Genotype imputation for indigenous beef cattle

10/31/2019

Genotype imputation as a genomic strategy for the South African Drakensberger beef breed

Industry Sector: Cattle And Small Stock

Research Focus Area: Livestock Production With Global Competitiveness: Breeding,Physiology And Management

Research Institute: Department Of Agriculture Forest And Fisheries (DAFF)

Year Of Completion : 2019

Researcher: Carina Visser

The Research Team

Title	Initials	Surname	Highest Qualification	Research Institution
Dr	M.M.	Scholtz	PhD	ARC-AP
Prof	E	van Marle-Koster	PhD	UP
Mrs	A.	Theunissen	MSc	Vaalharts Research Station

Executive Summary

The SA Drakensberger is a medium-framed breed with a sleek, black coat. Considering its history as one of the oldest indigenous breeds, its prominent role in the present beef industry and its potential for improving the beef cattle gene pool in the future; there is value in characterizing the SA Drakensberger on the genomic level. There has recently been interest in incorporating genomic information into selection strategies for this breed. Apart from the fact that the implementation of genomic technologies relies on diligent phenotyping efforts, accurate and complete pedigree recording; genomic selection also requires adequate SNP genotyping profiles (Meuwissen *et al*., 2001). The SA Drakensberger meets the requirements for genomic selection with 100% participation in SA Stud Book's *Logix Beef* performance recording scheme as well as an extensive recorded pedigree profile (SA Stud Book, 2017). Theoretically, current EBVs can therefore be enhanced with the use of genomics if financial resources allow the generation of adequate high-density genotypic profiles. 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). The main advantage of this methodology is the reduction in genotyping costs by allowing genotyping to be undertaken using lower density SNP panels. The utility of such low-density panels for applications such as

genomic selection will depend on the accuracy with which un-genotyped SNPs can be imputed to higher density from such lower density panels. Even though imputation is integrated into routine genomic evaluations internationally, the utility of this methodology has not been evaluated for indigenous cattle resources. Considering that these breeds often have admixed genomes, applying imputation requires optimization for such breeds and this includes the SA Drakensberger.

Objective Statement

The objective of this research project was to comprehensively study the validity of genotype imputation, from lower-density single nucleotide polymorphism (SNP) panels to higher density, for the economically-important SA Drakensberger beef cattle breed towards cost-effectively implementing genomically-enhanced breed improvement strategies such as genomic selection for this indigenous breed in the future.

Project Aims

1. To evaluate whether the Celtic mutation on the POLL locus is the causative mutation for polledness in Bonsmara and Drakensberger
2. To perform a genome wide association study of the Polled and Scur genes based on phenotypic data and genotypic data from the GGP Bovine 150K SNP bead chip
3. To apply sequence data available from the Bovine Genomics Program to finemap the suspected regions for the Polled and Scur genes

Results

Results generated from the first part of this study indicated that differences in genomic characteristics such as minor allele frequency (MAF), linkage disequilibrium (LD) and runs of homozygosity (ROH) exists between chromosomes. Mean genome-wide MAF was, for example, estimated to be 0.26 with chromosome-specific MAF ranging from 0.24 (Bos Taurus Autosome; BTA14) to 0.28 (BTA21). This was supported by the proportion of low-MAF (< 5%) SNPs estimated, which indicated 16.0% of SNPs to be classified as low-MAF SNPs on BTA14. The inter-SNP LD was generally weak, ranging from mean 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<30 kb. LD was weaker between SNP pairs including low-MAF SNPs. Consensus ROH segments were identified and the most prevalent of these occurred on BTA14 and was identified in ~23% of the sampled population. The ROH length characteristics furthermore pointed towards more ancient inbreeding, reflecting known historic bottleneck events.

For the second and main object preliminary results were generated to understand the necessary dynamics, in terms of size and composition, of an appropriate sub-population to use as a reference for estimation of haplotypes to be imputed from. Initial results indicated that a larger reference population would improve imputation accuracy. For example, it was observed that a 4% increase in imputation accuracy could achieved when the ratio of reference:test population was 90:10 versus 75:25; imputation accuracy improved from 0.981 (range: 0.895-0.997) to 0.985 (range: 0.905-0.996) when the former versus the latter scenario was used. It was further observed that using a reference population consisting of animals with closer genetic relatedness to the test population would also improve imputation accuracy. A strong correlation of 0.817 (P<0.001) was observed between the mean genetic relatedness of animals in the test population, with animals in the reference population, and their resulting imputation accuracy.

This was supported by estimates showing mean imputation accuracy of 0.994 as opposed to 0.982 for animals that had both as opposed to no parents in the reference population. The influence of using different low-density SNP panels, consisting of varying density and SNP content, on more specifically animal-wise and SNP-wise imputation accuracy was then determined. Animal-wise imputation accuracy improved when the SNP density of the lower-density panel improved; correlation-based imputation accuracy 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 when SNPs were randomly chosen. The variation between animals, as well as the degree of improvement in accuracy, became smaller with increasing SNP density. Improvements of 0.043 units were seen when SNPs were doubled from 2 500 to 5 000 SNPs, as opposed to an improvement 0f only 0.007 units when SNPs were (more than) doubled from 20 000 to 50 000 SNPs. Selection of SNPs based on both MAF and LD attributes proved to be the best selection strategy to maximize imputation accuracy and random selection produced the worst imputation accuracy. Mean imputation accuracy exceeding 97% (less than 3% errors) could be achieved by using only 5 000 SNPs when this method of selection was used; using other methods of selection this accuracy was only achieved when double the amount of SNPs (10 000) was used. In terms of SNP-wise

imputation accuracy, accuracy estimates were lower for SNPs located on the chromosomal extremes and if the MAF of these SNPs was low. For chromosome 19, which was the chromosome with the worst mean imputation accuracy for most scenarios, SNPs located in the first (n=32), middle (n=42) and last (n=64) 1Mb of this chromosome, for example, had mean SNP-wise correlation-based accuracy measures of 0.640, 0.810 and 0.577. The difference in SNP-wise imputation accuracy moreover was 0.071 between SNPs in the highest (0.4<MAF≤0.5) and lowest MAF bins (0.01<MAF≤0.1); imputation accuracy was better for SNPs with higher MAF.

Results generated to achieve the final aim of this study are still preliminary and in the process of being analyzed. Preliminary results, however, shows strong correlations between conventionally-estimated EBVs and GEBVs, with the inclusion of genomic information being advantageous to breeding value estimation. The difference in GEBV accuracies estimated from true- versus imputed genotypes was small thus far, depending on the per animal imputation accuracy; the discrepancy is expected to be larger for animals with lower mean imputation accuracy.

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 heterogenous genomic architecture. This variation across the genome allowed variation in imputation accuracy between different chromosomes and genomic regions within chromosomes to be pre-empted. Genotype imputation is a valid genomic strategy for the SA Drakensberger breed and this study concluded that a genotyping panel consisting approximately 10 000 SNPs would suffice in achieving less than 3% imputation errors. Results presented further suggests that if such a panel were to be designed, that the SNPs considered for inclusion would have to be selected based on selection criteria, such as MAF and LD, specific to the SA Drakensberger breed. Considering that no Sanga-specific genotyping panel currently exists, it would be recommended that these SNPs be chosen from re-sequencing efforts, 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. The reason for this is that low MAF, because of ascertainment bias, was the most influential factor affecting achievable imputation accuracy and therefore poses a concern. This study showed that it will be a valid strategy to integrate genotype imputation routinely into future genomic evaluation pipelines for the SA Drakensberger breed as imputation errors are expected to have a negligible effect on resulting GEBV accuracies. Finally, the inferences made from this study may be transferable to other Sanga breeds and may provide guidelines for consideration in future genomic endeavours for these breeds.

Popular Article

Genotype Imputation As A Genomic Strategy For The South African Drakensberger Beef Breed By SF Lashmar, C Visser And FC Muchadeyi

The Drakensberger is a medium-framed breed of cattle with a sleek, black coat. It is believed to be one of South Africa's oldest Sanga breeds and was developed from an ancestral population of cattle that was first sighted in 1659 in the Bredasdorp area of the Western Cape province. These cattle ancestors, also described as black in colour, belonged to native tribes and were crossbred with Dutch cattle of the Groningen breed, which were imported by European settlers in the 1700s. By this introduction of European *Bos Taurus* genetics, the development of the SA Drakensberger was initiated. The modern SA Drakensberger, as it is presently known, was however only recognized in 1947 when the SA Drakensberger Breeders' Society was established. The breed therefore underwent a process of development that spanned centuries, whereby it withstood many harsh challenges in its history and this has led to the hardy breed it is today. Nicknamed the "profit breed", the Drakensberger is both adapted and highly productive within SA's beef producing environment and has a long history of diligent performance recording. In fact, it was the first breed to receive estimated breeding values (EBVs) using best linear unbiased prediction (BLUP) methodology, as performance testing was made compulsory to all breeders since 1980. Participation by Drakensberger breeders in SA Stud Book's *Logix Beef* performance recording scheme is still at 100% today (SA Stud Book, 2017) and extensive pedigree records are available. Considering all of this, there has recently been interest in further enriching breed improvement strategies for the SA Drakensberger with genetic information in the form of genomic selection.

To implement genomic selection can significantly improve the efficiency of selection processes, and hence accelerate genetic progress, for the SA Drakensberger breed. This selection strategy, however, requires large numbers of animals to be "tested", referred to as "genotyped", for a high density of single nucleotide polymorphism markers (SNPs) in order to make reliable scientific deductions and to produce accurate

genomic estimated breeding values (GEBVs) for farmers or breeders. From experience, international researchers have suggested 1 000 animals to be included in a training- or reference population to deduce the prediction equations that will be used in calculating GEBVs for selection candidates. Generating the amount of data to fulfill the number of genotyped animals necessary in the training population alone can become unfeasibly expensive, especially in developing countries, considering that the cost of genotyping an animal for about 150 000 SNPs is currently approximated at ZAR200 per animal. The cost of genotyping can, however, be alleviated by genotyping animals with SNP chips containing lower numbers of SNPs and “imputing” to higher density.

In statistical terms, imputation refers to the process of replacing missing data with substituted values. In the context of genomics, genotype imputation refers to a method of predicting SNP genotypes for SNPs that are either missing or were not physically genotyped. The genotypes are predicted based on patterns observed from a more complete data set of SNPs that are available for a group of animals that are representative of a specific breed. Consider for example that 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, and the fact that these animals share large parts of the DNA, we can “impute” or infer the “missing” 90 000 markers for the young animal by making certain statistical assumptions using the principles of genetics. On a larger scale: if a “reference” population (consisting of older, high-impact animals with many offspring in the national herd) is genotyped for a high density of genetic markers (let’s say 150 000 SNPs) and a “test” population (younger, commercial animals in the national herd) is genotyped for a smaller subset of these SNPs (let’s say 50 000 SNPs), the 150 000-SNP genotype profile can be imputed for the “test” animals. The prerequisite is, however, that the animals in the reference- and test populations need to be related in some way, in other words they need to share underlying genetic patterns. These shared patterns can be used to fill in the gaps in SNP information. The imputed SNPs i.e. the 100 000 “missing” SNPs not included on the lower density panel, can however only be used in downstream application such as genomic selection if they were accurately imputed or assigned otherwise inaccurate scientific deductions will be made.

Imputation is now almost routinely included in genomic evaluation processes abroad because this methodology has been optimized, through trial-and-error and studying the factors influencing “imputability” of SNPs, for the most popular international breed. To be able to make use of this methodology within the South African beef industry, and more specifically for local breeds, requires a process of validation and this has not yet been performed for breeds such as the SA Drakensberger. The aim of the study was therefore to comprehensively evaluate genotype imputation for the SA Drakensberger breed so that it can be routinely applied in a GS pipeline.

The first step in the process of validation was to investigate the genomic characteristics of the breed. The genomic characteristics of SNPs have previously been shown to have an influence on the accuracy with which genotypes can be imputed. The genome of each animal is subdivided into different structures, called chromosomes, and on each of these chromosomes differences may furthermore exist between different DNA segments depending on the origin of these segments i.e. from which animal in the pedigree that part of DNA was inherited. As a result of the history of the SA Drakensberger, the genomes of animals belonging to this breed are expected to be composite i.e. containing genomic segments from both *Bos taurus* and *Bos indicus*. Certain parameters can provide more information on the SNPs within each of these segments and these include the minor allele frequency (MAF) and linkage disequilibrium (LD). The MAF gives an indication of the value of a specific SNP to the breed in question; if the MAF is high, it is an indication that both alleles of the SNP are present amongst the animals in the breed i.e. the SNP is informative. The LD provides an indication of the relationship between adjacent SNPs; if SNPs are in high LD, a “block” of SNPs can be inherited together and animals share larger parts of the genome with one another. This improves the ability of SNPs in these regions to be imputed. The software, Plink, was used to quantify these parameters on a per chromosome basis. Results showed that there was variation in these genomic characteristics between different chromosomes and this led us to expect differences in imputation accuracy between chromosomes.

The logical next step was to calculate the actual achievable imputation accuracy. The accuracy of imputation was calculated for imputation from several custom-derived low-density panels. To achieve this objective, different sets of SNPs were extracted from the SNP data available (150K SNP data) to mimic possible lower-density SNP panels. Panels containing 2 500, 5 000, 10 000, 20 000 and 50 000 SNPs were tested. The choice of SNPs to be included on each of these panels were based on certain SNP selection strategies i.e. different criteria were used to select the SNPs. The different strategies of selection included 1) selecting SNPs randomly, 2) selecting SNPs so that they were approximately evenly spaced, 3) selecting only SNPs with the highest MAF and 4) selecting SNPs based on a score combining its MAF and relationship to neighbouring SNPs (LD). Imputation was done using a software called FImpute and our findings suggested that a low-density SNP panel consisting of approximately 10 000 SNPs that were

selected based on their MAF and LD information will be optimal. Using such a panel resulted in less than 3% imputation errors.

The final step was to determine the influence of mistakenly imputed SNPs on the accuracy of GEBVs and hence on genomic selection. The “single-step” approach to GS was tested using software called Mix99. Breeding values were calculated using 1) only pedigree information (traditional), 2) using true genotypic data (GEBV) and 3) using imputed genotypic data (imputed GEBV). These different breeding values were compared to determine whether imputation accuracy had an effect. Our preliminary findings suggest that the inclusion of genomic data is advantageous and that there is a minimal effect on GEBV accuracy estimates if imputation accuracy was good.

To conclude, results from this study indicated that imputation is a valid genomic strategy towards cost-effectively implementing GS for an indigenous breed such as the SA Drakensberger despite the uniqueness and complexity of its genome. The outcomes of this study may moreover be transferable to other Sanga breeds and may provide a set of guidelines for genomic studies requiring imputation in the future. Even though this study has shown that a more affordable lower-density panel can be developed from choosing SNPs with high MAF in indigenous breeds from currently available genotyping platforms, it would be invaluable for future genomic endeavours to develop a Sanga-specific panel using breed-specific SNPs identified from re-sequencing efforts.

Please contact the Primary Researcher if you need a copy of the comprehensive report of this project on : – mmakgahlela@arc.agric.za

Cattle and Small Stock, Livestock Production, Uncategorized, with global competitiveness

2019, CSS, Online, UP, Visser

‹ Genetic markers for Haemonchus contortus in sheep

› Genetic diversity of landrace cattle breeds

DEADLINES for RESEARCHERS 2021

Proposals for 2021: TBC
Progress reports: 28 Jan 21
Final reports: 29 Jan 21 Final includes comprehensive report and popular article

COMMITEE MEETINGS for 2021

RMRDSA CSS Planning - TBC
Project Committee - TBC
Pork Planning - TBC

Calendar

< Apr 2021 >

Sun	Mon	Tue	Wed	Tur	Fri	Sat
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

PORK Priority Areas

Cattle & Small Stock Programmes

1 Sustainable natural resource utilisation

2 Improvement of Livestock production and forage

3 Management of agricultural risk to create a resilient Red Meat sector

4 Sustainable health and welfare for the Red Meat sector

5 Enhancement of production and processing of Animal Products

6 Consumer and market development of the Red Meat sector

7 Commercialisation of the emerging sector