

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

Chapter 1.	General Introduction	1
Chapter 2.	Review of Literature	6
2.1.	Parentage assignment and pedigree reconstruction.....	6
2.1.1.	Likelihood and frequency based methods for parentage assignments.....	6
2.1.2.	Using opposing homozygotes for parentage assignment and pedigree reconstruction	7
2.1.3.	Probability methods for parentage exclusion.....	8
2.1.4.	Application of parentage assignment.....	8
2.2.	Linkage disequilibrium between markers	9
2.2.1.	Haplotype blocks	10
2.3.	Haplotype inference	11
2.3.1.	Monte Carlo method for haplotype reconstruction in a half-sib family	11
2.3.2.	Optimal haplotype reconstruction in half-sib families.....	12
2.4.	Genomic selection	12
2.4.1.	Haplotype strategies for genomic selection	16
2.5.	Conclusions	17
Chapter 3.	A fast method for evaluating opposing homozygosity in large SNP data sets	19
3.1.	Abstract	19
3.2.	Introduction	20
3.3.	Method	21
3.4.	Result and discussion	23
3.5.	Conclusion.....	24
3.6.	Conflict of interest.....	24
3.7.	Acknowledgements	24
3.8.	Appendix A	24
Chapter 4.	Detection of recombination events, haplotype reconstruction and imputation of sires using half-sib SNP genotypes.....	28
4.1.	Abstract	29
4.1.1.	Background.....	29
4.1.2.	Methods.....	29

4.1.3.	Results.....	29
4.1.4.	Conclusions.....	30
4.2.	Background.....	30
4.3.	Methods.....	33
4.3.1.	Algorithm.....	34
4.3.2.	Phasing of genotypes of half-sib families.....	39
4.3.3.	Estimation of phasing accuracy using strand of origin.....	40
4.3.4.	Performance comparison.....	42
4.3.5.	Calculation of R^2 and switch error rates.....	43
4.3.6.	Implementation.....	43
4.4.	Results and discussion.....	44
4.4.1.	Simulated data.....	44
4.4.2.	Real data.....	55
4.5.	Competing interests.....	59
4.6.	Authors' contributions.....	59
4.7.	Acknowledgements.....	59
Chapter 5.	<i>hsphase</i> : an R package for pedigree reconstruction, detection of recombination events, phasing and imputation of half-sib family groups.....	63
5.1.	Abstract.....	64
5.1.1.	Background.....	64
5.1.2.	Results.....	64
5.1.3.	Conclusion.....	64
5.2.	Background.....	65
5.3.	Implementation.....	67
5.4.	Main functions in <i>hsphase</i>	73
5.4.1.	Input data.....	73
5.4.2.	Pedigree reconstruction and parentage assignment.....	73
5.5.	Results.....	80
5.5.1.	Pedigree reconstruction.....	80
5.5.2.	Accuracy of sire inference and imputation.....	84
5.5.3.	Recombination events.....	84

5.6.	Conclusion.....	85
5.7.	Availability and requirements	86
5.8.	Competing interests.....	86
5.9.	Authors' contributions.....	86
5.10.	Acknowledgements	87
Chapter 6. Evaluation of haplotype reconstruction accuracy and correction for switch errors in half-sib families		91
6.1.	Summary	92
6.2.	Introduction	93
6.3.	Materials and Methods.....	95
6.3.1.	Identification of paternal strand with phased genotypes on sire and progeny	95
6.3.2.	Identification of paternal haplotypes in half-sibs without sire's genotype	96
6.3.3.	Method validation with real and simulated data	99
6.3.4.	Identification of switch errors.....	100
6.3.5.	Evaluation of <i>BEAGLE</i> software	100
6.4.	Results and Discussion.....	100
6.4.1.	Analyzing switch errors in <i>BEAGLE</i> results.....	100
6.4.2.	Accuracy of switch detection with simulated data	102
6.5.	Conclusion.....	105
6.6.	Acknowledgements	105
Chapter 7. Identification of optimum haplotype length to build the genomic relationship matrix 108		
7.1.	Abstract	109
7.1.1.	Background	109
7.1.2.	Methods.....	109
7.1.3.	Results.....	109
7.1.4.	Conclusions.....	110
7.2.	Introduction	110
7.3.	Materials and methods	112
7.3.1.	Data	112
7.3.2.	Haplotypes	113

7.3.3.	Relationships.....	114
7.3.4.	Variance components.....	117
7.3.5.	Cross-validation.....	117
7.4.	Results.....	118
7.4.1.	Brahman haplotype diversity.....	118
7.4.2.	Similarity of G and H_{11}	118
7.4.3.	Simulated data:.....	120
7.4.4.	Real data:.....	121
7.5.	Discussion.....	126
7.6.	Conclusions.....	130
7.7.	Competing interests.....	130
7.8.	Acknowledgements.....	130
Chapter 8.	General Discussion.....	133
8.1.	Future research.....	137
8.2.	Conclusion.....	138

List of tables:

Table 3-1 Computation time and real time of method MA and LOOP as a function of the number of genotyped individuals and number of SNPs.23

Table 4-1 R^2 +/- standard deviation across replicates between inferred and true results, and percentage of assigned results using simulated data (dataset B).....48

Table 4-2 Accuracy of haplotype inference in comparison to the true haplotype.....51

Table 6-1 mean of R^2 +/- standard deviation across replicates between true haplotype and fixed haplotype.....103

Table 6-2 difference between average of R^2 's after and before fixing switches105

Table 7-1 Number of animals (N), mean (μ), standard deviation (SD) and heritability (h^2) for different traits113

Table 7-2 Log-likelihood, residual variance (σe^2), additive variance (σa^2) and intercept (μ) utilizing simulated data and different methods.....121

Table 7-3 Intercept, coefficient and R^2 of linear regression of elements of inverse relationship matrices from models using the different relationship matrices for the simulated data.....121

Table 7-4 Windows size, log-likelihood, residual variance (σe^2), additive variance (σa^2) and intercept (μ) with the best log-likelihood utilizing real data and different methods.....123

Table 7-5 Windows size and amount of improvement of the different methods in comparison to G with the best accuracy utilizing real datasets – AC: Windows size for the best accuracy, LL: Accuracy and windows size for the best log-likelihood125

List of figures

Figure 4-1 **Description of the hspbase algorithm for eight SNPs and three half-sibs.** **A.** Genotype matrix (0, 2 are homozygotes, 1 are heterozygotes), blue genotypes highlight informative loci (opposing homozygotes – heterozygous in the sire, homozygous in the offspring). **B.** Identification of the paternal origin of strands and recombination events (x) by using the Forward Memory Vector. (P and M are arbitrarily assigned paternal and maternal strands of the sire). **C.** Final result of blocking after filling in the loci for which the sire strand could be determined. P and M refer to the two strands in the sire; - (dash) is used for unknown strand. **D.** Imputation and phasing of the sire by combining the original offspring genotypes and the block structure.36

Figure 4-2 **Removal of genotyping errors.** **A.** Genotyping errors in the half-sibs. A highlighted genotype is indicative of a genotyping error because only one marker supports its change to another block. **B.** Fixing the genotyping errors in the half-sibs. The block structure is used to reject the recombination suggested by the genotype (it is not supported by downstream markers). **C.** Fixing the genotyping errors in the imputed sire. Based on the blocking structure in the genotypes, individuals 2, 3, and 4 received the marker from the sire’s M strand (blue); the average number of markers (haplotype) in this location is 0.7, which is closer to 1 than 0 and this value is used as the sire’s imputed SNP genotype.39

Figure 4-3 **The sire’s imputed haplotypes and block structure information are used to phase the offspring.** 40

Figure 4-4 **Example of paternal strand blocking structure in a half-sib family with 40 individuals, using simulated data on 10 000 markers (dataset C).** **A.** true paternal strands of origin from the sire; **B.** strand assignment (blocks) with hspbase, empty spaces indicate unassigned regions; **C.** half-sib phasing with hspbase; **D.** half-sib phasing with AlphaPhase with use of pedigree; **E.** half-sib phasing with AlphaPhase without use of pedigree; **F.** half-sib phasing with BEAGLE; **G.** half-sib phasing with PedPhase; the red and blue colours indicate the paternal and maternal strands of the sire within each offspring, obtained by comparing the phased data with the sire’s true haplotypes; empty spaces indicate unknown strand or haplotype.47

Figure 4-5 **Accuracy of haplotype reconstruction (hspbase, BEAGLE and AlphaPhase).** **A.** Boxplot of R^2 between inferred and true haplotypes for 20 half-sib families. Dataset C includes 10 000 SNP markers and 20 half-sib families with 40 individuals per family. **B.** Boxplot of R^2 between inferred and true haplotypes for the same data with 1% random genotyping error.53

Figure 4-6 **Accuracy of haplotype reconstruction (hspbase and PedPhase).** **A.** Boxplot of R^2 between inferred and true haplotypes. **B.** Boxplot of switch error rates. H = hspbase; P = PedPhase with 5, 10 and 20 half-sibs per family; E = with 1% random genotyping error.54

Figure 4-7 **Example of paternal strand blocking structure built from a phased half-sib family with 23 individuals using real data genotyped on the 50k Ovine Illumina array.** **A.** hspbase; **B.** AlphaPhase with pedigree; **C.** pedigree free AlphaPhase; **D.** BEAGLE; **E.** PedPhase; red and blue colours indicate the sire’s strands inherited by each offspring, obtained by comparing the phased data with the sire’s phased haplotypes, as inferred by the respective methods (the sire’s genotypes were part of the dataset); empty spaces indicate unknown strand or haplotype.56

Figure 4-8 **Accuracy of sire imputation utilising real data.** R^2 values (in blue) between imputed and observed sire genotypes (for sheep chromosome 1), with different numbers of half-sibs per family (110 groups, mean = 0.95, median = 0.98) and percentage of imputed SNPs (in red) in the sires. The vertical line is for families of size 10.58

Figure 5-1 **Density plot of opposing homozygotes.** The three distributions show the number of opposing homozygotes between parents and offspring (black), between half-sib families (blue) and between unrelated individuals (red). The distributions are genome wide and based on 290 Hanwoo cattle from 36 sires genotyped on the 700 k Illumina BeadChip. The separation between parent-offspring and other relationships is very clear but there

is some level of overlap between the distributions of half-sibs and unrelated animals which can make it not possible to perfectly characterize family groups. The level of separability is conditional on the overall genetic variation in the population and is better with genetically diverse populations. For reference purposes, Hanwoo cattle have small effective population size (~100) and are subject to some ascertainment bias in the array design which further constrains detectable variation.....69

Figure 5-2 **Separation value between true and false parent-offspring relations.** The separation value is the difference between the smallest number of opposing homozygotes found across all false sire-offspring relations (i.e. all pairwise combinations except the real sire-offspring pairs) and the maximum number of opposing homozygotes in the correct sire-offspring pairs divided by the maximum number of opposing homozygotes found in the dataset. Positive values allow reliable identification and exclusion of parent-offspring relationships. The figure shows sorted pair-wise combinations of opposing homozygotes for 326 Hanwoo cattle genotyped on the Illumina 700 k BeadChip (290 offspring and their 36 sires); true sire-offspring relations in black and false in red. The jagged line shows a cut off threshold of 1% genotyping errors.....71

Figure 5-3 **Block structure for a half-sib family.** The block structure shown is for chromosome 29 of eight half-sib Hanwoo cattle genotyped on the 700 k Illumina BeadChip array. The two colours indicate the paternal and maternal strands from the paternal line each individual inherited (i.e. the individual's relationship with the two haplotypes of the sire); white regions are areas where the phase could not be determined.72

Figure 5-4 **Correction of pedigree errors.** The heatmap shows the relationships between individuals based on opposing homozygotes. Half-sib families are colour coded. On the vertical bar (left-hand side), 4 individuals are misclassified in the phenotypic pedigree. The horizontal bar (top) shows the reconstructed pedigree using the rpoH function in hspHase. Data is for 106 Hanwoo cattle from 14 family groups genotyped on the 700 k Illumina BeadChip array. Four pedigree records were purposely swapped. The darker blocks on the diagonal help identify half-sib groups.....74

Figure 5-5 **Sorted pairwise numbers of opposing homozygotes with ohplot.** The plot is useful to guide decisions for parameter settings to reconstruct the pedigree. The separation value is the maximum separation found between sorted value pairs divided by the maximum number of opposing homozygotes. The cut off value is the number of SNP at the mid-point of the largest separation (red line). The other lines are the average number of opposing homozygotes expected in full-sib families (cyan – not shown here), half-sib families (green) and in unrelated individuals (blue) according to (Calus et al., 2011); plus the 90% threshold used for pedigree reconstruction (pink). The function can also be used with pedigree information to detect inconsistencies (relations are colour coded according to the pedigree to facilitate visualization). The plot is from 106 Hanwoo cattle from 14 family groups genotyped on the 700 k Illumina BeadChip array.....76

Figure 5-6 **Phasing error detection using blocks.** The two colours indicate the paternal and maternal strands from the paternal line each individual inherited. The large number of recombinations and unresolved areas (in white) is indicative of poor phasing results. The figure was simulated by using 31 random and unrelated individuals (bovine chromosome one).....78

Figure 5-7 **Distribution of recombination events.** The plot shows the probability of recombination along chromosome one (46495 SNP) for 106 Hanwoo cattle; this is useful to identify local variation in recombination rates. In this case the population size is too small to reliably detect these regions. Note however the high value in the first part of the chromosome, these are not true recombinations but mapping errors instead, probably due to an assembly error in the reference sequence.....79

Figure 5-8 **Comparison of pedigree reconstruction methods.** Left-hand side colour bars for true family relationships, top-side bars for inferred pedigree and pedigree matching. The top colour bars are respectively, **A.** Manually determined threshold for number of opposing homozygotes in a half-sib family (31,662). The pedigree is 100% accurate and correctly assigned to the sires. **B.** Regression coefficients derived from a large sheep population. Two families did not separate (yellow and red – top bar) and were classified as a two groups instead of four. **C.** The third method based on (Calus et al., 2011) failed to assign 9 individuals to their correct sire (unassigned individuals

shown in white). Method four uses number of recombinations to sort family groups. Here chromosome one was used with a maximum of 10 recombinations allowed. The method failed to group 5 individuals from one family (unassigned individuals shown in white)..... 82

Figure 5-9 **Error detection using blocks.** Plots generated using the imageplot function. **A.** The last individual is wrongly assigned to the family group (excessive number of short recombinations). Note that the inclusion of a wrong individual causes problems with block assignment for the whole family (compare the block structure for the same group in Figure 3). Data shown is for bovine chromosome 29 with 8 half-sib and an unrelated individual intentionally included. **B.** Putative map error on chromosome 1. A map error has a cumulative effect and causes problems downstream evidenced by chromosome blocks consistently breaking across all individuals..... 83

Figure 5-10 **Recombination pattern on chromosome 6 in sheep.** The figure shows regions of high and low recombination in a large sheep population genotyped on the Ovine 50 k Illumina BeadChip. White areas are regions without SNP coverage on the array. At the end of the chromosome there is a region with over twice the average number of recombinations. 85

Figure 6-1 **Description of the algorithm for a group matrix creation.** **A.** Matrix of sire haplotypes and offspring' haplotypes (haplotype matrix). **B.** Covering partition of haplotype matrix (Here for simplicity the partition sizes are 3 SNPs but the algorithm uses the partition size of 100 SNPs – there are only 7 partitions for 10 SNPs – the partitions show inside the horizontal square bracket). **C.** Grouping each partition based on their similarities to create the group matrix (partitions 1, 4 and 7 are shown). 96

Figure 6-2 **Observation of sire's haplotypes in the strands of its progeny – red and blue strands originated from the haplotype of the sire – the white area are the dam's strand (frequency less than two in the partition) or the recombination regions in last case their length should be equal to 100 columns in the group matrix (simulated data – dataset A).** **A.** Sire's haplotypes were used to create the group matrix (the first and second strands are sire's haplotype). **B.** The group matrix was created utilizing the progeny haplotypes only. 97

Figure 6-3 **Fixing the group matrix (simulated data – dataset A - genotyping errors and mapping errors can cause some rotations that can be easily detected by imageplots but they do not occur in the same locus).** **A:** Imageplot of the group matrix before fixing rotations. **B:** Imageplot of the group matrix after fixing the rotations... 98

Figure 6-4 **Imageplot of group matrix to illustrate the detection of switch errors (chromosome 1 – A half-sib family of Dataset C phased with BEAGLE – 10 half-sibs – 20 haplotypes)** **1.** Switch error. **2.** Phasing or genotyping errors. **3.** Mapping error. **4.** Genotyping or phasing errors (This can be also because of IBD regions that are common in both dam's and sire's haplotypes). **5.** Recombination event..... 101

Figure 7-1 **Description of methods for creation of relationship matrix.** **A.** Haplotypes for two example individuals with 5 SNPs. **B.** Each haplotype section is made up of SNPs of the same length (2 SNPs), with the last window potentially using SNPs that may have been used in the penultimate window. **C.** Haplotypes of length 2 are constructed from adjacent pairs of SNPs with SNPs appearing in more than one window. **D.** Total similarity. The total value of contiguous sections that are identical in pairs of haplotypes and at least the minimum length are counted. 115

Figure 7-2 **The haplotype diversity for chromosome 1 (Brahman Cattle) using two methods and different windows size for all segments.** **A.** Number of unique haplotypes requires to explain 60 percent of the haplotypes. **B.** Number of unique haplotypes requires to explain 90 percent of the haplotypes..... 118

Figure 7-3 **The log-likelihood of three methods to build different relationship matrices for three traits (SC, AGECL and WTCL) and different window sizes.** The black horizontal represents the G result. DW – Distinct Windows, SW – Sliding Windows and TMS – Total Minimum Similarity 122

Figure 7-4 **Additive and residual variances of three methods to build different relationship matrices for three traits (SC, AGECL and WTCL) and different window sizes.** The black horizontal represents the G result. DW – Distinct Windows, SW – Sliding Windows and TMS – Total Minimum Similarity. 124

Figure 7-5 **The accuracy of three methods to build different relationship matrices for three traits (SC, AGECL and WTCL) and different window sizes.** The black horizontal represents the G result. DW – Distinct Windows, SW – Sliding Windows and TMS – Total Minimum Similarity. 126