COMBINING HIGH AND LOW DENSITY SNP PANELS TO IMPUTE GENOTYPES FOR NON-TYPED ANIMALS WITH APPLICATION IN GENOMIC BREEDING VALUE ESTIMATION

by

HUIYU WANG

(Under the Direction of Romdhane Rekaya)

ABSTRACT

Availability of reliable genotyping platforms for single nucleotide polymorphism markers (SNP) has made genomic breeding value (GBV) estimation a reality. Unfortunately, genotyping is still expensive and its use at large scale requires SNPs genotypes of non-typed animals to be inferred from genotyped animals. However, relationships and allele frequency information could be limited. To overcome this problem we proposed combining genotyping information from high and low density SNP panels. This low density and low cost chip will provide an additional source of information, linkage disequilibrium, in inferring missing genotypes.

The proposed procedure was successful in increasing the probability of inferring true SNP genotypes for the non-typed animals by 12 to 18% depending of the simulation parameters. It increased accuracy of estimated GBVs by 3 to 12% depending on the number of SNPs and genotyped animals. These results suggest that this procedure could provide a cost effective tool for large genomic evaluation.

INDEX WORDS: SNP, Low density, High density, Genotype imputation, Genomic selection

COMBINING HIGH AND LOW DENSITY SNP PANELS TO IMPUTE GENOTYPES FOR NON-TYPED ANIMALS WITH APPLICATION IN GENOMIC BREEDING VALUE ESTIMATION

by

HUIYU WANG

B.S., China Agricultural University, China, 2007

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIE

2009

COMBINING HIGH AND LOW DENSITY SNP PANELS TO IMPUTE GENOTYPES FOR NON-TYPED ANIMALS WITH APPLICATION IN GENOMIC BREEDING VALUE ESTIMATION

by

HUIYU WANG

Major Professor:	Romdhane Rekaya
Committee:	Ignacy Misztal J. Keith Bertrand

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2009

TABLE OF CONTENTS

Page

LIST OF TABLES v

LIST OF FIGURES vi

CHAPTER

1 INTRODUCTION 1

2 LITERATURE REVIEW 4

3 IMPUTING MISSING GENOTYPES USING AN ANT COLONY APPROACH 31

4 COMBINING HIGH AND LOW DENSITY SNP PANELS TO IMPUTE GENOTYPES FOR NON-TYPED ANIMALS WITH APPLICATION IN GENOMIC BREEDING VALUE ESTIMATION 49

5 CONCLUSIONS 63

APPENDIX

A ABBRIVATIONS 64

LIST OF TABLES

Page

Table 2.1: Technical requirements and characteristics 26

Table 3.1: Mean, minimum and maximum of the minor allele frequency (MAF) and linkage disequilibrium (LD) for the 50 K SNP panel used in the simulation..44

Table 3.2: Percentages of genotypes correctly imputed (FK), and those identified with a probability greater then .8 (Pr.>.8) after the ant colony step 45

Table 3.3: Percentages of genotypes imputed with a probability smaller then .2 (Pr.<.2) after the ant colony step 46

Table 3.4: Percentages of genotypes correctly imputed (FK), and those identified with a probability greater then .8 (Pr.>.8) at the end of the sampling process 47

Table 3.5: Percentages of genotypes imputed with a probability smaller then .2 (Pr.<.2) at the end of the sampling process 48

Table 4.1: Increase in accuracy for partially genotyped animals using 1000 fully genotyped and phenotyped animals (average over 5 replicates) 61

Table 4.2: Increase in accuracy for partially genotyped animals using 2000 fully genotyped and phenotyped animals (average over 5 replicates) 62

LIST OF FIGURES

Page

Figure 2.1: Milk for Holstein or Red & White 27

Figure 2.2: QTL effect distribution for pig and dairy cattle and estimated proportion of variance..28

Figure 2.3: Average linkage disequilibrium (r^2) as a function of average genomic distance for multiple breeds of cattle 29

Figure 2.4: Genotyping on DNA microarrays procedure 30

CHAPTER 1

INTRODUCTION

Whole genome selection (GS) has recently become available due to the feasibility of high-throughput genotyping technologies and commercial platforms. Following completion of the human genome sequence in 2001, genome databases of several livestock species went under construction. For example, the map of genome sequence variation in chicken has become available, in 2004, with more then of 2.8 million single-nucleotide polymorphisms (SNPs) in 2004.

A two step procedure is currently used to implement the genomic selection approach which: 1) estimation of the effects of chromosome segments of the whole genome in a reference population, 2) prediction of genomic breeding values (GEBV) of selection candidates by applying estimated marker effects from training data to their marker genotypes. However, the two step procedure requires genotypes for both the reference dataset and selection candidates because GEBVs are only available for animals with genotypes. Thus, SNP genotypes for non-typed animals must be inferred from three types of information: allele frequencies, linkage disequilibrium and relationships with genotyped animals. Unfortunately, the relationship information is often limited because pedigrees span multiple generations for domestic animals. Moreover, GS using dense marker maps that require expensive high density chips that can not be used at the population level due to limited budgets.

Low density SNP panels have been developed through feature selection to reduce high dimension datasets. There are three advantages of using low density marker panels. Firstly, low density chips are much less expensive than high density chips thus they are relatively more cost effective. Secondly, with reduced genotyping cost, larger numbers of individuals of a population can be genotyped including some females. Lastly, due to the larger sample size being available, low cost panels will provide an additional source of information through linkage disequilibrium resulting in an increase in the detection of association between specific alleles and target phenotypes. Consequently, low density panels could potentially be an efficient approach to estimate GEBV for non-typed animals.

To optimize parameters critical to successful GS, numerous strategies have been developed. Meuwissen et al. investigated stepwise testing approaches to select parameters arbitrarily based on stringent significance tests. Habier et al. regarded methods of searching informative SNP subsets as “variable selection” and presented evenly-spaced SNP panels (ELD-panel) with broad application of GS but more loss of accuracy in estimating breeding values. Bayesian feature selection can be applied by adding indicator variables whose prior follows an independent Bernoulli distribution. Machine learning techniques select influential features efficiently employing the classifier, and are superior to other methods because of their robustness and power of information reduction. More recently, a two-step feature selection scheme was proposed in supervised learning (classification) using a naïve Bayesian classifier for mortality traits in broilers. It consisted of filtering and wrapping steps aiming to reduce the number of markers from all features then optimizing the filtered SNPs.

An alternative strategy recently being investigated in genomic selection is the Ant Colony Algorithm (ACA). It is an optimization algorithm capable of incorporating prior information, allowing it to search the sample space efficiently. The ACA has been proved efficient in high-dimension and was first applied to solve traveling salesman problems based on a mechanism by which ant colonies uncover the shortest/best route to a food source. Relying on a positive feed back system, ants communicate through the use of chemicals called pheromones which are deposited along the trail an ant travels. The shorter route will take ants less time thus more pheromone is accumulated generating a stronger signal to attract more ants. Artificial ants of ACA work analogously to real ants with the pheromone function updated by prediction accuracy.

The objective of this study was to develop a procedure that employ the Ant Colony Algorithm to select low density panels then combine high and low density panels to estimate GEBV. Different scenarios were simulated to evaluate the efficiency of the procedure.

CHAPTER 2

LITERATURE REVIEW

Traditional animal breeding: Development and Achievements

The concept of “traditional animal breeding” has been in used for genetic evaluation by animal breeders since the middle of the last century. Its development has taken much longer and is still under improvement. With the primary aim in maximizing the rate of genetic improvement (i.e., a combination of genetic values for several economic traits), data, which represent the observable records of performance, have become critical for analysis together with sophisticated and often complex statistical and computational procedures. There are several crucial landmarks that have shaped the history of animal breeding and genetics. Habier and Gianola (2000) presented a comprehensive review of major statistical developments. From selection index to Henderson’s mixed model equation (MME) procedure, restricted maximum likelihood (REML), and Bayesian approach, appropriate estimation and computational tools have lead to an unprecedented level of accuracy in predicting genetic merit. Furthermore, computer power has brought new opportunity for analyzing high-dimensional datasets in a reasonable computational cost. Through the integration of several braches of science, which include genetics, statistics, computer science and reproductive physiology, selective breeding has stepped into a new level. The achievement of traditional animal breeding is dramatic and it has been proved effective. In dairy cattle, for example, breeding values for milk yield in Holstein and Red & White cows calculated by USDA (April 2009) show a clear positive trend from one generation to the other and a spectacular increase in the genetic potential of the dairy population (Figure 1). Although the traditional animal breeding approach has developed dramatically fast and brought remarkable economic benefits, several limitations still exist: 1) Time consuming: Traditional breeding schemes are largely based on progeny or sib-families testing which is often a very lengthy process especially for animals with large generational intervals. The average generation interval for dairy cattle, for instance, is about five years. 2) Inefficient for specific traits: For traits that are sex-linked, difficult or expensive to measure including those that could be collected only late in life, such as longevity, traditional breeding methods are proved to be less effective and sometimes inappropriate (Goddard and Hayes, 2007). This is in part due to the limited amount of phenotypic information as a result of measurement difficulties or cost, low heritabilities, and often the inadequate statistical procedure used for analysis. 3) Confounded variation: The available sources of information of traditional breeding strategy are all observable data as pedigree and phenotypes. However, the objective of breeding process is looking for the genetic variation which is confounded with environmental part thus impossible to be observed directly. What complicated the situation even further is the fact that these so called “secondary traits” have been neglected for long period of time as the profitability of livestock operations was largely dominated by production traits. This is no more the situation in part due to the huge success in selecting for primary traits which has lead to an ever shrinking differences between the top animals for these traits, the negative indirect correlative responses in some secondary traits (health and fertility traits), and the emerging concerns about animal welfare and environmental impact.

Development in Animal breeding after the availability of molecular information

All these limitations together have created the need to strengthen the current genetic evaluation procedures and to find new tools and sources of information that will allow animal breeders to examine more closely the source of genetic variation. As the genetic variation is the inherited and consistent source that causes performance difference, research on genes or chromosome segments which affect traits of interest has gained a lot of popularity. This interest in molecular approaches in the field of animal breeding has evolved for several decades simultaneously with the effort in exploring better statistical and computational methods. Early molecular research often based on “naïve” simulation scenarios has placed high expectations on the potential impact of molecular information for animal selection. Unfortunately that was not true and the extent of the practical use has fallen short of those initial expectations (Dekkers, 2004). However, the sequencing of the human genome and short after of several livestock species together with the recent advances in high through put technologies such as gene expression profiling and large scale single nucleotide polymorphism (SNP) genotyping have created an unprecedented opportunity for a more direct and feasible dissection of complex traits.

Candidate gene and QTL mapping

Early molecular methods used in the field of animal breeding and genetics consisted on candidate gene and QTL mapping approaches. A quantitative trait locus (QTL) is defined as a chromosomal segment with a Mendelian transmission pattern and with an effect on a trait of interest (Boichard et al., 2003). Unlike typical Mendelian traits, these economical traits in livestock are under the control of more than one single gene or chromosome region. In fact, the distribution of QTLs is “moderately leptokurtic” indicating that most genes have small effects and only few have large effect (i.e., leading QTL or major QTL) as suggested by Hayes and Goddard (Figure 2A). In dairy cattle and pigs experiments it was reported that 17% and 35% of the leading QTLs could explain almost 90% of genetic variance respectively (Figure 2B). If the genes which contribute to quantitative trait variation can be detected and located on the different chromosomes, a deeper understanding of the impact of quantitative traits on the performance would be gained with more confidence. That was the original premise of the QTL detection approach. Unfortunately, such endeavor proved to be much more complex and even unrealistic. There are primarily two methods to detect QTLs: candidate gene approach and QTL mapping.

A candidate gene is a gene that is suspected to be involved in the expression of a particular trait. Although several mutant traits and disease phenotypes are likely to be caused by mutation in candidate genes, for quantitative genes, at least two main problems arose for testing variation association between DNA sequence and phenotypic traits which are the insufficient number of animal samples and the false positive rate (Hayes, 2007).

QTL mapping approach, in contrast, assumes that the actual genes impacting the quantitative trait are unknown. It involves linkage maps construction and QTL analysis indentifying polymorphic makers associated with target traits (Collard et al., 2005). A significant association between traits and markers may implicate that a QTL is close to the marker. QTL mapping is a combination of traditional quantitative genetics method and linkage mapping and has demonstrated the potential to model quantitative traits at the individual gene level (Liu, 1998). Single-marker analysis and interval mapping (including simple and composite interval mapping) are two basic methods to detect QTL. QTL mapping statistics often involve the logarithm of odds (LOD) scores and some permutation testing in order to account for the false positive detection (Liu, 1998). Methodological concerns for QTL mapping approach are important for accuracy. On one hand, large number of progeny per family or half-sib family is required to reduce the confidence interval for QTL location. On the other hand, dense maker maps, in excess of 200 markers, are often needed.

Marker-Assisted Selection (MAS)

Following a positive QTL identification in the mapping step using linked DNA markers, this information becomes the foundation for the implementation of a marker assisted selection (MAS) procedure. Furthermore, DNA makers have been deemed as the third data resource for selection strategy besides pedigree and phenotypic information (Hayes, 2007). Development in maker technology was notable. Conversion from low reproducibility and requirement for complex, time- and cost-consuming marker technique (e.g. RAPDs, RFLPs or AFLPs) to reliable and robust makers (e.g. SCARs or STSs) has been achieved (Collard et al., 2005). There are three types of marker for MAS defined as direct makers, LD (linkage disequilibrium) markers and LE (linkage equilibrium) makers (Dekkers, 2004). Direct makers are functional mutation loci and selection on them will be referred to as gene-assisted selection (GAS); LD and LE makers are loci in population–wide linkage disequilibrium and equilibrium with functional mutations on which selection will be called as LD-MAS and LE- MAS, respectively (Dekkers, 2004). It was originally speculated that the extraordinary development of MAS techniques will imply a revolutionary era in the field of animal breeding especially after dense maker maps has become available. However, that was not the case because the reliability for MAS depends on the accuracy of the utilized QTLs which are obtained in QTL mapping experiment. Unfortunately, QTL mapping, which is often based on linkage mapping that needs huge amount of progeny and dense markers on chromosomes, in order to retain recombination across generations, failed to detect real or even major QTLs instead of false positive errors and misestimate the QTL effects. Additionally, LE based mapping being the easiest method for detection, provided the hardest molecular information to use in a breeding program as the phase of the resulting QTLs are family specific. Moreover, even in the more favorable scenario, MAS could not outperform the traditional selection methods in the long term although it could lead to a faster genetic progress in the short term. These results are not surprising as the favorable alleles frequencies increase towards fixation over time (Hayes, 2007).

Microarray and high throughput makers

Since 1986 when Santa Fe conference accouchement was posted, the Human Genome Project (HGP) has officially started and more than five million dollars were allocated to develop the needed resources and technologies. Soon after, low-resolution to moderate and even high-resolution genetic linkage maps were made available for several chromosomes culminating in the publishing of the first working draft of the human genome (Venter et al., 2001) that was shortly followed by the sequencing of several laboratory and domestic animal genomes (Hillier et al., 2004; Humphray et al., 2007; Nadeau et al., 2001; Womack, 2005). Such milestone has provided the necessary resource for the feasibility and identification of structural variation and polymorphisms. Microarray technology has been blooming since late 1990's due to HGP. It allows for the simultaneous profiling of a massive number of genes or proteins on a tiny chip. Generally speaking, there are two types of DNA microarrays depending on the to nucleic acid printed on the chip: cDNA array and oligonucleotide array (Huang et al., 2001). Oligonucleotide arrays have not gained popularity until the improvement in better specificity and sensitivity. They are more robust because cDNA arrays need to use a completely uncharacterized library for expression profiling and may raise the risk of array errors (Pennington and Dunn, 2001). More importantly, Microarrays are not only useful for gene or protein profiling but also for genotyping or analysis of DNA sequence variation. For high-throughput markers, single nucleotide polymorphism (SNP) markers are one of the most abundant and available markers in genome-wide scale. The SNP maker genotyping could be achieved mainly either by PCR or array-based technologies. It is worth mentioning that the later does not require the use of PCR products and both methods have comparable accuracy (Gunderson et al., 2005).

Introduction to microsatellites and SNPs

Microsatellites, which are also named as Simple Sequence Repeats (SSRs), are tandem repeated of DNA of up to six base pairs. They were the dominant type of DNA markers for animal and human genetics applications before SNP genotyping had become feasible due to cost reduction and advances in high throughput techniques. Compared with other types of markers (Table 1), microsatellite markers are more appropriate for human genetic maps. This is because microsatellite markers possess a higher level of polymorphism than other types of markers due to multiple alleles at a single microsatellite locus. Furthermore, microsatellite markers have higher variability and mutation rate attributable to codominance (Thuillet et al., 2002). This characteristic makes genetic mapping and detection of heterogeneity in a population possible (Brinkmann et al., 1998). Finally, microsatellites are easily genotyped using PCR procedure in the laboratory.

Single Nucleotide Polymorphism (SNPs) must occur in more than 1% of the population of interest. It represents a single nucleotide variation in the genome sequence. SNPs can be categorizes into three categories according to their location and function: 1) Coding-region SNPs (cSNPs): They are variations located in coding regions which can alter amino acid encoding procedure and therefore affect protein sequence; 2) Perigenic SNPs (pSNPs): Located either inside or labeled relative to the nearest gene, pSNPs may affect transcription or other DNA functions; 3) Random non-coding SNPs (rSNPs): They occur in intragenic region but are deemed as nonfunctional DNA sequences in genome even though they have possession of similar sequence with other types of SNPs which own the capability to regulate gene expression (Nebert, 1999).

Although SNPs have much less variability compared with microsatellites, their unsurpassed characteristics have made them widely applied to construct high-density genetic maps for genome-wide association studies (GWAS) or genomic selection (GS). Their advantages are mainly derived from the following three characteristics. First, is their abundance in a given genome. In humans, for instance, almost 18 million SNPs are available in a public database "dbSNP" maintained by National Center for Biotechnology Information (NCBI) with more than 6 million of them have been already validated. Second, SNPs are often assumed as biallelic markers, although theoretically any one of the four nucleotide bases could be present at each locus, the mutation mechanisms are able to explain the favorable bias on the ratio of transition (i.e., purine to purine or pyrimidine to pyrimidine) over transversion (i.e., purine-pyrimidine or pyrimidine to purine) (Vignal et al., 2002). Although microsatellites are multi-allelic and might be the most informative markers, the number for highly polymorphic microsatellites is barely up to 30 K until recently for human genome and in some other species (e.g., chicken) only limited numbers are available (Bahram and Inoko, 2007; Vignal et al., 2002). Third, SNPs are considered as stable markers with relatively low mutation rate (i.e., merit of evolutionary conservation). As to autosomes and the X chromosome, for example, the average mutation rate for each nucleotide is only 2.5×10^{-8} per generation (Tishkoff and Verrelli, 2003). This feature is the foundation for SNPs as qualifiedly stable markers. Last, the cost of genotyping using high throughput methods has been decreasing constantly (just over $200 per animal in 2008 for the 50K chip).

SNPs discovery and genotyping

As individual genotypes are the main source of information for association or prediction studies, efficient and cost-effective SNP genotyping procedure are of crucial importance. There are three criteria for an ideal genotyping scheme: high-throughput capability, cost and computational efficiency and availability for reliability and robustness of results (Peltz, 2005). SNP genotyping protocols are different according to variant requirements but microarrays are one of the most robust methods for investigating allelic variation. Additionally, microarray-based SNP genotyping system which is widely applied currently involve three fundamental components: allele distinction, signal detection and assay format (Lorincz, 2006).

For allele distinction, there are mainly five techniques to identify single base difference at a specific locus, which are allelic-specific hybridization, restriction enzyme cleavage, enzymatic ligation, primer extension by polymerization and structure-specific cleavage (Lorincz, 2006). Changes in Physical properties are used to detect SNPs, as radioactivity, molecular mass, and luminescence phenomena. Moreover, assay formats are usually commercialized as beads, glass slides, and gel matrices which can by easily visualized using a standard microarray scanner. Commercial platforms are offered by genomic service providers to make very reliable and efficient high-throughput SNP genotyping data available like Taqman, Illumina, and Affymetrix.

Application of SNP technologies

There are mainly two aspects for application of high density SNP genotyping assays which are referred as GWAS and GS. On one hand, combination of candidate gene approach and association-based fine mapping is becoming available to identify human disease causing gene(s) across the whole genome (Lai, 2001; Sladek et al., 2007). On the other hand, appropriate models and algorithms are under development aiming to predict accurate total genetic values for animals or plants (Meuwissen et al., 2001).

Genome Wide Association Studies

Genome wide association studies use high density genetic maps as tools mainly for classification, class prediction and diagnostics. In medical and pharmacogenetic research, it is used to classify heterogeneous diseases (e.g., cancer or type II diabetes) for diagnosis and treatment (e.g., test of individual metabolic responses to a given medicine). It is revolutionary to use common genetic variation across the complete genome to discover genetic associations with observable traits without segregation information (Pearson and Manolio, 2008; Van Ommen, 2008).

Pearson and Manolio (2008) proposed a procedure for in a typical GWAS. First, a reasonable population structure of samples is needed as the foundation for further analyses. A large number of individuals with disease or traits of interested should be collected as well as reasonable quantity of samples for comparison group (e.g., case-control design). Moreover, three steps, which are DNA isolation, genotyping and data review, are needed to ensure high genotyping quality. Additionally, appropriate statistical methods to uncover association between potentially useful SNPs (i.e., SNPs which passed a certain threshold) and disease/traits are needed. Finally, the results have to be validated in another independent population.

GWAS is a highly robust approach to classify the genetic variants which impact the disease/trait of interest. However, there are also problems that need to be resolved after the design of the experiment and the implementation of the statistical analysis. These issues are mainly centered on the potentially large the false-positive disvovery rate and the gene by gene and gene by environment interactions (Ziegler et al., 2008).

Genomic Selection

The completion of a working draft of the human genome sequencing project had lead to the proliferation of several genomic technologies with applications in several fields of science including animal breeding and genetics. Genomic selection is a MAS tool of predicting the total genetic breeding values using dense marker maps, especially SNPs which are able to cover the whole genome and they are often in linkage disequilibrium with a neighborhood QTL (Calus, 2008; Goddard and Hayes, 2007; Meuwissen et al., 2001). Dense marker maps would likely to guarantee that a QTL would be close to at least one marker and in sufficient linkage disequilibrium with them to successfully perform the prediction work. Linkage disequilibrium is usually caused by limited effective population size in livestock (i.e. as low as 100; for human, it is almost 10,000) and used for MAS across populations and generations (Hayes, 2007). Measure of LD is frequently based on r^2 which is less dependent on allele frequencies. Figure 3 presents a plot of the relationship between the pair-wise linkage disequilibrium (r^2) and the physical distance, in Kb, between markers.

According to the “moderately leptokurtic” distribution of QTLs, significance test is too stringent to accurately accept true QTL with small effects. Therefore, the scheme of GS is to skip the stringent significance testing and as an alternative simultaneously predicting the marker effects for all chromosomal regions or genes (Meuwissen, 2003; Meuwissen et al., 2001). The two step procedure of implementing GS consists of 1) predict the effects of markers according to genotypes and modified phenotypes (i.e. pseudo-phenotypes) of individuals for trait(s), and 2) calculate the genomic breeding values (GEBV) based on the estimated marker effects and their marker genotypes .

Compared with traditional breeding approach such as parent average (PA), genomic prediction could increase the reliability (squared correlation of true BV and GEBV) depending on trait and sample size in training and validation dataset. For example, for chicken feed conversion efficiency trait, the accuracy of genomic selection is almost four times higher than PA (Gonzalez-Recio et al., 2009; VanRaden et al., 2008). For progeny test based BVs whose reliability is already reasonably high, genomic prediction could increase the genetic gain by decreasing generation interval while keeping or slightly increasing accuracy of selection (Meuwissen, 2003; VanRaden et al., 2008).

Mixed linear methodology is often used to implement GS (Meuwissen et al., 2001).

$$y = \mu 1_n + \sum_i X_i g_i + e \qquad [1]$$

where: y is the corrected data vector often referred to as pseudo-phenotypes, μ is the overall mean, 1_n is a $n \times 1$ vector of ones; summation term is all chromosomal segments (or markers) or genes effects, and e is residual.

The implementation of model in [1] could be carried out using different methods and approaches. Meuwissen et al. (2001) presented four approaches consisting of Least-Squares (LS), Best Linear Unbiased Prediction (BLUP), BayesA and BayesB. Other approaches have followed soon after including Partial Least Square (PLS), BayesC (π), Bayesian LASSO (Least Angle Shrinkage Selection Operator), semiparametric approach and other machine learning methods (e.g., support vector machine).

1. Least Square estimation

Least-squares approach is the simplest method to estimate the effect of markers. It assumed markers as fixed effects. In matrix notation, the system of equations could be presented as:

$$\begin{bmatrix} 1_n{'}1_n & 1_n{'}X \\ X{'}1_n & X{'}X \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 1_n{'}y \\ X{'}y \end{bmatrix}$$

However, this approach is ararely used due to its multiple disadvantages (Goddard and Hayes, 2007). Using any meaneful high-density marker map, the number of explonatory variable (markers effects) is much higher than the number of observations leading to the well known small “n” large “p” problem. Even in the situation where LS could be used, significance test will allow only the detection of markers with large variance limiting the amount of total genetic variance that can be captured by the markers and consequently the accuracy of the GEBV (Meuwissen et al., 2001; Goddard and Hayes, 2007).

2. BLUP

Marker effects in the model are assumed to be independent and randomly distributed with a constant variance. The resulting mixed model allows the simultaneous estimation of the marker effects and to a higher accuracy of breeding values estimations (Goddard and Hayes, 2007). Furthermore, the model could be implemented using Henderson’s mixed model equations:

$$\begin{bmatrix} 1_n{'}1_n & 1_n{'}X \\ X{'}1_n & X{'}X + I\lambda \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 1_n{'}y \\ X{'}y \end{bmatrix}$$

Where λ refers to ratio of the residual and marker effects variances. Although this approach could lead to a more accurate estimates, several reports (Meuwissen et al., 2001, and ???) found that it could be out performed using Bayesian approaches which employ prior information on QTL effects.

3. Bayesian approach

The Bayesian approach differs from the previous method (BLUP) mainly through its ability to assign prior information for the marker effects. Depending on that prior information several Bayesian methods have been proposed and implemented. Additionally, the Bayesian approach relies of inferring the model parameters including marker effects through the repetive updating from their respective conditional distributions. In the animal breeding field, three main Bayesian implementations: BayesA, BayesB and BayesC (Dekkers et al., 2009; Fernando et al., 2007; Jack Dekkers, 2009; Meuwissen et al., 2001).

Bayesian methods assume that marker effects are random and depending on the prior information, varying number of elements of the vector g_i are assigned large probability of being zero or very close to zero. Contrarily to the BLUP approach were only one source of information is used, the Bayesian approach combines the data and prior information and all inferences are based on the resulting joint posterior distribution. For the model in [1], the Bayesian implementation requires prior specification for μ, g_i and e and sometimes their hyper-parameters. It is often assumed that:

Parameters	Priors
μ	constant
g_i	$(g_i \mid \sigma^2_{gi}, \pi)$ ~ $\mathrm{N}(0, \sigma^2_{gi})$; σ^2_{gi} ~ scaled inverse chi-square distribution
e_j (the j-th element of e)	~ $\mathrm{N}(0, \sigma^2_e)$; σ^2_e ~ scaled inverse chi-squared distribution

For BayesA, the genetic variances of marker are assumed locus specific and they would be updated in each iteration by Gibbs sampler. But although most of the marker effects are very close to zero, their variances are not zero. In contrast, BayesB assumes that not all of the marker could capture QTL variances; therefore for a large portion of SNP markers, their genetic variances are exactly equal zero thereby no effects. The prior for BayesB is that with a high proportion π, loci have no variance and the rest of the loci with density of $(1-\pi)$ have nonzero variance (Meuwissen et al., 2001). It can be referred as:

$$\sigma_{gj}^2 = 0 \qquad \text{with probability } \pi,$$

$$\sigma_{gj}^2 \sim \mathrm{X}^{-2}(\nu, S) \qquad \text{with probability } 1-\pi,$$

For parameters of the prior which follows scaled inverse chi-square distribution, Meuwissen et al. (2001) presented $\nu = 4.012$ and $S = 0.0020$ as the degree of belief and scale factor of σ_{gj}^2 when σ_{gj}^2 is nonzero. Comparatively, Xu (2003) defined as no prior information which is $\nu = 0$ and $S = 0$. π is defined as 0.95 in most situations. Furthermore, Gibbs algorithm is not appropriate no longer due to the zero variances for most markers. Metrololis-Hastings sampler is implemented for Bayesian analysis (Meuwissen et al., 2001).

The choice of priors could be either arbitrary or according to experiential learning. However, the results from BayesA and BayeB methods are dependent on prior which means the accuracy of genomic selection based on such approaches is affected by values of π, ν and S. Especially for the value of π, if we use incorrect scaled factor as the prior, the correlation between TBV and GEBV would be largely dependent on how close the π chosen to the true value(Dekkers et al., 2009). To solve this problem, BayeC approach was implemented to search for an appropriate π. BayesC approach is different from the other two Bayesian methods in two aspects. On the one hand, the priors for BayesC are changed as:

$$(g_i \mid \sigma_g^2, \pi) \sim \mathrm{N}(0, \sigma_g^2)$$

π is a prior but no longer with a fixed value of BayesB but following uniform distribution which generated samples from 0 to 1. Therefore, for each locus π is specific. And π is sampled from beta distribution in the conditional likelihood (Dekkers et al., 2009). On the other hand, the genetic variance for each marker is the same. This means, genetic variances of markers are not locus specific as BayeA analysis. Consequently, for BayesA and BayesB approach, shrinkage thought is applied and one would fail to know which markers were really informative and able to capture QTL variances. But BayeC method will be less dependent on priors and can estimate the true value of π to detect the markers with effects. The comparison of different methods is presented in one paper of Meuwissen et al. (2001) as:

Methods	$r_{TBV;EBV}+SE$	$b_{TBV;EBV}+SE$
LS	0.318 ± 0.018	0.285 ± 0.024
BLUP	0.732 ± 0.030	0.896 ± 0.045
BayesA	0.798	0.827
BayesB	0.848 ± 0.012	$0.946+0.018$

4. Bayesian LASSO

LASSO method is presented as “posterior mode estimates when the regression parameters have independent and identical Laplace priors”; it combines the good features of variable selection with the coefficient shrinkage produced by Bayesian regression and acknowledges at most (number of animals -1) nonzero regression coefficients (de los Campos et al., 2009; Park and Casella, 2008). Bayesian LASSO (i.e. BL) is its Bayesian version implemented with Gibbs sampling; compare with ridge regression of Gaussian prior, Bayesian LASSO, which uses Laplace (i.e. double-exponential) prior can make a faster shrinkage of weakly related parameters to 0 (de los Campos et al., 2009). The most essential parameter of BL is λ which is selected by marginal maximum likelihood (Park and Casella, 2008).

GS is a straightforward approach to predict the young animals but still have some problems (Dekkers et al., 2009; Hayes et al., 2008). On one side, the knowledge for informative and uninformative loci is not always complete and one can not promise given SNPs in sufficient LD with QTL; this means the correct model equation is a precondition of all the other analysis procedures and it will be difficult to prove its correctness. On the other, if the marker effects are assumed random, the variances for the informative marker loci and residual are actually unknown. However, in recent study, they are just arbitrarily assigned some values from experiences. And Meuwissen et al. (2001) had referred the accuracy of GS will be decreasing in the following generation. Therefore, whether it will be benefitable or profitable from a long-term point of view is still pending.

Low-density panel and feature selection

High density SNP markers are available due to feasibility of reliable and cost efficient genotyping platform at commercial level. But genotyping all candidates of GS with density makers is still expensive for animals. Actually, not all the makers are strongly associated with loci which could affect corresponding traits. According to BayesC methodology, only a small part of SNPs are informative and able to capture some additive variances due to QTL. Therefore, separated low-density panels can be selected with subset of relevant SNPs in strong LD with QTL for different traits and populations (Habier et al., 2009a; Habier et al., 2009b). There are several approaches to attain the low-density panel selection: variable selection strategy in least squares regression, evenly-spaced low-density panel , Bayesian analysis and machine learning approaches (Habier et al., 2009b; Long et al., 2007).

Habier et al. (2009) presented that for evaluating loss of accuracy of low-density panels selected based on high-density SNPs, results from BayesB analysis showed reasonably sound quality with smallest loss and was superior to other methods of evenly-spaced low-density panel and forward least squares regression which is the worst always. But evenly-spaced low-density panel has advantage in accordance with reality and independent of number of QTL; moreover, it can be used across the traits and population instead of single trait or only within family. For machine learning approach, Long et al. (2007) concluded two-step feature selection method which consisted of filter and wrapper. Filtering and wrapping steps are executed sequentially to remove irrelevant features to a small arbitrary number then use a classifier to achieve the optimization of feature subset performance. It is promising and effective but limited due to potentially inappropriate model.

Summary

Genomic selection is gaining a lot of popularity in the field of animal breeding and genetics due to it efficiency, ease of implementation and its ability to produce a breeding value that can be used directly in selection. Schaeffer (2006) concluded that compared with current progeny test approach, 3 to 4 times more genetic gain would be obtained from GS scheme however the cost was only 3% of today's expense. International dominance will be achieved if more countries implemented extensively genotyping and GS approach. More packages and software will be available for estimating the GEBV and validating selection work. Because nucleus/consortium herds with high quality and complete data would be feasible and widely used for all manner of traits in future, genetic evaluation will be implemented with different strategies in distinct herds (Schaeffer, 2006).

REFERENCES

Bahram, S., and H. Inoko. 2007. Microsatellite markers for genome-wide association studies. Nature Reviews Genetics 8.

Boichard, D. et al. 2003. Detection of genes influencing economic traits in three french dairy cattle breeds. Genetics Selection Evolution 35: 77-101.

Brinkmann, B., A. Junge, E. Meyer, and P. Wiegand. 1998. Population genetic diversity in relation to microsatellite heterogeneity. Hum Mutat 11: 135-144.

Buzdin, A., and S. Lukyanov. 2007. Nucleic acids hybridization: Modern applications. 1st ed. Springer.

Calus, M. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178: 553.

Collard, B., M. Jahufer, J. Brouwer, and E. Pang. 2005. An introduction to markers, quantitative trait loci (qtl) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica 142: 169-196.

de los Campos, G. et al. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigrees. Genetics.

Dekkers, J. 2004. Commercial application of marker-and gene-assisted selection in livestock: Strategies and lessons. Journal of Animal Science 82: E313.

Dekkers, J., D. Garrick, and R. Fernando. 2009. Use of high-density SNP genotyping for genetic improvement of livestock No. 2009. Accessed on Dec. 07, 2009. http://www.ans.iastate.edu/stud/courses/short/

Fernando, R., D. Habier, C. Stricker, J. Dekkers, and L. Totir. 2007. Genomic selection. Acta Agriculturae Scandinavica, Section A-Animal Sciences 57: 192-195.

Goddard, M., and B. Hayes. 2007. Genomic selection. Journal of Animal Breeding and Genetics 124: 323-330.

Gonzalez-Recio, O., D. Gianola, G. Rosa, K. Weigel, and A. Kranis. 2009. Genome-assisted prediction of a quantitative trait measured in parents and progeny: Application to food conversion rate in chickens. Genetics Selection Evolution 41: 3.

Gunderson, K., F. Steemers, G. Lee, L. Mendoza, and M. Chee. 2005. A genome-wide scalable snp genotyping assay using microarray technology. Nature genetics 37: 549-554.

Habier, D., R. Fernando, and J. Dekkers. 2009a. Genomic selection using low-density marker panels. Genetics.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2009b. Genomic selection using low-density marker panels. Genetics 182: 343-353.

Hayes, B. 2007. Lecture notes, qtl mapping, mas, and genomic selection, june 4-6, 2007.

Hayes, B., and M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet Sel Evol 33: 209-229.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2008. Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 92: 433-443.

Hillier, L. et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432: 695-716.

Huang, J. et al. 2001. High-throughput genomic and proteomic analysis using microarray technology. Clinical Chemistry 47: 1912-1916.

Humphray, S. et al. 2007. A high utility integrated map of the pig genome. Genome Biology 8: R139.

Jack Dekkers, D. G., Rohan Fernando. 2009. Use of high-density snp genotyping for genetic improvement of livestock No. 2009, Ames.

Lai, E. 2001. Application of snp technologies in medicine: Lessons learned and future challenges No. 11. p 927-929. Cold Spring Harbor Laboratory Press.

Liu, B. 1998. Statistical genomics: Linkage, mapping, and qtl analysis. CRC Press.

Long, N., D. Gianola, G. Rosa, K. Weigel, and S. Avendano. 2007. Machine learning classification procedure for selecting snps in genomic selection: Application to early mortality in broilers. Journal of Animal Breeding and Genetics 124: 377-389.

Lorincz, A. 2006. Nucleic acid testing for human disease. 1st ed. CRC Press.

Meuwissen, T. 2003. Genomic selection: The future of marker assisted selection and animal breeding In: Marker assisted selection: a fast track to genetic gain in plant and animal breeding? Session II: MAS in animals. GAP Eletronic Forum on Biotechnology in Food and Agriculture: Conference 10 No. 2009.

Meuwissen, T., B. Hayes, and M. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819-1829.

Nadeau, J. H. et al. 2001. Sequence interpretation: Functional annotation of mouse genome sequences. Science 291: 1251-1255.

Nebert, D. 1999. Pharmacogenetics and pharmacogenomics: Why is this relevant to the clinical geneticist? Clinical Genetics 56: 247.

Park, T., and G. Casella. 2008. The bayesian lasso. Journal of the American Statistical Association 103: 681-686.

Pearson, T., and T. Manolio. 2008. How to interpret a genome-wide association study. Jama 299: 1335.

Peltz, G. 2005. Computational genetics and genomics: Tools for understanding disease. 1st ed. Humana Press Inc.

Pennington, S., and M. Dunn. 2001. Proteomics: From protein sequence to function. Taylor & Francis.

Schaeffer, L. 2006. Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics 123: 218-223.

Sladek, R. et al. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 445: 881-885.

Thuillet, A. C. et al. 2002. Direct estimation of mutation rate for 10 microsatellite loci in durum wheat, triticum turgidum (l.) thell. Ssp durum desf. Mol Biol Evol 19: 122-125.

Tishkoff, S. A., and B. C. Verrelli. 2003. Patterns of human genetic diversity: Implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet 4: 293-340.

van Ommen, G. 2008. Popper revisited: Gwas here, last year. European Journal of Human Genetics 16: 1-2.

VanRaden, P., G. Wiggans, T. Sonstegard, and C. Van Tassell. 2008. Using genomic data to improve dairy cattle genetic evaluations No. 2009.

Venter, J. C. et al. 2001. The sequence of the human genome. Science 291: 1304-1351.

Vignal, A., D. Milan, M. SanCristobal, and A. Eggen. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genetics Selection Evolution 34: 275-305.

Womack, J. 2005. Advances in livestock genomics: Opening the barn door No. 15. p 1699-1705. Cold Spring Harbor Laboratory Press.

Ziegler, A., I. Konig, and J. Thompson. 2008. Biostatistical aspects of genome-wide association studies. Biometrical Journal 50(1):8-28.

Table 1. Technical requirements and characteristics. (Vignal, 2002 }

	Technical requirements				Technical characteristics			
Marker name	Restriction Enzyme	PCR	Specific Primers	Gel	Development Effort	Genotyping Effort	Reproducibility[1]	Accuracy[2]
RFLP	yes	no	no[3]	yes	High	High	High	Very High
RAPD	no	yes	no	yes	Very Low	Very Low	Low	Very Low
AFLP	yes	yes	no	yes	Low	Very Low	High	Medium
Microsatellite	no	yes	yes	yes	High	Low	High	High
SNP	no	yes	yes	yes/no[4]	High	Variable[4]	High	Very High

[1] Refers to the genotyping error rate of the method: results may vary from one experiment to another.

[2] Refers to the precision at which true allele recognition can be performed.

[3] However, the RFLP technique relies on the use of a specific probe for the Soutnern-blot techinique. Nowadays, RFLPs are ususally genotyped by PCR-RFLPs, requiring specific primers.

[4] According to the genotyping technique used.

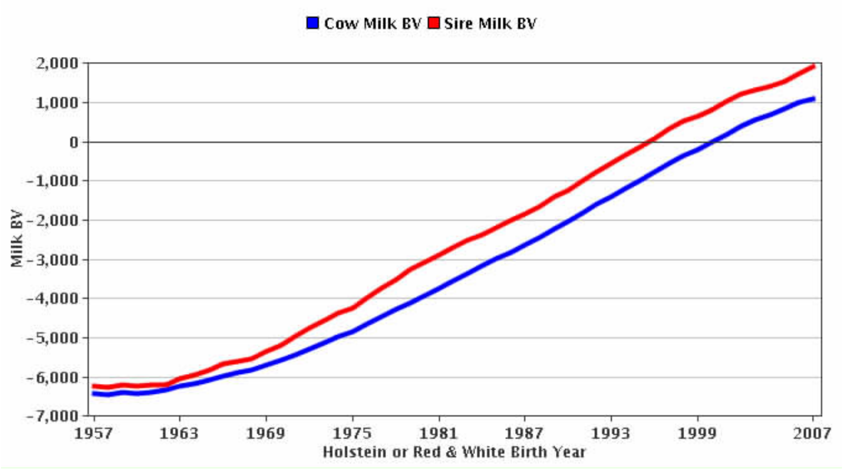
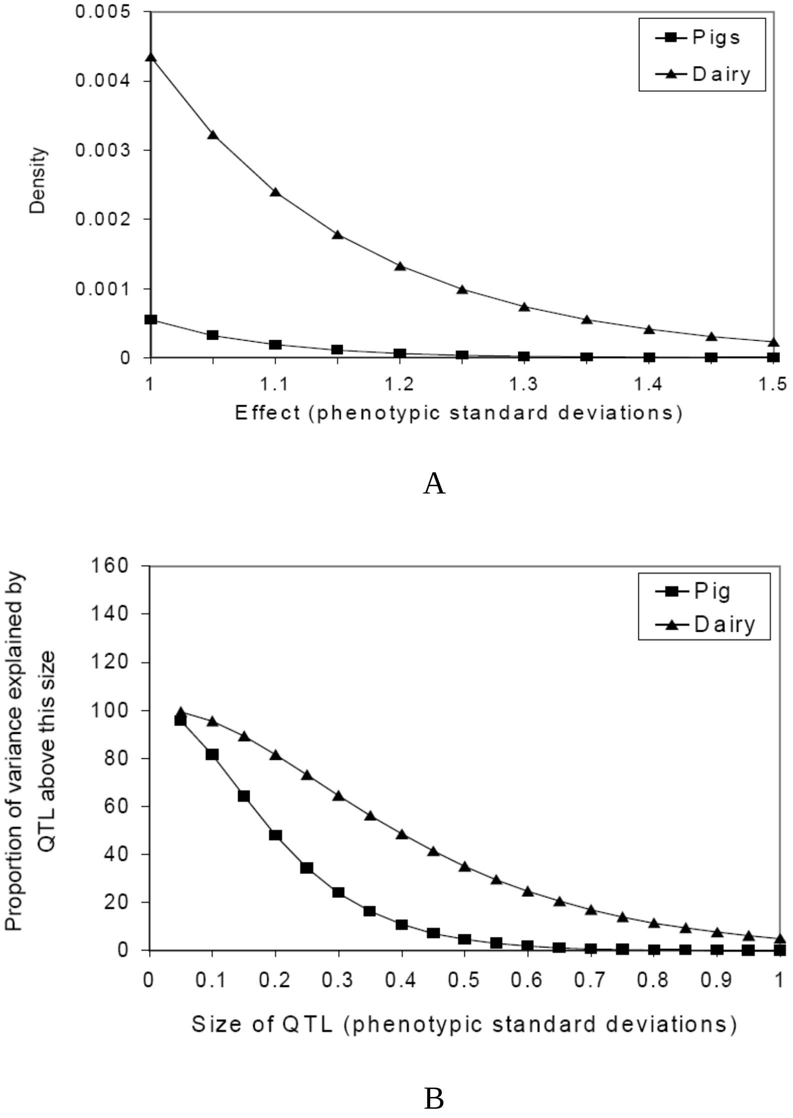


Figure 1. Milk for Holstein or Red & White (USDA, April 2009)



A

B

Figure 2. QTL effect distribution for pig and dairy cattle and estimated proportion of variance. A. Gamma distribution of QTL effect distribution for pig and dairy cattle (fitted with maximum likelihood). B. Estimated proportion of variance contributed by the QTL above a size of the true QTL effects (Hayes, 2001).

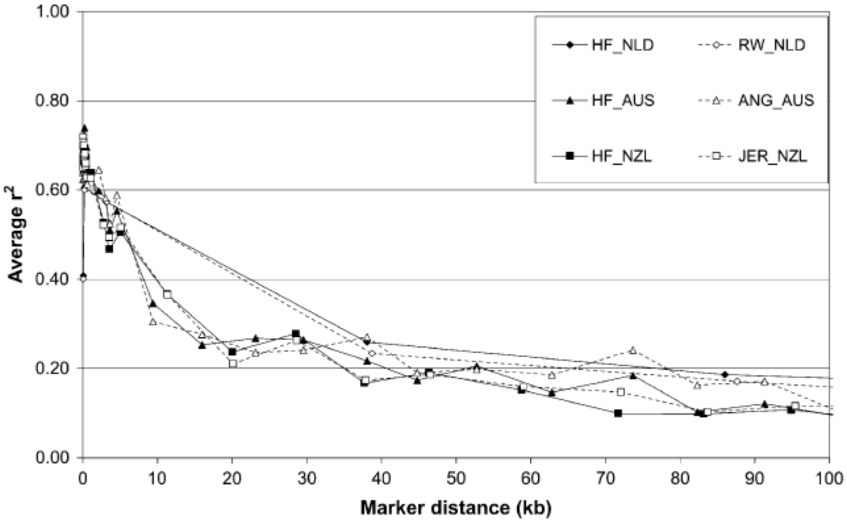


Figure 3. Average linkage disequilibrium (r^2) as a function of average genomic distance for Dutch black-and-white Holstein–Friesian bulls (HF_NLD), Dutch red-and-white Holstein–Friesian bulls (RW_NLD), Australian Holstein–Friesian bulls (HF_AUS), Australian Angus animals (ANG_AUS), New Zealand Friesian cows (HF_NZL), and New Zealand Jersey cows (JER_NZL) for distances between 0 and 100 kb. Each data point was based on 200 marker pairs, resulting in standard errors 0.03. (A. P. W. de Roos et al. 2008)

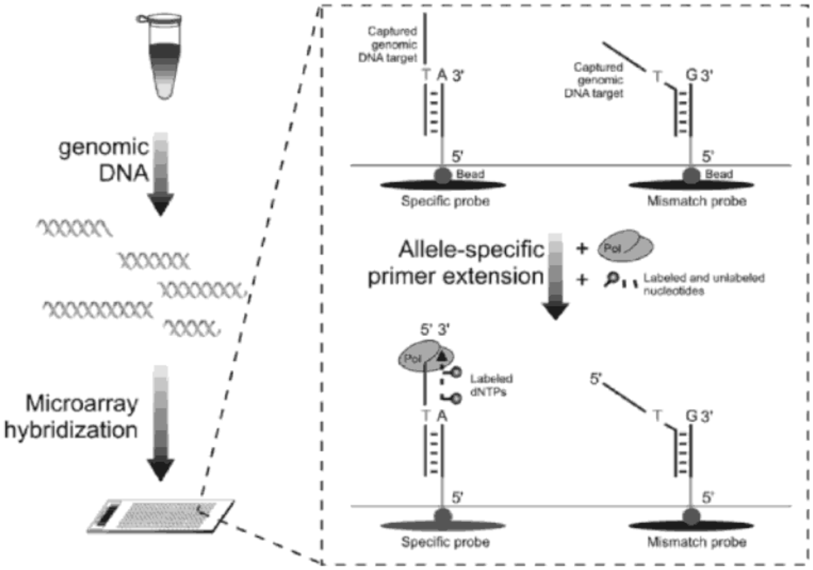


Figure 4. Genotyping on DNA microarrays procedure (Buzdin, 2007): Amplification of genomic DNA (gDNA) generates fragments that are hybridized to specific and sensitive oligonucleotide probes on microarray. An allele-specific primer extension (ASPE) reaction scores the captured SNP targets by incorporating multiple biotin-labeled dNTP (deoxynucleotide triphosphate*)* nucleotides into the appropriate allelic probe. For a given SNP on a give strand, two or more different allele-specific oligonucleotide probes are designed to capture different SNPs, since polymerase extension occurs preferentially from matched 3’-termini, enabling appropriate scoring of the SNP.

CHAPTER 3

IMPUTING MISSING GENOTYPES USING AN ANT COLONY APPROACH[1]

[1] Huiyu Wang, R. Rekaya. To be submitted to the Journal of Animal Science

Abstract

Availability of reliable genotyping platforms for single nucleotide polymorphism markers (SNP) has made the possibility for genomic breeding values estimation in several livestock species a reality. Unfortunately, at above two hundred dollars per animal this technology is still too expensive for massive use at the commercial level. Currently, this technology is mainly used for genotyping top animals and then used in two step procedure for estimating genomic breeding values. For its use at the population level, the SNP genotypes of non-typed animals have to be inferred somehow from the already genotyped animals and their relationships. In fact, several attempts have been proposed ranging from the calculation of gene content to the construction of a covariance matrix similar to the classical additive relationship matrix. However, in all cases the only information used is the one available from the allele frequencies and the relationships between animals. This information could be limited especially in pedigrees that span several generations as it is the case in food producing animals. In this study, we propose using low-density chips with few hundred SNPs for large scale genotyping. This low cost chip will provide an additional source of information, linkage disequilibrium, in inferring genotypes of non-typed animals. For that purpose a simulation was conducted where 1,000 and 2,000 animals were genotyped for the 50 K SNPs and 500, 1,000 and 2,000 SNPs were selected jointly using an ant colony algorithm. The results showed an increase in the probability of predicting the true SNP genotypes. In fact, the percentage of alleles known after the sampling process (AK) and the average probability of the true genotype being identified for every animal and locus (APTG) were 0.79 and 0.61, respectively.

INDEX WORDS: SNP, Ant Colony, Low density, High density, Genotype imputation

Introduction

Many algorithms have been developed to predict breeding values in livestock applications, to classify disease types and to identify causative mutations based on the SNP genotypes and significant gains have been made in the accuracy of prediction and of disease classification. In addition many studies have shown that improved performance can be achieved when using a selected subset of features, as opposed to using all available data. Increases in accuracy achieved through the selection of predictive features can complement and enhance the performance of classification algorithms, improve the understanding of biologically relevant features, and reduce the cost. Ideally, one would like to select an optimal sub-set of features that would yield maximum predictive power. In the case of high-dimensional data sets, such as SNP genotypes, this can be very computationally demanding. Furthermore, several studies (Coutinho et al., 2007; Barendse et al., 2007) have found that gene interactions may play important roles in many complex traits. Unfortunately, due to the high density of SNP maker maps, it is computationally infeasible to examine all possible interactions. As a result studies examining gene interactions tend to focus on a small number of SNP, previously identified as having strong marginal associations.

While this approach has shown some success, simulation studies conducted by Marchini et al. (2005) and Pickrell et al. (2007) showed that, in the presence of several types of gene interactions, there is reduced power to detect causative loci with models estimating only marginal effects. Using an exhaustive search of all two-way interactions, Marchini et al. (2005) achieved greater power to detect causative mutations when compared to models estimating only marginal effects. However, due to the high computational cost of this approach, a two-stage model was proposed, in which SNP were selected in the first stage based on marginal effects and then tested for interactions in the subsequent stage [Marchini, et al. 2005]. Such an approach represents a compromise that could result in the failure to detect important regions of the genome in the first stage of the model. As such, there is a need for methodologies capable of identifying important genomic regions in the presence of potential gene interactions when large numbers of markers are genotyped.

In order to overcome the problems associated with the classical approach based on the marginal effects for reducing the dimensionality of SNP data and its inability to account properly for the potential interactions, we propose a two stage procedure based on (1) SNP tagging and (2) marker identification using Ant colony algorithm. The specific objective of this study was to develop a combined ant colony algorithm and peeling process to select a low density SNP marker panel that will maximize the probability for imputing the non-typed loci.

Material and Methods

SNP Tagging and Genotyping

SNP tagging provides a mean to both reduce collinearity and dimensions of genotype data. Use of these methods can reduce genotyping cost without substantially decreasing the coverage of the genome, in terms of LD between adjacent SNP. For applications in association studies, a multiple linear regression (MLR) approach developed by He and Zelikovsky (2006) was modified and integrated into our optimization and feature selection software suite. The modified procedure was implemented in a wrapper scheme using the ACA for selection of tagged SNP. The selected SNP will be evaluated using MLR:

$$\mathbf{y}_i = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \qquad \mathbf{e} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \otimes \mathbf{I}) \qquad (1)$$

where **X** is an n by k incidence matrix with n rows corresponding to n genotypes or haplotype blocks and k rows corresponding to k tagged SNP; $\mathbf{y}_i$ is a vector containing genotypes for the non-tagged SNP i; and e is a vector of random residuals. The accuracy in which the tagged SNP predict the untagged SNP will be used to update pheromone in the ACA to find subsets of SNP that reduce dimension to an acceptable degree while minimizing the loss of LD.

Ant colony algorithm for SNP data

The ACA employs artificial ants that communicate through a probability density function (PDF) that is updated each iteration with weights or “pheromone levels”, which are analogous to the chemical pheromones used by real ants. In the case of SNP association studies, the weights can be determined by the strength of the association between selected genotypes and the response of interest (in case the ability to impute missing genotypes). Using the notation of Dorigio and Gambardella (1997) and Ressom et al. (2006), the probability of sampling SNP m at time t is defined as:

$$P_m(t) = \frac{(\tau_m(t))^{\alpha} \eta_m^{\beta}}{\sum_{m=1}^{nf} (\tau_m(t))^{\alpha} \eta_m^{\beta}} \qquad (2)$$

where $\tau_m(t)$ is the amount of pheromone for SNP m at time t; η_m is some form of prior information on the expected performance of SNP m; α and β are parameters determining the weight given to pheromone deposited by ants and a priori information on the features, respectively

Using the PDF as defined in equation (2), each of j artificial ants selects a subset S_k of n SNP from the sample space S containing all SNPs. Given the relationship between adjacent SNP, ants randomly change SNP selections following a multinomial distribution, with changes being limited to the three adjacent SNP on either side of the originally selected SNP marker. The pheromone level of each feature m in S_k is then updated according to the performance of S_k as:

$$\tau_m(t+1) = (1-\rho) * \tau_m(t) + \Delta\tau_m(t) \qquad (3)$$

where ρ is a constant between 0 and 1 representing the rate at which the pheromone trail evaporates; $\Delta\tau_m(t)$ is the change in pheromone level for feature m based on the sum of accuracy of all S_k containing SNP m, and is set to zero if SNP m was not selected by any of the artificial ants.

The procedure can be summarized in the following steps:

1) Each ant selects a predetermined number of SNP markers.
2) Using the selected SNP markers, accuracies are computed using regression on genotypes.
3) The pheromone for each selected group of SNP, S_k, is calculated as:

   $\text{pheromone}_k\text{=acc}^{(1\text{-acc})}$ (acc=accuracy)
4) The change in pheromone at time *t* is then calculated using equations (3) .
5) Following the update of pheromone levels according to equations (3), the PDF was updated and the process is repeated until pheromone levels have converged.

Once the optimal small set of SNPs of the low density panel have been identified by the ACA procedure, it will be combined with the allele frequencies and the pedigree relationship in a peeling algorithm similar to Qian and Beckmann (2002), Tapadar *et al.* (2000) and Spangler et al. (2009). Once the peeling process is completed, the number of animals with one or two alleles known and the probability of inferring the true genotype at each marker locus, PTG_{igj}, will be computed as:

$$\mathrm{PTG}_{ig_j} = \frac{\text{number of times genotype } g_{ij} \text{ was assigned at locus j}}{\text{total number of samples}}$$

where genotype g_{ij} is the true genotype for animal i at locus j.

Peeling process

Given that genotypes in this study were assigned at random in the population, it is possible to extract additional genotypic information from the pedigree. Animals with missing genotypic information can be assigned one or both alleles given parental, progeny, or mate information. Given this trio of information sources and following an algorithm similar to Qian and Beckmann (2002) and Tapadar et al. (2000), imputation on missing genotypes were made and additional genotypic information was garnered. As indicated before, it was assumed that there were no errors in the recorded pedigree resulting in all animals having known paternity and maternity.

For genotypes that were correctly inferred by the ACA, homozygous provided the major additional information for the peeling process as the origin of both alleles is known. However, for the heterozygous genotypes the origin of both alleles was unknown after ACA step and their usefulness in the peeling process depended on the ability of resolving the phase situation. For that purpose, information on relatives, mates and Mendelian inconsistency was used to determine the origin of each allele. Similarly, genotypes imputed with a probability greater than .8 during the ACA step were used in the peeling process. Those genotypes were assumed correct unless they conflict with Mendelian rules of segregation. In that case, that genotype was ignored and the next best genotype (the second highest probability) will be assumed as true and the process described earlier was repeated. In the cases where the parental origin of an allele was unclear, then allele was arbitrarily assigned as either the paternal or maternal allele.

At the end of the peeling process those animals that had either one or two alleles known were retained to be used in an iterative procedure in order to infer the remaining unknown alleles in the population. A Gibbs sampling procedure using those animals having one or two known alleles, used as prior information, was carried out in order to impute the remaining unknown alleles in the population.

For the base population animals, the unknown allele(s) were randomly sampled given the allele frequencies in the population and the assumption of Hardy-Weinberg equilibrium. Unknown alleles for non-base population animals were randomly sampled from the parent's genotypes according to Mendelian rules. An equal probability was assumed for inheriting either the first or second allele from a parent. For a non-base population animal that had only one unknown allele, the unknown allele was sampled approximately half of the time from the sire's genotype and the remaining time from the dam's genotype.

At the end of the sampling process and following Spangler et al. (2009), a benefit function that described the total number of alleles known in the population was computed. This function combines known alleles and the probability of unknown alleles assigned during the sampling process. In order to be included in the benefit function, it is not enough that the genotype at a locus include the allele in question. In fact, an allele in a particular position had to be equal to the true allele of the same position The probability of allele $a_{i,j}$, (*j* = 1 or 2) being assigned as the true allele j for animal *i* was calculated as:

$$p(a_{i,j}) = \frac{\text{number of times } a_{i,j} \text{ was assigned}}{\text{number of iterations}}.$$

Using $p(a_{i,j})$ and the number of known alleles, the benefit function at a locus *k* was computed as:

$$Benefit_k = n_1 \times 2 + \sum_{i=1}^{n_2}[1 + p(a_{i,j})] + \sum_{i=1}^{n_3}[p(a_{i,1}) + p(a_{i,2})],$$

where n_1, n_2, and n_3 were the number of animals with 2, 1 or 0 alleles known, respectively. The percentage of alleles known after the Gibbs sampling process, AK , was computed as:

$$\text{AK} = [\sum_{i=1}^{q} \frac{benefit_i}{2 * na} \times 100] / q,$$

where $benefit_i$ was the benefit function for locus *i*, n_a was the total number of animals in the pedigree, and q was total number of loci in the panel.

At every round of the sampling process only one true genotype was assigned at each locus and each animal. Thus, at the end of the sampling process every animal had a probability of having the true genotype at locus *j*, PTG_{igj}, assigned as

$$\text{PTG}_{ig_j} = \frac{\text{number of times genotype } g_j \text{ was assigned}}{\text{total number of samples}},$$

where genotype g_j was the true genotype for animal i for locus *j*. The average probability of the true genotype being identified for every animal and locus (APTG) was computed using the following equation:

$$\text{APTG} = \frac{\sum_{i=1}^{n_a}\sum_{\text{j}=1}^{\text{q}} \text{PTG}_{ig_j}}{n_a x q},$$

where n_a and q are the total number of animals and loci, respectively.

Data Simulation

A simulation was carried out where all animals in the pedigree were assigned SNP genotypes. No errors or missing genotypes were assumed. A pedigree with four over-lapping generations was simulated. The base population included 500 unrelated animals and subsequent generations consisted of 5000 animals with a total of 15,500 animals generated. Genotypes of the base population animals were assigned based on allele frequencies and linkage disequilibrium. For the subsequent generations, genotypes were randomly assigned using the parent's genotype, where an equal chance of passing either the first or second allele was assumed.

Two experimental factors were investigated: 1) number of animals (1,000 or 2,000) fully genotyped using the 50 K SNP chip; and 2) number of SNPs in the low density panel being either 500, 1,000 or 2,000 SNPs. In total there have been 6 (2x3) simulation scenarios. Five replicates of the simulated data were generated for each simulation scenario. A full description of the simulation parameters could be found in Table 1.

Results and Discussion

The percentages of SNP genotypes correctly imputed and those identified with a probability greater than 0.8 at the end of the ant colony step are presented in Table 2. They range from 7 to 21% for the completely known genotypes and from 24 to 42% for genotypes identified with probability greater than 0.8. As expected, both percentages increased with the increase of the number of SNPs in the low density panels and the number of fully genotyped animals. Furthermore, it seems that the increase of the number of SNPs in the low density panels has more influence in imputing the true genotypes than the number of fully genotyped animals. In fact, the percentage of correctly imputed genotypes almost doubled (7% vs. 12%) when the number of SNPs in the low density panel was increased from 500 to 1,000 SNPs. This could be due to the fact that the larger the number of SNPs in the low density panel, the higher is their linkage disequilibrium with the missing genotypes.

Table 3 presents the percentages to genotypes imputed with a probability smaller than 0.2 at the end of the ant colony step. They ranged between 5 and 14% with a clear tendency of decreasing with the increase of the number of SNPs in the low density panel. However, very little change was observed when the number of fully genotyped animals was increased from 1,000 to 2,000. This could be due to some SNPs with high minor allele frequency that the ant colony algorithm could not converge on a good solution.

After adding the peeling and the sampling steps, the percentages of correctly imputed genotypes have increased significantly as presented in Table 4. The percentages of correctly imputed and those identified with a probability greater than 0.8 ranged between 13 to 32% and 47 to 63%, respectively, with greater increase with the increase of the number of SNPs in the low density panel. These results indicate, as before, that an increase of the number of SNPs in the low density panel is more beneficial in imputing the missing genotypes. For the genotypes imputed with a probability smaller than 0.2 after the sampling step (Table 5), there has been only modest improvement likely indicating that there are few loci that are hard to impute either due to their weak linkage disequilibrium with the SNPs in the low density panel or/and their high minor allele frequencies. Finally, the percentage of alleles known after the sampling process (AK) and the average probability of the true genotype being identified for every animal and locus (APTG) were 0.79 and 0.61, respectively.

Conclusions

The results of this study suggest that the ant colony algorithm is a powerful tool for selecting low density SNP markers panel that will maximize the probability of imputing the missing genotypes, especially when combined with a peeling step. Furthermore, it appears that the probability of imputing missing SNP genotypes is more influenced by the number of alleles in the low density panel than the number of fully genotyped animals very likely due to the effect of linkage disequilibrium. With almost a third of the missing genotypes imputed correctly and two thirds with probability greater than 0.8, the proposed procedure could provide a cost effective tool for large scale genomic evaluation in livestock. However, it is necessary to evaluate the effects of the imputed genotypes in the accuracy of the estimated breeding values in order to access the practical usefulness of the proposed procedure.

REFERENCES

Barendse, W., B. E. Harrison, R. J. Hawken, D. M. Ferguson, J. M. Thompson, M. B. Thomas, and R. J. Bunch. 2007. Epistasis between Calpain 1 and its inhibitor Calpastatin within breeds of cattle. Genetics 176:2601-2610

Coutinho A, I. S. Sousa, M. Martins, C. Correia, T. Morgadinho, C. Bento, C. Marques, T. Miguel, J. Moore J and others. 2007. Evidence for epistasis between SLC6A4 and ITGB3 in autism etiology and in the determination of platelet serotonin levels. Human Genetics 121(2):243-256

Dorigio, M., and L. M. Gambardella. 1997. Ant colonies for the travailing salesman problem. BioSystems. 43:73-81.

Dorigio, M., G. Di Caro and L. M. Gambardella, 1999. Ant algorithms for discrete optimization. Artificial Life, 5, 2, pages 137-172.

He J., A. Zelikovsky. 2006. MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression. Bioinformatics 22(20):2558-61

Marchini J., P. Donnelly, L. R. Cardon. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics 37(4):413-417.

Pickrell, J., F. Clerget-Darpoux, and C. Bourgain. 2007. Power of genome-wide association studies in the presence of interacting loci. Genet. Epidemiol. [Epub ahead of print].

Qian, D., and L. Beckmann. 2002. Minimum-recombinant haplotyping in pedigrees. Am. J.Hum. Genet. 70: 1434-1445.

Ressom, H. W., R. S. Varghese, E. Orvisky, S. K. Drake, G. L. Hortin, M. Abdel-Hamid, C. A. Loffredo, and R. Goldman, 2006. Ant colony optimization for biomarker identification from MALDI-TOF mass spectra. 28th Annual International Conference IEEE Engineering in Medicine and Biology Society (EMBS) SaB03.6.

Spangler, M. L., K. R. Robbins, J. K. Bertrand, M. MacNeil, and R. Rekaya. 2009. Ant colony optimization as a method for strategic genotype sampling. Animal Genetics 40: 308 – 314.

Tapadar, P., S. Ghosh, and P. P. Majumder, 2000 Haplotyping in pedigrees via a genetic Algorithm. Hum Hered 50: 43-56.

Table 1: Mean, minimum and maximum of the minor allele frequency (MAF) and linkage disequilibrium (LD) for the 50 K SNP panel used in the simulation

	Mean	Min	Max
MAF	.37	0.11	0.49
LD	0.54	0.32	0.93

Table 2: Percentages of genotypes correctly imputed (FK), and those identified with a probability greater then .8 (Pr.>.8) after the ant colony step.

Fully genotyped	1000 animals		2000 animals	
Selected SNPs	FK	Pr>.8	FK	Pr.>.8
500	7	24	8	27
1000	12	32	14	36
2000	18	39	21	42

Table 3: Percentages of genotypes imputed with a probability smaller then .2 (Pr.<.2) after the ant colony step.

Fully genotyped	1000 animals	2000 animals
Selected SNPs	Pr.<.2	Pr.<.2
500	14	12
1000	8	7
2000	5	5

Table 4: Percentages of genotypes correctly imputed (FK), and those identified with a probability greater then .8 (Pr.>.8) at the end of the sampling process.

Fully genotyped	1000 animals		2000 animals	
Selected SNPs	FK[1]	Pr>.8	FK[1]	Pr.>.8
500	13	47	15	49
1000	21	52	24	57
2000	29	61	32	63

[1] defined based on a probability of having the true genotype at locus j, PTG_{igj} greater than 0.95

Table 5: Percentages of genotypes imputed with a probability smaller then .2 (Pr.<.2) at the end of the sampling process.

Fully genotyped	1000 animals	2000 animals
Selected SNPs	Pr.<.2	Pr.<.2
500	13	12
1000	6	6
2000	5	5

CHAPTER 4

COMBINING HIGH AND LOW DENSITY SNP PANELS TO IMPUTE GENOTYPES FOR NON-TYPED ANIMALS WITH APPLICATION IN GENOMIC BREEDING VALUE ESTIMATION[1]

[1] Huiyu Wang, R. Rekaya. To be submitted to the Journal of Animal Science

Abstract

The cost of genotyping using high density markers panel has precluded its use at a large scale in livestock applications. In order to reduce cost and extend the use of this new technology, a procedure that optimizes the use of existing information and with the ability to impute missing genotypes for non-typed animals is of interest to the industry. For that purpose we proposed combining genotyping information from high and low density SNP panels for genomic evaluation. Under several simulation conditions, the proposed procedure was evaluated on its ability of estimating the true breeding values. The results showed that the proposed procedure was successful in increasing accuracy of estimated GBVs by 3 to 12% depending on the number of SNPs and genotyped animals. These results suggest that this procedure could provide a cost effective tool for large genomic evaluation.

INDEX WORDS: SNP, Low density, High density, Genotype imputation, Genomic selection

Introduction

The advent of new technologies makes it possible to efficiently genotype animals for thousands of SNP in order to identify genomic regions associated with phenotypes of interest. Through the identification of genomic 'hot spots', marker coverage could be allocated accordingly, with markers placed more densely on 'hot regions' and more sparsely in regions with no apparent associations. Genetic variation present in sparse marker areas could then be accounted for by including a polygenic effect (Calus and Veercamp, 2007). One issue associated with this approach is the lack of power due to high-dimensions and small additive effects. When dealing with complex traits, the underlying genetic mechanisms are often complex involving several interacting genes (Barense et al., 2007; Coutinho et al., 2007). Under such scenarios, several simulation studies have show there is little power to detect important genomic regions when looking at marginal effects only (Marchini et al., 2005; Pickrell et al., 2007). Piyasatian et al. (2007) found that when including only markers with significant effects, there was reduced accuracy for prediction of expected breeding values when compared to genomic selection utilizing full genome coverage. Given the high dimensions of SNP data sets traditional statistical methodologies cannot estimate the parameters needed to explore all potential interactions. In such scenarios, machine learning methodologies have been shown to be effective. These methodologies are capable of efficiently searching large sample spaces for optimal solutions. Robbins et al. (2008a) showed that, when using genomic features selected by the ant colony algorithm, substantial increases in prediction accuracy were obtained when compared to methodologies selecting features base on marginal effects. In applications to SNP studies we have found that the ACA has far greater power to detect important genomic regions for traits controlled by interacting genes.

When we applied to the Ant colony Algorithm (ACA) to several high-dimensional data sets, it was able to identify small subsets of highly predictive and biologically relevant features without the need for extensive pre-selection. Using the selected features to train a predictive model yielded substantial increases in prediction accuracy when compared to several rank based methods and results obtained in previous studies. When applied to the often noisy and high dimensional expression data, our results (Robbins et al., 2008a) showed that the performance of the ACA was superior, not only to the filter based methods but to several reported results using the GCM data set that has been a benchmark to compare the performance of classification and feature selection and it consists of 198 samples collected from 14 tumor types. The ACA consistently yielded higher accuracies than the filter based methods, for which ranks varied across replications. Furthermore, the breaks in pheromone levels observed with the most predictive genes also provided more objective selection criteria for identifying top features, unlike the filter methods used in this study, in which truncation points were somewhat arbitrary.

More recently, our group has developed and tested a modified ACA for use in association analysis involving complex traits (Robbins et al., 2008b). The ACA has been applied to genotype data from the HapMap ENCODE project and a simulated binary trait under the control of two interacting loci.

The ACA was implemented with logistic regressions on haplotypes and genotypes as the evaluation function for pheromone deposition and compared to single locus genotype and sliding window haplotype methods of analysis. As with previous applications, the ACA's ability to account for complex data structures, in this case epistasis, allows it to achieve substantial increases in power over models accounting for only marginal effects.

Furthermore, plots of the pheromone can then be used to select relevant markers for use in evaluations. Since markers are analyzed in groups, as opposed to evaluating a single marker region at a time, the ACA is able to efficiently account for gene interactions without any prior knowledge of gene networks underlying traits of interest.

Our group has an extensive expertise in developing alternative methods of selecting animals to be genotyped using Ant Colony algorithms (ACA). Using simulated and real beef cattle data, Spangler et al. (2009) suggest that ACA is the most desirable method of selecting candidates for genotyping, particularly after a peeling step. From these results it appears that the number of offspring and the number of mates along with the homozygosity of the genotyped animals is critical in the selection process. Consequently, in application it will be critical to have good estimates of allele frequencies prior to implementing our genotype sampling strategy.

Differences in performance of ACA do exist between the explored pedigrees. This is due to the proportion of sires and dams that have large numbers of offspring and/or mates. In the dairy industry, for example, there may be only a small number of sires in a pedigree but they may all be used heavily as in the case of the simulated pedigrees in the current study. In contrast, a pedigree from the beef industry may have a larger proportion of sires but a large number of them may be used less frequently. Furthermore, pedigrees from field data or from research projects will also have innate structural differences. Research projects may be limited by the size of the population and thus only use a small number of sires. In this scenario it would also be possible for higher rates of inbreeding and larger numbers of loops in a pedigree due to a large number of full sibs. Ant colony optimization offers a new and unique solution to the optimization problem of selecting individuals for genotyping. However, the heuristics, such as the number of ants, number of iterations, and the evaporation rate have to be chosen carefully.

Recent interest in marker-assisted selection (MAS) and genome wide selection for livestock populations has increased greatly due to improvement in genotyping techniques and development of theoretical framework for dealing with such information. Yet, it may not be viable to genotype each animal due to cost, time or lack of availability of DNA. Furthermore, for some traits such as N and P retention, the situation is even more complex due to the fact that the collection of those phenotypes requires individual feeding and fecal collection protocols which will present logistical problems. Consequently, we believe that it is of crucial practical importance to carefully address two main issues: 1) identification of a set of animals to genotype that maximizes the information on non-genotyped ones, and 2) to develop a low density and low cost marker panel that will maximize the information on imputing non-genotyped loci or/and animals. By combining both procedures, we will be in a position to maximize the use of information, reduce cost and genotype a much large portion of the population. As such, a method that could select a small subset of size “n” of features (animals or loci) for genotyping, which in turn, could be used to infer, with high probability, the genotypes for the remaining animals in the population or loci high density panel could be beneficial. By using such a method, fewer animals in a population and SNPs would be needed for genotyping which would decrease the time and cost of genotyping. Theoretically the problem at hand is simple to solve. If it were possible to evaluate every possible subset of animals in the population or loci in the low density panel equal to the desired size (e.g. 5% of the total number of features) the optimal solution could be found. Unfortunately, such an approach is computationally impossible at the current time. Consequently a more feasible solution is needed. Spangler et al. (2007 and 2009) presented an alternative approach for selective genotyping, viewing the problem as one of optimization using an ant colony approach. The “path” chosen by an artificial ant is a subset of features selected from a larger sample space, and the “distance” traveled is some measure of the features performance. In the case of genotyping, the ACA should select a subset of animals that, when genotyped, should give an optimal performance in terms of extrapolating the alleles of non-genotyped animals.

In this study we will extend the ant colony algorithm presented by Spangler et al. (2007 and 2009) with the specific objectives of: 1) selecting a low density panel from a high density chip and evaluate its effectiveness in imputing missing genotypes and 2) evaluate the accuracy of GBV estimated by combining low and high density panels. The proposed procedures will be evaluated using simulated data.

Material and Methods

Data Simulation

A simulation was carried where all animals in the pedigree were assigned 50 K SNP genotypes. A pedigree with four over-lapping generations was simulated. The base population included 500 unrelated animals and subsequent generations consisted of 5000 animals with a total of 15,500 animals generated. Genotypes of the base population animals were assigned based on allele frequencies. For the subsequent generations, genotypes were randomly assigned using the parent's genotype, where an equal chance of passing either the first or second allele was assumed.

Three experimental factors were investigated: 1) number of animals (1,000 or 2,000) fully genotyped using the 50 K SNP chip; 2) number of SNPs in the low density panel being either 500, 1,000 or 2,000 SNPs; and 3) the percentage of animals genotyped using the low density chip being either 10 or 25%. In total there have been 12 (2x3x2) simulation scenarios. Five replicates of the simulated data were generated for each simulation scenario.

Estimation of the SNP effects

The following mixed linear model was used to estimate the SNP effects

$$\mathbf{y} = \mu\mathbf{1} + \sum_{i=1}^{q}\mathbf{X_i a_i} + \mathbf{e} \qquad [1]$$

where $\mathbf{y}$ is the vector of phenotypes of size n, μ is the overall mean, $\mathbf{a}_i$ is the SNP effects at position i, $\mathbf{X}_i$ is a known incidence matrix, and $\mathbf{e}$ is the vector of residual terms. The summation term, Σ_i, is over all possible SNPs.

For a Bayesian implementation of the model in equation [1], prior distributions for all parameters are required. In this study, the following prior distributions were assumed

$$p(\mu) \propto cons\tan t$$

$$p(\mathbf{a}_i \mid \sigma^2_{ai}) \sim N(\mathbf{0}, \sigma^2_{ai})$$

$$p(\sigma^2_{ai} \mid \nu_a, s^2_a) \sim \chi^{-2}(\nu_a, s^2_a)$$

$$p(\sigma^2_e \mid \nu_e, s^2_e) \sim \chi^{-2}(\nu_e, s^2_e)$$

Where ν_{ai}, and ν_e are the prior degrees of belief for the SNP and residual variance respectively, and s^2_a and s_e are the correspondent scaling factors. This formulation is similar to the BayesA method presented by Meuwissen et al. (2001).

The resulting joint posterior distribution of all unknown parameters in the model is the product of densities in equation [1] and the prior distributions. Because conjugate priors were used, the joint posterior distribution was in closed form. For implementation via Gibbs sampling the full conditional distributions were needed. In the case, all conditional distributions were in closed form being normal for the position parameters (overall mean and the SNP effects) and scaled inverted Chi square for the variance components

$$p(\sigma^2_{ai} \mid \mathbf{a}_i, \mu, \sigma^2_e, y) \sim \chi^{-2}(\nu_a + n_i, \mathbf{a}^{'}_{\mathbf{i}}\mathbf{a}_i + s^2_a)$$

$$p(\sigma^2_e \mid \mathbf{a}_i, \mu, \sigma^2_{ai}, y) \sim \chi^{-2}(\nu_e + n, \mathbf{e}'\mathbf{e} + s^2_e)$$

Analyses

Training: Fully genotyped animals were used in the training step. Depending on the simulation scenario, either 1,000 or 2,000 animals were used. In order to evaluate the adequacy of combining low and high density panels and the imputation of missing genotypes, the training was conducted using both the high and low density panels. In other words, if 1,000 animals were considered to be fully genotyped, two training analyses were conducted: 1) using the high density panel, and 2) using the low density panel (500 -2,000 SNPs).

In all training analyses, Gibbs sampler with a single chain of 10,000 to 20,000 iterations was implemented. Convergence was assessed by graphical inspection of the trace plot of the sampling process. The burn-in period ranged from 2,000 to 3,000 iterations.

Validation: Animals genotyped using the low density panel (10 or 25% of the population) were used in the validation step. When the high density panel was used in the training step, the missing genotypes for the animals genotyped with the low density panel were imputed.

Comparison of estimated breeding values

Genomic breeding values were computed as the sum of all the SNP effects. Accuracies of estimated breeding values for animals in the validation set obtained using the low density panel and the high density panel with imputed genotyped were computed and compared in order to quantify the potential advantage of combining low and high density panels.

Results

The proposed procedure was successful in increasing the probability of inferring the true SNP genotypes for the non-typed animals by 12 to 18% depending of the simulation parameters. This increase was observed even for animals with loose to no relationships with fully genotyped animals due to the use of linkage disequilibrium between loci in the low and high density panels. Furthermore, the proposed method increased the accuracy of the estimated breeding values by 2 to 12% depending on the number of SNPs in the low density panel and the number of genotyped animals. When only 1,000 animals were fully genotyped using the 50 K panels, the percentage increase in accuracy for breeding value estimation for the non-fully typed animals ranged from 2 to 9% as indicated in Table 1. As expected, increase in accuracy improved with the increase of the number of SNPs in the low density panel and the number of partially genotyped animals (10 or 25%). However, it worth mentioning that the marginal increase from 1,000 to 2,000 SNPs is smaller than from 500 to 1,000 SNPs. This could be due to fact that 1,000 SNPs will contain the majority of most influential genome segments and increase the number to 2,000 will add only marginally to the accuracy of breeding value estimation. When 2,000 animals were fully genotyped a similar trend was observed. However, with the increase in the accuracy of breeding values estimation is higher than in the case when only 1,000 animals were genotyped (Table 2).

Conclusions

The results of this study suggest that for some livestock industries, such as dairy cattle, the proposed procedure could offer a practical and cost effective tool for large scale use of genomic information in the genetic evaluation where animals of high use (i.e. bulls) will be genotyped with the high density SNP panel and the rest of animals (especially females) will be genotyped using the low density panel. It seems that genotyping a small fraction of the population with a high density panel (50 K SNPs) could provide reasonable information to infer the missing genotypes from animals genotyped with low density panels or non genotyped at all. It is worth mentioning that two issues have to be considered while contemplating an approach similar to the one presented on this study: 1) the SNPs in the low density panel have to selected based on either ability in inferring the missing genotyped for the non-typed loci mainly through the linkage disequilibrium and 2) the subset of animals to be fully and partially genotyped have to be selected carefully in order to maximize the information content about non-typed SNPs. For both issues, the Ant colony algorithm proved to be robust yet practical and cost effective.

REFERENCES

Barendse, W., B. E. Harrison, R. J. Hawken, D. M. Ferguson, J. M. Thompson, M. B. Thomas, and R. J. Bunch. 2007. Epistasis between Calpain 1 and its inhibitor Calpastatin within breeds of cattle. Genetics 176:2601-2610

Calus M.P.L., R. F. Veerkamp. 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal of Animal Breeding & Genetics 124(6):362-368.

Coutinho A, I. S. Sousa, M. Martins, C. Correia, T. Morgadinho, C. Bento, C. Marques, T. Miguel, J. Moore J and others. 2007. Evidence for epistasis between SLC6A4 and ITGB3 in autism etiology and in the determination of platelet serotonin levels. Human Genetics 121(2):243-256

Marchini J., P. Donnelly, L. R. Cardon. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics 37(4):413-417.

Pickrell, J., F. Clerget-Darpoux, and C. Bourgain. 2007. Power of genome-wide association studies in the presence of interacting loci. Genet. Epidemiol. [Epub ahead of print].

Piyasatian N, R. L. Fernando RL, J. C. Dekkers. 2007. Genomic selection for marker-assisted improvement in line crosses. Theor Appl Genet 115(5):665-74.

Robbins K. R., W. Zhang, J. K. Bertrand, and R. Rekaya. 2008a. The ant colony algorithm for feature selection in high-dimension gene expression data for disease classification. Math Med Biol. 24(4):413-426.

Robbins K. R., W. Zhang, J. K. Bertrand, and R. Rekaya. 2008b. The use of the ant colony algorithm for the detection of marker associations in the presence of gene interactions. Genetics (Submitted)

Spangler, M. L., J. K. Bertrand, and R. Rekaya. 2007. Combining molecular test information and correlated phenotypic records for breeding value estimation. J. Anim. Sci. 85:641-649.

Spangler, M. L., K. R. Robbins, J. K. Bertrand, M. MacNeil, and R. Rekaya. 2009. Ant colony optimization as a method for strategic genotype sampling. Animal Genetics 40: 308 – 314.

Table 1: Increase in accuracy for partially genotyped animals using 1000 fully genotyped and phenotyped animals (average over 5 replicates).

Percentage	Low density panel		
genotyped animals	500	1,000	2,000
10	2.42	4.28	5.71
25	5.18	7.44	8.57

Table 2: Increase in accuracy for partially genotyped animals using 2000 fully genotyped and phenotyped animals (average over 5 replicates).

Percentage	Low density panel		
genotyped animals	500	1,000	2,000
10	2.96	5.70	8.57
25	7.14	10.0	11.42

CHAPTER 5

CONCLUSIONS

The results of this study showed that the ant colony algorithm is an efficient tool for selecting a low density SNP markers panels to impute the non-typed loci. When combined with a peeling like procedure, it increased the probability of predicting the true SNP genotypes. In fact, the percentage of alleles known after the sampling process (AK) and the average probability of the true genotype being identified for every animal and locus were 0.79 and 0.61, respectively. When combining low and high density SNP markers panels with missing genotyped being imputed, the proposed method increased the accuracy of the estimated breeding values by 3 to 12% depending on the number of SNPs in the low density panel and the number of genotyped animals. These results suggest that for some livestock industries, such as dairy cattle, the proposed procedure could offer a practical and cost effective tool for large scale use of genomic information in the genetic evaluation where animals of high use (i.e. bulls) will be genotyped with the high density SNP panel and the rest of animals (especially females) will be genotyped using the low density panel.

APPENDIX A

ABBREVIATIONS

AFLP:	Amplified Fragment Length Polymorphism
GS:	Genomic Selection
GWAS:	Genome-Wide Association Studie s
HGP:	Human Genome Project
LOD:	Logarithm of Odds
RAPD:	Random Amplified Polymorphic DNA
RFLP:	Restriction Fragment Length Polymorphism
SCAR:	Sequence Characterized Amplified Region
STS:	Sequence Tagged Site
SNP:	Single Nucleotide Polymorphism
SSRs:	Simple Sequence Repeats