments of equal size and hence the similarity and slight superiority in resulting accuracy of this strategy to DISTMAF was expected. Moreover, the variability in imputation accuracy between all strategies that grouped candidate SNPs in some way, either within segments or clusters, was expected to diminish with increasing density of the reduced panel. For higher increments of panel density, each autosome was divided into a higher number of segments of shorter length. The probability of selecting the same SNP across different selection strategies therefore became higher because there were fewer options, candidate SNPs, to choose from within smaller segments.

SNP selection methods combining both attributes, i.e. MAF and LD, have previously been reported to generate the most accurate imputation in cattle compared to other methods such as random selection, selection using distance scores or selection using machine-learning algorithms (Carvalheiro *et al*., 2014; Judge *et al*., 2016). Other studies have found that placing emphasis on either one of these attributes, i.e. MAF (Corbin *et al*., 2014; He *et al*., 2018) or LD (Badke *et al*., 2013; Ogawa *et al*., 2016), while maintaining even genomic dispersion is more accurate than assuring even genomic dispersion alone. Improvements of 0.05 units proportion correctly imputed (1 000 SNP low-density panel; Corbin *et al*., 2014) and 0.6 percentage correctly imputed (6 000 SNP low-density panel; He *et al*., 2018) have been reported when MAF was optimized; while 0.02, 0.01 and 0.005 unit accuracy improvements (concordance rate) were achieved when selection for LD was performed within genomic segments, similar to the present study, to construct 1 000, 4 000 and 10 000 SNP panels, respectively (Ogawa *et al*., 2016). Results presented here were therefore consistent with previous research and suggests that if a reduced panel is to be designed for the SA Drakensberger, MAF and LD need to both be considered. This makes sense considering the weak estimates of genome-wide LD and high proportions of lower MAF documented in Chapter 3 of this thesis.

4.4.2. Imputation accuracies per SNP

4.4.2.1. Differences in imputation accuracy between and within autosomes

Several imputation studies in cattle have reported differences in imputation errors per chromosome (e.g. Berry & Kearney, 2011; Sun *et al*., 2012; Chud *et al*., 2015; Judge *et al*., 2016; Bernardes *et al*., 2018). Larger autosomes are expected to harbour more SNPs, which facilitates the capturing of stronger LD and subsequently enables more accurate inference of haplotypes (Sun *et al*., 2012). In agreement with previous research, imputation accuracy was superior for larger autosomes, as previously shown for dairy cattle (Judge *et al*., 2016), and was higher for autosomes that displayed stronger average LD between SNPs (Chapter 3; Lashmar *et al*., 2018), as previously shown for crossbred- (Chud *et al*., 2014) and *Bos indicus* beef cattle (Boison *et al*., 2015). In Chapter 3 of this thesis, however, it was shown that stronger LD estimates for certain autosomes might be an artefact of closer distances between SNPs and are not necessarily a reflection of stronger relationships between evenly dispersed SNPs. There may therefore be a higher probability that the greater imputation accuracy estimated for longer autosomes is because the low SNP-level accuracies in poorly imputed regions, such as the chromosomal extremities, are averaged out more over greater lengths (Judge *et al*., 2016). Conversely, shorter chromosomes are disadvantaged because these poorly imputed SNPs proportionally comprise a greater proportion of the total number of SNPs mapped to these chromosomes.

A further contributor to variability in accuracy between autosomes could be the gaps in mapped SNPs identified on specific autosomes. These gaps are indicated in Addendum 4. The largest identified genomic region that did not harbour any SNPs was a 2.3-megabase pairs (Mb) region on BTA12. These gaps between SNPs on the high-density panel (i.e. 120 608 SNP panel) resulted in certain segments in those regions not harbouring any candidate SNPs to select from. Adjacent segments were enriched with a second SNP and the dispersion of SNPs was therefore uneven flanking these gap regions.

Differences in correlation-based imputation accuracy were observed for SNPs located on the two autosomal extremes versus in the middle of autosomes. This was in agreement with studies such as Badke *et al*. (2013) that showed 0.023 units improvement in imputation accuracy for the middle 10% of SNPs versus SNPs within 5% of the chromosomal peripheries. Because the MAFLD selection methodology enriched the two extremes on each autosome with an extra SNP, higher mean imputation accuracies (+0.040 and +0.020 units higher than RAN for the first and last 1Mb regions across chromosomes) were achieved for SNPs located in these regions when this method was used. Furthermore, despite selecting additional SNPs on the chromosomal extremities, Judge *et al*. (2016) still documented poor imputation accuracy for shorter chromosomes due to the peripheral regions making up a larger proportion of the chromosome. Variation in SNP imputation accuracy within chromosomes was reduced in the present study as the SNP density of the panel increased, suggesting that the effect of a chosen SNP’s location within chromosome is more detrimental to imputation accuracy when SNP density is already sparse.

4.4.2.2. Imputation of low MAF markers

The Sanga subspecies has been shown to suffer a bias when commercial genotyping panels have been employed and this has been evidenced by a higher proportion of low-MAF, less informative SNPs, relative to taurine breeds (e.g. Qwabe *et al*., 2013). This has resulted in what is referred to as an ascertainment bias, introduced because of the exclusion of these breeds from SNP discovery processes involved in the panel design. Although commercially available and custom SNP panels available may have some utility for SA Sanga breeds, albeit sub-optimal, no Sanga-specific SNP panel exists at present (Zwane *et al*., 2019). Larger proportions of low MAF SNPs have therefore been observed for these breeds when these commercial SNP panels have been employed in genomic studies (e.g. Zwane *et al*., 2016; Lashmar *et al*., 2018). The impact of low MAF on imputation accuracy has been extensively studied and reported results have varied depending on the parameter(s) used to define accuracy (e.g. Mulder *et al*., 2012; Brøndum *et al*., 2014; Calus *et al*., 2014). Results from the present study corroborate the sensitivity of accuracy relative to MAF being a factor of the used measure of imputation accuracy. Concordance-based measure of accuracy (GCR_{SNP} and ACR_{SNP}) decreased with increasing MAF, whilst the correlation-based accuracy measure (COR_{SNP}) increased with increasing

MAF. Correlation-based measures have been suggested to minimize the dependency of imputation accuracy on SNP allele frequencies (Sargolzaei *et al*., 2014; Ventura *et al*., 2014). The difference in imputation accuracy was more pronounced for the RAN method than the MAFLD method as SNPs were not specifically chosen to maximize MAF; the difference in mean ACR_{SNP} between the lowest ($0.01<MAF\leq0.02$) and highest ($0.4<MAF\leq0.5$) MAF classes was for example 0.050 units for RAN as opposed to 0.017 for MAFLD. The mean COR_{SNP} increased from 0.821 to 0.955 for the same intervals with RAN. Similarly Ma *et al*. (2012) showed allele correct rates for imputation carried out with FImpute to decrease from >99% to ~93% and the correlation coefficient to increase from ~80% to ~85% when MAF increased from 0.1 to ~0.5 in European Red cattle.

The lower mean correlation-based imputation accuracy for SNPs within MAF intervals below 10% suggests that imputation of rare variants was more challenging (Calus *et al.*, 2014), which is concerning considering that these variants may be associated with unique or complex traits such as those pertaining to adaptability (Zwane *et al.*, 2019). Difficulty in imputing these markers may furthermore negatively influence the applicability of imputation towards GWAS as these studies use a SNP-by-SNP approach to test for association as opposed to using all SNPs collectively as in GS (Marchini & Howie, 2010). The independence of the COR accuracy measure to the influence of MAF, has therefore deemed this measure more robust when low-MAF SNPs are more abundant (Hickey *et al.*, 2012; Ogawa *et al*., 2016).

4.5 Conclusion

Genomic selection is currently undertaken in many countries globally using medium to high SNP density panels. Genotyping at higher densities is costly and financially unfeasible within the SA beef industry; this has been evidenced by the smaller number of genotyped animals compared to more developed countries globally. Genotyping selection candidates for lower densities and imputing to higher density will, however, significantly reduce the costs involved in routinely applying genomic selection strategies for locally important breeds in the commercial sector. The development of such panels generally relies on the identification of the most informative SNPs for a breed from a higher density genotypic platform, whether it is high-density genotyping arrays or sequencing data. Because SNP panels with specific utility in indigenous SA cattle are non-existent, the employment of a low-density panel in genomic technologies is contingent on identifying the most informative SNPs for these breeds. Results from this study indicate that a custom low-density panel consisting of at least 10 000 SNPs can be generated by using a combination of MAF and LD as selection criteria. Imputation accuracies from this panel were comparable with accuracies achieved internationally; a mean imputation error rate of <3% per animal could be achieved for the SA Drakensberger. Enrichment of the peripheral regions on chromosomes, by selecting more SNPs in these regions, will aid in improving mean chromosome-wide imputation accuracy.

References

Badke, Y.M., Bates, R.O., Ernst, C.W., Schwab, C., Fix, J., Van Tassell, C.P. & Steibel, J.P., 2013. Methods of tagSNP selection and other variables affecting imputation accuracy in swine. BMC Genet. 14, 8.

Bernardes, P.A., Al-Mamun, H.A., Suarez, M., Lim, D., Park, B. & Gondro, C., 2018. Imputation accuracy of whole-genome sequence data in Hanwoo cattle. In: Proceedings of the 11th World Congress of Genetics Applied to Livestock Production. Auckland, New Zealand, 6-11 February 2018.

Berry, D.P. & Kearney, J.F., 2011. Imputation of genotypes from low-to high-density genotyping platforms and implications for genomic selection. Animal 5, 1162-1169.

Berry, D.P., Garcia, J.F. & Garrick, D.J., 2016. Development and implementation of genomic predictions in beef cattle. Anim. Front. 6(1), 32-38.

Bonsma, J.C., 1980. Cross-breeding, breed creation and the genesis of the Bonsmara. Livestock production. A Global Approach. Ed. J.C. Bonsma. Tafelberg, Cape Town, South Africa. pp. 126-136.

Boison, S.A., Santos, D.J.A., Utsunomiya, A.H.T., Carvalheiro, R., Neves, H.H.R., Perez O'Brien, A.M., Garcia, J.F., Sölkner, J. & da Silva, M.V.G.B., 2015. Strategies for single nucleotide polymorphism (SNP) genotyping to enhance genotype imputation in Gyr (*Bos indicus*) dairy cattle: Comparison of commercially available SNP chips. J. Dairy Sci. 98, 4969-4989.

Brøndum, R.F., Guldbrandtsen, B., Sahana, G., Lund, M.S. & Su, G., 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. BMC Genomics,15(1), 728.

Calus, M.P.L., Bouwman, A.C., Hickey, J.M., Veerkamp, R.F. & Mulder, H.A., 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: A review of livestock applications. Anim. 8, 1743-1753.

Carvalheiro, R., Boison, S.A., Neves, H.H., Sargolzaei, M., Schenkel, F.S., Utsunomiya, Y.T., O'Brien, A.M.P., Sölkner, J., McEwan, J.C., Van Tassell, C.P. & Sonstegard, T.S., 2014. Accuracy of genotype imputation in Nelore cattle. Genet. Sel. Evol. 46(1), 69.

Chud, T.C., Ventura, R.V., Schenkel, F.S., Carvalheiro, R., Buzanskas, M.E., Rosa, J.O., de Alvarenga Mudadu, M., da Silva, M.V.G., Mokry, F.B., Marcondes, C.R. & Regitano, L.C., 2015. Strategies for genotype imputation in composite beef cattle. BMC Genet. 16(1), 99.

Corbin, L.J., Kranis, A., Blott, S.C., Swinburne, J.E., Vaudin, M., Bishop, S.C. and Woolliams, J.A., 2014. The utility of low-density genotyping for imputation in the Thoroughbred horse. Genet. Sel. Evol., 46(1), 9.

Ferraz, J.B.S., Wu, X., Li, H., Xu, J., Ferretti, R., Simpson, B., Walker, J., Silva, L.R., Garcia, J.F., Tait Jr., R.G. & Bauck, S., 2018. Design of a low-density SNP chip for Bos indicus: GGP indicus technical characterization and imputation accuracy to higher density SNP genotypes. In: *Proceedings* *of the 11th World Congress of Genetics Applied to Livestock Production*. Auckland, New Zealand, 6-11 February 2018.

García-Ruiz, A., Ruiz-Lopez, F.J., Wiggans, G.R., Van Tassell, C.P. & Montaldo, H.H., 2015. Effect of reference population size and available ancestor genotypes on imputation of Mexican Holstein genotypes. J. Dairy Sci. 98, 3478-3484.

He, J., Xu, J., Wu, X.L., Bauck, S., Lee, J., Morota, G., Kachman, S.D. & Spangler, M.L., 2018. Comparing strategies for selection of low-density SNPs for imputation-mediated genomic prediction in US Holsteins. Genetica 146(2), 137-149.

Hickey, J.M., Crossa, J., Babu, R. & de los Campos, G., 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52, 654-663.

Judge, M.M., Kearney, J.F., McClure, M.C., Sleator, R.D. & Berry, D.P., 2016. Evaluation of developed low-density panels for imputation to higher density in independent dairy and beef cattle populations. J. Anim. Sci. 94, 949-962.

Kaufman L. & Rousseeuw, P.J., 2009. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons; 2009 September 25.

Lashmar, S.F., Visser, C., Van Marle-Köster, E. & Muchadeyi, F.C., 2018. Genomic diversity and autozygosity within the SA Drakensberger beef cattle breed. Livest. Sci. 212, 111-119.

Ma, P., Brøndum, R.F., Zhang, Q., Lund, M.S. & Su, G., 2012. Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish red cattle. J. Dairy Sci. 96, 4666-4677.

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.

Marchini, J. & Howie, B., 2010. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11, 499-511.

Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS & Van Tassell CP, 2009. Development and characterization of a high density SNP genotyping assay for cattle. PloS one 24; 4(4):e5350.

Mulder, H.A., Calus, M.P.L., Druet, T. & Schrooten, C., 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95, 876-889.

Mullen, M.P., McClure, M.C., Kearney, J.F., Waters, S.M., Weld, R., Flynn, P., Creevey, C.J., Cromie, A.R. & Berry, D.P., 2013. Development of a custom SNP chip for dairy and beef cattle breeding, parentage and research. Interbull Bulletin, 47.

Nicolazzi, E.L., Biffani, S., Biscarini, F., Orozco ter Wengel, P., Caprera, A., Nazzicari, N. & Stella, A., 2015. Software solutions for the livestock genomics SNP array revolution. Anim. Genet. 46(4), 343-353.

Nicolazzi, E., Marras, G. & Stella, A., 2016. SNPConvert: SNP array standardization and integration in livestock species. Microarrays, 5(2), 17.

Niemand, M., 2013. Feedlot performance of the Drakensberger in comparison with other cattle breeds: A Meta-analysis. PhD thesis, University of Pretoria.

Ogawa, S., Matsuda, H., Taniguchi, Y., Watanabe, T., Takasuga, A., Sugimoto, Y. & Iwaisaki, H., 2016. Accuracy of imputation of single nucleotide polymorphism marker genotypes from low-density panels in Japanese Black cattle. Anim. Sci. J. 87, 3-12.

Qwabe, S.O., Van Marle-Köster, E., Maiwashe, A. & Muchadeyi, F.C., 2013. Evaluation of the BovineSNP50 genotyping array in four South African cattle populations. S. Afr. J. Anim. Sci. 43, 64-67.

SA Drakensberger Breeders' Society, 2011. Drakensberger Handbook (1st Ed.). Available online at URL: http://www.drakensbergers.co.za/pdf/manual.pdf.

SA Stud Book. http://www.sastudbook.co.za/?CID=2. Accessed 12 March 2019.

Sargolzaei, M., Chesnais, J.P. & Schenkel, F.S., 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15, 478.

Sun, C., Wu, X.L., Weigel, K.A., Rosa, G.J., Bauck, S., Woodward, B.W., Schnabel, R.D., Taylor, J.F. and Gianola, D., 2012. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genet. Res. 94(3), 133-150.

van Marle-Köster, E. & Visser, C., 2018. Genetic improvement in South African livestock: can genomics bridge the gap between the developed and developing sectors? Front. Genet., 9.

Tsai, H.Y., Matika, O., Edwards, S.M., Antolín–Sánchez, R., Hamilton, A., Guy, D.R., Tinch, A.E., Gharbi, K., Stear, M.J., Taggart, J.B. & Bron, J.E., 2017. Genotype imputation to improve the cost-efficiency of genomic selection in farmed Atlantic salmon. G3: Genes, Genomes, Genet. 7(4), 1377-1383.

Ventura, R.V., Lu, D., Schenkel, F.S., Wang, Z., Li, C. & Miller, S.P., 2014. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. J. Anim. Sci. 92, 1433-1444.

Wiggans, G.R., Cole, J.B., Hubbard, S.M. & Sonstegard, T.S., 2017. Genomic selection in dairy cattle: the USDA experience. Annual Rev. Anim. Biosci. 5, 309-327.

Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M., 2011. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88(1), 76-82.

Yoshida, G.M., Carvalheiro, R., Lhorente, J.P., Correa, K., Figueroa, R., Houston, R.D. & Yáñez, J.M., 2018. Accuracy of genotype imputation and genomic predictions in a two-generation farmed Atlantic salmon population using high-density and low-density SNP panels. Aquaculture, 491, 147-154.

Zhang, Z. & Druet, T., 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. J. Dairy Sci. 93, 5487-5494.

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.

Zwane, A.A., 2017. Genome-wide marker discovery in three South African indigenous cattle breeds (Afrikaner, Drakensberger and Nguni) using whole genome sequencing. PhD thesis, University of Pretoria.

Zwane, A.A., Schnabel, R.D., Hoff, J., Choudhury, A., Makgahlela, M.L., Maiwashe, A., Marle-Koster, V. & Taylor, J.F., 2019. Genome-wide SNP discovery in indigenous cattle breeds of South Africa. Front. Genet., 10, 273.

CHAPTER FIVE

Evaluating the accuracy of single-step genomic prediction using imputed genotypes of the SA Drakensberger beef cattle breed

Extracts of this chapter will be prepared as a manuscript to be submitted for publication.

5.1. Introduction

Until recently, genome-enhanced breed improvement programs were not a realistic option for South African beef cattle. The inauguration of the Beef Genomics Program (BGP) in 2015 assisted in accelerating efforts towards implementing genomic selection for a handful of the approximately 30 local beef breeds (van Marle-Köster & Visser, 2018a). Considerable focus was placed on indigenous cattle resources and state funding facilitated training populations to be assembled for the most numerous breeds, including the SA Bonsmara composite (SA Stud Book, 2017a) and the SA Beefmaster (Beefmaster Breeders' Society of SA & SA Stud Book, 2017). With training populations comprising more or less 2 200 genotyped animals for the SA Bonsmara and 800 genotyped animals for the SA Beefmaster, single-step genomic prediction methodology was implemented to estimate and release genomic-estimated breeding values (GEBVs) for these breeds (van Marle-Köster & Visser, 2018b). Due to the abundance of performance records, sufficient phenotypic data for a number of economically important traits was deemed adequate to consider genomic evaluation. Despite the fact that the number of genotyped animals per breed in SA is regarded as small (Mrode *et al*., 2018), especially in relation to the numbers acquired internationally (e.g. 1.6 million Holstein cattle in North America and >1 million total cattle in Ireland by 2017; ICBF, 2017; Weller *et al*., 2017); accuracies could potentially be improved by between 15 to 30% for lowly heritable traits using training populations of these sizes (van der Westhuizen *et al*., 2017). The SA Drakensberger is a prominent breed within the SA beef industry, displaying growth performance that is superior to most Sanga breeds and is comparible with exotic breeds within the commercial beef-producing sector of SA. With a history of diligently kept trait records, it is a prime candidate breed for the implementation of genomic selection.

Genomic evaluations were conventionally carried out using multi-step methods, whereby estimated breeding values (EBVs) were de-regressed to firstly estimate SNP effects and subsequently direct genomic values (DGVs) (Lourenco *et al*., 2015). These DGVs can be estimated for selection candidates based on their genotypes alone i.e. using the sum of their SNP effects (Lourenco *et al*., 2015). Relatively recently a quicker and simpler alternative to this methodology, the single-step genomic BLUP strategy, was proposed and has gained popularity (ssGBLUP; Misztal *et al*., 2009; Aguiler *et al*., 2010). What mainly differentiates this strategy of genomic evaluation from others is the extension of the relationship matrix into an H-matrix, which combines the genomic relationship matrix (G) of genotyped animals with a conventional numerator relationship matrix (A) that includes all non-genotyped animals in the pedigree (Christensen *et al*., 2012). Genotypic profiles are thereby essentially "imputed" for non-genotyped animals, using the genotypes available (Fernando *et al*., 2014). Because this strategy does not rely on prior de-regression of EBVs, a major advantage is computational time efficiency as GEBVs can be estimated for young, newly genotyped animals quicker (Legarra *et al*., 2014). The bias introduced in selective genotyping, by genotyping only animals based on the availability of information, is also overcome using ssGBLUP (Abdalla *et al*., 2019). A main challenge with this strategy, however, is ensuring that the A and G matrices are compactable in the H-matrix, and this generally requires adjustments in the form of scaling. National beef genomic evaluations are currently undertaken using single-step methodology in for example Europe and North America (Berry *et al*., 2016).

The single-step method is ideal for application in situations such as those pertaining to the SA Drakensberger where the prediction of GEBVs will rely jointly on the genotypes and phenotypes available. The number of genotyped animals for this breed is currently relatively small (approximately 1 200 animals). The pedigree completeness of the breed up to a pedigree depth of six generations is, however, approximately at 70% (Abin *et al*., 2016) and performance recording has been made compulsory since the 1980's (SA Drakensberger Breeders' Society, 2017). Emphasis of trait measurement has been on growth performance and the initial focus of GS will therefore be limited to these traits.

To routinely implement single-step genomic evaluation will require relatively frequent updating of the H-matrix used and this will include the addition of more genotyped animals. Despite the suitability of ssGBLUP for the breed, sustaining this technology can become financially unfeasible when considering current genotyping costs because high-density genotypes will still have to be acquired for major seedstock animals, that have a high genetic impact within the national herd, to accurately predict GEBVs of selection candidates. Uptake of genomic technologies by farmers, in terms of genotyping key animals within their herds, will furthermore need to be improved; to improve breed representation of the genotyped animals in the H-matrix, not only high-impact seedstock animals but animals across the spectrum of performance, i.e. both good and bad performing animals from both sexes, will have to be genotyped at higher density. Imputation will therefore have to be integrated into such a pipeline so that genotyping can be undertaken for selection candidates at a more affordable, lower density of SNPs, as this will be the responsibility of participating breeders. The achievable imputation accuracy for the SA Drakensberger from low-density genotyping panels has been addressed in Chapter 4 of this thesis.

The utility of imputed genotyped for GEBV estimation has been studied for beef cattle (e.g. Berry & Kearney, 2011; Cleveland *et al*., 2011; Mulder *et al*., 2012) and Wu *et al*. (2016) suggested that the bias from wrongly imputed genotypes, even when imputation error rates are high, will not significantly influence the accuracies of genomic predictions made from these genotypes. In the previous chapter of this study, imputation accuracy from several custom-derived low-denity panels were evaluated. It was concluded that a genotyping panel consisting of approximately 10 000 SNPs that were selected based on a combination of their minor allele frequncy (MAF) and linkage disequilibrium (LD), would be sufficient to achieve a mean±standard deviation animal-wise imputation accuracy of 0.972±0.024 for the SA Drakensberger.

The aim of this chapter was therefore to investigate the impact of using imputed genotypes on subsequent GEBV accuracy, using the ssGBLUP approach. The objective was to quantify the improvements in breeding value accuracy possible by using either true or imputed genotypes in addition to traditional pedigree information.

5.2. Material and methods

The Ethics committee of the University of Pretoria's Faculty of Natural and Agricultural Science granted ethics clearance (ethics number: EC151106-024).

5.2.1. Pedigree and phenotypic data for the study population

Genotypic and pedigree data for 1 135 SA Drakensberger cattle (214 male and 921 female animals) were used in the present study. The complete pedigree of the entire SA Drakensberger breed consisted of 232 169 animals. A cohort of 6 074 animals had direct pedigree relationships to the 1 135 animals with genotypes available. The average pedigree depth of the genotyped animals was 11.5 generations. Phenotypic records for two growth traits namely birth weight (BW) and weaning weight (WW) were studied; both the direct- and maternal breeding values were evaluated for these traits. Details pertaining to the phenotypic information that was available are depicted in Table 5.1. SA Stud Book provided pedigree and performance data with the signed consent of the SA Drakensberger Breeders' Society.

Table 5.1 Descriptive statistics of performance records for SA Drakensberger growth traits.

Trait	**Observations**	**Mean (kg)**	**Standard deviation (kg)**
Birth weight	152 931	35.04	4.55
Weaning weight	133 236	211.28	37.52

5.2.2. Single nucleotide polymorphism data for the study population

The SNP data used in the present study was generated using the GeneSeek® Genomic Profiler^{TM} uHD panel, which features 139 480 SNPs with a mean SNP density of 1 SNP per 19 kilobase pairs (kb). Standard quality control procedures in PLINK (Purcell *et al*., 2007) were followed to derive the final set of SNPs used in analysis; this included filtering out SNPs based on call rate (<95%), MAF (<1%) and Hardy-Weinberg Equilibrium *P*-values ($<0.01 \times 10^{-6}$). Any SNPs with unknown genomic positions or that were located on non-autosomal chromosomes were discarded. Sporadically missing genotypes per animal were imputed using FImpute (Sargolzaei *et al*., 2014). A total of 120 608 SNPs were available for the 1 135 genotyped individuals after edits.

5.2.3. Breeding value estimation

EBVs for each studied trait were estimated using three distinct sets of information. Firstly, breeding values were estimated traditionally, using a pedigree-based relationship matrix only. Secondly, genomic breeding values were estimated using single-step methodology using the true set of complete genotypes (120 608 SNPs). Lastly, genomic breeding values were estimated as in the previous step but using imputed genotypes; imputation was undertaken using a custom 10 000-SNP low-density panel i.e. 110 608 SNPs were imputed genotypes. The custom panel was derived by choosing from the pool of 120 608 SNPs, the 10 000 SNPs with the highest combinative score of their standardized MAF and LD within segments of equal length. This selection strategy is detailed in Chapter 4 (4.2.3.4. Segment-based selection combining MAF and LD; MAFLD). Imputation of the masked genotypes, i.e. the genotypes not in common between the custom low-density panel and the 120 608-SNP high-density panel, was undertaken for 235 validation animals, which were the youngest of the entire 1 135-animal data set, with FImpute software (Sargolzaei *et al.*, 2014).

Each of these strategies for breeding value estimation were carried out as follows:

5.2.3.1. Best linear unbiased prediction (BLUP)

A conventional animal model, using a pedigree-based numerator relationship matrix (A) was used to estimate breeding values for each trait. The model implemented to estimate these breeding values was defined as follows:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Z_1a} + \boldsymbol{Z_2m} + \boldsymbol{Z_3pe} + \boldsymbol{e} \tag{1}$$

where $\boldsymbol{y}$ was the vector of phenotypic values, $\boldsymbol{b}$ was a vector of fixed effects, $\boldsymbol{a}$ was a vector of random genetic effects and e was the vector of random residuals. The matrices $\boldsymbol{X}$, $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$ and $\boldsymbol{Z_3}$ were the incidence matrices associating the fixed ($\boldsymbol{b}$) and random ($\boldsymbol{a}$=sire x herd interaction, $\boldsymbol{m}$=maternal additive genetics and $\boldsymbol{pe}$ =permanent environment) effects with the phenotypic values ($\boldsymbol{y}$), respectively. For both traits, the fixed effects included were the sex, dam status i.e. whether the dam was a heifer or a cow, and contemporary group. For BW, the number of days into the calving season that the animal was born was also considered as a fixed effect; for WW, the age at weaning was considered. Random genetic effects included the sire x herd interaction as well as the additive genetic merit of the dam and the animal itself. For WW, an addition random effect considered was the permanent environment of the dam. Zero genetic correlation between direct and maternal effects was assumed.

Variance components for BW and WW traits were estimated by using VCE software (Variance Component Estimation; Groeneveld *et al.*, 2008) and were provided by SA Studbook. The estimated variance components were used to calculate heritabilities (h^2) for each trait using the equations V_A/V_T for direct- and V_M/V_T for maternal heritability, respectively. In the afore-mentioned equations V_T represented the total variance for the specific trait and, V_A and V_M represented the variance components for direct additive and maternal additive genetics, respectively. Heritability estimates for these traits are depicted in Table 5.2.

Table 5.2 Estimates of heritability for birth- and weaning weights of SA Drakensberger cattle.

Trait	**Heritability (h^2)**	
	Direct	**Maternal**
Birth weight	0.34	0.12
Weaning weight	0.25	0.17

5.2.3.2 Single-step genomic BLUP (ssGBLUP)

The ssGBLUP animal models were executed using Mix99 software (Lidauer *et al.*, 2015) to estimate GEBVs using phenotypic information of both genotyped and non-genotyped animals and incorporating both SNP and pedigree information in the process. The animal model employed had the same elements as the model employed for traditional BLUP, however, replaced the relationship matrix with a matrix, H, which was the product of incorporating the genomic relationship matrix (G) into the pedigree-based numerator relationship matrix. As described by Legarra *et al.* (2009), Christensen & Lund (2010) and Aguilar *et al.* (2011); the inverse of the H-matrix, which was used in the model, was constructed as follows:

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

[2]

where G represented the genomic relationship matrix, A_{22} represented the pedigree-based relationship matrix between genotyped animals only and A represented the pedigree-based relationship matrix between all animals, genotyped and non-genotyped. The numerator relationship matrix was constructed using *Relax*2 software (Strandén & Vuori, 2006), using the complete SA Drakensberger pedigree as input. The G and H^{-1} matrices were then prepared using the *hginv* program of Mix99 software (Lidauer *et al.*, 2015). The method using Euclidean distances between animals was used in the preparation of the G and hence H^{-1} matrices (Garcia-Baccino *et al.*, 2017). Two separate H^{-1} matrices were prepared - one using the G constructed from actual data and one using the G constructed from the data set including imputed genotypes. The former data set consisted of the true genotypes of the 120 608 SNPs genotyped per animal whereas the latter data set consisted of 10 000 true genotypes, selected using MAFLD SNP selection (Chapter 4), and 110 608 imputed genotypes.

5.2.3.3. Prediction accuracy

Reliabilities of GEBVs were approximated using the ApaX99 software (Lidauer *et al.*, 2015) and by means of the calculation method suggested by Misztal & Wiggans (1988). The accuracy of GEBVs per animal was calculated as $\sqrt{reliability}$. The pedigree-based EBVs were used as a benchmark for comparison and the Pearson correlation, *r(EBV, GEBV)*, between EBVs and GEBVs were estimated. The GEBV accuracies using actual genotypes (GEBV$_{\text{TRUE}}$) were calculated for the entire set of 1 135 animals, whilst the GEBV accuracies using a proportion of imputed genotypes ($GEBV_{IMPUTED}$) were calculated for the 235 youngest animals that were used as the validation population in imputation carried out in Chapter 4.

5.3. Results

5.3.1. Accuracy of genomic predictions without imputation

The relationship between EBV and $GEBV_{TRUE}$ accuracies was first estimated to determine whether the addition of genomic information to traditional pedigree-information increased the accuracy of EBVs for the growth traits studied. The mean (±standard deviation) accuracy of EBVs is summarized in Table 5.3.

Table 5.3 Mean EBV and GEBV accuracies for the 1 135 genotyped SA Drakensberger cattle.

Trait	EBV accuracy±SD	GEBV accuracy±SD*
BW_{direct}	0.783±0.081	0.799±0.065
$BW_{maternal}$	0.696±0.107	0.723±0.091
WW_{direct}	0.749±0.081	0.769±0.066
$WW_{maternal}$	0.725±0.115	0.747±0.100

*GEBVs were estimated from the complete set of 120 608 SNPs (without imputation)

The mean improvement in breeding value accuracy that was made using genomic information was 0.016, 0.026, 0.019 and 0.021 units for BW_{direct}, $BW_{maternal}$, WW_{direct} and $WW_{maternal}$, respectively. Across the 1 135 genotyped animals, the maximum per-animal improvement in accuracy was 0.294, 0.336, 0.247 and 0.311 units for BW_{direct}, $BW_{maternal}$, WW_{direct} and $WW_{maternal}$. The relationship between the animal-wise EBV and $GEBV_{TRUE}$ accuracies for birth- and weaning-weight traits are illustrated in Figure 5.1 and 5.2.

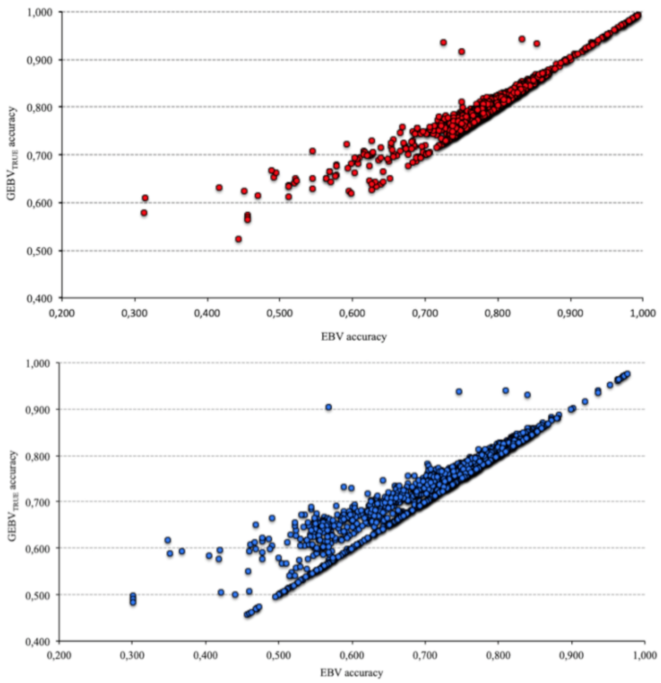


Figure 5.1 Scatter plot indicating the relationship between accuracies of EBV and GEBV$_{TRUE}$ for direct- (top) and maternal (bottom) birth weight traits of the SA Drakensberger cattle sampled.

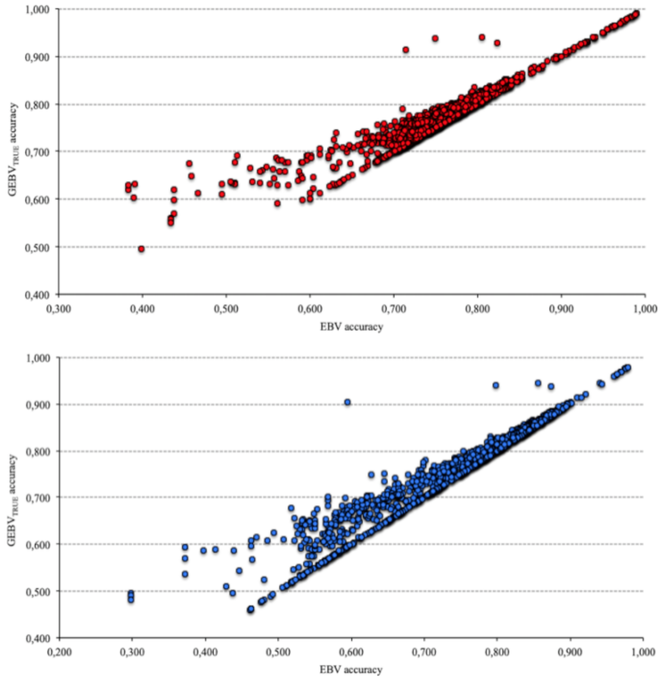


Figure 5.2 Scatter plot indicating the relationship between accuracies of EBV and GEBV$_{TRUE}$ for direct- (top) and maternal (bottom) weaning weight traits of the SA Drakensberger cattle sampled.

5.3.1. Accuracy of genomic predictions with imputed genotypes

The accuracy of predicted GEBVs were estimated from genotypic data sets consisting of either actual or imputed SNP genotypes. The mean GEBV accuracies achieved are summarized in Table 5.4.

Table 5.4 Mean GEBV accuracies for four growth traits estimated using true- versus imputed SNP genotypes.

Trait	**GEBV prediction accuracy**[*]		**Correlation**[**]
	True±SD (range)	**Imputed±SD (range**	
Birth weight (direct)	0.774±0.056 (0.564-0.951)	0.773±0.055 (0.570-0.951)	1.000 (P<0.001)
Birth weight (maternal)	0.657±0.068 (0.456-0.776)	0.656±0.068 (0.460-0.778)	1.000 (P<0.001)
Weaning weight (direct)	0.739±0.054 (0.551-0.926)	0.739±0.054 (0.558-0.926)	1.000 (P<0.001)
Weaning weight (maternal)	0.670±0.075 (0.461-0.814)	0.669±0.074 (0.461-0.814)	1.000 (P<0.001)

[*] Mean ± standard deviation; [**] Pearson correlation

As depicted in Table 5.4, minimal differences were observed between GEBV accuracies produced from the genomic evaluations using either actual or imputed genotypes; the accuracies produced were mostly the same, with a 0.001 units difference observed in mean GEBV accuracy for $\mathrm{BW}_{\mathrm{direct}}$, $\mathrm{BW}_{\mathrm{maternal}}$ and $\mathrm{WW}_{\mathrm{maternal}}$.

To investigate the influence of true versus imputed genotypes on the actual breeding value produced, and not the accuracy, the unit difference (in kilograms) per trait was quantified between the different analyses. Comparisons were made between breeding values produced using pedigree data versus true genotypes ($\Delta_{GEBV_{TRUE}-EBV}$), pedigree data versus imputed genotypes ($\Delta_{GEBV_{IMP}-EBV}$) and true versus imputed genotypes ($\Delta_{GEBV_{TRUE}-GEBV_{IMP}}$). Because animals with lower EBV accuracies are expected to gain more in accuracy from genomic evaluation, as opposed to animals with high EBV accuracies, validation animals were seperated into two groups of below average versus above average EBV accuracies to compare the kilogram difference in breeding values estimated. The unit difference in breeding values produced for the four growth traits are illustrated in Figure 5.3.

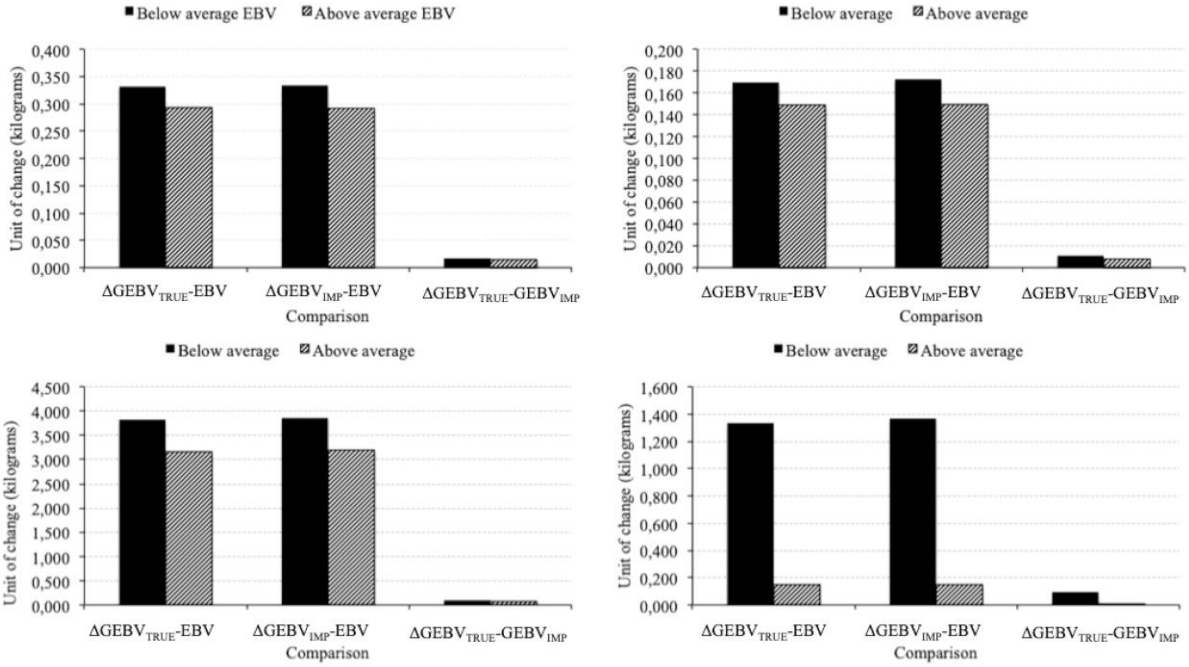


Figure 5.3 Bar plots illustrating the unit difference between mean breeding values produced with and without genomic information and the use of imputation for the four growth traits studies (top left: BW$_{\mathrm{direct}}$, top right: BW$_{\mathrm{maternal}}$, bottom left: WW$_{\mathrm{direct}}$, bottom right: WW$_{\mathrm{maternal}}$).

Larger differences between traditional- and genomic breeding values were always observed for animals that traditionally had lower EBV accuracy i.e. the addition of genomics produced a breeding value that was more different than its traditional breeding value for these animals. The kilogram breeding value difference was always larger between EBV and GEBV$_{\mathrm{IMPUTED}}$ than between EBV and GEBV$_{\mathrm{TRUE}}$. For BW$_{\mathrm{direct}}$, BW$_{\mathrm{maternal}}$, WW$_{\mathrm{direct}}$ and WW$_{\mathrm{maternal}}$ the root mean square error values (RMSE) were 0.154, 0.206, 3.960 and 1.782kg for $\Delta_{GEBV_{TRUE}-EBV}$ versus 0.158, 0.209, 3.994 and 1.810kg for $\Delta_{GEBV_{IMP}-EBV}$. The RMSE values for $\Delta_{GEBV_{TRUE}-GEBV_{IMP}}$ were always the smallest; RMSE values were 0.023, 0.013, 0.104 and 0.104kg for BW$_{\mathrm{direct}}$, BW$_{\mathrm{maternal}}$, WW$_{\mathrm{direct}}$ and WW$_{\mathrm{maternal}}$. The Spearman's rank correlation (P<0.001) between the EBV and GEBV$_{\mathrm{TRUE}}$ (EBV and GEBV$_{\mathrm{IMPUTED}}$) was r_S=0.965 (r_S=0.965), r_S=0.980 (r_S=0.980), r_S=0.961 (r_S=0.961) and r_S=0.982 (r_S=0.982) for BW$_{\mathrm{direct}}$, BW$_{\mathrm{maternal}}$, WW$_{\mathrm{direct}}$ and WW$_{\mathrm{maternal}}$. These correlations were always 1 (0.9998, P<0.001) between GEBV$_{\mathrm{TRUE}}$ and GEBV$_{\mathrm{IMPUTED}}$ for all four traits studied.

5.4. Discussion

5.4.1. Genomic predictions without imputation

In this chapter, a single-step genomic evaluation was applied to the SA Drakensberger breed for two growth traits, having both direct and maternal components. To evaluate firstly the value of using the single-step approach to genomic evaluation, the accuracy of ssGBLUP GEBVs were compared to the accuracy of EBVs estimated by means of traditional pedigree-based BLUP. The single-step approach to genomic evaluations takes advantage of information from both non-genotyped and genotyped animals by combining the numerator relationship matrix and genomic relationship matrix into one H-matrix (Forni *et al*., 2011). The ssGBLUP approach to genomic prediction has previously been shown to produce highly accurate GEBVs (e.g. Gao *et al*., 2015; Mäntysaari *et al*., 2017; Nayee *et al*., 2018). Considering that the sample of genotyped animals used in this study was numerically relatively small, but consisted of animals with extensive pedigree records available (mean=11.5 generations depth), it was expected that single-step methodology would be the most appropriate strategy for applying GS.

The results generated in this study were consistent with expectations and previous reports; higher breeding value accuracies were observed when ssGBLUP was undertaken as opposed to pedigree-based BLUP (e.g Aliloo *et al*., 2018; Johnston *et al*., 2018). The improvements in accuracy observed were, however, marginal with a mean increase of approximately 0.02 units across the four traits studied. Authors Song *et al*. (2018) found similarly small improvements (~0.01 units) when GEBV estimates across seven body measurement traits were based on simulated data of a small number of genotyped animals. Song *et al*. (2018) observed that even when the number of genotyped animals reached 3 000 animals, the improvements with ssGBLUP was small when this number was proportionately small in comparison to the number of animals in the pedigree (e.g. 26 000), which was used in BLUP. For the SA Drakensberger breed, 0.005% of the animals in the complete pedigree are genotyped; considerable increases in the number of genotyped animals will therefore need to be achieved to significantly increase GEBV accuracies from ssGBLUP.

In the current study, the improvements that were observed for the maternal traits were superior to those observed for traits considering only the animal's additive genetics. For birth weight, for example, the improvement in accuracy from ssGBLUP was 0.01 units higher for the maternal component than the improvement for the direct genetic component. Even though few studies have been performed in beef cattle using the single-step approach, Lourenco *et al*. (2013) showed that with simulation data, ssGBLUP prediction accuracies were always superior to BLUP prediction accuracies, and this was especially true for maternal traits (Legarra *et al*., 2014). Because the focus of breeding value estimation in the present study was on growth traits, which have been more diligently measured over a longer period of time, traditional EBVs were expected to have relatively high accuracies already and subject to potentially only small improvement from the addition of genomic information.

The larger difference in EBV and GEBV-based accuracies between the direct and maternal components of birth weight compared to weaning weight can also be related to a larger discrepancy in heritabilities of these components. Many previous studies have reported lower predictive ability of GEBVs for lower heritable traits, however, these traits are expected to gain the most from GS methodology; GEBV accuracy for lowly heritable traits have been shown to improve with a larger number of phenotypic records included in the prediction model (Behmaram *et al*., 2013). Using a training population of approximately 2 200 animals, van der Westhuizen *et al*. (2017) observed up to 15% improvements in prediction accuracy for lowly heritable and hard-to-measure traits. It is therefore expected that further improvements can be made for these traits, once more animals are genotyped and these genotypes are included in the constructed H-matrix. Emphasis should be placed on phenotyping lowly heritable and hard-to-measure traits for the SA Drakensberger as many of these traits are proxy-indicators for the overall ability of indigenous breeds to adapt (Scholtz *et al*., 2014).

5.4.2. Genomic predictions with imputation

The utility of imputed genotypes in genomic prediction equations have been studied previously in other beef cattle breeds (e.g. Weigel *et al*., 2010; Berry & Kearny, 2011; Cleveland *et al*., 2011; Mulder *et al*., 2012; Cleveland & Hickey, 2013; Aliloo *et al*., 2018) and results generally suggest a benefit. Studies on intensively raised livestock such as dairy cattle and pig populations have obtained correlations of ~0.95 between $\text{GEBV}_{\text{TRUE}}$ and $\text{GEBV}_{\text{IMPUTED}}$ (e.g. Weigel *et al*., 2010; Cleveland & Hickey, 2013) and this is expected because these livestock populations tend to have strong LD and thus share large genomic segments within-population. Imputation studies on these livestock species often include a substantial number of genotyped animals with high imputation accuracies being achieved. The effect of a small percentage of incorrectly imputed genotypes is therefore expected to be negligible, especially considering the fact that the GS algorithms generally estimate genome-wide SNP effects (Wu *et al*., 2016).

Results presented in the present study indicated marginal differences (approximately 0.001 units) between GEBV accuracies produced from true- versus imputed SNP genotypes. Furthermore minor differences existed in kilogram breeding values produced; RMSE ranged from 0.013kg ($\text{BW}_{\text{direct}}$) to 0.104kg (both $\text{WW}_{\text{direct}}$ and $\text{WW}_{\text{maternal}}$). This suggests that the substitution of imputed genotypes for actual genotypes had minimal influence on the breeding values generated. For comparison, results by Aliloo *et al*. (2018) were of specific interest, as a similar correlation coefficient (r=0.95) was observed for East African crossbred dairy cattle that also have admixed genomes. Comparably, a smaller number of genotyped animals were included in the ssGBLUP H-matrix and considerably fewer SNPs were to be imputed (imputation from 10K to ±120K in this study versus imputation from 4K to 777K by Aliloo *et al*., 2018). Considering that, despite these limitations, results presented here compared favourably; one can deduce that, with improvements in the number of genotyped animals available and more diligent phenotyping efforts, ssGBLUP will be a valid strategy for admixed cattle such as the SA Drakensberger. Lastly, the fact that the correlation observed by Aliloo *et al*. (2018) was obtained despite poor animal-wise imputation accuracies (correlation as low as 0.65) and that results presented here could be achieved by using imputed genotypes with a mean±standard deviation (minimum) correlation-based imputation accuracy of 0.972±0.024 (0.862) indicates that imputation mediated GS is moreover a viable solution for these breeds. It is, however, recommended that GEBVs for these traits be generated and evaluated routinely as more genotyped animals become available as the results presented here are based on limited sample size.

5.5. Conclusion

This study provided a first insight into the application of genomic selection methodology for the South African Drakensberger breed using imputed genotypes. The results presented here should be interpreted with the consideration that a relatively small population of genotyped animals was available in comparison with global studies and should therefore be regarded as preliminary. The achieved prediction accuracies were, however, promising and suggest that the inclusion of imputed genomic information in breed improvement strategies for the breed will be beneficial. The accuracy of GEBVs were not sensitive to imputation errors less than 3%, indicating that the development of a low-density panel for the South African Drakensberger would be an invaluable addition towards realizing low-cost, genomic selection for the breed.

References

Abdalla, E.E., Schenkel, F.S., Emamgholi Begli, H., Willems, O.W., Van As, P., Vanderhout, R., Wood, B.J. & Baes, C.F., 2019. Single-step methodology for genomic evaluation in turkeys (Meleagris gallopavo). Front. Genet. 10, 1248.

Abin, S.A., 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.

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S. & Lawlor, T.J., 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93(2), 743-752.

Aguilar, I., Misztal, I., Legarra, A. & Tsuruta, S., 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128(6), 422-428.

Aliloo, H., Mrode, R., Okeyo, A.M., Ni, G., Goddard, M.E. & Gibson, J.P., 2018. The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa. J. Dairy Sci. 101(10), 9108-9127.

Beefmaster Breeders' Society of SA & SA Stud Book, 2017. Joint media release: Genomic breeding values for South African Beefmaster cattle. Available online at: http://www.sastudbook.co.za/images/photos/News-Beefmaster-Genomic-EBVs.pdf.

Berry, D.P. & Kearney, J.F., 2011. Imputation of genotypes from low-to high-density genotyping platforms and implications for genomic selection. Anim. 5, 1162-1169.

Berry, D.P., Garcia, J.F. & Garrick, D.J., 2016. Development and implementation of genomic predictions in beef cattle. Anim. Front. 6, 32-38.

Cleveland, M.A., Hickey, J.M. & Kinghorn, B.P., 2011. Genotype imputation for the prediction of genomic breeding values in non-genotyped and low-density genotyped individuals. BMC Proc. 5, S6.

Cleveland, M.A. & Hickey, J.M., 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J. Anim. Sci. 91(8), 3583-3592.

Christensen, O.F. & Lund, M.S., 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42(1), 2.

Christensen, O.F., Madsen, P., Nielsen, B., Ostersen, T. & Su, G., 2012. Single-step methods for genomic evaluation in pigs. Anim. 6(10), 1565-1571.

Fernando, R.L., Dekkers, J.C. & Garrick, D.J., 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet. Sel. Evol. 46(1), 50.

Forni, S., Aguilar, I. & Misztal, I., 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genet. Sel. Evol. 43(1), 1.

Gao, H., Koivula, M., Jensen, J., Strandén, I., Madsen, P., Pitkänen, T., Aamand, G.P. & Mäntysaari, E.A., 2018. Genomic prediction using different single-step methods in the Finnish red dairy cattle population. J. Dairy Sci. 101(11), 10082-10088.

Garcia-Baccino, C.A., Legarra, A., Christensen, O.F., Misztal, I., Pocrnic, I., Vitezica, Z.G. & Cantet, R.J., 2017. Metafounders are related to F st fixation indices and reduce bias in single-step genomic evaluations. Genet. Sel. Evol. 49(1), 34.

Groeneveld, E., Kovac, M. & Mielenz, N., 2008. VCE 6.0.2. Co-variance components estimation package. Institute of Farm Animal Genetics, Mariensee, Germany.

ICBF, 2017. Ireland reaches 1 million genotype milestone. Available online at URL: https://www.icbf.com/wp/?p=8703.

Johnston, D.J., Ferdosi, M.H., Connors, N.K., Boerner, V., Cook, J., Girard, C.J., Swan, A.A. & Tier, B., 2018. Implementation of single-step genomic BREEDPLAN evaluations in Australian beef cattle. Proc. World Congr. Genet. Appl. Livest. Prod., Auckland, New Zealand. 269.

Legarra, A., Christensen, O.F., Aguilar, I. & Misztal, I., 2014. Single Step, a general approach for genomic selection. Livestock Sci. 166, 54-65.

Lidauer, M., Matilainen, K., Mäntysaari, E., Pitkänen, T., Taskinen, M. & Strandén, I., 2015. MiX99-Solving large mixed model equations. MiX99 Development Team, Biometrical Genetics. Natural Resources Institute Finland (Luke), FI-31600 Jokioinen, Finland.

Lourenco, D.A.L., Misztal, I., Wang, H., Aguilar, I., Tsuruta, S. & Bertrand, J.K., 2013. Prediction accuracy for a simulated maternally affected trait of beef cattle using different genomic evaluation models. J. Anim. Sci. 91(9), 4090-4098.

Lourenco, D.A.L., Tsuruta, S., Fragomeni, B.O., Masuda, Y., Aguilar, I., Legarra, A., Bertrand, J.K., Amen, T.S., Wang, L., Moser, D.W. & Misztal, I., 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. J. Anim. Sci. 93(6), 2653-2662.

Legarra, A., Christensen, O.F., Aguilar, I. & Misztal, I., 2014. Single Step, a general approach for genomic selection. Livestock Sci. 166, 54-65.

Martini, J.W., Schrauf, M.F., Garcia-Baccino, C.A., Pimentel, E.C., Munilla, S., Rogberg-Muñoz, A., Cantet, R.J., Reimer, C., Gao, N., Wimmer, V. & Simianer, H., 2018. The effect of the H− 1 scaling factors τ and ω on the structure of H in the single-step procedure. Genet. Sel. Evol. 50(1), 16.

Misztal, I. & Wiggans, G.R., 1988. Approximation of prediction error variance in large-scale animal models. J. Dairy Sci. 71, pp.27-32.

Misztal, I., Legarra, A. & Aguilar, I., 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92(9), 4648-4655.

Mrode, R., Ojango, J.M., Okeyo, A.M. & Mwacharo, J.M., 2018. Genomic selection and use of molecular tools in breeding programs for indigenous and crossbred cattle in developing countries: current status and future prospects. Frontiers Genet. , 9.

Mulder, H.A., Calus, M.P.L., Druet, T. & Schrooten, C., 2012. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J. Dairy Sci. 95, 876-889.

Nayee, N.G., Su, G., Gajjar, S.G., Sahana, G., Saha, S., Trivedi, K.R. & Guldbrandtsen, B., 2018. Genomic prediction by single-step genomic BLUP using cow reference population in Holstein crossbred cattle in India. In: Proceedings of the World Congress on Genetics Applied to Livestock Production. Volume 11, 411.

Purcell, S., Neale, B. & Todd-Brown, K., 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81, 559-575.

SA Stud Book, 2017. Media Release: Genomic Breeding Values for Bonsmara. Available online at: http://www.sastudbook.co.za/n16/general-news/media-release:-genomic-breeding-values-for-bonsmara.html.

Sargolzaei, M., Chesnais, J.P. & Schenkel, F.S., 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15, 478.

Song, H., Zhang, J., Zhang, Q. & Ding, X., 2018. Using different single-step strategies to improve the efficiency of genomic prediction on body measurement traits in pig. Frontiers Genet., 9, 730.

Strandén, I. & Vuori, K., 2006. RelaX2: pedigree analysis programme. In: *Proceedings of the 8th World Congress on Genetics Applied to Livestock Production*, Belo Horizonte, Minas Gerais, Brazil, 13-18 August 2006 (pp. 27-30).

van der Westhuizen, R.R., van der Westhuizen, J. & van Marle-Köster, E., 2017. Estimation of genomically enhanced estimated breeding values for SA beef cattle. In: Proceedings of the 50th congress of the South African Society for Animal Science. Port Elizabeth, South Africa, 18-21 September 2017.

van Marle-Köster, E. & Visser, C., 2018a. Genomics for the advancement of livestock production: A South African perspective. S. Afr. J. Anim. Sci. 48(5), 808-817.

van Marle-Köster, E. & Visser, C., 2018b. Genetic improvement in South African livestock: can genomics bridge the gap between the developed and developing sectors? Frontiers Genet., 9.

Weigel, K.A., Van Tassell, C.P., O'Connell, J.R., VanRaden, P.M. & Wiggans, G.R., 2010. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J. Dairy Sci. 93, 2229-2238.

Weller, J.I., Ezra, E. & Ron, M., 2017. Invited review: a perspective on the future of genomic selection in dairy cattle. J. Dairy Sci. 100(11), 8633-8644.

Wu, X.L., Xu, J., Feng, G., Wiggans, G.R., Taylor, J.F., He, J., Qian, C., Qiu, J., Simpson, B., Walker, J. & Bauck, S., 2016. Optimal design of low-density SNP arrays for genomic prediction: algorithm and applications. PloS one 11(9), e0161719.

CHAPTER SIX

CRITICAL REVIEW AND CONCLUSION

6.1 General discussion and recommendations

Efficient breed improvement ultimately relies on the ability to predict the genetic merit of individual animals so that the desired combination of alleles can be utilized to facilitate genetic progress in economically important traits (Jonas & de Koning, 2015). Traditionally, estimated breeding values (EBV) have been used as indicators of an animal's genetic merit for recorded traits within a breed. These EBVs are calculated using best linear unbiased prediction (BLUP; Henderson, 1984) methodology, which relies purely on performance data and pedigree-based relationship estimates. Meuwissen *et al*. (2001) first proposed the enrichment of EBVs with genomic information and the acceptance of this proposition was driven by the promise of more precise selection and accelerated genetic gain. The implementation of genomic selection was made possible largely because of advances made in genomic technologies. These advances enabled the sequencing of draft genomes, which were subsequently used as references to develop SNP panels consisting of the most informative markers. Currently, these SNP panels form the basis of genomic selection pipelines and for various other research purposes (Gurgul *et al*., 2014).

The genotyping panels that have been released commercially have been designed in such a way that their SNP content has optimal utility in specific breeds and, needless to say, this included the most popular and abundant breeds worldwide. The SNPs that were used to construct the first SNP panel for cattle, the Illumina® Bovine SNP50 (Matukumalli *et al*., 2009), were selected to be highly informative in *Bos Taurus* breeds. Subsequent commercial panels improved on this panel, in terms of SNP density and content, and later genotyping panels focusing specifically on the most prominent *Bos Indicus* breeds were also developed (e.g. GeneSeek® Genomic Profiler^{TM} Indicus; Ferraz *et al*., 2018). Higher density genotyping panels facilitated research ventures into high-resolution genomic characterization of taurine- and indicince breeds including the profiling of genome-wide patterns of linkage disequilibrium (LD; e.g. Khatkar *et al*., 2008), runs of homozygosity (ROH; e.g Purfield *et al*., 2012) and signatures of selection (SoS; e.g. Zhao *et al*., 2015). Although genotyping at such high SNP densities (e.g. Illumina® 777K) can provide significantly more information in certain situations, for instance in genome characterization and GWAS, the cost of genotyping, despite its reduction over the years, can render it unfeasible in routinely applied methods such as GS. Research ventures into GS methodology led to the realization that not all of the SNPs included on higher density panels are necessary to achieve accurate genomic breeding values (GEBVs). This in turn led to the exploration of genotype imputation from more affordable lower density genotyping panels with the view of implementing low-cost GS.

Applying low-cost GS will be an invaluable addition to traditional breed improvement programs involving indigenous cattle, especially in the developing world where most of these breeds are located (Mrode *et al*., 2018). A major concern, and the main driving force behind the motivation of this thesis, was that no genotyping panel with specific utility in globally uncommon and unique, indigenous breeds existed for such purposes. Many of these breeds are genetically distinct and harbour genomic diversity originating from both taurine and indicine ancestors; this genomic diversity facilitates efficient adaptation to different environments (Kim *et al*., 2019). Breeds of the Sanga subspecies are prime examples. Many inferences have been made about the history of these breeds on the African continent, including possible migration routes (e.g. Hanotte *et al*., 2002) but, despite this, the precise genetic composition of many of these breeds remains unresolved (Mwai *et al*., 2015). Previous efforts to characterize these cattle on the genome level have typically included too few animals genotyped on too sparse genotype densities to make detailed deductions on the genomic architecture of Sanga breeds. Inferences made from these efforts have, however, been sufficient to confirm expected genomic heterogeneity by studying shared germplasm with exotic breeds (e.g. Makina *et al*., 2016).

One of the explicit objectives of the present research was to determine whether imputed SNP genotypes of the SA Drakensberger can be reliably used in genome-based breeding programs. The first objective was therefore to elucidate the usefulness of SNPs included on a commercial 150 000-SNP panel for the breed. Makina *et al*. (2015) suggested this number of SNPs to be more appropriate for the breed than the SNPs on the initial 50K Illumina® panel in order to adequately capture linkage disequilibrium (LD) patterns to enable genomic selection (r^2=0.2; Meuwissen *et al*., 2001). Breeds participating in the BGP were therefore routinely genotyped on this density. Albeit this density was sufficient to capture LD persisting for SNP pairs separated by larger genomic distances than previously suggested using 50 000 SNPs (~30kb versus <20kb), the persistence of LD was still weak. This was not surprising as shorter haplotypes are shared within breeds that are characterized by genomic diversity (Toosi *et al*., 2010). Admixture in the SA Drakensberger genome was introduced centuries ago; for older breeds, more opportunities have occurred for recombination events to disrupt LD that was present in the breed's ancestors (Toosi *et al*., 2010). This was supported by the abundance of short ROH ($ROH_{2\text{-}4Mb}$ = 35.7%) observed in the SA Drakensberger genome (Chapter 3), which could have originated from the disruption of longer segments by recombination events (Purfield *et al*., 2012). The higher frequency of short homozygous haplotypes and weak genome-wide LD raised concerns about the achievable imputation accuracy (as well as the accuracy of genomic evaluations).

It can be a difficult endeavour to pinpoint the precise SNP density and characteristics that a lower density genotyping panel should comprise of to facilitate accurate imputation to higher densities, as there are many determining factors involved. In terms of SNP density, however, rough estimates or ideals can be determined based on the extent of LD. Because larger genomic segments are shared between animals, populations characterized by strong LD may be genotyped on considerably fewer SNPs. As few as 300 to 400 SNPs have, for example, been deemed adequate to achieve minimal loss in imputation accuracy for swine (Grossi *et al*., 2018) and poultry (Wang *et al*., 2013) breeds, however, this could also be as a result of their population structure; larger segment are shared within sub-populations because of line breeding for instance. Taking into account the weak persistence of LD observed in Chapter 3, despite improvements in SNP density and sample sizes compared to previous research (Makina *et al*., 2015), a higher minimum requirement of SNPs was expected. Using the calculated imputation accuracies as a guideline, it could be concluded that a genotyping panel consisting of at least 10 000 SNPs would be appropriate to achieve a mean of <3% imputation errors for the SA Drakensberger. A panel of this density should not, however, be derived by selecting SNPs at random or using solely their locality as a selection criterion if the aim is to achieve high imputation accuracy. Using random selection to derive a 10 000 SNP genotyping panel resulted in the worst mean animal-wise imputation accuracy, compromising 3.9% in imputation accuracy, and this was partly because it had the most variable SNP distribution, with consecutive SNPs separated by up to 4477kb, and partly because of lower mean MAF. The acceptable mean imputation accuracy will vary between studies, and will depend on the target genomic application in which imputed genotypes will be employed. The mean imputation accuracy should, however, at least be sufficiently high to minimize the influence on downstream estimates, such as GEBVs. In this study, a threshold of 3% imputation inaccuracies was deemed acceptable. Using a panel whereby a mean imputation accuracy of 97% could be achieved, had minimal influence on GEBV estimation.

The development process of a 10K low-density panel for the SA Drakensberger needs to prioritize the selection of highly polymorphic SNPs to insure that the selected SNPs are indeed segregating within the breed. Minor allele frequency is an important selection criterion; when considering only SNPs of moderate to high MAF (>20%) a higher proportion (+7%) of SNP pairs displayed LD ≥ r^2=0.2 (Chapter 3) i.e. low MAF has a diminishing effect on localized LD (Qanbari *et al*., 2010). Attributable to the ascertainment bias in the development of commercial SNP panels, a high proportion of low-MAF SNPs were observed in this study across the genome (up to 16.6% on BTA14) and this supported previous results for Sanga cattle that included small numbers of SA Drakensberger samples (Qwabe *et al*., 2013; Zwane *et al*., 2016). Identifying informative SNPs can become a challenge if the pool of candidate SNPs to select from is limited to those only included on commercially available higher density panels. The selection of SNPs should ideally be distributed evenly across chromosomes and the genome in general. To achieve this, whilst maximizing MAF, the selection of SNPs were limited to markers occurring within genomic segments of a fixed length, with the size of segments becoming smaller for denser genotyping panels.

In the present study, the challenge faced with methods selecting SNPs within segments of pre-defined length was firstly that certain chromosomes harboured relatively large gaps i.e. regions where no SNPs were mapped. Prior to any quality control procedures employed, a genomic region spanning approximately 2.3 megabase pairs (Mb) and harbouring no SNPs was observed on BTA12. This meant that no SNPs could be selected in this region. The second concern with this selection strategy was that certain segments only harboured a single SNP and this SNP was therefore chosen by default, with no regard to MAF or LD with neighbouring SNPs. Across the segment-based strategies of selecting SNPs for the low-density panel, the same SNP was chosen for higher density panels and this would explain some of the reduced variation in imputation accuracy observed at higher densities. The number of segments containing either a single SNP or no SNPs at all increased significantly for the selection of SNP densities higher than 10 000 SNPs (e.g. 20 000 SNP: 167 no-SNP segments; 50 000 SNP: 2 682 no-SNP segments). This suggests that even though the higher density genotyping panel employed in this study had utility, it was not optimal as a resource of SNPs to select from; it would be more useful to use breed-specific SNPs identified in genome re-sequencing efforts for future panel design process. The SA Drakensberger was included in a recent study that aimed to re-sequence the genomes of three indigenous cattle breeds and data generated may serve as a valuable resource in SNP selection endeavours. The partitioning-around-medoids (PAM; Kaufman & Rousseeuw, 2009) method of selection undertaken in this study has not been previously evaluated as a possible SNP selection strategy. The employment of this method can be used to bypass the limitation of selection to fixed segments of equal size and serve as an alternative method of selecting evenly spaced SNPs. Attributes of MAF and LD can furthermore be calculated within clusters produced. Implementation of the PAM algorithm was, however, computationally demanding and it is recommended that this method be employed on a per chromosome basis for the selection of more that 20 000 SNPs.

The final part of this thesis aimed to validate the utility of the proposed lower density panel in imputation-driven estimation of genomic breeding values. Two separate genotypic data sets were used, one consisting of the 120 608 SNPs that were actually genotyped i.e. true genotypes and another data set that consisted of the SNPs included on the 10K low-density panel and the remaining 110 608 SNPs imputed, to derive GEBVs. In addition to the genomic enhanced breeding values, traditional breeding values were also estimated using pedigree information alone. Results indicated that the single-step genomic prediction strategy was a valid strategy for the SA Drakensberger, and that improvements in the accuracy of breeding value estimation could be achieved when genomic information was included. Improvements were more significant for lowly heritable traits, i.e. the maternal components of the growth traits studied (e.g. 3.4% for $BW_{maternal}$ and 3.1% for $WW_{maternal}$), and further improvements are expected when larger sample sizes of genotyped animals become available. Previous research have documented greater effects of genomic predictions for traits that are measured post-weaning or through indicator traits as these traits are less frequently recorded (MacNeil, 2016) and therefore the improvements for the traits studied here were expected to be small. Van der Westhuizen *et al*. (2017) observed slightly higher improvements for low-heritability and hard-to-measure traits (between 5% and 15%) when a sample of approximately 2 500 locally developed SA Bonsmara animals were genotyped. It should, however, be noted that the phenotypic records for the SA Bonsmara breed is more extensive for lowly heritable traits and, being the most numerous breed in SA, that the population size of the SA Bonsmara breed is ten times that of the SA Drakensberger (SA Studbook, 2016).

Lastly, because prediction equations estimate the effect of all SNPs simultaneously, a loss in predictive accuracy because of few wrongly imputed SNPs was expected to be negligible based on previous research done on admixed dairy cattle. This impacts the downstream practicality of incorporating imputation into GS pipelines routinely. If minor imputation inaccuacies are truely negligible, and a mean imputation accuracy of 97% can be maintained, selection candidates can routinely be genotyped on a low-density panel and imputation performed with each routine BLUP evaluation. This was corroborated in this study by perfect correlations (P<0.001) estimated between GEBVs derived from true versus imputed SNPs. Imputation is expected to have a more pronounced effect on applications such as GWAS that rely on SNP-by-SNP association testing (Badke *et al*., 2013). The presence of large numbers of rare SNPs, which are expected to be the basis of adaptive mechanisms, in the SA Drakensberger genome and its influence on imputation will also need further consideration.

6.2. Recommendations

Results from this study may be used as guidelines to aid in the process of designing a low-density panel for the SA Drakensberger in the future. Inferences made from this study may be transferable to other Sanga breeds and may serve in guiding future genomic endeavours for these breeds and other breeds that have admixed genomes. If the development of a custom low-density genotyping panel can be set into motion, it will enable routine genomic evaluations for this breed without incurring a high cost. Alternatively, the utility of existing commercial low-density panels (e.g. avaiable through Illumina® and GeneSeek®, listed in Chapter 2) should be examined by quantifying the accuracy of imputation from these panels to the 120 608 SNP panel; this might exclude the cost of developing a new panel if an existing panel is efficient. Re-genotyping subsets of animals with these lower density panels will, however, present a financial constraint as initially, financial resources would rather be focused towards improving the sample of animals genotyped at high densities to assemble an appropriate training population. A further recommendation in such an endeavour would be to compare the achievable imputation accuracy of alternative imputation software (also listed in Chapter 2), as previous studies have reported noticeable differences in the performance of software especially for breeds with complex genetic structure. Lastly, given the lower imputation accuracy in certain regions of the genome, including the autosomal extremities, it might be beneficial to explore alternative SNP selection strategies such as the multiple-objective, local optimization (MOLO) algorithm (Wu *et al*., 2016) as well, or to design a new selection strategy.

Although this study provided a baseline demonstration of the practicality of using imputation, it is recommended that the short- and long-term economic benefits be relayed to farmers in monetary values. Certain breeders are still resistant to the uptake of genomic technologies. Providing a proven genotyping strategy that is more cost-effective may therefore serve as an incentive to increase the utilization of these technologies to benefit not only individual breeders, but also the entire breed. If more animals are genotyped in the future: 1) imputed genotypes will become more accurate and 2) the

H-matrix applied in single-step genomic evaluation can be extended to improve GEBV accuracies. Both of these advantages will accelerate genetic progress within the breed. Initially, genotyping efforts should be focused towards high-density genotyping of animals that are representative of the breed, i.e. high-impact animals, and reduced-density genotyping of selection candidates; a higher degree of relatedness between animals genotyped on high- versus reduced-density panels will further improve imputation accuracy for selection candidates.

Although reduced densities of SNPs could be chosen from the high-density panel in this study, and resulted in accurate imputation, whole-genome sequncing data would provide a more valuable resource to identify breed-specific SNPs that might assist in locating adaptive genes. Sequencing data generated by the BGP could serve as a resource for a more detailed investigation into the unique genetic composition of breeds such as the SA Drakensberger. Future research efforts should aim to elucidate and quantify the proportion of genomic segments within the SA Drakensberger genome that are derive from taurine and indicine ancestor. In the meantime, it is advised that the utility of imputed SNPs be explored for genome-wide association studies; this methodology can be used to boost SNP numbers, which will add power to these studies performed on the SA Drakensberger breed. These studies might assist in uncovering the genetic architecture of traits of interest to the SA beef industry and these include traits involved in adaptation, which will be important amidst climate change and consequent changes to beef production environments. Considering that no Sanga-specific genotyping panel currently exists, it would be recommended that lower-density SNPs be chosen from re-sequencing efforts in the future, i.e. from a pool of SNPs that are identified as specific to the breed, and not necessarily from a pool of SNPs that are available on taurine- and/or indicine-derived genotyping platforms. This will insure the sampling of SNPs that are segregating within the breed(s) of interest and might alleviate the influence of low-MAF on estimates of LD and imputation accuracy. The obstacle of gaps in the genome, identified between mapped SNPs during the execution of selection strategies, can also hereby be bypassed, as there will be a higher coverage of markers to choose from.

The validity of imputed genotypes in genomic prediction should be properly validated when more genotypes become available; proper validation could not be carried out in this study because of limited sample size. To properly validate the accuracy of genomic prediction, a cross-validation approach is usually followed. In a typical five-fold cross-validation, for example, five different sets of reference and validation animals are generated, with no overlap between animals in the separate validation sets, using the animals with imputed genotypes (Tsai *et al.*, 2017). Phenotypic records are then masked for the validation animals and subsequently predicted from each animal's GEBVs (Tsai *et al.*, 2017). Due to the small number of genotyped animal available for the SA Drakensberger, compared to the numbers available internationally for more popular breeds, this type of cross-validation was not possible considering that no overlapping between validation sets are allowed; validation sets would have been too small for this method of validation to be scientifically sound. For the purpose of this thesis, and given the amount of data available, the procedures followed here were adequate to investigate the utility of imputed genotyped in GEBV estimation.

6.3. Conclusion

The variation observed in genomic characteristics such as MAF and LD conformed to expectations and supported previous research suggesting that the SA Drakensberger is a composite breed with an admixed genome and heterogeneous genomic architecture. This variation across the genome enabled the variability in imputation accuracy across chromosomes and genomic regions within chromosomes to be pre-empted. Because negligible gain was observed for MAF and LD estimates using a higher density panel, in comparison to the 50K panel used in previous studies, it was concluded that a lower density panel, combined with imputation to higher density, would suffice towards implimenting GS. Genotype imputation is a valid genomic strategy for the SA Drakensberger breed and this study concluded that a genotyping panel consisting of approximately 10 000 SNPs would suffice in achieving less than 3% imputation errors. Results also suggests that if such a panel were to be designed, that the SNPs considered for inclusion should be selected based on selection criteria, such as MAF and LD, specific to the SA Drakensberger breed in order to maximize achievable imputation accuracy. This study showed that it would be a valid strategy to integrate genotype imputation routinely into single-step genomic evaluation pipelines for the SA Drakensberger breed as imputation errors proved to have a negligible effect on resulting GEBV accuracies for growth traits.

References

Badke, Y.M., Bates, R.O., Ernst, C.W., Schwab, C., Fix, J., Van Tassell, C.P. & Steibel, J.P., 2013. Methods of tagSNP selection and other variables affecting imputation accuracy in swine. BMC Genet. 14, 8.

Grossi, D.A., Brito, L.F., Jafarikia, M., Schenkel, F.S. & Feng, Z., 2018. Genotype imputation from various low-density SNP panels and its impact on accuracy of genomic breeding values in pigs. Anim. 12(11), pp.2235-2245.

Gurgul, A., Semik, E., Pawlina, K., Szmatoła, T., Jasielczuk, I. & Bugno-Poniewierska, M., 2014. The application of genome-wide SNP genotyping methods in studies on livestock genomes. J. Appl. Genet. 55(2), 197-208.

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. Sci. 296(5566), 336-339.

Henderson, C.R., 1984. Applications of linear models in animal breeding. Volume 462. Guelph: University of Guelph.

Jonas, E. & Koning, D.J.D., 2015. Genomic selection needs to be carefully assessed to meet specific requirements in livestock breeding programs. Frontiers Genet. 6, 49.

Kaufman L. & Rousseeuw, P.J., 2009. Finding groups in data: an introduction to cluster analysis. John Wiley & Sons; 2009 Sep 25.

Khatkar M., Nicholas F., Collins A., Zenger K., Cavanagh J., Barris W., Schnabel R., Taylor J. & Raadsma H., 2008. Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genomics 9, 187.

Kim, J., Hanotte, O., Mwai, O.A., Dessie, T., Bashir, S., Diallo, B., Agaba, M., Kim, K., Kwak, W., Sung, S. & Seo, M., 2017. The genome landscape of indigenous African cattle. Genome Biol. 18(1), 34.

MacNeil, M.D., 2016. Value of genomics in breeding objectives for beef cattle. Revista Brasileira de Zootecnia, 45(12), 794-801.

Makina, S.O., Taylor, J.F., Van Marle-Köster, E., Muchadeyi, F.C., Makgahlela, M.L., MacNeil, M.D. & Maiwashe, A., 2015. Extent of linkage disequilibrium and effective population size in four South African Sanga cattle breeds. Front. Genet. 6, 1-12.

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.

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, e5350-5063.

Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genet. 157, 1819-1829.

Mrode, R., Ojango, J.M., Okeyo, A.M., & Mwacharo, J.M., 2018. Genomic selection and use of molecular tools in breeding programs for indigenous and crossbred cattle in developing countries: current status and future prospects. Front. Genet. 9, 694.

Mwai, O., Hanotte, O., Kwon, Y.J. & Cho, S., 2015. African indigenous cattle: unique genetic resources in a rapidly changing world. Asian-Australasian J. Anim. Sci. 28(7), 911.

Purfield, D.C., Berry, D.P., McParland, S. & Bradley, D.G., 2012. Runs of homozygosity and population history in cattle. BMC Genet. 13, 70.

Qanbari, S., Pimentel, E.C., 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.

Qwabe, S.O., Van Marle-Köster, E., Maiwashe, A. & Muchadeyi, F.C., 2013. Evaluation of the BovineSNP50 genotyping array in four South African cattle populations. S. Afr. J. Anim. Sci. 43, 64-67.

SA Stud Book, 2016. SA Stud Book annual report. Available online at:http://www.sastudbook.co.za/images/photos/Annual_Report_2016_a.pdf.

Toosi, A., Fernando, R.L. & Dekkers, J.C.M., 2010. Genomic selection in admixed and crossbred populations. J. Anim. Sci. 88(1), 32-46.

Tsai, H.Y., Matika, O., Edwards, S.M., Antolín–Sánchez, R., Hamilton, A., Guy, D.R., Tinch, A.E., Gharbi, K., Stear, M.J., Taggart, J.B. & Bron, J.E., 2017. Genotype imputation to improve the cost-efficiency of genomic selection in farmed Atlantic salmon. G3: Genes Genom. Genet. 7, 1377-1383

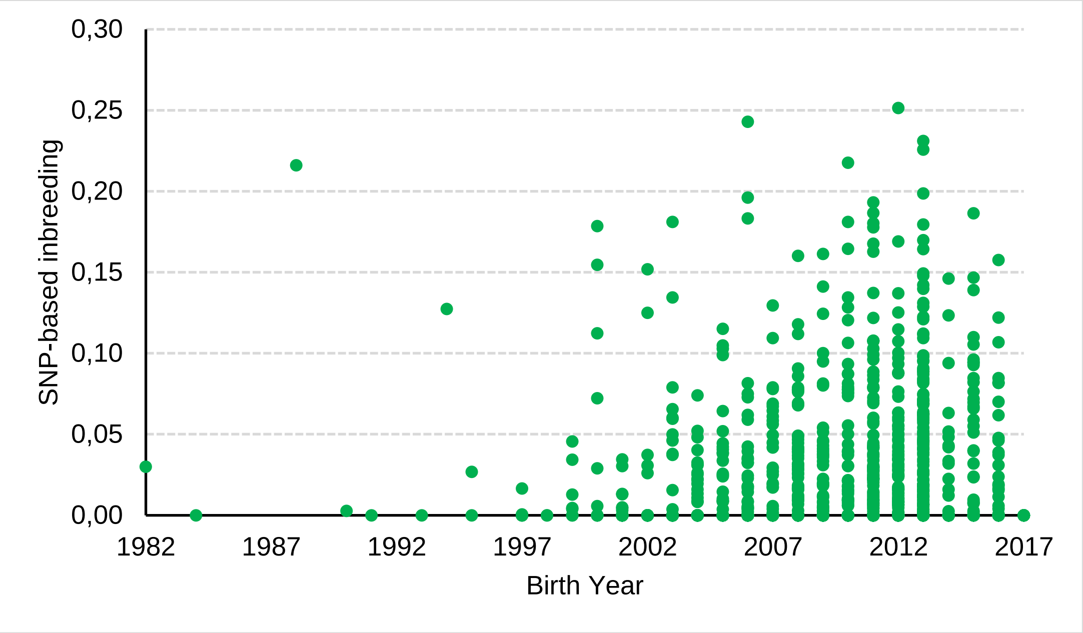
Van der Westhuizen, R.R., van der Westhuizen, J. & van Marle-Köster, E., 2017. Estimation of genomically enhanced estimated breeding values for SA beef cattle. In: Proceedings of the 50th congress of the South African Society for Animal Science. Port Elizabeth, South Africa, 18-21 September 2017.

Wang, Y., Lin, G., Li, C. & Stothard, P., 2016. Genotype imputation methods and their effects on genomic predictions in cattle. Springer Sci. Rev. 4, 79-98.

Zhao, F., McParland, S., Kearney, F., Du, L. & Berry, D.P., 2015. Detection of selection signatures in dairy and beef cattle using high-density genomic information. Genet. Sel. Evol. 47(1), 49.

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.

Addendum 1 Genome-wide single nucleotide polymorphism based inbreeding coefficients by birth year across the SA Drakensberger population sampled



Addendum 2 The number of single nucleotide polymorphisms that were selected per chromosome for the 2 500 (2.5K), 5 000 (5K), 10 000 (10K), 20 000 (20K) and 50 000 (50K)-marker low-density genotyping panels.

Chromosome	Chromosome length (Mb)	Number of single nucleotide polymorphisms					
		HD[1]	2.5K	5K	10K	20K	50K
1	158.21	7371	152	305	610	1222	3166
2	136.62	6455	133	267	535	1070	2734
3	121.38	5796	120	240	481	961	2428
4	120.55	5582	115	231	463	925	2414
5	121.14	6041	125	250	501	1001	2417
6	119.40	6626	137	274	549	1098	2385
7	112.60	5705	118	236	473	946	2230
8	113.35	5157	106	213	428	855	2254
9	105.59	4948	102	205	410	820	2096
10	104.23	4891	101	202	406	811	2075
11	107.24	5028	104	208	417	833	2137
12	91.10	4211	87	174	349	698	1797
13	84.20	3924	81	162	325	650	1678
14	84.03	4701	97	194	390	779	1679
15	85.20	3984	82	165	330	660	1700
16	81.65	3741	78	156	310	621	1623
17	75.11	3446	72	143	286	572	1491
18	65.87	3073	64	128	255	510	1313
19	63.89	2935	61	122	243	487	1269
20	71.88	3799	78	158	315	630	1440
21	71.53	3338	70	139	277	554	1416
22	61.24	2901	61	121	241	482	1226
23	52.45	2469	52	103	205	410	1045
24	62.59	3377	70	140	280	560	1254
25	42.76	2003	42	84	166	333	861
26	51.58	2469	52	103	205	410	1029
27	45.35	2127	45	89	176	353	899
28	46.24	2131	45	89	177	354	922
29	51.17	2379	50	99	197	395	1022

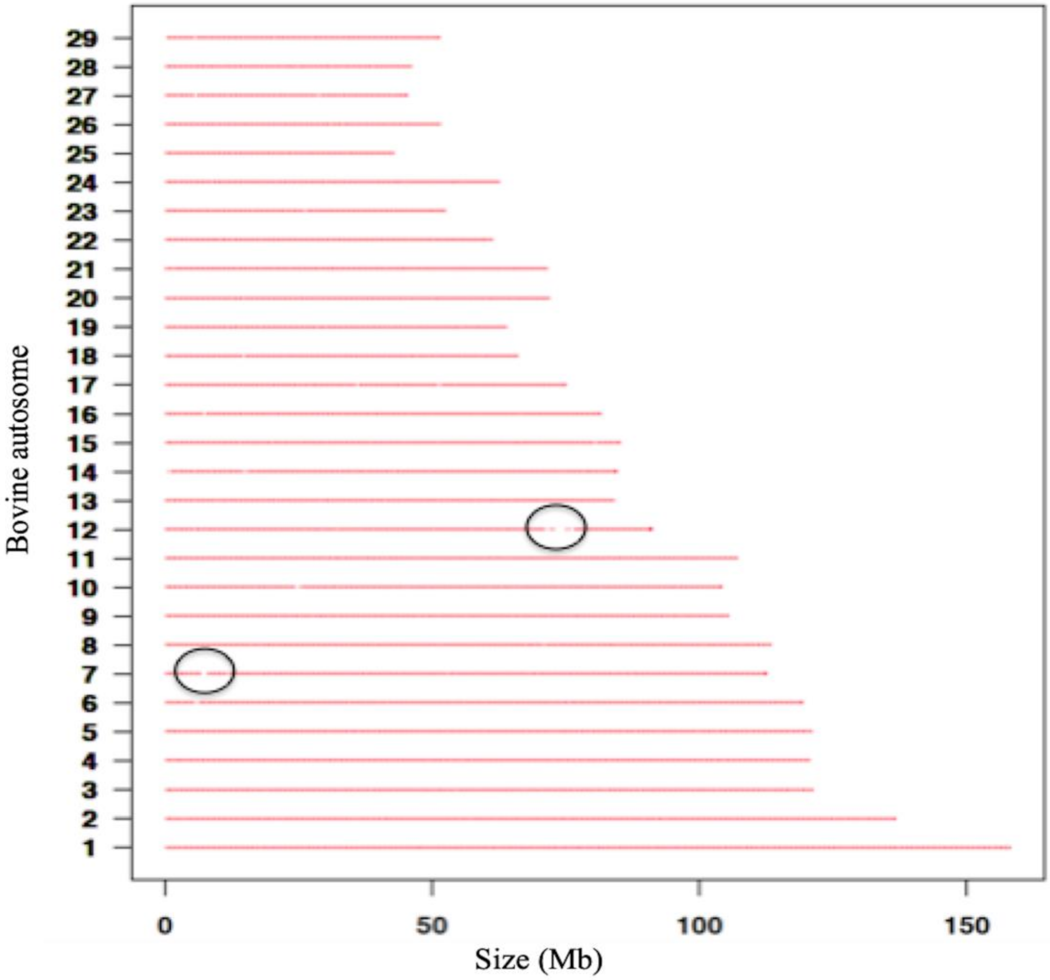
[1]HD=120 608 SNPs that remained after quality control procedures

Addendum 3 Mean COR_{SNP} per autosome for SNPs located on the autosomal extremities and within the centre of the autosomes when the worst (RAN) and best (MAFLD) SNP selection methods were used to derive 10 000 SNPs. Autosomal extremities were defined as first and last 1Mb of each autosome whereas autosomal centre was defines as 0.5Mb to either side of the physical midpoint of each autosome.

Chromosome	RAN[1]			MAFLD[2]		
	First	Centre	Last	First	Centre	Last
1	0.829	0.904	0.901	0.875	0.942	0.918
2	0.937	0.934	0.882	0.945	0.950	0.910
3	0.861	0.966	0.928	0.916	0.978	0.914
4	0.826	0.956	0.879	0.866	0.943	0.912
5	0.916	0.940	0.863	0.953	0.945	0.923
6	0.973	0.963	0.899	0.963	0.954	0.919
7	0.872	0.950	0.898	0.845	0.955	0.930
8	0.898	0.943	0.743	0.938	0.970	0.888
9	0.947	0.910	0.903	0.986	0.937	0.927
10	0.884	0.942	0.946	0.910	0.941	0.945
11	0.880	0.949	0.769	0.930	0.961	0.802
12	0.909	0.871	0.916	0.926	0.938	0.911
13	0.928	0.936	0.881	0.939	0.944	0.929
14	0.283	0.919	0.918	-[3]	0.945	0.904
15	0.883	0.919	0.917	0.881	0.926	0.951
16	0.861	0.960	0.844	0.914	0.976	0.894
17	0.946	0.915	0.913	0.947	0.959	0.933
18	0.876	0.937	0.907	0.879	0.916	0.858
19	0.961	0.907	0.928	0.952	0.941	0.915
20	0.879	0.891	0.912	0.935	0.968	0.934
21	0.844	0.932	0.896	0.874	0.966	0.898
22	0.897	0.911	0.905	0.925	0.939	0.936
23	0.988	0.908	0.874	0.983	0.970	0.911
24	0.911	0.927	0.960	0.885	0.955	0.935
25	0.893	0.885	0.904	0.955	0.924	0.911
26	0.907	0.775	0.924	0.908	0.809	0.919
27	0.921	0.894	0.915	0.921	0.912	0.932
28	0.940	0.883	0.942	0.934	0.931	0.929
29	0.944	0.937	0.789	0.956	0.958	0.858

[1] RAN=random selection; [2] MAFLD=combinative selection for MAF and LD; [3] No SNPs were mapped to the first 1Mb of BTA14 after quality edits.

Addendum 4 Figure illustrating the SNP density per chromosome and indicating gaps in the genome prior to quality edits.



Addendum 5 Popular science article: Using imputation to save on genotyping costs for indigenous cattle. Published in: Drakensberger Breeders' Society Newsletter, December 2017 (available online: https://www.drakensbergers.co.za/files/150383_DrakensbergerNuusbrief%20-%20Finaal.pdf.)

The South African beef cattle industry is a complex one, consisting of highly diverse breeds, production systems and, subsequently, breeding goals. In conjunction with the rest of the world, the country is faced with two major future challenges - global warming and growing human population sizes. Indigenous cattle breeds, such as the Drakensberger, have endured the diverse and sometimes harsh environmental conditions of South Africa and therefore display superior adaptability. Their ability to thrive in these conditions make them an asset in assuring future sustainability of the beef industry and it will become increasingly important to conserve and use these animal resources during tough times. Despite their importance locally, there have been few efforts to delve into the genetic architecture of these cattle for the purpose of improving individual breeds until recently.

The evolution of "genomics" has provided an opportunity to increasingly incorporate DNA information into breed improvement programs. In genetics, the area of genomics involves looking at the DNA content of an animal in its entirety. Going back to the basics: DNA is a molecule carrying the genetic make-up of an animal. These molecules are passed on from parents to offspring and explain why one individual appears or performs different from another. Looking at an animal on the DNA-level therefore provides us with a high-resolution view of why that animal grows, reproduces, behaves or generally performs in a certain way. If we have this most basic information, we can significantly improve how accurately we choose which animals are the best for specific situations.

With the advent of genomics, came improvements in technology that have allowed scientists to holistically determine the genetic code of an animal by means of whole-genome sequencing. This in turn has allowed the identification of specific genetic markers in the genome, in the form of single nucleotide polymorphisms (SNPs), which can be incorporated into marker panels to test animals for specific purposes. SNPs explain genetic variation between animals and can be used for several applications such as to assign parents or to identify DNA regions responsible for variation within certain animal traits (such as growth, milk yield or the presence/absence of horns). These marker panels, referred to as "SNP chips", have progressively been modified and updated for specific purposes either by including more markers or by retaining smaller numbers of markers for specific purposes. Each unique panel has been released as either commercially- or privately-available products for research or personal use.

In recent years, two SNP-based genomic applications have received a lot of focus in the field of animal breeding and genetics, and these are genome-wide association studies (GWAS) and genomic selection (GS). GWAS is a method that uses associations between SNP- and performance data to find regions of the genome harboring genes that are of importance to specific traits. This can, for example, assist researchers in identifying which parts of the animal's DNA causes that animal to grow faster, give more milk or be more adapted to a certain environment. GS, on the other hand, is a method whereby animals can be selected based on an estimate genomic breeding value (GEBV). A GEBV is the genetic merit of an animal for a specific trait based on all of the SNPs included on a dense maker panel.

The above-mentioned applications, amongst others, can help us to significantly improve the Drakensberger breed on the genetic level; these strategies, however, require high densities of SNP data for large numbers of animals in order to make reliable scientific deductions. Generating the amount of data required on this scale can be costly, especially in developing countries, and this would call for multiparty collaborations for funding or cost saving alternatives. Genotype imputation presents one such strategy.

Imputation is a statistical method that predicts missing information. This method uses specific patterns observed in a data set containing complete information to fill in the gaps within another data set containing incomplete information. For example: we have a young animal tested for 10 000 markers (which would be referred to as a "low-density SNP panel") and the parents of this animal are tested for 100 000 markers (which would be referred to as a "high-density SNP panel"). Given the genetic relationship between the parents and the offspring, we can "impute" or infer the "missing" 90 000 markers for the young animals by making certain statistical assumptions. On a larger scale: If a "reference" population (consisting of high-impact animals with many offspring in the national herd) is genotyped for a high density of genetic markers (let's say 150K) and a "test" population (commercial animals in the national herd) is genotyped for a smaller subset of these SNPs (let's say 50K), the assumption is that these populations should, if they are related in some way, share an underlying genetic pattern. This shared genetic structure, which would be unique to the Drakensberger breed, can be used to predict missing genetic information in the test population from the genetic structure of the reference population. It will therefore only be necessary to genotype the test population (commercial animals) with a small subset of the SNPs. Given the fact that the low-density panel will be a lot cheaper, imputation is therefore a cost-saving approach and provides the means of getting more information at lower costs.

This method will enable researchers to obtain high densities of genetic marker information that would allow applications such as GWAS and GS in indigenous cattle breeds. Before this strategy can be applied, however, we need to assure that these predictions can be done accurately and reliably. The project currently underway that is testing the validity of the strategy for Drakensberger cattle will therefore aim to optimize this method for indigenous breeds. This will entail testing the effects that the number of animals, genetic relatedness between animals and certain inherent genetic characteristics of the breed will have on how accurately marker genotypes can be imputed.