12-13-2018

Comparing strategies for selection of low-density SNPs for imputation-mediated genomic prediction in U.S. Holsteins

Jun He
University of Nebraska–Lincoln

Jiaqi Xu
University of Nebraska-Lincoln, jxu15@unl.edu

Xiao-Lin Wu
University of Wisconsin, xwu8@wisc.edu

Stewart Bauck
GeneSeek (a Neogen Company)

Jungjae Lee
University of Nebraska–Lincoln

See next page for additional authors

Published in *Genetica*, 2017. doi 10.1007/s10709-017-0004-9

Submitted 18 September 2017; accepted 8 December 2017; published 14 December 2017.

Comparing strategies for selection of low-density SNPs for imputation-mediated genomic prediction in U. S. Holsteins

Jun He,[1,2,3] Jiaqi Xu,[3,4] Xiao-Lin Wu,[3,5] Stewart Bauck,[3] Jungjae Lee,[1] Gota Morota,[1] Stephen D. Kachman,[4] and Matthew L. Spangler[1]

1 Department of Animal Science, University of Nebraska–Lincoln, Lincoln, NE 68583, USA
2 College of Animal Science and Technology, Hunan Agricultural University, Changsha 410128, Hunan, China
3 Biostatistics and Bioinformatics, Neogen GeneSeek, Lincoln, NE 68504, USA
4 Department of Statistics, University of Nebraska–Lincoln, Lincoln, NE 68583, USA
5 Department of Animal Sciences, University of Wisconsin, Madison, WI 53706, USA

Jun He and Jiaqi Xu contributed equally and are co-first authors.

Corresponding authors — Xiao-Lin Wu, nwu@neogen.com , xwu8@wisc.edu ; Matthew L. Spangler, mspangler2@unl.edu

Abstract

SNP chips are commonly used for genotyping animals in genomic selection but strategies for selecting low-density (LD) SNPs for imputation-mediated genomic selection have not been addressed adequately. The main purpose of the present study was to compare the performance of eight LD (6K) SNP panels, each selected by a different strategy exploiting a combination of three major factors: evenly-spaced SNPs, increased minor allele frequencies, and SNP-trait associations either for single traits independently or for all the three traits jointly. The imputation accuracies from 6K to 80K SNP genotypes were between 96.2 and 98.2%. Genomic prediction accuracies obtained using imputed 80K genotypes were between 0.817 and 0.821 for daughter pregnancy rate, between 0.838 and 0.844 for fat yield, and between 0.850 and 0.863 for milk yield. The two SNP panels optimized on the three major factors had the highest genomic prediction accuracy (0.821–0.863), and these accuracies were very close to those obtained using observed 80K

http://orcid.org/0000-0002-0414-0209 June He
http://orcid.org/0000-0002-5604-9220 Xiao-Lin Wu

genotypes (0.825–0.868). Further exploration of the underlying relationships showed that genomic prediction accuracies did not respond linearly to imputation accuracies, but were significantly affected by genotype (imputation) errors of SNPs in association with the traits to be predicted. SNPs optimal for map coverage and MAF were favorable for obtaining accurate imputation of genotypes whereas trait-associated SNPs improved genomic prediction accuracies. Thus, optimal LD SNP panels were the ones that combined both strengths. The present results have practical implications on the design of LD SNP chips for imputation-enabled genomic prediction.

Keywords: Holstein, Imputation, Genomic prediction, Low-density SNP chips

Abbreviations

ANOVA	Analysis of variance
DPR	Daughter pregnancy rate
FY	Fat yield
GEBV	Genomic-estimated breeding value
GER	Genotype (imputation) error rate
GPA	Genomic prediction accuracy
GS	Genomic selection
HD	High-density
LD	Low-density
LGPA	Loss in genomic prediction accuracy
MAF	Minor allele frequencies
MCMC	Markov chain Monte Carlo
MD	Moderate-density
MOLO	Multiple-objective, local-optimization
MY	Milk yield
PTAs	Predicted transmitting abilities
RGPA	Relative genomic prediction accuracy
RTMGL	Relative total maximum gap length
TMGL	Total maximum gap length

Introduction

The availability of whole-genome DNA information has opened the door for genome-enabled genetic improvement in agricultural animals (Hayes and Goddard 2001; van der Werf 2013), and SNP arrays are commonly used for genotyping animals in genomic selection. Though genotyping cost per SNP has been drastically decreased in the past 10 years, use of moderate-density (**MD**) or high-density (**HD**)

SNP chips is still expensive for animal breeding and selection programs in practice. Consequently, when the advantage of GS is compared to traditional genetic selection in terms of genetic gain per unit of cost, it is clear that low-density (**LD**) SNP chips are preferred in order to fully exploit the genetic gain advantages of GS because they are cost-effective (Habier et al. 2009; Weigel et al. 2009; Biochard et al. 2012; Bolormaa et al. 2015).

Often, LD-SNP chips are selected either based on their map locations, such as evenly-spaced SNPs (Habier et al. 2009; Wiggans et al. 2012), or selected based on their associated effects (Weigel et al. 2009). Recently, a multiple-objective, local-optimization (**MOLO**) algorithm was proposed to select LD SNPs, which is capable of selecting SNPs to meet multiple objectives, which included map coverage, minor allele frequency (**MAF**), map gaps, and many more criteria (Wu et al. 2016). Nevertheless, genomic prediction using LD SNP genotypes directly can suffer from information loss due to insufficient genome coverage, which in turn can result in substantially decreased prediction accuracy (Weigel et al. 2009). Besides, it has been discovered that selected SNPs based on a certain statistical cut-off tend to explain only a small portion of its total genetic variation for a quantitative trait of polygenic inheritance (Manolio et al. 2009; Eichler et al. 2010; Zuk et al. 2012). Alternatively, MD or HD genotypes can be imputed based on a set of known LD SNP genotypes and then used for genomic prediction with increased accuracy (Erbe et al. 2012; Pimentel et al. 2013). This type of approaches is referred to as imputation-mediated genomic prediction hereafter. Unlike genomic prediction using trait-specific LD SNP chips, imputation-mediated genomic prediction allows the use of a common, multiple-trait SNP chip, which not only saves over-head costs associated with chip design and manufacturing, thus simplifying the practicality of genotyping by providing one assay for multiple economically relevant traits, but it also can minimize the loss of genomic prediction accuracy (LGPA) as compared to that using observed MD or HD SNP genotypes.

Consider GS in dairy cattle in the USA, for example. The genomic prediction system (i.e., linear prediction equations with SNPs as the predictors) was built on 50K (now 66K) SNP genotypes (Wiggans et al. 2009). With the genomic prediction system for Holsteins in place, it is possible to genotype these candidate animals using a LD SNP chip and then impute to 50K genotypes for these animals, instead of

genotyping all candidate animals using the bovine 50K SNP chip. Finally, genomic-estimated breeding values (**GEBV**s) are computed using imputed 50K genotypes for candidate animals, according to SNP effects estimated on observed 50K genotypes in the training population. Therefore, selection of optimal LD SNPs is central to imputation-mediated genomic prediction. Although there were previous studies on the accuracies of imputation from LD SNP genotypes to MD- and HD-SNP genotypes (e.g., Boichard et al. 2012), selection of LD SNPs for imputation-mediated GS has not been addressed adequately. Wu et al. (2016) investigated the effects of imputation-mediated genomic prediction using trait-associated LD SNPs, but there are still many important pieces missing in the portrait of imputation-mediated GS, such as lacking of a direct comparison of trait-association panels to map-optimal panels relative to prediction accuracy and of an understanding as to how genomic prediction accuracies respond to imputation accuracies.

The objectives of the current study were to evaluate the performance (i.e., imputation and genomic prediction accuracies) of eight sets of imputed 80K SNP genotypes from LD SNPs, each derived using a different strategy, and further explore genomic prediction errors in relation to imputation errors in a U.S. Holstein population.

Materials and methods

Genotype and phenotype data

The data consisted of 6,988 Holstein animals (approximately 54% males and 46% females), each genotyped by the GeneSeek Genomic Profile (**GGP**) HD 80K (77,376) SNP chip: http://www.neogen.com/en/geneseek-announces-next-generation-of-dna-technology-geneseek-genomic-profilerbovine-hd . The phenotypes included predicted transmitting abilities (**PTA**s) for daughter pregnancy rate (**DPR**), fat yield (**FY**), and milk yield (**MY**). DPR was defined as the percentage of cows eligible to become pregnant in a 21-day period that actually become pregnant. Distributions of these traits showed that they were approximately normally distributed, yet skewed slightly toward large values (**Fig. 1**). Data cleaning steps of genotypes included the following. Firstly, unmapped SNPs and those on mitochondrial and Y

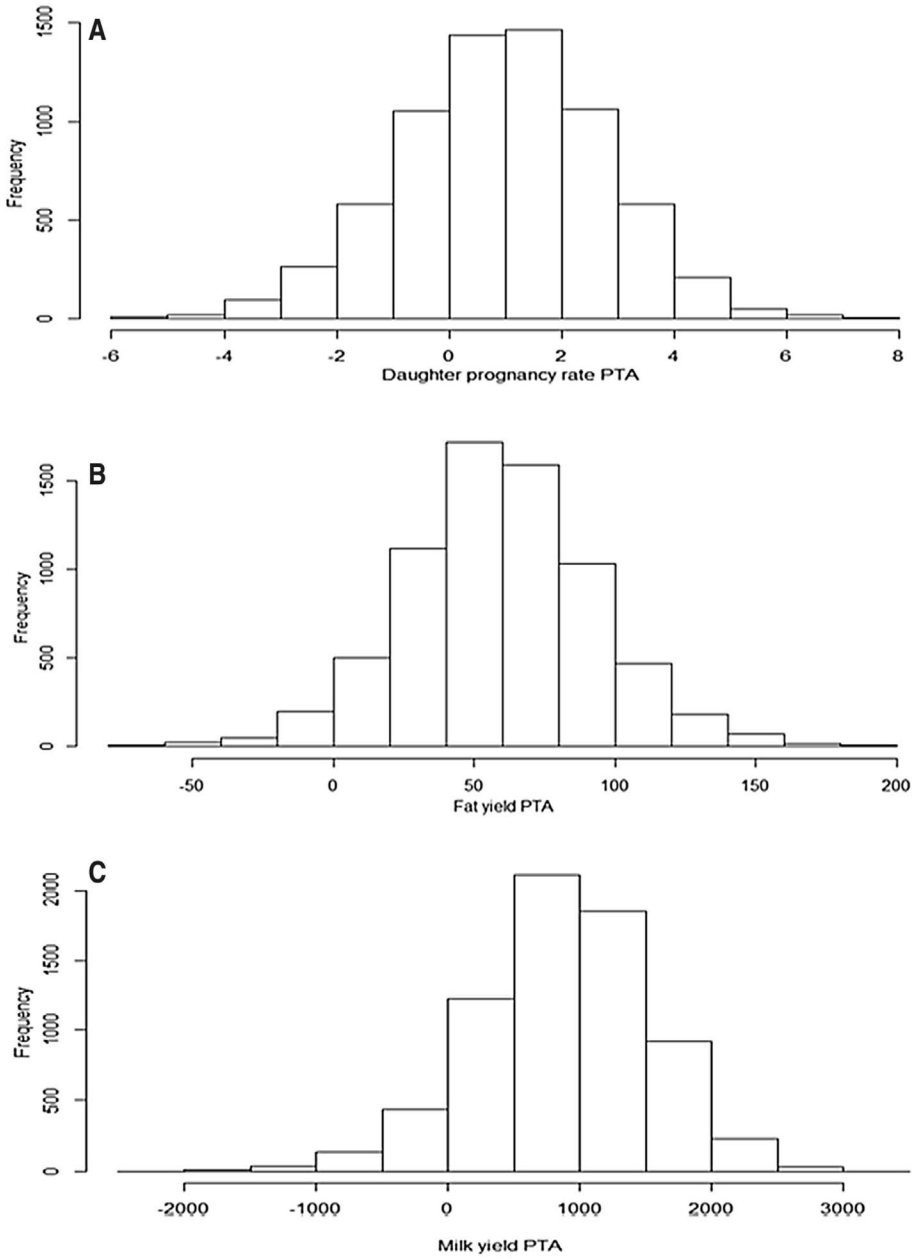

Fig. 1. Distributions of three phenotypes: **a)** daughter pregnancy rate, **b)** fat yield, and **c)** milk yield.

chromosomes were removed. Secondly, monomorphic SNPs and SNPs with MAF < 0.05%, and SNPs with > 10% missing genotypes were all removed. Finally, co-linearity among SNP genotypes was a concern when fitting a genomic model in which all SNPs were evaluated simultaneously. To reduce co-linearity between SNP loci, percentage of

genotype sharing was computed on a moving window of 20 neighboring SNPs on each chromosome. For SNPs with > 99% genotype sharing, only the one with the greatest MAF, and closest to the central location of each moving window if there were ties, were kept and all the remaining SNPs were deleted. These data editing and cleaning steps retained 68,748 SNPs for subsequent genomic prediction.

To mimic the scenario for forward genomic prediction, these animals were sorted by their dates of birth (**Table 1**), and SNP effects were estimated in 5593 older animals (born on and before 2014-08-18) as the training set and validated in the remaining 1395 younger animals (born after 2014-08-18). The sex ratios (males:females) were 55.6:44.4% and 44.3:55.7%, respectively, in the training and validation sets. For the validation animals, GEBV were computed based on the observed and imputed 80K genotypes, respectively, according to the estimated SNP effects from the training set.

Selection of LD SNPs

Eight LD SNP panels were formed using various strategies for selecting SNPs. These strategies attempted to optimize on each or a combination of three major factors, which are optimal map coverage (i.e.,

Table 1. Distribution (by years of birth) of the Holstein animals used in the present study

Year of birth	*Number of animals*
2000	1
2001	2
2003	2
2004	1
2005	2
2006	3
2007	5
2008	2
2009	7
2010	16
2011	58
2012	552
2013	2,647
2014	3,527
2015	163
SUM	6,988

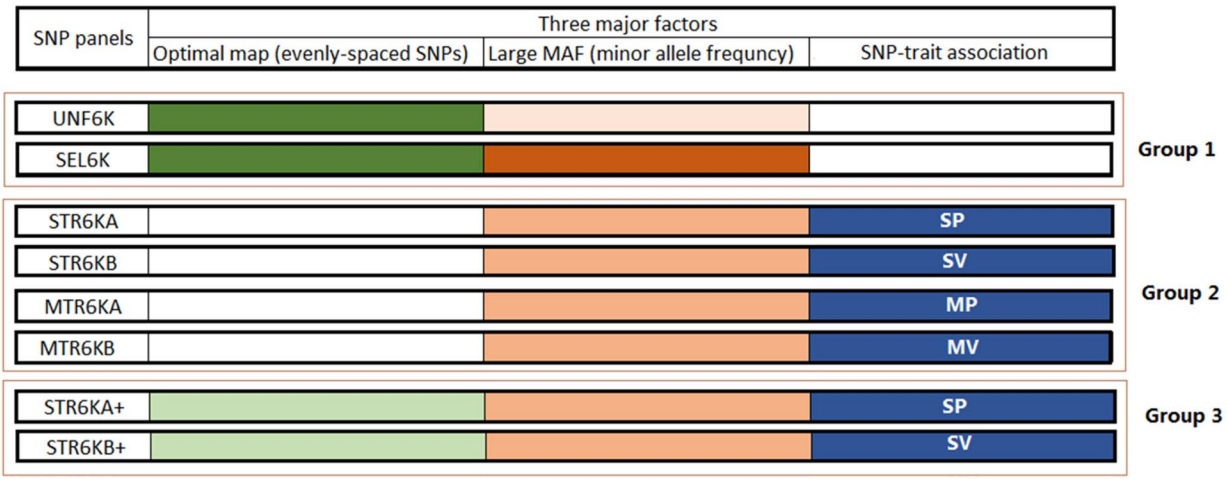

Fig. 2. Schematic diagram illustrating eight SNP panels in three groups with SNP selected by exploring various optimization criteria. *SP* SNPs with the largest posterior model probability of inclusion based on a single-trait BayesCπ model, *SV* SNPs with the largest SNP variance based on a single-trait BayesCπ model, *MP* SNPs with the largest posterior model probability of inclusion based on a multiple-trait model, *MV* SNPs with the largest weighted SNP variance based on a multiple-trait BayesCπ model. The intensity of color represents the optimization intensity on each of the three major factors.

evenly-spaced SNPs), large MAF, and significant SNP-traits associations (**Fig. 2**). Accordingly, the eight SNP panels can be classified into three groups. The first group consisted two panels, namely **UNF6K** and **SEL6K**, which are 6K SNP panels optimized for SNP map coverage, but SEL6K were also optimized for large MAF. The second group consisted of four SNP panels which were optimized to have SNPs with large SNP-trait associations, either for single traits (**STR6KA** and **STR6KB**) or for the three traits jointly (**MTR6KA** and **MTR6KB**). It turned out that, by selecting SNPs with large association effects, it led to having SNPs with large MAF as well. The third group included two enhanced panels of STR6KA and STR6KB, respectively, by including SNPs which are optimal selected for map coverage. The resulting two panels were denoted by **STR6KA**+ and **STR6KB**+, respectively.

Selection of SNPs for these eight panel are discussed in more detail as follows. UNF6K consisted of 6,000 approximately uniform-distributed SNPs. SEL6K had 6,000 SNPs optimally selected based on map coverage and MAF, and minimized for maximum gaps. These two 6K SNP panels were selected by the selectSNP package according to different optimization objectives (Wu et al. 2016). Single-trait

BayesCπ (Habier et al. 2011) was used to select trait-specific LD SNPs. The STR6KA panel was formed by pooling three sets of trait-specific LD SNP panels, each consisting of 2000 SNPs with the largest posterior model probability of inclusion (i.e., posterior probability for each SNP to have nonzero effects) for each trait. The STR6KB panel was formed by pooling three single-trait subsets, each consisting of 2000 SNPs with the largest SNP variance on each trait. Because there were common SNPs among the three trait-specific sets of 2000 SNPs, leaving space for a few hundreds of SNPs on each pooled panel, two enhanced LD panels (namely STR6KA+ and STR6KB+) were made by adding optimally selected SNPs to these two panels till the 6,000 slots of SNPs were filled. The multiple-trait BayesCπ (Jia and Jannink 2012) was used to selection LD SNPs of importance to the three traits jointly. There were two multiple-trait LD SNP panels: MTR6KA consisted of 6000 SNPs which were selected according to their posterior model probability of inclusion evaluated using a multiple-trait BayesCπ model and MTR6KB consisted of 6000 SNPs with the largest weighted SNP variances, with the weights being the averages of standardized SNP variances of each SNP on the three traits. Selection of SNPs using single-trait and multiple-trait BayesCπ were conducted using in-house software (Wu et al. 2012a, b).

Multiple-objective, local-optimization

The MOLO algorithm was used to optimally select SNPs for the SEL6K SNP panel. This algorithm centers on an objective function, $f(\boldsymbol{x})$, which maximizes the adjusted system information (Shannon entropy) and non-gap map length for a set of selected SNPs under multiple constraints (e.g., on MAFs, location distribution of SNPs, inclusion of obligatory SNPs, and number and size of gaps). That is,

$$max\ \{ f(\boldsymbol{x})\} \mid g(\boldsymbol{x}),\ \ h(\boldsymbol{x}),\ \ i(\boldsymbol{x}|\boldsymbol{o}),\ r\ \} \tag{1}$$

where $g(\boldsymbol{x})$ collectively includes all equality constraints, $h(\boldsymbol{x})$ includes all inequality constraints, $i(\boldsymbol{x}|\boldsymbol{o})$ represents constraints given the set of obligatory SNPs, and $0 \leq r \leq 1$ is a tunable parameter for the bin width that is used in the heuristic search for local optima.

Briefly, the putative distributions of SNPs were initialized uniformly. Gaps were minimized given the number of SNPs on each chromosome. The SNP quality and fidelity criteria, such as call rate and

Mendelian inconsistency, were resolved prior to optimization and hence were not included in the MOLO algorithm. Information for a chip were computed based on multi-loci frequencies of all involving SNPs, adjusted by the uniformness of SNP distribution on each chromosome. The objective function in Eq. (1) was highly non-linear and a heuristic search algorithm was used to find local optima in an attempt to approximate the global optimum.

Single-trait BayesCπ

For each trait, the phenotype data were described by the following linear model:

$$y_i = \mu + \sum_{j=1}^{k} x_{ij} b_j + e_i \quad (2)$$

where y_i was a PTA for the i-th individual, μ is the overall mean, x_{ij} was the genotype (which were coded as -1, 0, 1, respectively) of the j-th SNP measured on the i-th individual, b_j was the additive association effect of the j-th SNP, k is the number of SNPs, and $e_i \sim N(0,\sigma_e^2)$ was a residual term.

The BayesCπ model (Habier et al. 2011) assumed *a priori* that each SNP effect was null with probability π, or it followed a normal distribution, $N(0,\sigma_b^2)$, with probability $1 - \pi$.

$$b_j|\pi, \sigma_b^2 \sim \begin{cases} N(0,\sigma_b^2), & \text{with probability } (1-\pi) \\ 0 & \text{with probability } \pi \end{cases} \quad (3)$$

In the above, σ_b^2 was a variance common to all non-zero SNP effects, which in turn was assigned a scaled inverse Chi square prior distribution, $\chi^{-2}(v_b, s_b^2)$. Similarly, the prior distribution for σ_e^2 was also taken to be a scaled inverse Chi-square distribution, $\chi^{-2}(v_e, s_e^2)$. Furthermore, the value of π in the model was unknown and was inferred with the prior distribution of π taken to be uniform between 0 and 1.

The BayesCπ model was implemented via Markov chain Monte Carlo (**MCMC**) with three parallel chains each consisting of 50,000 iterations after a burn-in of 5,000 iterations, thinned at every one-tenth. The posterior inference on each unknown parameter was made on the pool of saved posterior samples from the three parallel chains after the burn-in period.

Multiple-trait BayesCπ

LD SNPs with the greatest impact on all the three traits were selected using multiple-trait BayesCπ model (Jia and Jannink 2012). Based on the following multiple-trait version of model (2):

$$\boldsymbol{Y} = 1\,\mu' + \sum_{j=1}^{k} \boldsymbol{x}_j\,\boldsymbol{b}'_j \tag{4}$$

where **Y** was a $n \times m$ matrix for m traits measured on n individuals, **1** was a $n \times 1$ vector of 1's, ($\mu = \mu_1\ \mu_2 \ldots \mu_m$) was a $m \times 1$ vector of overall means for the m traits, $\boldsymbol{x}_j = (x_{1,j} \ldots x_{i,j} \ldots x_{n,j})'$ was a $n \times 1$ vector of genotypes for the j-th SNP, $\boldsymbol{b}_j = (b_{j1}\ b_{j2} \ldots b_{jm})$ was a $m \times 1$ vector of genetic effects of marker j on the m traits, and $\boldsymbol{E} = (\boldsymbol{e}_1\ \boldsymbol{e}_2 \ldots \boldsymbol{e}_n)'$ was a $n \times m$ residual matrix.

The multiple-trait BayesCπ model was computed via MCMC. Three parallel MCMC chains were run each with 50,000 iterations after a burn-in of 5000 iterations, thinned at every one-fifth. The saved posterior samples were pooled after the burn-in period and then used to make inference on unknown model parameters.

Weighted SNP variances were computed as follows. Consider the j-th SNP selected for the t-th trait, for $j = 1, 2, \ldots, k$ and $t = 1, 2, 3$. Then, the standardized variance of association effects of this SNP on the t-th trait was computed to be:

$$\hat{\sigma}^2_{j(t)} = \frac{2p_j q_j \hat{b}^2_{j(t)}}{\sum_{j=1}^{k} 2p_j q_j \hat{b}^2_{j(t)}} \tag{5}$$

where p_j and q_j were the observed frequencies of the two alleles for the j-th SNP and $\hat{b}^2_{j(t)}$ is an estimate of the corresponding additive association effects, both pertaining to the t-th trait in the training population. Then, standardized SNP variances were averaged across the three traits for each SNP, as follows:

$$\overline{\sigma^2_j} = \tfrac{1}{3} \sum_{t=1}^{3} \hat{\sigma}^2_{j(t)} \tag{6}$$

Estimation of SNP effects using ridge-regression BLUP for genomic prediction

SNP effects were estimated for each trait independently using the following linear model:

$$\mathbf{y} = 1\mu + \boldsymbol{Xb} + \boldsymbol{e} \tag{7}$$

where $\boldsymbol{y}$ was a $n \times 1$ vector of PTA for all the animals in the training population, $\boldsymbol{X}$ was an $n \times p$ matrix of SNP genotypes, $\boldsymbol{b}$ was a $p \times 1$ vector of unknown allelic substitution effects of all the SNPs, and $\boldsymbol{e}$ was the residual term.

The ridge regression estimator solved the above linear regression using ℓ_2 penalized least squares:

$$\hat{\beta}(\text{ridge}) = \arg\min_{\beta} \| \boldsymbol{y} - \mathbf{1}\mu - \boldsymbol{Xb}\|^2 + \lambda\|\boldsymbol{b}\| \tag{8}$$

In the above, $\| \boldsymbol{y} - \mathbf{1}\mu - \boldsymbol{Xb}\|^2$ was the ℓ_2–norm (quadratic) loss function (i.e., residual sum of squares), $\|\boldsymbol{b}\|^2 = \sum_{j=1}^{k} b_j^2$ was the ℓ_2–norm penalty on $\boldsymbol{b}$, and $\lambda \geq 0$ is the tuning parameter, which regulated the strength of the penalty (linear shrinkage). *A priori*, we set $\lambda = \hat{\sigma}_e^2/\hat{\sigma}_b^2$, where $\hat{\sigma}_e^2$ was the estimated residual variance, and $\hat{\sigma}_b^2$ was the estimated variance of regression coefficients given by $Var(\boldsymbol{b}) = \boldsymbol{I}\sigma_b^2$. Let $\hat{\sigma}_a^2$ and $\hat{\sigma}_e^2$ be the estimated additive genetic variance and the estimated residual variance, respectively, from an equivalent animal model. The initial values for σ^2_b was set up to be

$$\hat{\sigma}_b^2 = \frac{\hat{\sigma}_a^2}{(k \times \overline{2pq})}$$

where $\overline{2pq} = k^{-1} \sum_{i=1}^{k} (2p_j q_j)$, and p_j and q_j are the observed frequencies of the two alleles at SNP j. A Bayesian version of the above ridge regression model was implemented via MCMC, which allowed for sampling the common SNP variance and the residual variance, in addition to the overall mean and SNP effects. Estimation of SNP effects were conducted using in-house genomic prediction pipelines (Wu et al. 2012a, b).

Preliminary determination of optimal trait-specific LD SNP panel size

Prior to the design of LD SNP chips, an optimal trait-specific LD SNP panel size was determined as such that the loss in genomic prediction accuracy (LGPA) using a subset of selected SNPs were at most 3% as compared to genomic prediction using the observed 80K SNPs. Briefly, eight subsets of SNPs were evaluated, each consisting of top 500, 1000, 1500, 2000, 2500, 3000, 3500, and 4000 SNPs, respectively, sorted in the descending order by the posterior probability of inclusion of a SNP as having non-zero effect on each trait in the single-trait BayesCπ model (Habier et al. 2011). Then, GPA was evaluated by three-fold cross-validation (Kohavi 1995). **LGPA** was measured by percent decrease of GPA using a subset of the 80K SNP genotypes compared to that using the whole set of observed 80K SNP genotypes. LGPA were roughly between 1 and 11% with between 500 and 4000 SNPs selected. The more SNPs were selected for genomic prediction, the less LGPA. Overall, LGPA was approximately ≤ 3% for each of the three traits with 2000 selected SNPs fitted in the genomic prediction model, and it began to plateau when more selected SNPs were fitted in the genomic prediction model (**Fig. 3**). Hence, this number (i.e., 2000 SNPs) was taken to be the optimal number of SNP for each trait to be included on the panels to guide the SNP selection in the following sections. Note that the optimal LD SNP panel size, as determined this way, is only empirical and it can vary with the actual data.

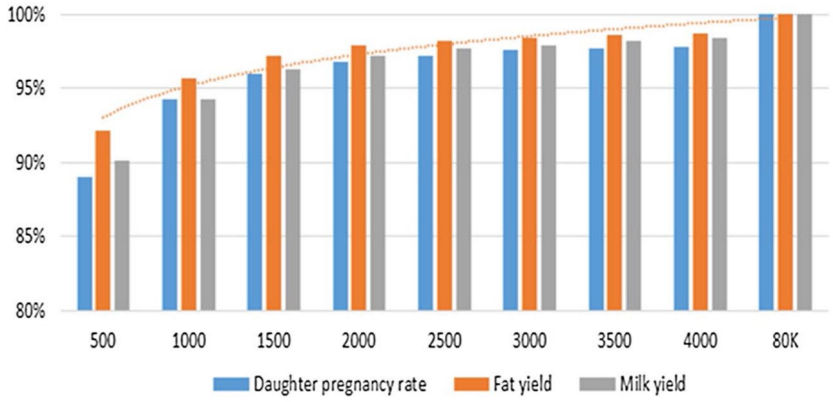

Fig. 3. Relative genomic prediction accuracies using subsets (i.e., from 500 to 4000) of selected SNPs with the largest association effects on each of the three traits over those using the whole 80K SNPs (i.e., 68,748 SNPs with MAF > 0.05).

Measurements of imputation accuracy and genomic prediction accuracy

Imputation accuracy rate was computed as the percentage of correctly imputed cases of genotypes for all SNPs that had been imputed, as compared to observed genotypes. Conversely, imputation error rate is the percentage of incorrectly imputed genotype cases for all SNPs. Calus et al. (2014) noted that imputation error rate (and hence imputation accuracy rate) depends on MAF. They also argued that a more appropriate measurement of imputation accuracy should be computed as the correlation between true and imputed genotypes, because the latter does not depend on MAF and therefore can be compared across loci with different MAF (Calus et al. 2014). In the present study, imputation error rates were compared among panels but not cross loci. Though MAF varied drastically with SNPs and with these eight LD SNP panels, there were very slight differences among the remaining sets of (~63K) SNPs to be imputed. Thus, we decided to use imputation accuracy rate.

SNP effects were estimated on the observed 80K SNP genotypes using ridge-regression BLUP for each of the three traits independently in the training population (5393 animals). In the validation set (1,395 animals), 80K SNP genotypes were imputed based on each set of 6K LD SNP genotypes using the FImpute package (Sargolzaei et al. 2014). Then, GEBV was computed for each validation animal with the observed and imputed 80K SNP genotypes, respectively, as the predictor variables according to SNP effects estimated on the observed 80K genotypes in the training set. GPA obtained using observed or imputed 80K genotypes were measured by the correlation between PTAs and GEBVs of animals in the validation set. Relative genomic prediction accuracy (**RGPA**) was also computed as a percentage of GPA using imputed 80K SNP genotypes over that obtained using observed 80K SNP genotypes in the validation set.

Results

Design of LD SNP panels

In the eight LD SNP panels, UNF6K consisted of 6000 approximately uniform-distributed SNPs. SEL6K had 6000 SNPs optimally selected based on map coverage and SNP information (i.e., MAF), and minimized for maximum gaps. After removing duplicated SNPs among the three trait-specific sets of 2000 SNPs, the STR6KA panel had 5373 unique SNPs and the STR6KB panel had 5218 unique SNPs. The STR6KA+ panel included all the unique SNPs in the STR6KA panel, plus 627 SNPs optimally selected by the selectSNP package (Wu et al. 2016), and the STR6KB+ panel included all the SNPs in the STR6KB panel plus 782 SNPs optimally selected by the selectSNP package (Wu et al. 2016). For convenience of discussion, UNF6K and SEL6K are also referred to as map-optimal panels because they were optimized for SNP distributions on the maps, and STR6KA+ and STR6KB+ are referred to as enhanced panels because they contained both trait-specific SNPs and map-optimal SNPs. There were two multiple-trait LD SNP panels: MTR6KA and MTR6KB, each consisting of 6000 SNPs selected by a multiple-trait BayesCπ model.

Average MAF was 0.45 for the SEL6K panel and 0.30 for the UNF6K panel, and it was 0.30 for the 80K SNPs (i.e., 68,748 SNPs with MAF > 0.05). Thus optimal selection of SNPs for both map coverage and MAF considerably elevated MAF (**Fig. 4a** vs. **c**), but selection of evenly-spaced SNPs did not change MAF relative to the 80K genotypes (Fig. 4a vs. b). Selection of 6K SNPs according to their association effects did not directly contemplate MAF but it elevated MAF indirectly (**Fig. 4a** vs. **d**). The means (standard deviations) of MAF for STR6KB and MTR6KB panels were 0.36 (0.12) and 0.34 (0.11), respectively, in this Holstein population. This was possibly because of the fact that SNPs with small MAF also had larger variances associated with their estimated effects, and were therefore more difficult to pass certain cutoffs imposed in the test of association effects than those with larger MAF.

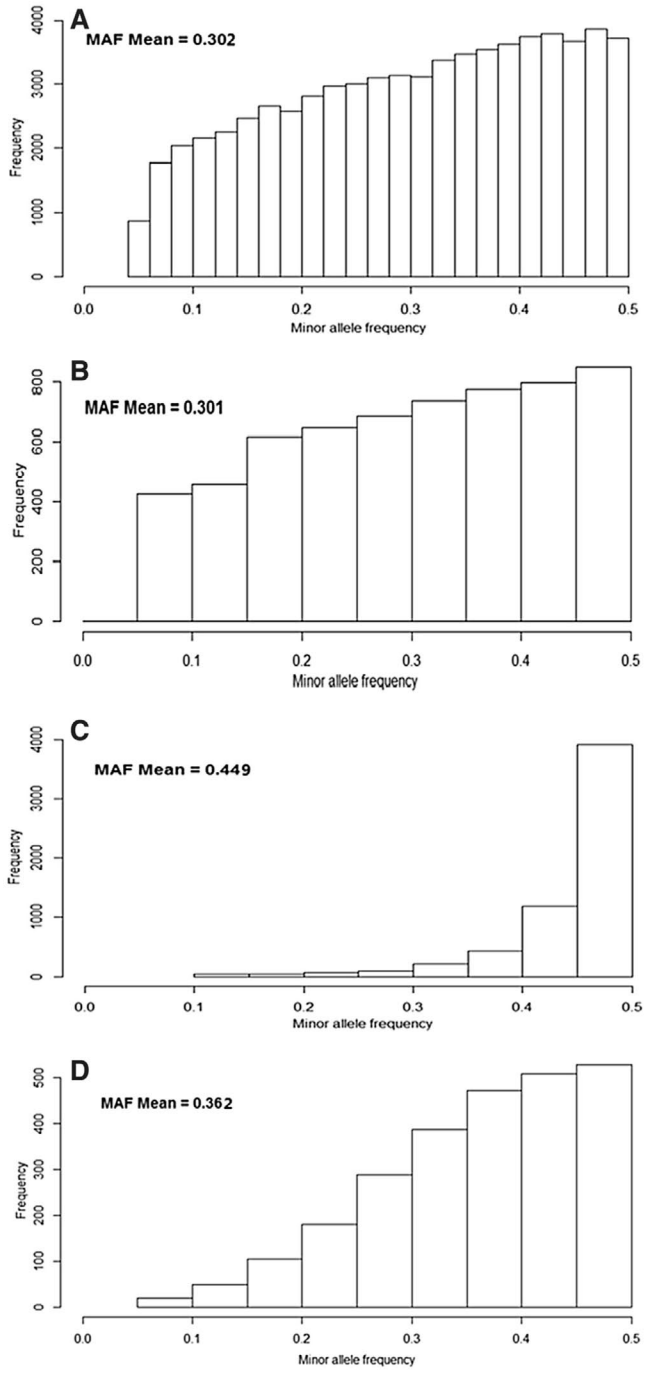

Fig. 4. Distributions of MAF computed for: **a)** 68,748 SNPs with MAF > 0.05, **b)** 6000 evenly-spaced SNPs with MAF > 0.05 (UNF6K), **c** 6,000 SNPs optimally-selected by the MOLO algorithm (SEL6K), and **d** 5218 unique SNPs selected the largest SNP effect variances on each of the three traits (STR6KB).

Imputation accuracy from 6K to 80K SNP genotypes

The average imputation accuracy rates were between 96.2% and 98.2% (**Table 2**). The SEL6K had the greatest imputation accuracy (98.2%), followed by UNF6K - the 6K panel of evenly-spaced LD SNPs (97.6%). For the two enhanced panels, STR6KA+ and STR6KB+, their imputation accuracy rates (97.4–97.5%) were only slightly lower than those of the two map-optimal panels (SEL6K and UNF6K).

Table 2. Summary statistics and imputation accuracy rate (SD) of the eight low-density 6K SNP panels

Panel	*Number of SNPs*	*Map optimal*	*MAF optimal*	*SNP-trait association*	*Imputation accuracy, %*
UNF6K	6000	Yes	No	No	97.6 (0.45)
SEL6K	6000	Yes	Yes	No	98.2 (0.27)
STR6KA	5373	No	Correlated	Yes	96.2 (1.30)
STR6KA+	6000	Yes	Yes	Yes	97.4 (0.94)
STR6KB	5218	No	Correlated	Yes	96.4 (1.36)
STR6KB+	6000	Yes	Yes	Yes	97.5 (0.90)
MTR6KA	6000	No	Correlated	Yes	96.4 (1.86)
MTR6KB	6000	No	Corrected	Yes	96.4 (1.88)

UNF6K = 6,000 evenly-spaced SNPs
SEL6K = 6,000 SNPs optimally-selected by the selectSNP package (Wu et al. 2016)
STR6KA = 5,373 unique SNPs pooled from three sets of trait-specific SNP panels, each consisting of 2,000 SNPs with the largest model probability of having non-zero association effects on each trait (i.e., selected by single-trait BayesCπ)
STR6KA+ = STR6KA plus 627 SNPs optimally-selected by the selectSNP package (Wu et al. 2016)
STR6KB = 5,218 unique SNPs pooled from three sets of trait-specific SNP panels, each consisting of 2,000 SNPs with the largest variance of SNP association effects on each trait (i.e., selected by single-trait BayesCπ)
STR6KB+ = STR6KA plus 782 SNPs optimally selected by the selectSNP package (Wu et al. 2016)
MTR6KA = 6,000 SNPs with the largest model probability of having non-zero association effects on the three traits (i.e., selected by multiple-trait BayesCπ)
MTR6KB = 6,000 SNPs with the largest weighted SNP variances on the three traits (i.e., selected by multiple-trait BayesCπ
SD = standard deviation of imputation accuracy rates by chromosomes

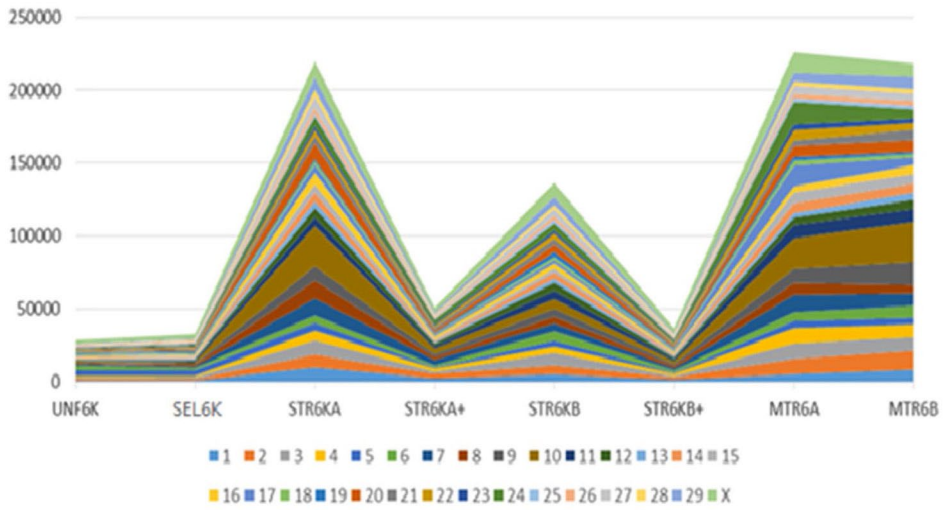

Fig. 5. Sum of maximum gap length, in 1000 base pairs, on the 29 autosomes and X chromosome, computed for each of the eight 6K low-density SNP panels.

Imputation accuracy rates were negatively associated with maximum map gaps. To illustrate this situation, the maximum gap length on each of 30 chromosomes (29 autosomes and X chromosome), denoted by the total maximum gap length (**TMGL**), were computed and summed up for each of the eight 6K LD SNP panels. As shown in **Fig. 5**, TMGL was the smallest for UNF6K and SEL6K, and the largest for STR6KA, STR6KB, MTR6KA and MTR6KB (Fig. 5). Nevertheless, the two enhanced 6K panels (STR6KA+ and STR6KB+) had considerably decreased TMGL, which were only slightly larger than those for the two map-optimal 6K SNP panels (UNF6K and SEL6K). Relative to TMGL for the UNF6K panel (which was set to be 100%), SEL6K had a relative TMGL (**RTMGL**) of 114.38%, and STR6KA+ and STR6KB+ had a RTMGL of 156.87 and 129.36%, respectively. These four LD SNP panels had comparable TMGL and their imputation accuracy rates were also comparable. However, the remaining four LD SNP panels (STR6KA, STR6KB, MTR6KA, and MTR6KB) had TGML which were approximately 4 to 7 times (434.14–759.76%) larger than UNF6K, and their imputation accuracy rates were the lowest. Thus, our results support adding a set of optimally-selected SNPs to association LD SNP panels (as in the cases of STR6KA+ and STR6KB+) in order to decrease map gaps and increase imputation accuracies.

Genomic prediction accuracy using observed vs. imputed 80K genotypes

Genomic prediction accuracies using observed 80K SNP genotypes were 0.825 for DPR, 0.847 for FY, and 0.868 for MY (**Table 3**). Genomic prediction accuracies using imputed 80K SNP genotypes were slightly lower than those using the observed 80K SNP genotypes, which were 0.817–0.821 for DPR, 0.838–0.844 for FY, and 0.850–0.863 for MY (Table 3). Relative genomic prediction accuracy (**RGPA**), which was defined as a percentage of genomic prediction accuracy (**GPA**) using imputed 80K SNPs over that using observed 80K SNPs, were 97.9–99.6% for all the eight 6K LD SNP panels, and 99.3–99.6% for the two map-enhanced panels (STR6KA+ and STR6KB+). Our results showed

Table 3. Genomic prediction accuracy using imputed 80K and observed 80K SNP genotypes, respectively

SNP panels	*DPR (%)*		*FY (%)*		*MY (%)*	
	GPA	*RGPA (%)*	*GPA*	*RGPA (%)*	*GPA*	*RGPA*
UNF6K->80K	0.817	99.0	0.838	98.9	0.850	97.9
SEL6K->80K	0.819	99.3	0.841	99.3	0.851	98.0
STR6KA->80K	0.819	99.3	0.842	99.4	0.855	98.5
STR6KA+->80K	0.821	99.5	0.844	99.6	0.862	99.3
STR6KB->80K	0.818	99.2	0.842	99.4	0.856	98.6
STR6KB+->80K	0.821	99.5	0.844	99.6	0.863	99.4
MTR6KA->80K	0.820	99.4	0.843	99.5	0.858	98.8
MTR6KB->80K	0.820	99.4	0.844	99.6	0.858	98.8
Observed 80K	0.825	100	0.847	100	0.868	100

GPA stands for genomic prediction accuracy, which was computed to be the correlation between PTA and genomic estimated PTA, and RGPA stands for relative genomic prediction accuracy, which was the percentage of GPA using imputed 80K genotypes over than using the observed 80K genotypes, both evaluated in the validation population (i.e., 2,639 U.S. Holstein animals)

See Table 2 for acronyms of the eight LD panels (UNF6K, SEL6K, STR6KA, STR6KA+, STR6KB, STR6KB+, MTR6KA, MTR6KB)

X->80K = 80K SNP genotypes imputed from the LD X SNP panel, where *X* stands for UNF6K, SEL6K, STR6KA, STR6KA+, STR6B, STR6KB+, MTR6KA, and MTR6B, respectively.

Observed 80K = observed 80K genotypes

that GPA using imputed HD-SNP genotypes were highly comparable to that using observed HD SNP genotypes, in particular when the LD SNP panel was constructed with optimized SNP coverage, MAF and SNP-trait associations.

Genomic prediction accuracies using imputed 80K genotypes did not show a parallel relationship with the corresponding imputation accuracies. For example, the two map-optimal panels, SEL6K and UNF6K, had the greatest imputation accuracy (97.6–98.2%) but their corresponding genomic prediction accuracies (0.817–0.851) were among the lowest. On the other hand, the two enhanced LD SNP panels (STR6KA+ and STR6KB+) had the highest genomic prediction accuracies (0.821–0.863), though their corresponding imputation accuracies (97.4–97.5%) were slightly lower than the two map-optimal panels (SEL6K and UNF6K). To probe into this situation, the results from three sets of imputed 80K SNP genotypes (derived from UNF6K, STR6KB and STR6KB+, respectively) were examined further. For each of the three LD panels, imputed 80K (68,748) SNPs were divided into two subsets: one subset consisting of 6000 SNPs with the largest SNP variance for each trait (Top6K) and the other subset including all the remaining 62,748 SNPs (R63K). In other words, all the 68,748 SNPs were assigned to two groups, one with SNPs having decisive impacts on genomic prediction and the other with SNPs whose impacts on genomic prediction were trivial. Then, genotype (imputation) error rate for each of the two subsets of SNPs (and their ratio) was computed (**Table 4**). Note that imputation error rates were computing by including all the SNPs, either reference SNPs or imputed SNPs, in this part of the search, which was collectively referred as genotype error rate (**GER**) hereafter. Our purpose was to compare how many SNPs had wrong genotypes, compared to the corresponding observed genotypes. For the uniform panel UNF6K, GER were comparable between these two groups, though slightly higher for SNPs in the Top6K group. Because SNPs on the UNF6K panel were map-optimally selected without considering SNP-trait associations, genotype (imputation) error rates were expected to be comparable between these two groups. The observed slight differences could be intrinsic or resulted from random sampling bias. For the two panels featuring SNP-trait associations (STR6KB and STR6KB+), GER for the 6,000 "influential" SNPs in the Top6K group was only 50.0–69.3% as much as that for SNPs in the R63K group. This coincided with the fact that a majority of these

Table 4. Comparing genotype (imputation) error rates for top 6000 (Top6K) SNPs with the largest SNP variance on each trait and those for the remaining 62,748 (R63K) SNPs

Traits	*SNP panel*	*Imputation error,%*			*Top6K/R63K*
		All	*Top6K*	*R63K*	*(%)*
DPR	UNF6K	1.99	2.5	1.96	127.6
	STR6KB	3.49	1.88	3.65	51.5
	STR6KB+	2.35	1.67	2.41	69.3
FY	UNF6K	1.99	2.32	1.94	119.6
	STR6KB	3.49	1.85	3.65	50.7
	STR6KB+	2.35	1.5	2.43	61.7
MY	UNF6K	1.99	1.89	1.66	113.9
	STR6KB	3.49	1.82	3.64	50.0
	STR6KB+	2.35	1.47	2.43	60.5

See Table 2 for acronyms of the three LD panels (UNF6K, STR6KB, STR6KB+)
Top6K/R63K (%) = Ratio of genotype (imputation) error rate for the SNPs in Top6K over that for the SNPs in R63K

6000 SNPs were included in the selected 6K LD SNPs and their genotypes were known (i.e., not imputed). Thus, UNF6K had a lower GER in general but not necessarily lower GER for SNPs of importance to genomic prediction. In contrast, selection of LD SNPs based on SNP-trait associations did not increase the overall imputation accuracy *per se*, but, by including most-influential SNPs into the reference SNPs for imputation, their genotypes were known (instead of being imputed) and the negative impact of imputation errors for this set of trait-associated SNPs on genomic prediction was minimized.

Furthermore, the association effect variance of each SNP on DPR was plotted against its imputation error for two SNP panels, UNF6K and STR6KB. For STR6KB, SNPs with large association effects mostly had zero imputation errors (**Fig. 6b**) whereas for the map-optimal LD SNP panel (UNF6K), very few SNPs with large association effect variances had non-zero GER (**Fig. 6a**). These results confirmed our assumption that SNPs selected according to SNP-trait association had smaller GER. For the two enhanced panels (STR6KA+ and STR6KB+), each had a number of map-optimal SNPs, in addition to SNPs with significant associations with the quantitative traits. Therefore, their GER were minimized for both "trait-influential" SNPs and for all SNPs

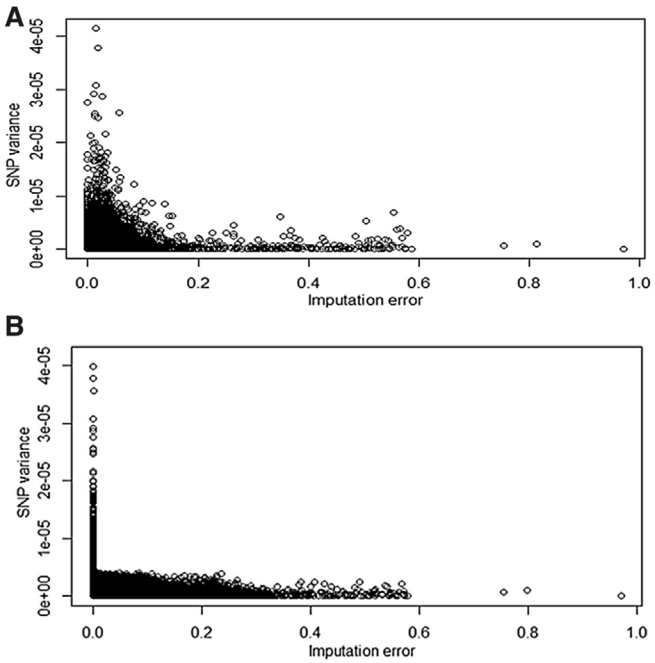

Fig. 6. Illustration of the relationship between single SNP variance contribution rate (%) for daughter pregnancy rate (DPR) and genotype (imputation) error rate (%) per SNP basis for two 6K LD SNP panels: **a)** UNF6K, and **b)** STR6KB.

in general, and these two enhanced panels had the greatest genomic prediction accuracies among the eight LD SNP panels.

Impact of imputation error rates on genomic prediction accuracy

Finally, analysis of variance (ANOVA) was conducted to determine the impact of GER on genomic prediction using imputed 80K SNP genotypes. Briefly, GER were split into two variables, one pertaining to top 6,000 SNPs with largest (weighted) association effect variance on the three traits (GER1) and the other attributable to the remaining 62,748 SNPs (GER2). Two ANOVA models were evaluated: Model 1 include traits and GER as the explanatory variables (treatments), in addition to the residuals; Model II included traits, GER1 and GER2, in addition to the residuals. In both models, RGPA was the dependent variable. The ANOVA results from model I showed that RGPA using imputed 80K genotypes was significantly different among the three traits ($P = 3.55e{-}06$) but it was not significantly affected by the overall GER for all LD SNPs panels ($P = 0.38$). The ANOVA results from model II

showed that GPA was significant among traits (P = 1.98e−06) and it was significantly affected by GER1 (P = 0.0308), but not significantly affected by GER2 (P = 0.5559). Therefore, it is concluded that the accuracy of imputation-mediated genomic prediction critically depends on genotype (imputation) accuracies of a set of SNPs with large impacts on genomic prediction.

Discussion

Imputation accuracy

Generally speaking, the two SNP panels with optimal SNP map coverage (group 1 SNP panels; Fig. 2) had greater imputation accuracies than the remaining six SNP panels (group 2 and 3 SNP panels; Fig. 2). Of these two map-optimal SNP panels, SEL6K had greater imputation accuracy rate than UNF6K. This possibly reflected the fact that SEL6K had highest MAF on average (0.449) than UNF6K (0.301). Similarly, Boichard et al. (2012) designed a LD array of 6,909 SNPs optimal both in map distributions and MAF, yet using a different optimization method and they obtained an average imputation accuracy of 98.9% in North American Holstein cattle. Their imputation accuracy rate was slightly higher than that of the SEL6K panel, because their LD SNP array had 909 more SNPs and they imputed to approximately 10,000 less SNPs than in the current study. It is important to note that imputation accuracy is decided by many factors including the relationships between the reference and the target imputation set, and the number of animals in the reference population, both of which varied between these two studies.

Imputation accuracy rates were negatively associated with map gaps. Evidently, UNF6K and SEL6K (group 1 SNP panels; Fig. 2) had the smallest gaps and therefore the greatest imputation accuracies. On the other hand, SNPs selected based on their association effects (group 2 SNP panels; Fig. 2) tended to be extremely unevenly-distributed, leaving large gaps on the genome. This also reflected the fact that causative variants for each of the traits were not evenly distributed. Thus, trait-association LD SNP panels tend to have lower imputation accuracies, as compared to map-optimal SNP panels, assuming that everything else is the same. Nevertheless, by including

map-optimal, informative SNPs to trait-specific SNP panels, these large gaps were filled, which in turn led to improved SNP coverage on the genome and therefore greater imputation accuracy rates. Thus, tor these two enhanced panels in group 3 (STR6KA+ and STR6KB+), their imputation accuracy rates were only slightly lower than those of the two map-optimal panels (SEL6K and UNF6K).

Minor allele frequency of a SNP was another factor affecting imputation accuracy. Here, we distinguish SNPs by their roles in the imputation: SNPs with missing genotypes to be imputed and SNPs with known genotypes as the reference for imputation. For a SNP with missing genotypes to be imputed, larger MAF indicates greater uncertainty in the determination of its genotypes and hence may be associated with large imputation error rate. Consider a frequency-based imputation approach and assume a complete linkage between the SNP with missing genotypes and the SNP with known genotypes as the reference, Calus et al. (2014) showed, both analytically and with empirically, that imputation error rates depended on MAF. However, in this part of discussion, we relax the assumption of complete linkage. A reference SNP (i.e., one with known genotypes to be used for inferring missing SNP genotypes) can be any SNP which is informative of the missing genotypes. This also included the situation in which population-wise linkage disequilibrium contributed to imputation (e.g., Sargolzaei et al. 2014). Thus, SNPs with greater MAF are more informative in the determination of the phases of a missing SNP genotype than those with lower MAF. Possibly, this could explain the situation with the SEL6K panel, which was optimized on MAF in addition to map positions. SEL6K outperformed the UNF6K panel in terms of imputation accuracy, because the former had more SNPs with high MAF than the latter.

Genomic prediction accuracy

Genomic prediction accuracies obtained using imputed 80K SNP genotypes were highly comparable to those obtained using observed 80K SNP genotypes, in particular for group 3 SNP panels (STR6KA+ and STR6KB+), which were was optimally constructed for SNP coverage, MAF and SNP-trait associations. Compared to previous studies, our genomic prediction accuracies were higher than those reported by VanRaden et al. (2009), who obtained genomic prediction accuracies

of 0.54 for DPR, 0.66 for FY and 0.70 for MY in a U.S. Holstein population. This was probably because they used fewer SNPs (i.e., 38,416 SNPs) and fewer calibration animals (3,576 bulls) for genomic prediction. Cooper et al. (2015) reported genomic prediction accuracies of 0.76 for DPR, 0.87 for FY and MY with a calibration set of 6,623 U.S. Holstein bulls, which were comparable to ours in both calibration/training population size and genomic prediction accuracies. Nevertheless, they had higher genomic prediction accuracies (i.e., 0.84 for DPR and 0.90 for FY and MY) than ours with a calibration set of 17,407 bulls. Genomic prediction accuracies in the present study were lower than those reported by Wu et al.(2016), because the PTAs for the three traits used by Wu et al.(2016) included genomic information whereas PTAs in the current study did not.

GPA obtained using imputed SNP genotypes were subject to imputation errors, but they did not show a parallel relationship in the present study. For example, the two map-optimal panels (group 1 SNP panels: SEL6K and UNF6K) had the greatest imputation accuracy but their corresponding genomic prediction accuracies were not the best. On the other hand, the two enhanced LD SNP panels (group 3 SNP panels: STR6KA+ and STR6KB+) had the best genomic prediction accuracies (0.821–0.863), though their corresponding imputation accuracies were slightly lower than the two map-optimal panels (SEL6K and UNF6K). Our results indicated that SNPs varied relative to their impacts on genomic prediction, and imputation errors that were projected through these SNPs onto genomic prediction errors could vary as well. Thus, by including "influential" SNPs in the LD SNP panels, genotype (imputation) errors relative to the set of "influential" SNPs could be reduced dramatically and therefore the corresponding imputed 80K SNPs could be highly predictive. In other words, selection of LD SNPs based on SNP-trait associations did not necessarily increase the overall imputation accuracy *per se*, but, by including most-influential SNPs into the reference SNP list, their genotypes were known (instead of being imputed) and the negative impact of imputation errors on genomic prediction was minimized. This assumption was affirmed by the ANOVA results, which showed that the accuracy of imputation-mediated genomic prediction critically depended on genotype (imputation) accuracies of a set of SNPs with large impact on genomic prediction. This is an interesting finding which has important implications to the design of LD panels for imputation-mediated genomic prediction.

In group 2 SNP panels, the two pooled, single-trait 6K LD SNP panels (STR6KA and STR6KB) performed slightly worse than the two multiple-trait 6K SNP panels, possibly because these two former LD SNP panels had a few hundred less SNPs than the two multiple-trait panels. Nevertheless, the two enhanced single-trait 6K LD SNP panels (STR6KA+ and STR6KB+), with the inclusion of map-optimal and informative SNPs, had better imputation accuracy and better GPA than the two multiple-trait 6K LD SNP panels. Our results, however, should not be used to suggest that the single-trait approach was worse or better than the multiple-trait approach to select LD SNPs, because the results from these two sets were not directly comparable. Nevertheless, these results justified the need to include SNPs associated with traits to be selected in LD SNP chips for imputation-mediated genomic prediction.

Conclusions

Genomic prediction using 80K genotypes imputed from 6K LD SNPs had accuracies which were comparable to (or slightly lower than) those using observed 80K SNPs in the Holstein population. The eight 6K LD SNP panels showed some differences in their imputation accuracies and prediction accuracies. Generally speaking, evenly-spaced, informative (e.g., large MAF) SNPs (group 1 SNP panels) were favorable for obtaining accurate imputation because they had a better coverage of genome than trait-associated SNPs. On the other hand, SNPs selected based on their association effects (group 2 SNP panels) were favorable for obtaining increased GPA because a majority of SNPs of importance to genomic prediction were included the LD panel and their genotypes were known (not imputed). Hence, optimal LD panels for imputation-mediated genomic prediction were the ones that combined both strengths (group 3 SNP panels). Our results justified the need to include SNPs associated with traits of interest in LD SNP chips for imputation-mediated genomic prediction.

Finally, it is worth mentioning that, in practice, however, it may not be possible to include all trait-specific SNPs in the design of LD SNP chips, but it is favorable to consider some major traits of interest in genomic selection. The differences in both imputation accuracy and genomic prediction accuracy, as were observed in the present study,

were obvious though not drastic, and they could vary with the size of LD panels. As we observed, the differences in imputation and genomic prediction accuracies tend to be diminished as the SNP panel size went beyond 20K (data not presented). Hence, the conclusions of this study are more relevant to the optimal design of LD SNP chips, rather than that for MD or HD SNP chips.

Author contributions — JH and JX analyzed the data. JH and XW drafted the manuscript. XW,JL,SB,GM,SK and MS participated in it's the design and discussions of this research. All authors have proof-read and approved the final manuscript.

Conflict of interest — The authors declare that they have no conflict of interests in this work.

Funding — JH, JX, and JL acknowledge the financial support by University of Nebraska–Lincoln, and GeneSeek (A Neogen company). HJ was also supported by the Bairen Plan of Hunan Province, China (XZ2016-08-07) and Hunan Co-Innovation center of Animal Production Safety, China.

References

Boichard D, Chung H, Dassonneville R, David X, Eggen A, Fritz S, Van Tassell CP (2012) Design of a bovine low-density SNP array optimized for imputation. *PLoS One* 7:e34130

Bolormaa S, Gore K, Werf JHJ, Hayes BJ, Daetwyler HD (2015) Design of a low density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. *Anim Genet* 46:544–556

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* 8:1743–1753

Cooper TA, Wiggans GR, VanRaden PM (2015) Analysis of genomic predictor population for Holstein dairy cattle in the United States—Effects of sex and age. *J Dairy Sci* 98:2785–2788

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11:446

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* 95:4114–4129

Habier D, Fernando RL, Dekkers JCM (2009) Genomic selection using LD marker panels. *Genetics* 182:343–353

Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. *BMC Bioinform* 12:186

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

Jia Y, Jannink JL (2012) Multiple-trait genomic selection methods increase genetic value prediction accuracy. *Genetics* 192:1513–1522

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. *IJCAI-1995*. Morgan Kaufmann, San Mateo, v. 2, pp 1137–1143

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, MacKay TFC, McCarroll SA, Visscher PM (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753

Pimentel EC, Wensch-Dorendorf M, König S, Swalve HH (2013) Enlarging a training set for genomic selection by imputation of ungenotyped animals in populations of varying genetic architecture. *Genet Sel Evol* 45:12

Sargolzaei M, Chesnais JP, Schenkel FS (2014) A new approach for efficient genotype imputation using information from relatives. *BMC Genom* 15:478

van der Werf J (2013) Genomic selection in animal breeding programs. In: Gondro C, van der Werf J, Hayes BJ (ed) *Genome-wide association studies and genomic prediction*. Springer, New York, pp 543–561

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* 92:16–24

Weigel KA, de los Campos G, Gonzalez O. Naya H, Wu XL, Long N, Rosa GJM, Gianola D (2009) Predicting ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *J Dairy Sci* 92:5248–5257

Wiggans GR, Sonstegard TS, VanRaden PM, Matukumalli LK, Schnabel RD, Taylor JF, Schenkel FS, Van Tassell CP (2009) Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. *J Dairy Sci* 92:3431–3436

Wiggans GR, Cooper TA, Vanraden PM, Olson KM, Tooker ME (2012) Use of the Illumina Bovine3K BeadChip in dairy genomic evaluation. *J Dairy Sci* 95:1552–1558

Wu XL, Sun C, Beissinger TM, Rosa GJ, Weigel KA, Gatti Nde L, Gianola D (2012a) Parallel Markov chain Monte Carlo bridging the gap to high-performance Bayesian computation in animal breeding and genetics. *Genet Sel Evol* 44:29

Wu XL, Hayrettin O, Duan H, Beissinger T, Bauck S, Woodward B, Rosa GJ, Weigel KA, de Leon Gatti N, Taylor J, Gianola D (2012b) Parallel-BayesCpC on OSG: Grid-enabled high-throughput computing for genomic selection in practice. *PAG XX,* San Diego

Wu XL, Xu J, Feng G, Wiggans GR, Taylor JF, He J, Qian C, Qiu J, Simpson B, Walker J, Bauck S (2016) Optimal design of low-density SNP arrays for genomic prediction: Algorithm and applications. *PLoS ONE* 11:e0161719

Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: Genetic interactions create phantom heritability. *PNAS* 109:1193–1198