Population based Genotype Imputation

by

Yining Wang

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

University of Alberta

Abstract

In this dissertation, I focus on the study of genotype imputation in population data. Genotype imputation is a process of inferring missing values for genotype data and has been extended to predicting "untyped" genotypes for samples in low-density chips with a reference population assayed using dense marker chips. It has been successfully and routinely applied to merge genotype datasets of different densities that arise from various genotyping and sequencing platforms. First, I examine and compare several influential imputation models that incorporate biological concepts, mine for associations among genetic markers and explore genetic relatedness. I further evaluate the effect of imputation on genomic prediction, which combines dense marker data with phenotypic data for improving quantitative traits. Additionally we propose a multi-step strategy that can work with any existing genotype imputation methods to boost the accuracy of imputation from low-density chips to high-density chips. Finally we describe a new hidden Markov model for genotype imputation based on an existing framework.

Preface

This thesis is an original work by Yining Wang under the supervision of Dr. Guohui Lin and Dr. Paul Stothard.

Chapter 2 and Chapter 3 of this thesis has been published as Y. Wang, G. Lin, C. Li and P. Stothard, "Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle," Springer Science Reviews, vol. 4, issue 2, pp 79-98. I was responsible for the the literature review in chapter 2, and the design of experiments and analysis as well as the manuscript composition. C. Li granted permission for the use of the data and C. Li, G. Lin and P. Stothard contributed to manuscript edits. Chapter 4 of this thesis has been published as Y. Wang, T. Wylie, P. Stothard, and G. Lin, "Whole genome SNP genotype piecemeal imputation," BMC Bioinformatics, 16:340, 2015. I carried out the computational experiments and was responsible for the manuscript composition of the methods section. T. Wylie participated in the experiments, the clustering part in particular. G. Lin and P. Stothard conceived of the study, and participated in its design and coordination and helped to draft and edit the manuscript. All authors read and approved the final manuscript.

To my parents

For their encouragement, love and support throughput my studies.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisors, Dr. Guohui Lin and Dr. Paul Stothard for their guidance and support throughout my PhD study. I am also very grateful to Dr. Changxi Li who has provided me with helpful comments, advice and kind suggestions on portions of the thesis work. Next, I would like to thank all dissertation committee members for their valuable time and effort to read the dissertation.

I benefited tremendously from many discussions with friends and colleagues whom I would like to thank: Dr. Xiaoping Liao, Dr. Tim Wylie, Dr. Liuhong Chen and Dr. Chunyan Zhang. I would also like to thank Steve Sutphen and all the kind staff at the CS Help Desk for their help, quick responses, and regular followups in troubleshooting technical problems and providing technical support.

Last but not the least, this thesis would not exist without the love and support of my family. My deepest thanks go to my mom and dad for their continuing encouragement, endless love and unwavering support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Genetics Background

The genome, organized in chromosomes, is made up of deoxyribonucleic acid (DNA) molecules composed of *bases* or *nucleotides*. There are four possible nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T). Each nucleotide can pair up with a complementary nucleotide in a double-stranded DNA molecule, A with T and C with G, to form *base pairs*. For diploid species such as cattle and humans, chromosomes come in pairs. The bovine genome contains 3×10^9 base pairs across 28 pairs of autosomes and two sex chromosomes, X and Y. Most of the nucleotides along the genome are identical between individuals of the same species. For example, human beings are 99.9 percent the same in their DNA makeup [35]. We refer to a position along a chromosome as a *locus* (plural, loci). A locus where variation occurs is said to be polymorphic. The alternative forms of sequence that occur at a polymorphic site in the genome are called alleles. Genetic variations that involve single nucleotides are called *single-nucleotide polymorphisms* (SNPs). Most SNPs are *biallelic*, meaning that there are only two alleles observed at that locus as opposed to all four possible forms (from the four types of nucleotides). In this dissertation, we deal with biallelic SNPs, which are considered to be common forms of genetic variation. We refer to the most common allele and the second most common allele at a locus in a given population as “major allele” and “minor allele” respectively. For any individual, the pair of alleles at a locus is referred to as the genotype at that locus. If the two alleles are identical, then the individual is known as “homozygous”; otherwise, the individual is “heterozygous”. A genotype at a locus does not specify which allele comes from which one of the two chromosomes. Thus, a genotype at a locus can be denoted as an unordered pair of two alleles, and the genotype of a pair of homologous chromosomes can be viewed as a sequence of unordered pairs of SNP alleles. In case of bi-allelic SNPs, the two alleles at that locus are sometimes coded as “0” and “1” respectively. As a result, the genotype at that locus can be represented as 0 for “00”, 1 for “01”, and 2 for “11”. If the allele “0” is the major allele at a locus in a given population comprised of N individuals and let $c(00)$, and $c(01)$, and $c(11)$ denote the counts of the three genotypes at the bi-allelic locus, then “minor allele frequency” (abbreviated as “MAF”), which is defined as the frequency of the less frequent allele “1”, can be calculated as $f(1) = c(11)/N + \frac{1}{2}c(01)/N$.

The Mendelian law of segregation states that, for each diploid individual, one of a pair of homologous autosomes is inherited from the paternal side and the other from the maternal side. Moreover, a child does not inherit a complete parental chromosome from each parent, as recombination (or crossover) occurs. That is, during the meiosis process that produces gametes, portions of DNA are exchanged between the two homologous chromosomes present in each of the parents. As a consequence, a child inherits a mosaic pattern of their parents’ chromosomes. Contrary to the definition of genotype, we refer to an ordered sequence of alleles that were inherited together along a chromosome as a *haplotype*. The parental origin of the alleles of genotypes across segments of the genome is not directly observable but can be inferred using pedigree information, which records biological relationships of individuals in a population. When genotypes at several loci are heterozygous, several feasible pairs of haplotypes can be formed given existing genotypes. For example, when k heterozygous genotypes are observed in an individual’s genotype, there are 2^{k-1} feasible pairs of haplotypes that could have produced the genotype if no additional information is provided. We are interested in inferring haplotypes (termed “phasing” or “haplotyping”) using SNP genotype data obtained from several individuals in a population, and a process that exploits correlated (termed “linked”) loci that tend to be inherited together. Those linked loci are known to be in “linkage disequilibrium" (LD), which is defined as the non-random association of alleles at different loci. Mathematically, let A_1 and A_2 be the two alleles at locus 1 and let B_1 and B_2 be the two alleles at locus 2. We would like to know whether the haplotype frequency of "A_1B_1" (denoted $f(A_1B_1)$) is equal to the product of the allele frequencies of A_1 and B_1 at the two loci. If $f(A_1B_1) = f(A_1) \cdot f(B_1)$, then it implies that two loci are independently inherited and also known as "linkage equilibrium"; otherwise, the given two loci are in LD. The success of population-based genotype imputation relies on mining LD patterns from the genotype data and uncovering haplotypes for each individual.

Currently, haplotype phasing can be determined through laboratory based methods such as sperm typing; however, laboratory methods are prohibitively expensive in large scale for tens of thousands of individuals. Computational methods are proposed as cost-efficient alternatives to laboratory methods. Phasing is closely related to genotype imputation because haplotyping identifies blocks of closely "linked" loci and due to correlation, one can fill in missing alleles of genotypes with the observed alleles at linked loci. Recombination is a key factor that leads to decay of LD between alleles at different loci. Mutation, which refers to alteration of an allele at any locus, is another event that introduces genetic variation into population. Although it is rare, mutation leads to a child inheriting an allele that is different from her parents. A mutation can be inherited if it is not harmful to cell viability and eventually may become what is termed a common polymorphism if it occurs in more than 1 percent of the population. The mutation rate is usually thought to be low for SNP per generation and is estimated to range between 10^{-8} and 10^{-9} [38]. Closely related individuals tend to share long segments of haplotypes, whereas distantly related individuals many generations apart share much shorter segments. The probability of the occurrence of a recombination event is not uniform across the genome, and varies region by region. Some regions with increased recombination rates are more likely to harbour recombination events and are known to be recombination hotspots, while others may have little or no recombination. The reconstruction of high-density genetic map through studies in both families and at the population level using genotype and sequencing datasets reveals a mapping between physical loci and recombination rates [54, 58]. With a genetic map, one can look up the recombination rates between any two loci in a population. The distance between two loci in a genetic map is measured in centimorgans (cM) where 1 cM is defined as 1% chance of observing a recombination in a single generation.

1.2 Genotypic Data Overview

The Human Genome Project has been successful in accelerating discoveries of human health related genetic variants and disease genes. The same strategies and technologies used in human genomics have been applied to livestock animals for uncovering important genetic variations and conducting genetic analyses [106]. In bovine genomics, the 1000 Bull Genomes Project (`http://www.1000bullgenomes.com/`) identified 28.3 million genetic variants including 26.7 million single nucleotide polymorphisms (SNPs) [23]. These dense SNPs that exhibit variations in regions along the whole genome have become a valuable tool for parental verification [71], for identification of potential disease risk genes [70] and for subsequent genome-wide selection (GWS) studies and genomic selection (GS) in the aim of improving genetic gains [14, 74].

Both Illumina (`https://www.illumina.com`) and Affymetrix (`http://www.affymetrix.com`) offer general purpose commercial SNP chips for genotyping. For example, the BovineSNP50 BeadChip (Bovine50K; Illumina Inc., San Diego, USA), a medium density SNP chip containing $54,609$ SNPs, has been successfully applied in dairy cattle for estimating breeding values [14, 46]. The high-density bovine SNP chips, the Illumina BovineHD BeadChip ("Illumina 770K") containing more than $777,000$ SNPs and the Affymetrix Axiom Genome-Wide BOS 1 Bovine Array containing more than $640,000$ SNPs ("Affymetrix 640K"), are available for accurate genetic merit evaluations and comprehensive genome wide association studies. Although SNP genotyping enjoys a lower typing error rate due to their bi-allelic nature, denser genomic coverage, lowering cost, and standardization among laboratories [72, 14], the price of genotyping remains a major challenge for large number of candidate animals to be typed for genomic selection, not to mention the more expensive genome sequencing. A commercially available "BovineLD Genotyping BeadChip" of 6,909 SNPs ("Illumina 6K"; Illumina Inc., San Diego, USA) has been developed as a cost-effective low-density alternative to the Illumina 50K with selected markers optimized for imputation [5] and was reported to contain lower genotyping errors than its LD predecessor the Illumina Golden Gate Bovine3K chip. Also, the Illumina 6K chip can be customized by adding SNPs. Previously, genotype imputation mostly refers to inferring the sporadic missing genotypes in the assays and now the term has been extended to the scenario in which we would like to infer untyped SNPs that are not directly assayed in a study sample of individuals genotyped in a low-density chip by use of a high-density SNP genotype dataset as a reference panel [68].

With the development of high-throughput DNA genotyping chips of various densities and the advance of sequencing technologies [92, 41, 70, 5], numerous genetic variants have become available for use in livestock improvement. Genomic prediction (GS), which combines high-density genotypic and phenotypic data, has become a new tool in the selection of above-average candidates that have better breeding values for traits of interest as parents of the next generation [34, 26]. Compared to traditional evaluations, which solely rely on phenotypic and/or pedigree information to extrapolate relatedness and identity-by-descent between animals, GS has revolutionized animal breeding by increasing the accuracy of estimates of genetic merit and shortening the generational interval. Various statistical approaches have been proposed for GS and differ in their assumptions of marker effects. For example, the genomic best linear unbiased prediction (GBLUP) model [36] assumes all markers contribute to genetic variance of the trait. On the other hand, some Bayesian alphabet methods including BayesB adopt a Bayesian inference framework for parameter estimation and assume that the trait is influenced by only a fraction of all markers, while others have no effect [74]. Genotype imputation traditionally is a procedure of inferring the small percentage of sporadic missing genotypes in the assays, but it now commonly refers to the process of using a reference population genotyped at a higher density to predict untyped genotypes that are not directly assayed for a study sample genotyped at a lower density [68]. Genotype imputation is expected to boost the statistical power because it equates the number of SNPs for datasets genotyped using different chips and leads to an increased number of SNPs in association studies, which in turn should result in higher persistence of linkage phase between quantitative trait loci (QTL) and SNPs, and potentially increase the accuracy of genomic predictions. Additionally, dense SNP markers will more likely contain some causative SNP markers, which can increase the statistical power for genome wide association studies.

1.3 Motivation and Definition of Genotype Imputation

Imputation is a well-studied statistical problem and the success of imputation depends on the missing value mechanism. That is, we would like to know under what circumstances, if any, our inference of missing values would lead to valid answers as if the data set were fully observed. One of the most stringent assumption one can make about missing values is called “missing completely at random” (MCAR) [63, 86]. This is equivalently saying that the probability of an observation being missing is independent of observed or unobserved variables. SNP genotypes missing because of random failures of laboratory samples can be considered MCAR. For example, Yu and Schaid [107] evaluated a total of eight imputation methods for imputing missing values within SNP genotype data from the HapMap project [22] under the assumption that missing genotypes were MCAR. A more general assumption under which imputation analyses can be done with observed data is called “missing at random” (MAR). That is, whether or not a value is missing is independent of unobserved missing variables and values although it can depend on observed data (such as allele intensities and neighbouring SNPs). The MAR hypothesis is considered to be realistic and reasonable if important predictors of the SNP with missing values are included in the imputation model [63, 37]. When the missingness depends on unobserved data, missing data are said to be “not missing at random” (NMAR) [86, 63].

Genotype calling programs, which convert raw instrument data into genotypes for downstream analyses, typically use a clustering algorithm to assign the genotype (“00”, “01” and “11”) to each SNP in each individual. The clustering algorithm is applied on a per-SNP basis to multiple samples and seeks to assign each sample to one of the three genotype classes, based on the intensity of signals corresponding to the two alleles generated by the instrument for each sample. If the clustering algorithm is unable to find an appropriate genotype class for a sample, then a missing outcome would be assigned to him at the SNP locus [37]. Both genome wide association studies (GWAS) for QTL fine-mapping of complex traits and genomic selection (GS) for livestock improvement require high-density genotypes from a large number of individuals, which are apt to contain a certain small percentage of missing values (termed “sporadic”) ranging from 0.05% to 5%. The missing mechanism for the small percentage of “sporadic” genotypes that have not been called is usually assumed to be MAR. Additionally, GWAS and GS tools usually assume no presence of missing values in genotype data. The most common approach for dealing with missing data is to remove samples with many missing values and/or loci with a large percentage of missing SNPs. Addressing missing values using this approach leads to reduced sample size and consequently power to detect QTL. Computationally inferring those sporadic missing values, otherwise known as imputation, is an alternative to re-genotyping or re-sequencing samples containing missing values and has advantages of saving both labour and cost. Also, genotype imputation can be used as a strategy to increase the coverage and resolution of SNPs beyond the original chips up to a reference panel of dense SNP markers. Thirdly, missing genotypes arise when we try to combine data from samples genotyped using different SNP arrays. It is not uncommon to have samples from different labs genotyped using different chips with only a small percentage of SNPs in common across the chips. Imputation methods are routinely applied to infer “untyped” SNPs that are unique in one chip and the missing mechanism in this scenario is usually assumed to be NMAR.

The population based genotype imputation can be formally defined as follows: given a reference panel of known, unrelated, and unphased high-density genotype data DG, our goal is to impute the untyped markers that are not directly assayed in a genetically similar data set SG, termed a "study sample," genotyped on a low-density chip. Strictly speaking, individuals are "related" to some degree in that even two distantly related individuals can be traced back to a common ancestor if we follow genealogy into the past. To clarify the context of "unrelatedness," we imagine that unrelated individuals are independent, identically distributed observations drawn from a population and they are not recently related, not related via close family relationships in a pedigree [42]. We use SG_{ij} to denote the genotype of study individual i at marker j, where SG_{ij} can be 0, 1, or 2 representing the number of copies of the alternative allele if observed and $SG_{ij} =?$ if untyped. Likewise, DG_{ij} denotes the genotype of individual i at marker j on the reference panel $\mathcal{R}$. DG and SG share an overlapping set of markers, denoted $\mathcal{T}$, representing the set of *"typed"* markers in both low-density and high-density chips. Assume that all markers of the two datasets are bi-allelic and they fall into two disjoint subsets: an overlapping set $\mathcal{T}$ of markers, typed in both the low-density study sample and high-density reference panel, and a set $\mathcal{U}$ of markers that are typed only in DG but untyped in SG. Information gain of imputation from low density chips to high density chips comes not only from linkage disequilibrium of SNPs in low-density study sample but mainly from haplotype information of reference panel. When unphased reference data is used, genotype imputation algorithms need to phase individuals for samples in DG to obtain a set of haplotypes and the quality of haplotypes has an impact on the imputation of the study sample.

In addressing this problem, I investigated several existing statistical models and proposed a new statistical hidden Markov model (HMM) based on Li and Stephens' "Product of Approximate Conditionals" (PAC) framework [61] and the clustering approach. The running time of HMM-based genotype imputation grows quadratically with the number of hidden states at each locus, representing available haplotypes. As a result, when the number of individuals in the reference panel becomes large, PAC-based HMM becomes slow; additionally, existing programs based on PAC use heuristic approaches to reduce the number of individuals at each locus for speeding up the sampling progress.

The key idea of the existing genotype imputation methods is to explore and hunt for shared “identical by descend” (IBD) haplotypes that exhibit high LD from a high-density reference panel of genotypes or haplotypes over a region of tightly linked markers and use them to fill untyped SNPs of any low-density study samples. The success of genotype imputation depends on the length of correlated markers in LD blocks. Markers common to both study samples and reference panels serve as anchors for guiding genotype imputation approaches imputing any unobserved haplotypes within the LD block. Because of domestication, selection and breeding in cattle, Matukumalli et al. [70] reported that the length of LD blocks of correlated markers in cattle is about three times greater than that of human populations. In human populations, substantial efforts have been made to produce accurately phased “haplotype” reference panels, available from the International HapMap project (International HapMap Consortium, 2005) [22] and the 1000 Genomes Project [50]. Yet, in cattle and many other livestock species, “unphased” SNPs from sequencing or in HD genotyping chips and medium-density genotyping chips are commonly used as reference panels for imputation.

1.4 Comparative Studies of Imputation Methods and Their Effects on Genomic Predictions

The discovery of millions of SNPs from genome sequencing and dramatic reduction in the cost of genotyping have enabled the adoption of a form of genomic selection known as genomic selection (GS) [74] as a popular tool for selecting breeding animals from populations of candidates [34]. Genetic improvement aims to select above-average candidates as parents of the next generation and to produce progenies with performance above-average of the current generation. In livestock species, performance of candidates is largely determined by complex traits, which are quantitative in nature and are likely controlled by many genes along with the influence of environmental factors [100, 26, 73]. For quantitative traits, a single locus only accounts for a limited proportion of the total genetic variance, most genes that contribute to the complex traits (quantitative trait loci or QTL) are still unknown and detected QTLs only explain a small fraction of the total genetic variance [73, 26]. As a result, previous marker assisted selections achieved limited successes for traits controlled by a few major genes. The theory underlying the genomic selection methods is that genetic effects must exist somewhere along the genome for any trait with a non-zero heritability and the effects of QTL are expected to be in LD with some SNP markers, although these common SNPs are unlikely to be causal variants for functional genetic differences [100, 74]. GS fits all SNPs covering the whole-genome in a linear prediction equation, estimates the effects of the SNP markers simultaneously and thus potentially captures all the genetic variance explained by these SNPs. GS proceeds in two steps. In Step 1, dense SNP markers divide the entire genome into smaller chromosome segments and GS estimates the effect of these segments in a training population in which each animal has both genotype and real-valued phenotype records. In Step 2, GS tries to calculate genomic estimated breeding values (GEBV) for selection candidates that are not in the training population and have only genotypes but no phenotypes, which can be obtained by combining their genotypes with the estimated effects (from step 1) of the segments they carry:

$$\text{GEBV}(\vec{\mathbf{X}}) = \sum_{j=1}^{m} \hat{\beta}_j \mathbf{X}_j,$$

where m is the total number of SNP markers across the entire genome, $\vec{\mathbf{X}} = (\mathbf{X}_1, \cdots, \mathbf{X}_m)$ is a vector consisted of coded genotypes (as the counts of allele "1") for an individual, and $\hat{\beta}_j$ is the estimated effect of the genotype at locus j from Step 1. It follows from Step 1 that once the marker effects are estimated they can be re-used for many selection candidates in a few generations and from Step 2 that GS enables evaluation and selection of candidates without phenotypic information and for traits that are expensive or difficult to measure as long as candidates are genotyped. The other benefits of GS include increasing genetic value prediction accuracies of selection by including DNA markers as additional information and shortening the generation interval. GS enables us to select animals before they are of productive and/or reproductive age, select female candidates on male traits and vice versa. The success of GS requires dense SNP markers because density of SNPs affects the LD between the causative QTLs and SNPs. In reality due to cost constraints usually only low-density or medium-density genotypes are available; therefore genotype imputation can be used as an approach to convert animals genotyped in low-density chips up to the high-density. In the comparative studies, I investigate to what extent the accuracy of imputation affects genomic breeding values in GS for a beef cattle trait.

1.5 Piecemeal Imputation

An important subproblem in genotype imputation is how to improve accuracies with existing imputation softwares. Although existing genotype imputation programs are largely successful in imputing untyped markers for individuals genotyped in low density chips, they are not perfect. Each imputation program has different assumptions in its model setting and sometimes sacrifices accuracies for efficiency in running time. Experimental results from several previous studies on bovine genotypes [97, 60, 55, 47] show that two-step imputation using programs such as Beagle and Impute 2 from a low density panel (e.g. Illumina 6K) to an intermediate density panel (e.g. Bovine 50K) and then to high density panel (e.g. Bovine 770K) yielded higher accuracies than a direct one-step imputation from the low density panel to the high density panel. One possible reason why there is such increase in imputation accuracy is related to obtaining accurate phasing of haplotypes in DG. An explanation offered by van Binsbergen *et al.* [97] is that imputation algorithm has problems with selecting the correct haplotypes since there are multiple possible matches between HD and LD panels, whereas there are fewer possible matches when an intermediate genotype chips is introduced in between. Our piecemeal strategy tries to evaluate the effect of each untyped marker on the accuracy of imputation along the genome by creating a pseudo intermediate panel with all the markers in the low-density chip plus the untyped one, and two-step imputation with any existing method goes from the low-density panel to the pseudo-intermediate panel, then to the high-density panel. A feature vector that incorporates the accuracy at each untyped locus from the two-step procedure is used for clustering intermediate pseudo-panels. Intermediate panels that fall into the same cluster have similar performance in affecting accuracies of markers along the genome. For each cluster, once we identify regions along the genome where two-step procedures via the associated intermediate panels unanimously increase the accuracies, it suffices to use one of the intermediate panel within each cluster for future two-step procedure to impute the identified "regions". Final results can be obtained via piecing together partial results from the two-step imputation and one-step imputation.

1.6 Overview

This dissertation is organized into six chapters. Chapter 2 reviews some important models in the development of genotype imputation. In Chapter 3, I evaluate the performance of several existing genotype imputation methods using beef cattle data genotypes and examined the effects of imputed results in genomic predictions. In Chapter 4, I investigate a step-wise strategy that can work with any existing genotype imputation program for boosting the accuracies of genotype imputation. In Chapter 5, I present a statistical model derived from an existing framework for modelling linkage disequilibrium and genotype imputation. In Chapter 6, I summarize my contributions and discuss some possible directions for future work.

The work in Chapter 2 and Chapter 3 was published in Springer Science Reviews as "Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle" [102]. The work in Chapter 4 was published in BMC bioinformatics as "Whole genome SNP genotype piecemeal imputation" [103]. Chapter 5 contains a result we have made great efforts in, but turns out not completely successful.

Chapter 2

Related Work

2.1 Imputation Models and Popular Methods

In this section, we review the most widely used computational models underlying several population-based genotype imputation methods. An overview of population based genotype imputation is given in Figure 2.1. Existing methods for genotype imputation can be categorized computationally into the linear regression model by Yu and Schaid [107], clustering models [90, 10, 91, 40], hidden Markov models and expectation-maximization (EM) algorithms [28]. More recent works have included "BLIMP" by Wen and Stephens [105] based on "Kriging" for imputation from summary data and "Mendel-Impute" via matrix completion [17]. Alternatively, imputation methods can be divided into two broad categories: the aforementioned "population-based" imputation methods that use LD information and the "family-based" imputation methods that use both pedigree and LD information such as rule-based AlphaImpute [49] and sampling-based GIGI [16]. In general, family-based imputation programs using Mendelian segregation rules and LD information result in higher accuracies than population-based ones for rare variants because pedigrees record patterns of relationship among individuals and performance of population-based imputation programs can be weakened by low LD of distant SNPs in sparse low-density chips [49, 16, 80, 87]. Our review focuses on population-based programs that do not require pedigree information because of the following three reasons. First, pedigree information is not always available for reasons of privacy or missing pedigree records. Second, population-based methods yield more accurate imputation for common variants than family-based imputation [16]. Thirdly, some family-based programs require availability of dense genotypes for all immediate ancestors [49].

There have been several excellent reviews on genotype imputation methods and applications to human genome wide association studies [9, 45, 68] as well as related reviews on haplotyping methods [11]. Several studies have investigated the performance of imputation methods in the context of livestock applications [13] and evaluated their effects on genomic predictions [76, 13, 79]. In this review, we attempt to survey and categorize various historical and more recent population-based genotype imputation methods that accept unphased reference panels as input and then evaluate effects of imputed data on feed efficiency genomic predictions for beef cattle. We focus on the most important population-based imputation methods that have been widely adopted and relevant to both human and bovine genomics and their underlying computational schemes for parameter estimations, including Beagle [10], the "PAC" model of Li and Stephens [61] and its variants [90], and a simple rule-based method called FImpute [88] inspired by "long range phasing" [59].

All existing population based genotype imputation methods, in essence, try to find matches of similar haplotypes over a short chromosomal region between the study sample and the reference panel [51]. That is, the population-based genotype imputation methods pool information from typed markers that are in linkage disequilibrium with the untyped markers , and due to correlation, untyped markers $\mathcal{U}$ in the study sample SG can be filled with observed genotypes from the reference genotype DG if there is a match at typed markers $\mathcal{T}$ [69, 105]. Most methods not only perform genotype imputation for the study sample but infer their haplotype phases as well [69, 51, 10].

2.1.1 The "Product of Approximate Conditionals" (PAC) Model

The statistical model of Li and Stephens [61] for population patterns of linkage disequilibrium (LD) and identification of recombination hotspots is a milestone in the development of genotype imputation methods, and a number of methods including Impute 1 [69], Impute 2 [51], MaCH [62] and fastPHASE [90] are all variants based

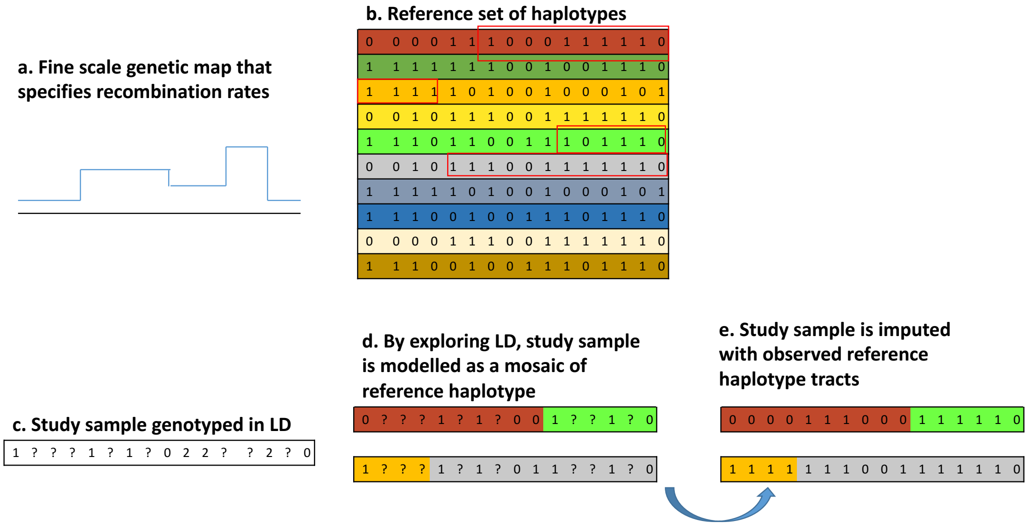


Figure 2.1: An illustration of how population based genotype imputation works for a study sample genotyped in low-density chip. A fine scaled genetic map (part a) is available for looking up how likely recombinations occur between two loci on the population level. The reference data (part b) consist of a set of haplotypes over a set of dense SNP markers derived from phasing algorithms. When a study sample comes in, its genotype is compared to the dense haplotypes in the reference panel (part b). Tracts of haplotypes in the reference from which the study sample copies are identified (coloured in red rectangles). Missing genotypes in the study sample are then imputed using those matching haplotypes in the reference panel (part e).

on this idea. Li and Stephens proposed "the product of approximate conditionals" (PAC) model for approximating coalescence with recombination and mutation in a population. Given n sampled diploid individuals at L markers, there are in total 2^L possible haplotypes. Due to the fact that recombination and mutation are both rare events, instead of considering the exponential number of haplotypes 2^L, one can narrow down the search list of candidate haplotypes and approximate a new haplotype as an imperfect mosaic of the observed n haplotypes, which represent the hidden states of a hidden Markov model (HMM). The "PAC" model approximates the recombination event as a Markov jump process along the genome: the new haplotype can copy from different haplotypes at two consecutive loci. Incorporation of recombination rates into the HMM significantly simplifies the transition probabilities and allows for transition from one marker to the next independent of the current hidden state from which the new haplotype copies. There is a chance that an allele of the new haplotype is close to but not exactly the same as the one from which it copies, reflecting that a mutation or a genotyping error occurs [61].

Discrete HMM Models – Impute 1, Impute 2 and MaCH

Impute 1 [69], Impute 2 [51] and MaCH [62] can be grouped together as they treat the observed genotypes as discrete counts of alleles and adopt a sampling scheme for estimating the posterior probabilities of missing genotypes in SG in a hidden Markov framework.

Impute 1 [69] assumes the availability of a high-density haplotype reference panel (denoted DH, which can be thought of as a "*phased*" version of DG), a fine-scale recombination map ρ that defines the probability of recombination occurring between two consecutive loci, an effective population size parameter N_e that is a scaling factor for genetic distance between two consecutive loci. It defines $P(SG_i|DH, \rho, \lambda)$ in the HMM framework of Li and Stephens [61], where λ is the mutation rate dependent on the number of individuals in the reference panel $\mathcal{R}$.

$$P(SG_i|DH, \rho, \lambda) = \sum_{Z_i} P(SG_i|Z_i, \lambda)P(Z_i|DH, \rho), \forall i = 1, \cdots, n$$

The hidden state $Z_{im} = \{k_1, k_2\}$ where $k_1 \in \mathcal{Z}$ and $k_2 \in \mathcal{Z}$ at each marker m

is an unordered pair of haplotypes in the reference panel from which two alleles of SG_i receive the copies, and therefore the number of hidden states is quadratic in the number of the haplotypes in $\mathcal{R}$. Posterior probabilities of untyped or missing genotypes SG_{im} are expressed via the forward-backward algorithm, and are estimated in a sampling process, and computation grows linearly in the number of markers and quadratically in number of haplotypes [69].

MaCH [62] further extends Impute 1's discrete-valued HMM model to the usage of the reference panel $\mathcal{R}$ containing unphased genotypes DG. Phasing in DG is obtained from a Monte Carlo Gibbs sampling precedure $P(DG_i|DG_{-i}, \rho, \lambda)$ and only a few rounds of updates are needed to obtain accurate consensus haplotype templates through empirical experiments [62]. The detailed path-sampling procedure of the HMM can be found in Appendix B of Scheet and Stephens [90]. The phasing procedure takes $\mathcal{O}(N^3)$ if all individuals in DG are used since each update needs to sample a path from N^2 hidden states and the number of updates grows linearly in N. The cubic running time for phasing becomes an issue when thousands of individuals are present in DG. To make MaCH scalable to large number of individuals in DG, Li *et al.* suggested using a randomly selected subset of DG for sampling phases of DG_i at a small cost of accuracy.

Impute 2 [51, 50] is considered as a major improvement over Impute 1 and is flexible with either "phased" or "unphased" reference panels. The major contribution of Impute 2 is a general strategy for HMM-based genotype imputation: first to resolve phasing in DG and SG then to impute alleles in haplotypes of SG. Computation is allocated more to the phasing step, as the accuracy of phased haplotype is key in obtaining accurate imputed alleles in $\mathcal{U}$ of SG. Impute 2 adopts MaCH's "Markov chain Monte Carlo" sampling strategy for phasing with modifications as follows:

- it initializes a set of haplotypes that are consistent with each individual of DG and SG respectively;
- it iteratively updates phasing in DG_i conditional on k "closest" haplotypes to obtain DH from $P(DG_i|DH_{-i}, \rho, \lambda)$;

- it iteratively updates phasing in SG_i at $\mathcal{T}$ typed markers conditional on "phased" DH and current guess of the rest of individuals from $P(SG_i^{\mathcal{T}}|SH_{-i}^{\mathcal{T}}, DH^{\mathcal{T}\cup\mathcal{U}}, \rho, \lambda)$;
- it imputes alleles at $\mathcal{U}$ untyped markers for $SH_{i,1}$ and $SH_{i,2}$ from $P(SH_{i,d}|DH, \rho, \lambda)$ via the forward-backward algorithm, where $SH_{i,1}$ and $SH_{i,2}$ are the two phased haplotypes that make up SG_i

Unlike MaCH, the phasing routine in Impute 2 is conditional on k closest haplotypes, which is determined by hamming distance to the current individual and computation burden of phasing grows quadratically with k closest neighbours $\mathcal{O}(N \cdot k^2)$ and increases linearly in the number of markers $\mathcal{O}(L)$. As phasing is resolved in the previous step, imputation step becomes haploid imputation and computation is linear in the number of individuals in DG and the number of markers L.

Continuous Local Cluster-Based HMM Models – fastPHASE and Bimbam

fastPHASE [90] is another HMM-based method that can estimate phasing and impute sporadic missing genotypes. The model is based upon the observation that haplotypes over tightly-linked regions tend to cluster into groups of similar patterns [90]. Each unobserved cluster can be viewed as a common haplotype from which underlying haplotype of genotype data originates. The transition probabilities in the HMM are modelled as a Markov jump process related to recombination events independent of the current state; however, the emission probabilities are no longer related to the mutation rate and but captured with regard to the real-valued "allele frequencies" of each cluster. The total number of clusters K is a parameter specified by users. We regard the underpinning HMM of fastPHASE as *continuous* in that at every marker m, each cluster is associated with a real-valued "relative frequency" α_{km} and a real-valued "allele frequency" θ_{km} of allele 1 with the constraints $\sum_{k=1}^{K} \alpha_{km} = 1$ and $\theta_{km} \in [0, 1]$. Structure 2.0 [31], a software developed for inference of population structure, was very similar to fastPHASE's local cluster HMM model, assumed that each cluster represents a sub-population, and used computationally-expensive Markov chain Monte Carlo sampling for parameter estimations.

Unlike its predecessors that employ MCMC for phasing and imputation, fastPHASE speeds up the process of estimating parameters via a maximum likelihood (ML) approach. An "expectation-maximization" (EM) algorithm is used for finding ML estimates of all parameters. It should be noted that Kimmel and Shamir [56] formalized a similar HMM model ("HINT") for disease association studies and proved that the genotype optimization problem is neither convex or conclave, and their exact form of maximization for updating θ_{km} does not exist. In HINT, Kimmel and Shamir [56] propose to update θ_{km} via a grid search in the neighbourhoods of 0 and 1 at the maximization step of the EM. In fastPHASE, Scheet and Stephens give a formula for approximating maximal θ_{km}, which updates the current value with the value at the previous step in the maximization step.

To obtain better parameter estimates, these authors suggested one set $K = 20$, run EM multiple times and take the average of estimates to overcome local maxima issues. The computational time is in $\mathcal{O}(n \cdot L \cdot K^2)$, which increases linearly in the number of individuals n in the dataset and number of markers L and quadratically in the number of clusters K. Missing genotypes are imputed by choosing the value that maximizes $P(G_i|\alpha, \theta, \rho)$.

The model was not originally designed for imputation with reference panels and special care must be taken to ensure the maximum likelihood approach does not yield higher error rate [68, 40]. When applying fastPHASE for imputation with a reference panel DG, Guan and Stephens [40] suggested using parameter estimates obtained from maximizing the likelihood for DG only, $P(DG|\theta, \alpha, \rho)$, rather than the full likelihood function $P(SG, DG|\theta, \alpha, \rho)$ as they believed inclusion of SG in the model fit for parameter estimation would reduce the number of clusters available to model DG.

The idea of fastPHASE has been incorporated into Bimbam [91, 40], a software for Bayesian imputation-based association mapping. Guan [39] extended fastPHASE's idea into a two-layered HMM for inference of population structure and local ancestry, and proposed an alternative to approximating and updating θ_{km} in EM by solving a linear system at the cost of $\mathcal{O}(K^3)$.

2.1.2 BLIMP

Following the suggestion by Guan and Stephens [40] on fitting the cluster-based HMM to only DG for estimating parameters and looking into the EM step, if we treat homozygous genotypes as known alleles and heterozygous genotypes as missing allele, we can further simplify the genotype-based Bimbam [91], derive EM updates for the haplotype-based Bimbam (all clusters collapse into identical ones) and obtain a much simplified linear model. Update for θ_{km} is only dependent on the frequencies of typed alleles – the summary level data mentioned by Wen and Stephens [105]. Wen and Stephens [105] developed a linear model called "BLIMP" based on Kriging by incorporation of recombination rate between two loci in the linear model. BLIMP requires as input a genetic map for information of recombination rates and is capable of not only untyped SNP loci frequency inference but individual level imputation as well. Imputation accuracy with BLIMP that uses summary data was comparable to that obtained from the current best available method Impute 2 [105].

2.1.3 Beagle 3.3.2 and Beagle 4.1

Beagle 3.3.2 is based on a flexible "localized haplotype-cluster" model [8] that groups locally similar haplotypes into clusters [10]. Beagle 3.3.2 is capable of imputing untyped genotypes, phasing haplotypes and handling multi-allelic markers. It allows users to incorporate the pedigree information as an option, and supports family-based genotype imputation. The underlying model of Beagle is an HMM that does not explicitly model recombination and mutation events, but adapts to data for local clusters at each marker and transitions [10]. The HMM of Beagle is a directed acyclic graph that has variable number of hidden states at each marker, representing local clusters as nodes. Each cluster only emits one possible allele. Also, Beagle allows at most two transitions coming out of each cluster. Compared with the HMMs based on the "PAC" model, which have fixed number of hidden states at each marker, Beagle has few number of hidden states (clusters) and transitions, which speeds up computations. Beagle achieves fewer number of hidden states (clusters) and transition through a pruning procedure. The pruning procedure detects the length of IBD segments shared among individuals by examing haplotype frequencies at each nodes Nodes at each level of Beagle's graph that are IBD are merged and combined. The other notable difference between Beagle's model and the "PAC" model lies in how they use haplotype information among individuals. Unlike Bimbam that only uses information from reference dataset in the model fit, Beagle 3.3.2 pools observed haplotypes from all individuals at each marker. The algorithm starts with randomly phasing genotypes and imputing missing values of individuals. An iterative EM-style update is repeated in subsequent steps for re-estimating phases and re-inferring missing values from current sampling of phasing information.

Browning and Browning (2013) [7] further improved the IBD detection algorithm (termed "Refined IBD") in Beagle in a two-step manner. In the first step, a linear time algorithm "GERMLINE" by Gusev *et al.* is used to find candidate sharing IBD segments. In the second step, Beagle uses a probabilistic approach to refine the candidate IBD segment to get consensus haplotypes. Such changes have been reflected in the latest version (4.0) of Beagle. Switch error rate is a commonly used metric for assessing the accuracy of inferred phases of haplotypes and is defined as the ratio of the count of possible switches from an inferred haplotype phase to obtain the true haplotype to the total number of of heterozygote loci in the individual's genotype minus one. O'Connell *et al.* [77] reported in their studies that the phasing results from Beagle 3.3.2 tended to have a much larger number of switch errors than SHAPEIT [27].

2.1.4 FImpute

FImpute [88] is an efficient, rule-based, and deterministic method for phasing and genotype imputation inspired by the idea of "long range phasing" [59]. Kong *et al.* [59] reasoned that the length of shared haplotypes reflects the degree of relatedness between two individuals. The closer two individuals are, the longer their shared haplotype is [88]. The algorithm first resolves phasing for homozygous genotypes of each individual, treats heterozygous genotypes as missing or wild card, and builds up a library of haplotypes with frequencies. Next, the algorithm iteratively looks for perfect or near perfect ($> 99\%$) matches at currently phased markers using an "overlapping sliding windows" from the maximum length of whole genome to the minimum of 2 SNPs, i.e. from close relatives to distant relatives. If a match is found, FImpute infers phasing for heterozygous genotypes, merges similar haplotypes in the library and updates their frequencies accordingly. If more than one match is found, FImpute uses match with higher frequencies for imputation and phasing. It imputes the remaining genotypes by random sampling of alleles based on observed frequencies.

Chapter 3

Genotype Imputation Methods and Their Effects on Feed Efficiency Genomic Predictions for Beef Cattle using 50K SNP Genotypes

The objective of this chapter is to present experimental results from our comparative studies of different population-based imputation methods and to investigate for two training scenarios the effects of imputation results on subsequent genomic predictions. An earlier version of the chapter was published in Springer Science Reviews [102]. In livestock, one of the direct applications of imputation results is in genomic predictions (GS) for genetic improvements of economic traits. GS has revolutionized animal breeding by accelerating the selection process of candidates of genetic superiority. It requires availability of dense SNP markers with known phenotypes for a reference population of selection candidates and exploits LD between SNPs that cover the entire genome and QTLs. Despite the continuing reduced costs, genotyping a large number of animals in high-density chips is still costly and sometimes not feasible. A combination of low, medium and high density chips can be used for lowering the cost of genotyping and imputation has become a common practice to bring animals genotyped with low-density chips up to the medium and high densities. In this study, we are concerned with beef cattle and a phenotype trait called “residual feed intake” because feed is a major expense for cattle producers and seeking genetic improvement of traits associated with the feed efficiency of a beef cattle is of great economic importance [82].

3.1 Methods

3.1.1 Genomic Predictions

We introduce two popular genomic prediction methods for predicting the genomic estimated breeding values (GEBV) in validation dataset. including an efficient GBLUP with a genomic relationship matrix $\mathbf{G}$ (VanRaden [99]) and a Bayesian method (BayesB [74]). Genomic prediction methods assume availability of a training dataset that have both the genotype data and the associated phenotypic records for each individual and tries to compute the genomic estimated breeding values (GEBV) for individuals that only have genotype data available in validation dataset.

The BayesB model proposed by Meuwissen *et al.* [74] fits all SNP effects simultaneously and assumes the following linear model

$$y_i = \mu + \sum_{j=1}^{p} \beta_j \mathbf{X}_{ij} + e_i, i = 1 \cdots n$$

where y_i is the adjusted RFI for animal i, μ is the overall mean, β_j is the regression coefficient (allele substitution effect) on the jth SNP, $\mathbf{X}_{ij}$ is the jth SNP genotype of animal i defined above, and e_i is the random residual effect for animal i, which is drawn from a normal distribution $\mathcal{N}(0, \sigma_e^2)$ and variance σ_e^2 is drawn from a scaled inverse chi-squared distribution with the degrees of freedom ν_e set to 10 and the scale parameter S_e^2 set to $\sigma_e^2(\nu_e - 2)/\nu_e$. The regression coefficient β_j has probability π to be exactly 0 (indicating no effect for the marker), denoted as $\delta(0)$, and probability $(1 - \pi)$ to be drawn from the normal distribution $\mathcal{N}(0, \sigma_j^2)$. That is, a mixture of a normal distribution and point mass at zero was used in the BayesB for β_j as shown below

$$\beta_j | \sigma_j^2 \sim \pi\delta(0) + (1 - \pi)\mathcal{N}(0, \sigma_j^2)$$

where π is our prior knowledge of the proportion of SNP that has no effects on the trait. The value of π is specified as an input by the user and the locus specific variance σ_j^2 is the unknown and is estimated from the data. Again, the prior for σ_j is assumed to be from a scaled chi-squared distribution with the degrees of freedom ν_j set to 4 and the scale S_j^2 set to $(\nu_j - 2)\sigma_a^2/\nu_j(1 - \pi)\sum 2p_j(1 - p_j)$, where σ_a^2 is the additive genetic variance component calculated by the phenotypic variance (after adjustment for fixed effects) on the training data, multiplied by heritability (h^2), and p_j and $(1 - p_j)$ are the two allele frequencies at SNP j. The unknowns including the regression coefficient β_j and its associated locus-specific variance σ_j^2 were estimated via a Markov chain Monte Carlo (MCMC) sampler. SNP effects were estimated by averaging all the samples after the burn-in period. The GEBV for animal i was predicted by adding up SNP effects over all loci: $\text{GEBV}_i = \sum_{j=1}^{m} \beta_j \mathbf{X}_{ij}$, where m is the total number of SNPs.

The GBLUP method (VanRaden [99]) assumes a linear model that uses a genomic relationship matrix $\mathbf{G}$ derived from the SNP dataset $\mathbf{X}_{n\times m}$ for estimating genomic breeding values (GEBV). The linear model can be written as

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{a} + \mathbf{e}$$

where $\mathbf{y}$ is the vector of adjusted RFI, μ is the overall mean, $\mathbf{a}$ is the vector of breeding values, $\mathbf{Z}$ is the incidence matrix relating $\mathbf{a}$ to $\mathbf{y}$, and $\mathbf{e}$ is the vector of random residuals. $\mathbf{G}$ measures genomic similarity between each pair of individuals based on allele frequencies. Let $\mathbf{p} \in \mathbb{R}^m$ be a vector whose ith component (denoted p_i) is the frequency of allele A at locus i. Define $\mathbf{P} = \mathbf{1}_{n\times 1}\mathbf{p}^\top$ to be the matrix of allele frequencies with n identical rows. Next, let $\mathbf{Z} = \mathbf{X} - 2\mathbf{P} + \mathbf{1}_n\mathbf{1}_m^\top$. Then, the genomic relationship matrix can be obtained via

$$\mathbf{G} = \frac{\mathbf{Z}\mathbf{Z}^\top}{\sum_{i=1}^{m} p_i(1-p_i)}.$$

GEBV are obtained by solving the following set of equations

$$\hat{\mathbf{a}} = \mathbf{G}(\mathbf{G} + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu})$$

where $\mathbf{R}$ is a diagonal matrix with entries $\mathbf{R}_{ii} = 1/h^2 - 1$, where h^2 is the parameter known as "heritability" and reveals the proportion of the phenotypic variation in a trait due to variation of the genetic factors. Heritability is a population-specific parameter, and depends on the allele frequencies, the effect of genetic variants as well as the environmental factors associated with the study population.

3.1.2 Genotypes and Phenotypic Records

A total of $1,800$ animals were used in this study, from a pool of $11,414$ beef cattle genotyped on the Illumina BovineSNP50 BeadChip (Illumina 50K) collated from various projects and research herds across Canada. Included in this study were animals from several breeds: a purebred Angus; a purebred Charolais; a composite population sired by Angus, Charolais, or hybrid bulls from the University of Alberta's Roy Berg Kinsella Research Ranch (Kinsella); a population of multibreed and crossbred cattle mainly Angus with proportions of Simmental, Piedmontese, Gelbvieh, Charolais, and Limousin from the University of Guelph's Elora Beef Cattle Research Station (Elora); a population of animals whose sire breeds were Angus, Charolais, Gelbveih and commercial crossbred from the the Phenomic Gap Project (PG1); a TX/TXX commercial population that is heavily influenced by Charolais with infusion of Holstein, Maine Anjou, and Chianina [64]. Quality controls (QC) were performed considering merged samples of all breeds simultaneously to filter out SNPs if one of the following holds: SNP (1) with minor allele frequency (MAF) < 0.01, (2) call rate < 0.90, and (3) heterozygosity excess > 0.15 [64]. A selected group of animals from the most influential beef cattle breeds and crossbred populations genotyped with both Illumina 50K and Affymetrix HD were used to further remove SNPs with conflicting alleles between the two panels. Exclusion of SNPs with missing, or duplicated coordinates and SNPs on sex chromosomes resulted in $33,911$ remaining SNPs with known physical positions on 29 autosomes for the Illumina 50K panel. Among the $33,911$ SNPs, we identified $5,088$ SNPs shared with the Illumina BovineLD Genotyping BeadChip (Illumina 6K). The physical map of the bovine genome used in this work was the UMD 3.1 assembly. From each of the six populations, 300 animals were randomly selected for our study. We refer to Kinsella, Elora, PG1 and TX/TXX as crossbred populations. All animals in this study are taurine breeds.

The phenotypic trait we considered in this study is residual feed intake (RFI), which is a measure of feed efficiency and is defined as the difference between an animal's actual daily feed intake and expected daily feed intake required for maintenance of body weight and growth, proposed by Koch et al. [57]. Values of RFI for all $1,800$ genotyped animals in the Illumina 50K panel were adjusted for contemporary groups including herd-year-sex, age at feedlot test and breed composition. The animal populations and traits are described in Basarab et al. [3], Chen et al. [15] and Lu et al. [64].

3.1.3 Scenario

Six imputation methods were investigated in this study, including Impute 2, FImpute 2.2, Beagle 4.1, Beagle 3.3.2, MaCH 1.0 and Bimbam 1.0. The imputation task was to impute genotypes from the Illumina 6K panel to the Illumina 50K panel. Five-fold cross validation was performed by randomly partitioning animals in each population into five non-overlapping groups. Each group consisted of 60 animals from each population, in total 360 across six populations. We simulated a low-density study sample by masking SNPs that belong to the 50K but not the 6K. About 15% $(5,088/33,911)$ of SNPs in a study sample were typed. In turn, each group was used as a study sample in the Illumina 6K while the rest of the four groups formed the reference set of Illumina 50K genotypes. That is, in each round of 5-fold CV, imputation was carried out for low-density target samples across six populations using a single reference panel composed of the $1,440$ animals across six populations. The partition of the dataset was used for both imputation and subsequent genomic predictions.

We applied two genomic prediction methods including an efficient GBLUP with a genomic relationship matrix (VanRaden [99]) and a Bayesian method (BayesB [74]) introduced in Chapter 2, together with imputed 50K genotypes from different methods and associated phenotypic values to predict the genomic breeding values (GEBV) in five-fold cross validation. In each round, actual 50K genotypes and associated adjusted RFI for animals in the reference panel were fit in the model as the training data, whereas a dataset containing imputed 50K genotypes was held for validation, assuming unknown phenotypic values. Additionally, we predicted the GEBV using BayesB and GBLUP based on actual 50K and 6K genotypes for comparisons.

3.1.4 Evaluation

To assess the qualities of imputed genotypes among various methods, a validation dataset is held with actual SNP genotypes assayed and by comparing the imputed genotypes against the actual ones one can get CR (also known as accuracy to the CS/ML audience). However, as Hickey et al. (2012) pointed out, concordance rates are allele-frequency dependent and do not reflect the power of any imputation method to infer rare allele variants with minor allele frequency (MAF) less than 1%. Additionally, Calus et al. [13] demonstrated that use of Pearson correlation coefficient between true and imputed genotypes is preferred to CR because it is more sensitive to errors at loci with lower MAF. Alternatively, the squared Pearson correlation coefficient (r^2) between the best guess (dosage) of genotypes and the actual genotypes can be used for imputation accuracy. The closer to 1 that r^2 is, the more power to detect an imputation method exhibits. We followed the notion of Howie et al. [50] by assigning undefined to 0 when imputation methods yielded all identical predictions for all individual at a marker. For programs (e.g. Impute 2) that report only marginal posterior probabilities $P(G = x)$, the best guess genotype (or imputed allele dosage) can be computed as $\sum_{x=0}^{2} x \cdot P(G = x)$. The accuracy of the genomic prediction for RFI in the validation population was calculated as Pearson's correlation coefficient between the estimated genomic breeding values (GEBV) using either GBLUP or BayesB and the adjusted phenotypic values of RFI.

3.1.5 Program Settings

We performed all the imputation experiments on a local computational cluster consisting of 15 identical nodes with dual quad core 64-bit CPUs run at 2.0 GHz and shared 8 GB memory. We ran all the programs using their population-based configurations without any pedigree information in the model fit. For Impute $2.3.1$, we followed its example commands under the scenario "imputation with one unphased reference panel", set the effective population size to 150 for cattle populations, calculated the recombination rates between two consecutive loci using the Haldane

(1919) recombination model by assuming that 1 million base pair approximately corresponded to 1 Morgan and used the default total MCMC iterations 30. For Beagle 4.1 (“09Nov15.d2a.jar”) and Beagle 3.3.2, the default numbers of iterations 15 and 10 were used in the study respectively. For MaCH 1.0, we first used MaCH’s haplotyping option to phase genotypes in the reference panel with two input files (a MERLIN formatted data file followed by the option “–d” and a pedigree file followed by the option “–p”) and the flags “–phase” and “–states 200”. It took 14 hours and 18 min on average for phasing the reference panel per fold. We did not provide with MaCH any map file in the all experiments. After completion of phasing unphased reference data, we used MaCH for imputing the study samples without any genetic map. For BimBam 1.0, we set the number of clusters “-c” to 15, and provided as inputs 1) a physical positions at each marker in each chromosome, 2) two unphased genotype files (one for the reference dataset and the other for the study sample), 3) default number of EM runs (“-e 10”), and 4) the default steps of each EM run “-s 1” of 5) the number of warm-up EM step runs (“-w 20”).

In this study for genomic predictions, heritability $h^2 = 0.2$ estimated on this dataset was used under all scenarios. The value of π in BayesB was set to 0.95. An implementation of the BayesB method by Fernando and Garrick, known as “Gensel” [33], was used in this study. Since Gensel requires no missing values in the Genotypic data $\mathbf{X}_{n\times m}$, Impute 2 with the option “-phase” was used to infer the small percentage of sporadic missing genotypes. In all BayesB experiments, we set the total number of iterations running the MCMC sampling to $150,000$ iterations and discarded first $20,000$ as burn-in. We examined Gensel’s output file ‘mcmcSample’ for trace plots of the residual variance in all experiments (results not shown), and confirmed all the chains had good mixings for the chosen chain length and burn-ins [14]. We used an implementation of GBLUP by Sargolzaei *et al.* in the software “GEBV” [66]. In GEBV, the genomic relationship matrix was efficiently computed using Colleau’s indirect method [20].

3.2 Results

3.2.1 Accuracy of Genotype Imputation

Table 3.1 shows the mean concordance rates (CR) in Column 2 and the mean squared correlation coefficients (r^2) in Column 3 respectively across all untyped SNPs for six different methods from the low density Illumina 6K panel to the medium-density 50K panel in 5-fold cross validation. A huge variance was observed among different methods in accuracies of imputation when a reference panel made up of composite populations was used. The overall mean CR and mean r^2 were the highest when Impute v2 was used for imputation, followed by FImpute 2 and Beagle 4.1 both of which yielded above 91% mean CR and above 66% mean r^2. Beagle 3.3.2 yielded a mean CR 87.38% and a mean r^2 0.5556. MaCH 1.0 and Bimbam 1.0 gave mean CRs 80.21% and 71.72% respectively and mean r^2 are 0.4180 and 0.2506 respectively.

The fifth column in Table 3.1 shows the average running time per fold for imputing 360 animals genotyped in the 6K chip while $1,440$ animals genotyped in the 50K chip were used as reference animals. In terms of speed, FImpute 2.2 was the fastest program yet achieved competitive imputation accuracies in terms of CR and r^2 to the currently best performing program Impute 2. FImpute 2.2 finished the whole-genome imputation only at a fraction of the latter's run time. Impute 2 was able to complete whole-genome imputation within a day for 360 animals. Beagle 4.1 had a great improvement over Beagle 3.3.2 in terms of imputation accuracies but had the longest running time 191 hours. Impute 2 overcame the quadratic running time with the number of animals by heuristically searching the closest reference haplotypes (defined by Hamming distances) [50]. However, the model-based imputation methods such as Impute 2 and Beagle 4.1 both suffer the scalability issue once we would like to impute from genotype chips up to the full sequence level. Table 3.2 shows the detailed imputation accuracies for each population. In Table 3.2, each method performed well with pure breed populations Angus and Charolais and the crossbred population Kinsella. Each method achieved the highest mean concordance rates with Angus, followed by Charolais and Kinsella. Due to differences

program	mean CR (%)	mean r^2	running time
Impute 2	**93.95**	**0.7545**	22 hrs 7 min 41 sec
FImpute 2.2	91.88	0.6626	**4 min 12 sec**
Beagle 4.1	91.70	0.6655	191 hrs 6 min 5 sec
Beagle 3.3.2	87.38	0.5556	31 hrs 22 min
MaCH 1.0	80.21	0.4180	16 hrs 53 min 46 sec
BIMBAM 1.0	71.72	0.2506	3 hrs 15min 50 sec

Table 3.1: Accuracy of genotype imputation from Illumina 6K to Illumina 50K for different methods. It took additional 14 hrs 18 min 14 sec for MaCH to pre-phase the unphased animals in the reference panel.

in their breeding programs, crossbred populations Elora, PG1 and TX/TXX exhibit high levels of genomic divergence in their population structure as evidenced by the number of genotypes that carry the minor allele in each class of MAF and as indicated by principal components in Figure 3.1. Impute v2 clearly outperformed all other methods in both mean CR and mean r^2 for the two purebred and four crossbred populations.

Population	Impute 2		FImpute 2.2		Beagle 4.1		Beagle 3.3.2		MaCH		BIMBAM	
	CR	r^2	CR	r^2	CR	r^2	CR	r^2	CR	r^2	CR	r^2
Angus	**97.75**	**0.7557**	96.47	0.7065	96.74	0.7152	94.69	0.6585	87.64	0.5288	77.89	0.3509
Charolais	**95.84**	**0.7523**	93.57	0.6616	93.00	0.6526	87.79	0.5207	78.13	0.3259	68.39	0.1543
Kinsella	**95.93**	**0.8458**	94.84	0.7875	94.51	0.7827	90.85	0.6787	83.16	0.5105	72.18	0.2895
Elora	**91.01**	**0.747**	88.21	0.6151	87.68	0.6091	82.15	0.4685	76.42	0.354	71.07	0.2518
PG1	**92.12**	**0.7738**	89.64	0.6595	89.87	0.6722	85.36	0.553	78.98	0.4178	72.17	0.2960
TX/TXX	**91.08**	**0.7319**	88.55	0.6132	88.39	0.6167	83.48	0.4962	76.95	0.3645	68.64	0.2068
All	**93.95**	**0.7545**	91.88	0.6626	91.70	0.6655	87.38	0.5556	80.21	0.418	71.72	0.2506

Table 3.2: Accuracy of genotype imputation from Illumina 6K to Illumina 50K for different methods and different populations

3.2.2 Effect of Minor Allele Frequency (MAF) on Accuracy of Genotype Imputation

We are also interested in the accuracy of each method for imputing genotypes that carry uncommon or rare variants as much of the causation of complex or quantitative traits is due to rare variants [19]. We evaluated imputation methods for their concordance rates on genotypes “AB” and “BB” carrying the minor allele “B” at each locus. To investigate the association between MAF and the accuracy of imputation among different methods, we classified the untyped markers into the

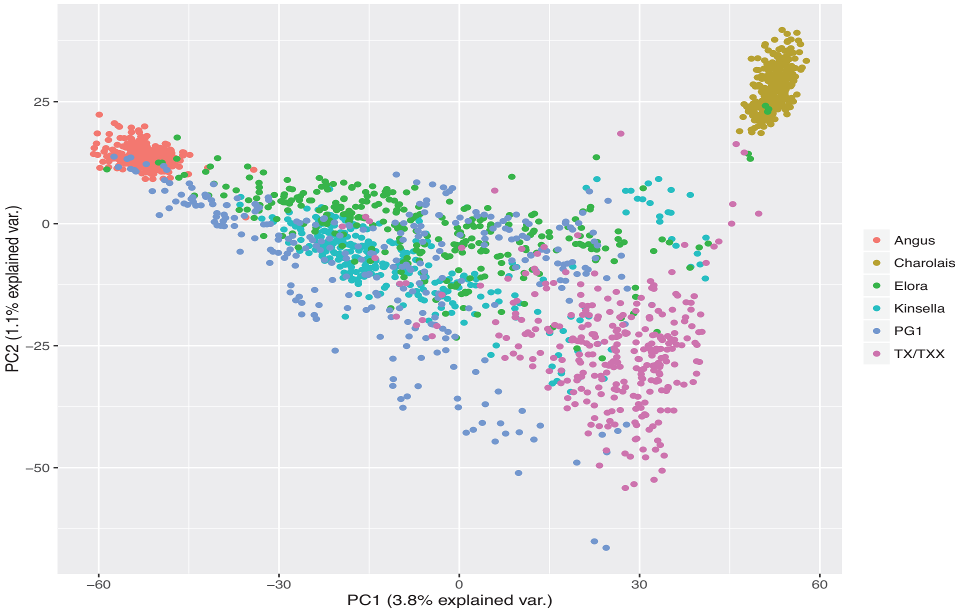
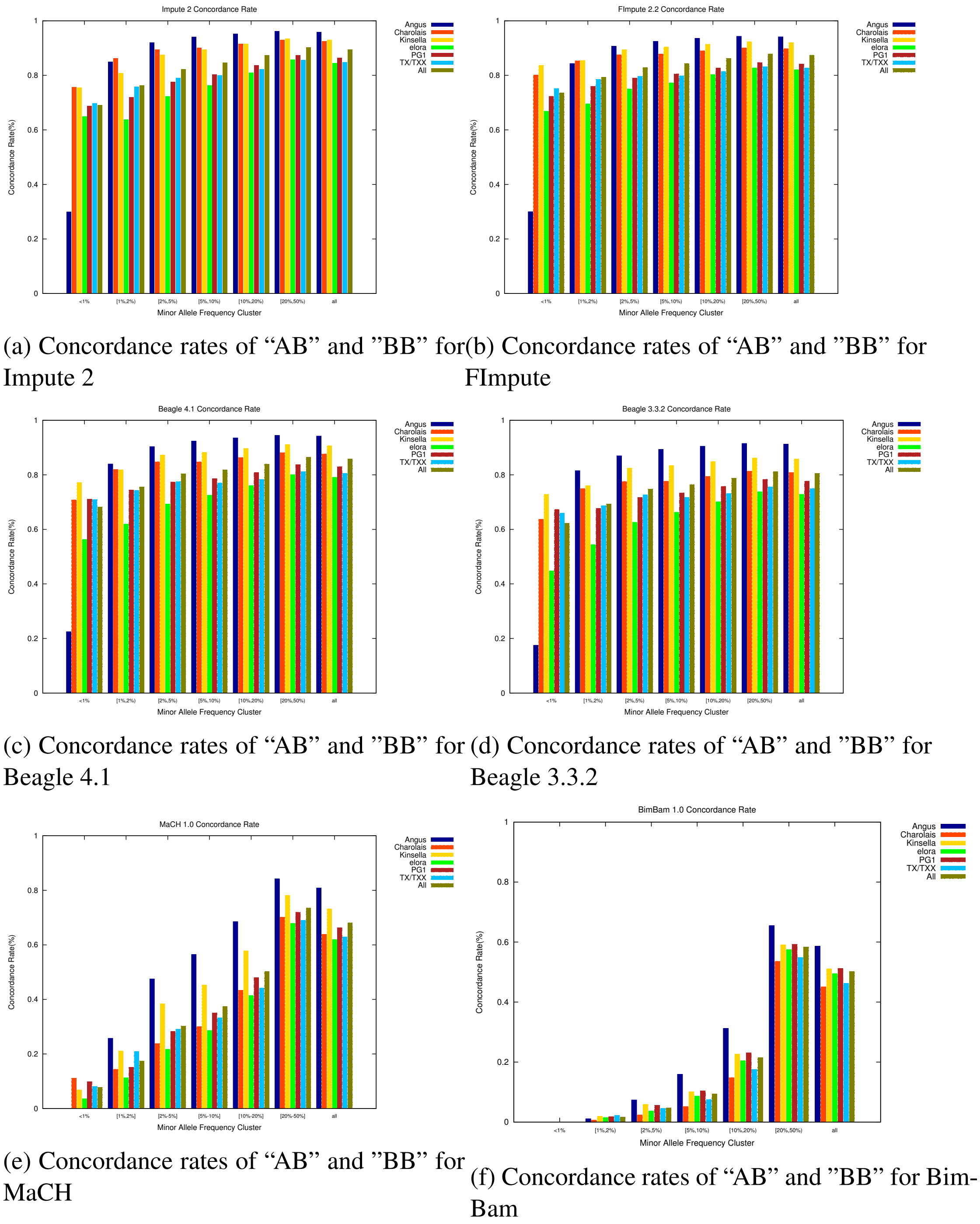


Figure 3.1: Principal component analysis (PCA) for population stratification using the top two principal components (PCs) obtained from 50K genotype data of all $1,800$ beef cattle. Individuals are grouped by their population, as described in Materials and Methods.

following six classes according to MAF, $(0, 1\%)$, $[1\%, 2\%)$, $[2\%, 5\%)$, $[5\%, 10\%)$, $[10\%, 20\%)$, and $[20\%, 50\%)$. Figure 3.2 (3.2a through 3.2f) shows the relationship between MAF and concordance rates of genotypes “AB” and “BB” for different methods. As MAF increased, concordance rates of all methods for imputing genotypes “AB” and ”BB” increased. The trends of imputation accuracy with MAF classes were consistent with reports from other studies in maize populations [48], and whole-genome sequencing Holstein Friesian cattle [97]. Greater differences among different methods were observed across variant MAF classes in the concordance rates of genotypes “AB” and “BB”. FImpute 2.2 outperformed Impute v2 for extremely rare variants (MAF class $(0, 1\%)$) across both pure and crossbred populations. For rare variants in MAF class $[1\%, 2\%)$ and $[2\%, 5\%)$, Impute v2 outperformed FImpute in purebred populations Angus and Charolais, but did worse than FImpute in crossbred populations Kinsella, Elora, PG1, and TX/TXX. Impute v2 had advantages over FImpute 2.2 for MAF greater 10%. The success of FImpute 2.2 was possibly due to their rule-based strategy for keeping haplotypes anchoring



(a) Concordance rates of “AB” and ”BB” for Impute 2

(b) Concordance rates of “AB” and ”BB” for FImpute

(c) Concordance rates of “AB” and ”BB” for Beagle 4.1

(d) Concordance rates of “AB” and ”BB” for Beagle 3.3.2

(e) Concordance rates of “AB” and ”BB” for MaCH

(f) Concordance rates of “AB” and ”BB” for BimBam

Figure 3.2: Effects of MAF of untyped SNPs on imputing genotypes “AB” and “BB” carrying the minor allele (MA) “B”

the rare allele in their update library. On the other hand, the model-based Impute v2 may ignore rare variants as mutations or errors when MAF was small. Beagle 4.1 and Beagle 3.3.2 performed worse than FImpute 2.2 and Impute 2 in each MAF class and were in the second tier. Beagle 4.1 outperformed Beagle 3.3.2 in each MAF class. MaCH did not yield comparable concordance rates in that we did not supply with the program an accurate haplotype reference. Although we applied MaCH's own phasing options in the first step for the reference data, no genetic map was provided and MaCH seemed to have difficulty in modelling the recombination and resolving phasing for the reference genotype data. Inaccurate haplotype data would have a significant impact on the subsequent genotype imputation process for MaCH as we observed in Figure 3.2e. A possible explanation for Bimbam's poor performance in imputation would be its over-generalization of the reference panel and its MLE for parameter inference. Bimbam was not designed for dealing with admixed populations and assumed that the reference data can be generalized through an MLE estimation with a local-clustered HMM. When the admixed population contained several breeds with distinct patterns of co-ancestry, the small number of clusters could result in MLE stuck in the local maxima as the distribution of the admixed data is likely to be multimodal.

The distribution of genotypes "AB" and "BB" in each MAF class for different populations in Table 3.3 clearly shows crossbred populations Kinsella, Elora, PG1 and TX/TXX in general contained more genetic variants than purebred populations Angus and Charolais. We can see from Table 3.3 the total number of genotypes that carry the minor allele across the genome was the fewest with Angus. Even though concordance rates of genotypes "AB" and "BB" were poorest for Angus with the MAF class $(0, 1\%)$, the number of such rare variants were extremely small and all methods were capable of imputing well for all MAF classes with Angus.

3.2.3 Accuracy of Genomic Predictions Using Actual 50K and Imputed 50K

We investigated two strategies of constructing training and validation datasets for genomic prediction. Across-breed training and validation datasets were constructed

Population	$(0, 1\%)$	$[1\%, 2\%)$	$[2\%, 5\%)$	$[5\%, 10\%)$	$[10\%, 20\%)$	$[20\%, 50\%)$	all
Angus	40	1429	21000	103384	490788	2792475	3409116
Charolais	226	4489	37597	135664	539437	2742995	3460408
Kinsella	306	4190	38420	134292	533583	2806099	3516890
Elora	444	3903	34136	127569	535479	2822135	3523666
PG1	661	5066	40973	137047	542973	2834610	3561330
TX/TXX	679	6402	49277	150388	559059	2798682	3564487
All	2356	25479	221403	788344	3201319	16796996	21035897

Table 3.3: Distribution of SNP genotypes (AB and BB) that carry the minor allele "B" among MAF classes

using animals across all six populations, whereas within-breed training and validations were constructed using animals of the same breed. That is, in the case of genomic predictions, in each round of 5-fold CV, the across-breed training dataset of actual 50K genotypes corresponded to our reference panel of $1,440$ animals across six populations whereas the within-breed training dataset was composed of only 240 animals of the same breed as the within-breed validation dataset.

Table 3.4 shows across-breed accuracies of genomic predictions between GEBV and adjusted RFI phenotypic values in Angus, Charolais, Kinsella, Elora, PG1, TX/TXX validation datasets using GBLUP and BayesB for actual 50K/6K SNP genotypes and imputed 50K genotypes. Columns with "actual 50K" and "actual 6K" show the genomic prediction results using actual 50K and actual 6K datasets as both training and validation datasets. Columns that have imputation methods-50K as titles report predication accuracies when using imputed 50K of the imputation method as validation datasets. A slight increase in Pearson correlation coefficient or accuracy of genomic prediction was observed for Angus, Charolais, Elora and TX/TXX via either BayesB or GBLUP when actual 50K training and validation datasets were compared with the actual 6K ones. However, there were no significant differences observed in correlation coefficients between the actual 50K and the actual 6K datasets for both BayesB and GBLUP methods when the standard errors were considered.

In comparison of genomic prediction accuracies of 50K to that of imputed 50K for across-breed genomic prediction, imputed 50K genomic prediction results from all the imputation methods except for Bimbam gave comparable accuracies to the

Population	Actual 50K		Actual 6K	
	BayesB	GBLUP	BayesB	GBLUP
AN	0.18 ± 0.05	0.21 ± 0.04	0.13 ± 0.03	0.16 ± 0.03
CH	0.22 ± 0.05	0.21 ± 0.06	0.14 ± 0.05	0.18 ± 0.06
KS	0.11 ± 0.08	0.08 ± 0.06	0.11 ± 0.06	0.08 ± 0.07
EL	0.09 ± 0.06	0.16 ± 0.05	0.05 ± 0.04	0.15 ± 0.04
PG	0.10 ± 0.05	−0.04 ± 0.06	0.12 ± 0.06	−0.01 ± 0.06
TX	0.19 ± 0.04	0.17 ± 0.04	0.14 ± 0.01	0.14 ± 0.04
All	0.15 ± 0.03	0.12 ± 0.03	0.11 ± 0.01	0.12 ± 0.03

Population	Impute2-50K		Fimpute2.2-50K	
	BayesB	GBLUP	BayesB	GBLUP
AN	0.18 ± 0.05	0.20 ± 0.04	0.16 ± 0.05	0.18 ± 0.02
CH	0.22 ± 0.05	0.20 ± 0.06	0.23 ± 0.07	0.22 ± 0.07
KS	0.10 ± 0.08	0.08 ± 0.06	0.09 ± 0.07	0.10 ± 0.05
EL	0.09 ± 0.06	0.15 ± 0.06	0.11 ± 0.06	0.16 ± 0.03
PG	0.10 ± 0.05	−0.04 ± 0.06	0.12 ± 0.06	−0.02 ± 0.07
TX	0.18 ± 0.04	0.16 ± 0.04	0.16 ± 0.03	0.18 ± 0.06
All	0.14 ± 0.03	0.11 ± 0.03	0.14 ± 0.03	0.12 ± 0.03

Population	Beagle4.1-50K		Beagle3.3.2-50K	
	BayesB	GBLUP	BayesB	GBLUP
AN	0.18 ± 0.05	0.20 ± 0.04	0.17 ± 0.05	0.20 ± 0.04
CH	0.21 ± 0.06	0.20 ± 0.07	0.22 ± 0.07	0.20 ± 0.08
KS	0.11 ± 0.08	0.08 ± 0.06	0.10 ± 0.06	0.07 ± 0.05
EL	0.07 ± 0.05	0.14 ± 0.05	0.07 ± 0.06	0.15 ± 0.05
PG	0.11 ± 0.06	−0.04 ± 0.06	0.10 ± 0.05	−0.04 ± 0.06
TX	0.17 ± 0.03	0.16 ± 0.04	0.19 ± 0.03	0.16 ± 0.03
All	0.14 ± 0.03	0.11 ± 0.03	0.14 ± 0.03	0.11 ± 0.03

Population	MaCH-50K		Bimbam-50K	
	BayesB	GBLUP	BayesB	GBLUP
AN	0.16 ± 0.05	0.19 ± 0.04	0.13 ± 0.04	0.15 ± 0.03
CH	0.24 ± 0.06	0.21 ± 0.07	0.18 ± 0.04	0.18 ± 0.05
KS	0.11 ± 0.06	0.08 ± 0.05	0.08 ± 0.05	0.05 ± 0.04
EL	0.05 ± 0.07	0.14 ± 0.06	0.06 ± 0.05	0.15 ± 0.05
PG	0.12 ± 0.07	−0.04 ± 0.06	0.10 ± 0.06	−0.07 ± 0.04
TX	0.15 ± 0.03	0.15 ± 0.03	0.12 ± 0.03	0.14 ± 0.03
All	0.13 ± 0.03	0.11 ± 0.03	0.11 ± 0.02	0.09 ± 0.03

Table 3.4: Across-breed accuracy of genomic estimated breeding values predicted with actual 6K panel, actual 50K panel, imputed 50K panels from Impute 2, FImpute 2, Beagle 4.1, Beagle 3.3.2, MaCH, and Bimbam for RFI using GBLUP and BayesB for Angus (AN), Charolais (CH), Kinsella (KS), Elora (EL), PG1 (PG), TX/TXX (TX) validation groups. Standard errors of the mean from the five-fold cross validation follow after and are defined as SEM $= \frac{\sigma}{\sqrt{5}}$ where σ is the sample standard deviation. Training groups consist of $1,440$ animals pooled from all six populations.

actual 50K results using both GBLUP and BayesB. For purebred Charolais, the highest mean correlation coefficients were 0.24 using BayesB on imputed 50K via MaCH although the mean concordance rate of MaCH was only 78.13%, 0.23 using BayesB on imputed 50K via FImpute, 0.22 using GBLUP on imputed 50K via FImpute, 0.22 using BayesB on imputed 50K via Impute 2, and 0.22 using BayesB on actual 50K genotypes. With Charolais on either imputed or actual 50K panels, BayesB gave slightly better or similar accuracies compared to GBLUP although the advantage was not statistically significant. With Angus on either imputed or actual 50K panels, GBLUP tended to give higher accuracies than BayesB and again the small advantage was not significant. While in crossbred cattle populations Kinsella, Elora, PG1, TX/TXX, the most highest mean correlation coefficients was 0.19 using BayesB on imputed 50K TX/TXX from Beagle 3.3.2 and actual 50K. Bimbam imputed 50K yielded slightly lower prediction accuracies in comparison to that of actual 50K for purebred Angus and Charolais. For across-breed genomic prediction based on either actual 50K or imputed 50K SNPs, BayesB and GBLUP had similar prediction accuracies for all the breed/populations except for PG1, for which BayesB yielded significantly higher prediction accuracies than that of GBLUP.

Within-breed accuracies of GEBV predictions for RFI using BayesB and GBLUP in all six populations are presented in Table 3.5. Similarly, genomic prediction accuracies of actual 50K, actual 6K and imputed 50K are similar. Unlike across-breed genomic prediction, Bimbam imputed 50K of within-breed genomic prediction had similar prediction accuracies to that of actual 50K. Moreover, within-breed GBLUP improved accuracies using either imputed 50K or actual 50K/6K for crossbred population PG1. However, GBLUP still yielded slightly lower prediction accuracies for Charolais than that of BayesB using either actual 50K, actual 6K and imputed 50K of various methods while for breeds including Angus, Kinsella, Elora, PG1, and TX/TXX, GBLUP and BayesB had comparable genomic prediction accuracy for the trait.

In comparison to the results of across-breed genomic predictions, the within-breed genomic prediction yielded relatively higher accuracies for purebred Angus under BayesB and for crossbred PG1 under GBLUP. For both across and within-

Population	Actual 50K		Actual 6K	
	BayesB	GBLUP	BayesB	GBLUP
AN	0.24 ± 0.03	0.25 ± 0.01	0.23 ± 0.05	0.26 ± 0.02
CH	0.21 ± 0.06	0.20 ± 0.06	0.19 ± 0.06	0.20 ± 0.05
KS	0.10 ± 0.06	0.12 ± 0.06	0.11 ± 0.07	0.13 ± 0.06
EL	0.17 ± 0.05	0.18 ± 0.05	0.16 ± 0.02	0.18 ± 0.04
PG	0.13 ± 0.06	0.16 ± 0.08	0.15 ± 0.03	0.14 ± 0.07
TX	0.17 ± 0.04	0.18 ± 0.04	0.13 ± 0.05	0.14 ± 0.04

Population	Impute2-50K		Fimpute2.2-50K	
	BayesB	GBLUP	BayesB	GBLUP
AN	0.25 ± 0.02	0.25 ± 0.01	0.24 ± 0.02	0.24 ± 0.01
CH	0.21 ± 0.06	0.20 ± 0.06	0.21 ± 0.06	0.21 ± 0.07
KS	0.11 ± 0.06	0.12 ± 0.06	0.10 ± 0.06	0.12 ± 0.06
EL	0.15 ± 0.05	0.16 ± 0.04	0.14 ± 0.05	0.17 ± 0.03
PG	0.13 ± 0.06	0.15 ± 0.07	0.14 ± 0.06	0.15 ± 0.07
TX	0.16 ± 0.04	0.18 ± 0.04	0.15 ± 0.04	0.18 ± 0.05

Population	Beagle4.1-50K		Beagle3.3.2-50K	
	BayesB	GBLUP	BayesB	GBLUP
AN	0.24 ± 0.03	0.25 ± 0.01	0.24 ± 0.02	0.25 ± 0.01
CH	0.21 ± 0.06	0.20 ± 0.06	0.21 ± 0.06	0.20 ± 0.07
KS	0.14 ± 0.04	0.12 ± 0.06	0.11 ± 0.06	0.13 ± 0.06
EL	0.04 ± 0.06	0.16 ± 0.04	0.13 ± 0.05	0.15 ± 0.04
PG	0.11 ± 0.06	0.16 ± 0.08	0.14 ± 0.06	0.16 ± 0.08
TX	0.16 ± 0.04	0.17 ± 0.04	0.15 ± 0.05	0.17 ± 0.05

Population	MaCH-50K		Bimbam-50K	
	BayesB	GBLUP	BayesB	GBLUP
AN	0.24 ± 0.02	0.24 ± 0.02	0.13 ± 0.04	0.26 ± 0.01
CH	0.22 ± 0.06	0.21 ± 0.06	0.18 ± 0.04	0.22 ± 0.06
KS	0.11 ± 0.06	0.13 ± 0.06	0.08 ± 0.05	0.14 ± 0.05
EL	0.15 ± 0.05	0.16 ± 0.05	0.06 ± 0.05	0.16 ± 0.05
PG	0.14 ± 0.07	0.16 ± 0.08	0.10 ± 0.06	0.16 ± 0.08
TX	0.14 ± 0.05	0.16 ± 0.04	0.12 ± 0.03	0.14 ± 0.05

Table 3.5: Within-breed accuracy of genomic estimated breeding values predicted with actual 6K panel, actual 50K panel, imputed 50K panels from Impute 2, FImpute 2, Beagle 4.1, Beagle 3.3.2, MaCH, and Bimbam for RFI using GBLUP and BayesB for Angus (AN), Charolais (CH), Kinsella (KS), Elora (EL), PG1 (PG), TX/TXX (TX) validation groups. Standard errors of the mean from the five-fold cross validation follow after and are defined as SEM $= \frac{\sigma}{\sqrt{5}}$, where σ is the sample standard deviation. Training groups consist of 240 animals from the within breed population while validation groups contain 60 animals from the same breed.

breed genomic predictions based on either actual 50K, actual 6K and imputed 50K SNPs, purebred populations (Angus and Charolais) had relatively high prediction accuracies than that of crossbred populations Kinsella, Elora, PG1, TX/TXX.

3.3 Discussion

Factors that affect the accuracy of imputation from previous studies include the number of genotyped immediate ancestors, the size of the reference panel, the linkage disequilibrium between typed and untyped SNPs, the composition of the reference panel, the relationship of individuals between the study sample and reference population, and minor allele frequencies [53, 45, 68, 6, 48, 13]. Bouman and Veerkamp [6] showed that combining animals of multiple breeds was preferred to a small reference panel comprised of animals of the same breed for imputation from high-density SNP panels to whole-genome sequence, especially for low MAF loci. In our study, we adopted this strategy to construct reference panels with animals across six populations, and observed that it was especially beneficial to FImpute for imputing rare variants. Since rare alleles might be under-represented in a single population, as shown in Table 3.3 under the column "$(0, 1\%)$" for Angus for example, and FImpute relies on observed alleles to build up its haplotype library, haplotypes carrying the rare variants can be borrowed from other breeds or populations. As we move from low MAF to high MAF, the accuracy of imputation for genotypes that carry the minor allele improves for all methods as shown in Figure 3.2a through Figure 3.2f because imputation methods have higher confidence in imputing untyped genotypes at higher MAF loci.

Genotype imputation methods such as fastPHASE and Bimbam that adopt maximum likelihood estimation (MLE) yielded poor accuracies of imputation likely due to their model-based estimation of the admixed population structure of our genotype data. Compared to Beagle 3.3.2's haplotype frequency based model, which builds up clusters based on the current estimates of haplotypes, fastPHASE and Bimbam derive clusters from the generalization of data. With fastPHASE and Bimbam, two haplotypes with two distinct alleles at the current locus could end up in the same cluster, whereas with Beagle 3.3.2 they are guaranteed to be in different clusters [9]. Therefore, at low-MAF loci, fastPHASE and Bimbam tend to cluster the rare allele and the major allele into the same cluster and mistake heterozygous genotypes carrying the rare allele as homozygous genotypes carrying the major allele [51], as evidenced in Figure 3.2f where Bimbam did not make any correct predictions for genotypes carrying rare alleles (MAF $< 1\%$). Figure 3.1 shows a plot of the principal component analysis (PCA) using the top two principal components (PCs). It has long been known that the MLEs of finite mixtures can lead to local maxima [84, 104]. Both fastPHASE and Bimbam rely on estimation of clusters in their model settings via the MLE. Recently, Feller et al. [32] examined pathological behaviours of the MLEs via a mixture of two normal distributions and showed the MLEs can wrongly estimate the component means to be equal when the mixture components are weakly separated and convergence of the parameters in the MLE setting sometimes can break down.

Previous studies on Holstein dairy cattle for imputation from 6K to 50K show an overall CR over 93% with Beagle 3.1.0 [4], over 97% with Fimpute [14], over 98% with Fimpute [89]; from 6K to 50K, our findings with several purebred/crossbred beef populations (overall mean CR 91.88% with FImpute) were similar to the ones from beef cattle reported by Piccoli et al. [78], Ventura et al. [101], and Chud et al. [18]. Accuracies of imputation were in general higher in Holstein dairy breeds than in beef breeds based on previous reports and our studies, as levels of LD were higher in Holstein dairy breeds than in beef breeds because Holsteins have a relatively small effective population size [52]. The design of the Illumina 6K chip is another factor that results in different accuracies of imputation in various breeds and populations [18]. The SNPs on this panel were selected to provide optimized imputation in dairy breeds [5] and thus lower performance in beef breeds is expected, as is lower performance in indicine breeds relative to taurine breeds.

We observed in this study that the accuracies of genomic prediction of RFI are not sensitive to imputation errors in general when the 6K SNPs were imputed to the 50K SNPs except for the Bimbam method, which yields lower genomic predict accuracies in across-breed genomic prediction. Also, genomic predictions based on actual 6K SNPs resulted in similar accuracies to that of actual 50K SNPs. However, in within-breed genomic prediction Bimbam imputed 50K achieved comparable genomic predictions to that of the actual 50K. Our results are in line with reports by Li et al. [12] where a larger number of beef cattle (over 5,000) from the same data pool as ours were used for evaluation of accuracy of genomic prediction for RFI based on imputed Affymetrix HD SNPs (428K SNPs used) and 50K SNPs under three different Bayesian methods. The imputed HD and actual 50K SNP data yielded similar accuracies under all three methods. van Binsbergen et al. [98] also reported no improvement in accuracy of genomic prediction was observed when using imputed sequence data over BovineHD data, suggesting that increases in density of imputed genotypes may not necessarily lead to an increase in accuracy of genomic prediction with the current SNP panel information and statistical methods.

Previous studies [75, 82, 15] have shown evidence that RFI is a complex trait likely to be controlled by many SNPs with small effects. Therefore, genotype imputation errors from 6K to 50K SNP as observed in this study may have minimal impacts on the accuracy of genomic prediction for RFI. However, when a trait is influenced by a few of SNPs with major effects, imputation error will likely affect the genomic prediction accuracy as shown in Chen et al.'s studies on genomic predictions of fat percentage using dairy cattle [14]. For RFI genomic prediction, FImpute was suggested as an imputation method as it is fast and has advantages over all other methods in imputing rare variants.

In our study, GBLUP and BayesB methods yielded comparable genomic prediction accuracies for the trait for across-breed and within-breed genomic prediction in most of the breed/populations, which is in agreement with the previous reports [46, 100, 75, 67]. GBLUP is believed to be less sensitive than BayesB to the genetic architecture of any trait as it relies mainly on pairwise relationship between individuals across the genome for prediction [96]. However, it was observed that GBLUP gave lower prediction accuracies than BayesB in the PG1 population for the across-breed training strategy under all the SNP types (actual 50K, actual 6K and imputed SNPs), but resulted in comparable prediction accuracies to BayesB when the within-breed strategy was adopted. PG1 is a crossbred population with animals being more widespread in the plot of PCA in Figure 3.1, indicating greater dissimilarity of animals in the population in comparison to other populations, which usually lead to a relatively lower prediction accuracy. Lund et al. [65] reported that there was little or no benefit when combining distantly related breeds such as Jersey and Holstein using GBLUP. Effects of across-breed genomic predictions have been studied by De Roos et al. [25] through simulation studies, which conclude that the across-breed training could lead to suboptimal marker effects for each population as linkage disequilibrium between markers and QTL would be unlikely to persist across populations and suggested high density marker set must be needed when across-breed training is applied. Therefore, the greater dissimilarity of animals in PG1 may lead to lower prediction accuracies of GBLUP. Moreover, the very low prediction accuracy of GBLUP in PG1 could also be attributed to a greater sampling error due to more genetic dissimilarity among animals as shown in Figure 3.1, coupled with a small validation population size ($N = 60$) in the study.

The level of relatedness between training and validation set has a determinant role on the accuracy of both imputation and genomic prediction. Previous authors including Habier et al. [43, 44] and Sun et al. [94] show the genetic relationship among animals as reflected in LD or linkage phase persistence or co-segregation (CS) of QTL with SNPs can contribute to accuracy of genomic predictions in SNP-based models. CS of alleles at two loci indicates that these alleles both originate from the same chromosome of a parent, and captures more recent close relationship between individuals. A closer relatedness between training and validation leads to higher persistency of CS among animals [44, 94], which will improve the accuracy of both imputation and genomic prediction. CS has advantages over LD because common SNPs usually have higher MAF whereas most QTLs are rare allelic variants and LD under such scenario becomes weak. When LD between QTL and SNPs is weak, which is believed to be the case for multiple beef cattle populations due to the difference in breeding and selection of different breeds, CS information therefore becomes a more dominant factor in affecting accuracy of genomic predictions for the across-breed strategy. Employing a within-breed training strategy improves the accuracies in purebred populations in that within-breed training and validation dataset comprised of more closely related individuals results in an increase of CS, and its persistence is higher than that of across-breed genomic prediction [94], which was shown by Chen et al. [15] and also is consistent with the results in this study for the purebred Angus and Charolais populations. Principal component analysis (PCA) has been widely applied to inferring genetic structure and exploring the level of relatedness in cattle. For more closely related individuals, the expected length of shared haplotypes is larger and population-based imputation methods have higher confidence to predict untyped genotypes if immediate ancestors are present in the reference panel [59, 48, 13]. From the plot of PCA in Figure 3.1, purebred Angus and Charolais cattle are positioned distantly from each other, but tend to have similar major components with animals of the same breed, and exhibit a greater genetic similarity and a closer relationship within each breed. However, crossbred animals within the same population are more dispersed, implying that crossbred animals within the same population are more genetically divergent. If the study sample is distantly related to the training population or the reference panel, the average accuracy of imputation and genomic prediction were lower, which has been demonstrated in previous studies with dairy cattle populations [43, 65].

The density of DNA markers is expected to affect accuracy of genomic predictions as use of genotypes in a higher-density SNP panel would on average result in an increase of the level of linkage disequilibrium (LD) between a SNP marker and a QTL. However, it is not unprecedented to observe no gain or a small gain between a low density 6K and a higher density SNP panel 50K as observed in this study in beef cattle, suggesting that increasing density of SNP panels by simply adding SNPs with high MAF will unlikely improve LD between SNPs and QTL of rare MAF [95], and further studies are needed to make a better use of existing higher density SNP panels and design better higher density SNP panels to improve genomic prediction accuracy. Previous genomic prediction studies of RFI and milk production traits in dairy cattle by Pryce et al. [82], Erbe et al. [30], and Ertl et al. [29] showed only a slight gain in accuracy as SNP marker density increased. However, it may be still worthwhile to investigate the impacts of imputation errors on genomic prediction for higher density SNPs or whole genome SNPs on other traits in larger populations of beef cattle.

Chapter 4

Piecemeal Imputation

In this chapter, we introduce a strategy called "piecemeal imputation" for boosting the accuracy of imputation based on existing imputation methods and marker panel information in a multi-step procedure. A version of this chapter has been published in BMC bioinformatics [103]. The goal is to improve the accuracy of imputation from a low-density chip to a high-density chip via an intermediate pseudo-chip. Usually, a set of animals genotyped in both the high-density (HD) chip and the low-density (LD) chip are held for validation purposes. In this study, the accuracy of imputation from LD to HD is computed as the percentage of correctly imputed genotypes assuming that actual genotypes in high-density are true.

We refer to running an existing imputation method one time to directly impute from a low density chip to a high density chip as the *one-step* imputation. Several studies in bovine genomics showed evidence that *two-step* imputation is generally more accurate than the one-step imputation, where the lower density genotyped animals are first imputed to a medium density SNP set and then further impute to the higher density [55, 47]. For instance, Larmer *et al.* [60] showed that for Beagle [10], FImpute [88], and Impute v2 [51], the two-step imputation from 6K to 50K then to 777K achieves higher accuracies than the one-step imputation from 6K directly to 777K. The exact reason why two-step imputation performs better than the one-step imputation is possibly . One possible explanation would be that some imputation algorithms (such as Beagle 3.3.2) have to choose from multiple matches or near matches of haplotypes between low-density and high-density chips for filling the unphased genotypes, whereas the number of choices could be reduced when a medium chip is introduced [97]. We conducted the *"add-one" two-step experiments*, in which the median density reference panel contains only one extra SNP than the low density SNP panel. While rotating this extra SNP from the pool of markers in the high density panel, we observed that a portion of them can individually boost the imputation accuracy in the add-one two-step experiment compared to the one-step direct imputation. We present a novel two-step piecemeal imputation framework, which essentially builds an intermediate pseudo array by mining the hidden relations between the lower and the higher density arrays. The pseudo array in the intermediate step is an artificial one derived from a learning procedure, which evaluates and selects some SNP markers based upon their add-one two-step imputation performance. Moreover, the pseudo-arrays are model-dependent. That is, different base imputation programs built upon different models could result in different selection of markers for our two-step piecemeal imputation. We demonstrate that by wrapping either Beagle or FImpute in our two-step piecemeal imputation framework, we are able to achieve higher genotype imputation accuracies.

4.1 Methods

Figure 4.1 shows a flow chart of our two-step imputation process, with the training process through the 5-fold cross validation on the left and the independent testing on the right. For ease of presentation, we use the Illumina 6 K gene chip to represent the lower density chip and the Illumina 50 K gene chip to represent the higher density one. Following previous definitions of $\mathcal{T}$ and $\mathcal{U}$, in the training process, we masked SNPs in $\mathcal{U}$ for 1-fold of the animals to form study sample denoted $\mathcal{S}$ while keep the remainder of 4 folds of the animals as the reference sample denoted $\mathcal{R}$. The genotype dataset thus can be represented as $(\mathcal{S} \cup \mathcal{R}, \mathcal{T} \cup \mathcal{U})$. The study samples are genotyped on the 6 K SNP set $\mathcal{T}$, and the reference samples are genotyped on the 50 K SNP set $\mathcal{T} \cup \mathcal{U}$. The goal is to impute the genotype values on U for the study samples. The top two lines in Figure 4.2 plot $\mathcal{T} \cup \mathcal{U}$ and $\mathcal{T}$, respectively, using their physical loci on the first half of chromosome 14 (BTA 14). Our goal is to impute the untyped genotypes in $\mathcal{U}$ for the study samples.

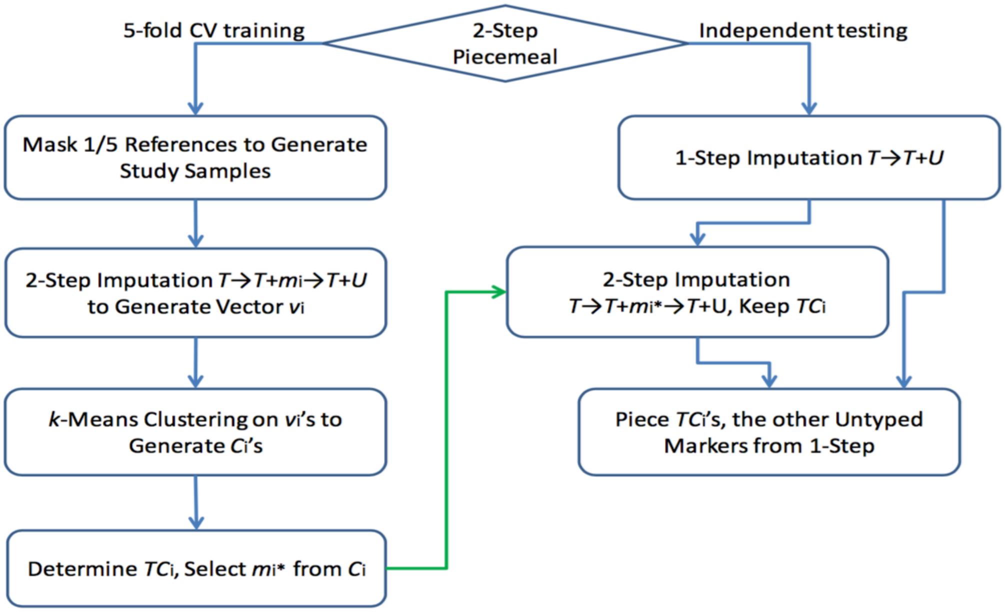

5-fold CV training
2-Step Piecemeal
Independent testing
Mask 1/5 References to Generate Study Samples
1-Step Imputation T→T+U
2-Step Imputation T→T+mi→T+U to Generate Vector vi
2-Step Imputation T→T+mi*→T+U, Keep TCi
k-Means Clustering on vi's to Generate Ci's
Piece TCi's, the other Untyped Markers from 1-Step
Determine TCi, Select mi* from Ci

Figure 4.1: A flow chart of the two-step piecemeal imputation framework, including both the training phase through a 5-fold cross validation and independent testing. $\mathcal{T}$ is the set of markers in the lower density chip and $\mathcal{T} \cup \mathcal{U}$ is the set of markers in the higher density chip; mi is a marker of $\mathcal{U}$; $\mathcal{S}$ is the set of study samples genotyped on $\mathcal{T}$ and $\mathcal{R}$ is the set of references genotyped on $\mathcal{T} \cup \mathcal{U}$. The goal is to impute the genotype for markers of U for the study samples

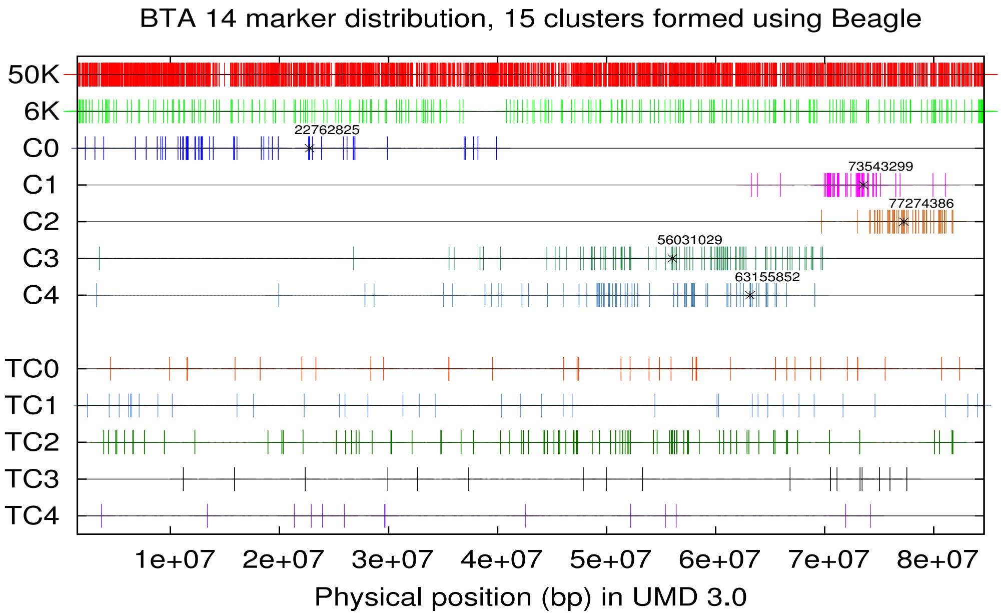


Figure 4.2: Untyped SNP genotype piecemeal imputation. Both the SNP set $\mathcal{T}$ of a lower density 6 K chip and the SNP set $\mathcal{T} \cup \mathcal{U}$ of a higher density 50K chip are shown, using their physical loci on BTA14. The second to the seventh lines plot the SNPs in the first five clusters, by the k-means algorithm ($k = 15$) on the marker feature vectors generated by the add-one two-step imputation using Beagle. The starred markers are the selected markers, one per cluster, and the associated target marker clusters are shown in the last five lines in the figure

One-Step Imputation We first present the training process. We chose Beagle and FImpute as our two base programs because of their relative fast speed for imputation. We ran either program on the simulated dataset ($\mathcal{S} \cup \mathcal{R}, \mathcal{T} \cup \mathcal{U}$) and collected the achieved CV accuracy denoted acc_1 as the proportion of genotypes correctly imputed in $\mathcal{U}$ assuming that the masked genotypes have no errors.

Add-One Two-Step Imputation For each untyped marker $m_i \in \mathcal{U}$, an add-one two-step imputation from $\mathcal{T}$ to $\mathcal{T} \cup \{m_i\}$, then from $\mathcal{T} \cup \{m_i\}$ to $\mathcal{T} \cup \mathcal{U}$ is conducted to evaluate its potential in imputing other untyped markers. Our goal in the training process is to select a relatively small portion of SNPs from $\mathcal{U}$, denoted $\mathcal{M}$ (in our case $\mathcal{M} = \{m_i\}$ in each iteration), and append them to $\mathcal{T}$ to create an intermediate pseudo array, in the hope that the subsequent two-step imputation from $\mathcal{T}$ to $\mathcal{T} \cup \mathcal{M}$, then to $\mathcal{T} \cup \mathcal{U}$ yielded higher imputation accuracy than acc_1. At the end of the two-step process, a *feature vector* $v_i = (a_{i1}, a_{i2}, \cdots, a_{i|U|})$ was obtained by calculating accuracy for each added marker m_i at each locus j of $\mathcal{U}$ across all *study* animals as the proportion of correctly imputed genotypes across 5-fold CVs.

Marker Clustering and Target Marker Cluster Intuitively, two markers of similar feature vectors have about the same performance to impute other untyped markers, when they are independently appended to $\mathcal{T}$ in 2-step imputation. Thus, it is sufficient to include only one of them. The k-means clustering algorithm was applied to cluster feature vectors, where k is the number of clusters that can be empirically determined. We examined $k = 5$ to 100 clusters with an increment of 5 when k-means was applied. The resultant clusters denoted $C_1, C_2, \cdots, C_k$ are k groups of SNPs in $\mathcal{U}$.

For each cluster C_i, if average accuracy at marker $m_j \in \mathcal{U}$ is consistently higher than or equal to the one-step imputation accuracy acc_1, then m_j would be a target marker for cluster C_i. The set of all target markers for cluster C_i form the target marker cluster TC_i for cluster C_i. Note that the target cluster associated with a cluster C_i could be empty and in case of empty target clusters, no markers would be selected to form the pseudo array; otherwise define the contribution of a marker m_j of C_i as the add-one two-step imputation accuracy from $T \cup \{m_j\}$ to $\mathcal{T} \cup \mathcal{U}$. Eventually, we have at most k selected markers $\mathcal{M} = \{m_{1^*}, m_{2^*}, \cdots, m_{k^*}\}$ for piecemeal imputation. When an independent dataset is available, we would apply add-one two-step with m_{i^*} to impute only untyped markers in TC_i. We call the target marker cluster TC_i one *piece* of final imputation result. Notice that TC_i's can overlap with each other and an untyped marker may not belong to any target marker cluster. In case of overlapping pieces TC_i's, majority voting scheme is used to resolve ambiguities, if any. In case of any untyped marker not belonging to any target cluster TC_i's, we use one-step imputation result at this particular marker for imputing. Piecing these tracts from add-one two-step imputation and one-step imputation together gives the final piecemeal imputation result. The final piecemeal imputation accuracy from CV is the average over all five folds and is denoted as acc_π.

4.2 Experimental Results

4.2.1 Datasets – Sequence Animals

The Canadian Cattle Genome Project [93] has contributed more than 350 animals to the 1000 Bull Genomes Project. From these projects we derived two datasets: a Holstein sequence collection containing 114 animals, and a Simmental sequence collection containing 82 animals. They were used for the piecemeal imputation method training through a 5-fold cross validation process (i.e. partitioned into a subset of study samples and another subset of reference samples). They also served as reference animals in subsequent independent testing experiments.

4.2.2 Genotyped Animals

From the Canadian Cattle Genome Project, we obtained 390 Simmental animals genotyped with the Affymetrix 660 K chip. Further investigation showed 23 of these 390 Simmental animals also appeared in our sequenced Simmental dataset. The genotyped animals that did not occur in the set of sequenced animals are used as our study samples in the independent testing experiments from a lower (than 660 K) density to impute their genotypes at the density 660 K. The 23 genotyped and sequenced animals were used as study samples in independent testings for imputation from a lower density up to their whole sequence.

4.2.3 SNP Sets

We used single chromosomes of small length (BTA 27) or medium length (BTA 14) in the development of the piecemeal imputation framework. BTA 27 was chosen for the Holstein data set while BTA 14 for the Simmental data set. The only challenge to deal with all 29 bovine chromosomes is the requirement for a huge amount of disk storage, see Discussion.

The numbers of SNPs included in the Illumina 6 K, 50 K, 777 K and the Affymetrix 660 K are summarized in Table 4.1, where the second column contains their formal chip names that one can look up on the Illumina and Affymetrix websites.

On BTA 27, the 114 sequenced Holstein animals have genotype values for $529,674$ SNPs. The Illumina 777 K chip contains $10,219$ of them, among which 664 are included in the 50 K chip, and 119 of these 664 SNPs are included in the 6 K chip, as summarized in Table 4.2. On BTA 14, the 82 sequenced Simmental animals have genotype values for $933,833$ SNPs. Table 4.2 shows that the Affymetrix 660 K chip contains $14,367$ of these $933,833$, among which $1,618$ are included in the 50 K chip, and a further 219 of these $1,618$ SNPs are included in the 6 K chip.

SNP chip	Chip Name	No. SNPs
Illumina 6 K	Illumina BovineLD BeadChip	$6,909$
Illumina 50 K	Illumina BovineSNP50 BeadChip	$54,001$
Illumina 777 K	777 K BovineHD BeadChip	$786,799$
Affymetrix 660 K	Axiom Genome-Wide BOS 1 Array	$648,875$

Table 4.1: Description of the different SNP chips and the SNP subsets

Chr	No. Animals	No. SNPs	HD	50 K	6 K
BTA 27	114	$529,674$	$10,219$	664	120
BTA 14	82	$933,833$	$14,367$	$1,618$	219

Table 4.2: Description of the different SNP chips and the filtered SNP subsets used in the study

4.2.4 5-fold cross validation

We use 5-fold cross validation to empirically examine our piecemeal imputation method, also to construct (a.k.a. "train") the staircase pseudo arrays to impute the genotyped animals to their whole genome. The cross validation results also suggest the possible levels of improvement compared with the one-step imputation.

Table 4.3 contains the cross validation results (Columns 3 to 7) on the Simmental datasets of 82 animals, where the lower density is either 6 or 50 K and the higher density refers to either 50 or 660 K (second column). The third and the fourth columns hold the one-step (acc_1) and piecemeal accuracies (acc_π) respectively for 5-fold cross validation, while the eighth and the ninth columns show the one-step (acc_1) and piecemeal accuracies (acc_π) respectively on the independent testing dataset. The improvement of piecemeal over one-step is shown in the fifth column. We conducted the statistical significance testing with the null hypothesis that the usual one-step accuracies and the two-step piecemeal imputation have equal mean accuracies. With Beagle, the p-values for the three 5-fold cross validation experiments are 0.0215, 0.0005 and 0.0004, respectively, indicating that the improvements by the two-step piecemeal imputation are statistically significant; with FImpute, the corresponding p-values are 0.64, 0.49 and 0.61 suggesting statistically insignificant improvements. Analogous results on the Holstein datasets of 114 animals are presented in Table 4.4. In Table 4.4 Columns 3 through 7 show a $1.5-3.0\%$ improvement net accuracy improvement in cross-validation with Beagle (the statistical significance testing p-values are 0.00369, 0.00003 and 0.00019, respectively) and a $0.5-1.0\%$ net accuracy improvement against FImpute (p-values 0.54, 0.38 and 0.31, respectively).

From 6 K to 50 K, 5 to 100 marker clusters, in increments of 5, were examined and the best piecemeal imputation results are included in the table, while in Figure 4.3 all of these accuracies are plotted (blue dots). From 6 or 50 K to 660 K, 100 to $1{,}000$ marker clusters, in increments of 100, were examined. We conducted the experiments on *bovine* two separate sequence datasets of 114 Holstein animals and 82 Simmental animals respectively in two chromosomes (BTA 14 and BTA 27) through a 5-fold cross validation training process. They also served as reference samples in all the independent testing experiments. BTA 27 was used for the Holstein data set and BTA 14 for the Simmental data set. On BTA 27, the 114 sequenced Holstein animals have genotype values for 529,674 SNPs. The Illumina 777 K chip contains $10,219$ of them, among which 664 are included in the 50 K chip, and 119 of these 664 SNPs are included in the 6 K chip. On BTA 14, the 82 sequenced Simmental animals have genotype values for $933,833$ SNPs. The Affymetrix 660 K chip contains $14,367$ of these $933,833$, among which $1,618$ are included in the 50 K chip, and a further 219 of these $1,618$ SNPs are included in the 6 K chip. Additionally, 390 Simmental animals from the Canadian Cattle Genome project genotyped in the Affymetrix 660 K chip were used for independent testing for the quality of the selected markers from the training process.

We used 5-fold cross validation to empirically examine our piecemeal imputation method, also to train the staircase pseudo arrays to impute the genotyped animals to their whole chromosome. The cross validation results also suggest the possible levels of improvement compared with the one-step imputation.

		5-fold cross validation					Independent testing		
Program	Imputation	acc_1	acc_π	$+$	#Clusters	#TClusters	acc_1	acc_π	$+$
Beagle	6K → 50K	69.35	70.81	$\mathbf{1.46}$	100	100	60.68	61.39	$\mathbf{0.71}$
	6K → 660K	72.37	74.92	$\mathbf{2.55}$	800	800	66.00	67.76	$\mathbf{1.76}$
	50K → 660K	86.61	88.89	$\mathbf{2.28}$	1000	1000	72.83	74.11	$\mathbf{1.29}$
FImpute	6K → 50K	75.95	76.70	$\mathbf{0.75}$	55	55	61.87	62.16	$\mathbf{0.29}$
	6K → 660K	79.11	80.11	$\mathbf{1.00}$	1000	1000	68.43	68.95	$\mathbf{0.52}$
	50K → 660K	90.31	90.74	$\mathbf{0.43}$	1000	1000	77.11	77.33	$\mathbf{0.22}$

Table 4.3: Accuracy comparisons between the two-step piecemeal and the classic one-step imputation on the Simmental datasets. Results are for markers on chromosome 14. Columns 3 through 7 contain the 5-fold cross validation results on the 82 animals, with the selected markers and their associated target marker clusters. Independent testing results on the 367 animals are in columns 8 –10, using the selected markers and their associated target marker clusters from the cross validation. In the independent testing from 50K to 660K, 8 markers of the Affymetrix 660K chip were filtered out due to their genotype disagreeing with the alternating alleles specified by sequencing, and consequently only 999 target marker clusters were used. The columns labelled with $+$ show the improvements, in bold, of the piecemeal imputation over the one-step imputation.

	5-fold cross validation						Independent testing		
Program	Imputation	acc_1	acc_π	$+$	#Clusters	#TClusters	acc_1	acc_π	$+$
Beagle	6K → 50K	86.98	89.81	$\mathbf{2.87}$	95	89	74.97	76.90	$\mathbf{1.94}$
	6K → 777K	82.35	85.27	$\mathbf{2.92}$	1000	963	71.29	73.25	$\mathbf{1.96}$
	50K → 777K	93.09	95.16	$\mathbf{2.07}$	1000	956	82.27	84.25	$\mathbf{1.97}$
FImpute	6K → 50K	91.11	91.64	$\mathbf{0.53}$	95	88	81.85	81.40	$\mathbf{0.25}$
	6K → 777K	89.22	90.14	$\mathbf{0.92}$	1000	942	82.80	82.81	$\mathbf{0.02}$
	50K → 777K	95.25	95.61	$\mathbf{0.36}$	800	765	87.72	87.83	$\mathbf{0.11}$

Table 4.4: Accuracy comparisons between the two-step piecemeal and the classic one-step imputation on the Holstein datasets. Results are for markers on chromosome 27. Columns 3–7 contain the 5-fold cross validation results on 114 animals, with the selected markers and their associated target marker clusters. Independent testing results on the 8 animals are in columns 8–10, using the selected markers and their associated target marker clusters from the cross validation. In the independent testing for Beagle 6, 37, and 44 target marker clusters are empty; for FImpute 7, 58, and 35 target marker clusters are empty. The columns labelled with $+$ show the improvements, in bold, of the piecemeal imputation over the one-step imputation.

4.2.5 Independent Testing

Independent testing examines the quality of the selected markers and the defined pieces learned from the training step. The study samples used in the testing are not involved in the training step. The piecemeal imputation accuracies are again compared to the corresponding one-step imputation accuracies, respectively. Columns 8–10 of Table 4.3 contain independent testing results on the 367 genotyped Simmental animals where the lower density represents either 6 or 50 K and the higher density refers to either 50 or 660 K (the second column). The 8th and 9th columns hold the one-step and piecemeal imputation accuracies (acc_1 and aac_π), respectively. The improvement of the piecemeal over the one-step is shown in the tenth column. For each imputation setting, the selected markers and the defined pieces are taken from the respective cross validation experiment. One exception is that there are 8 markers in the Affymetrix 660 K chip for which the two alleles (i.e. nucleotides) do not agree with the alternating alleles identified through genome sequencing; these 8 markers were excluded and one target marker cluster was discarded in the testing.

Analogous independent results on the 8 genotyped Holstein animals genotyped on Illumina 777K are shown in Table 4.4. Both tables show an accuracy improvement in all settings, though the improvement is about 40% lower than the 5-fold cross validation. 367 Simmental animals genotyped with Affymetrix 660K were used for the independent testing in Table 4.3.

4.2.6 Multi-Step Imputation: Independent Testing

With the selected markers and their associated target marker clusters from the training step, we experimented with the usual two-step imputation from 6 K to 50 K to 660 K on the 367 genotyped Simmental animals, and the four-step piecemeal imputation from 6 to 660 K. The four-step piecemeal imputation is a result of replacing each usual one-step imputation by a potentially promising two-step piecemeal imputation. The usual two-step imputation accuracy is denoted as acc_2; the four-step piecemeal imputation accuracy is still denoted as acc_π. Similar experiments were done on the 8 Holstein animals on BTA 27 genotyped using the 777 K chip.

For the 23 genotyped and sequenced Simmental animals, we experimented with the usual two-step imputation from 50 to 660 K to Sequence and the four-step piecemeal imputation from 50 K to Sequence, and the usual three-step imputation from 6 K to 50 K to 660 K to Sequence and the five-step piecemeal imputation from 6 K to Sequence. Here "Sequence" refers to all the $529,674$ SNPs on BTA 14. The usual three-step imputation accuracy is denoted as acc_3; the five-step piecemeal imputation accuracy is denoted as acc_π. Note that since we do not have a 660 K to Sequence training step to select markers (because first the usual one-step imputation is very good leaving little room for further improvement and second the training phase requires storage beyond our capacity), the last step in the five-step piecemeal imputation is a direct one-step imputation. All these usual two/three-step imputation accuracies and the corresponding four/five-step piecemeal imputation accuracies are summarized in Table 4.5, where there is accuracy improvement in all settings. We note that these 23 animals were used in the training step, and thus the results reported here could be slightly biased.

Program	Imputation	acc_1	acc_2	acc_3	acc_π	+
Beagle	8 Holstein BTA 27	71.29%	74.25%		74.43%	0.18%
FImpute	6 K →50 K →777 K	82.80%	82.74%		82.92%	0.18%
Beagle	367 Simmental BTA 14	66.00%	65.51%		66.59%	1.08%
FImpute	6 K →50 K →660 K	68.43%	68.54%		68.56%	0.02%
Beagle	23 Simmental BTA 14	84.91%	89.88%		90.17%	0.29%
FImpute	50 K →660 K →Sequence	87.95%	90.47%		90.50%	0.03%
Beagle	23 Simmental BTA 14	81.19%		83.94%	86.26%	2.32%
FImpute	6 K →50 K →660K →Sequence	82.23%		84.58%	84.67%	0.09%

Table 4.5: Results are on the Holstein datasets for markers on BTA27 and for the Simmental datasets for markers on BTA14, respectively. 8 Holstein and 367 Simmental genotyped animals are used in the two-step independent testing (6K→50K→HD), with results in columns 4, 6 and 7. The piecemeal imputation uses the selected markers and their associated target marker clusters from the training step. Additional 23 Simmental sequenced and genotyped animals are used in the two/three-step imputation to sequence (50K→660K→sequence, 6K→50K →660K→sequence). All one-step imputation accuracies are included in column 3. The last column labelled with + shows the improvements, in bold, of the piecemeal imputation over the two- or three-step imputation.

4.3 Discussion

4.3.1 Rationale Behind the Two-Step Piecemeal Imputation

Several recent studies in cattle have shown that two-step imputation can be more accurate than the classic one-step imputation [55, 60, 97]. A possible explanation for this phenomenon is likely due to frequency-based imputation algorithms for phasing correct haplotypes when there are multiple possible matches between the LD and HD panels whereas there are fewer matches once an intermediate panel is added in between [97]. The role the intermediate panel (in our case, each added marker) plays is similar to “long range phasing”; that is, markers in the intermediate panel encourages selecting long range haplotype over short haplotypes in the first step, reflecting closely related individuals should be considered than distant ones, thus improving the accuracy of imputation. Also in our preliminary study, we observed that some markers in the add-one two-step imputation experiments are able to boost the overall accuracy. These results have led us to put efforts into finding a set of markers that would perform the best in the subsequent two-step imputation. However, such an optimal set of markers is not assayed in any existing chips,

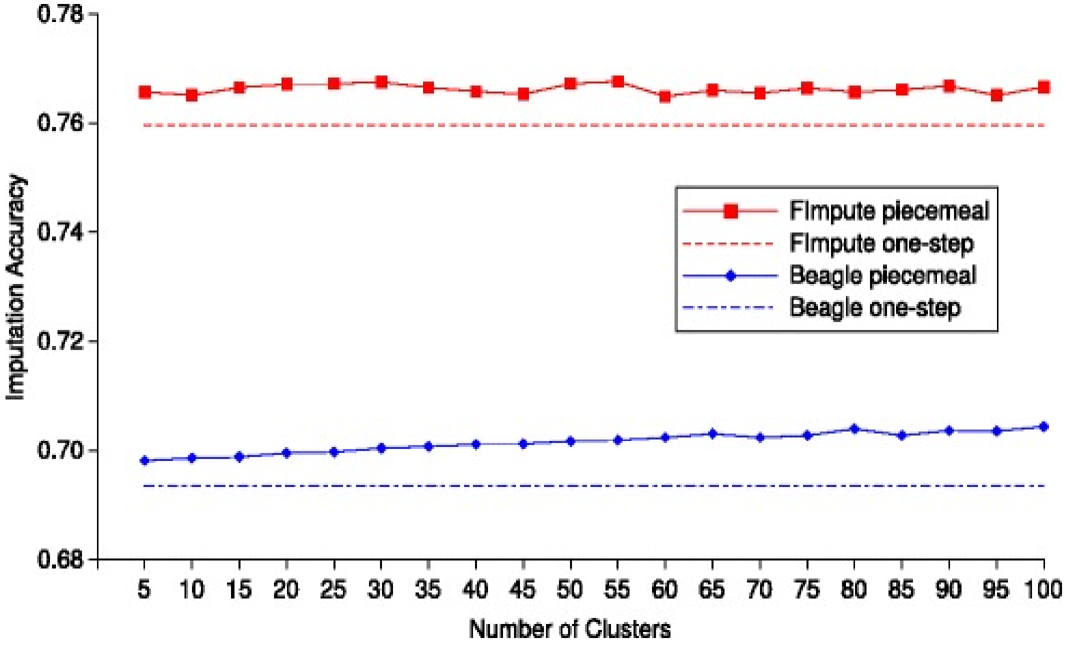


Figure 4.3: The Beagle/FImpute-based two-step piecemeal imputation accuracies against the number of SNP clusters

nor easy to obtain in reasonable computational time. Besides the selection scheme in our piecemeal imputation framework, we also tried several other approaches including sequential forward selection, which did not result in any significant improvement. We thus proposed an alternative to partition the higher density SNP set into multiple pieces, which are learned through the add-one two-step imputation experiments. Each piece is then imputed by the corresponding add-one two-step imputation experiment. This procedure laid the foundation for our two-step piecemeal imputation strategy. Nevertheless, our partition scheme is not necessarily optimal, as we adopted the k-means only because it outperformed other clustering methods slightly. In addition, we also experimented with the linkage-disequilibrium (LD) blocks produced by Haploview [2] for finding closely linked markers, but again the increase in accuracy was insignificant and the results are often inferior to those of the our marker selection scheme (detailed results not shown).

4.3.2 Marker Clusters and Their Effects

From our 5-fold cross validation results, it seems as though the number of marker clusters does not affect the final piecemeal imputation accuracy much. For example, for genotype imputation from 6 to 50 K on the Simmental dataset, the piecemeal imputation accuracies of all the 20 different clustering results are plotted in Figure 4.3, where the dashed blue/red lines are the Beagle/FImpute one-step imputation accuracies, and the solid dots represent the two-step piecemeal imputation accuracies. Despite FImpute performing better than Beagle, the connected dots for both FImpute and Beagle do not vary much with different numbers of clusters. A simple guideline would be to have an average cluster size of 10 – 100. We also look into the content of a marker cluster. For example, when $k = 15$, the first five of the 15 marker clusters are plotted in Figure 4.2 where the x-axis represents the physical locus. It is interesting to see that the markers of a cluster are not necessarily close to each other, though they have very similar imputation potentials. The LD between pairs of these markers, by Haploview, are insignificant.

4.3.3 Imputation Result Sensitivity to the Selected Markers

The imputed genotype for the study samples at a selected marker m_i is used in the second step, of the two-step piecemeal imputation, to impute the other untyped markers of $\mathcal{U} - \{m_i\}$. Comparing the add-one two-step imputation result to the usual one-step imputation, we have seen subtle changes at many untyped markers of $\mathcal{U}$ for different selected markers. Indeed, some of them exhibit a gain in accuracy whereas some have a loss in accuracy and yet others are unaffected. This has led us to use the overall gain in accuracy to measure the imputation potential of a selected marker.

By setting up a feature vector for a candidate marker to keep a record of the accuracy gains and losses at each untyped marker, we observed from our preliminary two-step imputation experiments (results not shown) that the candidate markers fall into three categories when used for creation of the pseudo-array: 1) those that yield an accuracy gain over the usual one-step imputation; 2) those that yield a net zero gain; 3) those that yield an accuracy loss from the usual one-step imputation. Through clustering these feature vectors, the impact of selecting different markers from a cluster is expected to be reduced to the minimum, as evidenced by our preliminary experiments (we did not re-examine this issue in all the experiments reported here).

4.3.4 Target Marker Clusters

All the markers from the same cluster have similar effects on accuracies of imputation at untyped markers when they are used to create the intermediate pseudo-array and impute in the add-one two-step procedure. Markers along the genome where the added markers from the same cluster unanimously perform better than the one-step imputation form the target markers associated with the cluster. We have looked into the content of such a target marker cluster. Similar to a marker cluster, it is interesting to see that the markers of a target cluster are not necessarily physically close to each other, nor are the LD between pairs of these markers by Haploview significant.

It is also interesting to observe that some target marker clusters are overlapping. Note that target clusters are formed after the marker clusters are determined, that is, in terms of the feature vectors, the marker clusters are formed using the whole vectors, but the target marker clusters are formed by using only the vector entries corresponding to the makers in a marker cluster. Therefore, such a phenomenon of an untyped marker being imputed with high accuracies by several selected markers can be explained.

4.3.5 Other Clustering Methods

The main reason for marker clustering is to avoid selecting redundant markers to form the intermediate pseudo array, here redundant means the similar potential in imputing the genotype for other SNPs. We had experimented with Haploview to construct the LD blocks for this purpose, which did not result in any conclusive accuracy increase (detailed results not shown). Other popular feature selection methods in machine learning, such as SFS and SBS, were also tested. Based on the feature vectors, we tried clustering methods other than k-means, with results not better than k-means. Thus we go with k-means in the final piecemeal imputation framework.

As discussed in the last paragraph, forming the marker clusters and the associated target marker clusters is more like a bi-clustering task, and it would be worthwhile to try some good bi-clustering algorithms. Coming back to the LD-based marker selection, though multiple experiments with different thresholds in Haploview did not give good results, we realize that such an approach avoids the add-one two-step imputation experiments in the training phase, and it can be substantially faster. This suggests the need for better LD block estimation/prediction by SNP genotype values.

4.3.6 Cattle Genomic Distance

In our current empirical experiments, we used the population-based option in our base programs. The underlying assumption for such an option is that individuals are unrelated. On the other hand, related animals can certainly bias towards the correct genotype. Therefore, if one would be able to define a degree of relatedness between two individuals based on their SNP genotype, then using only closely related sequenced animals to a study animal as references may potentially lead to more accurate genotype imputation.

Animals from different breeds are deemed more distantly related than the same breed animals. We therefore separated the datasets by breeds. In fact, earlier research suggests cattle whole genome SNP genotype imputation should be done breed by breed [60], which is also confirmed by our preliminary testing that inter-breed imputation has slightly lower performance (detailed results not shown).

4.3.7 Computational Time

The running time of the two-step piecemeal imputation depends on the number of study animals, the number of reference animals, and the number of SNPs. We were able to use the high-performance computing facilities and partition, and submit the experiments in parallel. The most time-consuming stage in the two-step piecemeal imputation is the training phasing, when the add-one two step imputation was performed to evaluate and select potential good markers. The major challenge is the need for a huge disk storage (more than 84 TB) when we were performing whole-genome SNP genotype imputation for the training phase. We used more than 3TB for storing all the intermediate data. The imputed SNP genotype values are expected to be useful in the downstream data analysis, such as genomic predictions, and thus the increased computation burden in the piecemeal imputation framework becomes worthy.

Chapter 5

A Statistical Model for Population Based Genotype Imputation

In this chapter, we introduce a statistical model based on Li and Stephens' "PAC" framework [61] for population based genotype imputation. This chapter contains a result we have made great efforts in, but turns out not completely successful. The "PAC" framework laid groundwork for many successful methods applied to a wide range of problems including phasing haplotypes (SHAPEIT 1 [27], SHAPEIT 2 [77]), inferring population structures (Structure 2.0 [31]), estimating recombination and ancestry reconstruction in admixed populations (HAPMIX [81], ELAI [39]), as well as genotype imputation (Impute 1 [69], Impute 2 [51], MaCH [62], fastPHASE [90], Bimbam [40] and BLIMP [105]), and it is considered to be major breakthrough that incorporates the biological concepts of "mutation" and "recombination" into a hidden Markov model (HMM) and used a "copying" process for approximating construction of a new haplotype from existing observed haplotypes. Mutations are modelled as copying errors and recombinations correspond to a switch of hidden states between two linked loci.

In Chapter 2, we reviewed several influential genotype imputation methods built on the "PAC" model and pointed out computational issues of scalability associated with the HMM framework. Under the scenario that we have a reference panel consisting of tens of thousands of animals and the "PAC" framework, the number of hidden states representing the origin of the two alleles of genotypes at each locus becomes large, and the running time would grow quadratically. Aforementioned continuous HMM-PAC methods including Bibam and fastPHASE that employ the idea of local clustering and EM for parameter estimations sometimes stuck with local maxima leading to poor imputation accuracies. We developed a statistical model that can circumvent the shortcomings of existing "PAC" models. Specifically, we are interested in developing a statistical HMM model that incorporates the idea of local clustering to reduce the number of hidden states at each loci, uses discrete genotype values to represent the "copying" process and builds on the existing "PAC" fundamentals. Our statistical model addresses "population-based" genotype imputation that uses "unphased" genotype data free of pedigree information as reference with the goal of imputing study data genotyped in low-density chip up to medium- (MD) and high-density (HD) levels.

Modern high throughput genotyping and sequencing technologies do not produce haplotype data directly but the combined sum of alleles (known as "genotypes") at tens of thousands of dense loci. Obtaining accurate estimation of haplotypes from high-density reference genotype data is considered to be a key step for the success of genotype imputation [69, 68]. In humans, the 1000 Genomes Project [21] and the legacy International HapMap Project [22] provide public accessible accurate haplotype for reference. These dense haplotypes were obtained from running softwares such Impute 2 [69, 51, 50]. However, in cattle and other livestock animals, due to privacy, data ownership and policy governing data sharing, animal data from unphased HD genotyping chips and MD genotyping chips are usually used as reference panels.

5.1 Hidden Markov models

Since its initial introduction in the 1960s and the 1970s, the HMM has gained successes in speech signal processing [1], speech recognition [83], image analysis [85] as well as genetic studies [24, 61]. In genotype imputation, we have as observed input genotype sequences spanning M loci $G = (g_1, \cdots, g_n)$, where $g_i = (g_{i1}, \cdots, g_{iM})$ and $g_{im} \in \{0, 1, 2\}$ and would like to recover a hidden sequence of "haplotypes" $Z = (Z_1, \cdots, Z_M)$ of the same length as G, which specifies that the tagging or the origin of the two alleles of genotypes at each locus. Each hidden state $Z_m = (Z_m^1, Z_m^2)$ is a pair, where Z_m^i can take a set of values $\{1, \cdots, N\}$, and in our example denotes from which haplotype each allele of a genotype is copied. The hidden sequence follows a Markov chain defined by initial state probabilities $P(Z_1 = (i, j))$ and the transition probabilities between two successive hidden states $P(Z_{m+1}|Z_m)$ at locus m and locus $m+1$, where $1 \leq i \leq N$ and $1 \leq j \leq N$. Additionally, there is a set of emission probabilities $P(G_{im}|Z_m)$, each of which defines the probability of observing G_{im} at a particular locus m given the state of the hidden variable Z_{im} at that time.

Once we have an HMM with its set of parameters (initial probabilities, transition probabilities and emission probabilities) denoted λ, there are three problems of interest.

- The Evaluation Problem: Given an HMM λ and a sequence of observations G, what is the probability that the observations are generated by the model, $P(G|\lambda)$?

- The Decoding Problem: Given an HMM λ and a sequence of observations G, what is the most likely state sequence in the model that generated the observations?

- The Learning Problem: Given an HMM λ and a set of observed sequences G, how do we set the model's parameters to maximize the probability of generating those sequences $P(G|\lambda)$?

The evaluation and learning problems can be solved using the recursive dynamic programming (DP) based "forward-backward" algorithms , whereas the decoding problem can be solved using a DP-based "Viterbi algorithm."

5.2 Methods

5.2.1 Notations and Background

In order to describe the statistical HMM model, we briefly re-introduce the key parameters, notations and the problem we attempted to target. Since bi-allelic markers are assumed throughout the dissertation, two alleles can be represented as "0" and "1" arbitrarily at each locus m. We have as input two genotype datasets, a reference panel $\mathcal{R}$ of N individuals genotyped in a high-density chip, denoted as $DG = \{DG_1, \cdots, DG_N\}$, where $DG_i = (DG_{i1}, \cdots, DG_{iM})$ defines a vector of genotypes for individual i over M loci and DG_{im} takes values from the set $\{0, 1, 2\}$ if observed or $DG_{im} = ?$ if missing or untyped, and a study sample $SG = \{SG_1, \cdots, SG_D\}$ of D individuals genotyped in a low-density chip, where there are many untyped genotypes $SG_{im} = ?$ in SG and the task of genotype imputation is to infer those "untyped" genotypes for SG. DG and SG share a set of "typed" markers, denoted $\mathcal{T}$. For a locus $m \in \mathcal{T}$, DG_{im} and SG_{jm} are typed meaning $SG_{jm} \in \{0, 1, 2\}$ and $DG_{im} \in \{0, 1, 2\}$. We would like to infer untyped genotypes in a set $\mathcal{U} = \mathcal{M} - \mathcal{T}$ of loci that are typed only in DG but untyped in SG, where $\mathcal{M}$ is the entire set of M markers.

Additionally, we have as an input a fine-scale genetic map $\rho = (\rho_1, \cdots, \rho_{M-1})$ where ρ_m defines the probability of recombination occurring between two consecutive loci (locus m and locus $m + 1$). In human species, genetic map can be downloaded from the HapMap project and in this dissertation, we used an approximation to calculate ρ_m for every two consecutive markers m and $m + 1$ in our cattle data. Physical locations of all markers are available and can be looked up in reference to a genome assembly. In genetics, a centimorgan (cM) is a unit for measuring the probability of two markers to be inherited together during the meiosis of sexual reproduction. In general, if two markers are distant apart, then it will be more likely that they get separated by recombination events. Since on average one centimorgan corresponds to about 1 million base pairs in cattle, how many cMs (termed "genetic distance", denoted by d_m) two markers are apart can be approximated by the difference of their physical locations divided by $1,000,000$. The Haldane model (1919) is further applied in this study for obtaining the probability of recombination occurring between two loci (also known as the"recombination rate"). Haldane assumes that recombination follows a Poisson process. That is, recombination takes place if there is an odd number of crossovers between two loci and no recombination if there is an even number of crossovers in between. Therefore, the probability that recombination takes place can be calculated via the formula

$$\rho_m = \frac{1}{2}(1 - e^{-2d_m}),$$

where d is the genetic distance in cM between marker m and marker $m + 1$ by finding the probability of an odd number of crossovers in a given interval length d_m in a Poisson process rate 1. As one can see from the formula, ρ_m is always a fraction between 0 and 0.5.

In order to infer missing values in SG, we need to obtain a set of accurate phased haplotypes DH derived from the reference panel DG and also restrict the number of values each hidden state can take at each locus. Inspired by the idea of local clustering in fastPHASE [90], we propose a localized haplotype cluster model H for representing the phased haplotypes DH derived from DG. H is a connected graph with following properties, see Figure 5.1:

- The graph is leveled with M levels, one level per locus.

- At each locus m, there are two nodes for representing two local clusters $H_m^{(0)}$ and $H_m^{(1)}$, corresponding to the two alleles "0" and "1".

- There are edges connecting the nodes between two consecutive loci m and $m + 1$ and on the edge there are weights keeping track of the number of estimated haplotypes DH that traverse the edge between nodes $H_m^{(k_1)}$ and $H_{m+1}^{(k_2)}$, denoted by $c_m(k_1, k_2)$. Weights on the edges are used to derive the transition probabilities in HMM.

- Likewise associated with each node (aka cluster) is the cluster frequency $w_m^{(k)} = c_m(k)/(2 * N)$ at the locus m, where $k = 0, 1$, keeping track of the number of haplotypes DH that traverses the node, denoted by $c_m(k)$.

- for each estimated haplotype h, there exists a unique path $Z = (Z_i, \cdots, Z_M)$ traversing on one node at each level and one edge between two consecutive loci. Each $Z_m = 0$ represents that the allele at locus m for h is copied from cluster $H_m^{(0)}$ and $Z_m = 1$ represents that the allele at locus m for h is copied from cluster $H_m^{(1)}$.

In Figure 5.1, we have nodes in circles for hidden states, representing the clusters at each loci and edges connecting two adjacent loci where weights indicate how many haplotypes of DG traverse such an edge. The nodes representing a local cluster for "allele 0" is also capable of emitting an allele "1" due to mutation with a small probability. Therefore, each hidden state can emit two alleles shown in two arrows on top or bottom of the node.

From the "PAC" framework, we can approximate a haplotype $P(h|H, \rho) = \sum_Z P(h|Z, \rho)P(Z|H, \rho)$using an HMM. The underlying assumption is that the allele of h_m at each locus m originates from one of the two defined clusters and Z_m specifies the cluster from which the allele h_m is copied. Then, the initial state probability is given by

$$P(Z_1 = k) = w_m^{(k)} = \frac{c_m(k)}{2\dot{N}}, k \in \{0, 1\}$$

and following SHAPEIT 1 [27], we define the transition probabilities between two consecutive loci by taking into account recombination events:

$$P(Z_{m+1} = k_2|Z_m = k_1) = (1 - \rho_m)\frac{c_m(k_1, k_2)}{c_m(k_1)} + \rho_m w_{(m+1)}^{(k_2)}, k_1, k_2 \in \{0, 1\}.$$

If no recombination event occurs with probability $1 - \rho_m$, the haplotype h would traverse the edge connecting nodes $H_m^{(k_1)}$ and $H_{m+1}^{(k_2)}$ and the transition probability incident on this edge is exactly $\frac{c_m(k_1,k_2)}{c_m(k_1)}$. If a recombination occurs, it happens with probability ρ_m and the quantity $w_{(m+1)}^{(k_2)}$ tells us how likely h traverses on the node $H_{m+1}^{(k_2)}$. The emission probability $P(h_m|Z_m = k)$ at each locus is given by

$$p(h_m|Z_m = k) := \begin{cases} \mu, & h_m = H_m^{(k)} \\ 1 - \mu, & h_m \neq H_m^{(k)} \end{cases}$$

The term $P(h_m|Z_m = k)$ concerns how the observed allele h_m will be close to but not exactly the same as the allele $H_m^{(k)}$ in cluster k being copied, and μ is the mutation rate that models the probability of a mutation that changes the copied allele $H_m^{(k)}$ to its complementary alleles [69]. Following Impute 1 [69] and the "PAC" model [61], we assume that the mutations are independent across loci and the formula for mutation rate is given by $\mu = \frac{\theta}{2(\theta + 2\dot{N})}$, where $\theta = (\sum_{i=1}^{2N-1} \frac{1}{i})^{-1}$.

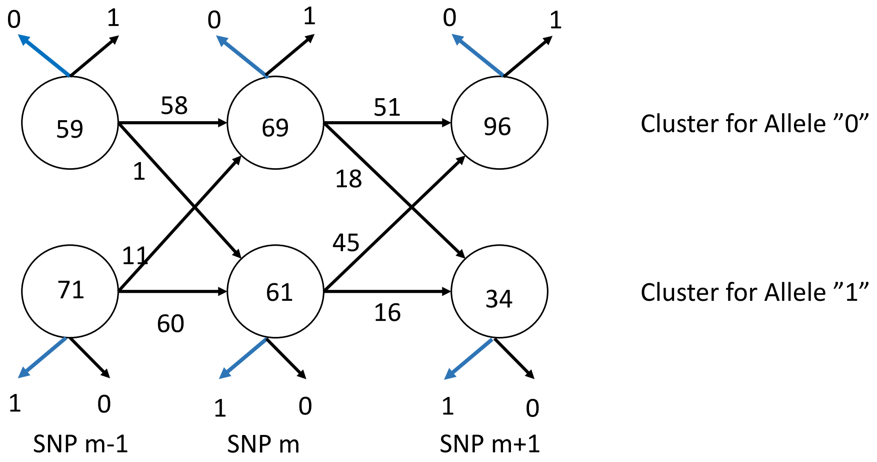


Figure 5.1: An illustration of localized haplotype cluster hidden Markov model for population based genotype imputation over three consecutive loci $m-1$, m and $m+1$. Each circle is a hidden state and represents a localized cluster for either allele “0” or allele “1”. Two arrows on top/bottom of each hidden state at each SNP represent the possible emissions. For hidden state that represents the local cluster for allele “0”, there is a higher chance of emitting allele “0” (in bold blue) and a lower chance of emitting allele “1” due to mutation. Edges between hidden states from one locus to the next are transitions of the HMM and the numbers incident on the edges and nodes count how many haplotypes in DG traverse it.

Once we have an accurate phasing H derived from DG and the genetic map ρ, forward-backward algorithms described in Li and Stephens [61] can be employed for computing $P(h|H, \rho) = \sum_Z P(h|Z, \rho)P(Z|H, \rho)$ with the above defined parameters. This concludes our introduction to notations and the "PAC" model for haplotype.

5.2.2 Extended HMM for Inferring Untyped Genotypes in SG

Next, instead of observing haplotype h, we assume that a study sample consisting of D individuals are observed. We wish to impute the untyped genotypes provided that we have our localized haplotype cluster model H derived from DG, mutation rate μ and our genetic map ρ. Since we are dealing with population data, we can assume that individuals in SG are independent and identically distributed (i.i.d) and by looking up the dependency structures between different variables in Figure 5.2, we can express the probability of SG conditional on the untyped reference data DG, the mutation rate μ and ρ as follows:

$$P(SG|DG, \mu, \rho) = \prod_{i=1}^{D} P(SG_i|DG, \mu, \rho),$$

and by Bayesian network in directed acyclic graph, one can factor a joint distribution into a product of conditional distributions,

$$\begin{aligned} P(SG_i|DG, \mu, \rho) &= \sum_H P(SG_i, H|DG, \mu, \rho) \\ &= \sum_H P(SG_i|H, \mu, \rho)P(H|DG, \mu, \rho) \end{aligned}$$

The first question we are interested in addressing is how we can compute $P(SG_i|H, \mu, \rho)$ efficiently if we obtain a good estimation H derived from DG using our genetic map ρ and defined mutation rate μ. The key idea is to extend the HMM for the haplotype copying process of the "PAC" model into an HMM for genotype. Each diploid individual SG_i in the study sample carries two copies of alleles (represented as an unordered pair of "haplotypes") at each locus, one from each of their parents, to form a long vector $SG_i = (SG_{i1}, \cdots, SG_{iM})$ over M loci. Two sequences of hidden states $Z_i^{(1)}, Z_i^{(2)}$ over M loci can be employed for computing $P(SG_i|H, \mu, \rho)$

in the HMM for genotype as follows

$$\begin{aligned} P(SG_i|H,\mu,\rho) &= \sum_{Z_i^{(1)},Z_i^{(2)}} P(SG_i, Z_i^{(1)}, Z_i^{(2)}|H,\mu,\rho) \\ &= \sum_{Z_i^{(1)},Z_i^{(2)}} P(SG_i|Z_i^{(1)}, Z_i^{(2)}, H,\mu,\rho)P(Z_i^{(1)}, Z_i^{(2)}|H,\mu,\rho) \end{aligned}$$

Again, the two hidden states $Z_{im}^{(1)}, Z_{im}^{(2)} \in \{0,1\}$ indicate the cluster origins of the two alleles for each genotype SG_{im}. The initial state probabilities for the extended HMM are given by:

$$P(Z_{1m}^{(1)} = k_1, Z_{1m}^{(2)} = k_2|H,\mu,\rho) = w_1^{(k_1)}w_1^{(k_2)}.$$

The transition probabilities for the extended HMM are given by

$$\begin{aligned} &P(Z_{i(m+1)}^{(1)} = k_3, Z_{i(m+1)}^{(2)} = k_4|Z_{i(m)}^{(1)} = k_1, Z_{i(m)}^{(2)} = k_2, H,\mu,\rho) \\ &= P(Z_{i(m+1)}^{(1)} = k_3, |Z_{i(m)}^{(1)} = k_1, H,\mu,\rho)P(Z_{i(m+1)}^{(2)} = k_4|Z_{i(m)}^{(2)} = k_2, H,\mu,\rho), \end{aligned}$$

where the transition probabilities $P(Z_{i(m+1)}^{(1)} = k_3, |Z_{i(m)}^{(1)} = k_1, H,\mu,\rho)$ are exactly the transition probabilities defined for haplotype HMM earlier.

Next, the emission probability $P(SG_{im}|Z_{i(m)}^{(1)} = k_1, Z_{i(m)}^{(2)} = k_2, H,\mu,\rho)$ at each locus can be looked up in Table 5.1. The initial state probabilities, transition probabilities, together with emission probabilities complete our definition of genotype-based HMM parameters H.

Our forward and backward algorithms are variants of the ones documented in Scheet and Stephens [90]. Let K be the number of clusters at each locus and by definition H of localized haplotype cluster HMM, $K = 2$. Our algorithm runs in $O(K^4 \cdot M)$ as opposed to $O(K^2 \cdot M)$ in fastPHASE's implementation. However many individuals there are in DG, in the derived HMM graph representation H, $K = 2$ guaranteed the running quartic in the number of clusters, and it was not an issue. Also, when genotype imputation was performed with this statistical model, special care must be taken to remove all homozygous loci from $\mathcal{T}$ as our computation of transition probabilities relied on the counts of haplotypes traversing on any node and we need to make sure that denominator $c_m(k_1)$ is non-zero.

$H_m^{(k_1)}$		$SG_{i(m)}$	
$+H_m^{(k_2)}$	0	1	2
0	$(1-\mu)^2$	$2\mu(1-\mu)$	μ^2
1	$\mu(1-\mu)$	$\mu^2+(1-\mu)^2$	$\mu(1-\mu)$
2	μ^2	$2\mu(1-\mu)$	$(1-\mu)^2$

Table 5.1: Emission probabilities $P(SG_{im}|Z_{i(m)}^{(1)} = k_1, Z_{i(m)}^{(2)} = k_2, H, \mu, \rho)$ based on mutation rates and the observed genotypes.

In our forward algorithm 1, we try to compute the joint probability

$$\alpha_H^i(m, \{k_1, k_2\}) = P(SG_{i1}, \ldots, SG_{i(m)}, Z_{i(m)}^{(1)} = k_1, Z_{i(m)}^{(2)} = k_2|H, \mu, \rho).$$

For ease of presentation, we use the following short notations for expressing emission probabilities, initial state probabilities and transition probabilities.

- Emission probabilities $e^i(SG_{im}|k_1, k_2) = P(SG_{im}|Z_{i(m)}^{(1)} = k_1, Z_{i(m)}^{(2)} = k_2, H, \mu, \rho)$;
- Initial probabilities $p_1^i(k_1, k_2) = P(Z_{i1}^{(1)} = k_1, Z_{i1}^{(2)} = k_2|H, \mu, \rho)$;
- Transition probabilities $p_m^i(k_1 \to k_3, k_2 \to k_4) = p_m(k_1 \to k_3) \cdot p_m(k_2 \to k_4)$ which is simply just the term

  $$(P(Z_{i(m+1)}^{(1)} = k_3, |Z_{i(m)}^{(1)} = k_1, H, \mu, \rho) \cdot P(Z_{i(m+1)}^{(2)} = k_4|Z_{i(m)}^{(2)} = k_2, H, \mu, \rho).$$

In our backward algorithm 2, we try to compute the joint probability

$$\beta_H^i(m, \{k_1, k_2\}) = P(SG_{i(m+1)}, \ldots, SG_{i(M)}|Z_{i(m)}^{(1)} = k_1, Z_{i(m)}^{(2)} = k_2, H, \mu, \rho).$$

In case of encountering a missing genotypes in the execution of either forward or backward algorithm, use 1 as the value of the emission probabilities.

After running the forward-backward algorithm, for a missing/untyped genotype $SG_{im} =?$ in SG_i, one can obtain the marginal probability of

$$P(Z_{i(m)}^{(1)} = k_1, Z_{i(m)}^{(2)} = k_2|SG_i, H, \mu, \rho) \propto \alpha_H^i(m, \{k_1, k_2\})\beta_H^i(m, \{k_1, k_2\})$$

with the normalizing constraint

$$\sum_{k_1=0}^{1}\sum_{k_2=0}^{1} P(Z_{i(m)}^{(1)} = k_1, Z_{i(m)}^{(2)} = k_2|SG_i, H, \mu, \rho) = 1.$$

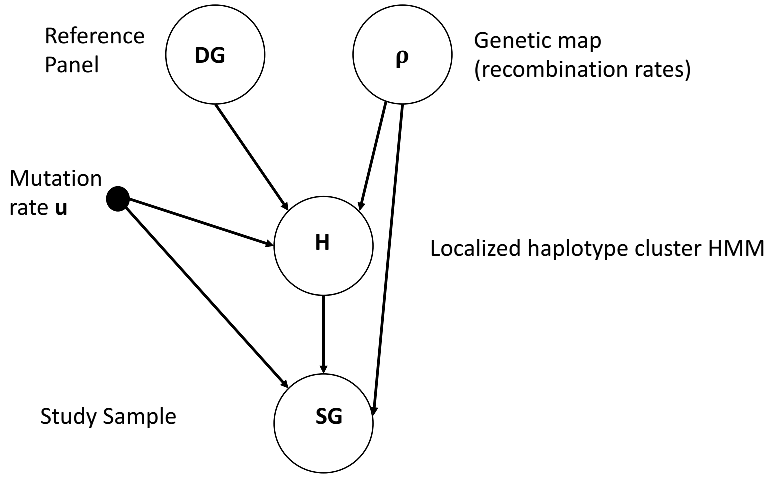


Figure 5.2: An illustration of the probabilistic graphical model (PGM) for population based genotype imputation. Each arrow indicates a dependency. SG depends on the localized cluster haplotype HMM H, mutation rate μ as well as the genetic that specifies recombination rates ρ.

Algorithm 1 The Forward Algorithm

```
Input a localized haplotype cluster HMM $H$, mutation rate $\mu$ and recombination rates $\rho$, a study sample $SG_i$
Output a matrix of joint probabilities $\alpha^i_H(m, \{k_1, k_2\})$ for $k_1, k_2 \in \{0, 1\}$ and $m = 1, \cdots M$.
for $k_1$ = 0 to 1 do
  for $k_2$ = 0 to 1 do
    initialize $\alpha^i_H(1, \{k_1, k_2\})$ at the first locus as follows:

    $\alpha^i_H(1, \{k_1, k_2\}) = P(SG_{i1}, Z^{(1)}_{i1} = k_1, Z^{(2)}_{i1} = k_2|H, \mu, \rho) \cdot P(Z^{(1)}_{1m} = k_1, Z^{(2)}_{1m} = k_2|H, \mu, \rho)$
    $= P(SG_{i1}|Z^{(1)}_{i1} = k_1, Z^{(2)}_{i1} = k_2, H, \mu, \rho) \cdot P(Z^{(1)}_{1m} = k_1, Z^{(2)}_{1m} = k_2|H, \mu, \rho)$
    $= e^i(SG_{i1}|k_1, k_2) \cdot w_1^{(k_1)} \cdot w_1^{(k_2)}$

  end for
end for

for $m$ = 1 to $M - 1$ do
  for $k_3$ = 0 to 1 do
    for $k_4$ = 0 to 1 do
      for $k_1$ = 0 to 1 do
        for $k_2$ = 0 to 1 do

          $\alpha^i_H(m + 1, \{k_3, k_4\})$
          $= e^i(SG_{i(m+1)}|k_3, k_4) \times \Big[\sum_{k_1=0}^{1} \sum_{k_2=0}^{1} p_m(k_1 \to k_3) \cdot p_m(k_2 \to k_4) \cdot \alpha^i_H(m, \{k_1, k_2\}))\Big]$

        end for
      end for
    end for
  end for
end for
```

Algorithm 2 The Backward Algorithm

```
Input a localized haplotype cluster HMM $H$, mutation rate $\mu$ and recombination
rates $\rho$, a study sample $SG_i$
Output a matrix of joint probabilities $\beta_H^i(m, \{k_1, k_2\})$ for $k_1, k_2 \in \{0, 1\}$ and
$m = 1, \cdots M$
for $k_1$ = 0 to 1 do
    for $k_2$ = 0 to 1 do
        initialize $\beta_H^i(M, \{k_1, k_2\}) = 1$ at marker $M$.
    end for
end for

for $m$ = $M - 1$ to 1 do
    for $k_1$ = 0 to 1 do
        for $k_2$ = 0 to 1 do
            for $k_3$ = 0 to 1 do
                for $k_4$ = 0 to 1 do

                    $\beta_H^i(m, \{k_1, k_2\})$
                    $= [\sum_{k_3=0}^{1} \sum_{k_4=0}^{1} e^i(SG_{i(m+1)}|k_3, k_4) \times p_m(k_1 \to k_3) \cdot p_m(k_2 \to k_4) \cdot \beta_H^i(m+1, \{k_3, k_4\}))]$

                end for
            end for
        end for
    end for
end for
```

Therefore,

$$P(Z_{i(m)}^{(1)} = k_1, Z_{i(m)}^{(2)} = k_2 | SG_i, H, \mu, \rho) = \frac{\alpha_H^i(m, \{k_1, k_2\})\beta_H^i(m, \{k_1, k_2\})}{\sum_{k_1=0}^{1}\sum_{k_2=0}^{1} \alpha_H^i(m, \{k_1, k_2\})\beta_H^i(m, \{k_1, k_2\})}.$$

Let $g \in \{0, 1, 2\}$, and we wish to choose a value g for genotypes SG_{im} in genotype sequence SG_i of individual i that maximizes

$$\begin{aligned}
&P(SG_{im} = g | SG_i, H, \mu, \rho) \\
&= \sum_{k_1=0}^{1}\sum_{k_2=0}^{1} P(SG_{im} = g, Z_{i(m)}^{(1)} = k_1, Z_{i(m)}^{(2)} = k_2 | SG_i, H, \mu, \rho) \\
&= \sum_{k_1=0}^{1}\sum_{k_2=0}^{1} P(SG_{im} = g | Z_{i(m)}^{(1)} = k_1, Z_{i(m)}^{(2)} = k_2, H, \mu, \rho) P(Z_{i(m)}^{(1)} = k_1, Z_{i(m)}^{(2)} = k_2 | SG_i, H, \mu, \rho)
\end{aligned}$$

The first term inside the summation is just the emission probability of observing a particular genotype g at the locus m and the second term is the result of computing the conditional distributions in forward-backward algorithms.

5.2.3 Estimation of Parameters in Localized Haplotype HMM H

Next we demonstrate how to compute the posterior probability $P(H|DG, \mu, \rho)$. Our aim is to estimate all the parameters associated with the localized haplotype HMM H from the observed high density genotype data DG, That is, we would like to obtain accurate phasings for our reference panel DG. If an individual's genotype sequence over M loci are all homozygous or has exact one heterozygous locus, then phasing of DG is trivial. For example, the genotype sequence 0020 can be phased as a pair of identical haplotype sequences in the form '0010; the genotype sequence 0012 over four loci can be phased as the pair of haplotypes 0001 and 0011. We say a pair of haplotypes $(h^{(1)}, h^{(2)})$ is compatible with an observed genotype sequence G over M loci, where $h_m^{(i)} \in \{0, 1\}$ denotes the allele at the m-th locus in the haplotype sequence $h^{(i)}$ and $G_m \in \{0, 1, 2\}$ is the corresponding genotype at the same locus, if and only if $h_m^{(1)} + h_m^{(2)} = G_m$. Homozygous genotype "0" or "2" can be phased with compatible identical alleles "$0|0$" and "$1|1$" in the localized haplotype HMM H. We update the counts in the corresponding nodes in H. Phasing of DG is in essence about updating the counts in edges between two consecutive nodes at levels m and $m+1$.

We achieve this via a Markov Chain Monte Carlo (MCMC) sampling from the conditional distribution of the haplotype, recombination rates and mutation rate. Let DH_i denote a pair of haplotypes (also known as "diplotype") that are compatible with DG_i. The MCMC sampling procedure starts with some random phasing DH_i that is compatible with DG_i. That is, at each heterozygous locus, the ordering of the alleles is randomly guessed, untyped genotypes (if any) are sampled according to the allele frequencies at the locus and weights of the edges are updated. We then perform a number of MCMC iterations. Each iteration updates phases of every reference diploid individual i (in some arbitrary order) in two steps: As we want to obtain $P(H|DG, \mu, \rho) \propto P(DG, H, \mu, \rho)$, which cannot be computed directly, we use the Gibbs sampler. In each iteration, phases DH_i of DG_i are updated in arbitrary order. H_{-i} keeps track of the current haplotypes of all individuals except i in the localized haplotype HMM. That is, H_{-i} contains the weights of the nodes and edges in the graph H by taking into account the current haplotype guesses of all individuals except i. Since the individuals in DG are assumed to be i.i.d, and it follows that $P(DH|DG, H, \mu, \rho) = \prod_{i=1}^{N} P(DH_i|DG_i, H, \mu, \rho)$. It follows from the fact that diplotype of each individual i can be sampled independently from $P(DH_i|DG_i, H_{-i}, \mu, \rho)$, which can be computed via the forward-backward

Algorithm 3 The Sampling Framework

```
Input: $DG$, $\rho$, $\mu$
Output: $H$

1. Start with some random phasing $DH_i$ that is compatible with $DG_i$ for individual $i = 1, \cdots, N$.
2. Update the counts on nodes and edges in the localized haplotype HMM graph $H$.
for iterations from 1 to 30 do
    for $i \in$ permutation of $\{1, \cdots, N\}$ do
        3. Sample a new pair of haplotypes $DH_i$ for reference individual $i$ from the conditional distribution $P(DH_i|DG_i, H_{-i}, \mu, \rho)$.
    end for
end for
```

algorithm introduced earlier. Accurate haplotypes of DG can then be obtained in small number of iterations typically less than 30 according to Impute 2 [51] and MaCH [62].

The detailed sampling procedure ("Step 3 in the Sampling framework") works as follows:

- sample $(Z^{(1)}_{iM} = k_1, Z^{(2)}_{iM} = k_2)$ from $P(Z^{(1)}_{iM} = k_1, Z^{(2)}_{iM} = k_2 | DG_i, H_{-i}, \mu, \rho) \propto P(Z^{(1)}_{iM} = k_1, Z^{(2)}_{iM} = k_2, DG_i | H_{-i}, \mu, \rho) = \alpha^i_{H_{-i}}(M, \{k_1, k_2\})$.

- sample $(Z^{(1)}_{im} = k_1, Z^{(2)}_{im} = k_2)$ recursively for loci $m = M - 1, \cdots, 1$ from $P(Z^{(1)}_{im} = k_1, Z^{(2)}_{m} = k_2 | Z^{(1)}_{i(m+1)} = k_3, Z^{(2)}_{i(m+1)} = k_4, DG_i, H_{-i}, \mu, \rho) \propto P(Z^{(1)}_{iM} = k_1, Z^{(2)}_{iM} = k_2, DG_{i1}, \cdots, DG_{im} | H_{-i}, \mu, \rho) \cdot P(Z^{(1)}_{im} = k_1, Z^{(2)}_{im} = k_2, | Z^{(1)}_{i(m+1)} = k_3, Z^{(2)}_{i(m+1)} = k_4, H_{-i}, \mu, \rho) = \alpha^i_{H_{-i}}(m, \{k_1, k_2\}) \cdot p_m(k_1 \to k3) p_m(k_2 \to k_4)$.

- sample DH_i from $P(DH_i | Z^{(1)}_{im}, Z^{(2)}_{im}, DG_i, H_{-i}, \mu, \rho) = \prod_{i=1}^{M} P(DH_{im} | Z^{(1)}_{im}, Z^{(2)}_{im}, H_{-i}, \rho, \mu)$ where $P(DH_{im} | Z^{(1)}_{im}, Z^{(2)}_{im}, H_{-i}, \rho, \mu)$ can be expressed as $P(DH^{(1)}_{im} | Z^{(1)}_{im}, H_{-i}, \rho, \mu) \cdot P(DH^{(2)}_{im} | Z^{(2)}_{im}, H_{-i}, \rho, \mu)$. Recognizing $P(DH^{(1)}_{im} | Z^{(1)}_{im}, H_{-i}, \rho, \mu)$ is the emssion probability in the haplotype version of HMM, we have

  $$\begin{aligned} & P(DH^{(1)}_{im}, DH^{(2)}_{im} | Z^{(1)}_{im} = k_1, Z^{(2)}_{im} = k_2, H_{-i}, \rho, \mu) \\ & \propto (1-\mu)^{I(DH^{(1)}_{im} = H^{k_1}_m)} \mu^{I(DH^{(1)}_{im} \neq H^{k_1}_m)} (1-\mu)^{I(DH^{(2)}_{im} = H^{k_2}_m)} \mu^{I(DH^{(2)}_{im} \neq H^{k_2}_m)}, \end{aligned}$$

  where $I(\cdot)$ is the identity function and $H^{k_1}_m$ is the allele (without mutation) associated with cluster k_1.

5.3 Experiments and Discussion

To assess accuracy of imputation, a total of 82 Simmental beef cattle from the 1000 Bull Genomes Project with their sequence data on BTA 14 were used in the simulation study. Through comparisons between the sequence data and Illumina 50K/6K chips, we identified $1,618$ SNP markers that belonged to Illumina 50K chip and 219 SNPs were shared between the Illumina 6K panel and the Illumina 50K panel. we further partitioned the 82 animals into two datasets, a reference panel consisting of 65 animals with SNPs in the Illumina 50K chip and a study sample of 17 animals for which only markers in the Illumina 6K panels were kept and the rest of markers were masked as “untyped”. The imputation task is to predict the “untyped” markers for the study sample.

Accuracy of imputation is the percentage of correctly imputed genotypes while assuming the masked genotypes are ground truths in the study sample. Table 5.2 shows the comparisons of accuracies and running time between Impute 2, our HMM and a baseline method. All experiments were conducted on the same computer with 2.2 GHz core and 4 GB memory. Impute 2 is currently the most accurate genotype imputation program and the baseline approach uses the most frequently observed genotypes in DG to fill untyped genotypes in SG. Our HMM model differs from Impute 2 and fastPHASE in several ways. First, we used the Haldane model to approximate the recombination rates between two consecutive marker, whereas Impute 2 used the formula $\rho_{m-1} = 4 \cdot N_e \cdot r_m$ for obtaining the recombination rates, where N_e is the effective population size for the population and r_m is the genetic distance between locus m and $m - 1$. Secondly, our HMM tried to represent the inferred and phased haplotypes of DG in a localized haplotype structure whereas Impute 2 did not cluster haplotypes in DG into clusters. Additionally, at each locus, our HMM used exactly two hidden states to represent the two clusters of alleles; however, in the model of fastPHASE/Bimbam, the number of clusters K is a parameter that can be specified by the user. fastPHASE and Bimbam used allele frequencies not the discrete allele to model clusters. Thirdly, Impute 2 used a heuristic approach for selecting a subset of closely related individuals for estimating the parameters in their models. Because of our compact representation of haplotypes in DG in a graph, our localized haplotype HMM graph took all the information in DG into account and was fast in forward-backward calculations. Fourthly, Impute 2 tried to resolve phasing in both SG and DG first and once accurate diplotypes were obtained for SG and DG, Impute 2 carried out haplotype based imputation based on “PAC” framework. In our HMM model, we only tried to phase haplotypes in DG.

One possible explanation for the poor accuracy achieved in our HMM model is its parsimonious representation of haplotypes in DG in a localized cluster HMM. For example, note the genotypes sequence "101" over three loci can be phased as either a diplotype pair "000" and "101" or a diplotype pair "100" and "001". However, the two diplotype pairs become indistinguishable in our graph representation. It implies that phasing in DG plays a vital role in genotype imputation. One possible way to improve our model is to relax the restriction that there are exactly two clusters at each locus. We tested a version of the localized cluster HMM model using the 130 haplotypes from the 65 reference animals to construct 130 clusters at each cluster in our HMM model. That is, we did not collapse the obtained haplotypes into two clusters but treated the obtained haplotypes from DG as clusters. The accuracy of imputation can be improved to 77.86% at the cost of much longer running time 24615.84 seconds. This suggests that one needs to find a balance between restricting the number of hidden states and retaining the phase information of DG to make the statistical model accurate and efficient. The phasing software SHAPEIT 2 [77] used a segmented approach to restrict the number of compatible haplotypes with ambiguous phases for a given genotypes and split the haplotypes in segments. A pruning strategy is applied to prune intra-segment edges. Although the number of clusters is larger than 2 in SHAPEIT 2, within each segment the number of possible haplotype segments is a constant less than the total number of individuals in DG. Therefore it is quick to obtain accurate haplotypes.

Program	#correctly imputed	#untyped genotypes	accuracy	running time
Impute2	19167	23783	80.59%	95.14 sec
HMM ($K = 2$)	16182	23783	68.04%	60.12 sec
baseline	13576	23783	57.08%	0.82 sec

Table 5.2: Comparison of methods measured by "accuracy of imputation" on the Simmental dataset BTA 14 for imputation from 6K to 50K. Impute 2 is currently the best imputation program and was run with the effective population size 200 and the default settings under the run type "Imputation with one unphased reference panel" and HMM is the statistical model we implemented. The running times are also reported for Impute 2 and HMM. Majority vote is the baseline methods that fills untyped genotypes in SG with the most frequently observed genotypes in DG.

Chapter 6

Conclusion

In this dissertation, I investigated genotype imputation for inferring missing genotypes and untyped markers in population genotype data. I examined a novel way to improve accuracy of imputation that can work existing genotype imputation methods in a multi-step procedure. Evaluation of untyped markers in a two-step setting along the genome yielded candidate markers and clustering was employed to narrow down the candidate list and group markers of similar effects in the two-step imputation. I studied existing genotype imputation models and categorized them according to their modelling parameters and the underlying biological concepts. Based on an existing popular and successful framework, I presented an HMM model that incorporates clustering for genotype imputation.

6.1 Summary of Contributions

In Chapter 2, I reviewed the recent developments in methods for population-based genotype imputation, and discussed in detail the underlying models for each method. In comparative studies of genotype imputation methods, I compared six current best population-based methods that use unphased reference panels for genotype imputation and investigated the effects of imputed 50K genotypes on feed efficiency genomic predictions for beef cattle data from both purebred and crossbred populations in Chapter 3. The six genotype imputation methods fall into three major categories: 1) methods based on Li and Stephens's "PAC" framework [61]; 2) Browning and Browning's IBD based HMMs (Beagle 3.3.2 and Beagle 4.0) and 3) a fast, efficient, and rule-based method called FImpute inspired by Kong et al.'s "long range phasing" [59]. HMMs based on the "PAC" framework can be further divided into two categories, one that models genotypes as discrete counts of alleles including Impute 2, MaCH and ones that use clustering and real-valued allele frequencies including fastPHASE and Bimbam. For HMM-based imputation methods, either Markov chain Monte Carlo sampling or EM-based maximum likelihood estimator is employed for parameter inference. In terms of efficiency, rule-based FImpute is the fastest method and is capable of yielding comparable accuracies to current best Impute 2. Computational burdens scale quadratically with the number of hidden states in "PAC"-based models. Our simulation studies confirmed that minor allele frequency plays a key role in the accuracy of imputation. As minor allele frequency increases, accuracies of all imputation methods to impute genotypes carrying the minor allele increase. Existing imputation methods have limitations in imputing rare alleles of frequencies less than 1%. FImpute exihited advantages over other methods in terms of running time and imputing rare alleles. Bimbam's lower performance is likely due to its use of MLE for cluster inference of the underlying architecture of the data. Accuracies of genomic predictions for RFI via either BayesB or GBLUP were higher on purebred populations than on crossbred populations, and no significant advantage of usage of 50K panel over 6K panel in genomic predictions was observed. Employing a within-breed training strategy has the potential to improve accuracies of genomic predictions for both BayesB and GBLUP, as observed in purebred populations because the level of relatedness plays a key role in the persistence of co-segregation of QTL with SNPs. Imputed 50K genotypes in the subsequence genomic predictions, via BayesB and GBLUP, in general yielded similar results for the trait to that using actual 50K genotypes in this study.

In Chapter 4, for genotype imputation from a lower density panel to a higher density panel, in order to boost accuracies of imputation, I presented a novel two-step strategy called "piecemeal genotype imputation," which essentially inserts a pseudo intermediate array in between the low-density chip and the high-density chip. I first demonstrated how to identify, evaluate and select untyped SNPs that can lead to accuracy improvement to construct the pseudo intermediate panel. Subsequently, I identified regions along the genome where accuracy of imputation can be further improved in a two-step manner with the selected markers, and lastly showed how the clusters of imputed genotype can be pieced together to form the final imputation result. Using the two-step piecemeal imputation, I showed how a stair-case of intermediate SNP arrays can be gradually built up for the whole genome SNP genotype imputation. I applied this strategy to chromosomes 14 and 27 of real cattle SNPs that arise from the whole genome sequencing, by carrying out extensive experiments using various density levels of bovine SNP chips, up to the sequence level. The results show preliminary success of our multi-step piecemeal imputation with an accuracy improvement compared to the classic one-step imputation by the state-of-the-art methods Beagle and FImpute. From a low-density chip to the whole sequence, intermediate pseudo-arrays can be computationally constructed by selecting the most informative SNPs for untyped SNP genotype imputation. Such pseudo-array staircases are able to boost accuracies of imputation compared to the classic one-step imputation.

In Chapter 5, I presented a statistical model based on Li and Stephens' "PAC" framework [61] for population based genotype imputation. It also incorporates the "local clustering" as the relative frequencies associated with the hidden states in the model in an HMM. The proposed method used a Gibbs sampler to estimate phasing in dense reference panels. It was fast and memory efficient for genotype imputation. However, compared to currently best performed program Impute 2, my model yielded lower accuracy of imputation. The poor accuracy of the model I proposed is likely due to its parsimonious representation of phasing information in DG. This drawback suggests that phasing plays in key role in the success of genotype imputation. One needs to find a balance between restricting the number of hidden states and representing the phasing information in DG properly.

6.2 Future Research Directions

The challenges of the piecemeal imputation strategy include its large disk space and computation time requirements. The training phase is the most time consuming step, when the add-one two-step experiments are conducted to select the good potential markers. The embedded imputation methods and the total number of markers that need to be evaluated in the add-one two-step training stage also affect the total running time. One way to speed up the training process is to parallelize tasks and submit jobs to large, powerful computing facilities. A huge disk storage of intermediate results (more than 84 TB) in the training phase is a major challenge and an overhead cost when we had to perform full-sequence SNP genotype imputation as we relied on accuracy-based feature vectors to evaluate and cluster added markers. Also, the use of mixed reference panels can result in increased imputation accuracy in all populations shown in previous studies [53]. As more individuals have the genome data sequenced and genotyped, if we incorporate them for re-training, it would be expected to increase the accuracies in one-step, two-step, piecemeal imputations. Evaluation of imputed SNP genotype along the genome is expected to be useful in the downstream data analysis, as well as for improvement of chip designs.

From the comparative studies, the experimental results demonstrated that existing genotype imputation methods all had limits in imputing rare variants. The problem persists especially for those statistical imputation models as these programs cannot distinguish a rare allele from genotype errors at loci of extremely low minor allele frequencies (MAF). The SNPs included in the SNP chips usually have high MAFs and are generally believed to be unlikely causal variants for complex phenotypic traits. Linkage disequilibrium between common SNPs and rare causal loci is not very strong or poor and then this could lead to low accuracy of genomic predictions as variation generated by the causal variants cannot be fully explained by the common SNPs. Therefore, sequencing all selection candidates has now been proposed as an alternative to overcome the mentioned problem because causal variants are in the data, although it is unlikely to happen in the near future mainly due to the cost. Also, design of new chips that include more low MAF SNPs is needed for beef cattle populations. Using imputation, obtaining accurate imputed rare variants can still be an issue depending on how many animals are sequenced. However, memory and computation time for sequence data remain an issue. Fast and efficient genomic prediction methods that can handle sequence data are needed.

For the development of the statistical model for genotype imputation, one of the challenges faced by many “PAC”-based methods is the number of individuals P in the haplotype reference panel. Although more individuals in the reference can potentially improve accuracies of imputation as shown in previous studies [6], the running time grows quadratically $O(P^2)$ in terms of the size of the reference P. MaCH [62] used a subset of individuals chosen in random to condition on in its MCMC sampling procedures whereas Impute 2 [51] used a heuristic “nearest neighbour” method to search for closely related samples for imputing study samples to overcome the quadratic issue. Since bi-allelic markers are considered in this dissertation, I attempted a parsimonious approach that collapses the number of haplotypes into two clusters representing the two alleles at each site. Such treatment stores information from the reference panel in a compact way and it is both efficient and fast in terms of space requirement and running time. However, my approach yielded poor accuracies of imputation and the cause of it is likely due to its parsimonious representation of the reference haplotypes. One can investigate more advanced approach such as “Dirichlet process” for obtaining a cluster of haplotypes from the genotype reference panel.

Bibliography

[1] JK Baker. The dragon system – an overview. Technical Report 1, 1975.

[2] JC Barrett, B Fry, JDMJ Maller, and MJ Daly. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21(2):263–265, 2005.

[3] JA Basarab, MG Colazo, DJ Ambrose, S Novak, D McCartney, and VS Baron. Residual feed intake adjusted for backfat thickness and feeding frequency is independent of fertility in beef heifers. *Canadian Journal of Animal Science*, 91(4):573–584, 2011.

[4] DP Berry, MC McClure, and MP Mullen. Within-and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low-density genotypes. *Journal of Animal Breeding and Genetics*, 131(3):165–172, 2014.

[5] D Boichard, H Chung, R Dassonneville, X David, A Eggen, S Fritz, KJ Gietzen, BJ Hayes, CT Lawley, TS Sonstegard, et al. Design of a bovine low-density snp array optimized for imputation. *PloS one*, 7(3):e34130, 2012.

[6] AC Bouwman and RF Veerkamp. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. *BMC genetics*, 15(1):1, 2014.

[7] BL Browning and SR Browning. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, 194(2):459–471, 2013.

[8] SR Browning. Multilocus association mapping using variable-length markov chains. *The American Journal of Human Genetics*, 78(6):903 – 913, 2006.

[9] SR Browning. Missing data imputation and haplotype phase inference for genome-wide association studies. *Human genetics*, 124(5):439–450, 2008.

[10] SR Browning and BL Browning. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *The American Journal of Human Genetics*, 81(5):1084 – 1097, 2007.

[11] SR Browning and BL Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12(10):703–714, 2011.

[12] Li C, Chen L, Vinsky M, Crowley J, Miller SP, Plastow G, Basarab J, and Stothard P. Genomic prediction for feed efficiency traits based on 50k and imputed high density snp genotypes in multiple breed populations of canadian beef cattle (abstract). volume 99, page 94, 2015.

[13] MPL Calus, AC Bouwman, JM Hickey, RF Veerkamp, and HA Mulder. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. *Animal*, 8(11):1743–1753, 2014.

[14] L Chen, C Li, M Sargolzaei, and F Schenkel. Impact of genotype imputation on the performance of gblup and bayesian methods for genomic prediction. *PloS one*, 9(7):e101544, 2014.

[15] L Chen, F Schenkel, M Vinsky, DH Crews, and C Li. Accuracy of predicting genomic breeding values for residual feed intake in angus and charolais beef cattle. *Journal of animal science*, 91(10):4669–4678, 2013.

[16] CYK Cheung, EA Thompson, and EM Wijsman. Gigi: an approach to effective imputation of dense genotypes on large pedigrees. *The American Journal of Human Genetics*, 92(4):504–516, 2013.

[17] EC Chi, H Zhou, GK Chen, DO Del Vecchyo, and K Lange. Genotype imputation via matrix completion. *Genome research*, 23(3):509–518, 2013.

[18] TCS Chud, RV Ventura, FS Schenkel, R Carvalheiro, ME Buzanskas, JO Rosa, M de Alvarenga Mudadu, MVGB da Silva, FB Mokry, CR Marcondes, et al. Strategies for genotype imputation in composite beef cattle. *BMC genetics*, 16(1):1, 2015.

[19] ET Cirulli and DB Goldstein. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Reviews Genetics*, 11(6):415–425, 2010.

[20] J-J Colleau. An indirect approach to the extensive calculation of relationship coefficients. *Genetics Selection Evolution*, 34(4):409–422, 2002.

[21] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[22] International HapMap Consortium et al. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[23] HD Daetwyler, A Capitan, H Pausch, P Stothard, R Van Binsbergen, RF Brøndum, X Liao, A Djari, SC Rodriguez, C Grohs, et al. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nature genetics*, 46(8):858–865, 2014.

[24] MJ Daly, JD Rioux, SF Schaffner, TJ Hudson, and ES Lander. High-resolution haplotype structure in the human genome. *Nature genetics*, 29(2):229–232, 2001.

[25] APW De Roos, BJ Hayes, and ME Goddard. Reliability of genomic predictions across multiple populations. *Genetics*, 183(4):1545–1553, 2009.

[26] JCM Dekkers. Application of genomics tools to animal breeding. *Current genomics*, 13(3):207–212, 2012.

[27] O Delaneau, J Marchini, and J-F Zagury. A linear complexity phasing method for thousands of genomes. *Nature methods*, 9(2):179–181, 2012.

[28] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[29] M Erbe, BJ Hayes, LK Matukumalli, S Goswami, PJ Bowman, CM Reich, BA Mason, and ME Goddard. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of Dairy Science*, 95(7):4114 – 4129, 2012.

[30] J Ertl, C Edel, R Emmerling, H Pausch, R Fries, and K-U Gtz. On the limited increase in validation reliability using high-density genotypes in genomic best linear unbiased prediction: Observations from fleckvieh cattle. *Journal of Dairy Science*, 97(1):487 – 496, 2014.

[31] D Falush, M Stephens, and JK Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003.

[32] A Feller, E Greif, L Miratrix, and N Pillai. Principal stratification in the twilight zone: Weakly separated components in finite mixture models. *arXiv preprint arXiv:1602.06595*, 2016.

[33] RL Fernando and DJ Garrick. Gensel - user manual for a portfolio of genomic selection related analyses. *Animal Breeding and Genetics, Iowa State University, Ames, IA, USA.*, 2009.

[34] DJ Garrick. The nature, scope and impact of genomic prediction in beef cattle in the united states. *Genetics Selection Evolution*, 43(1):17, 2011.

[35] RA Gibbs, JW Belmont, P Hardenbol, TD Willis, F Yu, H Yang, L-Y Ch'ang, W Huang, B Liu, Y Shen, et al. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

[36] ME Goddard, BJ Hayes, and THE Meuwissen. Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics*, 128(6):409–421, 2011.

[37] J Graffelman, M Sánchez, S Cook, and V Moreno. Statistical inference for hardy-weinberg proportions in the presence of missing genotype information. *PLoS One*, 8(12):e83316, 2013.

[38] AJF Griffiths. *An introduction to genetic analysis*. Macmillan, 2005.

[39] Y Guan. Detecting structure of haplotypes and local ancestry. *Genetics*, 196(3):625–642, 2014.

[40] Y Guan and M Stephens. Practical issues in imputation-based association mapping. *PLoS Genet*, 4(12):e1000279, 12 2008.

[41] KL Gunderson, FJ Steemers, G Lee, LG Mendoza, and MS Chee. A genome-wide scalable snp genotyping assay using microarray technology. *Nature genetics*, 37(5):549–554, 2005.

[42] A Gusev, JK Lowe, M Stoffel, MJ Daly, D Altshuler, JL Breslow, JM Friedman, and I Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome research*, 19(2):318–326, 2009.

[43] D Habier, RL Fernando, and JCM Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397, 2007.

[44] D Habier, RL Fernando, and DJ Garrick. Genomic blup decoded: a look into the black box of genomic prediction. *Genetics*, 194(3):597–607, 2013.

[45] E Halperin and DA Stephan. Snp imputation in association studies. *Nature biotechnology*, 27(4):349–351, 2009.

[46] BJ Hayes, PJ Bowman, AJ Chamberlain, and ME Goddard. Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science*, 92(2):433–443, 2009.

[47] M Heidaritabar, MPL Calus, H-J Megens, A Vereijken, MAM Groenen, and JWM Bastiaansen. Accuracy of genomic prediction using imputed whole-genome sequence data in white layers. *Journal of Animal Breeding and Genetics*, 2016.

[48] JM Hickey, J Crossa, R Babu, and G de los Campos. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop science*, 52(2):654–663, 2012.

[49] JM Hickey, BP Kinghorn, B Tier, JHJ van der Werf, and MA Cleveland. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. *Genetics Selection Evolution*, 44(1):1, 2012.

[50] B Howie, J Marchini, and M Stephens. Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*, 1(6):457–470, 2011.

[51] BN Howie, P Donnelly, and J Marchini. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529, 06 2009.

[52] C Hozé, M Fouilloux, E Venot, F Guillaume, R Dassonneville, S Fritz, V Ducrocq, F Phocas, D Boichard, and P Croiseau. High-density marker imputation accuracy in sixteen french cattle breeds. *Genetics Selection Evolution*, 45(1):1, 2013.

[53] L Huang, Y Li, AB Singleton, JA Hardy, G Abecasis, NA Rosenberg, and P Scheet. Genotype-imputation accuracy across worldwide human populations. *The American Journal of Human Genetics*, 84(2):235–250, 2009.

[54] RR Hudson. Two-locus sampling distributions and their application. *Genetics*, 159(4):1805–1817, 2001.

[55] MS Khatkar, G Moser, BJ Hayes, and HW Raadsma. Strategies and utility of imputed snp genotypes for genomic analysis in dairy cattle. *BMC genomics*, 13(1):538, 2012.

[56] G Kimmel and R Shamir. A block-free hidden markov model for genotypes and its application to disease association. *Journal of Computational Biology*, 12(10):1243–1260, 2005.

[57] Robert M Koch, Lo A Swiger, Doyle Chambers, and KE Gregory. Efficiency of feed use in beef cattle. *Journal of Animal Science*, 22(2):486–494, 1963.

[58] A Kong, DF Gudbjartsson, J Sainz, GM Jonsdottir, SA Gudjonsson, B Richardsson, S Sigurdardottir, J Barnard, B Hallbeck, G Masson, et al. A high-resolution recombination map of the human genome. *Nature genetics*, 31(3):241–247, 2002.

[59] A Kong, G Masson, ML Frigge, A Gylfason, P Zusmanovich, G Thorleifsson, PI Olason, A Ingason, S Steinberg, T Rafnar, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*, 40(9):1068–1075, 2008.

[60] S Larmer, M Sargolzaei, R Ventura, and F Schenkel. Imputation accuracy from low to high density using within and across breed reference populations in holstein, guernsey and ayrshire cattle.

[61] N Li and M Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165(4):2213–2233, 2003.

[62] Y Li, CJ Willer, J Ding, P Scheet, and GR Abecasis. Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8):816–834, 2010.

[63] RJA Little and DB Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2002.

[64] D Lu, EC Akanno, JJ Crowley, F Schenkel, H Li, M De Pauw, SS Moore, Z Wang, C Li, P Stothard, G Plastow, SP Miller, and J A Basarab. Accuracy of genomic predictions for feed efficiency traits of beef cattle using 50k and imputed hd genotypes. *Journal of Animal Science*, 94(4):1342–1353, 2016.

[65] MS Lund, G Su, L Janss, B Guldbrandtsen, and RF Brøndum. Genomic evaluation of cattle in a multi-breed context. *Livestock Science*, 166:101–110, 2014.

[66] Sargolzaei M, VanRaden PM, Kistemaker GJ, and Schenkel FS. Gebv software. *L'alliance boviteq, sainthyacinthe, quebec and centre for genetic improvement of livestock*, 2011.

[67] KA Macdonald, JE Pryce, RJ Spelman, SR Davis, WJ Wales, GC Waghorn, YJ Williams, LC Marett, and BJ Hayes. Holstein-friesian calves selected for divergence in residual feed intake during growth exhibited significant but reduced residual feed intake divergence in their first lactation. *Journal of dairy science*, 97(3):1427–1435, 2014.

[68] J Marchini and B Howie. Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511, 07 2010.

[69] J Marchini, B Howie, S Myers, G McVean, and P Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–913, 07 2007.

[70] LK Matukumalli, CT Lawley, RD Schnabel, JF Taylor, MF Allan, MP Heaton, J O'Connell, SS Moore, TPL Smith, TS Sonstegard, et al. Development and characterization of a high density snp genotyping assay for cattle. *PloS one*, 4(4):e5350, 2009.

[71] M McClure, TS Sonstegard, G Wiggans, and CP Van Tassell. Imputation of microsatellite alleles from dense snp genotypes for parental verification. *Frontiers in genetics*, 3:140, 2012.

[72] MC Mcclure, TS Sonstegard, GR Wiggans, AL Van Eenennaam, KL Weber, MCT Penedo, D Berry, J Flynn, JF Garcia, AS Carmo, LCA Regitano, M Albuquerque, MVGB Silva, MA Machado, M Coffey, K Moore, M-Y Boscher, L Genestout, R Mazza, JF Taylor, RD Schnabel, B Simpson, E Marques, J McEwan, A Cromie, LL Coutinho, L Kuehn, J Keele, E Piper, J Cook, R Williams, and C Van Tassell. Imputation of microsatellite alleles from dense snp genotypes for parentage verification across multiple bos taurus and bos indicus breeds. *Frontiers in Genetics*, 4(176), 2013.

[73] TH Meuwissen, BJ Hayes, and M Goddard. Accelerating improvement of livestock with genomic selection. *Annu. Rev. Anim. Biosci.*, 1(1):221–237, 2013.

[74] TH Meuwissen, BJ Hayes, and ME Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

[75] FDN Mujibi, JD Nkrumah, ON Durunna, P Stothard, J Mah, Z Wang, J Basarab, G Plastow, DH Crews, and SS Moore. Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle. *Journal of animal science*, 89(11):3353–3361, 2011.

[76] HA Mulder, MPL Calus, Tom Druet, and C Schrooten. Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in dutch holstein cattle. *Journal of dairy science*, 95(2):876–889, 2012.

[77] J O'Connell, D Gurdasani, O Delaneau, N Pirastu, S Ulivi, M Cocca, M Traglia, J Huang, JE Huffman, I Rudan, R McQuillan, RM Fraser, H Campbell, O Polasek, G Asiki, K Ekoru, C Hayward, AF Wright, V Vitart, P Navarro, J-F Zagury, JF Wilson, D Toniolo, P Gasparini, N Soranzo, MS Sandhu, and J Marchini. A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet*, 10(4):e1004234, 04 2014.

[78] ML Piccoli, J Braccini, FF Cardoso, M Sargolzaei, SG Larmer, and FS Schenkel. Accuracy of genome-wide imputation in braford and hereford beef cattle. *BMC genetics*, 15(1):1, 2014.

[79] ECG Pimentel, C Edel, R Emmerling, and K-U Götz. How imputation errors bias genomic predictions. *Journal of dairy science*, 98(6):4131–4138, 2015.

[80] ECG Pimentel, M Wensch-Dorendorf, S König, and HH Swalve. Enlarging a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture. *Genetics Selection Evolution*, 45(1):1, 2013.

[81] AL Price, A Tandon, N Patterson, KC Barnes, N Rafaels, I Ruczinski, TH Beaty, R Mathias, D Reich, and S Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, 5(6):e1000519, 2009.

[82] J.E. Pryce, J. Arias, P.J. Bowman, S.R. Davis, K.A. Macdonald, G.C. Waghorn, W.J. Wales, Y.J. Williams, R.J. Spelman, and B.J. Hayes. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. *Journal of Dairy Science*, 95(4):2108 – 2119, 2012.

[83] LR Rabiner and B-H Juang. Fundamentals of speech recognition. 1993.

[84] RA Redner and HF Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM review*, 26(2):195–239, 1984.

[85] JK Romberg, H Choi, and RG Baraniuk. Bayesian tree-structured image modeling using wavelet-domain hidden markov models. *IEEE Transactions on image processing*, 10(7):1056–1068, 2001.

[86] DB Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[87] M Saad and EM Wijsman. Combining family-and population-based imputation data for association analysis of rare and common variants in large pedigrees. *Genetic epidemiology*, 38(7):579–590, 2014.

[88] M Sargolzaei, JP Chesnais, and FS Schenkel. A new approach for efficient genotype imputation using information from relatives. *BMC genomics*, 15(1):478, 2014.

[89] M Sargolzaei, F Schenkel, and J Chesnais. Accuracy of imputed 50k genotypes from 3k and 6k chips using fimpute version 2. pages 1–9, 2011.

[90] P Scheet and M Stephens. A fast and flexible statistical model for large-scale population genotype data: Applications to inferring missing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78(4):629 – 644, 2006.

[91] B Servin and M Stephens. Imputation-based analysis of association studies: Candidate regions and quantitative traits. *PLoS Genet*, 3(7):e114, 07 2007.

[92] FJ Steemers, W Chang, G Lee, DL Barker, R Shen, and KL Gunderson. Whole-genome genotyping with the single-base extension assay. *Nature methods*, 3(1), 2006.

[93] P Stothard, X Liao, AS Arantes, M De Pauw, C Coros, GS Plastow, M Sargolzaei, JJ Crowley, JA Basarab, F Schenkel, et al. A large and diverse collection of bovine genome sequences from the canadian cattle genome project. *GigaScience*, 4(1):49, 2015.

[94] X Sun, RL Fernando, and J Dekkers. Contributions of linkage disequilibrium and co-segregation information to the accuracy of genomic prediction. *Genetics Selection Evolution*, 48(1):77, 2016.

[95] X Sun, RL Fernando, DJ Garrick, and J Dekkers. Improved accuracy of genomic prediction for traits with rare qtl by fitting haplotypes. *Animal Industry Report*, 661(1):86, 2015.

[96] F Tiezzi and C Maltecca. Accounting for trait architecture in genomic predictions of us holstein cattle using a weighted realized relationship matrix. *Genetics Selection Evolution*, 47(1):1, 2015.

[97] R van Binsbergen, MCAM Bink, MPL Calus, FA van Eeuwijk, BJ Hayes, I Hulsegge, and RF Veerkamp. Accuracy of imputation to whole-genome sequence data in holstein friesian cattle. *Genetics Selection Evolution*, 46(1):1, 2014.

[98] R van Binsbergen, MPL Calus, MCAM Bink, FA Eeuwijk, C Schrooten, and RF Veerkamp. Genomic prediction using imputed whole-genome sequence data in holstein friesian cattle. *Genetics Selection Evolution*, 47(1):1–13, 2015.

[99] PM VanRaden. Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11):4414–4423, 2008.

[100] PM VanRaden, CP Van Tassell, GR Wiggans, TS Sonstegard, RD Schnabel, JF Taylor, and FS Schenkel. Invited review: Reliability of genomic predictions for north american holstein bulls. *Journal of dairy science*, 92(1):16–24, 2009.

[101] RV Ventura, D Lu, FS Schenkel, Z Wang, C Li, and SP Miller. Impact of reference population on accuracy of imputation from 6k to 50k single nucleotide polymorphism chips in purebred and crossbreed beef cattle. *Journal of animal science*, 92(4):1433–1444, 2014.

[102] Y Wang, G Lin, C Li, and P Stothard. Genotype imputation methods and their effects on genomic predictions in cattle. *Springer Science Reviews*, pages 1–20, 2017.

[103] Y Wang, T Wylie, P Stothard, and G Lin. Whole genome snp genotype piecemeal imputation. *BMC bioinformatics*, 16(1):340, 2015.

[104] L Wasserman. Mixture models: The twilight zone of statistics. `https://normaldeviate.wordpress.com/2012/08/04/mixture-models-the-twilight-zone-of-statistics/`. Accessed: 2016-09-30.

[105] X Wen and M Stephens. Using linear predictors to impute allele frequencies from summary or pooled genotype data. *The annals of applied statistics*, 4(3):1158–1182, 09 2010.

[106] JE Womack. Advances in livestock genomics: opening the barn door. *Genome Research*, 15(12):1699–1705, 2005.

[107] Z Yu and DJ Schaid. Methods to impute missing genotypes for population data. *Human genetics*, 122(5):495–504, 2007.