GenTORE

Genomic management Tools to Optimise Resilience and Efficiency

Grant agreement n°: 727213

H2020 - Research and Innovation Action

D4.2
Paper on breed admixture in the multi-breed data

Due date: M42

Actual submission date: M42

Project start date: 1st June 2017 **Duration:** 60 months

Workpackage concerned: WP4

Concerned workpackage leader: Mogens Sandø Lund

Lead Beneficiary: AU

Dissemination level:

- ■ **PU:** Public (must be available on the website)
- ☐ **CO:** Confidential, only for members of the consortium (including the Commission Services)
- ☐ **CI:** Classified, as referred to in Commission Decision 2001/844/EC

Table of content

1. Summary

New genomic prediction methods were developed and tested with simulated data. Genotypic data from real dairy populations, i.e. Danish Holstein, Swedish Red and Danish Jersey, were used as base populations to start genotype simulations for pure breeds and an admixed population. Phenotypes were also simulated, considering two traits which differ in their heritability levels, and different levels of genetic correlations between the breeds.

2. Introduction

Crossbreeding is an efficient strategy in dairy cattle breeding, to achieve better productivity and robustness at the animal and system level. Systems relying on crossbreeding, e.g. ProCROSS system, results in crossbred animals being highly admixed in terms of their breed of origin. Moreover, genomic evaluations in dairy cattle are generally carried out separately for pure breeds, and neither crossbred individuals' data are used, nor they get evaluations. This is partially due lack of methods which can efficiently handle data from pure breeds and admixed individuals. This deliverable provides and discusses such methodologies, which can handle data from multiple pure breeds and admixed individuals, allowing simultaneous evaluation of pure breeds and crossbred animals. The developed methodology relies on accurate estimation of breed origin of each genome segment (represented with a single marker), whose use in genomic prediction models has been proposed earlier, and generally known as BOA (breed origin of alleles).

3. Results

Results for proposed methods using simulated data are presented in Annex, in the format that it was submitted to a peer-reviewed journal, where it is currently under review.

4. Conclusions

The use of admixed individuals' data together with pure breeds' data in genomic prediction has two main consequences: (i) it increases the data size for all pure breeds, particularly for the breed with a small population size, allowing more accurate estimation of breeding values, (ii) it increases the prediction accuracy for admixed individuals.

5. Partners involved in the work

AU, ALLICE (current work)

6. Annexes

Starts on next page.

RESEARCH

Genomic prediction using a reference population of multiple pure breeds and admixed individuals

Emre Karaman[1*], Guosheng Su[1], Iola Croue[2] and Mogens S Lund[1]

[*]Correspondence: emre@qgg.au.dk
[1]Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark
Full list of author information is available at the end of the article

Abstract

Background: In dairy cattle populations where crossbreeding has been used, animals show some level of diversity in their origins. In rotational crossbreeding, for instance, crossbred dams are mated with purebred sires from different pure breeds, and the genetic composition of crossbred animals is an admixture of the breeds included in the rotation. How to use the data of such individuals in genomic evaluations is still an open question. In this study, we aimed at providing methodologies for the use of data from crossbred individuals with admixed genetic background together with data from multiple pure breeds, for the purpose of genomic evaluations for both purebred and crossbred animals. A three-breed rotational crossbreeding system was mimicked using simulations.

Results: For purebred populations, within-breed genomic predictions generally led to higher accuracies than those from multi-breed predictions using combined data of pure breeds. Adding admixed population's (MIX) data to the combined pure breed data as of a different breed led to higher accuracies. When prediction models were able to account for breed origin of alleles, accuracies were generally higher than those from combining all available data, dependent on the correlation of QTL effects between the breeds. Accuracies varied when using SNP effects from any of the pure breeds to predict the breeding values of MIX. Using the breed-specific SNP effects estimated separately in each pure breed, while accounting for breed origin of alleles for the selection candidates of MIX, accuracies were generally improved compared to using SNP effects from multiple pure breed's reference data, but dependent on the correlation of QTL effects between the breeds. Models able to accommodate MIX data with breed origin of alleles approach generally led to higher accuracies than those from the models without considering breed origin of alleles.

Conclusions: Combining all available data, pure breed's and admixed population's data, in a multi-breed reference population is beneficial for the estimation of breeding values for pure breeds with a small reference population. For MIX, such an approach can lead to higher accuracies than considering breed origin of alleles, and using breed-specific SNP effects estimated separately in each pure breed. Including MIX data in the reference population of multiple breeds considering the breed origin of alleles, accuracies can be further improved. Our findings are relevant for breeding programs where crossing is systematically applied, and also for populations involving different subpopulations where exchange of genetic material among those became a routine.

Keywords: admixed population; crossbreeding; breed of origin; multi-breed prediction; genomic selection

Background

Genomic evaluation facilitates the accurate selection of genetically superior individuals as early as their DNA samples are obtained [1]. Genetic progress by selection depends on the accuracy of prediction. For genomic prediction, it depends on the proportion of genetic variance explained by genome-wide single nucleotide polymorphisms (SNPs), and the accuracy with which the effect of those SNPs is estimated [2, 3]. Both factors are conditional on the linkage disequilibrium (LD) between SNPs and quantitative trait loci (QTL) [1–3].

For an accurate genomic prediction, a large population of individuals with both phenotypes and genotypes is needed, which may not be possible for all traits and/or all breeds [4–6]. In such cases, remedies would be to use SNP effects from another breed (a strategy known as across-breed prediction) with a large reference population, or to add data from other breeds (multi-breed prediction) to improve the accuracy of SNP effect estimates. However, the accuracy of across-breed predictions were generally around zero, and combining data from multiple breeds did not notably improve accuracies in empirical studies [7–10].

When multiple breeds are combined to form a reference population, predictions rely on the SNP-QTL LD across breeds. However, LD may be different [11, 12], or the phase of the SNP and QTL alleles may be reversed [13] among the breeds, due to selection and genetic drift [9]. The QTL, or SNPs in high linkage disequilibrium with QTL can be integrated into marker panels for genomic prediction with a multi-breed reference population [14] or performing across-breed predictions [15]. Although this may alleviate the issue of SNP-QTL LD being different in different breeds, it includes an implicit assumption that QTL effects are the same across the breeds. This may not be true if, for instance, QTL by genetic background interactions exist [10, 11]. Hence, it may be more appropriate to assume that QTL, and therefore SNP effects are different but correlated than that they are the same across the breeds.

Crossbreeding emerges as an efficient strategy for dairy cattle breeding to achieve better productivity and robustness at the animal as well as the system level. The improved performance is due to utilization of specific combining abilities and heterosis [16]. In dairy cattle populations, where crossbreeding has been used, animals show different levels of diversity in their origins [11, 17]. In rotational crossbreeding, for instance, where crossbred dams are mated to purebred sires from different pure breeds, the genetic composition of crossbred animals is an admixture of the breeds included in the rotation. At each rotation cycle, depending on the breed of the sires used, admixture proportions of crossbred individuals changes greatly [18]. On the other hand, gene pool of some "purebred" populations may also contain fractions from other breeds, because bulls are used across the breeds to some extent [19]. A prerequisite for a well-structured crossbreeding system is to have an efficient breeding plan within the pure breeds, as well as crossbred population. Because, a sufficient number of purebred bulls are required for the system, and genetic gain in the pure breeds should be maintained so as to ensure that overall economical benefit over time is not negatively affected [20]. Nonetheless, genomic evaluations in dairy cattle are mostly carried out separately for each breed, and neither cross breed data is utilized nor breeders get genomic evaluations for their crossbred animals. It is, therefore, required in some breeding programs that genomic prediction models are able to accommodate a reference population including admixed individuals, as well as multiple pure breeds, allowing simultaneous evaluation of all selection candidates.

An appealing approach to make use of data of admixed individuals in genomic prediction is to incorporate breed proportions in genomic prediction models. Makgahlela et al. [11] extended the random regression model to account for interactions between marker effects and breed proportions, where the breed proportions were inferred from pedigree in Nordic Red Dairy cattle. They reported that prediction accuracy can be higher if breed proportions are considered. Thomasen et al. [21] performed genomic predictions in Danish Jersey dairy cattle, and showed that a model that accounts for breed proportions, estimated either from pedigree or markers, does not improve genomic predictions compared to a model that ignores it. There are at least two limitations with both [11, 21] approaches. First, a single measure of breed proportion may not be appropriate, because two individuals with exactly the same breed proportions may have very different pattern of admixture over their genome depending on which chromosomal region is inherited from which pure breed [22]. Second, the correlations between the breeds were assumed to be homogenous across the whole genome [21], or those correlations were even set to zero due difficulties in estimation [11].

In this article, we propose a methodology suitable for genomic prediction using a reference population of multiple purebred and admixed individuals. Using simulations, we investigated the impact of correlation of QTL effects among the breeds, and the heritability of the trait on accuracy of genomic prediction using different approaches: (i) treating the combined data as of a single homogeneous population, (ii) considering breed-specific SNP effects with/without accounting for correlations between the breeds, and (iii) considering priors that lead to the use of region specific correlations among the breeds.

Methods

Data Simulation

Genotypes

Genotype data at 51,477 loci were available for animals from each of the three dairy cattle breeds: Danish Holstein (HOL), Swedish Red (RED) and Danish Jersey (JER), from which a subset of 1,050 (HOL and RED) or 220 (JER) individuals were formed the base populations for this study. The SNPs which were fixed for the same allele in all three breeds were removed. For computational reasons, only the SNPs (12,664) on first 5 chromosomes were considered. A plot summarising the principle component analysis of genomic relationships among all animals was depicted to assess the genetic relationships among the pure breeds (Additional File 1). In order to establish a data set including multiple pure breeds (i.e., HOL, RED and JER) and an admixed population (hereafter, MIX), a rotational crossbreeding system was mimicked using simulations, considering three cycles of rotation (Table 1) for nine generations. Using the same sets of base population genotype data, a total of 10 replicates were generated.

Simulations started with 1,050 (HOL and RED) or 220 (JER) individuals (generation 0 - G0), of which 50 (HOL and RED) or 20 (JER) were assigned to be males and the rest as females. The purebred populations were generated by mating sires and dams from the same breed (Table 1). Population sizes and the number of males and females were kept constant at each of the simulated generations for HOL, RED and JER. This was achieved by mating 20 dams with the same sire, each mating yielding one offspring, except for one mating which yielded two offsprings, for the simulations of HOL and RED. In simulation of JER population, each sire was mated with 10 dams, where each mating yielding one offspring, except for one mating which yielded two offsprings.

The MIX in G1 was generated by mating sires from JER and dams from HOL of G0. The MIX in G2 was generated by mating sires from RED and dams from MIX of G1. Finally, one rotation cycle was completed with generating MIX in G3 by mating sires from HOL and dams from MIX of G2. The following generations of MIX were generated by mating sires from a pure breed, where the pure breed was dependent on the rotation cycle, with the dams from the MIX (Table 1). Population size and the number of males ($n = 50$) and females ($n = 1,050$) were also kept constant at each of the simulated generations for MIX. When MIX individuals were mated with HOL or RED, the mating structure was similar to that in those pure breeds, whereas when MIX (or HOL) individuals were mated with JER, each JER sire was mated with 50 dams, where 2 or 3 matings per sire was replicated to retain the population size of MIX at 1,050. Selection was not considered, and mating was completely at random.

The number of recombinations on each chromosome was determined using a random variable drawn from a Poisson distribution, under the assumption that the length of a chromosome in the Morgan's unit (we assumed 1 cM $\sim$ 1 Mb) is the lambda parameter [23]. Recombination positions were sampled from a uniform distribution, and interference was ignored [23]. Mutation was not considered in the simulations.

Phenotypes

The total number of QTL was set at 250, which were selected randomly among the SNPs that satisfied $0.01 < \text{MAF} \leq 0.30$, where MAF is the minor allele frequency computed as follows. First, allele frequency at each locus (p_k) was computed for each breed, and then averaged over the breeds ($\bar{p_k}$), to avoid population sizes effecting allele frequencies. Second, MAF of each locus was computed as $min(\bar{p_k}, 1-\bar{p_k})$. The selection of QTL with $0.01 < \text{MAF} \leq 0.30$ ensured that the QTL were segregating with a lower MAF compared to SNPs, for the combined population at G0 (Table 2). The QTL were excluded from the final data set of SNPs. It should be noted that although G0 was common to all 10 replicates, and therefore, the SNPs that met the criteria to be selected as QTL were the same, the QTL or SNP sets did not fully overlap among the replicates due to randomised selection of QTL. The effects (explained below) of QTL were also simulated separately for each replicate.

Even if additive and dominance effects of QTL are the same in different breeds, the difference in QTL allele frequencies may cause substitution effects of QTL [16] to differ among the breeds, as well as genetic (co)variances. In this study, the QTL effects were simulated directly from a multivariate normal distribution for varying levels of correlations among the QTL effects of different breeds, that is correlations of 1.00, 0.50 or 0.25.

Each individual had two alleles (maternal and paternal alleles) at each locus, inherited from its sire and dam. Breed origin of each allele for all loci were traced back to pure breeds at G0, and were known without error. Breeding value of each individual i (u_i) across G0-G9 were generated with:

$$u_i = \sum_{k=1}^{250} \left[Q^M_{ijk} * \alpha^M_{jk} + Q^P_{ijk} * \alpha^P_{jk}\right]$$

where Q^M_{ijk} and Q^P_{ijk} are the number of copies (0 or 1) of an arbitrarily chosen allele A at QTL locus k, inherited from its dam and sire breed j (j=H,R,J), respectively. The α^M_{jk} and α^P_{jk} are the simulated QTL effects for locus k, in breed j. The QTL effects were scaled such that the mean of the breed specific genetic variances (computed as the variance of breeding values) is 100 at G0. A random residual e_j drawn from a normal distribution, $e_j \mid \sigma^2_e \sim N(0, \sigma^2_e)$, was added to each animal's breeding value to form phenotype. The size of σ^2_e was determined according to the simulated heritabilities (explained later) and the mean genetic variance (100) over the breeds. It worths to note that, due to fixing the size of the residual variance across the breeds, heritabilities fluctuated around their mean values over the breeds. The same value of σ^2_e was used in all generations for all individuals.

True (simulated) genetic correlations between the breeds were computed from the genetic variances, $\sigma^2_{u,j} = \sum_{k=1}^{250} 2p_{jk}\left(1-p_{jk}\right)\sigma^2_{\beta_j}$, and covariances, $\sigma_{u,jj'} = \sum_{k=1}^{250} \sqrt{2p_{jk}\left(1-p_{jk}\right)2p_{j'k}\left(1-p_{j'k}\right)}\sigma_{\beta_{jj'}}$ (j=H,S,J and $j \neq j'$) [24] at k QTL. The genetic correlations between HOL-RED, HOL-JER and RED-JER were 0.88, 0.75 and 0.78, respectively, for QTL correlation of 1.00, over 10 replicates and at G0. Those were 0.45, 0.38 and 0.38 for QTL correlation of 0.50, and 0.22, 0.19 and 0.19 for QTL correlation of 0.25, respectively. The differences between QTL effect correlations and genetic correlations were due to the difference in QTL allele frequencies between the breeds. The correlations between QTL allele frequencies of HOL-RED, HOL-JER and RED-JER were 0.33, 0.22 and 0.41, respectively. The correlations between SNP allele frequencies were 0.47, 0.32 and 0.46. The QTL effect correlations of 0.50 and 0.25 are consistent with the reported genomic correlations (genetic correlations estimated based on available SNP set) among some cattle breeds for milk [14, 25] and fat [14], respectively. Two levels of heritability were considered for each scenario of correlations. Those were 0.40 and 0.05, which are of the same magnitude as the reported heritabilities of milk production and fertility traits, respectively (e.g.,[6]).

Reference and Validation Populations

Generations 6,7 and 8 (G6-G8) were used to form reference populations, while generation 9 (G9) was used to form validation populations. Hence, 660 JER individuals, and 3,150 individuals from each of the HOL, RED and MIX were available in forming reference populations to estimate SNP effects.

The SNP effects were estimated using different reference populations: (i) a single pure breed (separate by breeds, i.e., **HOL**, **RED** or **JER**), (ii) combined data of multiple pure breeds **(HOL+RED+JER)**, and (iii) combined data of multiple pure breeds and admixed (MIX) individuals. The MIX data was either used as of a different "breed", assuming homogeneous SNP effects across all breeds **(HOL+RED+JER+MIX)**, or truly treated as an admixed population considering breed origin of alleles (BOA) approach **(HOL+RED+JER+MIX uncor/cor)**.

The prediction of breeding values for each pure breed were performed using: (1) the estimated SNP effects from their own breed (within-breed prediction), (2) the estimated SNP effects from each of the other breeds (across-breed prediction), (3) the estimated SNP effects from a combined reference population (multi-breed prediction) and (4) the estimated SNP effects from a combined reference population considering BOA approach. The breeding values were predicted by multiplying SNP effects with allele dosages, with (4) or without (1-3) considering breed origin of alleles. These same strategies (1-4) were used to predict the breeding values of admixed individuals. For the admixed individuals, SNP effects estimated separately using pure breed reference populations **(HOL/RED/JER)** were additionally used to predict breeding values, considering the BOA approach only for the validation animals (hereafter, pure-BOA). That is, breed origin of each SNP allele was traced back to its pure breed population only for the validation population, and the number of counted alleles were multiplied by the breed specific estimate of SNP effects of the pure breeds.

We classified the methods using only a single breed's data in model **training** to estimate SNP effects as **pure** (also includes pure-BOA as explained above), multiple breeds data without considering breed origin of alleles as **combined**, and multiple breed's and MIX data considering breed origin of alleles as **BOA**.

Statistical Models

Pure and combined

A simple approach for genomic prediction using a combined reference population of multiple pure breeds and/or admixed individuals is to assume that the marker effects are the same across the breeds [26]. For this simple approach, when the data consisted of multiple breeds treated as of a single homogeneous population (Combined), we used the following model:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{X}\mathbf{b} + \mathbf{M}\boldsymbol{\beta} + \mathbf{e} \tag{1}$$

In the above equation, $\mathbf{y}$ is a vector of phenotypes ($n \times 1$), $\mathbf{1}$ is the vector of 1s, μ is the general mean, $\mathbf{X}$ is the matrix of breed proportions ($n \times 3$), $\mathbf{b}$ is the vector of fixed breed effects (3×1), $\mathbf{M}$ is the matrix of centered genotypes ($n \times l$) where centering was based on the current allele frequencies in the combined data, $\boldsymbol{\beta}$ is the vector of SNP effects, and $\mathbf{e}$ is the vector of residuals ($n \times 1$). The value of n depends on the reference population size, and l is the number of SNPs. Model (1) was used without the breed proportions component $\mathbf{Xb}$ when the SNP effects were estimated separately for each pure breed (Pure and Pure-BOA).

BOA

Admixed breeds' data was utilised by extending the existing linear model proposed for simple 2-way crosses (e.g., [27]) to accommodate more than two pure breeds:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{M}_H\boldsymbol{\beta}_H + \mathbf{M}_R\boldsymbol{\beta}_R + \mathbf{M}_J\boldsymbol{\beta}_J + \mathbf{e} \tag{2}$$

where $\mathbf{y}$ is the vector of phenotypes $(n \times 1)$ of all animals, that is, both purebred and admixed animals. The $\mathbf{1}$ is a vector of 1s, μ is the general mean, $\mathbf{M}_H$, $\mathbf{M}_R$ and $\mathbf{M}_J$ are the matrices of breed specific allele content of SNPs $(n \times l)$ for HOL, RED and JER, respectively. The entry at a locus in, for instance $\mathbf{M}_H$, for an animal were the number (0,1 or 2) of counted allele A originated from HOL. That is, when the animal had no allele originating from HOL, or when a HOL animal had an aa genotype, the corresponding entry was zero. The same applied to matrices $\mathbf{M}_R$ and $\mathbf{M}_J$. The matrices were column centered prior to analysis. The $\boldsymbol{\beta}_H$, $\boldsymbol{\beta}_R$ and $\boldsymbol{\beta}_J$ are vectors of SNP effects for HOL, RED and JER, respectively, and $\mathbf{e}$ is the vector of residuals.

Bayesian Analysis

Bayesian approach was considered in parameter estimation, which requires assigning prior distributions to the unknowns of the model. Analyses were carried out separately for each trait. To investigate the impact of assuming heterogeneous (co)variance of SNP effects among different genome regions, three region sizes were considered based on a fixed number of SNPs; 1 SNP, 100 SNPs and the whole genome (WG). Regions sizes of 1 SNP and WG can be regarded as BayesA and SNP-BLUP [1] (or equivalently GBLUP [28]) when using model (1), and extensions of them for multiple components (breeds) when using model (2), respectively. In BayesA it is assumed that each SNP (1 SNP) follows a normal distribution with null mean and a locus-specific variance, while in GBLUP it is assumed that all SNPs (WG) have null means and a common variance. To consider heterogeneous variance of SNP effects among different genome regions using model (1), the matrix of genotypes and vector of SNP effects were partitioned into S subsets each with k_s loci $(s = 1, \ldots, S)$, and priors were assigned to each sub-vector of $\boldsymbol{\beta}$: $\boldsymbol{\beta}_s \mid \sigma_s^2 \sim N(\mathbf{0}, \mathbf{I}\sigma_s^2)$ [29, 30]. The σ_s^2(s) were further assigned a scaled inverse chi-square prior with a degrees of freedom (df) and a scale parameter (S): $\sigma_s^2 \mid df, S \sim \chi^2(df, S)$. The values of hyper-parameters will be explained later.

In the analyses using model (2), all genotype matrices and vectors of SNP effects were also partitioned into S subsets each with l_s loci. A normal distribution prior was assigned for each sub-vector of SNP effects for population j (j=H,R,J): $\boldsymbol{\beta}_{j,s} \mid \sigma_{j,s}^2 \sim N(\mathbf{0}, \mathbf{I}\sigma_{j,s}^2)$. Hence, the SNP effects were breed-specific and uncorrelated across the breeds. That is, the genetic correlations among the breeds were assumed to be zero. The $\sigma_{j,s}^2$(s) were further assigned a scaled inverse chi-square prior with a degrees of freedom (df_j) and a scale (S_j) parameter: $\sigma_{js}^2 \mid df_j, S_j \sim \chi^2(df_j, S_j)$. Using model (2), priors were also assigned such that the marker effects were breed-specific, but correlated between the breeds. That is, a multivariate normal distribution was assigned for each sub-vector of SNP effects: $\left[\boldsymbol{\beta}_{H,s}\, \boldsymbol{\beta}_{R,s}\, \boldsymbol{\beta}_{J,s}\right]' \mid \mathbf{B}_s \sim N\Big(\mathbf{0}, \mathbf{B}_s \otimes \mathbf{I}\Big)$, where $\mathbf{I}$ is an identity matrix of size equal to l_s if $l_s > 1$ or a scalar of 1 if $l_s = 1$.

$$\mathbf{B}_s = \begin{bmatrix} \sigma^2_{H,s} & \sigma_{HR,s} & \sigma_{HJ,s} \\ \sigma_{RH,s} & \sigma^2_{R,s} & \sigma_{RJ,s} \\ \sigma_{JH,s} & \sigma_{J,s} & \sigma^2_{J,s} \end{bmatrix}$$

The diagonals of $\mathbf{B}_s$ are the breed-specific SNP variances, while the off-diagonals are SNP covariances between the breeds. The $\mathbf{B}_s$ was assumed to follow an inverted Wishart distribution with a shape (v_B) and a scale ($\mathbf{V}_B$) parameter: $\mathbf{B}_s \mid v_B, \mathbf{V}_B \sim IW(v_B, \mathbf{V}_B)$.

In both model (1) and (2), residuals were assigned a univariate normal prior, $e_i \mid \sigma^2_e \sim N(0, \sigma^2_e)$, and variance σ^2_e were assigned a scaled inverse chi-square prior with a degrees of freedom (df_e) and a scale (S_e) parameter: $\sigma^2_e \mid df_e, S_e \sim \chi^2(df_e, S_e)$. Fixed effects were assigned flat priors.

The hyper-parameters of the prior distributions for the variance components were driven from the simulated genetic (co)variances and residual variances at G0 as follows. For the analysis using model (2) assuming independent SNP effects among the breeds, $df_j = 4$ and $S_j = \frac{\sigma^2_{\beta_{j,s}}(df_j - 2)}{df_j}$, where $\sigma^2_{\beta_{j,s}} = \frac{\sigma^2_{u,j}}{\sum 2p_{j,l}(1-p_{j,l})}$ [31]. Here, $\sigma^2_{u,j}$ is the genetic variance for breed j, and p_{jl} is the allele frequency of l'th SNP in breed j. Only one S_j was required for the analysis using model (1), which was computed using $\sigma^2_{u,j}$ (pure breed analysis) or the mean value of $\sigma^2_{u,j}$ over the breeds (combined analysis). For the analysis using model (2) assuming correlated SNP effects between the breeds, $\mathbf{V}_B = (v_B - 3 - 1)\mathbf{B}$ where $v_B = 6$, and

$$\mathbf{B} = \begin{bmatrix} \frac{\sigma^2_{u,H}}{\sum 2p_{H,j}(1-p_{H,j})} & \frac{\sigma_{u,HR}}{\sum \sqrt{2p_{H,j}(1-p_{H,j})}\sqrt{2p_{R,j}(1-p_{R,j})}} & \frac{\sigma_{u,HJ}}{\sum \sqrt{2p_{H,j}(1-p_{H,j})}\sqrt{2p_{J,j}(1-p_{J,j})}} \\ \frac{\sigma_{u,RH}}{\sum \sqrt{2p_{R,j}(1-p_{R,j})}\sqrt{2p_{H,j}(1-p_{H,j})}} & \frac{\sigma^2_{u,R}}{\sum 2p_{R,j}(1-p_{R,j})} & \frac{\sigma_{u,RJ}}{\sum \sqrt{2p_{R,j}(1-p_{R,j})}\sqrt{2p_{J,j}(1-p_{J,j})}} \\ \frac{\sigma_{u,JH}}{\sum \sqrt{2p_{J,j}(1-p_{J,j})}\sqrt{2p_{H,j}(1-p_{H,j})}} & \frac{\sigma_{u,JR}}{\sum \sqrt{2p_{J,j}(1-p_{J,j})}\sqrt{2p_{R,j}(1-p_{R,j})}} & \frac{\sigma^2_{u,J}}{\sum 2p_{J,j}(1-p_{J,j})} \end{bmatrix}$$

In the above equation, $\sigma^2_{u,j}$ and $\sigma_{u,jj'}$ (j=H,R,J and $j \neq j'$) are genetic variances and covariances respectively. For residual variances, $df_e = 4$ and $S_e = \frac{\sigma^2_e(df_e - 2)}{df_e}$, where σ^2_e is the residual variance at G0.

Markov-chain Monte Carlo (MCMC) algorithm with Gibbs sampling method was used to obtain samples of each parameter from its full conditional posterior distribution. Chain length for the analyses consisted of 50,000 cycles, of which the first 10,000 were discarded as burn-in. Every 10th sample of the post burn-in cycles were stored for posterior analysis, yielding 4,000 posterior samples. Mean value of the posterior samples was used as the estimate of each parameter. All analysis were performed using self-written scripts in Julia [32].

Prediction accuracy

Prediction accuracy was assessed as the correlation between true and predicted breeding values of validation individuals (1,050 individuals for HOL, RED, MIX, and 220 individuals for JER) at G9. Accuracy of prediction using different data sets and models to estimate SNP effects were compared for each trait, QTL correlation and region size separately. Accuracy of prediction for different region sizes were compared for each data set and model, trait and QTL correlation, separately. All comparisons were performed using a two-sided paired t-tests, for which accuracies were paired across each replicate for the same validation population. A Bonferroni correction was used to control the Type 1 error rate of 0.05.

Results

Accuracies for all scenarios and all region sizes are given in Supplementary Tables 1-4, Additional file 2. For readability, only the core results from QTL effect correlation of 0.5 are presented in the main text. Accuracies were higher for high heritability trait than those for low heritability trait (Figures 1 and 2). Within-breed predictions in breeds with large reference populations (HOL and RED) were more accurate than in a breed with small reference population (JER). For the high heritability trait, within-breed predictions for HOL, RED and JER were 0.785, 0.747 and 0.629, respectively, when the region size was 1 SNP (Figure 1). For this high heritability trait, combining data from multiple pure breeds (HOL+RED+JER) assuming homogenous SNP effects (multi-breed prediction) did not improve, or even decreased (though not always significant) the accuracies for all breeds. Including the admixed population's (MIX) data in multi-breed prediction, as if it belongs to a different breed (HOL+RED+JER+MIX), yielded higher accuracies compared with combining only the data from pure breeds, and similar to or higher accuracies than using the single breed data alone (within-breed prediction), for genomic prediction of JER. When prediction models were able to accommodate data of admixed individuals by accounting for breed origin of alleles (HOL+RED+JER+MIX uncor/cor), accuracies were generally improved compared to combining all available data, but dependent on the correlation scenario. Across-breed predictions yielded accuracies much lower than within-breed predictions.

Accuracies were the lowest when using SNP effects from any of the pure breeds to predict the breeding values of admixed individuals. For the high heritability trait, predictions using SNP effects of HOL, RED and JER yielded accuracies of 0.411, 0.275 and 0.114, respectively, when the region size was 1 SNP (Figure 1). For the same scenario and region size, estimating SNP effects separately for each breed, but accounting for breed origin of alleles in prediction of breeding values (HOL/RED/JER) of MIX, improved accuracy up to 0.531. Combining MIX data with pure breeds' data assuming common SNP effects for all breeds (HOL+RED+JER+MIX), improved accuracies over combining only three pure breeds' data (HOL+RED+JER) for the accuracy of admixed individuals (0.792 vs 0.501). Models able to use MIX data with breed origin of alleles (HOL+RED+JER+MIX uncor/cor), improved accuracies over combining all available data, i.e., combining all purebred data or all purebred and admixed individuals' data, though dependent on the correlation of QTL scenario. For QTL correlation of 1.0, and predictions in MIX, (HOL+RED+JER+MIX) led to higher accuracies than (HOL+RED+JER+MIX uncor). Accounting (0.877) or not (0.876) for correlations between the SNP effects of different pure breeds did not make any difference (Figure 1). Among different region sizes considered here, region size of whole genome generally yielded the lowest accuracies for pure breeds and admixed population (Figure 3).

The importance of the methods considering breed origin of alleles in model training became more apparent as the correlation of the true QTL effects between the breeds decreased (See Suplementary Tables 1-4, Additional File 2). For high heritability trait and purebred populations, accuracies for (HOL+RED+JER+MIX uncor/cor) were significantly higher than those for (HOL+RED+JER+MIX) in QTL effect correlation of 0.25. For MIX population, (HOL+RED+JER+MIX uncor/cor) yielded significantly higher accuracies than those for (HOL+RED+JER+MIX), for QTL effect correlation of 0.25, and for both traits.

Discussion

Within- and across-breed predictions

The simple approach for avoiding the unfavourable impact of the difference in marker effects among the different purebred populations is to carry out separate evaluations for each of those pure breeds, as is the case for genomic evaluations in many countries [19]. Such an approach, however, comes with the cost of a potential loss of data information resource, and therefore, in the accuracy of SNP effect estimation. This is a limitation for genetic improvement in populations with a small reference population. In this study, accuracies from within breed predictions were higher for HOL and RED, compared to JER. Although there could be other reasons, one explanation is the low reference population size (660 vs 3,150) set for JER. The accuracies for pure breeds differed between the two heritability levels for any QTL effect correlation scenario, with the high heritability trait having higher accuracies than the low heritability trait. The fact that genomic prediction accuracy is higher with large reference populations and/or for a high heritability trait was reported in many other studies [3, 5, 33–35]. It should be noted that the accuracies for the same heritability level fluctuated slightly for different QTL effect correlation scenarios, because QTL effects were simulated using different multivariate normal distributions (the covariance matrices differed) for those scenarios.

Using SNP effects of one pure breed to predict the breeding values of individuals of other breeds (across-breed prediction), yielded much lower accuracies than within-breed predictions. This was true even when the simulated QTL effects had a correlation of one. This is in line with the study of Steyn et al. [36] where several breeds were simulated assuming identical QTL effects, but across-breed predictions were poor. Studies using real data also showed that using data from one breed to predict breeding values in other breeds results in accuracies as low as zero (e.g., [9, 10]). The prediction accuracy of MIX generally reflected the expected breed proportions of the validation individuals. Using SNP effects from HOL, for instance, led to the highest prediction accuracies for MIX, as HOL was the most recent ancestor population for MIX, and therefore, MIX individuals had a higher proportion of their genome from HOL.

For within-breed predictions, both family relationships and linkage disequilibrium (LD) between SNPs and QTL contribute to accuracy [37–39]. For across-breed prediction, the relationships of the individuals of the target breed with the individuals in the reference population are lower than those with the members of the target breed. The relative contributions of the two factors, family relationships and LD, to accuracy of breeding value estimation were not studied as it was not in the scope of this paper. If we rely on the argument that low across-breed prediction accuracy is due to differences in LD patterns among the breeds, that is the differences in LD or the phase of the SNP and QTL alleles, then across-breed prediction can not compete with within-breed prediction, even for closely related breeds. In addition to LD patterns, it is also possible that QTL effects and/or QTL allele frequencies differ among the breeds, while some QTL may only segregate in one breed [25, 40]. Needless to say, even if the QTL properties were the same among the breeds, SNP effects would still be different to the extent to which LD between SNPs and QTL differs between them [11, 12, 30, 41].

Although the simulated traits in this study were relatively polygenic, the variance structure at the SNP level may be different from that of at the QTL level across the genome [42, 43], favouring models that can accommodate such heterogeneity [30, 44, 45]. The SNP panels tend to include SNPs with high minor allele frequency (MAF), while the QTL have generally low MAF [46, 47]. The LD between the two sets, SNPs and QTL, can not be perfect if their MAF differs. Because the SNPs in a genome region are likely inherited together, and also likely to be in LD with the same QTL, they may collectively capture the genetic variance at the QTL [29, 45, 48]. Hence, assuming a common variance for groups of adjacent SNPs is reasonable, while it allows more Bayesian learning compared to assuming variance specific to every single SNP [49]. For region sizes larger than an optimum level, on the other hand, the advantage of grouping adjacent SNPs will start to disappear as the assumption on (co)variance will approach to that of whole genome region size (WG).

For high heritability trait and purebred analysis, accuracies from different region sizes were generally ranked as 100 SNPs $>$ 1 SNP $>$ WG. It was shown earlier by simulations [30, 35] and real data analysis that assigning priors to groups of SNPs may improve accuracies [44, 45] compared to assigning a common prior for all SNPs. In a recent study, on the other hand, Liu et al., [50] reported negligible differences between several region sizes, 1- 30- or 100 SNPs and WG, for milk production and fertility traits in Danish Jersey, and using a model which is nearly identical to our model (1).

Combined data of multiple pure breeds

If the studied population is small, it might be challenging to establish a large reference population, and in turn the accuracy of genomic prediction might also be limited [6]. For breeds with a limited reference population size, incorporating data from other breeds may yield higher accuracies [26, 40, 51], though dependent on the relatedness of those breeds [9, 19]. When HOL and RED individuals were included in the reference population of JER (HOL+RED+JER), accuracies generally dropped. Similarly, using that combined reference population, accuracies for HOL and RED also generally dropped, but less compared to those for JER. When multiple purebred populations are combined to form a reference population, SNP effects are dominated by the breeds contributing more to the reference population. This may cause prediction models to pick up only the effect of SNPs that are in LD with QTL in all breeds, and/or only in the largest population, but not the effect of SNPs specific to small populations [14]. We had additional simulations where all breeds had the same number of individuals in the reference population (3,150 for each), which led accuracies for JER also to be high and get less affected from the joint analysis, as HOL and RED (results not given). These imply that the proportion of each single breed in a combined reference population of multiple breeds is important to achieve a sufficient accuracy for each breed, particularly when the breeds are genetically distant. This was more formally investigated in [40] using a high density SNP chip ($\sim 600,000$ SNPs), where one of the two breeds (Holstein and Jersey) that formed a joint reference population had varying sizes, 0, 100, 500 or 2,000 animals, while the size of the other breed kept constant at 2,000 animals. As the number of individuals of a breed in the joint reference population decreased, accuracies for the candidates of the same breed also decreased [40].

In a study based on real genotypes of imputed sequence variants ($\sim$ 1 million SNPs), van den Berg et al. [52] simulated phenotypes for four dairy cattle breeds using identical QTL effects. They reported generally higher accuracies for multi-breed predictions, compared to within-breed predictions. In our scenario of QTL effect correlation of 1.0, the difference in the accuracies from within- and multi-breed predictions were smaller compared to other (lower) QTL effect correlation scenarios. At long distances of genome, LD differs between species and also between different cattle breeds, whereas it is relatively consistent at short distances [3]. The standard SNP sets, such as the one used here, are not sufficient to include all such SNPs that are in high LD with QTL across the breeds. Moreover, we selected QTL such that they had relatively low MAF compared to SNPs, whereas QTL were randomly selected in [52], which have an impact on LD between QTL and SNPs. These may partially explain why multi-breed genomic predictions generally had lower accuracies than within-breed predictions even when the simulated QTL effects were identical, compared to the findings of [52].

For the analysis of data consisting of multiple breeds (or lines, populations), an appealing strategy is to apply multi-trait methods where the same trait in different breeds is considered as different but correlated traits, e.g. [8, 25]. In those applications of multi-breed genomic prediction, however, a homogeneous genomic correlation was assumed across the genome, for pairs of breeds. Lehermeier et al. [41] applied a multivariate modelling approach, which is flexible in that both marker effects and their (co)variances are allowed to differ among multiple breeds, but still assumes a homogenous correlation across the genome of breed pairs. Chen et al. [53] proposed a method which allows the estimation of SNP effects specific to each breed while accounting for heterogenous (co)variances across the genome. Their method, however, applies a variable selection procedure aiming to pinpoint the SNPs that have an effect in all breeds involved, leaving out the SNPs with effect only on one or a subset of the breeds. It was further extended by Calus et al. [10] so as to accommodate also the selection of SNPs that are breed-specific. Nevertheless, both methods [10, 53], make limited use of the correlated information in the data, because, regardless of how the SNPs to be included in the model are selected, their effects are estimated separately within each breed. Furthermore, all those multi-trait approaches are pertained to situations where individuals can be assigned to certain pure breeds, and are not able to accommodate data of individuals with admixed genetic background.

Genomic prediction including data from admixed individuals

If a large number of commercial farm data for admixed populations becomes available, it can help to improve selection accuracy by expanding the data size for each pure breed population. Such data can also allow to exploit heterosis due dominance, which would not be possible with purebred data [12]. How to use those data in genomic evaluations is still an open question. Naturally, all purebred and admixed individual data can be combined together, when homogeneous SNP effects are assumed.

Including the data of admixed population (hereafter, MIX) along with the data of pure breeds in the reference population led to higher accuracies than the combined reference population of pure breeds. The JER benefited relatively more from adding MIX data. Because we mimicked a rotational crossing system, at each generation, admixed population individuals were sired by a purebred individual. Consequently, when an admixed female was mated with a purebred male, the offspring had an entire paternal chromosome from a pure breed, and a maternal chromosome including large chunks of (i) admixture of all breeds and (ii) the pure breed of the maternal grand-sire. This means that, at each generation following G1, pure breeds were not equally represented in the genome of admixed individuals. Consider a single admixed individual at generation 6. That individual has expected breed composition for a maternal chromosome of roughly 28% JER, 16% HOL, and 56% RED, and for a paternal chromosome of %100 HOL. Those proportions change to be 14% JER, 58% HOL, and 28% RED for a maternal chromosome, and 100% JER for a paternal chromosome at generation 7, and to 57% JER, 29% HOL, and 14% RED for a maternal chromosome, and 100% RED for a paternal chromosome at generation 8. Because a full rotation cycle of three generations (G6-G8) was considered when forming the reference populations, each pure breed was represented in the MIX data almost equally. Thereby, the reference population size indeed increased almost equally for all breeds by adding MIX data to the combined data of three breeds, HOL+RED+JER+MIX. As one would expect, JER benefited more from this increase in data size, as it is the breed with the smallest pure breed reference population. It should be noted that the validation individuals of RED had the grand-sires which were also the sires of MIX at G8, and G8 was included in the reference population. Hence, although the data size increased almost equally for each breed, the information in the data may not be equally informative for all the breeds.

More elaborative ways to include individuals with admixed genetic background in the genomic evaluations, were proposed. Makgahlela et al. [11] fitted a multi-trait random regression model to account for interactions between marker effects and breed proportions, where the breed proportions were inferred from pedigree in Nordic Red Dairy cattle. They reported, for some traits, higher prediction accuracies for the model accounting for breed proportions, than a GBLUP model treating the data as of a single homogeneous population. Another example of admixture is admixture due to different populations, instead of breeds. Danish Jersey dairy cattle, for instance, includes animals with different proportions of their genome from original Danish and US Jersey populations [19, 21]. Although both originate from a single breed, they have been separated long ago, and the persistency of phase were shown to differ between the two, particularly at long distances of loci [21]. Hence, the accuracy of genomic prediction for Danish Jersey may not only be challenged by the small reference population size, but also by its admixed population structure. In order to overcome the negative impact of admixed population structure in Danish Jersey on genomic prediction accuracy, Thomasen et al. [21] applied a set of random regression models that included proportions of population origin for each animal. Contrariwise, they [21] did not find any strong evidence that a model which accounts for proportions of population origin, estimated either from pedigree or markers, is superior to a model which ignores it. A possible explanation could be that admixture due to different breeds may be a more important problem than admixture due to subpopulations of the same breed, in genomic prediction. Nevertheless, there are at least two limitations with both [11, 21] approaches. First, breed proportions of an individual were average values over their whole genome, because they were computed based solely on pedigree or markers. This may not be appropriate, as two individuals with exactly the same breed proportions may have very different admixture patterns over their genome depending on which chromosomal region is inherited from which pure breed [21, 22]. Second, their models are somewhat restricted in that the correlations between the breeds were assumed to be homogenous across the whole genome [21], or those correlations were even set to zero for difficulties in estimation [11]. When the breeds are in different SNP-QTL LD, the (co)variances of SNP effects are expected to differ over the genome, and across the breeds [11, 21, 22, 41].

Genomic prediction considering breed origin of alleles

Models accounting for breed origin of each SNP allele, rather than genome-wide breed proportions estimated from pedigree or markers, have been proposed, and were shown to improve genomic predictions for simple 2 or 3-way crosses. Those studies applied either univariate whole genome regression models at the SNP level ignoring that the SNP effects might be correlated between the pure-breeds [27, 54], or rather computationally demanding multi-trait genomic BLUP models with "partial" relationship matrices at the individual level [22, 55, 56]. It was claimed that considering genomic correlations between pure-breed populations had limited relevance in models for predicting crossbred performance [22, 55, 56].

Our results did not show any clear evidence of the benefit of accounting for correlations among the breeds when MIX data were used with BOA approach, even for the breed (JER) with a small reference population, for which one would expect more gain in accuracy compared with breeds with a large reference population (HOL and RED). A possible explanation of unobserved benefit for JER could be due to this breed being genetically distinct from HOL and RED [57], and therefore, the pattern of SNP effects over the genome being different from HOL and RED. Additionally, the information in the data may be weak to estimate correlations among the breeds. The MIX data also increased the within-breed data size to some extent, which may lower the importance of correlated information from other breeds [41]. For the scenario of QTL effect correlation 1.00, analysis with HOL+RED+JER+MIX was competitive with or even superior to analysis using BOA without accounting for correlations between the breeds, particularly in predicting breeding values MIX. This is may be due including MIX individuals in the reference population simply increasing the data size in a joint analysis, whereas BOA with uncorrelated analysis utilize only the information in a single breed.

The differences in LD pattern and phase persistency across different breeds [43] may result in marker effects to be highly correlated for regions, where LD and SNP-QTL phase is constant between the breeds [41]. Hence, we have anticipated that correlations among the populations at the region level might improve the accuracy of genomic predictions, even though the correlations at the whole genome level do not. In this study, the differences in accuracies from 100 SNPs and 1 SNP region sizes were generally negligible, whereas WG generally yielded the lowest accuracies. It worths to note that, however, the fixed-length of 100 SNPs as region size was arbitrarily chosen to give an insight on the impact of grouping SNPs in within-, across- and multi-breed genomic prediction accuracy, and there may exist other region sizes to yield higher accuracies than 100 SNPs. In analysis aiming to utilize correlations between the breeds, such as analysis using BOA approach, the knowledge of the LD patterns and persistence of phase among the breeds may be useful in grouping SNPs.

van den Berg et al., [14] showed that prediction of breeding values and genomic correlations across populations can be more accurate if a carefully selected set of causal variants or SNPs that are very close to causal variants from sequencing data are used together with commercial SNP panels. Doing so may alleviate the issue of SNP-QTL LD being different in different breeds. In a recent study, Liu et al. [6] showed that integrating additional selected sequence variants to the standard 54K SNP chip led to significant improvements of reliabilities for the genomic evaluation of milk production traits in Danish Jersey. They reported that the benefits of using selected sequence variants in genomic prediction for milk and protein remained significant even in the scenario of the largest reference population consisting of animals from Danish and US Jersey populations. In order to eliminate the impact of LD differences among the breeds on the comparison of accuracy for using the two BOA approaches (correlated and uncorrelated SNP effects), we ran additional analyses for the QTL correlation scenario of 0.50 and the low heritability trait ($h^2 = 0.05$), using only the 250 QTL as SNPs and the region size of 1 SNP. Analyses using BOA approach assuming correlated SNP effects were higher than those assuming uncorrelated SNP effects between the breeds (Figure 4). In light of these, one can argue that integrating selected sequence variants may be an efficient way of using correlated information from the breeds, and that in this case taking into account the correlation of SNP effects between breeds may allow for greater accuracy, in genomic evaluations with data from multiple purebred and admixed individuals using BOA approach.

Estimation of the breed composition of individuals with admixed genomic background is of relevance for genomic prediction, because if not accounted for it may lead to spurious estimates of SNP effects [58]. In real life applications, pedigree records and/or parentage validation can be used to distinguish purebred and admixed animals, but any error in the pedigree may lead to inaccurate consideration of individuals as pure or admixed [18]. Nevertheless, genomic prediction should rely on local ancestry (i.e., breed of origin) for each of the SNP alleles, rather than a genome-wide (global) ancestry computed from pedigree or markers [59]. Methods exist to estimate local ancestry in a population of admixed individuals (e.g., [60]). In this simulation study, breed origin of admixed individuals were known without error, but those could also be estimated from the data of purebred individuals. Due to mimicking a systematic crossing scheme in our simulations with well-defined purebred individuals, such estimates are expected to be highly accurate

(Ana C. Guillenea, personal communication). For populations, where admixture is more complex, however, one first needs to find the number of pure breeds in the gene pool, and then to assign breed origin to each SNP allele for all animals in the population. This may introduce another source of error, and the models requiring breed origin of alleles, with or without accounting for correlations, may suffer from such errors to the extent where simply combining all available data (multiple pure and admixed breeds data) might become highly competitive. It was shown that a higher number of animals would be required to distinguish closely related breeds than to distinguish distantly related breeds [61], when the breed origin of an animal is needed to be inferred from the genotypic data. To the best of our knowledge, there is no information on the number of purebred animals required to correctly assign breed-origin of alleles of the crossbred animals.

Genome scaling

Approximations for genomic prediction accuracy [3, 62] use the size of the reference population (n_R), trait heritability (h^2), and the effective number of chromosomal segments segregating in the population (M_e), where M_e is a function of the genome length and the effective population size (N_e). Following those studies [3, 62], within-breed prediction accuracy can be estimated with $\sqrt{h^2 n_R / (h^2 n_R + M_e)}$. In this study, only the first five chromosomes were simulated, which is roughly a quarter of the cattle genome. Those approximations suggest that, if we scale up the genome size (and the number of QTL) to that of the whole genome, and the size of the reference populations accordingly, our results will still hold, in within-breed predictions. For across-breed prediction, Wientjes et al. [63], suggested to use of $r_g \sqrt{h^2 n_R / (h^2 n_R + M_e)}$, where r_g is the genetic correlation between the breeds. They further suggested that M_e values of 20,000 and 40,000 may be used when the populations are closely and distantly related, respectively. On the other hand, combining different breeds together will increase N_e [64], and thereby M_e, requiring a larger reference population size to compensate this increase in M_e, to avoid a reduction in accuracy [36, 52]. Models accounting for BOA, on the other hand, make use of single-breed data, while taking advantage of an increase in n_R by using data from admixed individuals. The BOA model with correlations further utilizes correlated information from other breeds. It worths to note that those approximations assume a single homogenous target (validation) population.

Conclusion

The aim of this simulation study was to provide a model allowing the inclusion of data from individuals with admixed genetic background in genomic evaluations, while accounting for the differences of marker effects for purebred populations in the gene pool. Combining pure breeds' and admixed population's data, in a multi-breed reference population was beneficial for the estimation of breeding values for pure breeds with a small reference population. For the admixed population, combining all available data (from purebred and admixed individuals) and realizing a combined genomic evaluation led to higher accuracies than considering BOA for selection candidates only, and using breed-specific SNP effects estimated separately in each pure breed. Including admixed individual's data in the reference population of multiple breeds considering the BOA approach, accuracies were further improved. Our findings are relevant for breeding programs where crossing is systematically applied (e.g., ProCROSS system, http://www.procross.info), and also for populations involving different subpopulations where exchange of genetic materials among those became a routine (e.g., Nordic Red dairy cattle).

Competing interests

The authors declare that they have no competing interests.

Availability of data and material

The datasets used during the current study are available from the corresponding author on reasonable request.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation programme - GenTORE - under grant agreement No 727213.

Author's contributions

EK simulated the data, contributed to the formulation of the methods, implemented the methods and performed the analysis, and drafted the manuscript. GS co-supervised the study, contributed to the formulation of the methods, and revised the manuscript. IC contributed to the design of data simulation and discussion of the results. MSL conceived and supervised the study, contributed to the formulation of the methods and discussion of the results. All authors read and approved the final manuscript.

Author details

[1]Center for Quantitative Genetics and Genomics, Aarhus University, 8830 Tjele, Denmark. [2]ALLICE, F-78350 Jouy-en-Josas, France.

References

1. Meuwissen, T., Hayes, B., Goddard, M.: Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**, 1819–1829 (2001)
2. Dekkers, J.: Prediction of response to marker assisted and genomic selection using selection index theory. J Anim Breed Genet **124**, 591–611 (2007)
3. Goddard, M.: Genomic selection: prediction of accuracy and maximization of long term response. Genetica **136**(245-257) (2009)
4. Hayes, B., Bowman, P., Chamberlain, A., Verbyla, K., Goddard, M.: Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet Sel Evol, 41–51 (2009)
5. Karaman, E., Cheng, H., Firat, M., Garrick, D., Fernando, R.: An upper bound for accuracy of prediction using GBLUP. PLoS ONE **11**, 0161054 (2016)
6. Liu, A., Lund, M., Boichard, D., Karaman, E., Fritz, S., Aamand, G., Nielsen, U., Wang, Y., Su, G.: Improvement of genomic prediction by integrating additional single nucleotide polymorphisms selected from imputed whole genome sequencing data. Heredity (2019)
7. Erbe, M., Hayes, B., Matukumalli, L., Goswami, S., Bowman, P., Reich, C., Mason, B., Goddard, M.: Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci **95**(7), 4114–29 (2012)
8. Olson, K., VanRaden, P., Tooker, M.: Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. J Dairy Sci **95**(9), 5378–5383 (2012)
9. Kachman, S., Spangler, M., Bennett, G., Hanford, K., Kuehn, L., Snelling, W., Thallman, R., Saatchi, M., Garrick, D., Schnabel, R., Taylor, J., Pollak, E.: Comparison of molecular breeding values based on within- and across-breed training in beef cattle. Genet Sel Evol **45**(1), 30 (2013)
10. Calus, M., Goddard, M., Wientjes, Y., Bowman, P., Hayes, B.: Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection. J Dairy Sci **101**(5), 4279–4294 (2018)
11. Makgahlela, M., Mantysaari, E., Stranden, I., Koivula, M., Nielsen, U., Sillanpää, M., Juga, J.: Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. J Anim Breed Genet **130**(1), 10–19 (2013)
12. Veroneze, R., Bastiaansen, J., Knol, E., Guimaraes, S., Silva, F., Harlizius, B., Lopes, M.S., Lopes, P.: Linkage disequilibrium patterns and persistence of phase in purebred and crossbred pig (*Sus scrofa*) populations. BMC Genet **15**(1), 126 (2014)
13. de Roos, A., Hayes, B., Spelman, R., Goddard, M.: Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics **179**(3), 1503–1512 (2008)
14. van den Berg, I., Boichard, D., Lund, M.: Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genet Sel Evol **48**(1), 83–83 (2016)
15. Raymond, B., Bouwman, A., Schrooten, C., Houwing-Duistermaat, J., Veerkamp, R.: Utility of whole-genome sequence data for across-breed genomic prediction. Genet Sel Evol **50**(1), 27 (2018)
16. Falconer, D., Mackay, T.: Introduction to Quantitative Genetics, 4th edn. Benjamin Cummings, ??? (1996)
17. Hess, M., Druet, T., Hess, A., Garrick, D.: Fixed-length haplotypes can improve genomic prediction accuracy in an admixed dairy cattle population. Genet Sel Evol **49**(1), 54 (2017)
18. Crum, T., Schnabel, R., Decker, J., Regitano, L., Taylor, J.: CRUMBLER: A tool for the prediction of ancestry in cattle. PLoS ONE **14**(8), 0221471 (2019)

19. Mogens, S., Guosheng, S., Luc, J., Bernt, G., Rasmus, F.: Genomic evaluation of cattle in a multi-breed context. Livest Sci **166**, 101–110 (2014)
20. Sørensen, M., Norberg, E., Pedersen, J., Christensen, L.: Invited review: Crossbreeding in dairy cattle: A Danish perspective. J Dairy Sci **91**(11), 4116–4128 (2008)
21. Thomasen, J., Sørensen, A., Su, G., Madsen, P., Lund, M., Guldbrandtsen, B.: The admixed population structure in Danish Jersey dairy cattle challenges accurate genomic predictions. J Anim Sci **91**(7), 3105–3112 (2013)
22. Sevillano, C.A., ten Napel, J., Guimaraes, S., Silva, F., Calus, M.: Effects of alleles in crossbred pigs estimated for genomic prediction depend on their breed-of-origin. BMC Genomics **19**(1), 740 (2018)
23. Yamamoto, E., Matsunaga, H., Onogi, A., Kajiya-Kanegae, H., Minamikawa, M., Suzuki, A., Shirasawa, K., Hirakawa, H., Nunome, T., Yamaguchi, H., Miyatake, K., Ohyama, A., Iwata, H., Fukuoka, H.: A simulation-based breeding design that uses whole-genome prediction in tomato. Sci Rep **6**(19454) (2016)
24. Li, X., Lund, M., Janss, L., Wang, C., Ding, X., Zhang, Q., Su, G.: The patterns of genomic variances and covariances across genome for milk production traits between Chinese and Nordic Holstein populations. BMC Genet **18**(26), 12 (2017)
25. Zhou, L., Lund, M., Wang, Y., Su, G.: Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. J Anim Breed Genet **131**(4), 249–257 (2014)
26. de Roos, A., Hayes, B., Goddard, M.: Reliability of genomic predictions across multiple populations. Genetics **183**, 1545–1553 (2009)
27. Ibánẽz-Escriche, N., Fernando, R., Toosi, A., Dekkers, J.: Genomic selection of purebreds for crossbred performance. Genet Sel Evol **41**(1), 12 (2009)
28. Stranden, I., Garrick, D.: Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci **92**(6), 2971–2975 (2009)
29. Zeng, J., Garrick, D., Dekkers, J., Fernando, R.: A nested mixture model for genomic prediction using whole-genome SNP genotypes. PLoS ONE **13**(3), 0194683 (2018)
30. Karaman, E., Lund, M., Su, G.: Multi-trait single-step genomic prediction accounting for heterogeneous (co)variances over the genome. Heredity **124**, 274–287 (2020)
31. Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J.: Extension of the Bayesian alphabet for genomic selection. BMC Bioinform **12**(1), 186 (2010)
32. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.: Julia: A fresh approach to numerical computing. SIAM Review **59**(1), 65–98 (2017)
33. Daetwyler, H., Pong-Wong, R., Villanueva, B., Woolliams, J.: The impact of genetic architecture on genome-wide evaluation methods. Genetics **185**(3), 1021–1031 (2010)
34. Cheng, H., Kizilkaya, K., Zeng, J., Garrick, D., Fernando, R.: Genomic prediction from multiple-trait Bayesian regression methods using mixture priors. Genetics **209**(1), 89–103 (2018)
35. Karaman, E., Lund, M., Anche, M., Janss, L., Su, G.: Genomic prediction using multi-trait weighted GBLUP accounting for heterogeneous variances and covariances across the genome. G3-Genes Genom Genet **8**(11), 3549–3558 (2018)
36. Steyn, Y., Lourenco, D., Misztal, I.: Genomic predictions in purebreds with a multibreed genomic relationship matrix. J Anim Sci **97**(11), 4418–4427 (2019)
37. Habier, D., Fernando, R., Dekkers, J.: The impact of genetic relationship information on genome-assisted breeding values. Genetics **177**, 2389–2397 (2007)
38. Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., Thaller, G.: The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol **42**(1), 5 (2010)
39. Wientjes, Y., Veerkamp, R., Calus, M.: The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction **193**(2), 621–631 (2013)
40. Wientjes, Y., Calus, M., Goddard, M., Hayes, B.: Impact of QTL properties on the accuracy of multi-breed genomic prediction. Genet Sel Evol **47**(42) (2015)
41. Lehermeier, C., Schön, C., de Los Campos, G.: Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. Genetics **201**(1), 323–337 (2015)
42. de los Campos, G., Sorensen, D., Gianola, D.: Genomic heritability: What is it? PLoS Genet **11**, 1005048 (2015)
43. Wang, L., Sørensen, P., Janss, L., Ostersen, T., Edwards, D.: Genome-wide and local pattern of linkage disequilibrium and persistence of phase for 3 Danish pig breeds. BMC Genet **14**(1), 115 (2013)
44. Brøndum, R.F., Su, G., Lund, M., Bowman, P., Goddard, M., Hayes, B.: Genome position specific priors for genomic prediction. BMC Genomics **13**(1), 543 (2012)
45. Gebreyesus, G., Lund, M., Buitenhuis, B., Bovenhuis, H., Poulsen, N., Janss, L.: Modeling heterogeneous (co)variances from adjacent-SNP groups improves genomic prediction for milk protein composition traits. Genet Sel Evol **49**(1), 89 (2017)
46. Goddard, M., Hayes, B.: Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nat Rev Genet **10**, 381–391 (2009)
47. Kemper, K., Goddard, M.: Understanding and predicting complex traits: knowledge from cattle. Hum Mol Genet **21**(1), 45–51 (2012)
48. Sørensen, L., Janss, L., Madsen, P., Mark, T., Lund, M.: Estimation of (co)variances for genomic regions of flexible sizes: application to complex infectious udder diseases in dairy cattle. Genet Sel Evol **44**(1), 18 (2012)
49. Gianola, D., de los Campos, G., Hill, W., Manfredi, E., Fernando, R.: Additive genetic variability and Bayesian alphabet. Genetics **183**(1), 347–363 (2009)
50. Liu, A., Lund, M., Boichard, D., Karaman, E., Guldbrandtsen, B., Fritz, S., Aamand, G., Nielsen, U., Sahana, G., Wang, Y., Su, G.: Weighted single-step genomic best linear unbiased prediction integrating variants selected from sequencing data by association and bioinformatics analyses. Genet Sel Evol **52**(1), 48 (2020)
51. Lund, M., de Roos, A., de Vries, A., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu,

Z., Reents, R., Schrooten, C., Seefried, F., Su, G.: A common reference population from four European Holstein populations increases reliability of genomic predictions. Genet Sel Evol **43**(1), 43 (2011)
52. van den Berg, I., Meuwissen, T., MacLeod, I., Goddard, M.: Predicting the effect of reference population on the accuracy of within, across and multibreed genomic prediction. J Dairy Sci **102**(4) (2019)
53. Chen, L., Li, C., Miller, S., Schenkel, F.: Multi-population genomic prediction using a multi-task Bayesian learning model. BMC Genet **15**, 53 (2014)
54. Esfandyari, H., Sørensen, A., Bijma, P.: A crossbred reference population can improve the response to genomic selection for crossbred performance. Genet Sel Evol **47**(1), 76 (2015)
55. Xiang, T., Christensen, O., Legarra, A.: Technical note: Genomic evaluation for crossbred performance in a single-step approach with metafounders. J Anim Sci **95**(4), 1472–1480 (2017)
56. Sevillano, C., Vandenplas, J., Bastiaansen, J., Bergsma, R., Calus, M.: Genomic evaluation for a three-way crossbreeding system considering breed-of-origin of alleles. Genet Sel Evol **49**(1), 75 (2017)
57. Gautason, E., Schönherz, A., Sahana, G., Guldbrandtsen, B.: Relationship of Icelandic cattle with Northern and Western European cattle breeds, admixture and population structure. Acta Agriculturae Scandinavica, Section A-Animal Science, 1–14 (2019)
58. Toosi, A., Fernando, R., Dekkers, J.: Genome-wide mapping of quantitative trait loci in admixed populations using mixed linear model and Bayesian multiple regression analysis. Genet Sel Evol **50**(1), 32 (2018)
59. Vandenplas, J., Calus, M., Sevillano, C., Windig, J., Bastiaansen, J.: Assigning breed origin to alleles in crossbred animals. Genet Sel Evol **48**(1), 61 (2016)
60. Sankararaman, S., Sridhar, S., Kimmel, G., Halperin, E.: Estimating local ancestry in admixed populations. Am J Hum Genet **82**(2), 290–303 (2008)
61. Connolly, S., Fortes, M., Piper, E., Seddon, J., Kelly, M.: Determining the Number of Animals Required to Accurately Determine Breed Composition Using Genomic Data (2014). 10th World Congress on Genetics Applied to Livestock Production
62. Daetwyler, H., Villanueva, B., Woolliams, J.: Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE **3**(10), 3395 (2008)
63. Wientjes, Y., Veerkamp, R., Bijma, P., Bovenhuis, H., Schrooten, C., MPL, C.: Empirical and deterministic accuracies of across-population genomic prediction. Genet Sel Evol **47**(5) (2015)
64. Pocrnic, I., Lourenco, D., Masuda, Y., Misztal, I.: Dimensionality of genomic information and performance of the algorithm for proven and young for different livestock species. Genet Sel Evol **48**(82) (2016)

Figures

Figure 1 Accuracies (horizontal axis) for high heritability trait ($h^2 = 0.4$) in QTL correlation scenario of 0.5, using different data sets or models (vertical axis). The predicted population was given on top of each plot. Letters in parenthesis stands for the significance tests.

Figure 2 Accuracies (horizontal axis) for low heritability trait ($h^2 = 0.05$) in QTL correlation scenario of 0.5, using different data sets or models (vertical axis). The predicted population was given on top of each plot. Letters in parenthesis stands for the significance tests.

Figure 3 Accuracies (horizontal axis) for high ($h^2 = 0.4$, left figure) and low ($h^2 = 0.05$, right figure) heritability trait using BOA model with correlated SNP effects and different region sizes (vertical axis), in QTL correlation scenario of 0.5. Letters in parenthesis stands for the significance tests.

Figure 4 Accuracies (horizontal axis) for low heritability trait ($h^2 = 0.05$) in QTL correlation scenario of 0.5, using different data sets or models (vertical axis), when only the QTLs are considered with region size of 1 SNP. The predicted population was given on top of each plot. Letters in parenthesis stands for the significance tests.

Tables

Table 1 Parents of each simulated generation

Generation/Population	HOL[1]	RED	JER	MIX
1[2]	$\mathrm{HOL}_0^{\mathrm{M}} \times \mathrm{HOL}_0^{\mathrm{F}}$	$\mathrm{RED}_0^{\mathrm{M}} \times \mathrm{RED}_0^{\mathrm{F}}$	$\mathrm{JER}_0^{\mathrm{M}} \times \mathrm{JER}_0^{\mathrm{F}}$	$\mathrm{JER}_0^{\mathrm{M}} \times \mathrm{HOL}_0^{\mathrm{F}}$
⋮	⋮	⋮	⋮	⋮
6	$\mathrm{HOL}_5^{\mathrm{M}} \times \mathrm{HOL}_5^{\mathrm{F}}$	$\mathrm{RED}_5^{\mathrm{M}} \times \mathrm{RED}_5^{\mathrm{F}}$	$\mathrm{JER}_5^{\mathrm{M}} \times \mathrm{JER}_5^{\mathrm{F}}$	$\mathrm{HOL}_5^{\mathrm{M}} \times \mathrm{MIX}_5^{\mathrm{F}}$
7	$\mathrm{HOL}_6^{\mathrm{M}} \times \mathrm{HOL}_6^{\mathrm{F}}$	$\mathrm{RED}_6^{\mathrm{M}} \times \mathrm{RED}_6^{\mathrm{F}}$	$\mathrm{JER}_6^{\mathrm{M}} \times \mathrm{JER}_6^{\mathrm{F}}$	$\mathrm{JER}_6^{\mathrm{M}} \times \mathrm{MIX}_6^{\mathrm{F}}$
8	$\mathrm{HOL}_7^{\mathrm{M}} \times \mathrm{HOL}_7^{\mathrm{F}}$	$\mathrm{RED}_7^{\mathrm{M}} \times \mathrm{RED}_7^{\mathrm{F}}$	$\mathrm{JER}_7^{\mathrm{M}} \times \mathrm{JER}_7^{\mathrm{F}}$	$\mathrm{RED}_7^{\mathrm{M}} \times \mathrm{MIX}_7^{\mathrm{F}}$
9	$\mathrm{HOL}_8^{\mathrm{M}} \times \mathrm{HOL}_8^{\mathrm{F}}$	$\mathrm{RED}_8^{\mathrm{M}} \times \mathrm{RED}_8^{\mathrm{F}}$	$\mathrm{JER}_8^{\mathrm{M}} \times \mathrm{JER}_8^{\mathrm{F}}$	$\mathrm{HOL}_8^{\mathrm{M}} \times \mathrm{MIX}_8^{\mathrm{F}}$

[1] HOL, RED, JER and MIX: Danish Holstein, Swedish Red, Danish Jersey and admixed population respectively.
[2] Subscripts denote the generation, and superscripts denote the sex, i.e., males (M) and females (F).

Table 2 Some descriptive statistics[1] on SNPs and QTL for each pure-breed population in the base population (Generation 0-G0)

	HOL[2]	RED	JER
Number of fixed QTL for reference(alternative) allele	9(0)	5(0)	58(1)
Number of fixed SNPs for reference(alternative) allele	564(2)	385(14)	2281(286)
Number of breed specific QTL	3	4	1
Number of breed specific SNPs	261	356	50
Average MAF of segregating QTL	0.17	0.16	0.16
Average MAF of segregating SNPs	0.23	0.23	0.22

[1] average over 10 replicates
[2] HOL, RED and JER: Danish Holstein, Swedish Red and Danish Jersey dairy cattle, respectively

Additional Files

Additional File 1-Result of principle component analysis (PCA)

The file includes plot of the first two principle components from the PCA analysis of genomic relationship matrices.

Additional File 2-Results of all scenarios

The file includes results of all scenarios including the QTL correlation scenario of 0.5, which was partially presented in the figures of main text.

Figure 1 Click here to access/download;Figure;Figure 1.png

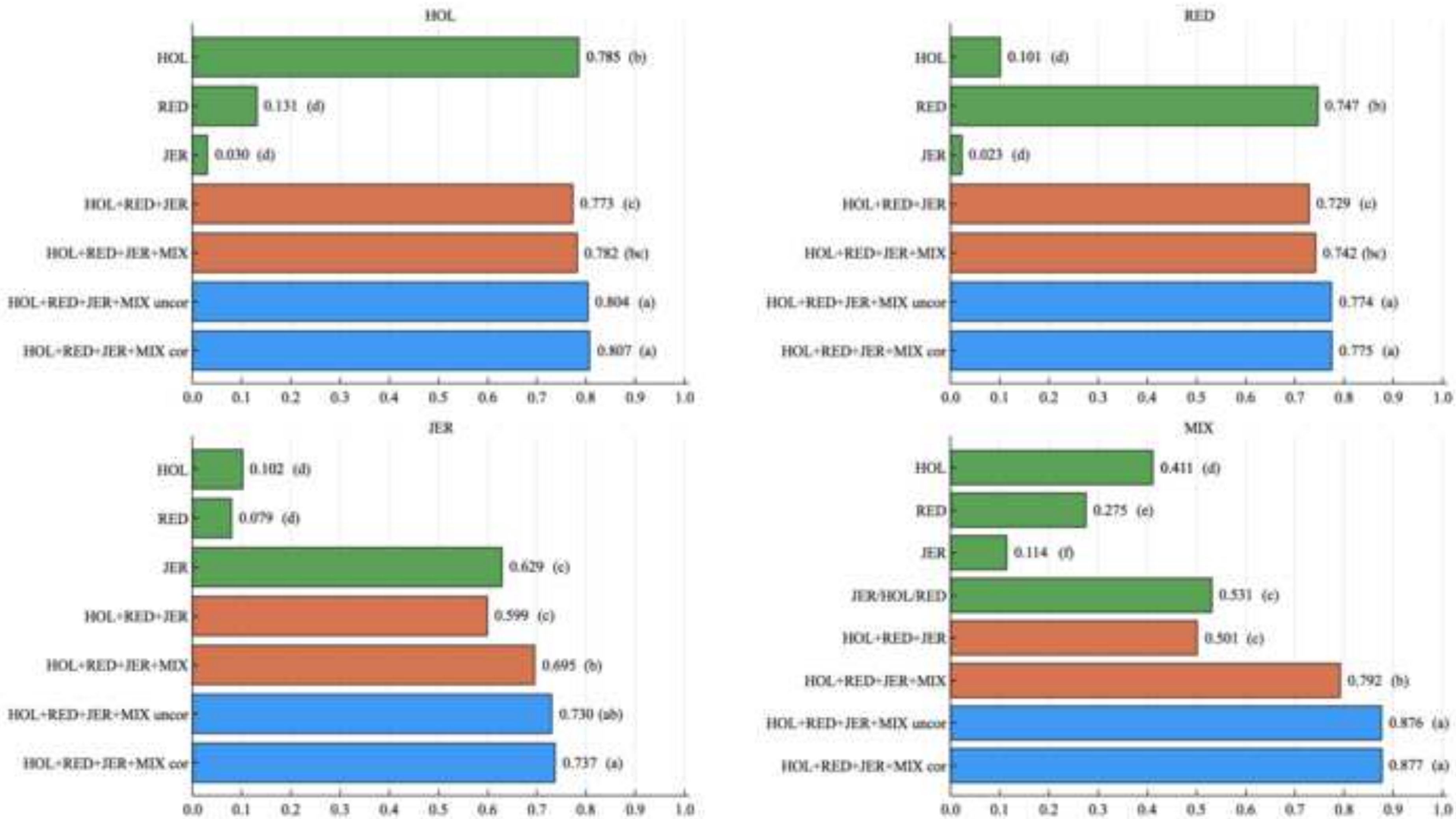


Figure 2

Click here to access/download;Figure;Figure 2.png

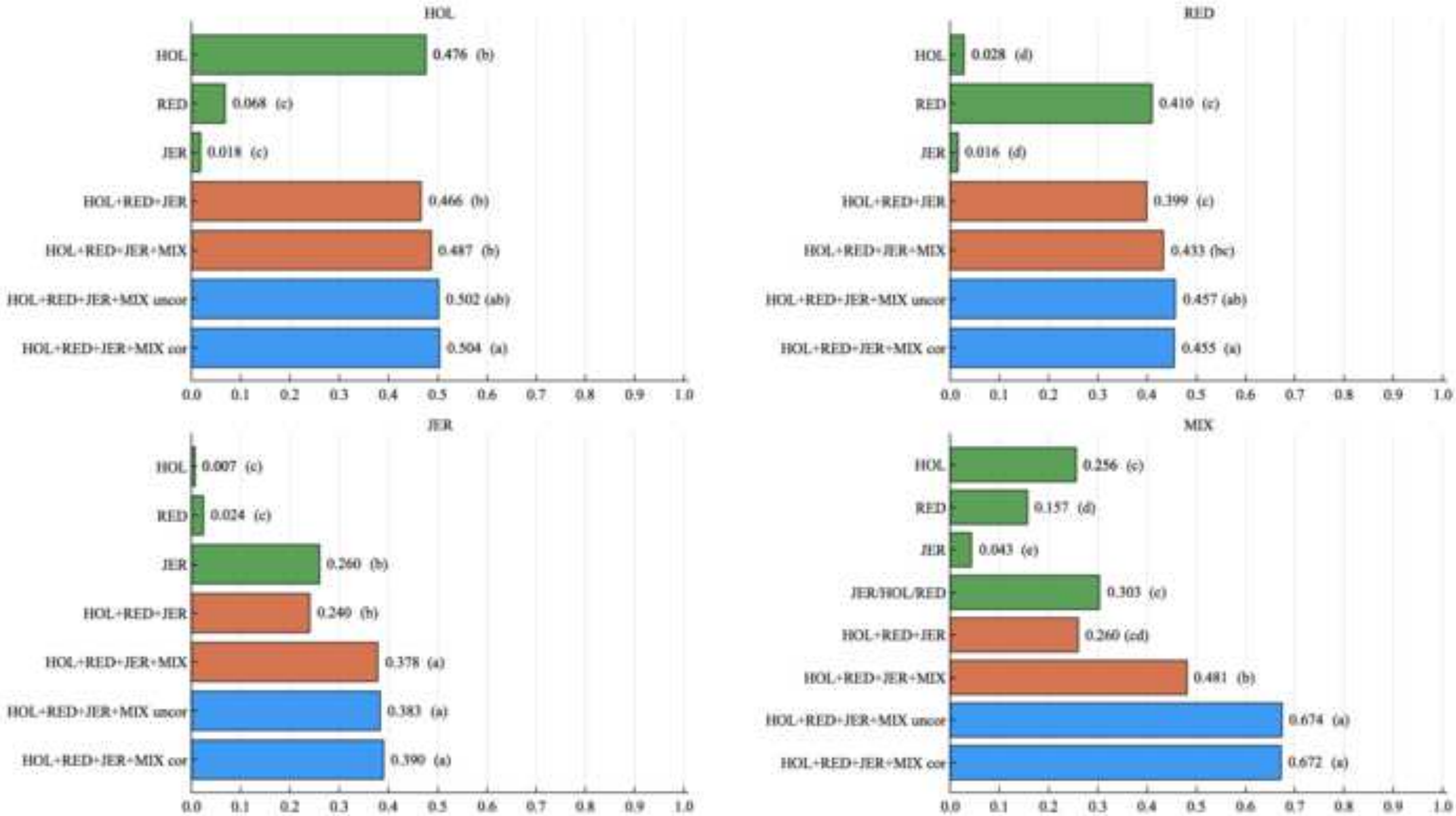


Figure 3

Click here to access/download;Figure;Figure 3.png

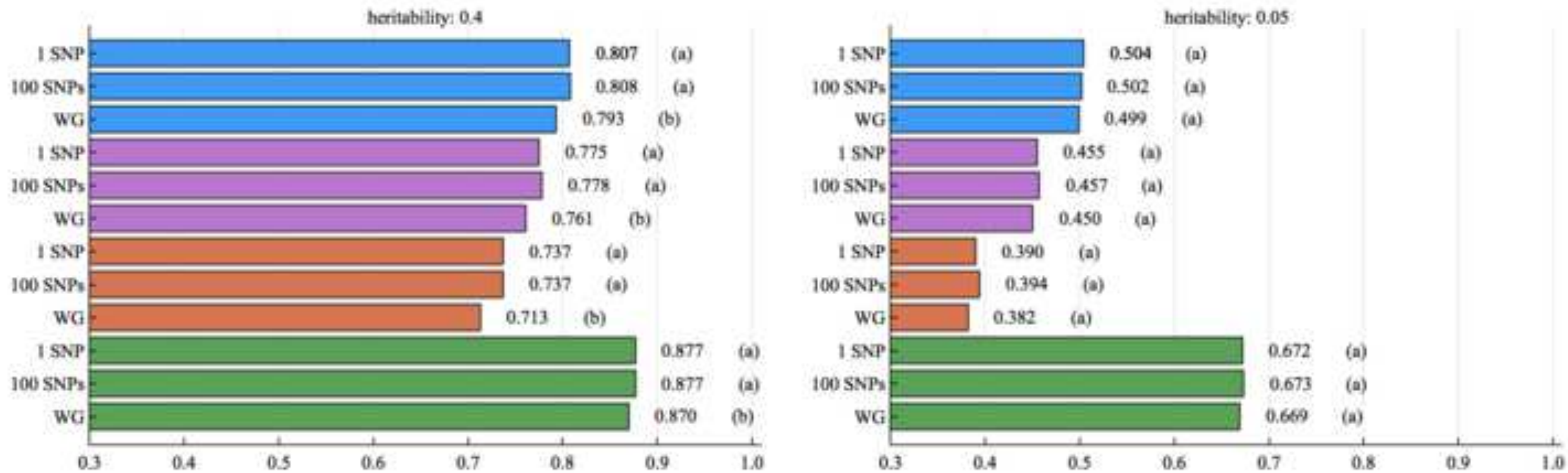
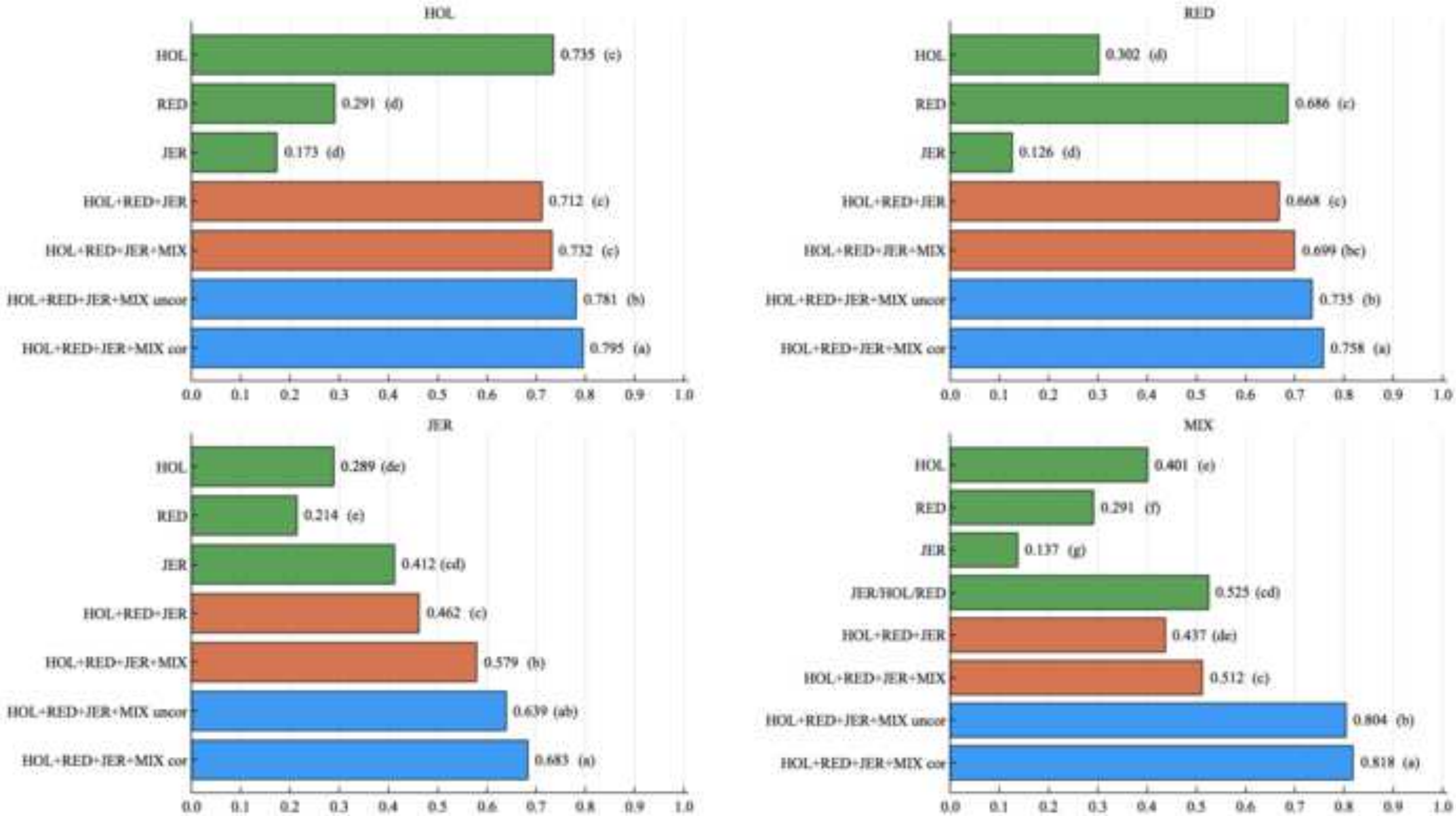
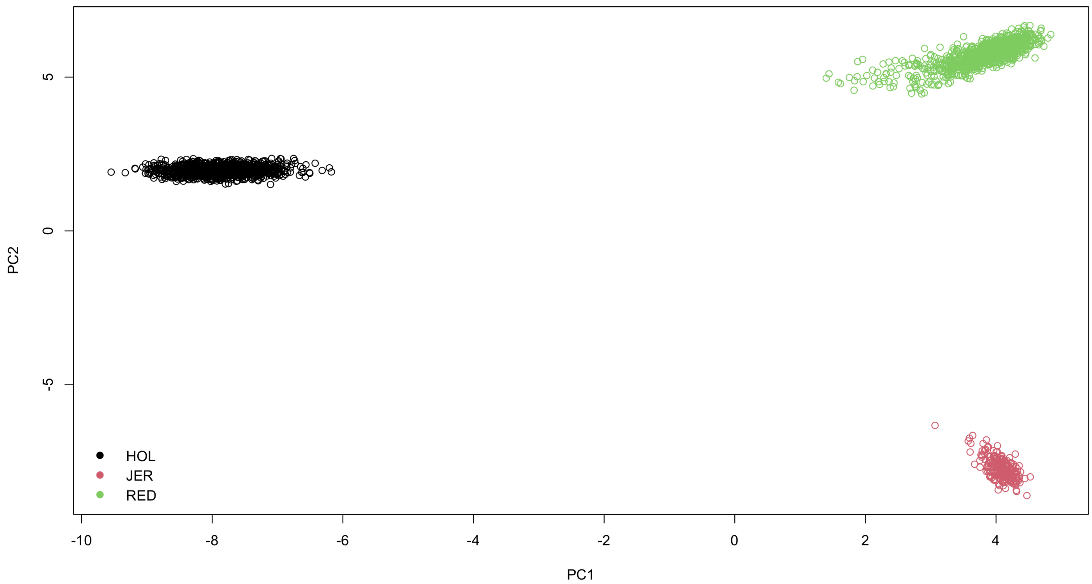


Figure 4

Click here to access/download;Figure;Figure 4.png





Supplementary Figure 1. Plot of the first two principle components from the PCA analysis of genomic relationship matrices. Genomic relationships were computed as described in [1], and analysis were carried out using R function prcomp() [2].

References

[1] Wientjes, Y., Veerkamp, R., Bijma, P., Bovenhuis, H., Schrooten, C., MPL, C.: Empirical and deterministic accuracies of across-population genomic prediction. Genet Sel Evol **47**(5) (2015)

[2] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2020). R Foundation for Statistical Computing. https://www.R-project.org/

Supplementary Table 1: Accuracies for purebred individuals for a trait with high heritability $(h^2 = 0.40)$

			HOL			**RED**			**JER**		
Correlation[1]	**Training[2]**	**Data/Region Size[3]**	**1 SNP**	**100 SNPs**	**WG**	**1 SNP**	**100 SNPs**	**WG**	**1 SNP**	**100 SNPs**	**WG**
1.00	Pure	HOL	${}_{b}0.771^{b}$	${}_{b}0.780^{a}$	${}_{b}0.760^{c}$	${}_{d}0.128^{a}$	${}_{d}0.095^{a}$	${}_{d}0.080^{b}$	${}_{d}0.120^{a}$	${}_{c}0.059^{b}$	${}_{c}0.036^{c}$
		RED	${}_{c}0.170^{a}$	${}_{c}0.162^{a}$	${}_{c}0.127^{b}$	${}_{c}0.751^{b}$	${}_{c}0.757^{a}$	${}_{bc}0.741^{c}$	${}_{d}0.170^{a}$	${}_{c}0.123^{b}$	${}_{c}0.111^{b}$
		JER	${}_{c}0.042^{a}$	${}_{d}0.035^{a}$	${}_{d}0.040^{a}$	${}_{d}0.060^{a}$	${}_{d}0.057^{a}$	${}_{d}0.057^{a}$	${}_{c}0.643^{b}$	${}_{b}0.652^{a}$	${}_{b}0.638^{c}$
	Combined	HOL+RED+JER	${}_{b}0.777^{a}$	${}_{b}0.779^{a}$	${}_{b}0.757^{b}$	${}_{c}0.749^{a}$	${}_{c}0.752^{a}$	${}_{c}0.730^{b}$	${}_{bc}0.652^{a}$	${}_{b}0.646^{a}$	${}_{b}0.617^{b}$
		HOL+RED+JER+MIX	${}_{a}0.802^{a}$	${}_{a}0.803^{a}$	${}_{a}0.781^{b}$	${}_{b}0.774^{a}$	${}_{b}0.777^{a}$	${}_{ab}0.754^{b}$	${}_{a}0.732^{a}$	${}_{a}0.734^{a}$	${}_{a}0.701^{b}$
	BOA	HOL+RED+JER+MIX no Cor	${}_{a}0.796^{b}$	${}_{a}0.804^{a}$	${}_{a}0.782^{c}$	${}_{ab}0.776^{b}$	${}_{ab}0.782^{a}$	${}_{a}0.763^{c}$	${}_{ab}0.723^{b}$	${}_{a}0.738^{a}$	${}_{a}0.713^{c}$
		HOL+RED+JER+MIX with Cor	${}_{a}0.803^{a}$	${}_{a}0.805^{a}$	${}_{a}0.783^{b}$	${}_{a}0.776^{a}$	${}_{a}0.779^{a}$	${}_{a}0.758^{b}$	${}_{a}0.734^{a}$	${}_{a}0.740^{a}$	${}_{a}0.703^{b}$
0.50	Pure	HOL	${}_{b}0.785^{b}$	${}_{b}0.790^{a}$	${}_{b}0.774^{c}$	${}_{d}0.101^{a}$	${}_{d}0.085^{ab}$	${}_{d}0.073^{b}$	${}_{d}0.102^{a}$	${}_{d}0.073^{b}$	${}_{e}0.067^{b}$
		RED	${}_{d}0.131^{ab}$	${}_{d}0.129^{a}$	${}_{d}0.116^{b}$	${}_{b}0.747^{b}$	${}_{b}0.754^{a}$	${}_{b}0.736^{c}$	${}_{d}0.079^{a}$	${}_{d}0.070^{a}$	${}_{e}0.043^{b}$
		JER	${}_{d}0.030^{a}$	${}_{d}0.039^{a}$	${}_{d}0.029^{a}$	${}_{d}0.023^{a}$	${}_{d}0.025^{a}$	${}_{d}0.024^{a}$	${}_{c}0.629^{ab}$	${}_{c}0.635^{a}$	${}_{cd}0.624^{b}$
	Combined	HOL+RED+JER	${}_{c}0.773^{a}$	${}_{c}0.774^{a}$	${}_{c}0.761^{b}$	${}_{c}0.729^{a}$	${}_{c}0.732^{a}$	${}_{c}0.716^{b}$	${}_{c}0.599^{a}$	${}_{c}0.587^{b}$	${}_{d}0.590^{ab}$
		HOL+RED+JER+MIX	${}_{bc}0.782^{a}$	${}_{bc}0.782^{a}$	${}_{bc}0.771^{b}$	${}_{bc}0.742^{a}$	${}_{bc}0.747^{a}$	${}_{bc}0.732^{b}$	${}_{b}0.695^{a}$	${}_{b}0.690^{a}$	${}_{bc}0.675^{b}$
	BOA	HOL+RED+JER+MIX no Cor	${}_{a}0.804^{a}$	${}_{a}0.807^{a}$	${}_{a}0.792^{b}$	${}_{a}0.774^{b}$	${}_{a}0.780^{a}$	${}_{a}0.762^{c}$	${}_{ab}0.730^{a}$	${}_{a}0.738^{a}$	${}_{ab}0.712^{b}$
		HOL+RED+JER+MIX with Cor	${}_{a}0.807^{a}$	${}_{a}0.808^{a}$	${}_{a}0.793^{b}$	${}_{a}0.775^{a}$	${}_{a}0.778^{a}$	${}_{a}0.761^{b}$	${}_{a}0.737^{a}$	${}_{a}0.737^{a}$	${}_{a}0.713^{b}$
0.25	Pure	HOL	${}_{b}0.782^{b}$	${}_{b}0.788^{a}$	${}_{b}0.772^{c}$	${}_{d}0.053^{a}$	$d0.041^{ab}$	${}_{d}0.034^{b}$	${}_{d}0.089^{a}$	${}_{d}0.066^{ab}$	${}_{d}0.058^{b}$
		RED	${}_{d}0.080^{a}$	${}_{d}0.081^{a}$	${}_{d}0.080^{a}$	${}_{b}0.742^{b}$	${}_{b}0.750^{a}$	${}_{b}0.730^{c}$	${}_{d}0.017^{a}$	${}_{d}0.012^{a}$	${}_{d}0.003^{a}$
		JER	${}_{d}0.026^{a}$	${}_{d}0.034^{a}$	${}_{d}0.025^{a}$	${}_{d}0.017^{a}$	${}_{d}0.014^{a}$	${}_{d}0.019^{a}$	${}_{bc}0.632^{a}$	${}_{b}0.639^{a}$	${}_{bc}0.627^{b}$
	Combined	HOL+RED+JER	${}_{c}0.764^{a}$	${}_{c}0.766^{a}$	${}_{c}0.755^{b}$	${}_{c}0.716^{b}$	${}_{c}0.721^{a}$	${}_{c}0.705^{c}$	${}_{c}0.587^{a}$	${}_{c}0.573^{b}$	${}_{c}0.581^{ab}$
		HOL+RED+JER+MIX	${}_{bc}0.769^{a}$	${}_{c}0.768^{ab}$	${}_{bc}0.761^{b}$	${}_{bc}0.725^{a}$	${}_{bc}0.731^{a}$	${}_{bc}0.718^{b}$	${}_{b}0.689^{a}$	${}_{b}0.680^{b}$	${}_{b}0.667^{c}$
	BOA	HOL+RED+JER+MIX no Cor	${}_{a}0.803^{a}$	${}_{a}0.805^{a}$	${}_{a}0.790^{b}$	${}_{a}0.771^{b}$	${}_{a}0.777^{a}$	${}_{a}0.759^{c}$	${}_{a}0.735^{a}$	${}_{a}0.743^{a}$	${}_{a}0.716^{b}$
		HOL+RED+JER+MIX with Cor	${}_{a}0.805^{a}$	${}_{a}0.806^{a}$	${}_{a}0.790^{b}$	${}_{a}0.772^{a}$	${}_{a}0.775^{a}$	${}_{a}0.758^{b}$	${}_{a}0.740^{a}$	${}_{a}0.741^{a}$	${}_{a}0.715^{b}$

[1] Correlation of simulated QTL effects. Different alphabets mean significantly different values at a Type 1 error rate of 0.05 with Bonferroni correction. Subscripts (within region size) and superscripts (within data) stand for comparisons within column and row, respectively, for each correlation scenario.
[2] The methods classified based on the data and model used to estimate SNP effects
[3] Data: Data included in reference population. Region Size: Number of SNPs assigned the same variance.

Supplementary Table 2: Accuracies for purebred individuals for a trait with low heritability ($h^2 = 0.05$)

Correlation[1]	Training[2]	Data/Region Size[3]	HOL			RED			JER		
			1 SNP	**100 SNPs**	**WG**	**1 SNP**	**100 SNPs**	**WG**	**1 SNP**	**100 SNPs**	**WG**
1.00	Pure	HOL	$_{c}0.469^{a}$	$_{c}0.470^{ab}$	$_{a}0.466^{b}$	$_{e}0.019^{a}$	$_{d}0.010^{b}$	$_{d}0.022^{ab}$	$_{c}$-0.023^{a}	$_{c}$-0.023^{a}	$_{b}$-0.021^{a}
		RED	$_{d}0.055^{a}$	$_{d}0.047^{a}$	$_{b}0.047^{a}$	$_{bd}0.421^{a}$	$_{bc}0.420^{ab}$	$_{bc}0.417^{b}$	$_{c}0.044^{a}$	$_{c}0.043^{a}$	$_{b}0.045^{a}$
		JER	$_{d}$-0.008^{a}	$_{d}$-0.002^{a}	$_{b}0.002^{a}$	$_{e}0.036^{a}$	$_{d}0.035^{a}$	$_{d}0.040^{a}$	$_{ab}0.319^{a}$	$_{ab}0.319^{a}$	$_{a}0.321^{a}$
	Combined	HOL+RED+JER	$_{bc}0.467^{a}$	$_{c}0.472^{a}$	$_{a}0.463^{a}$	$_{cd}0.415^{a}$	$_{c}0.411^{a}$	$_{c}0.408^{a}$	$_{b}0.291^{a}$	$_{b}0.294^{a}$	$_{a}0.289^{a}$
		HOL+RED+JER+MIX	$_{ab}0.501^{a}$	$_{b}0.503^{ab}$	$_{a}0.494^{b}$	$_{ab}0.453^{a}$	$_{ab}0.450^{ab}$	$_{ab}0.445^{b}$	$_{ab}0.389^{a}$	$_{a}0.385^{a}$	$_{a}0.381^{a}$
	BOA	HOL+RED+JER+MIX no Cor	$_{abc}0.500^{a}$	$_{abc}0.501^{a}$	$_{a}0.497^{a}$	$_{ac}0.451^{a}$	$_{a}0.451^{ab}$	$_{a}0.447^{b}$	$_{ab}0.387^{a}$	$_{ab}0.382^{a}$	$_{a}0.383^{a}$
		HOL+RED+JER+MIX with Cor	$_{a}0.503^{a}$	$_{a}0.506^{ab}$	$_{a}0.495^{b}$	$_{ab}0.454^{a}$	$_{ab}0.450^{ab}$	$_{ab}0.444^{b}$	$_{a}0.392^{a}$	$_{a}0.389^{a}$	$_{a}0.382^{a}$
0.50	Pure	HOL	$_{b}0.476^{a}$	$_{a}0.478^{a}$	$_{a}0.474^{a}$	$_{d}0.028^{a}$	$_{c}0.024^{a}$	$_{c}0.030^{a}$	$_{c}0.007^{a}$	$_{c}0.017^{a}$	$_{c}0.010^{a}$
		RED	$_{c}0.068^{a}$	$_{b}0.080^{a}$	$_{b}0.071^{a}$	$_{c}0.410^{a}$	$_{b}0.412^{a}$	$_{b}0.410^{a}$	$_{c}0.024^{a}$	$_{c}0.030^{a}$	$_{c}0.028^{a}$
		JER	$_{c}0.018^{a}$	$_{b}0.015^{a}$	$_{b}0.012^{a}$	$_{d}0.016^{ab}$	$_{c}0.013^{b}$	$_{c}0.026^{a}$	$_{b}0.260^{a}$	$_{b}0.260^{a}$	$_{b}0.264^{a}$
	Combined	HOL+RED+JER	$_{b}0.466^{a}$	$_{a}0.467^{a}$	$_{a}0.464^{a}$	$_{c}0.399^{a}$	$_{b}0.393^{a}$	$_{b}0.397^{a}$	$_{b}0.240^{a}$	$_{b}0.238^{a}$	$_{b}0.243^{a}$
		HOL+RED+JER+MIX	$_{b}0.487^{a}$	$_{a}0.483^{a}$	$_{a}0.482^{a}$	$_{bc}0.433^{a}$	$_{b}0.432^{a}$	$_{ab}0.432^{a}$	$_{a}0.378^{a}$	$_{a}0.374^{a}$	$_{a}0.372^{a}$
	BOA	HOL+RED+JER+MIX no Cor	$_{ab}0.502^{a}$	$_{a}0.498^{a}$	$_{a}0.499^{a}$	$_{ab}0.457^{a}$	$_{a}0.459^{a}$	$_{a}0.455^{a}$	$_{a}0.383^{a}$	$_{a}0.388^{a}$	$_{a}0.384^{a}$
		HOL+RED+JER+MIX with Cor	$_{a}0.504^{a}$	$_{a}0.502^{a}$	$_{a}0.499^{a}$	$_{a}0.455^{a}$	$_{a}0.457^{a}$	$_{a}0.450^{a}$	$_{a}0.390^{a}$	$_{a}0.394^{a}$	$_{a}0.382^{a}$
0.25	Pure	HOL	$_{ab}0.473^{a}$	$_{a}0.474^{a}$	$_{ab}0.471^{a}$	$_{c}0.012^{a}$	$_{d}0.007^{a}$	$_{e}0.014^{a}$	$_{c}0.005^{a}$	$_{c}0.016^{a}$	$_{c}0.007^{a}$
		RED	$_{c}0.050^{a}$	$_{b}0.059^{a}$	$_{c}0.054^{a}$	$_{b}0.407^{a}$	$_{b}0.409^{a}$	$_{c}0.407^{a}$	$_{c}$-0.006^{a}	$_{c}$-0.006^{a}	$_{c}0.000^{a}$
		JER	$_{c}0.017^{a}$	$_{b}0.014^{a}$	$_{c}0.012^{a}$	$_{c}0.024^{b}$	$_{d}0.017$b	$_{e}0.032^{a}$	$_{b}0.260^{a}$	$_{b}0.260^{a}$	$_{b}0.265^{a}$
	Combined	HOL+RED+JER	$_{b}0.459^{a}$	$_{a}0.458^{a}$	$_{b}0.458^{a}$	$_{b}0.392^{a}$	$_{c}0.385^{a}$	$_{d}0.391^{a}$	$_{b}0.228^{a}$	$_{b}0.226^{a}$	$_{b}0.233^{a}$
		HOL+RED+JER+MIX	$_{b}0.476^{a}$	$_{a}0.472^{a}$	$_{b}0.473^{a}$	$_{b}0.423^{a}$	$_{bc}0.417^{a}$	$_{bcd}0.424^{a}$	$_{a}0.372^{a}$	$_{a}0.364^{a}$	$_{a}0.368^{a}$
	BOA	HOL+RED+JER+MIX no Cor	$_{a}0.499^{a}$	$_{a}0.494^{a}$	$_{a}0.495^{a}$	$_{a}0.454^{b}$	$_{a}0.458^{a}$	$_{ab}0.453^{ab}$	$_{a}0.388^{a}$	$_{a}0.392^{a}$	$_{a}0.389^{a}$
		HOL+RED+JER+MIX with Cor	$_{a}0.501^{a}$	$_{a}0.497^{a}$	$_{ab}0.495^{a}$	$_{a}0.454^{ab}$	$_{a}0.458^{a}$	$_{a}0.451^{b}$	$_{a}0.393^{a}$	$_{a}0.397^{a}$	$_{a}0.384^{a}$

[1] Correlation of simulated QTL effects. Different alphabets mean significantly different values at a Type 1 error rate of 0.05 with Bonferroni correction. Subscripts (within region size) and superscripts (within data) stand for comparisons within column and row, respectively, for each correlation scenario.

[2] The methods classified based on the data and model used to estimate SNP effects

[3] Data: Data included in reference population. Region Size: Number of SNPs assigned the same variance.

Supplementary Table 3: Accuracies for admixed individuals for a trait with high heritability $\left(h^2 = 0.40\right)$

			MIX		
Correlation[1]	**Training[2]**	**Data/Region Size[3]**	**1 SNP**	**100 SNPs**	**WG**
1.00	Pure	HOL	${}_{e}0.578^{a}$	${}_{e}0.576^{a}$	${}_{e}0.562^{b}$
		RED	${}_{f}0.395^{a}$	${}_{f}0.395^{a}$	${}_{f}0.372^{b}$
		JER	${}_{g}0.178^{a}$	${}_{g}0.179^{a}$	${}_{g}0.175^{a}$
		HOL/RED/JER	${}_{d}0.687^{b}$	${}_{d}0.704^{a}$	${}_{d}0.685^{b}$
	Combined	HOL+RED+JER	${}_{c}0.726^{a}$	${}_{c}0.725^{a}$	${}_{c}0.704^{b}$
		HOL+RED+JER+MIX	${}_{a}0.788^{a}$	${}_{a}0.792^{a}$	${}_{ab}0.770^{b}$
	BOA	HOL+RED+JER+MIX no Cor	${}_{b}0.772^{b}$	${}_{b}0.782^{a}$	${}_{b}0.762^{c}$
		HOL+RED+JER+MIX with Cor	${}_{a}0.794^{b}$	${}_{a}0.799^{a}$	${}_{a}0.776^{c}$
0.50	Pure	HOL	${}_{d}0.411^{a}$	${}_{d}0.415^{a}$	${}_{d}0.410^{a}$
		RED	${}_{e}0.275^{a}$	${}_{e}0.274^{a}$	${}_{e}0.268^{a}$
		JER	${}_{f}0.114^{a}$	${}_{f}0.117^{a}$	${}_{f}0.114^{a}$
		HOL/RED/JER	${}_{c}0.531^{a}$	${}_{c}0.533^{ab}$	${}_{c}0.521^{b}$
	Combined	HOL+RED+JER	${}_{c}0.501^{a}$	${}_{c}0.498^{ab}$	${}_{c}0.493^{b}$
		HOL+RED+JER+MIX	${}_{b}0.792^{a}$	${}_{b}0.783^{a}$	${}_{b}0.774^{b}$
	BOA	HOL+RED+JER+MIX no Cor	${}_{a}0.876^{a}$	${}_{a}0.877^{a}$	${}_{a}0.870^{b}$
		HOL+RED+JER+MIX with Cor	${}_{a}0.877^{a}$	${}_{a}0.877^{a}$	${}_{a}0.870^{b}$
0.25	Pure	HOL	${}_{d}0.358^{a}$	${}_{d}0.363^{a}$	${}_{d}0.360^{a}$
		RED	${}_{e}0.231^{a}$	${}_{e}0.231^{a}$	${}_{e}0.231^{a}$
		JER	${}_{e}0.106^{a}$	${}_{e}0.109^{a}$	${}_{e}0.106^{a}$
		HOL/RED/JER	${}_{c}0.482^{a}$	${}_{c}0.482^{ab}$	${}_{c}0.469^{b}$
	Combined	HOL+RED+JER	${}_{c}0.434^{a}$	${}_{c}0.433^{ab}$	${}_{c}0.430^{b}$
		HOL+RED+JER+MIX	${}_{b}0.793^{ab}$	${}_{b}0.788^{a}$	${}_{b}0.777^{b}$
	BOA	HOL+RED+JER+MIX no Cor	${}_{a}0.897^{a}$	${}_{a}0.898^{a}$	${}_{a}0.892^{b}$
		HOL+RED+JER+MIX with Cor	${}_{a}0.897^{a}$	${}_{a}0.897^{a}$	${}_{a}0.892^{b}$

[1] Correlation of simulated QTL effects. Different alphabets mean significantly different values at a Type 1 error rate of 0.05 with Bonferroni correction. Subscripts (within region size) and superscripts (within data) stand for comparisons within column and row, respectively, for each correlation scenario.

[2] The methods classified based on the data and model used to estimate SNP effects

[3] Data: Data included in reference population. Region Size: Number of SNPs assigned the same variance.

Supplementary Table 4: Accuracies for admixed individuals for a trait with low heritability $\left(h^2 = 0.05\right)$

				MIX	
Correlation[1]	**Training**[2]	**Data/Region Size**[3]	**1 SNP**	**100 SNPs**	**WG**
1.00	Pure	HOL	${}_{c}0.363^{a}$	${}_{d}0.364^{a}$	${}_{d}0.362^{a}$
		RED	${}_{d}0.229^{a}$	${}_{e}0.225^{ab}$	${}_{e}0.221$b
		JER	${}_{e}0.075^{a}$	${}_{f}0.077^{a}$	${}_{f}0.078^{a}$
		HOL/RED/JER	${}_{b}0.433^{a}$	${}_{bc}0.436^{a}$	${}_{bc}0.429^{a}$
	Combined	HOL+RED+JER	${}_{b}0.436^{a}$	${}_{c}0.435^{a}$	${}_{c}0.429^{a}$
		HOL+RED+JER+MIX	${}_{a}0.491^{a}$	${}_{ab}0.488^{a}$	${}_{ab}0.484^{a}$
	BOA	HOL+RED+JER+MIX no Cor	${}_{bc}0.430^{ab}$	${}_{cd}0.434^{a}$	${}_{cd}0.424^{b}$
		HOL+RED+JER+MIX with Cor	${}_{a}0.500^{a}$	${}_{a}0.499^{a}$	${}_{a}0.493^{a}$
0.50	Pure	HOL	${}_{c}0.256^{a}$	${}_{c}0.256^{a}$	${}_{c}0.255^{a}$
		RED	${}_{d}0.157^{a}$	${}_{d}0.162^{a}$	${}_{de}0.155^{a}$
		JER	${}_{e}0.043^{a}$	${}_{e}0.043^{a}$	${}_{e}0.043^{a}$
		HOL/RED/JER	${}_{c}0.303^{ab}$	${}_{c}0.308^{a}$	${}_{c}0.299^{b}$
	Combined	HOL+RED+JER	${}_{cd}0.260^{a}$	${}_{c}0.260^{a}$	${}_{cd}0.258^{a}$
		HOL+RED+JER+MIX	${}_{b}0.481^{a}$	${}_{b}0.471$b	${}_{b}0.477^{ab}$
	BOA	HOL+RED+JER+MIX no Cor	${}_{a}0.674^{ab}$	${}_{a}0.676^{a}$	${}_{a}0.672^{b}$
		HOL+RED+JER+MIX with Cor	${}_{a}0.672^{a}$	${}_{a}0.673^{a}$	${}_{a}0.669^{a}$
0.25	Pure	HOL	${}_{cd}0.226^{a}$	${}_{cd}0.227^{a}$	${}_{cd}0.226^{a}$
		RED	${}_{de}0.136^{a}$	${}_{de}0.138^{a}$	${}_{de}0.134^{a}$
		JER	${}_{e}0.040^{a}$	${}_{e}0.041^{a}$	${}_{e}0.041^{a}$
		HOL/RED/JER	${}_{c}0.271^{ab}$	${}_{c}0.277^{a}$	${}_{c}0.266^{b}$
	Combined	HOL+RED+JER	${}_{cd}0.222^{a}$	${}_{cd}0.222^{a}$	${}_{cd}0.221^{a}$
		HOL+RED+JER+MIX	${}_{b}0.494^{a}$	${}_{b}0.485^{a}$	${}_{b}0.490^{a}$
	BOA	HOL+RED+JER+MIX no Cor	${}_{a}0.729^{a}$	${}_{a}0.730^{ab}$	${}_{a}0.727^{b}$
		HOL+RED+JER+MIX with Cor	${}_{a}0.728^{a}$	${}_{a}0.728^{a}$	${}_{a}0.725^{a}$

[1] Correlation of simulated QTL effects. Different alphabets mean significantly different values at a Type 1 error rate of 0.05 with Bonferroni correction. Subscripts (within region size) and superscripts (within data) stand for comparisons within column and row, respectively, for each correlation scenario.

[2] The methods classified based on the data and model used to estimate SNP effects

[3] Data: Data included in reference population. Region Size: Number of SNPs assigned the same variance.

Supplementary Table 5: Accuracy for low heritability $\left(h^2 = 0.05\right)$ trait, using only 250 QTL and region size of 1 SNP

			HOL	**RED**	**JER**	**MIX**
Correlation[1]	**Training**[2]	**Data/Region Size**[3]	**1 SNP**	**1 SNP**	**1 SNP**	**1 SNP**
0.50	Pure	HOL	$_{c}$0.735	$_{d}$0.302	$_{de}$0.289	$_{e}$0.401
		RED	$_{d}$0.291	$_{c}$0.686	$_{e}$0.214	$_{f}$0.291
		JER	$_{d}$0.173	$_{d}$0.126	$_{cd}$0.412	$_{g}$0.137
		HOL/RED/JER				$_{cd}$0.525
	Combined	HOL+RED+JER	$_{c}$0.712	$_{c}$0.668	$_{c}$0.462	$_{de}$0.437
		HOL+RED+JER+MIX	$_{c}$0.732	$_{bc}$0.699	$_{b}$0.579	$_{c}$0.512
	BOA	HOL+RED+JER+MIX no Cor	$_{b}$0.781	$_{b}$0.735	$_{ab}$0.639	$_{b}$0.804
		HOL+RED+JER+MIX with Cor	$_{a}$0.795	$_{a}$0.758	$_{a}$0.683	$_{a}$0.818

[1] Correlation of simulated QTL effects. Different alphabets mean significantly different values at a Type 1 error rate of 0.05 with Bonferroni correction. Subscripts (region size of 1 SNP) and superscripts (within data) stand for comparisons within column and row, respectively, for each correlation scenario.

[2] The methods classified based on the data and model used to estimate SNP effects

[3] Data: Data included in reference population.