Springer Science Reviews (2016) 4:79–98
DOI 10.1007/s40362-017-0041-x

ANALYTICAL STUDENT REVIEW

Genotype Imputation Methods and Their Effects on Genomic Predictions in Cattle

Yining Wang[1] · Guohui Lin[1] · Changxi Li[2,3] · Paul Stothard[2]

Received: 28 July 2016 / Revised: 19 December 2016 / Accepted: 29 December 2016 / Published online: 21 February 2017

Abstract In this study, we reviewed six imputation methods (Impute 2, FImpute 2.2, Beagle 4.1, Beagle 3.3.2, MaCH, and Bimbam) and evaluated the accuracy of imputation from simulated 6K bovine SNPs to 50K SNPs with 1800 beef cattle from two purebred and four crossbred populations and the impact of imputed genotypes on performance of genomic predictions for residual feed intake (RFI) in beef cattle. Accuracy of imputation was reported in both concordance rate (CR) and allelic r^2 and assessed via fivefold cross-validations. Running times of different methods were compared. Impute 2, FImpute and Beagle 4.1 yielded the most accurate imputation results (with CR $>$ 91%). FImpute was the fastest and had advantages over all other methods in imputing rare variants. Minor allele frequency (MAF) and genetic relatedness between individuals in reference and validation populations can affect accuracy of imputation. For all methods, imputation accuracy for genotypes carrying the minor allele increases as the MAF increases. Impute 2 outperformed all other methods on MAF $>$ 5% and onwards. FImpute and Impute 2 that adopted the nearest neighbour scheme coped better with individuals of distant relativeness. Bimbam yielded the poorest CR (76%) due to admixed reference panels. Imputed genotypes and actual 50K/6K genotypes were employed to predict genomic breeding values (GEBVs) of RFI using a Bayesian method and GBLUP. Accuracies of GEBV were similar using actual 50K genotypes or imputed genotypes, except those from Bimbam, and the imputation errors had minimal impact on the genomic predictions.

Keywords Genotype imputation · Genomic prediction · Residual feed intake

Endorsed by Guohui Lin and Paul Stothard.

✉ Changxi Li
changxi.li@ualberta.ca

✉ Paul Stothard
stothard@ualberta.ca

Yining Wang
yining@ualberta.ca

Guohui Lin
guohui@ualberta.ca

[1] Department of Computing Science, University of Alberta, Edmonton, AB, Canada

[2] Livestock Gentec, Department of Agricultural, Food and Nutritional Science, University of Alberta, Edmonton, AB, Canada

[3] Lacombe Research and Development Centre, Agriculture and Agri-Food Canada, 6000 C & E Trail, Lacombe, AB, Canada

Introduction

With the development of high-throughput DNA genotyping chips of various densities and the advance of sequencing technologies [1–4], numerous genetic variants have become available for use in livestock improvement. 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) and 1.6 million INDELs [5]. These types of dense SNPs that exhibit variations in regions along the whole genome have become a valuable tool for parental verification [6], identification of potential disease-risk genes [3] and genomic selection (GS) with the aim of improving genetic gains [7, 8].

Various statistical approaches have been proposed for genomic predictions, and they differ in their assumptions about marker effects. For example, the genomic best linear unbiased prediction (GBLUP) model [9] assumes all markers contribute to 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 [8].

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 [10]. 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 and genomic predictions.

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 [7, 11]. 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 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 [6, 7], the price of genotyping of high-density chips remains a major challenge for a 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 6909 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 [1] and was reported to contain lower genotyping errors than its low density predecessor the Illumina Golden Gate Bovine3K chip. In addition, the Illumina 6K chip can be customized by adding SNPs.

The key idea of the existing genotype imputation methods is to explore and hunt for shared "*identical by descend*" (IBD) haplotypes that exhibit high linkage disequilibrium (LD) measured in r^2 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. [3] 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 [12] and the 1000 Genome Project [13]. 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.

Existing methods for genotype imputation can be categorized computationally into the linear regression model by Yu and Schaid [14], clustering models [15–18], hidden Markov models (HMMs) and expectation–maximization (EM) algorithms. More recent works have included "BLIMP" by Wen and Stephens [19] based on "Kriging" for imputation from summary data and "Mendel-Impute" via matrix completion [20].

Alternatively, imputation methods can be divided into two broad categories: "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 [21] and sampling-based GIGI [22]. In general, family-based imputation programs using Mendelian segregation rules and LD information result in better 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 [21–24]. We focus 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 [22]. Third, some family-based programs require availability of dense genotypes for all immediate ancestors [21].

There have been several excellent reviews on genotype imputation methods and applications to human genome-wide association studies [10, 25, 26] as well as related reviews on haplotyping methods [27]. Several studies have investigated the performance of imputation methods in the context of livestock applications [28] and evaluated their effects on genomic predictions [28–30]. 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 [15], the "PAC" model of Li and Stephens [31] and its variants [17], and a simple rule-based method called FImpute [32] inspired by "long range phasing" [33]. We also evaluate the impact of genotype imputation accuracy on genomic predictions based on real beef cattle data.

Imputation Models and Popular Methods

In this section, we review the most widely used computational models underlying several population-based genotype imputation methods. The population-based genotype imputation problem can be formally defined as follows: given a 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 dataset SG, termed a "study sample", genotyped on a low-density chip. Strictly speaking, individuals are "related" to some degree in that even two distant 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 [34]. 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 $\mathrm{SG}_{ij} = ?$ 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 of markers $\mathcal{T}$ 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.

All existing 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 [35]. 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 SG can be filled with observed genotypes from DG if there is a match at typed markers $\mathcal{T}$ [19, 36]. Most methods not only perform genotype imputation for the study sample but infer haplotype phases as well [15, 35, 36].

The "Product of Approximate Conditionals" (PAC) Model

The statistical model of Li and Stephens [31] 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 [36], Impute 2 [35], MaCH [37], fastPHASE [17] and Bimbam [16] are all variants based on this idea. Li and Stephens [31] 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 for each sample. Due to the fact that recombination and mutation are both rare events and individuals are related in some degree, 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 N observed haplotypes, which represent the hidden states of a 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 that is 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 [31].

Discrete HMM Models—Impute1, Impute 2 and MaCH

Impute 1 [36], Impute 2 [35] and MaCH [37] 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 an HMM framework.

Impute 1 [36] 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. For each individual SG_i genotyped with low-density chip, it defines $P(\mathrm{SG}_i|\mathrm{DH}, \rho, \lambda)$ in the HMM framework of Li and Stephens [31], where λ is the mutation rate dependent on the number of individuals in the reference panel $\mathcal{R}$.

$$P(\mathrm{SG}_i|\mathrm{DH}, \rho, \lambda) = \sum_{Z_i} P(\mathrm{SG}_i|Z_i, \lambda)P(Z_i|\mathrm{DH}, \rho),$$
$$i = 1, \ldots, n,$$

where ρ_m is the probability that a recombination event occurs between loci m and $m+1$. The hidden state $Z_{im} = \{k_1, k_2\}$ at each marker m is an unordered pair of haplotypes k_1, k_2 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 computed via the forward–backward algorithm and are estimated in a sampling process, and computation grows linearly in the number of markers and quadratically in the number of haplotypes [36].

MaCH [37] further extends Impute 1's discrete-valued HMM model to the usage of unphased genotypes of DG for the reference panel $\mathcal{R}$. Phasing of DG is obtained through a Monte Carlo Gibbs sampling procedure $P(\mathrm{DG}_i|\mathrm{DG}_{-i}, \rho, \lambda)$, and only a few rounds of updates are needed to obtain accurate consensus haplotype templates [37]. The detailed path-sampling procedure of the HMM can be found in Appendix B of Scheet and Stephens [17]. The phasing procedure takes $O(N^3)$ if all individuals in DG are used since each update needs to sample a path from $O(N^2)$ 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 a large number of individuals in DG, Li et al. [37] suggested using a randomly selected subset of DG for sampling phases of DG_i at a small cost of accuracy. Howie et al. [38] proposed a two-step strategy named "minimac 2" that relies on MaCH for estimating haplotypes for target samples in a pre-phasing step, which can handle large reference panels of tens of thousands of individuals. In the second step, minimac 2 then imputes them using a selected set of reference haplotypes.

Impute 2 [13, 35] 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 in each iteration 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_i from $P(\mathrm{DG}_i|\mathrm{DH}_{-i}, \rho, \lambda)$;
- it iteratively updates phasing in SG_i at typed markers $\mathcal{T}$ conditional on current "phased" DH and current guess of the rest of individuals from $P(\mathrm{SG}_i^{\mathcal{T}}|\mathrm{SH}_{-i}^{\mathcal{T}}, \mathrm{DH}^{\mathcal{T}\cup\mathcal{U}}, \rho, \lambda)$;
- it imputes two alleles at $\mathcal{U}$ untyped markers for $\mathrm{SH}_{i,1}$ and $\mathrm{SH}_{i,2}$ from $P(\mathrm{SH}_{i,d}|\mathrm{DH}, \rho, \lambda)$ via the forward–backward algorithm, where $\mathrm{SH}_{i,1}$ and $\mathrm{SH}_{i,2}$ 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 are determined by hamming distance to the current individual, and computation burden of phasing grows quadratically with k closest neighbours $O(k^2N)$ and increases linearly in the number of markers $O(L)$. As phasing is resolved in the preceding 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 [17] 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 [17]. 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 dependent on the mutation rate but modelled with regard to the real-valued "allele frequencies" of each cluster. The total number of clusters K is a parameter specified by users. The relative frequencies of clusters α_{km} are initialized to be drawn from a Dirichlet prior distribution, and θ_{km} are initialized to be drawn from uniform distribution on [0.01, 0.99]. 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 $\theta_{km} \in [0.01, 0.99]$. Structure 2.0 [39], a software developed for inference of population structure, shares similarity to fastPHASE's local cluster HMM model, assuming that each cluster represents a sub-population, and using 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 employed for finding ML estimates of all parameters. It should be noted that Kimmel and Shamir [40] formalized a similar HMM model ("HINT") for disease association studies and proved that the genotype optimization problem is neither convex nor concave, and their exact form of maximization for updating θ_{km} does not exist. In HINT, Kimmel and Shamir [40] proposed 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 gave a formula for approximating maximal θ_{km}, which updates the current value of θ_{km} with the value in the preceding step of the maximization step. To obtain better parameter estimates, authors suggested setting $K = 20$, running EM multiple times and taking the average of estimates to overcome local maxima issues. The computational time is in $O(n \cdot L \cdot K^2)$, 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 [10, 16]. When applying fastPHASE for imputation with a reference panel DG, Guan and Stephens [16] suggested using parameter estimates obtained from maximizing the likelihood for DG only, $P(\mathrm{DG}|\alpha, \theta, \rho)$, rather than the full likelihood function $P(\mathrm{DG}, \mathrm{SG}|\alpha, \theta, \rho)$ 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 [16, 18], a software for Bayesian imputation-based association mapping. Guan [41] 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 $O(K^3)$.

Blimp

Following the arguments by Guan and Stephens [16] 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 [16], 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 [19]. Wen and Stephens [19] 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 [19].

Beagle 3.3.2 and Beagle 4.1

Beagle 3.3.2 is based on a flexible "localized haplotype-cluster" model [42] that groups locally similar haplotypes into clusters [15]. It 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 [15]. 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 to the HMMs based on the "PAC" model, which has the fixed number of hidden states at each marker, Beagle has fewer 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 examining haplotype frequencies at each node. 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 [43] 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. [34] 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. O'Connell et al. [44] 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 [45].

FImpute

FImpute [32] is an efficient, rule-based, and deterministic method for phasing and genotype imputation inspired by "long range phasing" [33]. Kong et al. [33] 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 [32]. The algorithm first resolves phasing for homozygous genotypes of each individual, treats heterozygous genotypes as missing or wild card, and gradually 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. If additional pedigree information is provided, FImpute starts with family-based imputation, and then performs population-based imputation.

Materials and Methods

Genotypes and Phenotypic Records

A total of 1800 animals were used in this study, from a large pool of 11,414 beef cattle genotyped on the Illumina BovineSNP50 BeadChip (Illumina 50K) collated from various projects and research herds across Canada including 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), and a TX/TXX commercial population that is heavily influenced by Charolais with infusion of Holstein, Maine Anjou and Chianina [46]. Quality controls (QC) were performed considering merged samples of all breeds simultaneously to filter out SNPs for the merged dataset of 11,414 animals 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 [46]. 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 because there are some genotyping discrepancies due to the design of the two genotyping chips. 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 5088 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. [47]. Values of RFI for all 1800 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. [48], Chen et al. [49] and Lu et al. [46].

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. Fivefold cross validation (CV) 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% (5088/33,911) 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 fivefold CV, imputation was carried out for low-density target samples across six populations using a single reference panel composed of the 1440 animals across six populations. The partition of the dataset was used for both imputation and subsequent genomic predictions.

We used two genomic prediction methods including an efficient GBLUP with a genomic relationship matrix G [50] and a Bayesian method (BayesB), together with imputed 50K genotypes from different methods and associated phenotypic values to predict the genomic breeding values (GEBVs) in fivefold 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.

Genomic Predictions Using BayesB and GBLUP

This section describes a Bayesian method (BayesB) and a genomic best linear unbiased prediction (GBLUP) method, both of which use information from a SNP dataset $X_{n \times L}$ containing genotypes for n animals over L SNP loci. Each element X_{ij} in X represents animal i's genotype coded as -1, 0 and 1 for the homozygote "AA", the heterozygote "AB", and the other homozygote "BB", respectively, assuming that two alleles at the jth locus are "A" and "B".

The BayesB model proposed by Meuwissen et al. [8] fits all SNP effects simultaneously and assumes the following linear model.

$$y_i = \mu + \sum_{j=1}^{L} \beta_j \cdot X_{ij} + e_i, \quad i = 1, \ldots, n,$$

where y_i is the adjusted RFI for animal i in the training population, μ is the overall mean, β_j is the regression coefficient (allele substitution effect) on the jth SNP, 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 χ^2 distribution with the degrees of freedom v_e set to 10 and the scale parameter S_e^2 set to $\frac{\hat{\sigma}_e^2 (v_e - 2)}{v_e}$, with $\hat{\sigma}_e^2$ being the estimated random residual effect variance. 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.

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

where π is our prior belief of the proportion of SNP that has no effects on the trait. In this study, the value of π was set to 0.99 under all scenarios, and the locus specific variance σ_j^2 is the unknown and is estimated from the data. Again, the prior for σ_j^2 is assumed to be from a scaled χ^2 distribution with the degrees of freedom set v_j to 4 and the scale S_j^2 set to $\left(v_j - 2\right)\hat{\sigma}_a^2 \Big/ \left[v_j (1 - \pi) \sum_{j=1}^{L} 2p_j (1 - p_j)\right]$, where $\hat{\sigma}_a^2$ 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. In this study, heritability $h^2 = 0.25$ was used under all scenarios. 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. An implementation of the BayesB method by Fernando and Garrick [51], known as "Gensel", was used in this study. Since Gensel requires no missing values in the Genotypic data $X_{n \times L}$, Impute 2 with the option "–phase" was used to infer the small percentage (0.36%) of sporadic missing genotypes. In all 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 [7]. SNP effects were estimated by averaging all the samples after the burn-in period. The GEBV for animal i in the validation population was then predicted by adding up SNP effects over all loci: $\mathrm{GEBV}_i = \sum_{j=1}^{L} \hat{\beta}_j X_{ij}$, where L is the total number of SNPs, and $\hat{\beta}_j$ is the estimated effect for marker j.

The GBLUP method [50] assumes a linear model that uses a genomic relationship matrix $\boldsymbol{G}$ derived from the SNP dataset $X_{n \times L}$ for estimating GEBVs. The linear model can be written as

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Za} + \boldsymbol{e},$$

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

$$\boldsymbol{G} = \frac{\boldsymbol{ZZ}'}{2\sum_{i=1}^{m}\boldsymbol{p}_i(1-\boldsymbol{p}_i)}.$$

GEBV are obtained by solving the following set of equations [52–54].

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

where $\boldsymbol{R}$ is a diagonal matrix with entries $\boldsymbol{R}_{ii} = \frac{1}{h^2} - 1$, and h^2 is the heritability and is set to 0.25. The genomic relationship matrix $\boldsymbol{G}$ was efficiently computed using Colleau's indirect method [55]. We used an implementation of GBLUP by Sargolzaei et al. [53] in the software "GEBV" to estimate GEBV for animals in the validation population.

Evaluation

To assess the qualities of imputed genotypes among various methods, a validation dataset is usually held with actual SNP genotypes assayed and by comparing the imputed genotypes against the actual ones one can get concordance rate (CR, aka accuracy), which is defined as proportion of imputed genotypes and actual ones at all untyped SNP loci. However, as Hickey et al. [21, 56] pointed out, CRs are allele frequency dependent and do not reflect the power of any imputation method to infer rare allele variants with MAF less than 1%. Additionally, Calus et al. [28] 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 [13]. The closer to one r^2 is the more power to detect an imputation method exhibits. We followed the notion of Howie et al. [13] by assigning undefined r^2 to 0 when imputation methods yielded all identical predictions for all individuals 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 GEBVs using either GBLUP or BayesB and the adjusted phenotypic values of RFI.

Program Settings

We performed all the 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 N_e to 150 for cattle populations, calculated the recombination rates between two consecutive loci using the Haldane [57] recombination model by assuming that 1 million base pair approximately corresponded to 1 Morgan and used the default total MCMC iterations 30. For FImpute 2, we adopted its default settings for population imputation. 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 h 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 inputs as 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 5) the default steps of each EM run "-s 1" of 4) the number of warm-up EM step runs ("-w 20").

Results

Accuracy of Genotype Imputation

Table 1 shows a huge variation among different methods in accuracies of imputation when a reference panel containing animals from all six populations was used. The overall

Table 1 Accuracy of genotype imputation from Illumina 6K to Illumina 50K for different methods

Program	Mean CR (%)	Mean r^2	Running time
Impute 2	93.95	0.7545	22 h 7 min 41 s
FImpute 2.2	91.88	0.6626	4 min 12 s
Beagle 4.1	91.70	0.6655	191 h 6 min 5 s
Beagle 3.3.2	87.38	0.5556	31 h 22 min
MaCH 1.0	80.21	0.4180	16 h 53 min 46 s[a]
Bimbam 1.0	71.72	0.2506	3 h 15 min 50 s

[a] Additional 14 h 18 min 14 s for pre-phasing the unphased animals in the reference panel for MaCH 1.0

Table 2 Accuracy of genotype imputation from Illumina 6K to Illumina 50K for different methods and different populations

Population	Impute 2		FImpute 2.2		Beagle 4.1	
	CR	r^2	CR	r^2	CR	r^2
Angus	97.75	0.7557	96.47	0.7065	96.74	0.7152
Charolais	95.84	0.7523	93.57	0.6616	93.00	0.6526
Kinsella	95.93	0.8458	94.84	0.7875	94.51	0.7827
Elora	91.01	0.747	88.21	0.6151	87.68	0.6091
PG1	92.12	0.7738	89.64	0.6595	89.87	0.6722
TX/TXX	91.08	0.7319	88.55	0.6132	88.39	0.6167
All	93.95	0.7545	91.88	0.6626	91.70	0.6655
Population	Beagle 3.3.2		MaCH		Bimbam	
	CR	r^2	CR	r^2	CR	r^2
Angus	94.69	0.6585	87.64	0.5288	77.89	0.3509
Charolais	87.89	0.5207	78.13	0.3259	68.39	0.1543
Kinsella	90.85	0.6787	83.16	0.5105	72.18	0.2895
Elora	82.15	0.4685	76.42	0.354	71.07	0.2518
PG1	85.36	0.553	78.98	0.4178	72.17	0.2960
TX/TXX	83.48	0.4962	76.95	0.3645	68.64	0.2068
All	87.38	0.5556	80.21	0.418	71.72	0.2506

mean CR and mean allelic r^2 were the highest when Impute 2 was used for imputation, followed by FImpute 2 and Beagle 4.1 both of which yielded above 91% mean CR and above 66% mean allelic r^2. Beagle 3.3.2 yielded a mean CR 87.38% and a mean allelic r^2 0.5556. MaCH 1.0 and Bimbam 1.0 gave mean CRs 80.21% and 71.72%, respectively, and mean allelic r^2 0.4180 and 0.2506, respectively.

In terms of speed, as shown in Table 1, FImpute 2.2 was the fastest program yet achieved competitive imputation accuracies in terms of CR and allelic 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 of 191 h. Impute 2 overcame the quadratic running time with the number of animals by heuristically searching the closest reference haplotypes (defined by humming distances) [13]. However, the model-based imputation methods such as Impute 2 and Beagle 4.1 both suffer scalability issue once we would like to impute from genotype chips up to the full sequence level.

In Table 2, each method performed well with pure breed populations Angus and Charolais and the crossbred population Kinsella. Each method achieved the highest mean CRs with Angus, followed by Charolais and Kinsella. Due to differences 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 measured by principal components in Fig. 1. Impute v2 clearly outperformed all other methods in both mean CR and mean allelic r^2 for the two purebred and four crossbred populations.

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 is due to rare variants [58]. We evaluated imputation methods for their CRs 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 following six classes according to MAF, (0, 1%), [1%, 2%], [2%, 5%], [5%, 10%], [10%, 20%] and [20%, 50%]. Figure 2a–f shows the relationship between MAF and CRs of genotypes "AB" and "BB" for different methods. As MAF increased, CRs 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 [56] and whole-genome sequencing Holstein–Friesian cattle [59]. Greater differences among different methods were observed across variant MAF classes in the CRs 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

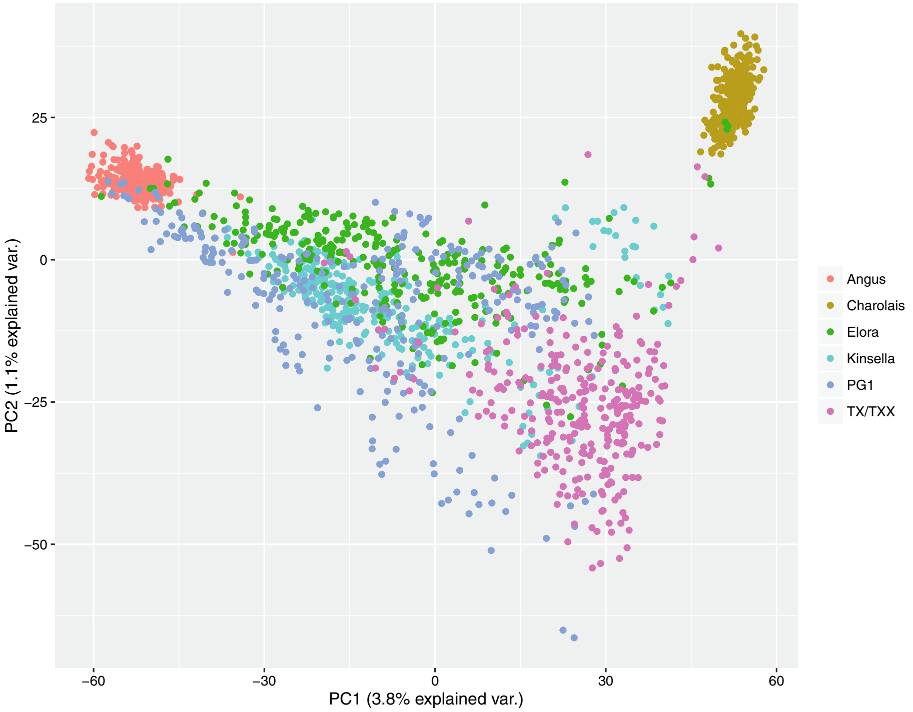


Fig. 1 Principal component analysis (PCA) for population stratification using the top two principal components (PCs) obtained from 50K genotype data of all 1800 beef cattle. Individuals are grouped by their population, as described in "Materials and Methods" section

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 the rare allele in their update library. On the other hand, 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 CRs 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 added 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 Fig. 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 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 the total number of variants was the fewest with Angus. Even though CRs of genotypes "AB" and "BB" were poorest for Angus with the MAF class (0, 1%), the number of such rare variants was extremely small and all methods were capable of imputing well for all MAF classes with Angus.

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 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 fivefold CV, the across-breed training dataset of actual 50K genotypes corresponded to our reference panel of 1440 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.

In Table 4, 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.

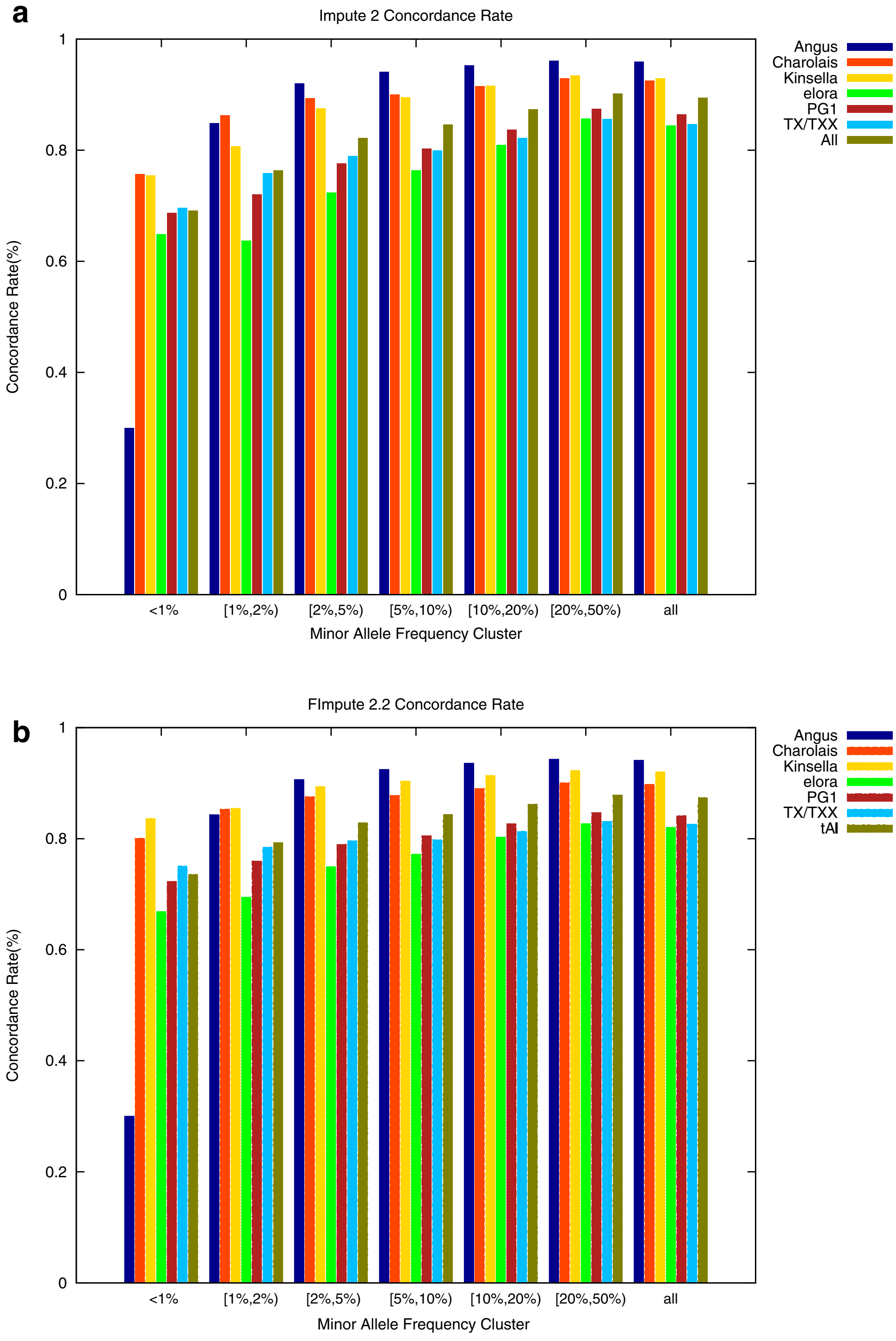
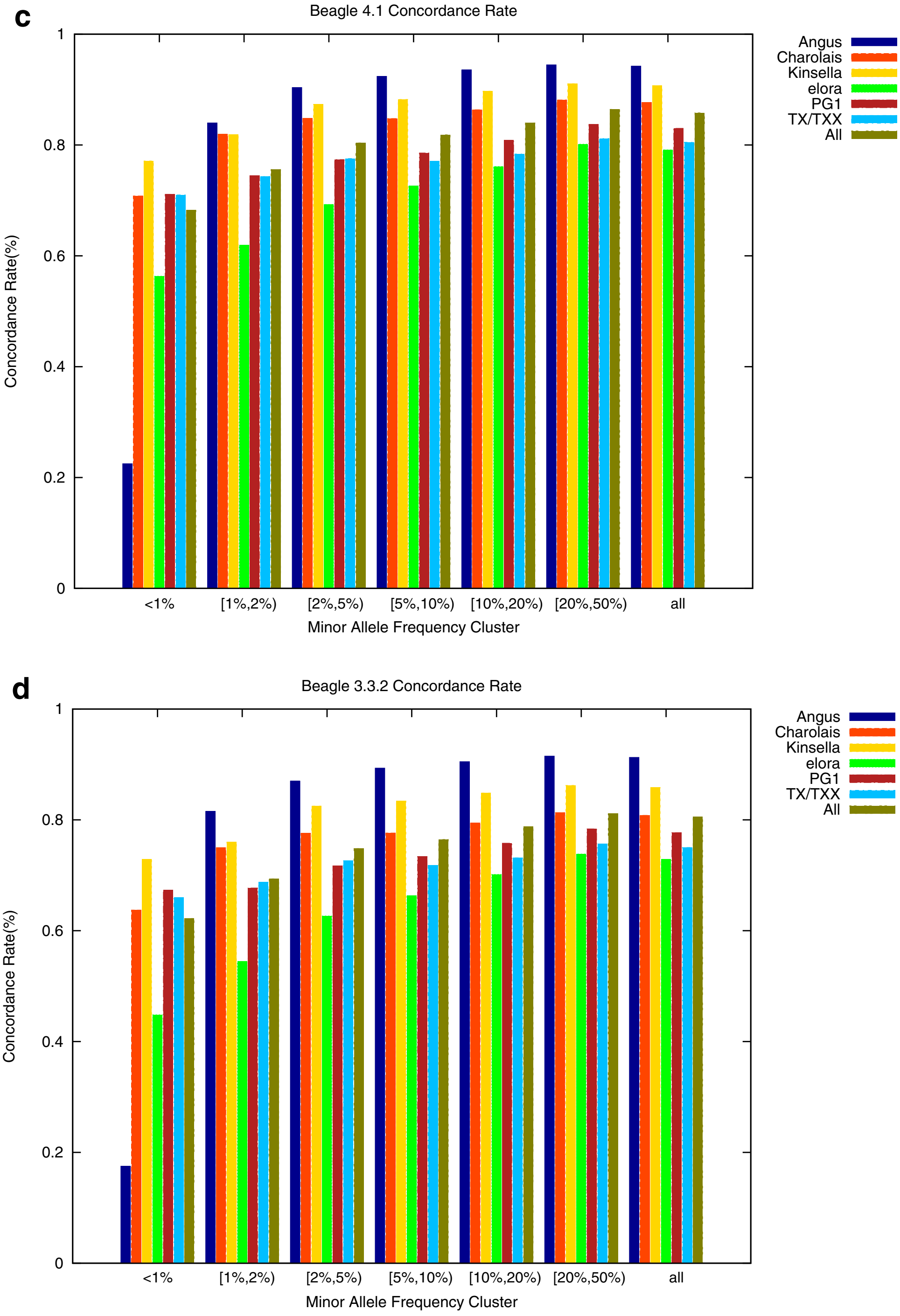


Fig. 2 Effects of MAF of untyped SNPs on imputing genotypes carrying minor allele (MA). **a** Concordance rates of genotypes “AB” and “BB” for Impute 2. **b** Concordance rates of genotypes “AB” and “BB” for FImpute. **c** Concordance rates of genotypes “AB” and “BB” for Beagle 4.1. **d** Concordance rates of genotypes “AB” and “BB” for Beagle 3.3.2. **e** Concordance rates of genotypes “AB” and “BB” for MaCH. **f** Concordance rates of genotypes “AB” and “BB” for Bimbam

Columns that have imputation methods-50K as titles report prediction accuracies when using imputed 50K of the imputation method as validation datasets. A slight increase in *r* or genomic prediction accuracy 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 for the across-breed genomic prediction. 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 are considered.

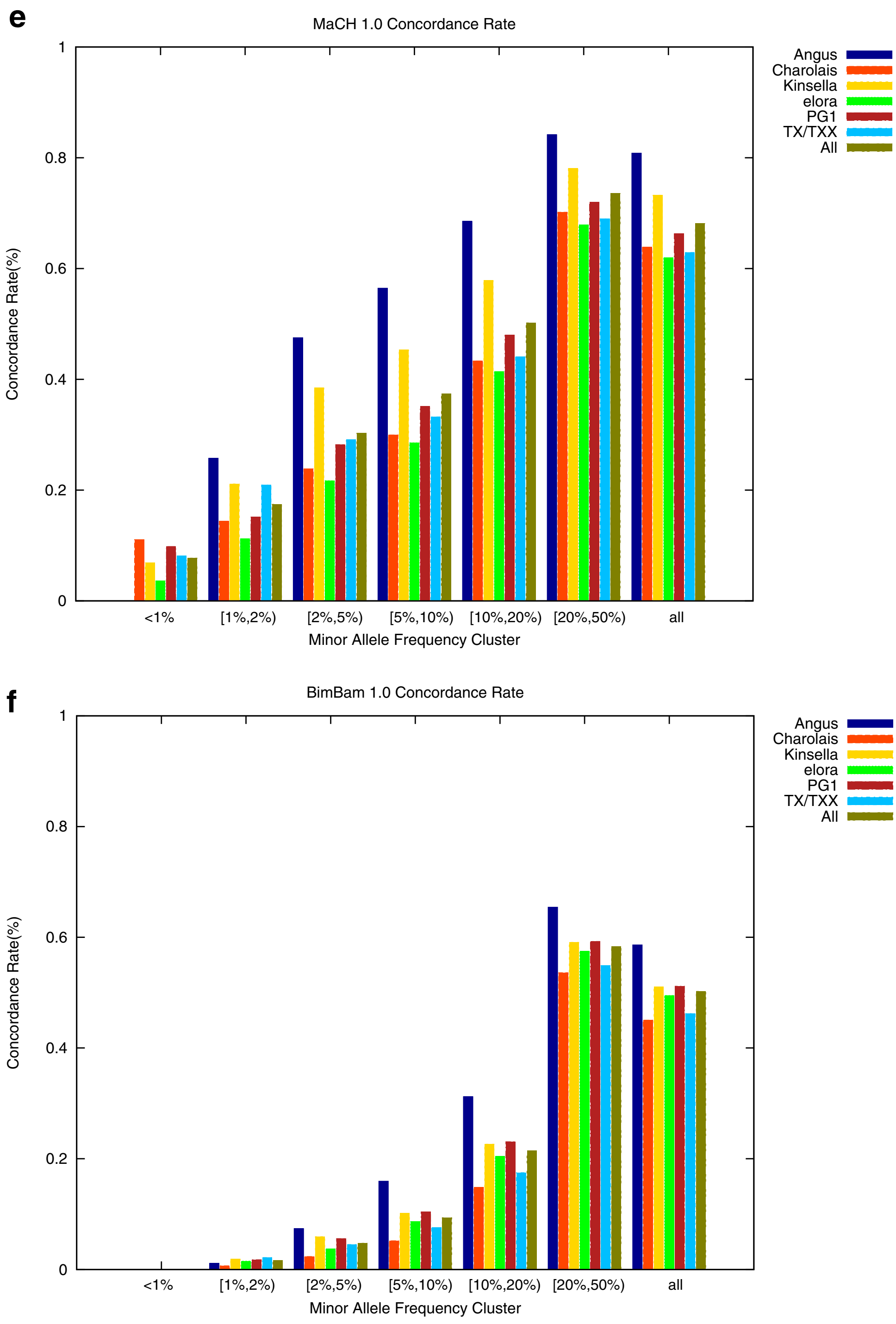
In comparison of genomic prediction accuracies of 50K to that of imputed 50K for across-breed genomic

Fig. 2 continued



prediction, imputed 50K genomic prediction results from all the imputation methods except for Bimbam gave comparable accuracies *r* to the actual 50K results using both GBLUP and BayesB. For purebred Charolais, the most accurate mean *r* were 0.24 using BayesB on imputed 50K via MaCH, although the mean CR 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 than 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 accurate mean *r* was 0.19 using BayesB on imputed 50K

Fig. 2 continued



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 5. Similarly, genomic prediction accuracies of actual 50K, actual 6K and imputed 50K are comparable. 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.

Table 3 Distribution of the number of genotypes (AB and BB) that carry the minor allele "B" among minor allele frequency classes in different populations

Population	MAF class						
	(0, 1%)	[1%, 2%]	[2%, 5%]	[5%, 10%]	[10%, 20%]	[20%, 50%]	All
Angus	40	1429	21,000	103,384	490,788	2,792,475	3,409,116
Charolais	226	4489	37,597	135,664	539,437	2,742,995	3,460,408
Kinsella	306	4190	38,420	134,292	533,583	2,806,099	3,516,890
Elora	444	3903	34,136	127,569	535,479	2,822,135	3,523,666
PG1	661	5066	40,973	137,047	542,973	2,834,610	3,561,330
TX/TXX	679	6402	49,277	150,388	559,059	2,798,682	3,564,487
All	2356	25,479	22,1403	788,344	3,201,319	16,796,996	21,035,897

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 better accuracies for purebred Angus under BayesB and for crossbred PG1 under GBLUP. For both across and within-breed genomic predictions based on either actual 50K, actual 6K and imputed 50K SNPs, purebred populations (Angus and Charolais) had relatively higher prediction accuracies than that of crossbred populations Kinsella, Elora, PG1, TX/TXX.

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 [10, 26, 28, 56, 60, 61, 62, 63]. Bouwman and Veerkamp [60] 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. Since rare alleles might be under-represented in a single population, as shown in Table 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 Fig. 2a–f, 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 [25]. 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 [35], as evidenced in Fig. 2f where Bimbam did not make any correct predictions for genotypes carrying rare alleles (MAF $< 1\%$). Figure 1 shows 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 [64, 65]. Both fastPHASE and Bimbam rely on estimation of clusters in their model settings via the MLE. Recently, Feller et al. [66] 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 [67], over 97% with Fimpute [7], over 98% with Fimpute [68]; 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

Table 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, Charolais, Kinsella, Elora, PG1, TX/TXX validation groups

Population	Actual 50K		Actual 6K	
	BayesB	GBLUP	BayesB	GBLUP
Angus	0.18 ± 0.05	0.21 ± 0.04	0.13 ± 0.03	0.16 ± 0.03
Charolais	0.22 ± 0.05	0.21 ± 0.06	0.14 ± 0.05	0.18 ± 0.06
Kinsella	0.11 ± 0.08	0.08 ± 0.06	0.11 ± 0.06	0.08 ± 0.07
Elora	0.09 ± 0.06	0.16 ± 0.05	0.05 ± 0.04	0.15 ± 0.04
PG1	0.10 ± 0.05	−0.04 ± 0.06	0.12 ± 0.06	−0.01 ± 0.06
TX/TXX	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	Impute 2-50K		FImpute 2.2-50K		Beagle 4.1-50K	
	BayesB	GBLUP	BayesB	GBLUP	BayesB	GBLUP
Angus	0.18 ± 0.05	0.20 ± 0.04	0.16 ± 0.05	0.18 ± 0.02	0.18 ± 0.05	0.20 ± 0.04
Charolais	0.22 ± 0.05	0.20 ± 0.06	0.23 ± 0.07	0.22 ± 0.07	0.21 ± 0.06	0.20 ± 0.07
Kinsella	0.10 ± 0.08	0.08 ± 0.06	0.09 ± 0.07	0.10 ± 0.05	0.11 ± 0.08	0.08 ± 0.06
Elora	0.09 ± 0.06	0.15 ± 0.06	0.11 ± 0.06	0.16 ± 0.03	0.07 ± 0.05	0.14 ± 0.05
PG1	0.10 ± 0.05	−0.04 ± 0.06	0.12 ± 0.06	−0.02 ± 0.07	0.11 ± 0.06	−0.04 ± 0.06
TX/TXX	0.18 ± 0.04	0.16 ± 0.04	0.16 ± 0.03	0.18 ± 0.06	0.17 ± 0.03	0.16 ± 0.04
All	0.14 ± 0.03	0.11 ± 0.03	0.14 ± 0.03	0.12 ± 0.03	0.14 ± 0.03	0.11 ± 0.03

Population	Beagle 3.3.2-50K		MaCH-50K		Bimbam-50K	
	BayesB	GBLUP	BayesB	GBLUP	BayesB	GBLUP
Angus	0.17 ± 0.05	0.20 ± 0.04	0.16 ± 0.05	0.19 ± 0.04	0.13 ± 0.04	0.15 ± 0.03
Charolais	0.22 ± 0.07	0.20 ± 0.08	0.24 ± 0.06	0.21 ± 0.07	0.18 ± 0.04	0.18 ± 0.05
Kinsella	0.10 ± 0.06	0.07 ± 0.05	0.11 ± 0.06	0.08 ± 0.05	0.08 ± 0.05	0.05 ± 0.04
Elora	0.07 ± 0.06	0.15 ± 0.05	0.05 ± 0.07	0.14 ± 0.06	0.06 ± 0.05	0.15 ± 0.05
PG1	0.10 ± 0.05	−0.04 ± 0.06	0.12 ± 0.07	−0.04 ± 0.06	0.10 ± 0.06	−0.07 ± 0.04
TX/TXX	0.19 ± 0.03	0.16 ± 0.03	0.15 ± 0.03	0.15 ± 0.03	0.12 ± 0.03	0.14 ± 0.03
All	0.14 ± 0.03	0.11 ± 0.03	0.13 ± 0.03	0.11 ± 0.03	0.11 ± 0.02	0.09 ± 0.03

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 animals pooled from all six populations (training groups consist of 1440 animals across six populations while validation groups contain 360 animals from each of the six populations)

Accuracy is measured by Pearson's correlation coefficient between the estimated genomic breeding values and the adjusted phenotype values in the validation group

cattle reported by Piccoli et al. [69], Ventura et al. [70], and Chud et al. [71]. 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 as Holsteins have a relatively small effective population size [72]. The design of the Illumina 6K chip is another factor that results in different accuracies of imputation in various breeds and populations [71]. The SNPs on this panel were selected to provide optimized imputation in dairy breeds [1] 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. [73] where a larger number of beef cattle (over 5000) from the same data pool as ours were used for evaluation of accuracy of genomic prediction for

Table 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 BayesB and GBLUP for Angus, Charolais, Kinsella, Elora, PG1, TX/TXX validation groups

Population	Actual 50K		Actual 6K	
	BayesB	GBLUP	BayesB	GBLUP
Angus	0.24 ± 0.03	0.25 ± 0.01	0.23 ± 0.05	0.26 ± 0.02
Charolais	0.21 ± 0.06	0.20 ± 0.06	0.19 ± 0.06	0.20 ± 0.05
Kinsella	0.10 ± 0.06	0.12 ± 0.06	0.11 ± 0.07	0.13 ± 0.06
Elora	0.17 ± 0.05	0.18 ± 0.05	0.16 ± 0.02	0.18 ± 0.04
PG1	0.13 ± 0.06	0.16 ± 0.08	0.15 ± 0.03	0.14 ± 0.07
TX/TXX	0.17 ± 0.04	0.18 ± 0.04	0.13 ± 0.05	0.14 ± 0.04

Population	Impute 2-50K		FImpute 2.2-50K		Beagle 4.1-50K	
	BayesB	GBLUP	BayesB	GBLUP	BayesB	GBLUP
Angus	0.25 ± 0.02	0.25 ± 0.01	0.24 ± 0.02	0.24 ± 0.01	0.24 ± 0.03	0.25 ± 0.01
Charolais	0.21 ± 0.06	0.20 ± 0.06	0.21 ± 0.06	0.21 ± 0.07	0.21 ± 0.06	0.20 ± 0.06
Kinsella	0.11 ± 0.06	0.12 ± 0.06	0.10 ± 0.06	0.12 ± 0.06	0.14 ± 0.04	0.12 ± 0.06
Elora	0.15 ± 0.05	0.16 ± 0.04	0.14 ± 0.05	0.17 ± 0.03	0.04 ± 0.06	0.16 ± 0.04
PG1	0.13 ± 0.06	0.15 ± 0.07	0.14 ± 0.06	0.15 ± 0.07	0.11 ± 0.06	0.16 ± 0.08
TX/TXX	0.16 ± 0.04	0.18 ± 0.04	0.15 ± 0.04	0.18 ± 0.05	0.16 ± 0.04	0.17 ± 0.04

Population	Beagle 3.3.2-50K		MaCH-50K		Bimbam-50K	
	BayesB	GBLUP	BayesB	GBLUP	BayesB	GBLUP
Angus	0.24 ± 0.02	0.25 ± 0.01	0.24 ± 0.02	0.24 ± 0.02	0.25 ± 0.03	0.26 ± 0.01
Charolais	0.21 ± 0.06	0.20 ± 0.07	0.22 ± 0.06	0.21 ± 0.06	0.22 ± 0.06	0.22 ± 0.06
Kinsella	0.11 ± 0.06	0.13 ± 0.06	0.11 ± 0.06	0.13 ± 0.06	0.12 ± 0.06	0.14 ± 0.05
Elora	0.13 ± 0.05	0.15 ± 0.04	0.15 ± 0.05	0.16 ± 0.05	0.13 ± 0.04	0.16 ± 0.05
PG1	0.14 ± 0.06	0.16 ± 0.08	0.14 ± 0.07	0.16 ± 0.08	0.14 ± 0.07	0.16 ± 0.08
TX/TXX	0.15 ± 0.05	0.17 ± 0.05	0.14 ± 0.05	0.16 ± 0.04	0.12 ± 0.05	0.14 ± 0.05

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. Both training and validation groups consist of animals from the same breed/population (training groups consist of 240 animals from the within-breed population, while validation groups contain 60 animals from the same breed)

Accuracy is measured by Pearson's correlation coefficient between the estimated genomic breeding values and the adjusted phenotype values in the validation group

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 [73]. Binsbergen et al. [74] 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 genotypes may not necessarily lead to an increase in accuracy of genomic prediction with the current SNP panel information and statistical methods [75].

Previous studies [49, 76, 77] have shown evidence that RFI is a complex trait likely to be controlled by many SNPs with small effects. Therefore, genomic 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. [7], studying genomic predictions of fat percentage using dairy cattle. 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 [11, 78, 76, 79]. 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 [80–82]. 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 Fig. 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. [83] 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. [84] 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 unlikely persist across populations and suggested high density marker set 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 Fig. 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 [85–87] 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. A closer relatedness between training and validation leads to higher persistency of CS among animals [86, 87], which will improve the accuracy of both imputation and genomic prediction. 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 which comprised more closely related individuals results in an increase of CS, and its persistence is higher than that of across-breed genomic prediction [87], which was shown by Chen et al. [49] 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 [21, 28, 33]. From the plot of PCA in Fig. 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 was lower, which has been demonstrated in previous studies with dairy cattle populations [83, 85].

The density of DNA markers is expected to affect accuracy of genomic predictions as use of genotypes in a high-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 the density of SNP panels by simply adding SNPs with high MAF will unlikely improve LD between SNPs and QTL of rare MAF [88, 89], and further studies are needed to make better use of existing higher density SNP panels and to design better high-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. [77], Erbe et al. [90] and Ertl et al. [91] 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 high-density SNPs or whole-genome SNPs on other traits in larger populations of beef cattle.

Conclusion

In this review, we 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. The six genotype imputation methods fall into three major categories: (1) methods based on Li and Stephens' "PAC" framework (2003); (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" (2008). 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 one that uses 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 v2. Computational burdens scale quadratically with the number of hidden states in "PAC"-based models. Our simulation studies confirmed that MAF plays a key role in the accuracy of imputation. As MAF 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 shows advantages over other methods in terms of running time and imputing rare alleles. Bimbam's poor performance is likely due to MEL 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. 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.

Acknowledgements The SNP genotype and phenotypic data were collected under various projects led by Dr. John Basarab (PG1 and TX/TXX), Dr. Stephen Miller (Elora), Dr. Stephen Moore (Kinsella) and Dr. Changxi Li (Angus and Charolais), and the data were consolidated through the Genome Canada Canadian Cattle Genome Project led by Drs. Paul Stothard, Stephen Moore, and Stephen Miller. This work was supported in part by the Alberta Livestock and Meat Agency (ALMA) (2014F047R) awarded to CL, and a NSERC Discovery Grant awarded to PS. This research was enabled in part by support provided by WestGrid (www.westgrid.ca) and Compute Canada Calcul Canada (www.computecanada.ca).

References

1. Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S, Gietzen KJ, Hayes BJ, Lawley CT, Sonstegard TS, Van Tassell CP (2012) Design of a bovine low-density SNP array optimized for imputation. PLoS ONE 7(3):e34130
2. Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS (2005) A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet 37(5):549–554
3. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, Sonstegard TS, Van Tassell CP (2009) Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE 4(4):e5350
4. Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL (2006) Whole-genome genotyping with the single-base extension assay. Nat Methods 3(1):31
5. Daetwyler HD, Capitan A, Pausch H, Stothard P, Van Binsbergen R, Brøndum RF, Liao X, Djari A, Rodriguez SC, Grohs C, Esquerré D (2014) Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. Nat Genet 46(8):858–865
6. McClure M, Sonstegard T, Wiggans G, Van Tassell CP (2012) Imputation of microsatellite alleles from dense SNP genotypes for parental verification. Front Genet 3(140):10–3389
7. Chen L, Li C, Sargolzaei M, Schenkel F (2014) Impact of genotype imputation on the performance of GBLUP and Bayesian methods for genomic prediction. PLoS ONE 9(7):e101544
8. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4):1819–1829
9. Goddard ME, Hayes BJ, Meuwissen THE (2011) Using the genomic relationship matrix to predict the accuracy of genomic selection. J Anim Breed Genet 128(6):409–421
10. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11(7):499–511
11. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Invited review: genomic selection in dairy cattle: Progress and challenges. J Dairy Sci 92(2):433–443
12. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437(7063):1299–1320
13. Howie BN, Marchini J, Stephens M (2011) Genotype imputation with thousands of genomes. G3 1(6):457–470
14. Yu Z, Schaid DJ (2007) Methods to impute missing genotypes for population data. Hum Genet 122(5):495–504
15. Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet 81(5):1084–1097
16. Guan Y, Stephens M (2008) Practical issues in imputation-based association mapping. PLoS Genet 4(12):e1000279
17. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78(4):629–644
18. Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3(7):e114
19. Wen X, Stephens M (2010) Using linear predictors to impute allele frequencies from summary or pooled genotype data. Ann Appl Stat 4(3):1158
20. Chi EC, Zhou H, Chen GK, Del Vecchyo DO, Lange K (2013) Genotype imputation via matrix completion. Genome Res 23(3):509–518
21. Hickey JM, Kinghorn BP, Tier B, van der Werf JH, Cleveland MA (2012) A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet Sel Evol 44(1):9
22. Cheung CY, Thompson EA, Wijsman EM (2013) GIGI: an approach to effective imputation of dense genotypes on large pedigrees. Am J Hum Genet 92(4):504–516

23. Pimentel EC, Wensch-Dorendorf M, König S, Swalve HH (2013) Enlarging a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture. Genet Sel Evol 45(1):12
24. Saad M, Wijsman EM (2014) Combining family-and population-based imputation data for association analysis of rare and common variants in large pedigrees. Genet Epidemiol 38(7):579–590
25. Browning SR (2008) Missing data imputation and haplotype phase inference for genome-wide association studies. Hum Genet 124(5):439–450
26. Halperin E, Stephan DA (2009) SNP imputation in association studies. Nat Biotechnol 27(4):349–351
27. Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. Nat Rev Genet 12(10):703–714
28. Calus MPL, Bouwman AC, Hickey JM, Veerkamp RF, Mulder HA (2014) Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: a review of livestock applications. Animal 8(11):1743–1753
29. Mulder HA, Calus MPL, Druet T, Schrooten C (2012) Imputation of genotypes with low-density chips and its effect on reliability of direct genomic values in Dutch Holstein cattle. J Dairy Sci 95(2):876–889
30. Pimentel ECG, Edel C, Emmerling R, Götz KU (2015) How imputation errors bias genomic predictions. J Dairy Sci 98(6):4131–4138
31. Li N, Stephens M (2003) Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics 165(4):2213–2233
32. Sargolzaei M, Chesnais JP, Schenkel FS (2014) A new approach for efficient genotype imputation using information from relatives. BMC Genom 15(1):478
33. Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, Olason PI, Ingason A, Steinberg S, Rafnar T, Sulem P (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. Nat Genet 40(9):1068–1075
34. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I (2009) Whole population, genome-wide mapping of hidden relatedness. Genome Res 19(2):318–326
35. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5(6):e1000529
36. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39(7):906–913
37. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34(8):816–834
38. Howie BN, Fuchsberger C, Stephens M, Marchini J, Abecasis GR (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44(8):955–959
39. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164(4):1567–1587
40. Kimmel G, Shamir R (2005) A block-free hidden Markov model for genotypes and its application to disease association. J Comput Biol 12(10):1243–1260
41. Guan Y (2014) Detecting structure of haplotypes and local ancestry. Genetics 196(3):625–642
42. Browning SR (2006) Multilocus association mapping using variable-length Markov chains. Am J Hum Genet 78(6):903–913
43. Browning BL, Browning SR (2013) Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194(2):459–471
44. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I, McQuillan R (2014) A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet 10(4):e1004234
45. Delaneau O, Marchini J, Zagury JF (2012) A linear complexity phasing method for thousands of genomes. Nat Methods 9(2):179–181
46. Lu D, Akanno EC, Crowley JJ, Schenkel FS, Li H, De Pauw M, Moore SS, Wang Z, Li C, Stothard P, Plastow G, Miller SP, Basarab JA (2016) Accuracy of genomic predictions for feed efficiency traits of beef cattle using 50K and imputed HD genotypes. J Anim Sci 94(4):1342–1353
47. Koch RM, Swiger LA, Chambers D, Gregory KE (1963) Efficiency of feed use in beef cattle. J Anim Sci 22(2):486–494
48. Basarab JA, Colazo MG, Ambrose DJ, Novak S, McCartney D, Baron VS (2011) Residual feed intakeadjusted for backfat thickness and feeding frequency is independent of fertility in beef heifers. Can J Anim Sci 91(4):573–584
49. Chen L, Schenkel F, Vinsky M, Crews DH, Li C (2013) Accuracy of predicting genomic breeding values for residual feed intake in Angus and Charolais beef cattle. J Anim Sci 91(7):4669–4678
50. VanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91(11):4414–4423
51. Fernando RL, Garrick DJ (2008) GenSel-User manual for a portfolio of genomic selection related analyses. Animal Breeding and Genetics, Iowa State University, Ames
52. Nejati-Javaremi A, Smith C, Gibson JP (1997) Effect of total allelic relationship on accuracy of evaluation and response to selection. J Anim Sci 75(7):1738–1745
53. Sargolzaei M, Schenkel FS, VanRaden PM (2009) GEBV: genomic breeding value estimator for livestock. In: Technical report to the Dairy Cattle Breeding and Genetics Committee, University of Guelph, Guelph
54. VanRaden PM, Van Tassell CP, Wiggans GR, Sonstegard TS, Schnabel RD, Taylor JF, Schenkel FS (2009) Invited review: reliability of genomic predictions for North American Holstein bulls. J Dairy Sci 92(1):16–24
55. Colleau JJ (2002) An indirect approach to the extensive calculation of relationship coefficients. Genet Sel Evol 34(4):409–422
56. Hickey JM, Crossa J, Babu R, de los Campos G (2012) Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci 52(2):654–663
57. Haldane JBS (1919) The combination of linkage values and the calculation of distances between the lociof linked factors. J Genet 8(29):299–309
58. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11(6):415–425
59. van Binsbergen R, Bink MCAM, Calus MP, van Eeuwijk FA, Hayes BJ, Hulsegge I, Veerkamp RF (2014) Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol 46(1):41
60. Bouwman AC, Veerkamp RF (2014) Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. BMC Genet 15(1):105
61. Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, Scheet P (2009) Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet 84(2):235–250
62. Huang Y, Hickey JM, Cleveland MA, Maltecca C (2012) Assessment of alternative genotyping strategies to maximize imputation accuracy at minimal cost. Genet Sel Evol 44(1):25
63. Khatkar MS, Moser G, Hayes BJ, Raadsma HW (2012) Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genom 13(1):538
64. Redner RA, Walker HF (1984) Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev 26:195–239
65. Wasserman L (2012) Mixture models: the twilight zone of statistics. https://normaldeviate.wordpress.com/2012/08/04/

mixture-models-the-twilight-zone-of-statistics/. Accessed 29 June 2016

66. Feller A, Greif E, Miratrix L, Pillai N (2016) Principal stratification in the Twilight Zone: weakly separated components in finite mixture models. arXiv preprint, arXiv:1602.06595
67. Berry DP, McClure MC, Mullen MP (2014) Within- and across-breed imputation of high-density genotypes in dairy and beef cattle from medium- and low- density genotypes. J Anim Breed Genet 131(3):165–172
68. Sargolzaei M, Schenkel FS, Chesnais J (2011) Accuracy of imputed 50k genotypes from 3k and 6k chips using FImpute version 2. In: Dairy Cattle Breeding and Genetics Committee Meeting, September, pp 1–9
69. Piccoli M, Braccini J, Cardoso FF, Sargolzaei M, Larmer SG, Schenkel FS (2014) Accuracy of genome-wide imputation in Braford and Hereford beef cattle. BMC Genet 15(1):157
70. Ventura RV, Lu D, Schenkel FS, Wang Z, Li C, Miller SP (2014) Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle. J Anim Sci 92(4):1433–1444
71. Chud TC, Ventura RV, Schenkel FS, Carvalheiro R, Buzanskas ME, Rosa JO, de Alvarenga Mudadu M, da Silva MVG, Mokry FB, Marcondes CR, Regitano LC (2015) Strategies for genotype imputation in composite beef cattle. BMC Genet 16(1):99
72. Hozé C, Fouilloux MN, Venot E, Guillaume F, Dassonneville R, Fritz S, Ducrocq V, Phocas F, Boichard D, Croiseau P (2013) High-density marker imputation accuracy in sixteen French cattle breeds. Genet Sel Evol 45(1):33
73. Li C, Chen L, Vinsky M, Crowley J, Miller SP, Plastow G, Basarab J, Stothard P (2015) Genomic prediction for feed efficiency traits based on 50K and imputed high density SNP genotypes in multiple breed populations of Canadian beef cattle (Abstract). J Anim Sci 94(E-Suppl. 5)/J Dairy Sci 99(E-Supp. 1)
74. van Binsbergen R, Calus MP, Bink MC, van Eeuwijk FA, Schrooten C, Veerkamp RF (2015) Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol 47(1):1–13
75. Saatchi M, Beever JE, Decker JE, Faulkner DB, Freetly HC, Hansen SL, Yampara-Iquise H, Johnson KA, Kachman SD, Kerley MS, Kim J (2014) QTLs associated with dry matter intake, metabolic mid-test weight, growth and feed efficiency have little overlap across 4 beef cattle studies. BMC Genom 15(1):1004
76. Mujibi FDN, Nkrumah JD, Durunna ON, Stothard P, Mah J, Wang Z, Basarab J, Plastow G, Crews DH, Moore SS (2011) Accuracy of genomic breeding values for residual feed intake in crossbred beef cattle. J Anim Sci 89(11):3353–3361
77. Pryce JE, Arias J, Bowman PJ, Davis SR, Macdonald KA, Waghorn GC, Wales WJ, Williams YJ, Spelman RJ, Hayes BJ (2012) Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. J Dairy Sci 95(4):2108–2119
78. VanRaden PM, Null DJ, Sargolzaei M, Wiggans GR, Tooker ME, Cole JB, Sonstegard TS, Connor EE, Winters M, van Kaam JBCHM, Valentini A (2013) Genomic imputation and evaluation using high-density Holstein genotypes. J Dairy Sci 96(1):668–678
79. Macdonald KA, Pryce JE, Spelman RJ, Davis SR, Wales WJ, Waghorn GC, Williams YJ, Marett LC, Hayes BJ (2014) Holstein-Friesian calves selected for divergence in residual feed intake during growth exhibited significant but reduced residual feed intake divergence in their first lactation. J Dairy Sci 97(3):1427–1435
80. Tiezzi F, Maltecca C (2015) Accounting for trait architecture in genomic predictions of US Holstein cattle using a weighted realized relationship matrix. Genet Sel Evol 47:24
81. Daetwyler HD, Pong-Wong R, Villanueva B, Woolliams JA (2010) The impact of genetic architecture on genome-wide evaluation methods. Genetics 185(3):1021–1031
82. Druet T, Macleod IM, Hayes BJ (2014) Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. Heredity 112(1):39–47
83. Lund MS, Su G, Janss L, Guldbrandtsen B, Brøndum RF (2014) Genomic evaluation of cattle in a multi-breed context. Livest Sci 166:101–110
84. De Roos APW, Hayes BJ, Goddard ME (2009) Reliability of genomic predictions across multiple populations. Genetics 183(4):1545–1553
85. Habier D, Fernando RL, Dekkers JCM (2007) The impact of genetic relationship information on genome-assisted breeding values. Genetics 177(4):2389–2397
86. Habier D, Fernando RL, Garrick DJ (2013) Genomic BLUP decoded: a look into the black box of genomic prediction. Genetics 194(3):597–607
87. Sun X, Fernando RL, Dekkers JCM (2016) Contributions of linkage disequilibrium and co-segregation information to the accuracy of genomic prediction. Genet Sel Evol 48:77
88. Sun X, Fernando RL, Garrick DJ, Dekkers J (2015) Improved accuracy of genomic prediction for traits with rare QTL by fitting haplotypes. Anim Ind Rep 661(1):86
89. Clark SA, Hickey JM, Van der Werf JH (2011) Different models of genetic variation and their effect on genomic evaluation. Genet Sel Evol 43(1):18
90. Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME (2012) Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci 95(7):4114–4129
91. Ertl J, Edel C, Emmerling R, Pausch H, Fries R, Götz KU (2014) On the limited increase in validation reliability using high-density genotypes in genomic best linear unbiased prediction: observations from Fleckvieh cattle. J Dairy Sci 97(1):487–496