The University of New England

AUSTRALIA

The Genetic Architecture of Complex Traits in Sheep

Hawlader Abdullah Al-Mamun

BSc (CSE) (Khulna University, Bangladesh)

MSc (Bioinformatics) (Chalmers University of Technology, Sweden)

A thesis submitted for the degree of Doctor of Philosophy at

The University of New England

School of Science and Technology and School of Environmental and Rural Science

October 2014

Declaration

I certify that the work has not been and is not being submitted for any other degree to this or any other university for any other degree or qualification.

I also certify that all help received in preparing the thesis and all sources used, are duly acknowledged.

Hawlader Abdullah Al-Mamun

Acknowledgements

Firstly, I would like to pay my sincere gratitude to the almighty Allah who created me and has been showering me with his tremendous blessings.

I am extremely grateful to my supervisors, Associate Professor Paul Kwan and Associate Professor Cedric Gondro, for their guidance, knowledge, ideas, suggestions, encouragements, and tired-less efforts in enabling this thesis to be written. I also like to express my appreciation to my co-supervisor, Dr Samuel Clark, who has provided tremendous support to me during the last six months of my thesis research.

I considered myself very fortunate to have been awarded the IPRS scholarship (tuition fees and stipend) by the Australian Government to conduct my study at UNE. In addition, I offer my sincere gratitude to the Bangladesh Government for providing me a free and solid education from primary through to the graduate level.

During my study, I lived far apart from my parents who are living in Bangladesh. However, I feel their constant emotional support, encouragement and good wishes. I would also like to thank my wife and my son. Without their continuous support, patience and sacrifices I will not have succeeded in completing this thesis.

List of publications from this thesis

Journal papers

Chapter 2

Al-Mamun, H. A., S. Clark, P. Kwan, and C. Gondro. *Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep.* (submitted to BMC Genetics on 13/10/14).

Chapter 3

Al-Mamun, H. A., Kwan, P., Clark, S., Ferdosi M. H., Tellam, R. L. and Gondro, C. *Genome-wide association study for body weight in Australian Merino sheep reveals an orthologous region on OAR6 to the genomic region of human and cattle affecting height and weight traits*. (submitted to Genetics Selection Evolution, initial comments from the reviewers received on 25/09/14).

Chapter 4

Al-Mamun, H.A., P. Kwan, S. Clark and C. Gondro. *Genomic best linear unbiased prediction using parallel differential evolution*. (in preparation for PLOS Computational Biology).

Chapter 5

Al-Mamun, H.A., P. Kwan, M.H. Ferdosi, R.L. Tellam, J.W. Kijas and C. Gondro. *Partitioning of gene expression variation in sheep half-sib families*. (in preparation for PNAS).

Conference papers

Chapter 3

Al-Mamun, H.A., S. Clark, P. Kwan and C. Gondro. *Genome-Wide Association Study on Body Weight Reveals Major Loci on OAR6 in Australian Merino Sheep*. 10th WCGALP, Vancouver, 18-22 August, 2014.

Chapter 5

Al-Mamun, H.A., P. Kwan, R.L. Tellam, J.W. Kijas and C. Gondro. *A Study on Effects of Family and Haplotype Blocks on Conservation of Gene Expression Traits in Half Sib Sheep Families*. 20th Conference of the AAABG, Napier, 20-23 October, 2013.

Abstract

Many important traits in biology, medicine and agriculture are complex and quantitative in that they exhibit continuous variation and non-trivial patterns of genetic inheritance. They are largely polygenic and influenced by factors such as gene-gene and gene-environment interactions. Important reasons to study complex traits include trying to understand how the genetic components operate on their own and how they relate to each other, quantification of the contributions of these elements to trait variation, and elucidation of the underlying genetic architecture behind a trait. An understanding of the sources and consequences of variation in complex traits and identification of the genes involved provides us with a handle to manipulate biological systems, which can have direct applications in medicine and agricultural production. From an agricultural standpoint there are huge economic benefits to be achieved by a better understanding and exploitation of the genetic architecture of complex production traits such as milk yield in dairy animals and meat quality in e.g. sheep or cattle.

This thesis is centred on making some inroads to better understand the genetic architecture of complex traits in sheep. The thesis progresses through a characterization of genetic structure and variability in Australian sheep populations, followed by a genome-wide association study for weight. Then a novel approach to improve estimates of genomic breeding values is discussed. Lastly, the inheritance and partitioning of gene expression variance is studied. A more detailed breakdown of the thesis follows.

Chapter 1 sets the scene with an overview of the genetics behind complex traits. Genome-wide association studies (GWAS) and genomic selection methods are reviewed followed by a brief discussion of heritability that discusses the notions of missing and phantom heritability. Next, an overview of differential evolution (DE) which is a heuristic optimization method used in chapter 4 to improve genomic selection is summarized.

Chapter 2 estimates and compares linkage disequilibrium (LD) and several population metrics, including gene diversity (H_e) and fixation index (F_{st}), in five Australian sheep populations – 3 independent breeds: Merino, Border Leicester, Poll Dorset and 2 crossbred populations: Merino and Border Leicester crosses and crosses of Merino, Border Leicester and Poll Dorset. In all five populations LD decayed rapidly with increased distances between marker pairs. Out of these populations, Merino exhibited higher rates of LD decay than the other pure breeds. Simultaneously, Merino was found to be the most diverse breed among the three pure breeds.

Results from this study are expected to provide an improved understanding of the genetic diversity among the three main Australian sheep breeds as well as insights on the effects of selection on these breeds and their crosses.

Chapter 3 presents a genome-wide association study for body weight (BW, 6 – 10 months) conducted on 1,781 Australian Merino sheep that led to the identification of a major QTL region on OAR6. Thirteen SNP on OAR6 were found to be associated with body weight and two neighbouring genetic loci at *NCPAG* and *LCORL* were identified. The syntenic regions in some other mammalian species are similarly associated with body size traits, thereby suggesting an ancient and common underlying biological mechanism. These findings are anticipated to facilitate the discovery of causative variants for body weight and will help inform marker-assisted selection.

Chapter 4 suggests a method to improve the predictive ability of genomic best linear unbiased prediction (GBLUP). In GBLUP, a genomic relationship matrix (GRM) estimated from markers is used to define the covariance between individuals based on observed similarity at the genomic level. Differential evolution was used to find a set of markers that maximized the prediction accuracy which were then used to build the GRM. The predictive ability of GBLUP with this new GRM was evaluated using simulated data that had different numbers of known QTL and various levels of heritability. Results with empirical data using this method had better predictive ability than conventional GBLUP.

Chapter 5 attempts to understand the basis of inheritance of gene expression in sheep half-sib families using SNP chips and gene expression microarrays. Heritability of gene expression was estimated and its variation partitioned into additive, haplotype and sire effects. Results indicate that gene expression is highly conserved within families but additive and haplotype effects only account for a small proportion of the variance. This suggests that conservation of expression levels is more dependent on higher order interactions than on allelic variants or local regulatory regions.

Chapter 6 provides a summary of the main findings of the thesis and discusses future research.

Contents

Declaration 2
Acknowledgements 3
List of publications from this thesis 4
Journal papers 4
Conference papers 4
Abstract 5
Contents 7
List of figures 10
List of tables 11
Chapter 1 Introduction 12
Complex traits and inheritance 12
Genome-Wide Association Studies (GWAS) 14
Marker-associated selection (MAS) and genomic selection (GS) 16
Missing heritability and phantom heritability 19
Evolutionary Algorithms (EA) and Differential Evolution (DE) for genomic analysis 20
Traits of interest and study aims 23
Chapter 2 Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep 25
Abstract 25
Background 26
Results 28
Linkage Disequilibrium Analysis 32
Discussion 36
Conclusion 41
Methods 41
Competing interests 44
Authors' contributions 44
Acknowledgements 44
Chapter 3 Genome-wide association study for body weight in Australian Merino sheep reveals an orthologous region on OAR6 to the genomic region of human and cattle affecting height and weight traits 47
Abstract 47
Background 48
Methods 49

Ethics statement 49
Phenotypic data 49
Genotyping and quality control 49
Statistical Analysis 49
Haplotype blocks estimation and regression analyses 50
Results 51
Descriptive statistics and quality control 51
Association analysis 51
Haplotype analyses 60
Discussion 60
Conclusion 65
Abbreviations 66
Competing interests 66
Authors' contributions 66
Acknowledgements 66
Chapter 4 Genomic best linear unbiased prediction using parallel differential evolution 69
Abstract 69
Introduction 70
Materials and Methods 72
Evolutionary Algorithms (EA) and Differential Evolution (DE) 72
Finding the optimum set of SNP by DE to build the relationships matrix 73
Parallel execution of DE 74
Constructing a trait specific relationship matrix (TRM) 75
Genomic prediction using G and T matrices 75
Genomic prediction using Bayesian Linear Regression (BLR) 76
Simulation data 76
Real data 77
Results 77
Simulated case studies 77
Case study 1: number of sample (*n*) is larger than number of markers (*m*) 77
Case study 2: number of samples (*n*) is less than number of markers (*m*) 82
True QTL discovered by DE in the simulated data 84
Real data 85
Discussion 86
Conclusion 88
Chapter 5 Partitioning of gene expression variation in sheep half-sib families 92

Abstract....92
Introduction....93
Methods....95
Animals....95
RNA extraction and cDNA synthesis....95
Microarray gene expression data....96
Microarray data processing....96
GRM and IBD estimates....97
Heritability estimates using IBD and GRM....97
Statistical Analysis....98
Results....98
Heritability of gene expression....98
Family plays an important role on gene expression value....99
Variation of gene expression is higher between individuals than within families....101
Gene expression varies between haplotype groups within families....102
Dissection of gene expression variation into variance components....103
Discussion....105
Chapter 6 Conclusions and future work....111
Bibliography....114
Appendix 1 Supplementary tables and figures for chapter 2....130
Appendix 2 Supplementary tables for chapter 3....143

List of figures

Figure 2.1 Average ROH per population 30

Figure 2.2 Number of ROH per chromosome and coverage percentage per chromosome 31

Figure 2.3 Incidence of each SNP in a ROH 32

Figure 2.4 Average r^2 values for each population 33

Figure 2.5 Average *D*' values for each population 34

Figure 2.6 Chromosome wise percentage of block coverage for the five populations 36

Figure 2.7 Percentage of chromosome length in haplotype blocks and average r^2 values for each chromosome 39

Figure 3.1 Manhattan Plot 53

Figure 3.2 Quantile-quantile plot 55

Figure 3.3 Plot of candidate genes in the QTL region 55

Figure 3.4 Linkage disequilibrium (LD) map 57

Figure 3.5 Ensembl alignments of the sequence for the 1 Mb region surrounding the NCAPG-LCORL gene 63

Figure 3.6 SNP effects 65

Figure 4.1 Plot of true genetic value (TGV) vs predicted breeding value for the four approaches 79

Figure 4.2 Accuracy of genomic breeding values (GEBV) using the four different approaches 80

Figure 4.3 Prediction accuracy for different heritabilities 81

Figure 4.4 Accuracy of genomic breeding values (GEBV) using the four different approaches for 10,000 and 40,000 SNP 83

Figure 4.5 Plots of true genetic value (TGV) versus predicted breeding value for 40,000 markers with 200 known QTL using the four approaches 84

Figure 5.1 Principal component analysis of gene expression data 100

Figure 5.2 Principal component analysis of SNP data 100

Figure 5.3 Scatter plot of total variance vs. within family variance 101

Figure 5.4 Scatter plot of gene expression between haplotypes in family 11 102

Figure 5.5 Bar plot of the prediction accuracies from LDA analysis using haplotype data 104

Figure 5.6 Gene expression variation explained by the different variance components 105

Figure 2.S1 Distribution of minor allele frequency (MAF) in five populations 142

List of tables

Table 2.1 Genetic diversity in five sheep populations 29

Table 2.2 Summary of haplotype blocks in the five breeds 35

Table 2.3 Average inbreeding coefficients and effective population sizes 35

Table 3.1 SNP on OAR6 showing significant association with body 54

Table 3.2 List of known genes in the 36.15-38.56 Mb region on OAR6 58

Table 3.3 QTLdb hits within vicinal regions surrounding the 36.15-38.56 Mb interval on OAR6 59

Table 3.4 Haplotype effects of the QTL for body weight 61

Table 4.1 Prediction accuracy using different methods for 1,000 simulated SNP 80

Table 4.2 Prediction accuracy using different methods for 10,000 and 40,000 simulated SNP with trait variance = 40 and h^2 = 0.5 82

Table 4.3 Number of true QTL discovered by DE 85

Table 4.4 Prediction accuracies for the three different approaches with sheep data 85

Table 5.1 Number of progeny in each family 95

Table 5.2 Estimation of heritability in gene expression data 99

Table 5.3 Percentages of variation in gene expression explained by different components 105

Table 2.S1 Summary of statistics for the SNP, average minor allele frequency and heterozygosity 131

Table 2.S2 Mean linkage disequilibrium in five populations over different map distances 132

Table 2.S3 Average linkage disequilibrium (r^2) between adjacent markers for each autosome (OAR) 133

Table 2.S4 Average linkage disequilibrium (|D'|) between adjacent markers for each autosome (OAR) 134

Table 2.S5 Chromosome wise average linkage disequilibrium (r^2) in five populations 135

Table 2.S6 Chromosome wise average linkage disequilibrium (D') in five populations 136

Table 2.S7 Chromosome wise haplotype analysis summary 137

Table 2.S8 Range of inbreeding coefficients in five populations 141

Table 3.S1 Distribution of SNP before and after quality control and the average distances between adjacent SNP on each chromosome 144

Table 3.S2 All 39 SNP showing significant association with body weight in 1,743 Merino sheep 145

Table 3.S3 The percentage of genetic variance explained by each chromosome 147