MULTI-GENERATIONAL IMPUTATION OF SNP MARKER GENOTYPES AND ACCURACY OF GENOMIC SELECTION

by

SAJJAD TOGHIANI

(Under the Direction of Romdhane Rekaya)

ABSTRACT

Recent Advances in DNA sequencing technologies and the increased availability of high density single nucleotide polymorphism (SNP) genotyping platforms provided unprecedented opportunities to enhance breeding programs in livestock, poultry and plant species and to better understand the genetic basis of complex traits. Using this genomic information, more accurate breeding values are obtained. The superiority of genomic selection is possible only when high density SNP panels are used to track genes and QTLs affecting the trait. Unfortunately, even with the continuous decrease in genotyping costs, only a small fraction of the population has been genotyped with these high density panels. In order to reduce the cost of genomic selection, it is often the case that a larger portion of the population is genotyped with low-density and low-cost SNP panels and then imputed a higher density. Accuracy of SNP genotype imputation tends to be high when minimum requirements are met. Nevertheless, a certain rate of genotype imputation errors is unavoidable. Furthermore, such rate of errors tends to increase with the increase of the generational interval between reference and testing generations. Thus, it is reasonable to assume that the accuracy of GEBVs will be affected by the imputation errors; especially

their cumulative effects over time. To evaluate the impact of multi-generational selection on the accuracy of SNP genotypes imputation on the reliability of resulting GEBVs, a simulation was carried out under varying updating of the reference population, distance between training and validation sets, and the approach used for the estimation of GEBVs. Using fixed reference populations, imputation accuracy decayed by around .5% per generations. In fact, after 25 generations, the accuracy was only 7% lower than the first generation. When the reference population was updated by either 1% or 5% of the top animals in the previous generations, decay of imputation accuracy was substantially reduced. These results indicate that low density panels are useful, especially when the generational interval between reference and testing population is small. As the generational interval increases, the imputation accuracies decay, although not at an alarming rate. In absence of updating of the reference population, accuracy of GEBVs decays substantially in one or two generations with a decrease rate of around 20-25% per generation. When the reference population is updated by 1 or 5% every generation, the decay in accuracy was only 8 to 11% for 7 generations using the true and imputed genotypes. These results indicate that imputed genotypes provide a viable alternative, even after several generations, as long the reference and training populations are appropriately updated to reflect the genetic change in the population.

INDEX WORDS: SNP genotype imputation, Genomic selection, Accuracy

MULTI-GENERATIONAL IMPUTATION OF SNP MARKER GENOTYPES AND ACCURACY OF GENOMIC SELECTION

by

SAJJAD TOGHIANI

BS, Shiraz University, IRAN, 2005

MS, The University of Guilan, IRAN, 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, GEORGIA

2014

MULTI-GENERATIONAL IMPUTATION OF SNP MARKER GENOTYPES AND ACCURACY OF GENOMIC SELECTION

by

SAJJAD TOGHIANI

Major Professor:	Romdhane Rekaya
Committee:	Ignacy Misztal
	Keith Bertrand

Electronic Version Approved:

Julie Coffield
Interim Dean of the Graduate School
The University of Georgia
December 2014

DEDICATION

This work is dedicated to my lovely wife, Maryam Forouhi, without whose caring support it would not have been possible and to my delightful parents, Abbas Toghiani and Nahid Toghiani, whose words of encouragement and push for tenacity ring in my ears and to my sister, Faezeh, and my brother, Amin, which have never left my side and are very special.

ACKNOWLEDGEMENTS

Sincere appreciation and gratitude is extended to my advisor, Dr. Romdhane Rekaya, for his constructive criticism, insightful guidance and immense help and support throughout the preparation and completion of this Master thesis. Thanks also go to other members of my advisory committee, Dr. Ignacy Misztal and Dr. Keith Bertrand, for their support and helpful discussions during my programs and to Dr. Sargolzaei for developing the QMSim software which accelerated my study. I would also like to thank all the faculty, staff and graduate students in The Department of Animal and Dairy Science UGA for providing a pleasant and stimulating environment. To my friends and faculties in the genetics and animal breeding group, thank you for listening, offering me advice, and supporting me through this entire process. Special thanks to all the people in the group, including: Shogo Tsuruta, Daniela Lino, Breno Fragomeni, El Hamidi Hay, Xinyue Zhang. To my friends scattered around the country and the world, thank you for your thoughts, well-wishes/prayers, phone calls, e-mails, texts, visits, editing advice, and being there whenever I needed as a friend. Finally, I thank the Almighty GOD for His Grace and Blessings without which this program would have been impossible.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

Genomic selection (GS) using dense single nucleotide polymorphism (SNP) markers covering the whole genome provides an innovative tool that allows the tracking of relevant genes and QTLs affecting traits of interests. Through linkage disequilibrium (LD), QTLs effects could be estimated using genotyped SNPs and their joint effects, or genomic breeding values (GEBVs), can be easily obtained. Accuracy of genomic selection depends on several factors including the size of the raining populations, population structure, method of analysis, and the density of SNP marker panel. Recent development in high-throughput genotyping technology has made the genotyping at an ever increasing density a reality with continuous decreases in costs. Unfortunately, in several livestock applications, the genotyping costs still too high for an extensive use in the population. A practical and cost effective alternative that is being adopted by the livestock industry is to genotype a fraction of the population with low-density and low-cost SNP panels. However, the density in the latter has to be increase in order to produce accurate GEBVs. This has been accomplished through the imputation of non-typed SNPs in the low density panel. Several methods have developed and high imputation accuracies are often obtained as long as minimum requirement are met (size of reference population, genetic distance, number and distribution of SNPs in the low density panel, allele frequencies, LD between SNPs in the low and high density panels). Although successful, imputation accuracy is expected to decay as the generational interval between the reference and testing populations increases.

In species with short generation interval, it does not take long to run several rounds of selection. Given the extensive use of SNP genotype imputation in livestock applications, it is of practical importance to evaluate in long term effectiveness in the implementation of genomic selection. The objectives of this study are: 1) to evaluate the decay of imputation accuracy over generations, and 2) to investigate the effect of multi-generational decay in imputation effectiveness on the accuracy of genomic selection. To reach these objectives, several data sets were simulation under varying conditions and were used in the analyses.

CHAPTER 2

LITERATURE REVIEW

Animal selection: From BLUP to Genomics

Animal breeding aims to improve economic productivity of future generations of domestic species through selection under changing costs and income scenarios. Most of the traits of economic interest in livestock are of a complex genetic nature and are under the influence for large number of genes, environmental factors, and their potential interactions. The introduction of best linear unbiased prediction (BLUP) represented a crucial milestone the field of animal breeding and genetics (Henderson 1975). Using BLUP machinery, unbiased estimates of breeding values with reasonable reliability can be obtained based on pedigree information across many generations and phenotype information collected on candidates to selection and their relatives. Elite individuals are selected according to the rank order of their estimated breeding values (EBVs) obtained using BLUP. Identifying genes affecting complex traits would greatly enhance our understanding of their genetic architecture and will undoubtedly enhance the selection process and ultimately the genetic response. Traditionally, the genetics of complex traits have been studied without identifying the genes involved. With all its limitations, this strategy has proven to be successful for most traits. However, for traits with low heritability, sex-limited traits, longevity traits, and traits that are difficult or expensive to measure (such as carcass traits), traditional selection methods were less successful.

With the advances in molecular genetics in the past 20 years, genetic markers in linkage disequilibrium within families or population wide with quantitative trait loci (QTLs) have been identified and several attempts were carried out for their potential use in genetic improvement programs (Goddard & Hayes 2009) through the so called marker-assisted selection (MAS). Unfortunately, such approach had little to no success due to multitude of factors. Chiefly, was the inability to precisely identify enough QTLs affecting the traits of interest. Early studies postulated that the number of loci influencing a specific trait will be small (Hayes & Goddard 2001) which turned out to be not true. (Reed *et al.* 2008). Additionally, although several statistical approaches have been developed to map QTLs (Sillanpaa & Corander 2002; Meuwissen & Goddard 2004), their associated effects have been often overestimated (Utz *et al.* 2000) precluding, thus, their confirmation in independent data sets. Additionally, the majority of identified QTLs using this approach were in population wide linkage equilibrium which complicated further their use in a marker assisted improvement program. Collectively, these and other factors have led to the quick demise of MAS in the field of animal breeding and genetics.

The sequencing of the human genome, completed in 2003, followed by those of several animal species as cattle (Bovine Genome *et al.* 2009), have paved the way to a new tool that uses genomic information for animal selection. The idea of using dense marker maps to predict the genetic merit was proposed first by Meuwissen *et al.* (2001) and has revolutionized the field of animal breeding and genetics in a way and at a speed not shown before. Modern sequencing techniques allow for the genotyping of thousands of variants. Currently, it is possible to efficiently generate high density SNP panels at a reasonable cost. Contrarily to MAS, the use of high density SNP markers does not require

the prior mapping of potential QTLs affecting the trait and thus illuminating all the problems indicated before. In fact, all potential QTLs are tracked by their associated SNP markers or haplotypes and genomic breeding values can be estimated as a function of their joint effects (Meuwissen *et al.* (2001). Although several authors (Nejati-Javaremi *et al.* 1997; Haley & Visscher 1998; Whittaker *et al.* 2000) have presented several idea for this type of genomic selection (GS), it was the groundbreaking article by Meuwissen *et al.* (2001) that laid the basis for using dense SNP markers in a simple yet powerful model to regress phenotypes on genomic variation. It did not take long since then for the proposed approach to become a standard tool for genomic selection in animal (VanRaden *et al.* 2009) and plant breeding (Bernardo & Yu 2007; Crossa *et al.* 2010).

SNP marker Arrays

The first decade of the 21st century has been a golden time for the advancement of genomics, driven by the completion of the Human Genome Project (HGP). Various methodologies and technologies have been developed during and after the process of building the human genetic blueprint that has been directly transferred into the studies of domestic animal genomics (Andersson 2009; Goddard & Hayes 2009). Single nucleotide polymorphisms (SNPs) are bi-allelic genetic markers, they are easy to assess and interpret, and they are widely distributed across the genome. With proper coverage and density, SNPs could capture the LD information embedded in the genome, which in turn could be used to pinpoint genes underlying complex traits. For domestic animals, these tools can be used in several applications including: i) a better understanding of evolution, domestication and breed formation, and the development of new theories of population

genetics; ii) to dissect the genetic mechanisms of complex agricultural traits; and iii) to enhance selection methods for genetic improvement of animal production. Currently, High density SNP arrays with varying number of SNP marker are available for the majority of domestic animal species. These SNP arrays are having considerable impacts on the theoretical and practical aspects of animal breeding and genetics and they will be gaining more importance in the years to come. When Meuwissen et al. (2001) published their paper, it was not clear when appropriate SNP arrays would be available to predict genomic breeding values with a level of reliability that is necessary for an application under practical circumstances in routine evaluations. In less than ten years and using new technologies, it is now possible to genotype individuals for many hundreds of thousands of SNPs in one step at reasonable costs. A key issue in a GS program in livestock is the optimization of genomic information in breeding programs (Pryce & Daetwyler 2012). Low density SNP panels have been developed with the objective of reducing genotyping costs. These low density and low cost platforms will increase the number of genotyped animals. However, their performance is less than desirable. Consequently, imputation methods have been developed to solve this problem.

Imputation of SNP genotypes

Genotype imputation is a powerful tool that allows the inclusion of animals genotyped with low density arrays in the genomic evaluation without having to genotype them with more expensive HD panels. Additionally, it can be used to impute missing SNP genotypes to hybridization problems or quality control issues which will help

increase the accuracy of genomic selection (Weigel et al. 2010; Zhang & Druet 2010). Imputation is a cost effective in-silico genotyping of missing genotypes.

Based on the sources of information used to infer the missing genotype, imputation methods can be implemented based on pedigree or/and population information. Family based imputation uses linkage and Mendelian segregation rules and is the most accurate for animals with genotyped relatives. Population imputation uses linkage disequilibrium information between the missing SNP and the observed flanking SNPs and is useful for unrelated animals or for animals without genotyped close ancestors. Pedigree based approach is more powerful (Hickey *et al.* 2012b). This is due for least two reasons: 1) the phase of HD genotyped individuals can generally be resolved more accurately using pedigree rules compared to linkage disequilibrium based phasing algorithms, and 2) pedigree information can be used during imputation to significantly reduce the number of plausible haplotypes that can be carried by an individual.

Various algorithms have been developed for imputation of missing genotypes. AlphaImpute (Hickey *et al.* 2012b), FImpute (Sargolzaei. *et al.* 2011), and findhap (VanRaden *et al.* 2011) use pedigree information, although the latter is not compulsory for FImpute. These imputation software were developed for animals and plants applications and can be efficiently implemented even with complex pedigrees. Mainly for human applications were pedigree information is seldom available, several population based imputation algorithms such as fastPHASE (Scheet & Stephens 2006), IMPUTE (Howie *et al.* 2009), MACH (Li *et al.* 2010), Beagle (Browning & Browning 2007) were developed. In a population of unrelated animals, the shared haplotype stretches are shorter, because common ancestors are more distant, and more complex algorithms are

required for accurate imputation. This often results in extremely long computing time when it is applied to larger data sets.

Imputation accuracy can be assessed using several criteria including the percentage of correctly imputed genotypes, the percentage of alleles imputed correctly, the correlation between true and imputed genotypes, and the proportion of variation contained within the true high-density genotypes that is explained by the imputed genotypes (R^2). Imputation accuracy is influenced by several factors, including the number and distribution of markers on the low-density genotyping panel, the number of individuals genotyped at high-density and their relationships with the individuals to be imputed, allele frequencies, and finally the local linkage disequilibrium between each low-density genotype and its surrounding high-density genotypes (Zhang & Druet 2010; Hickey *et al.* 2012b; Huang *et al.* 2012). Apart from the choice of the program, the size and the composition of the reference set are the two factors that mainly influence the accuracy of imputation (Pausch *et al.* 2013). Larger reference sets and a larger number of close relatives increase the accuracy; however, the more animals have to be genotyped with the higher marker density; the greater is the cost. Thus, one of the strategies that is often used is to select key ancestors in a manner to maximize their contribution to the gene pool of the actual population (Goddard & Hayes 2009) and to genotype these ancestors with the HD chip. Imputation of missing marker genotypes is based on available marker data from a given population. The population structure and the frequencies of marker genotypes in the given population have influence on the imputation accuracy (Druet *et al.* 2010; Dassonneville *et al.* 2011; Hickey *et al.* 2012a). Because of differences in algorithms and source of information, the superiority of an imputation

method could be data dependent. Therefore, it is necessary to find the optimal imputation method and strategy to be used in the population of interest.

Genomic Selection

Genomic selection was the major innovation in the field of animal breeding and genetics in the past decade. Rapid development in high throughput sequencing and genotyping technologies played major role on its success. Efficiency of GS derives for its potential effects of the factors affecting genetic response. Genetic gain (ΔG) depends on the intensity of selection (i), the accuracy of predictions (r), the additive genetic standard deviation of the trait (σ_g), and the generation interval (L):

$$\Delta G = (i \times r \times \sigma_g\ )/L$$

It is clear that GS has effects on the factors affecting the genetic gain. Using genomic information, candidates to selection could be evaluate at early age or even at the embryonic stage reducing substantially the generation interval compared to classical approach. Theoretically, more young animals could be tested which in turn will increase the pool of potential candidates allowing hence for an increase in selection intensity. It increases the prediction accuracy of genetic merit, shortening the generation interval, and therefore increasing the rate of genetic gain. Using dense marker maps, it is possible to track even QTLs with small effects and to improve the estimation of the additive relationships between relatives. This will lead to more accurate estimates of breeding values. Although the accuracy of GS depends, among other factors, on the LD between SNPs (Calus *et al.* 2008), high or even medium density arrays often provide enough LD between genome segments to trace QTLs affecting traits of interest.

Genomic Prediction

Deriving accurate predictions of complex traits requires the implementation of whole-genome regression models where phenotypes are regressed simultaneously on thousands of markers. The principle of genomic selection is to first estimate the effects of all markers in a training population consisting of phenotyped and genotyped individuals and then use these estimates in a validation population to evaluate their efficiency (Meuwissen et al. 2001). Genomic estimated breeding values (GEBVs) are then calculated as the sum of estimated marker effects for genotyped individuals. The training population consists typically of individuals having reliable dependent variables as well as genomic information. Dependent variables could be actual phenotypes, estimated breeding values or de-regressed proofs among others (de Roos *et al.* 2007), (VanRaden *et al.* 2009); (Gonzalez-Recio *et al.* 2008). By regressing the dependent variables on the genomic information, estimates for SNP effects are obtained. The estimates are then used on genotyped young selection candidates whose GEBVs are obtained by summing up all the relevant SNP effects. Alternatively, GS could be implemented by regressing the phenotypes directly on the GEBVs as suggested by (Legarra *et al.* 2009; Misztal *et al.* 2009).

Accuracy of GS depends on several factors including the size and structure of the training population, the heritability of trait, the density of SNP marker map, quality of the dependent variable and genomic information, the genetic relatedness between training and validation sets, the LD between marker and QTLs, and the effective population size. There is a huge literature on the importance of these factors on GS (Habier *et al.* 2007; Muir 2007; Calus *et al.* 2008; Solberg *et al.* 2008; Meuwissen 2009).

Implementation of genomic prediction

Genomic prediction approaches can be categorized as either direct or indirect methods. Indirect methods, which include Bayesian approaches, estimate marker effects in a reference population and then calculate the GEBVs of genotyped individuals by summing the effects of all relevant markers. Direct methods calculate the GEBVs directly using mixed model methodology within BLUP framework. Phenotype information of the reference population and the genotype information for all the markers from both reference and candidate populations are used, and a marker-derived relationship matrix is constructed using the high density markers (Zhang et al. 2011) to replace the average additive relationship matrix.

Using indirect methods, that estimate the SNP effects first, often run into the well-known large (p) and small (n) problem when high density SNP panels are used. The number of markers (p) can greatly exceed the number of records (n) where only a few thousand individuals are phenotyped and genotyped. To overcome this over-parameterization, several approaches have been developed and implemented.

Bayesian Methods

Bayesian methods, through the prior information on the unknown parameters, provide a nature way of dealing with the over-parameterization of the genomic model. In fact, genomic selection was first introduced using a Bayesian approach (Hayes & Goddard 2001; Meuwissen *et al.* 2001). Since, several variations of the original approach have been presented under different labelling (Gianola *et al.* 2009; Habier *et al.* 2010;

Fernando *et al.* 2014). At the theoretical level, all these Bayesian methods are the same and differ only on the prior information specified for the unknown parameters.

The General model for regressing the phenotype (or Pseudo-phenotype) on the genomic information could be written as:

$$y_i = \mu + \sum_{j=1}^{p} X_{ij} b_j + e_i$$

where y_i is the phenotype for animal i, μ is an overall mean, X_{ij} is the genotype for SNP j (j=1,2,…,p) for animal i coded as 0,1, or 2, b_j and e_i is the random residual term.

In a Full Bayesian implementation, prior distributions are needed for all unknown in the model. For the different Bayesian methods used in genome wide associations and GS, their only difference is on the prior specified for the SNP effects. Specifically, assuming a normal with the same variance for all SNPs in the array leads to the BLUP like implementation.

$$b_i \sim N(0, \sigma_b^2)$$

where σ_b^2 is a dispersion parameter comment to all SNPs. When the prior variance is specific to each SNP, such specification results in the BayesA implementation.

$$b_i \sim N(0, \sigma_{bi}^2)$$

Additionally, it is reasonable to assume that not all SNPs in the array are in LD with QTLs. Thus, not all SNPs will have non-zero effects. To convey such stage about the SNP effects of belief a priori, the following prior could be specified

$$b_i \sim N(0, V) \text{ with probability } (1-\pi)$$

$$b_i \sim I_0 \text{ with probability } \pi$$

where I_0 is the degenerate distribution with mass at zero, π is the portion of SNPs with zero effects, and V is the prior variance for non-zero effect SNPs. If a specific variance is specified for each SNP ($V = \sigma_{bi}^2$), such prior results in the BayesB implementation. When all non-zero effect SNPs have the same prior variance ($V = \sigma_b^2$), it will lead to the BayesC implementation. Additionally, if π was assumed unknown, it will result in the BayesCπ method.

Several other modifications of the prior have proposed in the literature leading to a myriad of Bayesian methods for the implementation of GWAS and GS (de los Campos *et al.* 2012; Vandenplas & Gengler 2012)

Single-Step Genomic selection

Single-step genomic selection is a unified approach eliminating the SNP markers effects estimation as in the multi-step (Misztal et al. 2009). This approach is based on an enhanced relationship matrix, called a genomic relationship matrix which combines genomic information and pedigree information as described by Legarra et al. (2009). The model used is as follows:

$$\mathbf{y = Xb + Zu + e}$$

y is a vector of observations, **b** is a vector of fixed effects and **u** is a vector of random animal effects and e is the residual. The relationship matrix can be modified to **H = A + $\mathbf{A_\Delta}$** to account for genomic information and $\mathbf{A_\Delta}$ is the deviation from expected relationships. Matrix G replaces the numerator relationship matrix for the genotyped animals (Legarra et al. 2009). Solving MME is exactly the same as in traditional mixed models. A detailed explanation on the construction of the G matrix could be found in VanRaden (2008).

$$G = ZZ'/[2\sum p_i q_i]$$

Where **Z** is $n \times m$ genotypes matrix, n is the number of animals and m is the number of genotypes and p_i and q_i are allelic frequencies. Division by $[2\sum p_i q_i]$ makes **G** analogous to **A**.

Aguilar *et al.* (2011) implemented a single-step procedure for genomic evaluation using national evaluation framework and compared its performance to a multiple-step procedure. The single step approach performed similarly to the multi-step approach and yielded similar accuracies. It is important to note that the single step approach has many advantages compared to the multiple step approach. Multiple step procedure requires 1) classical animal model evaluation 2) generation of pseudo phenotypes such as de-regressed proofs or daughter deviations 3) estimation of a large number of parameters (Misztal *et al.* 2009; Aguilar *et al.* 2011). Single step eliminates all these steps.

References

Aguilar I., Misztal I., Legarra A. & Tsuruta S. (2011) Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *J Anim Breed Genet* **128**, 422-8.

Andersson L. (2009) Genome-wide association analysis in domestic animals: a powerful approach for genetic dissection of trait loci. *Genetica* **136**, 341-9.

Bernardo R. & Yu J. (2007) Prospects for Genomewide Selection for Quantitative Traits in Maize All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher. *Crop Science* **47**, 1082-90.

Bovine Genome S., Analysis C., Elsik C.G., Tellam R.L., Worley K.C., Gibbs R.A., Muzny D.M., Weinstock G.M., Adelson D.L., Eichler E.E., Elnitski L., Guigo R., Hamernik D.L., Kappes S.M., Lewin H.A., Lynn D.J., Nicholas F.W., Reymond A., Rijnkels M., Skow L.C., Zdobnov E.M., Schook L., Womack J., Alioto T., Antonarakis S.E., Astashyn A., Chapple C.E., Chen H.C., Chrast J., Camara F., Ermolaeva O., Henrichsen C.N., Hlavina W., Kapustin Y., Kiryutin B., Kitts P., Kokocinski F., Landrum M., Maglott D., Pruitt K., Sapojnikov V., Searle S.M., Solovyev V., Souvorov A., Ucla C., Wyss C., Anzola J.M., Gerlach D., Elhaik E., Graur D., Reese J.T., Edgar R.C., McEwan J.C., Payne G.M., Raison J.M., Junier

T., Kriventseva E.V., Eyras E., Plass M., Donthu R., Larkin D.M., Reecy J., Yang M.Q., Chen L., Cheng Z., Chitko-McKown C.G., Liu G.E., Matukumalli L.K., Song J., Zhu B., Bradley D.G., Brinkman F.S., Lau L.P., Whiteside M.D., Walker A., Wheeler T.T., Casey T., German J.B., Lemay D.G., Maqbool N.J., Molenaar A.J., Seo S., Stothard P., Baldwin C.L., Baxter R., Brinkmeyer-Langford C.L., Brown W.C., Childers C.P., Connelley T., Ellis S.A., Fritz K., Glass E.J., Herzig C.T., Iivanainen A., Lahmers K.K., Bennett A.K., Dickens C.M., Gilbert J.G., Hagen D.E., Salih H., Aerts J., Caetano A.R., Dalrymple B., Garcia J.F., Gill C.A., Hiendleder S.G., Memili E., Spurlock D., Williams J.L., Alexander L., Brownstein M.J., Guan L., Holt R.A., Jones S.J., Marra M.A., Moore R., Moore S.S., Roberts A., Taniguchi M., Waterman R.C., Chacko J., Chandrabose M.M., Cree A., Dao M.D., Dinh H.H., Gabisi R.A., Hines S., Hume J., Jhangiani S.N., Joshi V., Kovar C.L., Lewis L.R., Liu Y.S., Lopez J., Morgan M.B., Nguyen N.B., Okwuonu G.O., Ruiz S.J., Santibanez J., Wright R.A., Buhay C., Ding Y., Dugan-Rocha S., Herdandez J., Holder M., Sabo A., Egan A., Goodell J., Wilczek-Boney K., Fowler G.R., Hitchens M.E., Lozado R.J., Moen C., Steffen D., Warren J.T., Zhang J., Chiu R., Schein J.E., Durbin K.J., Havlak P., Jiang H., Liu Y., Qin X., Ren Y., Shen Y., Song H., Bell S.N., Davis C., Johnson A.J., Lee S., Nazareth L.V., Patel B.M., Pu L.L., Vattathil S., Williams R.L., Jr., Curry S., Hamilton C., Sodergren E., Wheeler D.A., Barris W., Bennett G.L., Eggen A., Green R.D., Harhay G.P., Hobbs M., Jann O., Keele J.W., Kent M.P., Lien S., McKay S.D., McWilliam S., Ratnakumar A., Schnabel R.D., Smith T., Snelling W.M., Sonstegard T.S., Stone R.T., Sugimoto Y., Takasuga A., Taylor J.F., Van

Tassell C.P., Macneil M.D., Abatepaulo A.R., Abbey C.A., Ahola V., Almeida I.G., Amadio A.F., Anatriello E., Bahadue S.M., Biase F.H., Boldt C.R., Carroll J.A., Carvalho W.A., Cervelatti E.P., Chacko E., Chapin J.E., Cheng Y., Choi J., Colley A.J., de Campos T.A., De Donato M., Santos I.K., de Oliveira C.J., Deobald H., Devinoy E., Donohue K.E., Dovc P., Eberlein A., Fitzsimmons C.J., Franzin A.M., Garcia G.R., Genini S., Gladney C.J., Grant J.R., Greaser M.L., Green J.A., Hadsell D.L., Hakimov H.A., Halgren R., Harrow J.L., Hart E.A., Hastings N., Hernandez M., Hu Z.L., Ingham A., Iso-Touru T., Jamis C., Jensen K., Kapetis D., Kerr T., Khalil S.S., Khatib H., Kolbehdari D., Kumar C.G., Kumar D., Leach R., Lee J.C., Li C., Logan K.M., Malinverni R., Marques E., Martin W.F., Martins N.F., Maruyama S.R., Mazza R., McLean K.L., Medrano J.F., Moreno B.T., More D.D., Muntean C.T., Nandakumar H.P., Nogueira M.F., Olsaker I., Pant S.D., Panzitta F., Pastor R.C., Poli M.A., Poslusny N., Rachagani S., Ranganathan S., Razpet A., Riggs P.K., Rincon G., Rodriguez-Osorio N., Rodriguez-Zas S.L., Romero N.E., Rosenwald A., Sando L., Schmutz S.M., Shen L., Sherman L., Southey B.R., Lutzow Y.S., Sweedler J.V., Tammen I., Telugu B.P., Urbanski J.M., Utsunomiya Y.T., Verschoor C.P., Waardenberg A.J., Wang Z., Ward R., Weikard R., Welsh T.H., Jr., White S.N., Wilming L.G., Wunderlich K.R., Yang J. & Zhao F.Q. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522-8.

Browning S.R. & Browning B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97.

Calus M.P., Meuwissen T.H., de Roos A.P. & Veerkamp R.F. (2008) Accuracy of genomic selection using different methods to define haplotypes. *Genetics* **178**, 553-61.

Crossa J., eacute, Campos G.d.L., eacute, rez P., Gianola D., Burgue, ntilde, o J., Araus J., eacute, Luis, Makumbi D., Singh R.P., Dreisigacker S., Yan J., Arief V., Banziger M. & Braun H.-J. (2010) *Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers*.

Dassonneville R., Brondum R.F., Druet T., Fritz S., Guillaume F., Guldbrandtsen B., Lund M.S., Ducrocq V. & Su G. (2011) Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. *J Dairy Sci* **94**, 3679-86.

de los Campos G., Hickey J.M., Pong-Wong R., Daetwyler H.D. & Calus M.P.L. (2012) Whole Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics*.

de Roos A.P., Schrooten C., Mullaart E., Calus M.P. & Veerkamp R.F. (2007) Breeding value estimation for fat percentage using dense markers on Bos taurus autosome 14. *J Dairy Sci* **90**, 4821-9.

Druet T., Schrooten C. & de Roos A.P.W. (2010) Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *Journal of Dairy Science* **93**, 5443-54.

Fernando R., Dekkers J. & Garrick D. (2014) A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genetics Selection Evolution* **46**, 50.

Gianola D., de los Campos G., Hill W., Manfredi E. & Fernando R. (2009) Additive genetic variability and the bayesian alphabet. *Genetics* **183**, 347 - 63.

Goddard M.E. & Hayes B.J. (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* **10**, 381-91.

Gonzalez-Recio O., Gianola D., Long N., Weigel K.A., Rosa G.J. & Avendano S. (2008) Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics* **178**, 2305-13.

Habier D., Fernando R., Kizilkaya K. & Garrick D. (2010) Extension of the Bayesian alphabet for genomic selection. *Proceedings of the 9th World Congress on Genetics applied to Livestock Production: 1-6 August 2010*, 468 -

Habier D., Fernando R.L. & Dekkers J.C.M. (2007) The Impact of Genetic Relationship Information on Genome-Assisted Breeding Values. *Genetics* **177**, 2389-97.

Haley C.S. & Visscher P.M. (1998) Strategies to utilize marker-quantitative trait loci associations. *J Dairy Sci* **81 Suppl 2**, 85-97.

Hayes B. & Goddard M.E. (2001) The distribution of the effects of genes affecting quantitative traits in livestock. *Genet Sel Evol* **33**, 209-29.

Henderson C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423-47.

Hickey J.M., Crossa J., Babu R. & de los Campos G. (2012a) Factors Affecting the Accuracy of Genotype Imputation in Populations from Several Maize Breeding Programs. *Crop Science* **52**, 654.

Hickey J.M., Kinghorn B.P., Tier B., van der Werf J.H. & Cleveland M.A. (2012b) A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genetics Selection Evolution* **44**, 9.

Howie B.N., Donnelly P. & Marchini J. (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet* **5**, e1000529.

Huang Y., Hickey J.M., Cleveland M.A. & Maltecca C. (2012) Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. *Genetics Selection Evolution* **44**, 25.

Legarra A., Aguilar I. & Misztal I. (2009) A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science* **92**, 4656-63.

Li Y., Willer C.J., Ding J., Scheet P. & Abecasis G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34**, 816-34.

Meuwissen T.H. (2009) Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. *Genet Sel Evol* **41**, 35.

Meuwissen T.H. & Goddard M.E. (2004) Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genetics Selection Evolution* **36**, 261-79.

Meuwissen T.H., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819-29.

Misztal I., Legarra A. & Aguilar I. (2009) Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *Journal of Dairy Science* **92**, 4648-55.

Muir W.M. (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *J Anim Breed Genet* **124**, 342-55.

Nejati-Javaremi A., Smith C. & Gibson J.P. (1997) Effect of total allelic relationship on accuracy of evaluation and response to selection. *J Anim Sci* **75**, 1738-45.

Pausch H., Aigner B., Emmerling R., Edel C., Gotz K.U. & Fries R. (2013) Imputation of high-density genotypes in the Fleckvieh cattle population. *Genetics Selection Evolution* **45**, 3.

Pryce J.E. & Daetwyler H.D. (2012) Designing dairy cattle breeding schemes under genomic selection: a review of international research. *Animal Production Science* **52**, 107-14.

Reed D.R., Lawler M.P. & Tordoff M.G. (2008) Reduced body weight is a common effect of gene knockout in mice. *BMC Genetics* **9**, 4.

Sargolzaei. M., Chesnais. J.P. & Schenkel. F.S. (2011) FImpute: An efficient imputation algorithm for dairy cattle populations. *J. Dairy Sci*. **(E-Suppl. 1):421;(Abstr.)**.

Scheet P. & Stephens M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**, 629-44.

Sillanpaa M.J. & Corander J. (2002) Model choice in gene mapping: what and why. *Trends in Genetics* **18**, 301-7.

Solberg T.R., Sonesson A.K., Woolliams J.A. & Meuwissen T.H. (2008) Genomic selection using different marker types and densities. *J Anim Sci* **86**, 2447-54.

Utz H.F., Melchinger A.E. & Schon C.C. (2000) Bias and Sampling Error of the Estimated Proportion of Genotypic Variance Explained by Quantitative Trait Loci Determined From Experimental Data in Maize Using Cross Validation and Validation With Independent Samples. *Genetics* **154**, 1839-49.

Vandenplas J. & Gengler N. (2012) Comparison and improvements of different Bayesian procedures to integrate external information into genetic evaluations. *J Dairy Sci* **95**, 1513-26.

VanRaden P.M. (2008) Efficient methods to compute genomic predictions. *J Dairy Sci* **91**, 4414-23.

VanRaden P.M., O'Connell J.R., Wiggans G.R. & Weigel K.A. (2011) Genomic evaluations with many more genotypes. *Genetics Selection Evolution* **43**, 11.

VanRaden P.M., Van Tassell C.P., Wiggans G.R., Sonstegard T.S., Schnabel R.D., Taylor J.F. & Schenkel F.S. (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci* **92**, 16-24.

Weigel K.A., Van Tassell C.P., O'Connell J.R., VanRaden P.M. & Wiggans G.R. (2010) Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. *J Dairy Sci* **93**, 2229-38.

Whittaker J.C., Thompson R. & Denham M.C. (2000) Marker-assisted selection using ridge regression. *Genetical Research* **75**, 249-52.

Zhang Z. & Druet T. (2010) Marker imputation with low-density marker panels in Dutch Holstein cattle. *J Dairy Sci* **93**, 5487-94.

Zhang Z., Zhang Q. & Ding X. (2011) Advances in genomic selection in domestic animals. *Chinese Science Bulletin* **56**, 2655-63.

CHAPTER 3

MULTI-GENERATIONAL EFFECT ON IMPUTATION ACCURACY OF SINGLE NUCLEOTIDE POLYMORPHISMS [1]

[1] Toghiani, S. and R. Rekaya. To be submitted to Animal.

Abstract

Genomic selection requires the direct or indirect estimation of the effect of each SNP on the panel within the framework of mixed linear methodology. Several tools are already available to implement this step and accuracies of resulting breeding values are 30 to 70% higher than those obtained using classical quantitative approaches. However, high accuracies are achieved only when a reasonable number of SNPs are genotyped for each individual. Medium and high density SNP panels are often sufficient to reach such objective. Unfortunately, their current cost precludes their extensive use in the majority of livestock and poultry applications. Therefore, low-density and low-cost SNP marker chips represent a practical viable alternative to reduce the genotyping costs without excessive lost in accuracies. However, to use these low density panels in genomic breeding values prediction (GEBVs), the missing (non-genotyped SNP markers) need to be inferred. Although SNP marker genotype imputation is well studied and resulting accuracies are sufficiently high, its performance over time (generations) received little attention. In this study, accuracy of imputation over time was evaluated under varying simulation scenarios. Using fixed reference populations, imputation accuracy decayed by around 0.5% per generations. In fact, after 25 generations, the accuracy was only 7% lower than the first generation. When the reference population was updated by either 1% or 5% of the top animals in the previous generations, decay of imputation accuracy was substantially reduced. These results indicate that low density panels are useful, especially when the generational interval between reference and testing population is small. As the generational interval increases, the imputation accuracies decay, although not at an alarming rate.

Introduction

Traditionally, phenotypic and pedigree information were used successfully by animal breeders to identify superior animals to be used as parents of the next generation. Using this information, unprecedented genetic progress was observed in every specie and trait of economic interest. Although very successful, this approach tends to be time consuming, costly and requires a sophisticated logistic network for data collection and processing. For many traits, especially those expressed late in the life of an animal or having low heritability, the success of the classical approach tends to be limited. More importantly, the classical approach is based on using the expected additive relationships between individuals to identify superior animals conditionally to the collected phenotypic data. As such, the expected additive relationship does not represent the true relationship between two individuals. Thus, there is no doubt that better estimates of breeding values could be obtained if realized, rather than expected, additive relationships are used in inferring the EBVs. High density genotyping for SNP markers provides such opportunity. Advances in DNA sequencing technology and the increased availability of high density single nucleotide polymorphism (SNP) genotyping platforms provided unprecedented opportunities to enhance breeding programs in livestock, poultry and plant species and to better understand the genetic basis of complex traits. Using this genomic information in a breeding program within the so called genomic selection requires the direct or indirect estimation of the effect of each SNP on the high-density panel within the framework of mixed linear methodology. In fact, several tools are already available to implement this step and accuracies of resulting breeding values are 30 to 70% higher than those obtained using classical quantitative approaches (Kapell *et al.* 2012; Pungpapong *et al.* 2012; Scutari

et al. 2013). In Genomic selection (GS) using dense marker maps covering the whole genome allows for the effective tracking of all potential QTLs affecting the trait through the linkage disequilibrium (LD) with genotyped SNP markers. Genetic merit of all genotyped individuals could be then calculated based on the estimation of marker effects (Meuwissen *et al.* 2001) or on the relationship derived from whole-genome dense markers (VanRaden 2008).

However, high accuracies are achieved only when a reasonable number of SNPs are genotyped for each individual. Medium and high density SNP panels are often sufficient to reach such objective. Unfortunately, their current cost precludes their extensive use in some livestock and poultry applications. For example, selection of replacement heifers in dairy and beef herds cannot be conducted using high density SNP panels and low-density SNP chips (e.g. 3K or 7K) may be the only cost effective option available (Pryce & Hayes 2012).

Therefore, low-density and low-cost SNP marker chips could represent a practical viable alternative to reduce the genotyping costs without excessive lost in accuracies. However, to use these low density panels in genomic breeding values prediction (GEBVs), the missing (non-genotyped SNP markers) need to be inferred. This could be accomplished using imputing techniques to predict the missing SNP genotypes resulting, thus, in a denser coverage. Genotype imputation technology, a cost effective in-silico genotyping of missing SNP markers, allows animal breeders to genotype animals with affordable low-density panels and predicts the un-typed genotypes from the high-density panel (Johnston. *et al.* 2011). The use of low-density marker panels for genomic evaluation based on the standard 50K chip requires statistical methods for transferring genotype information from

individuals genotyped at a higher density. Imputation can be performed to fill in SNP genotypes from a higher density reference panel to lower SNP panel genotypes (Weigel *et al.* 2009; Druet *et al.* 2010; Zhang & Druet 2010; VanRaden *et al.* 2011). Based on the genomic information used to derive the missing genotypes, imputation methods could be divided into two categories: 1) pedigree-based imputation, and 2) population-based imputation. Pedigree based imputation uses the correlation of genotypes among relatives derived from the sharing of genomic segments identical by descent (IBD) within pedigrees and linkage information. In contrast, population-based imputation leverages information from the correlation among dense markers due to LD. Population-based imputation is useful especially for sets of unrelated individuals or for animals without genotyped close ancestors (Cheung *et al.* 2013). Such features make it possible to impute genotypes at un-typed markers in a high-density panel of markers from genotypes obtained with a low-density panel. According to the two different categories of genotype imputation methods, there are different genotype imputation programs available. fastPHASE (Scheet & Stephens 2006), MaCH (Li *et al.* 2010), Beagle (Browning & Browning 2007) and IMPUTE2 (Howie *et al.* 2009), which were designed specifically for human populations, using linkage disequilibrium information. AlphaImpute (Hickey *et al.* 2012), FImpute (Sargolzaei. *et al.* 2011) and findhap (VanRaden *et al.* 2011), which were developed for animals and plants applications, use pedigree and linkage information. It is worth mentioning that FImpute has an option to impute missing genotypes based on population and/or pedigree information. For both approaches, accuracy of genotype imputation is influenced by several factors; including the number and distribution of markers on the low-density panel, the number of individuals genotyped using the high-density chip (reference

population) and their genetic relationships with the animals to be imputed, allele frequencies at the SNP markers, and the local LD between each low-density genotype and its surrounding high-density genotypes (Zhang & Druet 2010; Hickey *et al.* 2012; Huang *et al.* 2012). Several studies (Badke *et al.* 2014; Chen *et al.* 2014; Sargolzaei *et al.* 2014) have been already conducted to evaluate the efficiency of SNP genotype imputation under different conditions with special emphasis on the accuracy of imputed genotypes and their impact on the quality of the genomic predictions. Results in Jersey cattle indicated that if a suitable reference population genotyped with a 50K chip is available, genotyping selection candidates with a 3K instead of a 50K chip and then imputing the remaining genotypes would result in a loss of predictive ability of only 5% (Weigel *et al.* 2010a). This result is assuring and clearly indicates the practical viability of SNP genotype imputation, at least when the high and low density genotyped sub-populations belong to the same generation or very close together generations. However, little attention has been paid to the imputation accuracy and its potential impact on genomic selection after several cycles of selection.

The objective of this study was to investigate the decay of imputation accuracy over time (generations) under several simulation scenarios mimicking actual livestock populations.

Materials and Methods

Simulation

Population structure: Genomic data were simulated using QMSim software (Sargolzaei & Schenkel 2009) and consisted of 42,000 SNP markers and 1500 QTLs distributed across a 30 Chromosomes genome. Simulating genomic data via QMSim

software consists of a two-step process as indicated in Figure 3.1. In the first step, a historical population is generated. In our case, a population of 600 individuals was maintained through random mating for 200 generations followed by an additional 205, 210 and 220 generations with population size of 5000, 4000 and 3000 respectively, in order to create initial LD and establish mutation-drift equilibrium in the historical generations. The sex ratio in the historical generations was maintained and the mating system was based on random union of gametes, randomly sampled from both the male and female gamete pools.

In the second step of the simulation, the founder population is generated and labeled as generation 0 (G0). In our case, such population was generated from the last historical generation, based on 300 males and 2700 females. The mating of these individuals was at random and no selection was considered at this step. Then after the recent population, 9 generations were simulated and later used to evaluate the imputation process.

Simulation parameters of the most recent generations were kept as close as possible to real beef cattle production system. It was assumed one progeny per dam per year, a sex ratio of 50% in the progeny; selection was based on EBV with a replacement rate of 50% for sires and 20% for dams. Selected sires and dams were randomly mated. A single trait with an overall heritability of 0.40 (QTL=0.2 plus polygenic=0.2) and phenotypic variance of 1.0 was simulated. The true breeding value of an individual was equal to the sum of the QTL additive effects and the polygenic effects. The phenotypes were generated by adding random residuals to the true breeding values. The whole simulation process was repeated 5 times. The parameters of simulation process are summarized in Table 3.1a.

Simulated genome: The simulated genome was carried out using QMSim and consisted of 30 pairs of chromosomes with 100 centi-Morgan (cM) in length each. Each

chromosome harbored 1,400 SNP markers that were evenly distributed. Additionally, 50 randomly distributed QTLs were simulated per chromosome. Both SNP markers and QTLs were assumed to be bi-allelic, and no marker loci overlapped with the QTLs. Further, it was assumed that both SNP markers and QTLs have the same allele frequency in the historical population. Effects of QTLs were sampled from a gamma distribution with shape parameter equal to 0.4. Complete linkage disequilibrium (LD) was simulated between markers, between QTLs and between markers and QTLs in the first historical population. The parameters used for simulating the genome are presented in Table 3.1b.

Imputation of SNP genotypes

Simulated genomic data consisted of the SNP marker genotypes of individuals in the last 9 generations. It included 24,300 genotyped animals with a panel of 42,000 SNP markers (42K) that were distributed equally (2700 genotyped animals) in each generation. To investigate the quality of genotype imputation over generations, genotyped animals were divided into a reference and testing (imputation) data sets. Depending on the simulation scenario, the reference population consisted of: 1) reference population was assumed fixed and included only animals in generations 1 and 2; 2) reference population included animals in generations 1 and 2 and was updated with an additional 1% of top animals in following generations, and 3) reference population included animals in generations 1 and 2 and was updated with an additional 5% of top animals in following generations. Thus, the reference population included at least 5,400 genotyped animals for 42K SNP marker panel. For the first scenario (fixed reference population), the testing populations were each of the remaining 7 generations (generations 3 to 9). For scenarios 2 and 3 (updated reference populations), the testing population consisted only of generation

9. In the testing populations, only 3,000 evenly distributed SNP markers were kept and the remaining 39,000 SNP genotypes were masked to mimic a scenario of low density genotyping with a 3K SNP marker panel. A number of imputation scenarios were carried out in this study. The primary goal of these imputation scenarios was to evaluate the population based imputation from the low density (3K) to high density (42K) over several rounds of selections (generations).

Imputation procedure

Imputation from low density (3K) to the high density panel (42K) was carried out using FImpute v2.2 (Sargolzaei. *et al.* 2011). FImpute was developed primarily to carry out genotype imputation for applications in animals and plants. It uses both population and pedigree information, although the latter is not always required. In absence of family information for an individual or if no pedigree file is included, FImpute will use only the population parameter to carry out the imputation. FImpute uses overlapping windows to reconstruct haplotypes and impute simultaneously. Unlike most population imputation software, FImpute assumes that all animals are related to some degree and uses these overlapping windows to find segments of haplotypes that are consistent between individuals having a common ancestor. The windows tend to be large at first in order to find segments of haplotypes derived from more recent ancestors. The window searches along each chromosome to find large segments consistent with the reference animals. After each chromosome has been covered using large windows, the same process is repeated numerous times although with smaller and smaller windows each time to capture consistent haplotypes from less recent ancestors. When multiple haplotypes are found in a certain window size, haplotype frequency in the reference population and high number are used to

determine the most likely haplotype, and fills that haplotype into the imputed animal's genotype.

Accuracy of Imputation

The original and the imputed genotypes at a given locus were coded 0, 1, and 2 for A_1A_1, A_1A_2, and A_2A_2, respectively. Original and imputed genotypes were deemed concordant if they matched perfectly. To access the accuracy of the imputation, the overall error rate (OER), and the average concordance rate (CR) were calculated over all imputed SNP marker genotypes. The OER consisted of calculating the percentage of the non-concordant genotypes. The CR was calculated based on the number of errors counted at an imputed genotype that was equal to 0 when true and imputed genotypes were identical, 1 if the true genotype was homozygous and the imputed genotype was heterozygous (or vice versa), and 2 if true and imputed genotypes were equal to the alternative.

$$OER = \left(1 - \frac{\#\ of\ non-concordant\ genotypes}{Total\ \#\ of\ masket\ SNPs\ genotypes}\right) x100$$

$$CR = \left(1 - \frac{\#\ of\ errors}{Total\ \#\ of\ masked\ SNPs\ genotypes\ x\ 2}\right) x100$$

Additionally, accuracy was dissected into the two types of imputation errors which could have different implications in association and selection studies: major allele homozygotes to minor allele homozygotes and vice versa (Miss 2), and homozygotes to heterozygotes and vice versa (Miss 1). The former type is potentially riskier as it most likely alters the allele frequencies or even switches major with minor alleles (Huang *et al.* 2009). All imputation accuracy measurements were based on an average of 5 replicates.

Results and Discussion

Imputation scenarios

Table 3.2 presents a summary statistic of the imputation parameters for the different simulation scenarios with the fixed reference population. In all cases, generations 1 and 2 were part of the reference population and consisted of 5,400 individuals genotyped with the high density SNP panel (42K chip). The testing data set consisted of generations 3 to 9. Each of these testing generations included 2,700 individuals. Only the low density SNP genotypes (3K panel) were kept for these individuals. The remaining 39,000 SNP markers were masked and later used in the evaluation of the imputation process. Consequently, the missing rate was 92.86% across the genome and in each one of the 30 chromosomes. Imputation was conducted separately for each of the seven testing data set (S3-S9). In other words, in each imputation scenario only the reference population (generations 1 and 2) and the one of the generations 3 to 9 were included in the analysis.

Imputation Accuracy

Imputation accuracy was evaluated based on the overall error rate (OER), the average rate of concordance (CR), and Miss and Miss 2, as defined earlier. Table 3.3 represents the imputation accuracy, when the reference population was fixed (generations 1 and 2), for the different scenarios and using the OER criteria. All results are based on the average of 5 replicates. As expected, the accuracy of imputing the true SNP marker genotypes decreased with the increase of the generational interval between the reference and testing data sets. This is largely due to the fact that successive mating events lead to a

reduction in size of the LD blocks which in turn will reduce the length of common haplotypes between progeny and ancestors. Thus, the remaining SNP markers in the low density panel will over time lose their predictive ability of predicting the missing genotypes due to the decay in LD. The latter plays a crucial role in the efficiency of the imputation performance (Sargolzaei *et al.* 2008). The results of imputation accuracy are in line with the results reported by (VanRaden *et al.* 2013). In their study, only one chromosome was simulated. The reference population consisted of 1,112 animals genotyped for the HP panel (777K SNPs). Imputation for lower density panels were carried out using FImpute program. Imputation accuracy was 99.96, 99.3, 94.7, and 91.1% when HD, 50K, 6K, and 3K SNP panels were used, respectively. The imputation from the 3K panel in VanRaden et al. (2013) study is similar to our first simulation scenario (S3) where generation 3 was imputed based on genotypes in generations 1 and 2. Within these parameters, our imputation accuracy (89.4%) was similar to their results (91.1%). The slight difference could be due to the size of the reference population, the level of genetic similarity between reference and validation sets, and the much large number of progeny per sire in VanRaden et al. (2013) study. Based on our results and those observed in VanRaden et al. (2013) study, it is clear that imputation of SNP genotype from a very low density panel (3K) leads to acceptable results when the reference and training populations are from adjacent generations; reflective of high genetic similarity. Furthermore, the imputation accuracy increases significantly with the increase of the coverage in the density panel (i.e. from 50K to 777K). This is expected because the higher LD in the low density panel and the shorter the haplotypes to be imputed. However, when the generational interval between the reference and testing population increases, the imputation accuracy decreases as indicated

in Figure 3.2. In this study, the imputation accuracy decreased by an average of 0.5% per generation for the first 7 generations following the reference population. Thus, even after 7 generations, the overall rate to decrease of imputation accuracy was only 2.5%. Similar trend was observed when imputation accuracy was evaluated using the CR as presented in Figure 3.3.

The proportion of Miss 1, as defined earlier, was always much greater than Miss 2, although the latter tends to increase as the generational interval between reference and testing populations increases (Figure 3.4). The increase in Miss 2 is not well captured by the OER criteria and could have a substantial effect on the genome wide association studies and genomic selection.

With updated reference populations, the decay in imputation accuracy was much smaller over generations as presented in Figure 3.5. In fact, when the reference population consisted of generation 1 and 2 plus an additional 1% of the top animals in the following 6 generations (generations 3 to 8), the imputation accuracy for the testing population (generation 9) increased to 87.97%. This increase translates in a 36% reduction in the decay of imputation accuracy between generations 3 and 9 compared to the fixed reference population scenario. When the updating of the reference population consisted of an additional 5% of the animals in generations 3 to 8, the imputation accuracy for the 9^{th} generation increased to 88.74%.

These results represent a snapshot of imputation accuracy and although they highlight a trend, the absolute numerical values depend on a multitude of factors including: 1) imputation algorithm and software package used (Browning & Browning 2007; Hao *et al.* 2009; Howie *et al.* 2009; Nothnagel *et al.* 2009; Ma *et al.* 2013), 2) population structure

and size and genetic similarity between the reference and testing populations (Druet & Georges 2010; Weigel *et al.* 2010b; Zhang & Druet 2010), and 3) density and distribution of SNP markers in the low density panel (Druet *et al.* 2010; Khatkar *et al.* 2012). Several authors (Berry & Kearney 2011; Hoze *et al.* 2013; Brondum *et al.* 2014) have reported improved imputation accuracy with the increase of the size of the reference population. Although it is the general trend that the more animals genotyped with the high density panel, the higher the imputation accuracy from 3k to 50k panels, it seems that the accuracy tends to reach a certain plateau after a specific proportion of population has been genotyped with the high density panel as reported by Wang *et al.* (2012). Thus, it does not seem that the reference population used in this study will have major impact of the results. However, it is very likely that the population structure and the genetic relatedness among animals in the reference and testing populations have impacted the results.

Finally, imputation accuracy is largely data dependent. Thus, trying to compare results across studies is difficult. Never the less general trends could be captured as a function of population size, level of messiness of genomic data, MAF distributions, and levels of LD between markers. In addition, accuracy reported as correct percentage cannot be compared across datasets with different MAF distributions.

Conclusions

Genotype imputation provides an attractive and cost effective tool for large scale implementation of GWAS and genomic selection in animal and plant applications. Even with low density panels with few thousand SNPs, high accuracies are possible, especially when the generational interval between reference and testing population is small. As the generational interval increases, the imputation accuracies decay, although not at an

alarming rate based on the results of this study. However, under different population structure the decay in imputation accuracy could be substantial. More importantly, a reasonable updating of the reference population will significantly reduce, or even eliminate, the decay in imputation accuracy over generations. At least currently, this tends to be the general practice in the industry.

References

Badke Y.M., Bates R.O., Ernst C.W., Fix J. & Steibel J.P. (2014) Accuracy of Estimation of Genomic Breeding Values in Pigs Using Low-Density Genotypes and Imputation. *G3: Genes|Genomes|Genetics* **4**, 623-31.

Berry D.P. & Kearney J.F. (2011) Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. *animal* **5**, 1162-9.

Brondum R., Guldbrandtsen B., Sahana G., Lund M. & Su G. (2014) Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. *BMC Genomics* **15**, 728.

Browning S.R. & Browning B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97.

Chen L., Li C., Sargolzaei M. & Schenkel F. (2014) Impact of Genotype Imputation on the Performance of GBLUP and Bayesian Methods for Genomic Prediction. *PLoS One* **9**, e101544.

Cheung C.Y., Thompson E.A. & Wijsman E.M. (2013) GIGI: an approach to effective imputation of dense genotypes on large pedigrees. *Am J Hum Genet* **92**, 504-16.

Druet T. & Georges M. (2010) A hidden markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. *Genetics* **184**, 789-98.

Druet T., Schrooten C. & de Roos A.P.W. (2010) Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *Journal of Dairy Science* **93**, 5443-54.

Hao K., Chudin E., McElwee J. & Schadt E.E. (2009) Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genetics* **10**, 27.

Hickey J., Kinghorn B., Tier B., van der Werf J. & Cleveland M. (2012) A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genetics Selection Evolution* **44**, 9.

Howie B.N., Donnelly P. & Marchini J. (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet* **5**, e1000529.

Hoze C., Fouilloux M.-N., Venot E., Guillaume F., Dassonneville R., Fritz S., Ducrocq V., Phocas F., Boichard D. & Croiseau P. (2013) High-density marker imputation accuracy in sixteen French cattle breeds. *Genet Sel Evol* **45**, 33.

Huang L., Wang C. & Rosenberg N.A. (2009) The relationship between imputation error and statistical power in genetic association studies in diverse populations. *Am J Hum Genet* **85**, 692-8.

Huang Y., Hickey J.M., Cleveland M.A. & Maltecca C. (2012) Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. *Genetics Selection Evolution* **44**, 25.

Johnston. J., Kistemaker. G. & Sullivan. P.G. (2011) Comparison of Different Imputation Methods. *INTERBULL BULLETIN*, 25-33.

Kapell D.N., Sorensen D., Su G., Janss L.L., Ashworth C.J. & Roehe R. (2012) Efficiency of genomic selection using Bayesian multi-marker models for traits selected to reflect a wide range of heritabilities and frequencies of detected quantitative traits loci in mice. *BMC Genetics* **13**, 42.

Khatkar M.S., Moser G., Hayes B.J. & Raadsma H.W. (2012) Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. *BMC Genomics* **13**, 538.

Li Y., Willer C.J., Ding J., Scheet P. & Abecasis G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology* **34**, 816-34.

Ma P., Brondum R.F., Zhang Q., Lund M.S. & Su G. (2013) Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. *J Dairy Sci* **96**, 4666-77.

Meuwissen T.H., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819-29.

Nothnagel M., Ellinghaus D., Schreiber S., Krawczak M. & Franke A. (2009) A comprehensive evaluation of SNP genotype imputation. *Human Genetics* **125**, 163-71.

Pryce J. & Hayes B. (2012) A review of how dairy farmers can use and profit from genomic technologies. *Animal Production Science* **52**, 180.

Pungpapong V., Muir W.M., Li X., Zhang D. & Zhang M. (2012) A Fast and Efficient Approach for Genomic Selection with High-Density Markers. *G3: Genes|Genomes|Genetics* **2**, 1179-84.

Sargolzaei M., Chesnais J. & Schenkel F. (2014) A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**, 478.

Sargolzaei M. & Schenkel F.S. (2009) QMSim: a large-scale genome simulator for livestock. *Bioinformatics* **25**, 680-1.

Sargolzaei M., Schenkel F.S., Jansen G.B. & Schaeffer L.R. (2008) Extent of Linkage Disequilibrium in Holstein Cattle in North America. *Journal of Dairy Science* **91**, 2106-17.

Sargolzaei. M., Chesnais. J.P. & Schenkel. F.S. (2011) FImpute: An efficient imputation algorithm for dairy cattle populations. *J. Dairy Sci*. **(E-Suppl. 1):421;(Abstr.)**.

Scheet P. & Stephens M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**, 629-44.

Scutari M., Mackay I. & Balding D. (2013) Improving the efficiency of genomic selection. *Statistical Applications in Genetics & Molecular Biology* **12**, 517-27.

VanRaden P.M. (2008) Efficient methods to compute genomic predictions. *J Dairy Sci* **91**, 4414-23.

VanRaden P.M., Null D.J., Sargolzaei M., Wiggans G.R., Tooker M.E., Cole J.B., Sonstegard T.S., Connor E.E., Winters M., van Kaam J.B.C.H.M., Valentini A.,

Van Doormaal B.J., Faust M.A. & Doak G.A. (2013) Genomic imputation and evaluation using high-density Holstein genotypes. *Journal of Dairy Science* **96**, 668-78.

VanRaden P.M., O'Connell J.R., Wiggans G.R. & Weigel K.A. (2011) Genomic evaluations with many more genotypes. *Genetics Selection Evolution* **43**, 11.

Wang H., Woodward B., Bauck S. & Rekaya R. (2012) Imputation of missing SNP genotypes using low density panels. *Livestock Science* **146**, 80-3.

Weigel K.A., de los Campos G., Gonzalez-Recio O., Naya H., Wu X.L., Long N., Rosa G.J. & Gianola D. (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J Dairy Sci* **92**, 5248-57.

Weigel K.A., de Los Campos G., Vazquez A.I., Rosa G.J., Gianola D. & Van Tassell C.P. (2010a) Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. *J Dairy Sci* **93**, 5423-35.

Weigel K.A., Van Tassell C.P., O'Connell J.R., VanRaden P.M. & Wiggans G.R. (2010b) Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. *J Dairy Sci* **93**, 2229-38.

Zhang Z. & Druet T. (2010) Marker imputation with low-density marker panels in Dutch Holstein cattle. *J Dairy Sci* **93**, 5487-94.

Table 3.1a. Parameters of the simulated population structure

Population structure	Simulation parameter
Step1:Historical population(HP)	
Size of historical populations (number of generations)	600(0) 600(200) 5000(205) 4000(210) 3000(220)
Number of males in the last HP generation	300
Step2:Recent (founder) population	
Selected males form historical population	300
Selected female form historical population	2700
Number of generations	100
Number of offspring per dam	1
Proportion of male progeny	0.50
Replacement ratio for males	0.50
Replacement ratio for females	0.20
Selection criteria	TBV
BV estimation method	BLUP animal model
Number of generations genotyped	9
Number of replicates	5
Overall heritability	0.4
OTL heritability	0.2
Phenotype variance	1.0

EBV: estimated breeding value; BV: breeding value; QTL: quantitative trait loci; TBV: True breeding value

Table 3.1b. Parameters of simulated genome

Genome structure	
Number of chromosomes	30
Chromosome length	100 cM
Number of markers (per chromosome)	1,400
Marker distribution	Evenly spaced
Marker allele frequency in the first HP	Equal
Number of QTL (per chromosome)	50
QTL distribution	Random
Additive allelic effects for QTL	Gamma distribution (shape = 0.40)
QTL allele frequency in the first HP	Equal

Table 3.2. Layout of the different simulation scenarios including the reference and testing populations.

	Chromosome	SNP chip	individuals
Reference Population			
Generation 1	30	42K	2700
Generation 2	30	42K	2700
Testing populations			
Generation 3 (S3)[1]	30	3K	2700
Generation 4 (S4)[1]	30	3K	2700
Generation 5 (S5)[1]	30	3K	2700
Generation 6 (S6)[1]	30	3K	2700
Generation 7 (S7)[1]	30	3K	2700
Generation 8 (S8)[1]	30	3K	2700
Generation 9 (S9)[1]	30	3K	2700

[1]S3-S9: Generations 3 to 9 were used separately as testing populations.

Table 3.3. Imputation accuracy from 3K to 42K SNP panels based on overall error rate (OER) for each of the 7 testing populations (S3-S9)[1]

	Imputation accuracy					
Testing population	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 4**	**Rep 5**	**Average**
Generation 3 (S3)	89.33	89.05	89.57	89.54	89.54	89.41
Generation 4 (S4)	89.00	88.70	89.23	89.25	89.23	89.08
Generation 5 (S5)	88.68	88.45	88.91	88.92	88.93	88.78
Generation 6 (S6)	88.42	88.19	88.72	88.69	88.63	88.53
Generation 7 (S7)	88.1	87.78	88.34	88.36	88.24	88.16
Generation 8 (S8)	87.74	87.40	87.95	87.97	87.90	87.79
Generation 9 (S9)	87.37	87.06	87.59	87.64	87.57	87.45

[1] results are based on 5 replicates

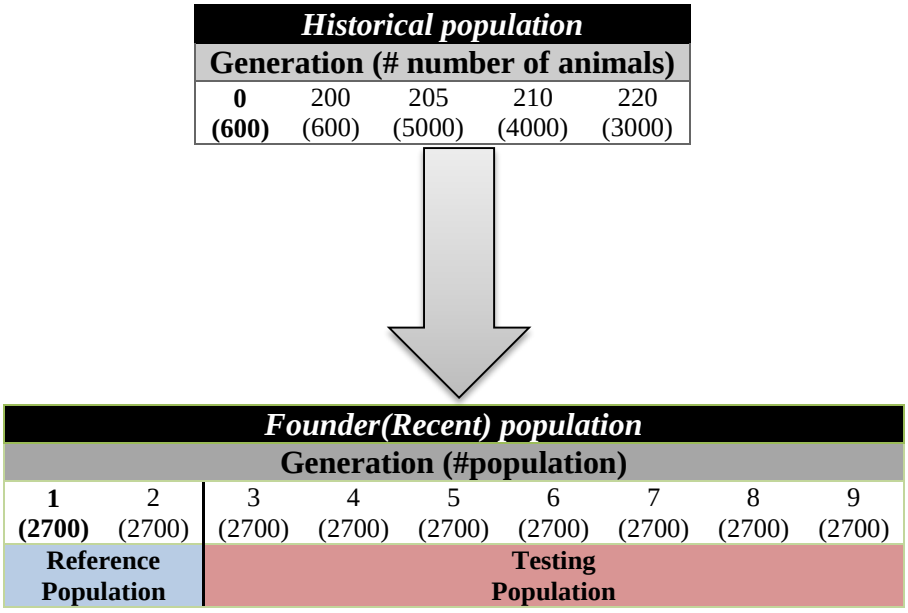

Figure 3.1. A schematic representation of the layout of the simulated populations

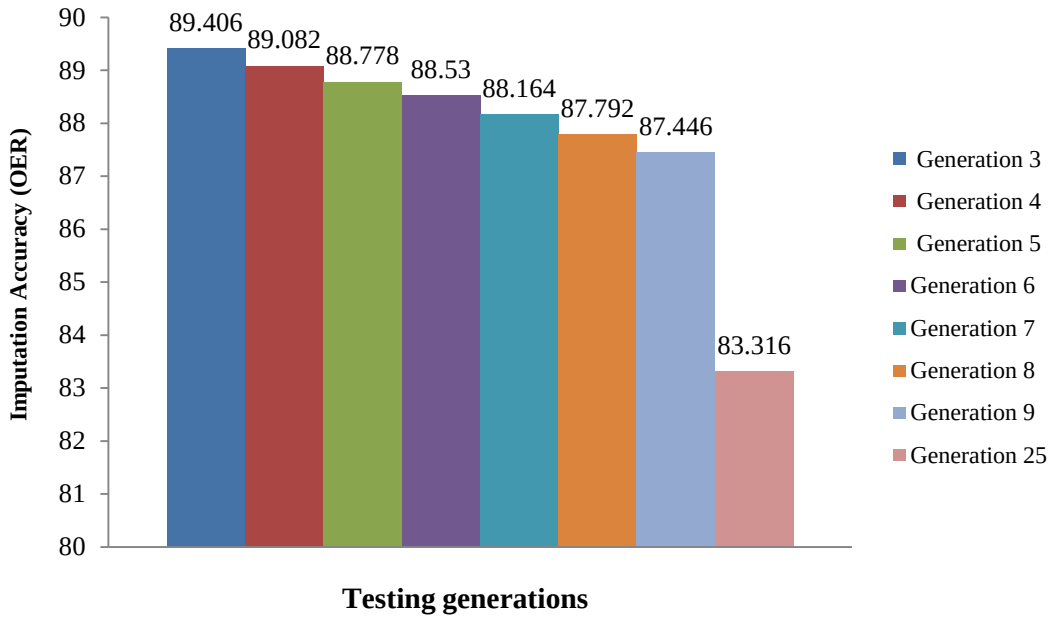

Figure 3.2. Imputation accuracy across generations using the overall rate of errors (ORE) criteria and fixed reference population (generations 1 and 2).

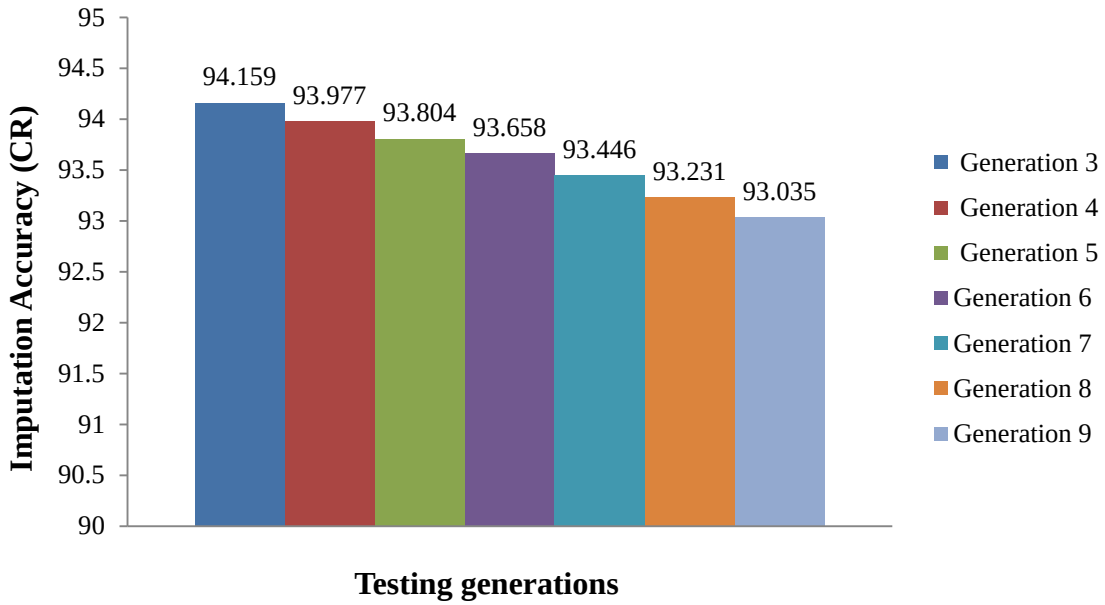

Figure 3.3. Imputation accuracy across generations using the concordance rate (CR) criteria and fixed reference population (generations 1 and 2).

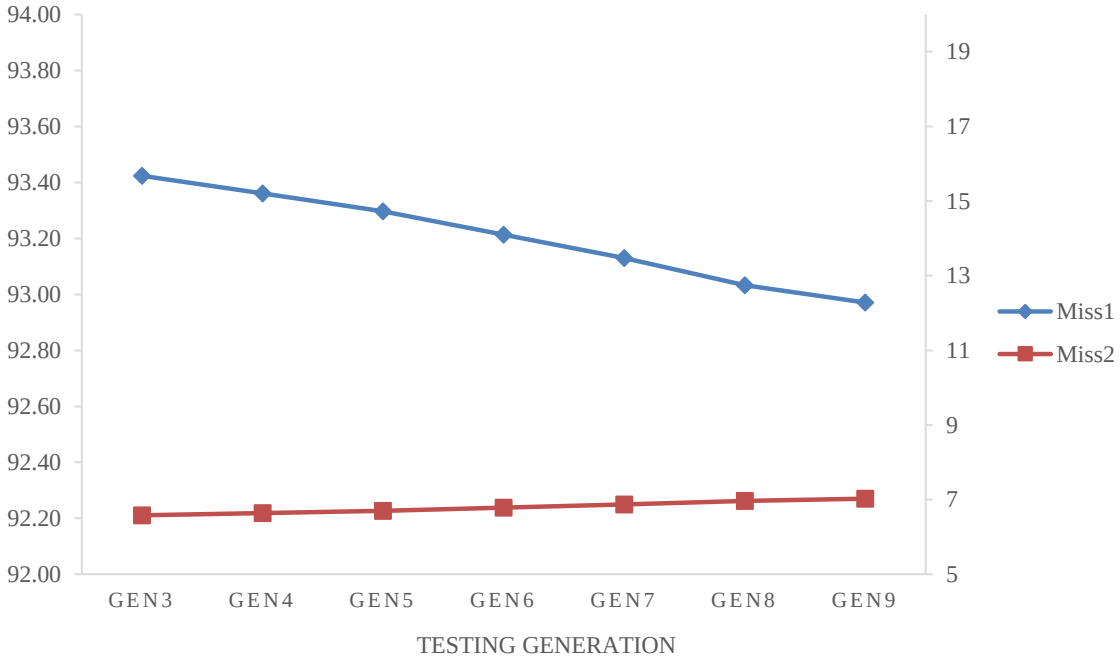

Figure 3.4. Percentage of one (Miss1) and two (Miss 2) allele errors across generations using fixed reference population (generations 1 and 2)

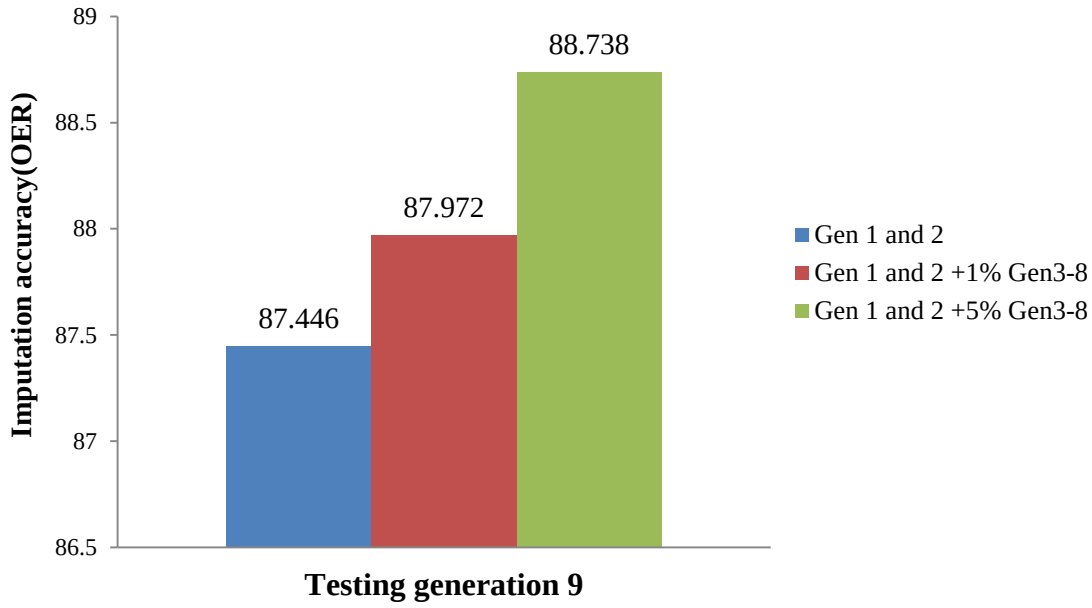

Figure 3.5. Imputation accuracy of overall rate of errors (ORE) in generation 9 using different reference populations

CHAPTER 4

ACCURACY OF GENOMIC SELECTION IN PRESENCE OF IMPUTED SNP GENOTYPES[2]

[2] Toghiani, S. and R. Rekaya. To be submitted to Animal.

Abstract

More accurate breeding values are obtained using genomic information. The superiority of genomic selection is possible only when high density SNP panels are used to track genes and QTLs affecting the trait. Unfortunately, even with the continuous decrease in genotyping costs, only a small fraction of the population has been genotyped with these high density panels. In order to reduce the cost of genomic selection, it is often the case that a larger portion of the population is genotyped with low-density and low-cost SNP panels and then imputed a higher density. Accuracy of SNP genotype imputation tends to be high when minimum requirements are met. Nevertheless, a certain rate of genotype imputation errors is unavoidable. Furthermore, such rate of errors tends to increase with the increase of the generational interval between reference and testing generations. Thus, it is reasonable to assume that the accuracy of GEBVs will be affected by the imputation errors; especially their cumulative effects over time. To evaluate the impact of multi-generational SNP genotypes imputation on the accuracy of GEBVs, a simulation was carried out under varying updating of the reference population, distance between training and validation sets, and the approach used for the estimation of GEBVs. In absence of updating of the reference population, accuracy of GEBVs decays substantially in one or two generations with a decrease rate of around 20-25% per generation. When the reference population is updated by 1 or 5% every generation, the decay in accuracy was only 8 to 11% for 7 generations using the true and imputed genotypes. These results indicate that imputed genotypes provide a viable alternative, even after several generations, as long the reference and training populations are appropriately updated to reflect the genetic change in the population.

Introduction

With the advance of new technologies it is now possible to efficiently genotype animals for thousands of single nucleotide polymorphisms (SNPs), generating high density markers maps. These high density maps can enable the identification of markers in population-wide disequilibrium with quantitative trait loci (QTLs). Using groups of markers, the effects of genomic regions can be estimated and combined to form genomic estimated breeding values (GEBV) as suggested by Meuwissen *et al.* (2001) . In livestock and poultry, major advances have been realized in the last few years, and genomic selection is becoming a routine technique, mainly due to the decreasing costs of genotyping for high density SNP markers panels. Furthermore, several simulation and real data based studies have shown that high accuracies for GEBV can be obtained in the absence of large numbers of progeny records (Meuwissen *et al.* 2001; Calus & Veerkamp 2007; Muir 2007). As work in developing commercial genotyping platforms continues, marker maps of increasing density will become available which in presence of appropriate size training data sets will further increase the accuracy of genomic selection.

The classical approach for estimation of breeding values is based on using the expected additive relationships between individuals to identify superior animals conditionally to the collected phenotypic data. As such, the expected additive relationship does not represent the true relationship between two individuals. Thus, there is no doubt that better estimates of breeding values could be obtained if realized, rather than expected, additive relationships are used in inferring the EBVs. High density genotyping for SNP markers provides such opportunity.

Using this genomic information in a breeding program requires the direct or indirect estimation of the effect of each SNP. Several analytical approaches have been proposed for genome-based prediction of genetic values. They main difference stems from the assumptions about the marker effects (Meuwissen *et al.* 2001; de Los Campos *et al.* 2009; Habier *et al.* 2011). Genomic selection is currently implemented either through a multiple regression (MR) or variance component (VC) based models. MR approach consists in a multiple step procedure where SNP effects are first estimated in a training population and then validated in separate data set. Several procedures including single marker analyses (Kim *et al.* 2009; Bohossian *et al.* 2014), ridge regression (Endelman 2011; Ogutu *et al.* 2012), non and semi parametric methods (Bennewitz *et al.* 2009; Perez-Rodriguez *et al.* 2012), and Bayesian approaches (Hayashi & Iwata 2013; Fernando *et al.* 2014) have been developed and used to implement the MR. Although these methods have different statistical and biological assumptions regarding the data generating process, they tend to yield similar results in the majority of the cases and differences are largely due to the genetic architecture of the trait, the genetic relationships between individuals in the sample, and the chosen prior information.

Accuracy of breeding values is 30 to 70% higher using genomic information. However, such superiority of genomic selection is possible only when a reasonable number of SNPs are genotyped for each individual. Unfortunately, high density SNP panels are costly which precludes their extensive use. Even with the continuous decrease in genotyping cost, it is likely that in the near future, marker genotype data will be sparse and only collected in limited number of animals.

A lot of effort has been placed in calculating genotype probabilities for non-typed animals conditionally on the genotype of typed animals. Iterative peeling, Markov chain Monte Carlo techniques, and regression on gene content (Gengler *et al.* 2007) are some of the common techniques being used with varying results depending on the complexity of the pedigree and the amount of available information. These genotype imputation methods have become a standard tool in the livestock industry in order to reduce genotyping costs.

In general, accuracies of SNP genotype imputation tend to be high when minimum requirements (reasonable size of reference population, numbers of SNPs in low and high density, genetic similarity between reference and testing populations) are met. Nevertheless, a certain rate of genotype imputation errors is unavoidable. Furthermore, such rate of errors tends to increase with the increase of the generational interval between reference and testing generations (Toghiani and Rekaya, 2014).

In presence of imputed SNP genotypes, it is reasonable to assume that the accuracy of GEBV will depend on the error rate of the imputation in addition to the method used to estimate marker effects (Luan *et al.* 2009), the heritability of the trait (Calus & Veerkamp 2007); and (Habier *et al.* 2007), the population structure (Hayes *et al.* 2009; Habier *et al.* 2010), and the size of the reference population (VanRaden *et al.* 2009). Furthermore, the impact of genotype imputation errors could be different between the different methods used in genomic selection (RR vs. VC).

The objective of this study is to evaluate the impact of multi-generational SNP genotypes imputation on the accuracy of GEBV obtained using either regression or variance based methods.

Material and Methods

Simulation:

Population structure: Genomic data were simulated using QMSim software (Sargolzaei & Schenkel 2009) and consisted of 42,000 SNP markers and 1500 QTLs distributed across a 30 Chromosomes genome as described by Toghiani and Rekaya (2014). The founder population (G0) consisted of 2,700 individuals generated through a random mating of 300 males and 2,700 females. Following the G0, 9 generations (1-9) were generated. It was assumed one progeny per dam per year, a sex ratio of 50% in the progeny, and selection was based on EBV with a replacement rate of 50% for sires and 20% for dams. Selected sires and dams were randomly mated. A single trait with an overall heritability of 0.40 and phenotypic variance equal to 1.0 was simulated. The true breeding value of an individual was set equal to the sum of the QTL additive effects and polygenic effects. The phenotypes were generated by adding random residuals to the true breeding values. A detailed description of the simulation process of the population structure is presented in Table 4.1.

Simulated genome: The simulated genome consisted of 30 pairs of chromosomes with 100 centi-Morgan (cM) in length each. Each chromosome harbored 1,400 SNP markers that were evenly distributed. Additionally, 50 randomly distributed QTLs were simulated per chromosome. Both SNP markers and QTLs were assumed to be bi-allelic, and no marker loci overlapped with the QTLs. Effects of QTLs were sampled from a gamma distribution with shape parameter equal to 0.4. Complete linkage disequilibrium (LD) was simulated between markers, between QTLs and between markers and QTLs in

the first historical population. The parameters used for simulating the genome are presented in Table 4.1.

Imputation of SNP genotypes: Simulated genomic data consisted of the SNP marker genotypes of individuals in the last 9 generations. It included 24,300 genotyped animals with a panel of 42,000 SNP markers (42K) that were distributed equally (2700 genotyped animals) in each generation. To investigate the effects of genotype imputation over generations on the accuracy of GEBVs estimation, the simulated data was divided into reference and testing populations. Three scenarios were evaluated: 1) reference population was assumed fixed and included only animals in generations 1 and 2; 2) reference population included animals in generations 1 and 2 and was updated with an additional 1% of top animals in following generations, and 3) reference population included animals in generations 1 and 2 and was updated with an additional 5% of top animals in following generations. SNP genotypes of testing populations (generations 3 to 9) were imputed from the low density panel (3K SNPs) to the high density panel (42K SNPs) using the FImpute (Sargolzaei *et al.* 2014) as described by Toghiani and Rekaya (2014).

*Estimation of GEBVs***:** In order to evaluate the effect of multigenerational imputation of SNP marker genotypes on the accuracy of genomic selection, GEBVs were estimated using true and imputed genotyped via regression based model (BayesA) and variance component based approach (GBLUP). For BayesA, the following model was used:

$$y_{ij} = \mu + \sum_{j=1}^{p} X_{ij} g_i + e_i \qquad [1]$$

where y_{ij} is the true BV of animal *i* in the reference (training) population, μ is the overall mean, X_{ik} is the genotype for animal *i* at locus k (k=1,2,..,p), g_k is the k^{th} SNP effect, and e_{ij} is the residual term.

In matrix notation, the model in [1] can be rewritten as:

$$\boldsymbol{y} = \mu\mathbf{1}_n + \boldsymbol{Xg} + e \qquad [2]$$

where **y** is the vector of observations , **g** is the vector of SNP effects; **X** is the matrix of SNP genotypes, **e** is the vector of residual terms, and *n* is the number of animals in the training population. A hierarchical Bayesian implementation was adopted. The first stage of the hierarchy consisted of the conditional distribution of the data given the parameters of the model and it was assumed to be normal:

$$\mathbf{y} \mid \mu, \mathbf{X}, \mathbf{g}, \sigma_e^2 \sim N(\mu\mathbf{1}_n + \mathbf{Xg}, \mathbf{I}_n\sigma_e^2)$$

In the second stage, the following priors will be assumed for the model parameters

$$\mu \sim cte$$
$$g_i \sim N(0, \sigma_i^2) \text{ for } i = (1,2,...,p)$$
$$\sigma_e^2 \sim \chi^{-2}(\upsilon_0, s_0^2)$$

Finally, prior distributions were specified for the variance of the SNP effects. A conjugate prior was assumed. Thus,

$$\sigma_i^2 \sim \chi^{-2}(\upsilon_i, s_i^2)$$

The hyper-parameters $\upsilon_0, \upsilon_i, s_0^2, and\ s_i^2$ were set equal to 2, 0.6, 4, and 1.41e^{-5}, respectively. The implementation of the proposed hierarchical model is straightforward as all conditional distributions are in closed form being normal for the position parameters and scaled inverted Chi square for the dispersion components.

When GBLUP was used, the following model was implemented to estimate GEBVs.

$$\boldsymbol{y} = \mu\mathbf{1}_n + \boldsymbol{Zu} + \boldsymbol{e} \qquad [3]$$

where $\boldsymbol{u}$ is the vector of genomic breeding values and $\boldsymbol{Z}$ is a known incidence matrix with the appropriate dimensions that relates the phenotype to the breeding values. Everything else is as defined before. It is worth mentioning that the vector $\boldsymbol{u}$ includes training and validation animals. Further, it was assumed that:

$$\mathbf{y} \mid \mu, \mathbf{u}, \sigma_e^2 \sim N(\mu\mathbf{1}_n + \mathbf{Zu}, \mathbf{I}_n\sigma_e^2)$$

and,

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}\sigma_u^2)$$
$$\sigma_e^2 \sim \chi^{-2}(\upsilon_0, s_0^2)$$
$$\sigma_u^2 \sim \chi^{-2}(\upsilon_u, s_u^2)$$

The hyper-parameters $\upsilon_0, \upsilon_u, s_0^2, and\ s_u^2$ were set equal to 2, 0.6, 2, and 0.4, respectively. G is the matrix of genomic (realized) additive relationship between animals in the training and validation sets. It was computed based on the observed SNP genotypes (coded 0, 1 and 2) using the following formulae (VanRaden 2008).

$$\boldsymbol{G} = \boldsymbol{W}'\boldsymbol{W}/[2\sum_{i=1}^{p} p_i\,(1-p_i)]$$

where $\boldsymbol{W}$ is the matrix $\boldsymbol{Z}$ of observed SNP genotypes adjusted by the minor allele frequencies.

The implementation of the model in equation [3] is identical to the classical mixed linear model, except that the average additive relationship matrix ($\boldsymbol{A}$) is replaced by $\boldsymbol{G}$. For both models, a full Bayesian implementation was adopted. A single chain of 100,000 iterations was implemented where the first 20,000 rounds were discarded as burn-in period. Using GBLUP, the GEBVs were directly obtained after solving the system of equations. Using BayesA, the estimated GEBV for each animal was computed as the sum (over all SNPs) of the product between each SNP effects and its associated genotypes

$$GEBV_i = \sum_{k=1}^{p} x_{ik}\,\hat{g}_k$$

Accuracy was calculated based on the correlation between true BVs and GEBVs and was averaged over 5 replicates.

Results and Discussions

Figure 4.1 presents the accuracies of estimated genomic breeding values of the validation animals when the training population was fixed (generations 1 and 2) using the true and imputed SNP genotypes implemented via the BayesA method. In all cases, the accuracy was, as expected, higher using the true SNP genotypes. Such superiority ranged from 12% for generation 3 to 15% for generation 7. However, using either the true or the imputed genotypes, the accuracy decayed with the increase of the generational interval

between the training and validation populations. In fact, the accuracy dropped from 0.60 to 0.26 and 0.52 to 0.22 between generations 3 and 9 using the true and imputed SNP genotypes, respectively. These results indicate that the decrease of accuracy is mainly due to the increase in genetic dissimilarity (distance) between training and validation sets, rather than the accuracy of the SNP genotypes. However, it is worth mentioning that in this study, the imputation accuracies (Table 4.2) were in the lower range of reported results of imputation performance. When the training population consisted of generations 1 and 2 and was updated with either 1% or 5% of top animals in following generations, the accuracy of estimated genomic breeding values in the 9^{th} generation was improved substantially (Figure 4.2). In fact, with 1% updating of the training population, the accuracy increased to 0.46 and 0.41 from 0.26 and 0.22 using the true and imputed genotypes, respectively. It increased to 0.54 and 0.46 when the training population was updated with an additional 5% of top animals in generations 3 to 8. These results indicate that the decay in accuracy was only 8% and 11% between generations 3 and 9 when a 5% updating scheme is adopted using the true and imputed genotypes, respectively. The small difference (8 vs 11%) again indicates that with reasonable imputation accuracies, true and imputed genotypes perform similarly, as long the training population is appropriately updated to reflect the genetic change in the population. Although these results show the importance of the size of the training population on the accuracies of GEBV as it has been shown in several other studies (Meuwissen *et al.* 2001; Goddard 2009), they highlight more importantly the effect of the genetic architecture of the trait and the LD between markers and QTLs. LD is primarily due to selection and recent drift rather than historic mutations and consequently the accuracy of unrelated individuals tend to be low. This will be the case

if selected candidates descend from a population having an LD structure that is different from that in the training data.

Comparing both estimation methods, it does not seem to be major difference in accuracy using BayesA or GBLUP approach (Figure 4.3). In fact, using either the true or imputed SNP genotypes, the accuracies are basically the same between the two methods. Theoretically one could postulate the possibility of difference between the two methods in their sensitivity of errors in the imputed genotypes due the manner in which they estimate SNP effects. However, that was not supported at least by the results of the present study.

Conclusions

Imputed SNP genotypes are being used routine by different segments of livestock industry in the implementation of genomic selections. Although several specific variations are used in each case, it seems that even when imputation is conducted base on low density panel of few thousand SNPs, the accuracy of imputation tends to be acceptable as long as the reference population is of sufficient size. More importantly, the use of these imputed genotypes in genomic selection will have little to no effects on the accuracy of estimated GEBVs as long a reasonable updating of the training population is implemented. This result holds true even after a number of selection rounds.

References

Bennewitz J., Solberg T. & Meuwissen T. (2009) Genomic breeding value estimation using nonparametric additive regression models. *Genetics Selection Evolution* **41**, 1-12.

Bohossian N., Saad M., Legarra A. & Martinez M. (2014) Single-marker and multi-marker mixed models for polygenic score analysis in family-based data. *BMC Proceedings* **8**, S63.

Calus M.P. & Veerkamp R.F. (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. *J Anim Breed Genet* **124**, 362-8.

de Los Campos G., Gianola D. & Rosa G.J. (2009) Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *J Anim Sci* **87**, 1883-7.

Endelman J.B. (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *Plant Gen.* **4**, 250-5.

Fernando R.L., Dekkers J.C. & Garrick D.J. (2014) A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genet Sel Evol* **46**, 50.

Gengler N., Mayeres P. & Szydlowski M. (2007) A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *animal* **1**, 21-8.

Goddard M. (2009) Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245-57.

Habier D., Fernando R.L. & Dekkers J.C.M. (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389-97.

Habier D., Fernando R.L., Kizilkaya K. & Garrick D.J. (2011) Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics* **12**, 186.

Habier D., Tetens J., Seefried F.R., Lichtner P. & Thaller G. (2010) The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genetics Selection Evolution* **42**, 5.

Hayashi T. & Iwata H. (2013) A Bayesian method and its variational approximation for prediction of genomic breeding values in multiple traits. *BMC Bioinformatics* **14**, 34.

Hayes B.J., Visscher P.M. & Goddard M.E. (2009) Increased accuracy of artificial selection by using the realized relationship matrix. *Genet Res (Camb)* **91**, 47-60.

Kim S., Sohn K.-A. & Xing E.P. (2009) A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformatics* **25**, i204-i12.

Luan T., Woolliams J.A., Lien S., Kent M., Svendsen M. & Meuwissen T.H. (2009) The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. *Genetics* **183**, 1119-26.

Meuwissen T.H., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819-29.

Muir W.M. (2007) Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *J Anim Breed Genet* **124**, 342-55.

Ogutu J., Schulz-Streeck T. & Piepho H.-P. (2012) Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC Proceedings* **6**, S10.

Perez-Rodriguez P., Gianola D., Gonzalez-Camacho J.M., Crossa J., Manes Y. & Dreisigacker S. (2012) Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3 (Bethesda)* **2**, 1595-605.

Sargolzaei M., Chesnais J. & Schenkel F. (2014) A new approach for efficient genotype imputation using information from relatives. *BMC Genomics* **15**, 478.

Sargolzaei M. & Schenkel F.S. (2009) QMSim: a large-scale genome simulator for livestock. *Bioinformatics* **25**, 680-1.

VanRaden P.M. (2008) Efficient methods to compute genomic predictions. *J Dairy Sci* **91**, 4414-23.

VanRaden P.M., Van Tassell C.P., Wiggans G.R., Sonstegard T.S., Schnabel R.D., Taylor J.F. & Schenkel F.S. (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci* **92**, 16-24.

Table 4.1. Simulation parameters of population structure and genomic data

Population structure	Information about simulation
Step1:Historical population (HP)	
Size of historical generation (number of generations)	600(0) 600(200) 5000(205) 4000(210) 3000(220)
Number of males in the last generation of HP	300
Step2:Recent (founder) population	
Selecting base male form historical population	300
Selecting base female form historical population	2700
Number of generations genotyped	9
Number of replicates	5
Overall heritability	0.4
Phenotype variance	1.0
Genomic structure	
Number of chromosomes	30
Chromosome length	100 cM
Number of markers (per chromosome)	1400
Marker distribution	Evenly spaced
Number of QTL(per chromosome)	50
QTL distribution	Random

EBV: estimated breeding value; BV: breeding value; QTL: quantitative trait loci; TBV: True breeding value;

Table 4.2. Imputation accuracy from 3K to 42K SNP panels for each of the 7 testing populations (S3-S9)[1]

	Imputation accuracy					
Testing population	**Rep 1**	**Rep 2**	**Rep 3**	**Rep 4**	**Rep 5**	**Average**
Generation 3 (S3)	89.33	89.05	89.57	89.54	89.54	89.41
Generation 4 (S4)	89.00	88.70	89.23	89.25	89.23	89.08
Generation 5 (S5)	88.68	88.45	88.91	88.92	88.93	88.78
Generation 6 (S6)	88.42	88.19	88.72	88.69	88.63	88.53
Generation 7 (S7)	88.1	87.78	88.34	88.36	88.24	88.16
Generation 8 (S8)	87.74	87.40	87.95	87.97	87.90	87.79
Generation 9 (S9)	87.37	87.06	87.59	87.64	87.57	87.45

[1] results are based on 5 replicates

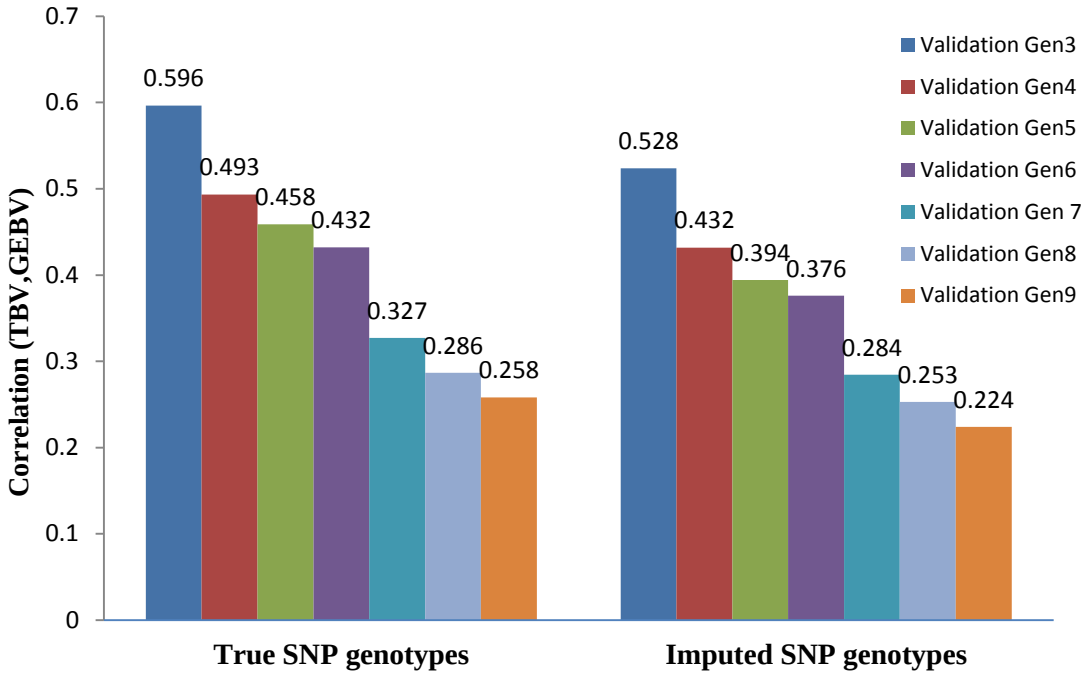

Figure 4.1. Accuracy of GEBVs with true and imputed SNP genotyped based on a fixed reference population and using BayesA method.

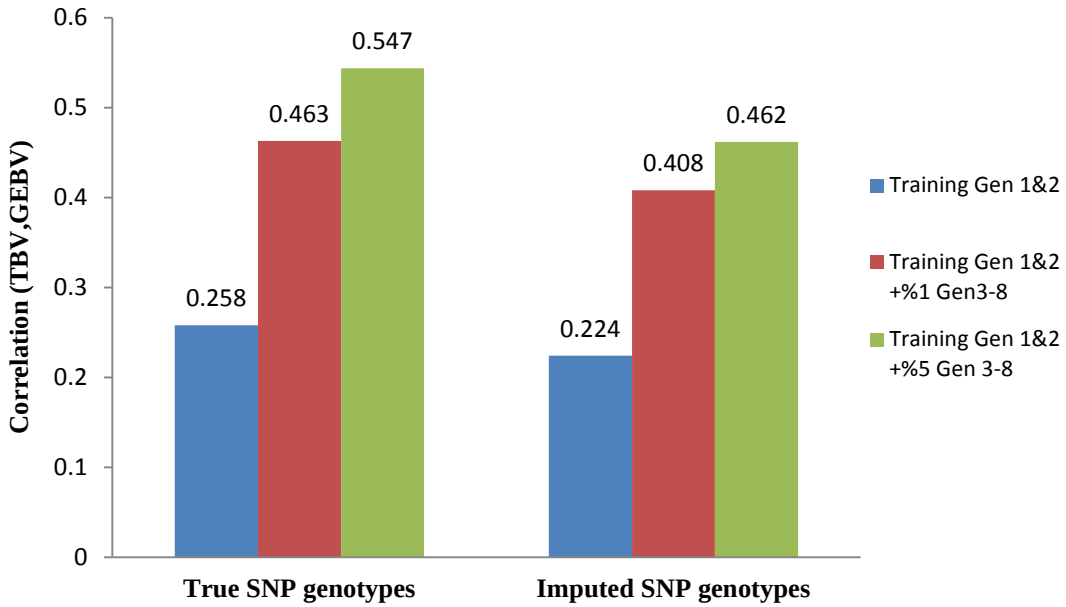

Figure 4.2. Accuracy of GEBV in generation 9 using true and imputed SNP genotypes and different training populations

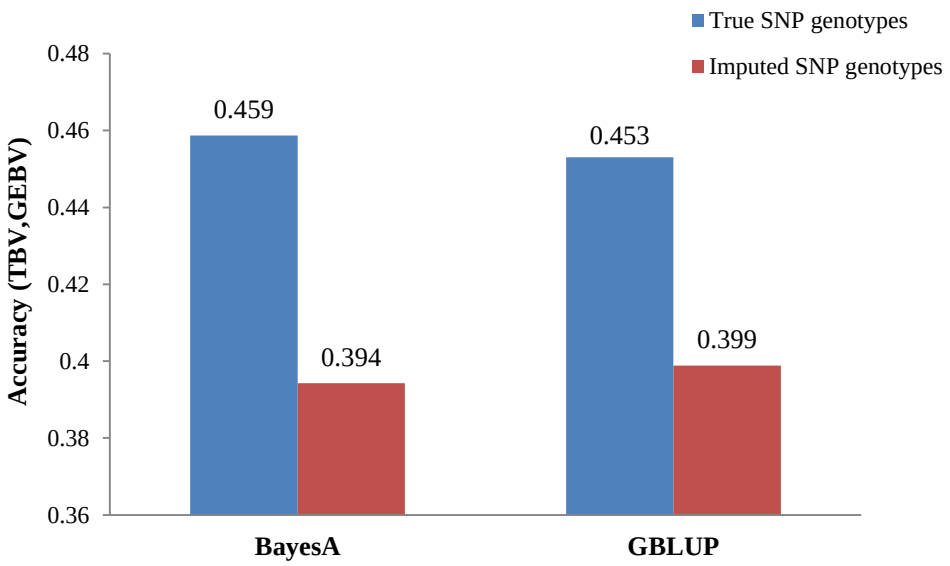

Figure 4.3. Accuracy of GEBVs in generation 5 with the true and imputed SNP genotypes using BayesA and GBLUP methods

CHAPTER 5

CONCLUSIONS

Genomic selection is fast becoming a standard tool in several livestock improvement programs. Its superior to classical breeding values estimation methods is unquestionable. However, these performances are possible only when a certain set of conditions are met. One of these conditions is the availability of dense SNP marker maps. In spite of the continuous decrease in genotyping cost, it remains too expensive for large scale use in several segments of the livestock industry. To balance these two conflicting requirements, imputation of un-typed SNP genotypes in low-density and low-cost SNP panels has become an acceptable option. Genotype imputation provides an attractive and cost effective tool for large scale implementation of GWAS and genomic selection in animal and plant applications. Even with low density panels with few thousand SNPs, high accuracies are possible, especially when the generational interval between reference and testing population is small. As the generational interval increases, the imputation accuracies decay, although not at an alarming rate based on the results of this study. However, under different population structure, the decay in imputation accuracy could be substantial. More importantly, a reasonable updating of the reference population will significantly reduce, or even eliminate, the decay in imputation accuracy over generations.

More importantly, the use of these imputed genotypes in genomic selection will have little to no effects on the accuracy of estimated GEBVs as long a reasonable updating of the training population is implemented. This result holds true even after several round

of selection. These results validate to large extend current industry practice of genomic selection. However, this will hold true only with a comprehensive updating of reference and training populations with new high density genotyped animals.