Use of Genetic Polymorphisms to Assess the Genetic Structure and Breed Composition of Crossbred Animals

WEERASINGHE MUDIYANSELAGE SHALANEE PRIYANGA WEERASINGHE

Bachelor of Veterinary Medicine and Animal Science, University of Peradeniya, 2003

Master of Philosophy in Animal Science, University of Peradeniya, 2010

A thesis submitted for the degree of Doctor of Philosophy of the

University of New England

June 2015

School of Environmental and Rural Science

Abstract

This thesis explores the accuracy of methods to estimate the breed composition of crossbred animals which have unknown pedigree. Herein I present the use of SNP technologies to estimate the breed composition of small-holder crossbred dairy cattle in developing countries for the first time. Before this could be done there was a need to determine: what are the accuracies of different methods for estimating breed composition? The genetic structure of animals, the design of reference populations, the number of SNP markers and the model selected has possible consequences for estimation of breed composition. Once the effect of the above factors on the accuracy of estimation of breed composition is identified, it is possible to estimate with confidence the breed composition of crossbred animals that have no recorded pedigree. The overall aim of this thesis was to investigate the use of high-density SNP data to understand the livestock breed's population structure and estimate the breed composition of crossbred animals.

The first block of research was to determine the genetic structure of Australian sheep breeds and estimate the accuracy and bias of SNP-based estimates of breed composition in F1 sheep, using Sheep-CRC data as a model. These results confirmed that principal component analysis (PCA) provided useful visualisation of the genetic diversity of Australian sheep breeds and crosses, and identified a variety of errors in the data. I therefore, recommend the use of PCA for initial examination of population structure of the data and identification of data errors before performing a detailed genomic analysis. Subsequent results indicated that for a simple cross the breed composition can be estimated with high accuracy and low bias when using high density SNP data. This was true even when breeds unrelated to the cross being analysed, were included in the analyses.

The second block of research examined the accuracy and bias of estimates of breed composition of alternative F1 crosses and three-way crosses in sheep. It examined the effect of population stratification of the crossbred population on the accuracy and bias of estimates. The results showed breed composition of composite crossbred sheep can also be estimated with high accuracy and low bias. The estimates of breed composition of crossbreds from purebred parents showed higher accuracy and less bias than the crossbreds from composite parent breeds. However, the accuracy was slightly increased and bias reduced when the ancestors of the composite breed were included in the admixture model. My results also showed that the design of the reference population has a considerable impact on the estimates of breed composition of crossbreds. The estimates of breed composition of crossbreds among genetically distant parent breeds were somewhat more accurate and less biased than those among genetically closer breeds. Overall, for crosses among breeds that are clearly genetically distinct, errors of estimation of breed proportion should have S.D. in the range of 0.02 to 0.03. My results can be extrapolated to predict that the S.D. could be substantially lower than I observed, when crosses are between highly diverse breeds, such as between *Bos indicus* and *Bos taurus* cattle. Finally, when all the crossbred data were combined into a single population and estimates of each crossbred group were obtained in a single analysis, estimates still had high accuracy and low bias.

The third and fourth blocks of research applied the best models from the first and second blocks of my research to estimate the breed composition of different populations of crossbred dairy cattle used by small-holder dairy farmers in East Africa. These crossbred populations have little information on the admixture levels, due to the lack of pedigree recording across the many generations of crossbreeding in the region. First, in both research blocks, the genetic structure of indigenous and crossbred cattle populations was examined. Then the dairy proportion and individual breed composition were estimated using the ADMIXTURE program. Chapter 3 focuses on the genetic structure and admixture levels of small-holder cattle in Kenya and Uganda. One of the most significant findings to emerge from this study was that the correlation of the estimated dairy proportion across the various types of analysis is greater than 0.99, indicating that these estimates are highly robust to model assumptions. The average dairy proportion of crossbred cows is 0.69 with S.D. 0.21, whereas the bulls sampled from the same area as crossbred cows show a comparatively higher dairy proportion (0.79 with S.D. 0.18). Another important finding was that the farmers' prediction of the dairy proportion of individual animals explains only 16% of the real variation of the dairy proportion. It demonstrates that many farmers have a poor understanding of the degree of crossbreeding of their animals. Also, the average dairy proportion is lower than expected and the variation much higher than expected in these populations of crossbred animals. The estimates of individual dairy breed proportion (Holstein vs Ayrshire vs Jersey, etc.) reflected what was expected in terms of the recorded history of different breed use in different locations. Although, the estimates obtained were dependent on the model used, and without an independent validation of the estimates, they were considered unreliable for use in analyses of breed performance.

The fourth research block examined the estimation of breed composition of crossbred dairy cattle in Tanzania and Ethiopia, which were believed to have different small-holder dairy cattle populations compared to Kenya and Uganda. Here, I explored in more detail than was possible with the data in the third research chapter, the genetic structure of East African indigenous cattle. The results here and in the third research chapter are consistent with what are believed to be the major historical migrations of *Bos indicus* and *Bos taurus* cattle on the African continent. I found that several indigenous breeds were contaminated with the recent introduction of European breed germplasm. I propose methods for using SNP-based breed composition estimates to aid genetic conservation programs. Furthermore, I discovered that the molecular estimates of breed composition of the Tanzania synthetic Mpwapwa breed confirmed its reported origins and also found that it constituted a relatively well-defined breed. I observed Holstein and Friesian breeds to have the most significant impact on small-holder crossbred cattle in Ethiopia and Tanzania. That the average and variance of total dairy proportion differed between the two countries and differed markedly between locations within the countries, has important implications for the design of programs to improve small-holder (SHD) management and genetics in different areas.

In conclusion, these results demonstrate the ability to identify the genetic structure and estimate the breed composition of crossbred cattle which do not have pedigree records. In this regards the design of the reference population and the number of SNP markers have a considerable effect on the accuracy of estimations. This knowledge has the potential to be used in livestock genetic improvement in the developing world. The further research could be carried out to design the cheap SNP test that could be used to genotypes the crossbred animals owned by small-holder farmers in these countries. Additionally, genotypes of more *Bos indicus* and African *Bos taurus* cattle breeds relevant to the sample collected regions could be included in the SNP panel to prevent the bias of estimations.

Declaration

I certify that the substance of this thesis has not been submitted previously for any degree and is not currently being submitted for any degree or qualification.

I also certify that any help received and all sources used in preparing this thesis have been acknowledged therein.

[redacted]

W.M.S.P.Weerasinghe

12th June 2015

Acknowledgments

First, I would like to thank my principal supervisor, Professor John Gibson, for all the encouragement and support he has given me. I thank him for his enthusiastic attitude and his knowledge which has enabled me to learn both the theoretical and applied aspects of scientific research, something I will value for my entire life. Every discussion and consultation with Professor Gibson left me both enlightened and compelled to think more deeply about my research. Second, I would like to express my heartfelt gratitude to associate supervisors, Associate Professor Cedric Gondro for his encouragement, support and never-ending patience. Thanks for showing me an inspirational journey in computer programming. I would also like to extend my gratitude to Dr Gilbert Jeyaruban who gave me strength and courage throughout my candidature especially when I was in difficult situations.

I would like to express my sincere gratitude to the Australian Government for granting me an Endeavour Postgraduate Research Scholarship which has provided me with the financial assistance to complete my PhD with so much ease and comfort. I also thank the University of New England, Australia, for providing all the infrastructure and facilities during my candidature period.

I would also like to thank my fellow scientists and friends who have helped me with my thesis.

My hard-working parents have sacrificed their lives for my sister, brother and me to provide us better education, and with unconditional love and care. I love them so much, and I would not have made it this far without them. I especially thank my dad for all the courage and care he has given me and my children. I will forever be grateful. I know you were devastated when we came to Australia, but you kept it from us. I miss you forever, I know wherever you are, you will be proud of me. I have completed your dream.

This thesis would not have been possible without the loving and selfless understanding and patience of my husband, Shantha. My lovely angels Ashini and Piyumi, thank you so much for your patience and unfailing love.

Table of Contents

Abstract ii

Declaration vi

Acknowledgments vii

List of Tables xiii

List of Figures xvi

Chapter 1. General Introduction 1

1.1. Scope of the thesis 5

Chapter 2. Review of literature 8

2.1. The main features of small-holder dairy systems in East African countries 9

2.2. Genetic markers 12

2.3. Measure of genetic diversity of livestock species 14

2.3.1. Principal component analysis (PCA) 15

2.3.2. Allele sharing 16

2.3.3. Fixation indices 17

2.3.4. Application to genetic diversity of sheep 18

2.3.5. Application to genetic diversity of African cattle 20

2.4. Statistical approaches to estimate the admixture proportion based on the genotype data using STRUCTURE, ADMIXTURE and FRAPPE 21

2.4.1. Regression methods 23

2.4.2. Example of application of molecular tools and statistical approaches to estimate the admixture proportion in non-livestock and livestock populations 24

2.5. Genetic diversity of sheep 26

2.6. Genetic Diversity of East African Cattle 29

2.7. Conclusion 31

Chapter 3. The accuracy and bias of estimates of breed composition and inference about genetic structure using high-density SNP markers in Australian sheep breeds 32

3.1. Abstract 33

3.1.1. Background 33
3.1.2. Methods 33
3.1.3. Results 34
3.1.4. Conclusions 34
3.2. Background 35
3.3. Methods 37
3.3.1. Data sets 37
3.3.2. Principal Component Analysis (PCA) 39
3.3.3. Genetic variation between selected reference breeds 41
3.3.4. Estimation of breed composition of Merino:Border Leicester F1 animals 41
3.3.5. Genetic variation within Merino:Border Leicester F1 animals 42
3.4. Results and Discussion 43
3.4.1. Genetic structure of selected reference sheep breeds 43
3.4.2. Estimation of breed composition of Merino:Border Leicester F1 51
3.4.3. Estimation of breed composition of F1 animals when including additional reference breeds 56
3.5. Conclusions 62
3.6. Competing interests 62
3.7. Authors' contributions 63
3.8. Acknowledgements 63
Chapter 4. Accuracy of estimation of individual animal breed composition in crossbred populations: a test case in sheep 66
4.1. Abstract 67
4.1.1. Background 67
4.1.2. Methods 67
4.1.3. Results 67
4.1.4. Conclusions 68
4.2. Background 68
4.3. Methods 71
4.3.1. Data-sets 71
4.3.2. Genetic structure of reference sheep 73
4.3.3. Estimation of breed composition 73

4.4. Results ... 76

4.4.1. Genetic structure of reference breeds ... 76

4.4.2. Estimation of breed composition of crossbred sheep ... 82

4.5. Discussion ... 94

4.6. Conclusions ... 100

4.7. Competing interests ... 100

4.8. Authors' contributions ... 100

4.9. Acknowledgments ... 100

Chapter 5. Genetic diversity and inference of breed composition of East African small-holder crossbred dairy cattle ... 103

5.1. Abstract ... 104

5.1.1. Background ... 104

5.1.2. Methods ... 104

5.1.3. Results ... 105

5.1.4. Conclusions ... 105

5.2. Background ... 105

5.3. Methods ... 109

5.3.1. Data sets ... 109

5.3.2. Genotyping ... 110

5.3.3. Clustering with principal component analysis (PCA) ... 111

5.3.4. F Statistics ... 112

5.3.5. Estimation of breed composition using ADMIXTURE analysis ... 112

5.4. Results ... 114

5.4.1. Genetic structure of East African cattle and reference breeds ... 114

5.4.2. Estimation of breed composition for DGEA crossbred cows ... 121

5.5. Discussion ... 131

5.5.1. Genetic structure of East African cattle and reference breeds ... 131

5.5.2. Estimation of breed composition of crossbred dairy cows ... 133

5.6. Conclusions ... 137

5.7. Competing interests ... 137

5.8. Authors' contributions ... 138

5.9. Acknowledgments ... 138

Chapter 6. Genetic structure and breed composition of small-holder crossbred dairy and indigenous cattle in Ethiopia and Tanzania ... 141

6.1. Abstract ... 142
6.1.1. Background ... 142
6.1.2. Methods ... 142
6.1.3. Results ... 142
6.1.4. Conclusions ... 143
6.2. Background ... 144
6.3. Methods ... 147
6.3.1. Data sets ... 147
6.3.2. Genotyping ... 150
6.3.3. Principal component analysis (PCA) ... 151
6.3.4. Genetic differentiation and relationship among the populations ... 152
6.3.5. ADMIXTURE analysis ... 152
6.4. Results ... 153
6.4.1. Genetic structure of East African indigenous cattle ... 153
6.4.2. Genetic structure of Mpwapwa cattle ... 162
6.4.3. Genetic structure of crossbred cattle in Ethiopia and Tanzania ... 163
6.5. Discussion ... 171
6.6. Conclusions ... 176
6.7. Competing interests ... 177
6.8. Authors' contributions ... 177
6.9. Acknowledgments ... 177
Chapter 7. General Discussion and Conclusions ... 180

7.1. General Discussion ... 180
7.1.1. The identification of genetic structures in reference breeds and crossbred populations. 181
7.1.2. The impact of the number of markers used in the estimation of breed compositions ... 183
7.1.3. The construction of reference populations for breed proportion estimations ... 184
7.1.4. Estimation of exotic dairy proportion and individual breed proportions in crossbred dairy cattle ... 185
7.2. Conclusions ... 188
Chapter 8. Consolidated Reference List ... 190

List of Tables

Table 2.1 Comparison of programs used in this study to estimate the breed proportion 23

Table 3.1 Number of reference animals from the International Sheep HapMap consortium and test animals from Sheep-CRC.. 39

Table 3.2 Summary of PCA for the first five principal components for the reference sheep breeds......... 44

Table 3.3 Pairwise estimates of FST between sheep breeds 50

Table 3.4 Standard deviation (accuracy) and average deviation from the pedigree proportions (bias) for the Merino:Border Leicester F1 animals using STRUCTURE program with different subsets of SNPs....... 54

Table 3.5 Accuracy and bias of estimation of Merino proportions for Merino:Border Leicester F1 animals using different subsets of SNPs and different programs 55

Table 3.6 Accuracy and bias of estimation of Merino proportions for, Merino:Border Leicester F1 animals with different models, with and without parental correction 56

Table 3.7 Accuracy and bias of Merino proportion of F1 animals with and without parental correction (P.C.) for analysis with K=4 60

Table 4.1 Number of animals sampled from International Sheep HapMap consortium and Sheep-CRC .. 72

Table 4.2 Pairwise estimates of F_{ST} between HapMap reference breeds and Sheep-CRC sires 79

Table 4.3 Standard deviation (accuracy) and average deviation from the pedigree proportion (bias) for the F1 and three-way crossbred animals using ADMIXTURE analyses 85

Table 5.1 Summary of the variance explained by the first five principal components for the DGEA cattle 117

Table 5.2 Pairwise genetic differentiation (FST) between reference and DGEA indigenous breeds and reference animals 119

Table 5.3 Mean and standard deviation of breed proportion estimates of crossbred cows for each site and an average across all sites, from the supervised analysis with K=7. Individual dairy breed proportions are expressed as a proportion of the total dairy proportion 127

Table 5.4 Correlations among estimated proportions of individual dairy breeds of crossbred cows sampled across all sites (upper diagonal) and sampled at Bushenyi (lower diagonal). Model K=7, unsupervised. Proportions expressed as proportion of total dairy proportion 128

Table 5.5 Mean and standard deviation of breed proportion estimates of bulls for each site and an average across all sites, when K=7, supervised analysis. Proportions expressed as proportion of total exotic proportions 130

Table 6.1 The data structure of Ethiopian and Tanzanian crossbred and local reference breeds sampled from DGEA 150

Table 6.2 Pairwise genetic differentiation (F_{ST}) between reference and EA indigenous breeds 159

Table 6.3 Mean S.D. for estimates of dairy breed proportion of Ethiopian crossbred for each district. Proportions of individual dairy breeds are expressed as a proportion of total dairy breed content. 166

Table 6.4 Mean and S.D. of breed proportion estimates of Ethiopian crossbreds classified by the farmer assessment of both the predominant dairy breed influence plus their assessment of degree of dairy influence (high grade vs lower grade). Proportions of individual dairy breeds expressed as proportion of total dairy proportion 167

Table 6.5 Mean and SD for various breed proportion estimates of Tanzanian crossbreds for each district. (ADMIXTURE, Supervised, K=7 –Ayrshire-CDN, Nelore, Holstein, Jersey, Friesian, N'Dama, Guernsey). Proportions for individual dairy breeds expressed as proportion of total dairy proportion 169

Table 6.6 Mean and S.D. of breed proportion estimates of Tanzanian crossbreds classified by the farmer assessment of both the predominant dairy breed influence plus their assessment of degree of dairy influence (high grade vs lower grade). Proportions expressed as proportion of total dairy proportion. 170

List of Figures

Figure 2.1 Possible migration routes of domestic cattle in Africa. 30

Figure 3.1 Plots of PC2 vs PC1 for individual animals colour-coded by breed 45

Figure 3.2 Phylogenetic tree for individual animals based on allele sharing 46

Figure 3.3 Distribution of the level of homozygosity for the reference sheep breeds 51

Figure 3.4 Estimation of Merino proportion in Merino:Border Leicester F1 using STRUCTURE program using 1 168 (A), 4 670 (B) and 46 700 (C) 52

Figure 3.5 Estimates of mean Merino proportions for Border Leicester (BRL), Merino (MEC) and Merino:Border Leicester F1 (BRL F1) using different subsets of SNPs.. 53

Figure 3.6 Plot of PC2 vs PC1 for an analysis including F1 MEC:BRL animals and subset of reference breeds 57

Figure 3.7 Estimated ancestral breed proportions with reference breeds APD, ISF, BRL, MEC and K=4, unsupervised analysis 58

Figure 3.8 Estimated ancestral breed proportions with reference breeds APD, ISF, BRL, MEC and K=4, supervised analysis 59

Figure 3.9 Estimated ancestral breed proportions with parent breeds having n=20, and K=2 unsupervised analysis.. 61

Figure 4.1 Plots of PC2 vs PC1 for individual reference animals colour-coded by breed 77

Figure 4.2 Plots of PC2 and PC1 for reference animals and Sheep-CRC sires colour-coded by breed 78

Figure 4.3 Estimation of admixture proportion of reference animals using ADMIXTURE program with K=2 to K=6 81

Figure 4.4 Plots of PC2 vs PC1 for individual reference and crossbred animals colour-coded by breed and cross-breed group 83

Figure 4.5 Estimates of Merino proportion in Suffolk:Merino F1 using the STRUCTURE program. The breeds are presented (A) - in the order of MEC, ISF and ISF F1 animals ordered according to sire families; (B) - in the order of ISF-c, MEC and ISF F1 animals ordered according to sire families; and (C) - in the order of ISF-c, MEC, ISF and ISF F1 animals ordered according to sire families 86

Figure 4.6 Estimation of Merino proportion in Coopworth:Merino F1 using STRUCTURE program. (A) Estimates without population information including Sheep-CRC Coopworth, MEC and CPW as reference samples. (B) Estimates including Sheep-CRC Coopworth, MEC and CPW as reference, and setting MEC, CPW and Sheep-CRC Coopworth as ancestor breeds in the model. (C) Estimates including BRL, ROM, Sheep-CRC Coopworth, MEC and CPW as reference, and BRL, ROM and MEC set as ancestor breeds in the model 88

Figure 4.7 Estimates of breed composition from unsupervised ADMIXTURE analysis with K=5, including all reference breeds, F1 and three-way crosses. Each individual is represented by a vertical bar, with coloured segments indicating estimated breed proportions. Breeds are separated by vertical black lines............ 89

Figure 4.8 S.D. (A) and bias (B) of Merino proportion for the F1 and three-way crosses estimated with an unsupervised ADMIXTURE analysis with K=5 and K=6 using 20 reference animals, all reference animals and corrected for parent breeds 91

Figure 4.9 Distribution of the standard deviation (A) and the bias (B) of estimates of the Merino proportion from STRUCUTURE supervised and unsupervised, and ADMIXTURE supervised analysis with K=5 and K=6 for F1 and three-way crosses 93

Figure 5.1 Sample location of the Kenyan and Ugandan crossbred cattle 110

Figure 5.2 Plots of PC2 versus PC1 for DGEA crossbred and indigenous cows, with crossbred cows coloured according to the site from which they were collected. B-Xbred (Bushenyi), W-Xbred (Wakiso), O-Xbred (Butere, Kabras, Kaptumo, Meteitei and Siongiroi) 116

Figure 5.3 Plots of PC2 vs PC1 for individual indigenous animals and Kenyan (K_Bull) and Ugandan bulls (U_Bull). PCA was conducted with the reference breeds, DGEA crossbred cows and bulls and indigenous animals. For clarity, crossbred cows are not shown 118

Figure 5.4 Distribution of the levels of homozygosity among the reference animals and DGEA cows with pruned genotypes 120

Figure 5.5 The relationship between levels of homozygosity and estimated dairy proportion for DGEA crossbred cows. 121

Figure 5.6 Bar plots of estimated breed proportions of crossbred cows, indigenous and reference breeds from unsupervised ADMIXUTURE analyses. Each individual is represent by a vertical bar and partitioned in to coloured segments according to its estimated breed proportions. . Breeds are separated by black lines. Reference breeds are presented in the order of Friesian (1), New Zealand Ayrshire (2), CDN Ayrshire (3), Nelore (4), Guernsey (5), Holstein (6), Jersey (7) N'Dama (8), Ankole (9), Nganda (10) and SEAZ (11). DGEA crossbred cows (12) are shown in in the order of location of samples, Bushenyi (Uganda), Kenyan and Wakiso (Uganda) 123

Figure 5.7 Estimates of mean exotic and relative Holstein-Friesian proportions across the sample collection sites for bulls and crossbred cows 125

Figure 5.8 Distribution of farmer-predicted dairy proportion and ADMIXTURE estimation for individual animals. 131

Figure 6.1 Sample locations of the Ethiopian crossbred animals 148

Figure 6.2 Sample locations of the Tanzanian crossbred animals 149

Figure 6.3 Plots of PC2 vs PC1 for individuals of East African indigenous cattle colour-coded by breed. 155

Figure 6.4 Inferred phylogenetic tree for individual East African indigenous and reference breeds, based on allele sharing.. 156

Figure 6.5 Estimation of admixture proportion of East African indigenous animals using ADMIXTURE program with K=2 to K=5 and K=11 161

Figure 6.6 Distribution of the relative *Bos indicus* proportion and absolute European Bos taurus proportion of indigenous breeds estimated with a supervised ADMIXTURE analysis with K=7............................... 162

Figure 6.7 Plots of PC2 vs PC1 for crossbred dairy cattle and reference breeds in Ethiopia, with individual animals colour-coded by breed. Crossbred are colour-coded according to the milk shed (zones). Zone-1 (Addis-Ababa, Addis-Ababa_Kality-DebreZeit and Nazreth_DireDawa), Zone-2 (Addis-Ababa-Selale, Mekelle and Dejen-BahirDar) and Zone-3 (Addis-Ababa-Ziway, Addis-Ababa-DebreBirhan and Shahemene-Awasa-Yirgalem) 164

Figure 6.8 Estimates of breed proportions from supervised ADMIXTURE analyses with K=7, for reference breeds, indigenous breeds and Ethiopian crossbreds. Each individual is represented by a vertical bar and partitioned into coloured segments according to the breed proportion of the individual. Breeds are separated by vertical black lines.. 165

Figure 6.9 Plots of PC2 vs PC1 for individual crossbred and reference animals in Tanzania, colour-coded by breed (A) and the same results for clarity, crossbred cows from Lushoto and Southern-Highland regions are not shown (B) 169