Efficient algorithms for using genotypic data

Mohammad H. Ferdosi

Master of Science in Genetics and Animal Breeding, Shahid Bahonar University of Kerman

Bachelor of Science in Animal Science, Shahid Bahonar University of Kerman

Diploma at National Organization for Development of Exceptional Talents

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

University of New England

June 2015

School of Environmental and Rural Science

Faculty of Arts and Science

Abstract

The aim of this thesis is to explore the specific structure in livestock populations to unravel hidden information such as recombination events and parental origin of markers in the genomic data. This information then can be used to improve the accuracy of prediction of breeding values which is one of the main aims of animal breeding.

In the first experimental chapter an efficient method for detecting opposing homozygotes was proposed. This method makes the detection of opposing homozygote for thousands of individuals and millions of markers feasible. An opposing homozygote matrix can be utilised to identify Mendelian inconsistency and to fix pedigree errors.

The second experimental chapter used opposing homozygotes between individuals in a half-sib family to identify recombination events in the sire, to impute sire haplotype and to reconstruct haplotype of offspring. The algorithm was compared with other frequently used methods, using both simulated and real data. The accuracy of detecting recombination events and of haplotype reconstruction was higher with this algorithm than with other algorithms, especially when there were genotyping errors in the dataset. For example, the accuracy of haplotype reconstruction was around 0.97 for a half-sib family size of 4 and the accuracy of sire imputation was 0.75 and 1.00 for a half-sib family size of 4 and 40, respectively.

In the third experimental chapter *hsphase* was developed which implements the algorithms used in the first two chapters into an efficient R package. In addition, an algorithm for grouping half-sib families utilising the opposing homozygote matrix was developed and verified with real datasets. The results show that the algorithm can group the half-sib families accurately, however the accuracy was depended on sample size and genetic diversity in the population. The package

includes several diagnostic functions to visualise and check half-sib's pedigree, parentage assignments, and phased haplotypes of offspring in a half-sib family.

The fourth experimental chapter utilised the half-sib population structure to fix switch errors. The switch error is a common problem in many haplotype reconstruction algorithms where the haplotype phase is locally correct but paternal and maternal strand are not consistently and correctly assigned across the longer segments (or across the entire genome). The algorithm partitions the genome into segments and creates a group matrix which is used to identify the switch points. Then the switches are fixed with a second algorithm. The results showed that this algorithm can fix the switch problems efficiently and increase the accuracy of genome-wide phasing.

In chapter five relationship matrices generated from haplotype segments were used to improve the accuracy of predicting breeding values. The haplotypes were partitioned in three ways and with various size. The new relationship matrices were evaluated with three sets of real data and with simulated data. In all cases the accuracy of prediction and log-likelihood were significantly increased although the amount of increase was trait dependent.

Declaration

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

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

Mohammad H. Ferdosi

Date: 8/6/2015

Acknowledgements

I would like to appreciate the support, flexibility, encouragement and patience of my principal supervisor Professor Julius van der Werf and my co-supervisors Associate Professor Cedric Gondro and Professor Bruce Tier. I am in huge debt for their advice.

I would like to thank Professor Brian Kinghorn and Dr Vinzent Borner for their advice, comments and valuable discussions.

I would like to thank postgraduate students and my friends.

I would like to acknowledge the staff of school of environment and rural science and animal breeding and genetic unit, especially, Ms Shirley Fraser and Mr Klint Gore.

Finally, this thesis has become feasible with great support and encouragement from my family, especially my kind parents.

1. Contents

List of tables:

List of figures