Genomic prediction using individual-level data and summary statistics from multiple populations

Vandenplas, J., Calus, M. P. L., & Gorjanc, G.

Vandenplas, J., Calus, M. P. L., & Gorjanc, G. (2018). Genomic prediction using individual-level data and summary statistics from multiple populations. Genetics, 210(1), 53-69. DOI: 10.1534/genetics.118.301109

Genomic prediction using individual-level data and summary statistics from multiple populations

Jeremie Vandenplas*, Mario P.L. Calus*, Gregor Gorjanc†

* Wageningen University & Research, Animal Breeding and Genomics, 6700 AH Wageningen, The Netherlands

† The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Research Centre, Midlothian EH25 9RG, UK

Running title: Multi-population genomic prediction

Key words: meta-analysis, quantitative trait, statistical method

Author information:

Jeremie Vandenplas
Wageningen University & Research
Animal Breeding and Genomics
P.O. box 338, 6700 AH Wageningen, the Netherlands
E-mail: jeremie.vandenplas @wur.nl
Phone: +31 06 83642304

ABSTRACT

This study presents a method for genomic prediction that uses individual-level data and summary statistics from multiple populations. Genome-wide markers are nowadays widely used to predict complex traits, and genomic prediction using multi-population data is an appealing approach to achieve higher prediction accuracies. However, sharing of individual-level data across populations is not always possible. We present a method that enables integration of summary statistics from separate analyses with the available individual-level data. The data can either consist of individuals with single or multiple (weighted) phenotype records per individual. We developed a method based on a hypothetical joint analysis model and absorption of population specific information. We show that population specific information is fully captured by estimated allele substitution effects and the accuracy of those estimates, i.e. the summary statistics. The method gives identical result as the joint analysis of all individual-level data when complete summary statistics are available. We provide a series of easy-to-use approximations that can be used when complete summary statistics are not available or impractical to share. Simulations show that approximations enables integration of different sources of information across a wide range of settings yielding accurate predictions. The method can be readily extended to multiple-traits. In summary, the developed method enables integration of genome-wide data in the individual-level or summary statistics form from multiple populations to obtain more accurate estimates of allele substitution effects and genomic predictions.

INTRODUCTION

Genome-wide markers are nowadays widely used in animal and plant breeding to predict complex traits. This prediction is based on a linear model that partitions for each individual the observed complex phenotype value into systematic effects, comprising at least a population mean, an individual genetic value and an environmental deviation (Fisher, 1918). With genome-wide markers, individual genetic values can be computed from allele substitution effects estimated from individual-level phenotype and genotype data (Meuwissen et al., 2001). Subsequently, genetic values can be also computed for individuals of interest that are genotyped, but not phenotyped. This process is commonly called genomic prediction. In animal and plant breeding, genetic values are used to identify genetically superior individuals and use them as parents of the next generation to improve complex traits like milk yield (Meuwissen et al., 2001; VanRaden, 2008) or grain yield (Schulthess et al., 2016). In human genetics, genetic values can be used to predict individual genetic risk for complex diseases to inform preventive and personalized medicine (Campos et al., 2010; Wray et al., 2013; Pasaniuc and Price, 2017).

Accuracy of estimated allele substitution effects and of resulting genetic values for complex traits are foremost a function of the number of individuals with available phenotypes and genotypes (Daetwyler et al., 2008). To maximize the prediction accuracy, use of all available data is recommended (Henderson, 1984; Wray et al., 2013; Vilhjálmsson et al., 2015). In some small populations, collecting large amounts of data is not possible, and a joint analysis across multiple populations is needed to achieve high accuracy (Hozé et al., 2014; Wientjes et al., 2016). However, such joint analysis is often impossible, because of logistic or privacy considerations (Powell and Norman, 1998; Maier et al., 2018). Therefore, several methods were proposed to enable analysis of data from multiple populations when individual-level data is not available (Pasaniuc and Price, 2017; Liu and Goddard, 2018; Maier et al., 2018). These methods, often called meta-analyses (Pasaniuc and Price, 2017), approximate a joint analysis by first obtaining summary statistics from separate analyses of individual-level data for each population and then combine these summary statistics to estimate genetic values. In human genetics, summary statistics usually consist of publically available allele substitution effects, i.e., genome-wide associations, together with their standard errors, estimated independently for each marker (Yang et al., 2012; Vilhjálmsson et al., 2015; Maier et al., 2018). In livestock, summary statistics more likely consist of allele substitution effects estimated jointly for all markers, together with prediction error (co)variances (Liu and Goddard, 2018). While these methods may increase prediction accuracy in comparison to separate analyses, a loss in prediction accuracy is expected relative to an analysis using all individual-level data due to approximations (Maier et al., 2018). Further, these methods are based on some assumptions that make them difficult to apply outside their context of development. For example, Maier et al. (2018) implicitly assumed that only a single phenotype record per trait was associated with an individual. While this is usually the case in human genetics, it is not in breeding populations where individuals may have repeated phenotype records for the same trait, e.g., repeated longitudinal production or reproduction records in livestock or replicated field trials in crops, or when phenotype records are measured on a group of individuals and linked to a genotyped relative, e.g., progeny tested bulls for dairy production. Also, these developed methods do not allow combining individual-level data from some and summary statistics from other populations in one analysis (Liu and Goddard, 2018; Maier et al., 2018).

The objective of this study was to develop a method that jointly analyses individual-level data and summary statistics from multiple populations with no or limited amount of approximation. The method assumes that individual-level data is composed of marker genotypes and phenotype records that potentially have a variable number of replicates per individual. Further, summary statistics are assumed to be composed of estimated allele substitution effects with an associated measure of accuracy. Different measures of accuracy can be used, which controls the amount of approximation. The developed method is validated with simulated data. The results show that the method enables accurate integration of different sources of information across a wide range of settings.

MATERIAL AND METHODS

The first part of this section describes the theory of (1) separate and joint analyses of two individual-level datasets, (2) an exact integration of estimated allele substitution effects from one population into the analysis of another, (3) approximate integrations, and (4) generalization for multiple populations. The second part describes simulations used for validation of the developed method.

Theory

Assume we have two populations with independent individual-level datasets of phenotyped and genotyped individuals. The two populations and their corresponding datasets are hereafter referred to as 1 and 2. Further assume that both datasets contain the same markers. From this data we want to obtain accurate estimates of allele substitution effects and genetic values for complex traits. We can achieve this by a joint analysis of the two datasets. When one of the datasets is not available, we can achieve this by integrating the results of a separate analysis of the unavailable data into the separate analysis of the available dataset. We show how to perform this integration exactly or approximately.

Separate and joint analyses

A standard marker model, using random regression on marker genotypes, for the separate analysis of dataset i ($i = 1, 2$) is:

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_i^* + \mathbf{Z}_i \mathbf{W}_i \boldsymbol{\alpha}_i^* + \mathbf{e}_i^*, \tag{1}$$

where $\mathbf{y}_i$ is a $n_{obs,i} \times 1$ vector of phenotypes, $\boldsymbol{\beta}_i^*$ is a $n_{f,i} \times 1$ vector of fixed effects that are linked to $\mathbf{y}_i$ by a $n_{obs,i} \times n_{f,i}$ incidence matrix $\mathbf{X}_i$, $\boldsymbol{\alpha}_i^*$ is a $n_{mar} \times 1$ vector of allele substitution effects that are linked to $\mathbf{y}_i$ by a $n_{obs,i} \times n_{ind,i}$ incidence matrix $\mathbf{Z}_i$ and a $n_{ind,i} \times$

n_{mar} matrix of genotypes $\mathbf{W}_i$, and $\mathbf{e}_i^*$ is the vector $n_{obs,i} \times 1$ of residuals. In this work we consider single-nucleotide polymorphism markers, which we code in $\mathbf{W}_i$ as 0 for homozygous aa, 1 for heterozygous aA or Aa, and 2 for homozygous AA. Other genotype coding and centering, that is of the form $(\mathbf{W}_i - \mathbf{1}\mathbf{v}_i')$ with $\mathbf{1}$ being a $n_{ind,i} \times 1$ vector of ones and $\mathbf{v}_i$ being a $n_{mar} \times 1$ vector, can be used with no difference in obtained estimates of allele substitution effects (Strandén and Christensen, 2011). We assume a prior multivariate normal (MVN) distribution for allele substitution effects for the separate analyis of the dataset i, $\boldsymbol{\alpha}_i^*$, with mean zero and covariance $\mathbf{B}_i\ \sigma_{\alpha_i}^2$, $\boldsymbol{\alpha}_i^* \sim MVN(\mathbf{0}, \mathbf{B}_i\ \sigma_{\alpha_i}^2)$, where $\mathbf{B}_i$ is a $n_{mar} \times n_{mar}$ diagonal matrix (e.g., an identity matrix **I**), and $\sigma_{\alpha_i}^2$ is the variance of allele substitution effects. We also assume that residuals are multivariate normally distributed with mean zero and covariance $\mathbf{R}_i \sigma_e^2$, $\mathbf{e}_i^* \sim MVN(\mathbf{0}, \mathbf{R}_i\ \sigma_e^2)$, where $\mathbf{R}_i$ is a $n_{obs,i} \times n_{obs,i}$ diagonal matrix (e.g., an identity matrix **I**), and σ_e^2 is the residual variance. For simplicity and without loss of generality, it is assumed in the following that residual variances are the same for all separate and joint analyses. Variance components $\sigma_{\alpha_i}^2$ and σ_e^2 are assumed known, as they will have been estimated from the data previously. This marker model is the ridge regression model (Hoerl and Kennard, 1976; Whittaker et al., 2000; Meuwissen et al., 2001; de los Campos et al., 2012) with optional different weights in $\mathbf{B}_i$ (to differentially shrink different loci) and $\mathbf{R}_i$ (to account for heterogeneous residual variance due to variable number of repeated phenotype records per individual).

Separate estimates of allele substitution effects $\widehat{\boldsymbol{\alpha}_i^*}$ are obtained by solving the following system of equations:

$$\begin{bmatrix} \mathbf{X}_i'\mathbf{R}_i^{-1}\sigma_e^{-2}\mathbf{X}_i & \mathbf{X}_i'\mathbf{R}_i^{-1}\sigma_e^{-2}\mathbf{Z}_i\ \mathbf{W}_i \\ \mathbf{W}_i'\mathbf{Z}_i'\mathbf{R}_i^{-1}\sigma_e^{-2}\mathbf{X}_i & \mathbf{W}_i'\mathbf{Z}_i'\mathbf{R}_i^{-1}\sigma_e^{-2}\mathbf{Z}_i\ \mathbf{W}_i\ + \mathbf{B}_i^{-1}\sigma_{\alpha_i}^{-2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}_i^*} \\ \widehat{\boldsymbol{\alpha}_i^*} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_i'\mathbf{R}_i^{-1}\sigma_e^{-2}\mathbf{y}_i \\ \mathbf{W}_i'\mathbf{Z}_i'\mathbf{R}_i^{-1}\sigma_e^{-2}\mathbf{y}_i \end{bmatrix}. \quad (2)$$

Separate estimates of genetic values for individuals in a dataset i (i = 1, 2) are obtained by $\widehat{\mathbf{g}_i^*} = \mathbf{W}_i \widehat{\boldsymbol{\alpha}_i^*}$.

A marker model for the joint analysis of two datasets 1 and 2 is:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_1\ \mathbf{W}_1 \\ \mathbf{Z}_2\ \mathbf{W}_2 \end{bmatrix} \boldsymbol{\alpha} + \begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix}, \qquad (3)$$

where phenotypes from the two populations are modelled with populations specific fixed effects $(\boldsymbol{\beta}_1\ , \boldsymbol{\beta}_2\ )$, but a joint set of allele substitution effects $(\boldsymbol{\alpha})$. We assume a multivariate normal prior distribution for allele substitution effects with mean zero and covariance $\mathbf{B}_J\ \sigma^2_{\alpha_J}$, $\boldsymbol{\alpha} \sim MVN\left(\mathbf{0}, \mathbf{B}_J\ \sigma^2_{\alpha_J}\right)$, where $\mathbf{B}_J$ is a $n_{mar} \times n_{mar}$ diagonal matrix, and $\sigma^2_{\alpha_J}$ is the variance of allele substitution effects in the joint analysis. We also assume that residuals are multivariate normally distributed, specifically $\begin{bmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{R}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_2 \end{bmatrix} \sigma^2_e\right)$ where $\mathbf{R}_i$ is a $n_{obs,i} \times n_{obs,i}$ diagonal matrix.

Joint estimates of allele substitution effects $\hat{\boldsymbol{\alpha}}$ are obtained by solving the following system of equations:

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{X}_1 & \mathbf{0} & \mathbf{X}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1 \\ \mathbf{0} & \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{X}_2 & \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{Z}_2\ \mathbf{W}_2 \\ \mathbf{W}_1'\mathbf{Z}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{X}_1 & \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{X}_2 & \mathbf{W}_1'\mathbf{Z}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1\ + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{Z}_2\ \mathbf{W}_2\ + \mathbf{B}_J^{-1}\sigma_{\alpha_J}^{-2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}_1} \\ \widehat{\boldsymbol{\beta}_2} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} =$$

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{y}_1 \\ \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{y}_2 \\ \mathbf{W}_1'\mathbf{Z}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{y}_1\ + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{y}_2 \end{bmatrix} \qquad (4).$$

Joint estimates of genetic values for individuals in a dataset i (i = 1, 2) are obtained by $\widehat{\mathbf{g}_i} = \mathbf{W}_i \widehat{\boldsymbol{\alpha}}$.

Exact integration

The integration of estimates of allele substitution effects from one dataset into the analysis of another can be performed by means of absorbing corresponding equations in the joint system of equations. We choose to integrate estimates from the dataset 1 into the analysis of dataset 2. Derivations in Appendix A1 lead to the following system of equations that performs such integration and gives equivalent estimates of allele substitution effects to the joint analysis (4):

$$\begin{bmatrix} \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{X}_2 & \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{Z}_2\ \mathbf{W}_2 \\ \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{X}_2 & \left(PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1} + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{Z}_2\ \mathbf{W}_2\ - \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2} + \mathbf{B}_J^{-1}\sigma_{\alpha_J}^{-2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}_2} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} =$$

$$\begin{bmatrix} \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{y}_2 \\ \left(PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1}\widehat{\boldsymbol{\alpha}_1^*} + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{y}_2 \end{bmatrix}, \qquad (5)$$

where $\widehat{\boldsymbol{\alpha}_1^*}$ are estimates of allele substitution effects from the separate analysis of dataset 1 using (2), and $\left(PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1}$ is the inverse of the corresponding prediction error covariance (PEC) matrix. The latter can be obtained as $\left(PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1} = \mathbf{W}_1'\mathbf{Z}_1'\mathbf{M}_1\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1\ + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}$ with $\mathbf{M}_1 = \left(\mathbf{R}_1^{-1} - \mathbf{R}_1^{-1}\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{R}_1^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{R}_1^{-1}\right)$. Note that only the individual-level dataset 2 and summary statistics from the dataset 1 (i.e., the estimated allele substitution effects and their PEC) are required. Individual-level dataset 1 is therefore not required.

It is worth noting that the integration of estimates of allele substitution effects from the dataset 1 into the analysis of dataset 2 can also be obtained from a Bayesian context. Bayes estimators for linear mixed models were discussed by several authors (Lindley and Smith, 1972; Dempfle, 1977; Gianola and Fernando, 1986). In a Bayesian context, we can assume the following prior multivariate normal distributions for the marker model (1) applied to dataset 2:

$[\boldsymbol{\beta}_2^*\ |\mathbf{b}_2, \mathbf{U}_2] \sim MVN(\mathbf{b}_2, \mathbf{U}_2)$, where $\mathbf{b}_2$ is a mean vector and $\mathbf{U}_2$ is a (co)variance matrix,

$$\left[\boldsymbol{\alpha}_2^*\middle|\mathbf{B}_2\sigma^2_{\alpha_2}\right]\sim MVN\left(\mathbf{0},\mathbf{B}_2\sigma^2_{\alpha_2}\right), \text{ and}$$

$$[\mathbf{e}_2^*|\mathbf{R}_2\sigma^2_e]\sim MVN(\mathbf{0},\mathbf{R}_2\sigma^2_e).$$

Assuming a noninformative prior for $\boldsymbol{\beta}_2^*$, the system of equations (2) for dataset 2 can be obtained by differentiating the joint posterior distribution of $\boldsymbol{\beta}_2^*$ and $\boldsymbol{\alpha}_2^*$ with respect to $\boldsymbol{\beta}_2^*$ and $\boldsymbol{\alpha}_2^*$, and setting the derivatives equal to 0 (Gianola and Fernando, 1986). Integration of estimates of allele substitution effects from dataset 1 into the analysis of dataset 2 can be therefore obained by defining a multivariate normal prior distribution for allele substitution effects in the analysis of dataset 2 using the posterior distribution for allele substitution effects from a separate analysis of dataset 1:

$$\left[\boldsymbol{\alpha}\middle|\widehat{\boldsymbol{\alpha}_1^*}, PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right), \mathbf{B}_1\ \sigma^2_{\alpha_1}, \mathbf{B}_J\ \sigma^2_{\alpha_J}\right]\sim MVN\left(\mathbf{Q}\left(PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1}\widehat{\boldsymbol{\alpha}_1^*}, \mathbf{Q}\right), \qquad (6)$$

$$\mathbf{Q}=\left(\left(PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1}-\mathbf{B}_1^{-1}\sigma^{-2}_{\alpha_1}+\mathbf{B}_J^{-1}\sigma^{-2}_{\alpha_J}\right)^{-1}.$$

The matrix $\mathbf{Q}$ can be considered as the PEC matrix of a hypothetical separate analysis of dataset 1 using the multivariate normal prior distribution for allele substitution effects of the joint analysis, that is $\boldsymbol{\alpha}_1^*\sim MVN\left(\mathbf{0},\mathbf{B}_J\ \sigma^2_{\alpha_J}\right)$ and $\mathbf{Q}=\left(\mathbf{W}_1'\mathbf{Z}_1'\mathbf{M}_1\sigma_e^{-2}\mathbf{Z}_1\mathbf{W}_1+\mathbf{B}_J^{-1}\sigma^{-2}_{\alpha_J}\right)^{-1}$, and the vector $\mathbf{Q}\left(PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1}\widehat{\boldsymbol{\alpha}_1^*}$ can be considered as the estimated allele subsitution effects of this hypothetical separate analysis. In animal breeding, a similar approach was used to integrate estimated genetic values and associated accuracies from one genetic evaluation into another genetic evaluation (Quaas and Zhang, 2006; Legarra et al., 2007; Vandenplas and Gengler, 2012).

Finally, it is worth noting that the term $\left(PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1}\widehat{\boldsymbol{\alpha}_1^*}$ can be interpreted as a vector of hypothetical or pseudo-phenotype records associated with allele substitution effects and as such summarize available information in dataset 1. In this sense, the system (5) is similar to approaches that compute pseudo-records associated with individuals, from available estimated genetic values where individual-level phenotypic information is not readily available, or is not measured on the individuals themselves but on close relatives. In animal breeding, these approaches are commonly known as deregression of estimated genetic values (Jairath et al., 1998).

Approximate integration

Exact integration requires the inverse of prediction error covariance matrix from the separate analysis, which could be approximated when unavailable. Genomic analyses of complex traits that combine different datasets commonly have access to estimated allele substitution effects and associated prediction error variances (in different forms), but not the whole prediction error covariance matrix $PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)$ required in (5). We propose several ways to accommodate this situation. We assume that we know, at least, the prediction error variances (PEV) of estimated allele substitution effects $\left(PEV\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)$, the number of individuals $\left(n_{ind,1}\right)$ and variance components used in the separate analysis of dataset 1 ($\sigma_{\alpha_1}^2$ and σ_e^2).

When only the prediction error variances of the estimated allele substitution effects $\left(PEV\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)$ are known, while PEC are not, then we can approximate $\left(PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1}$ with $\left(PEV\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1}$. This approximation would be accurate if the matrix product $\mathbf{W}_1'\mathbf{W}_1$ has (close to) zero off-diagonal elements, which is dependent on the characteristics of genotypes in dataset 1 (e.g., allele frequencies, linkage disequilibrium (LD), and population/family structure). If this is not the case, the approximation will bias the analysis by ignoring off-diagonal elements.

When allele frequencies and LD correlations in dataset 1 are known, we can obtain a good approximation of $PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)$ under some conditions (one phenotype record per individual, homogenous residual variance, overall mean is the only fixed effect, and Hardy-Weinberg equilibrium). Derivations in Appendix A2 show that under these conditions we can approximate $PEC(\widehat{\boldsymbol{\alpha}_1^*})$ with $\left(\mathbf{W}_1'\mathbf{W}_1\ \sigma_e^{-2} + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)^{-1}$ with the unknown matrix $\mathbf{W}_1'\mathbf{W}_1$ approximated from commonly available population parameters (i.e., allele frequencies and LD correlation) as $4n_{ind,1}\mathbf{pp}' + \mathbf{V}^{\frac{1}{2}}\mathbf{C}\mathbf{V}^{\frac{1}{2}}$, where $\mathbf{p}$ is a $n_{mar} \times 1$ vector of allele frequencies, $\mathbf{V}$ is a $n_{mar} \times n_{mar}$ diagonal matrix of expected genotype sum of squares with the i-th diagonal element equal to $n_{ind,1}2p_{i,1}(1 - p_{i,1})$, and $\mathbf{C}$ is a $n_{mar} \times n_{mar}$ matrix of pairwise genotype correlations between markers. In practice, the matrix $\mathbf{C}$ for dataset 1 could be unknown, but we can approximate it by using a reference panel that includes, for example, available genotypes of non-phenotyped individuals originating from this population (Yang et al., 2012; Vilhjálmsson et al., 2015; Maier et al., 2018).

Finally, we relax the assumption of having a single phenotype record per individual in the preceding approximations. This is relevant when individuals have repeated phenotype records, e.g., repeated longitudinal production or reproduction records in livestock or replicated field trials in crops. A related issue is the violation of assumption of homogenous residual variance when phenotype records are first pre-processed and then used in genomic analyses, e.g., deregressed progeny proofs in livestock (e.g., Garrick et al., 2009) or adjusted field trial means in crops (e.g., Schulz-Streeck et al., 2013; Oakey et al., 2016; Damesa et al., 2017). For these situations, we show in Appendix A3 that we can approximate $PEC(\widehat{\boldsymbol{\alpha}_1^*})$ with $\left(\boldsymbol{\Lambda}_1\left(4\mathbf{pp}' + \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{C}\boldsymbol{\Psi}^{\frac{1}{2}}\right)\boldsymbol{\Lambda}_1\sigma_e^{-2} + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)^{-1}$ where $\boldsymbol{\Psi}$ is a $n_{mar} \times n_{mar}$ diagonal matrix with the j-th diagonal element equal to $2p_{j,1}(1 - p_{j,1})$, and $\boldsymbol{\Lambda}_1$ is a $n_{mar} \times n_{mar}$ diagonal matrix with the j-th diagonal element representing the square root of effective number of records for the j-th marker. The matrix $\boldsymbol{\Lambda}_1$ can be obtained by solving the nonlinear system of equations

$$diag\left(\left(\boldsymbol{\Lambda}_1\left(4\mathbf{pp}' + \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{C}\boldsymbol{\Psi}^{\frac{1}{2}}\right)\boldsymbol{\Lambda}_1\sigma_e^{-2} + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)^{-1}\right) = PEV\left(\widehat{\boldsymbol{\alpha}_1^*}\right)$$

through a fixed-point iteration algorithm (Burden and Faires, 2010) detailed in Appendix A3. It is worth noting that the proposed algorithm requires the inversion of a $n_{mar} \times n_{mar}$ dense matrix at each iteration. This computational cost can be reduced by performing the algorithm for each chromosome separately.

Integration with multiple populations

When more than two populations or datasets are available the developed methods can be easily extended. With n datasets, the prior distribution for allele substitution effects in the separate analysis of the n-th dataset is defined using the posterior distributions for allele substitution effects from the separate analyses of $n - 1$ datasets:

$$\left[\boldsymbol{\alpha}|\widehat{\boldsymbol{\alpha}_1^*}, \widehat{\boldsymbol{\alpha}_2^*}, \dots, \widehat{\boldsymbol{\alpha}_{n-1}^*}\right] \sim MVN\left(\mathbf{Q}\sum_{i=1}^{n-1}\left(\left(PEC(\widehat{\boldsymbol{\alpha}_i^*})\right)^{-1}\widehat{\boldsymbol{\alpha}_i^*}\right), \mathbf{Q}\right),$$

$$\mathbf{Q} = \left(\mathbf{B}_J^{-1}\sigma_{\alpha_J}^{-2} + \sum_{i=1}^{n-1}\left(\left(PEC(\widehat{\boldsymbol{\alpha}_i^*})\right)^{-1} - \mathbf{B}_i^{-1}\sigma_{\alpha_i}^{-2}\right)\right)^{-1}.$$

Simulations

We tested developed methods with simulated data that either had low or high genetic diversity. The data was simulated in 5 replicates with the AlphaSim program, which uses the coalescent method for simulation of base population chromosomes and the gene drop method for simulation of chromosome inheritance within a pedigree (Hickey and Gorjanc, 2012; Faux et al., 2016).

A diploid genome was simulated with 30 chromosomes, each 10^8 base pairs long. Coalescent mutation and recombination rate per base pair were set to 10^{-8}, while effective population size was modelled over time to mimic population history of a livestock population in line with the values reported by MacLeod et al. (2013). Specifically, for the low diversity scenario effective population size of the base population was set to 100 and increased to 120, 250, 350, 1,000, 1,500, 2,000, 2,500, 3,500, 7,000, 10,000, 17,000, and 62,000 at respectively 6, 12, 18, 24, 154, 454, 654, 1,754, 2,354, 3,354, 33,154, and 933,154 generations ago. For the high diversity scenario, effective population size of the base population was set to 10,000 and increased above this value in the same way as in the low diversity scenario; to 17,000 and 62,000 at 33,154, and 933,154 generations ago. For each chromosome 10,000 whole chromosome haplotypes were sampled, which on average hosted about 700,000 markers (21 million per genome) for the low diversity scenario and 1,400,000 markers (42 million per genome) for the high diversity scenario. Out of these loci 100 per chromosome (3,000 per genome) were sampled as causal loci affecting a complex trait. The allele substitution effect of causal loci was sampled from a normal distribution with mean zero and variance 1/3,000. The effects were used to simulate a complex trait with additive genetic architecture. In addition, 2,000 loci per chromosome (60,000 per genome) were selected as markers with the restriction of having minor allele frequency above 0.05.

From the base population, founder genomes for four populations (A, B, C, and D) were obtained by random sampling of chromosomes with recombination. The populations were ancestrally related through the common base population, but otherwise maintained independently, i.e., there was no migration between the four populations. Each population was initiated with 10,000 founders (half males and half females) and maintained for 7 generations with constant size. In the low diversity scenario, with the effective population size of 100, 25 males and 5,000 females were selected as parents of each generation, while in the high diversity scenario, with the effective population size of 10,000, all 5,000 males and 5,000 females were used. The 25 males were selected on true genetic value, assuming accurate progeny test was available.

For every individual in the population we simulated two types of phenotypes. First, an own single phenotype was simulated as the sum of the true genetic value and a residual sampled from a normal distribution with mean zero and residual variance scaled relative to the variance of true genetic value in the base population such that heritability was 0.3. These simulated single phenotype records mimic records measured on the individual. Second, a weighted phenotype was simulated as the sum of the true genetic value and the mean of n_{weight} residuals. Each residual was sampled from a normal distribution with mean zero and residual variance scaled relative to the variance of true genetic value in the base population such that heritability was 0.3. The weight n_{weight} was equal to $n_{weight} = 1 + val$ where the real value val was sampled from a geometric distribution with a probability p of 0.15 and a probability mass function of $Pr(x) = p(1-p)^x$ with $x \in \{0, 1, 2, \dots\}$. The average n_{weight} was 6.6. These weighted phenotypes mimic either repeated records of an individual or records on multiple progeny of an individual. To satisfy the assumption of identical residual variance across all analyses, phenotype records were divided by the residual standard deviation specific for each population, such that $\sigma_e^2 = 1$. For every individual in each population we stored the true genetic value, own single and weighted phenotype records, associated weight, and 60,000 marker genotypes.

Analysis

The data was analysed in several ways to evaluate the developed methods. In each case the aim was to obtain accurate genetic values utilizing all the available information. Specifically, we integrated results from separate analysis of populations B, C, and D, into the analysis of population A. We assumed throughout that variance components were known and equal to the rescaled variances. We analysed three scenarios in total. The first and second scenario used population specific training data of randomly sampled 30,000 individuals with single phenotype record from generations 1 to 6 under low and high diversity settings. The third scenario used population specific training data of randomly sampled 10,000 individuals with weighted phenotype record from generations 1 to 6 under low diversity setting. In all scenarios all of the 10,000 individuals from generation 7 of each population were considered as validation individuals. The following analyses were performed:

1) A joint analysis of four populations. This was the reference that the other analyses were compared against;
2) A separate analysis for each of the four populations;
3) An exact integration of separate analyses of populations B, C, and D, into the analysis of population A;
4) The same as 3), but approximating the PEC matrix with a partial PEC matrix for each chromosome, i.e., PEC between markers on different chromosomes were set to zero;
5) The same as 3), but approximating the PEC matrix with a diagonal PEV matrix, i.e., PEC between all markers were set to zero;
6) The same as 3), but approximating the PEC matrix with PEV, allele frequencies, and LD correlations between markers obtained from the training sets. For the scenario with weighted phenotype records, the algorithm for estimating the effective number of records per marker was performed for each marker separately and for each chromsome separately.
7) The same as 6), but with LD correlations between markers computed from validation individuals instead of the training data.

For each analysis we calculated genomic prediction accuracy as the Pearson correlation between the true and estimated genetic value in validation individuals. Further, we evaluated the different integrations by comparing estimated genetic values of validation individuals against the estimated genetic values obtained from the joint analysis, which was considered as the reference because it used information from all populations. If integration was fully accurate, there should be no difference between the joint analysis and the analysis with integration. We assessed this by (a) accuracy of integration as a Pearson correlation between estimated genetic values from the joint analysis and the analysis with integration (desired value equals 1), (b) calibration of integration as a regression of estimated genetic values from the joint analysis on estimated genetic values from analysis with integration, and (c) magnitude of error in integration as a mean square error (MSE) between estimated genetic values from the joint analysis and from the analysis with integration (desired value equals 0). By calibration, we mean the slope of relationship of the estimates from the integration analysis onto the estimated genetic values from the joint analysis. The desired slope value is 1, which indicates a well calibrated model. Values above or below 1 indicate an uncalibrated model.

Data availability

Supplemental figures are available in File S1. A description of the simulated genotype and phenotype datasets for each scenario is provided in File S2. Simulated genotype and phenotype datasets for the 5 replicates of each scenario are provided in Files S3, S4, and S5. Data simulation scripts and Fortran codes developed to perform the different analyses, as well as a short description of each of them, are provided in File S6. All files were uploaded to Figshare: https://figshare.com/s/473dc83a7b154cfd610c.

RESULTS

Genomic prediction accuracy of separate and joint analyses

Joint analysis increased genomic prediction accuracy in comparison to separate analyses. This is shown in Table 1. Analysing separately the four datasets gave accuracies of about 0.71 (low diversity) and 0.53 (high diversity) with single phenotype records, and of about 0.73 (low diversity) with weighted phenotype records. Analysing jointly the four datasets increased accuracy by at least 0.09 absolute points with single phenotype records and by at least 0.12 absolute points with weighted phenotype records.

Integration based on PEC, partial PEC, or PEV matrices

For all scenarios the developed method enabled exact integration when complete PEC matrices were used. Integration of estimated allele substitution effects by means of the complete PEC matrix led to the same estimated genetic values as with the joint analysis, as shown by correlation and regression coefficients of 1, and MSE close to 0 (Figures 1-4; Figures S1-S8). For comparison, correlations between estimated genetic values from separate analyses and joint estimated genetic values were about 0.87 (low diversity) and 0.77 (high diversity) with single phenotype records, and 0.85 (low diversity) with weighted phenotype records.

Approximate integration by means of partial PEC matrices for each chromosome, that is ignoring PEC between markers on different chromosomes, gave almost as accurate and calibrated estimated genetic values as the exact integration. This is illustrated in Figures 1-4 and Figures S1-S8 with correlations higher than 0.96, regression coefficients close to 1, and MSE close to 0. Increasing the diversity slightly deteriorated accuracy and calibration of genomic predictions (Figures 1-2; Figures S1-S4).

Approximate integrations by means of PEV matrices, that is ignoring PEC between all markers, gave quite accurate, but not calibrated estimated genetic values. This is shown in Figures 1-4 and in Figures S1-S8. Correlations between joint estimated genetic values and estimated genetic values with integration by means of PEV were between 0.95 and 0.98 with single phenotype records and between 0.93 and 0.95 with weighted phenotype records. Despite these correlations close to 1, estimated genetic values were not well calibrated, as depicted by regression coefficients below 0.77 for the low diversity scenarios with single and weighted phenotype records, and below 0.86 for the high diversity scenario with single phenotype records (Figures 2, 4, S2, S6).

Integration based on PEV, allele frequencies, and LD information

When LD information was derived from training data of other populations, approximate integrations by means of PEV, allele frequencies, and LD information, resulted in highly accurate and well calibrated estimated genetic values with single phenotype records. This is shown in Figures 1-2 (Figures S1-S4). Correlation and regression coefficients were equal to 1 for the low diversity scenario. Slightly lower values, but still close to 1, were observed for the high diversity scenario. For both low and high diversity scenarios, MSE were close to 0. In contrast, when LD information was derived from validation data of other populations, approximate integrations gave less accurate and calibrated estimated genetic values. This is shown in Figures 1-2 (Figures S1-S4). For these scenarios, correlations were equal to at least 0.94, and regression coefficients varied between 0.87 and 1.05.

For the scenario with weighted phenotype records, approximate integrations by means of LD information from training data of other populations resulted in highly accurate and well calibrated estimated genetic values when sets of markers per chromosome were used to estimate the effective number of records for each marker. Correlations between joint estimated genetic values and estimated genetic values with integration were about 0.99 (Figure 3, Figure S5), regression coefficients were about 0.95 (Figure 4, Figure S6), and MSE were close to 0 (Figures S7-S8). Using LD information from the validation data of other populations, instead from the training data of other populations, gave slightly less accurate (correlations higher than 0.95), and moderately less calibrated estimated genetic values (regression coefficients between 0.87 and 1.04; Figures 3-4; Figures S5-S8). For both cases, estimating the effective numbers of records per marker, instead of for all markers per chromosome simultaneously, reduced accuracy and calibration of estimated genetic values (Figures 3-4; Figures S5-S6).

Comparison of estimated allele substitution effects

Correlation and regression coefficients between estimated allele substitution effects from the joint analysis and analysis with integration largely followed patterns of the corresponding values for estimated genetic values (Tables 2-3). Correlation and regression coefficients were close to 1 when the integration of estimated allele substitution effects was by means of the complete PEC matrices. Ignoring PEC between markers on different chromosomes, or ignoring PEC between all markers, reduced correlations to between 0.92 and 0.99 (Tables 2-3). Using LD information with PEV led to correlations between joint estimates of allele substitution effects and estimates with integration ranging from 0.71 to 0.83 for the scenario with weighted phenotype records (Tables 2-3).

DISCUSSION

The results show that the developed method enables accurate and well calibrated estimated genetic values for complex traits using both individual-level data and summary statistics. As expected from theory, the analysis of individual-level data and estimated allele substitution effects from other analyses by means of PEC matrices, yielded the same estimates as the joint analysis of all individual-level data. To our knowledge, this is the first time that individual-level data and summary statistics were analysed simultaneously for genomic predictions. As illustrated by simulations, the combined analysis of multiple datasets may increase genomic prediction accuracy over separate analyses of a single dataset. Unfortunately, combining individual-level data from several sources is generally not feasible for several reasons, e.g., political roadblocks, data protections concerns, or data inconsistencies (Powell and Sieber, 1992; Vilhjálmsson et al., 2015; Maier et al., 2018). However, summary statistics, such as estimates of allele substitution effects and associated measures of accuracy (e.g., PEV), are usually available for exchange in human genetics, or are discussed to be shared, e.g., at an international level for dairy cattle breeding (Liu and Goddard, 2018). The developed method enables increase in genomic prediction accuracy of complex traits by means of jointly analysing the available individual-level data and summary statistics.

Accurate integration of estimated allele substitution effects is possible also when the complete PEC matrix is not available. This is important because computing the exact PEC matrix and exchanging it between analyses might be challenging in some cases. For the vast majority of used marker arrays in animal and plant breeding the calculations and data transfers should be doable. For example, most arrays have between 10,000 and 100,000 markers, for which we need between ~1 and ~80 GB of memory to store the PEC matrix and between a minute and a day to invert it on current computers. For a larger number of markers, commonly used in human genetics, the memory requirements and computing time become prohibitive. The results show that in such cases we can still obtain accurate genomic predictions when the integration is done by means of partial PEC matrices for each chromosome. This is expected since high LD between markers mostly occurs within chromosomes. High LD between markers on different chromosomes may especially occur in structured populations and populations under selection (Farnir et al., 2000; Flint-Garcia et al., 2003; Rostoks et al., 2006). Both of these conditions are present in breeding populations. However, the results suggest that LD between chromosomes can be ignored for the purpose of integration for populations with both low and high diversity. The results also show that we can succesfully integrate estimated allele substitution effects when only PEV and allele frequencies from each population are available together with LD information of a reference genotype panel representative of each population. Assuming that such reference genotype panels are available, only estimated allele substitution effects, associated PEV, and allele frequencies need to be exchanged between populations for such analyses. Similar conclusions were drawn from studies combining only summary statistics obtained from genome-wide association studies to perform multi-trait genomic predictions (Maier et al., 2018).

Accurate integration of estimated allele substitution effects is possible irrespective of the diversity of the populations and characteristics of genotypes (e.g., allele frequencies, LD). This is obvious, and confirmed by our results, when integration is perfomed by means of complete PEC matrices. When complete PEC matrices are unavailable, accurate integration is possible if the inverses of the PEC matrices can be approximated accurately from available population parameters (i.e. LD and allele frequency information), whatever the level of diversity and characteristics of the populations, as shown by our results or a study combining summary statistics in human genetics (Maier et al., 2018). In our study, the population parameters obtained from the reference panels adequately reflected the characteristics of the training sets. We expect that this would be the case for populations with substantial migration, such as, for example, Holstein dairy cattle populations. Future studies should be conducted to assess the impact of suboptimal reference panels. Therefore, the developed method is expected to perform well on any type of data, from animal and plant breeding to human genetics, provided accurate information is available.

The developed method has some simplifying assumptions that can be readily relaxed. For example, we assumed that the same genotype coding was used in all populations. This assumption can be relaxed when centered genotype coding (i.e., of the form of $(\mathbf{W}_i - \mathbf{1}\mathbf{v}_i')$) is used because variance component estimates, estimates of allele substitution effects and PEC are the same irrespective of the centering of the genotype coding, provided that the model has a fixed general mean, which is considered in the integration (Strandén and Christensen, 2011). Also, centered and scaled (standardised) genotype coding is often used in human genetics, instead of only centered genotype coding (Yang et al., 2010; Speed et al., 2012; Maier et al., 2018). In practice, estimates of genetic values are only slightly influenced by scaling of centered genotype coding (Strandén and Christensen, 2011; Bouwman et al., 2017). Therefore, assuming that the same estimated genetic values are obtained with different scaling, allele substitution effects estimated using one type of genotype scaling could be obtained from a post-analysis by converting estimated genetic values computed for a reference genotype panel into allele substitution effects for another genotype scaling. Converting estimated genetic values into allele substitution effects is often referred to as back-solving of allele substitution effects (Strandén and Garrick, 2009; Strandén and Christensen, 2011; Wang et al., 2012; Bouwman et al., 2017). Prediction error covariances associated with the converted estimated allele subsitution effects could be derived from the (prediction error) covariances of the estimated genetic values (see derivations in Appendix A4).

Allele substitution effects estimated from analyses using different sets of markers or different residual variances, can be used in the integration as well. The assumption that all individuals were genotyped at the same loci could be considered as fullfilled if small differences in the sets of markers are corrected by assuming zero allele substitution effect and zero accuracy for markers not used in an analysis. When large differences between sets of markers are observed, this assumption can be accomodated following two approaches. A first, post-analysis, approach consists of assuming that estimated genetic values are the same for two different sets of markers, allowing the conversion of estimated allele substitution effects from one set of markers to another set of markers (Liu and Goddard, 2018). The conversion can be performed by back-solving estimated allele substitution effects from estimated genetic values, as proposed previously for different genotype codings, or by applying a marker model to the estimated genetic values with the reference set of markers (Liu and Goddard, 2018). A second approach consists of harmonizing genotype data across populations. This approach must be performed before the analyses, and requires therefore coordination between populations. Harmonization of genotype data could be performed by identifying a subset of markers for which all populations are genotyped, or by genotype imputation (e.g., Marchini and Howie, 2010). Finally, the assumption that residual variances were the same in all populations, can be relaxed by noting that separate estimates of allele substitution effects $\widehat{\boldsymbol{\alpha}_i^*}$, obtained by the system of equations (2), can be also obtained by the following different formulations:

$$\begin{aligned}\widehat{\boldsymbol{\alpha}_i^*} &= \left(\mathbf{W}_i'\mathbf{Z}_i'\mathbf{M}_i\sigma_{e_i}^2\mathbf{Z}_i\mathbf{W}_i + \mathbf{B}_i^{-1}\sigma_{\alpha_i}^{-2}\right)^{-1}\mathbf{W}_i'\mathbf{Z}_i'\mathbf{M}_i\sigma_{e_i}^2\mathbf{y}_i \\ &= (\mathbf{W}_i'\mathbf{Z}_i'\mathbf{M}_i\mathbf{Z}_i\mathbf{W}_i + \mathbf{B}_i^{-1}\lambda)^{-1}\mathbf{W}_i'\mathbf{Z}_i'\mathbf{M}_i\mathbf{y}_i \\ &= \left(\mathbf{W}_i'\mathbf{Z}_i'\mathbf{M}_i\sigma_{e_f}^{-2}\mathbf{Z}_i\mathbf{W}_i + \mathbf{B}_i^{-1}\lambda\sigma_{e_f}^{-2}\right)^{-1}\mathbf{W}_i'\mathbf{Z}_i'\mathbf{M}_i\sigma_{e_f}^{-2}\mathbf{y}_i\end{aligned}$$

where $\sigma_{e_i}^2$ ($\sigma_{e_f}^2$) is the residual variance used for the i-th (focal) analysis, and $\lambda = \sigma_{e_i}^2\sigma_{\alpha_i}^{-2}$.

For integration of $\widehat{\boldsymbol{\alpha}_i^*}$, $\left(PEC\left(\widehat{\boldsymbol{\alpha}_i^*}\right)\right)^{-1}$ must be approximated using the residual variance of the focal population ($\sigma_{e_f}^2$) and the effective numbers of records per marker estimated using variance components of the i-th analysis. Another way to relax this assumption is to extend our univariate model to a bivariate model, similarly to methods developed to combine different genetic evaluations in animal breeding (Schaeffer, 1994; Vandenplas et al., 2015). In a bivariate model, one trait would represent individual-level data, while the other trait would represent summary statistics. The genetic correlation between the two traits could be estimated based on a subset of individual-level data available for both datasets or based on summary statistics (Bulik-Sullivan et al., 2015). Such an approach would also allow the integration of summary statistics expressed on a different scale (e.g., different measure units, trait definitions) than the scale of the focal population (Vandenplas et al., 2015).

The developed method can be readily generalized to multi-trait models and is therefore a generalization of previous works that were based on several (implicit) assumptions (Liu and Goddard, 2018; Maier et al., 2018). For example, previous works assumed that no individual-level data were available. It was also (implicitly) assumed that only single phenotype records with homogeneous residual variance (Maier et al., 2018), or that the least-squares part of the separate analyses (Liu and Goddard, 2018), were available for integrating estimated allele substitution effects. Both assumptions lead to simple and accurate approximations of PEC matrices as shown in our study. However, we relax all these assumptions, such that our method can jointly analyse individual-level data and summary statistics, with possibly multiple phenotype records per individual.

With all the proposed generalizations, the developed method could be used in different contexts. For example, in human genetics, allele substitution effects with associated standard errors are publicly available (Yang et al., 2012; Vilhjálmsson et al., 2015; Maier et al., 2018). In animal breeding, individuals' genetic values with associated reliabilities are publicly available and in the case of dairy cattle extensively combined across multiple populations (Schaeffer, 1994; VanRaden and Sullivan, 2010; Jorjani et al., 2012; Vandenplas et al., 2017). The developed method can be used in both contexts, but in the latter case individuals' genetic values must be first back-solved to allele substitution effects (Strandén and Garrick, 2009; Strandén and Christensen, 2011; Wang et al., 2012; Bouwman et al., 2017). It is worth noting that our method assumes that summary statistics from one population are free of information from other populations. This suggest that it can be used when there is no, or limited, sharing of information between populations, as is for example the case in beef cattle, but not in dairy cattle populations such as Holstein, where pseudo-phenotypes summarising information from multiple populations are used extensively (VanRaden and Sullivan, 2010; Jorjani et al., 2012). This assumption can be relaxed by performing separate analyzes free of information from other populations, or by correcting for double-counting of information, which has bee developed for the integration of estimated genetic values from different populations (Vandenplas et al., 2014, 2017; VanRaden et al., 2014). This correction for double-counting of information is not yet developed for the integration of summary statistics, and should be investigated in future studies.

CONCLUSIONS

We developed a method for genomic prediction that accurately integrates summary statistics obtained from analyses of separate populations into an analysis of individual-level data. The method accommodates use of multiple phenotype (pseudo-)records per individual, and further extensions have been presented to accommodate for differences in residual variances or genotype codings used in the populations. When complete summary statistics information is available the method gives identical genomic predictions as the joint analysis of individual-level data from all populations. When summary statistics information is not complete we can use a series of approximations that give very accurate and well calibrated genomic predictions.

AUTHORS' CONTRIBUTIONS

JV derived the equations, wrote the programs to do the analyses, performed the analyses, and drafted the outline of the manuscript. GG performed the simulations. All authors discussed the design of the simulations. JV and GG wrote the first version of the manuscript. All authors provided valuable insights throughout the analysis and writing process.

ACKNOWLEDGMENTS

This study was financially supported by the Dutch Ministry of Economic Affairs (TKI Agri & Food project 16022), the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin, and UK Biotechnology and Biological Sciences Research Council (BBSRC) ISPG to The Roslin Institute BBS/E/D/30002275. The use of the HPC cluster has been made possible by CAT-AgroFood (Shared Research Facilities Wageningen UR).

LITERATURE CITED

Bouwman, A.C., B.J. Hayes, and M.P.L. Calus. 2017. Estimated allele substitution effects underlying genomic evaluation models depend on the scaling of allele counts. Genet. Sel. Evol. 49. doi:10.1186/s12711-017-0355-9.

Bulik-Sullivan, B., H.K. Finucane, V. Anttila, A. Gusev, F.R. Day, et al. 2015. An atlas of genetic correlations across human diseases and traits. Nat. Genet. 47:1236–1241. doi:10.1038/ng.3406.

Burden, R.L., and J.D. Faires. 2010. Numerical Analysis. 9 edition. Brooks Cole, Boston, MA.

Campos, G. de los, D. Gianola, and D.B. Allison. 2010. Predicting genetic predisposition in humans: the promise of whole-genome markers. Nat. Rev. Genet. 11:880–886. doi:10.1038/nrg2898.

de los Campos, G., J.M. Hickey, R. Pong-Wong, H.D. Daetwyler, and M.P.L. Calus. 2012. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345. doi:10.1534/genetics.112.143313.

Daetwyler, H.D., B. Villanueva, and J.A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PLoS ONE 3.

Damesa, T.M., J. Möhring, M. Worku, and H.-P. Piepho. 2017. One step at a time: Stage-wise analysis of a series of experiments. Agron. J. 109:845–857. doi:10.2134/agronj2016.07.0395.

Dempfle, L. 1977. Relation entre BLUP (Best Linear Unbiased Prediction) et estimateurs bayésiens. Genet. Sel. Evol. 9:27–32.

Farnir, F., W. Coppieters, J.-J. Arranz, P. Berzi, N. Cambisano, et al. 2000. Extensive genome-wide linkage disequilibrium in cattle. Genome Res. 10:220–227. doi:10.1101/gr.10.2.220.

Faux, A.-M., G. Gorjanc, R.C. Gaynor, M. Battagin, S.M. Edwards, et al. 2016. AlphaSim: Software for breeding program simulation. Plant Genome 9.

Fisher, R.A. 1918. The correlation between relatives on the supposition of Mendelian inheritance. Philos. Trans. R. Soc. Edinb. 52:399–433.

Flint-Garcia, S.A., J.M. Thornsberry, and E.S. Buckler. 2003. Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54:357–374. doi:10.1146/annurev.arplant.54.031902.134907.

Garrick, D.J., J.F. Taylor, and R.L. Fernando. 2009. Deregressing estimated breeding values and weighting information for genomic regression analyses. Genet. Sel. Evol. 41:55. doi:10.1186/1297-9686-41-55.

Gianola, D., and R.L. Fernando. 1986. Bayesian methods in animal breeding theory. J. Anim. Sci. 63:217–244.

Henderson, C.R. 1984. Applications of Linear Models in Animal Breeding. 2nd ed. University of Guelph, Guelph, ON, Canada.

Hickey, J.M., and G. Gorjanc. 2012. Simulated data for genomic selection and genome-wide association studies using a combination of coalescent and gene drop methods. G3 2:425–427. doi:10.1534/g3.111.001297.

Hoerl, A.E., and R.W. Kennard. 1976. Ridge regression iterative estimation of the biasing parameter. Commun. Stat. - Theory Methods 5:77–88. doi:10.1080/03610927608827333.

Hozé, C., S. Fritz, F. Phocas, D. Boichard, V. Ducrocq, et al. 2014. Efficiency of multi-breed genomic selection for dairy cattle breeds with different sizes of reference population. J. Dairy Sci. 97:3918–3929. doi:10.3168/jds.2013-7761.

Jairath, L., J.C.M. Dekkers, L.R. Schaeffer, Z. Liu, E.B. Burnside, et al. 1998. Genetic evaluation for herd life in Canada. J. Dairy Sci. 81:550–562.

Jorjani, H., J. Jakobsen, E. Hjerpe, V. Palucci, and J. Dürr. 2012. Status of genomic evaluation in the Brown Swiss populations. Interbull Bull. 46:46–54.

Legarra, A., J.K. Bertrand, T. Strabel, R.L. Sapp, J.P. Sanchez, et al. 2007. Multi-breed genetic evaluation in a Gelbvieh population. J. Anim. Breed. Genet. 124:286–295.

Lindley, D.V., and A.F.M. Smith. 1972. Bayes estimates for the linear model. J. R. Stat. Soc. Ser. B Methodol. 34:1–41.

Liu, Z., and M.E. Goddard. 2018. A SNP MACE model for international genomic evaluation: technical challenges and possible solutions. Page 11.393 in Proceedings of the 11th World Congress on Genetics Applied to Livestock Production, Auckland, New Zeland.

MacLeod, I.M., D.M. Larkin, H.A. Lewin, B.J. Hayes, and M.E. Goddard. 2013. Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. Mol. Biol. Evol. 30:2209–2223.

Maier, R.M., Z. Zhu, S.H. Lee, M. Trzaskowski, D.M. Ruderfer, et al. 2018. Improving genetic prediction by leveraging genetic correlations among human diseases and traits. Nat. Commun. 9:989.

Marchini, J., and B. Howie. 2010. Genotype imputation for genome-wide association studies. Nat. Rev. Genet. 11:499–511. doi:10.1038/nrg2796.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Misztal, I., and G.R. Wiggans. 1988. Approximation of prediction error variance in large-scale animal models. J. Dairy Sci. 71(Suppl. 2):27–32.

Oakey, H., B. Cullis, R. Thompson, J. Comadran, C. Halpin, et al. 2016. Genomic selection in multi-environment crop trials. G3 Bethesda Md 6:1313–1326. doi:10.1534/g3.116.027524.

Pasaniuc, B., and A.L. Price. 2017. Dissecting the genetics of complex traits using summary association statistics. Nat. Rev. Genet. 18:117–127. doi:10.1038/nrg.2016.142.

Powell, R.L., and H.D. Norman. 1998. Use of multinational data to improve national evaluations of Holstein bulls. J. Dairy Sci. 81:2257–2263. doi:10.3168/jds.S0022-0302(98)75805-9.

Powell, R.L., and M. Sieber. 1992. Direct and indirect conversion of bull evaluations for yield traits between countries. J. Dairy Sci. 75:1138–1146.

Quaas, R.L., and Z. Zhang. 2006. Multiple-breed genetic evaluation in the US beef cattle context: Methodology. Page CD-ROM Comm. 24-12 in Proceedings of the 8th World Congress on Genetics Applied to Livestock Production, Belo Horizonte, Brazil.

Rogers, A.R., and C. Huff. 2009. Linkage Disequilibrium between loci with unknown phase. Genetics 182:839–844. doi:10.1534/genetics.108.093153.

Rostoks, N., L. Ramsay, K. MacKenzie, L. Cardle, P.R. Bhat, et al. 2006. Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc. Natl. Acad. Sci. U. S. A. 103:18656–18661. doi:10.1073/pnas.0606133103.

Schaeffer, L.R. 1994. Multiple-country comparison of dairy sires. J. Dairy Sci. 77:2671–2678.

Schulthess, A.W., Y. Wang, T. Miedaner, P. Wilde, J.C. Reif, et al. 2016. Multiple-trait- and selection indices-genomic predictions for grain yield and protein content in rye for feeding purposes. TAG Theor. Appl. Genet. Theor. Angew. Genet. 129:273–287. doi:10.1007/s00122-015-2626-6.

Schulz-Streeck, T., J.O. Ogutu, and H.-P. Piepho. 2013. Comparisons of single-stage and two-stage approaches to genomic selection. Theor. Appl. Genet. 126:69–82. doi:10.1007/s00122-012-1960-1.

Speed, D., G. Hemani, M.R. Johnson, and D.J. Balding. 2012. Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. 91:1011–1021. doi:10.1016/j.ajhg.2012.10.010.

Strandén, I., and O.F. Christensen. 2011. Allele coding in genomic evaluation. Genet. Sel. Evol. 43:25. doi:10.1186/1297-9686-43-25.

Strandén, I., and D.J. Garrick. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J. Dairy Sci. 92:2971–2975. doi:10.3168/jds.2008-1929.

Vandenplas, J., F.G. Colinet, and N. Gengler. 2014. Unified method to integrate and blend several, potentially related, sources of information for genetic evaluation. Genet. Sel. Evol. 46:59.

Vandenplas, J., F.G. Colinet, G. Glorieux, C. Bertozzi, and N. Gengler. 2015. Integration of external estimated breeding values and associated reliabilities using correlations among traits and effects. J. Dairy Sci. 98:9044–9050. doi:10.3168/jds.2015-9894.

Vandenplas, J., and N. Gengler. 2012. Comparison and improvements of different Bayesian procedures to integrate external information into genetic evaluations. J. Dairy Sci. 95:1513–1526.

Vandenplas, J., M. Spehar, K. Potocnik, N. Gengler, and G. Gorjanc. 2017. National single-step genomic method that integrates multi-national genomic information. J. Dairy Sci. 100:465–478. doi:10.3168/jds.2016-11733.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423. doi:10.3168/jds.2007-0980.

VanRaden, P.M., and P.G. Sullivan. 2010. International genomic evaluation methods for dairy cattle. Genet. Sel. Evol. 42:7. doi:10.1186/1297-9686-42-7.

VanRaden, P.M., M.E. Tooker, J.R. Wright, C. Sun, and J.L. Hutchison. 2014. Comparison of single-trait to multi-trait national evaluations for yield, health, and fertility. J. Dairy Sci. 97:7952–7962.

Vilhjálmsson, B.J., J. Yang, H.K. Finucane, A. Gusev, S. Lindström, et al. 2015. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am. J. Hum. Genet. 97:576–592. doi:10.1016/j.ajhg.2015.09.001.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W.M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. 94:73–83. doi:10.1017/S0016672312000274.

Whittaker, J.C., R. Thompson, and M.C. Denham. 2000. Marker-assisted selection using ridge regression. Genet. Res. 75:249–252.

Wientjes, Y.C.J., P. Bijma, R.F. Veerkamp, and M.P.L. Calus. 2016. An equation to predict the accuracy of genomic values by combining data from multiple traits, populations, or environments. Genetics 202:799–823. doi:10.1534/genetics.115.183269.

Wray, N.R., J. Yang, B.J. Hayes, A.L. Price, M.E. Goddard, et al. 2013. Pitfalls of predicting complex traits from SNPs. Nat. Rev. Genet. 14:507–515. doi:10.1038/nrg3457.

Yang, J., B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, et al. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat. Genet. 42:565–569. doi:10.1038/ng.608.

Yang, J., T. Ferreira, A.P. Morris, S.E. Medland, G.I. of An.T. (GIANT) Consortium, et al. 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat. Genet. 44:369–375. doi:10.1038/ng.2213.

Table 1 – Genomic prediction accuracy for joint and separate analyses in scenarios with single or weighted phenotype records and low or high diversity (values are averages across the five replicates[1])

Phenotypes	Diversity	Analysis	Populations			
			A	B	C	D
Single	Low	Joint	0.811	0.811	0.823	0.815
		Separate	0.705	0.708	0.718	0.718
	High	Joint	0.687	0.686	0.687	0.684
		Separate	0.536	0.537	0.528	0.528
Weighted	Low	Joint	0.860	0.865	0.865	0.862
		Separate	0.720	0.739	0.724	0.727

[1] Standard errors are between 0.003 and 0.016.

Table 2 - Comparison of estimated allele substitution effects from different analyses with estimates from the joint statistical analysis using single phenotype records in scenarios with low and high diversity (values are averages across the five replicates[1])

Analysis	Low diversity		High diversity	
	Correlation	Regression	Correlation	Regression
Separate A	0.71	1.09	0.65	1.10
Separate B	0.71	1.09	0.65	1.10
Separate C	0.71	1.09	0.65	1.11
Separate D	0.71	1.09	0.64	1.10
PEC	1.00	1.00	1.00	1.00
$\text{PEC}_{\text{within chromosome}}$	0.99	0.98	0.97	0.95
PEV	0.96	0.80	0.96	0.89
$\text{LD}_{\text{training}}$	1.00	1.00	0.98	0.97
$\text{LD}_{\text{validation}}$	0.96	0.88	0.93	0.84

[1] Standard errors are between 0.00 and 0.01.

Table 3 - Comparison of estimated allele substitution effects from different analyses with estimates from the joint statistical analysis using weighted phenotype records in the scenario with low diversity (values are averages across the five replicates with standard errors between brackets)

Analysis	Correlation	Regression
Separate A	0.61 (0.10)	0.88 (0.13)
Separate B	0.58 (0.15)	0.62 (0.12)
Separate C	0.56 (0.12)	0.93 (0.23)
Separate D	0.33 (0.08)	0.65 (0.18)
PEC	1.00 (0.00)	0.99 (0.01)
$PEC_{\text{within chromosome}}$	0.96 (0.01)	1.01 (0.02)
PEV	0.92 (0.02)	0.80 (0.05)
LD_{training} (1 marker)	0.77 (0.09)	0.83 (0.10)
LD_{training} (1 chromosome)	0.83 (0.09)	0.95 (0.11)
$LD_{\text{validation}}$ (1 marker)	0.73 (0.11)	0.75 (0.13)
$LD_{\text{validation}}$ (1 chromosome)	0.71 (0.15)	0.74 (0.18)

FIGURES

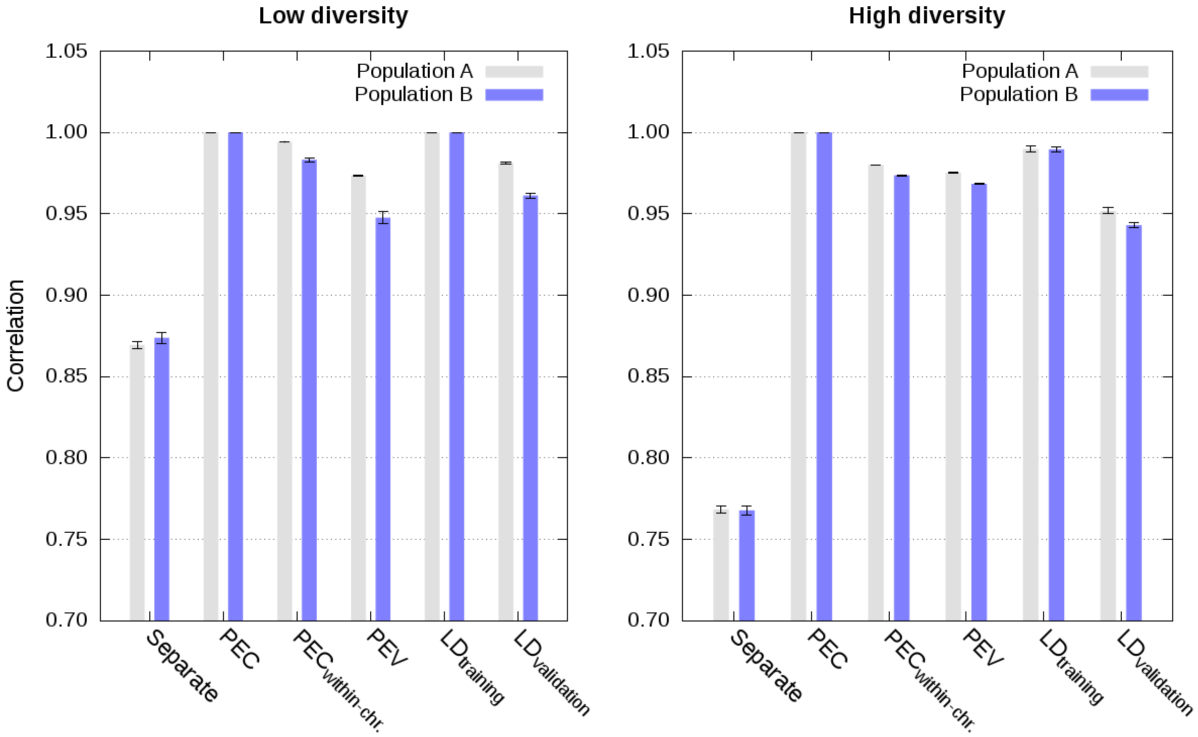

Figure 1 - Correlation between estimated genetic values from the joint analysis and from different analyses in populations A and B using a single phenotype record per individual in scenarios with low and high diversity (values are averages across the five replicates with standard errors).

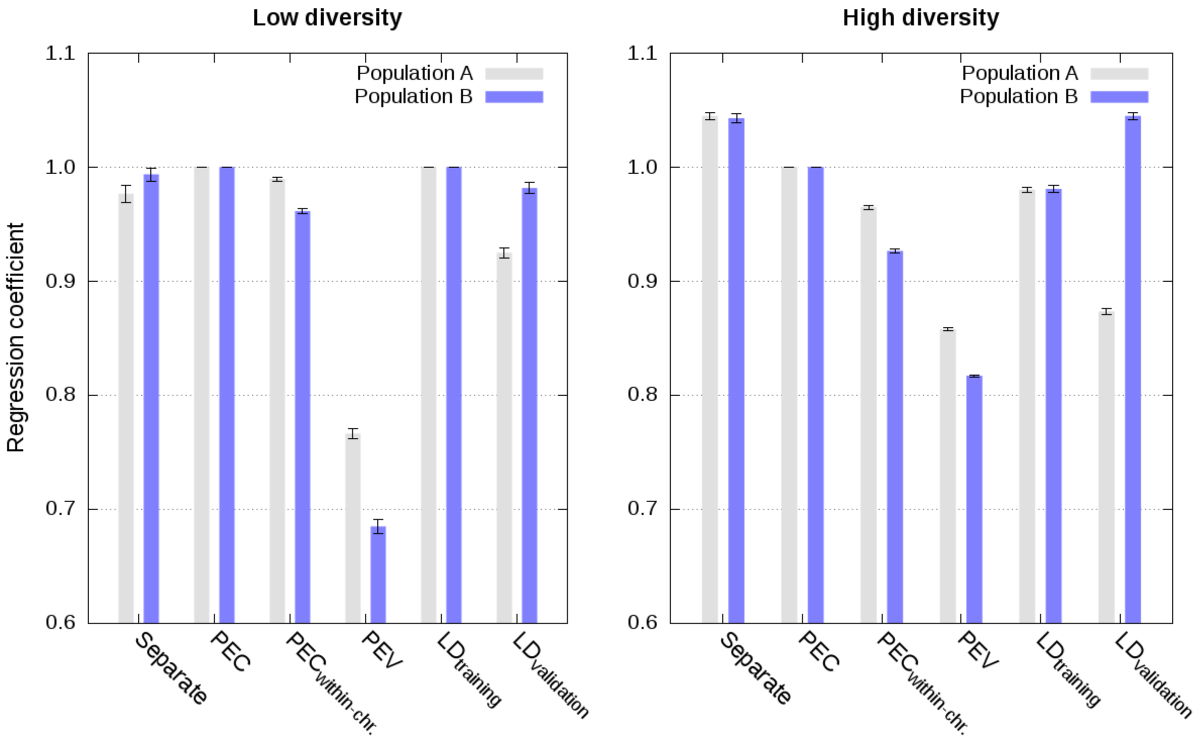

Figure 2 – Regression of estimated genetic values from the joint analysis on estimated genetic values from different analyses in populations A and B using a single phenotype record per individual in scenarios with low and high diversity (values are averages across the five replicates with standard errors).

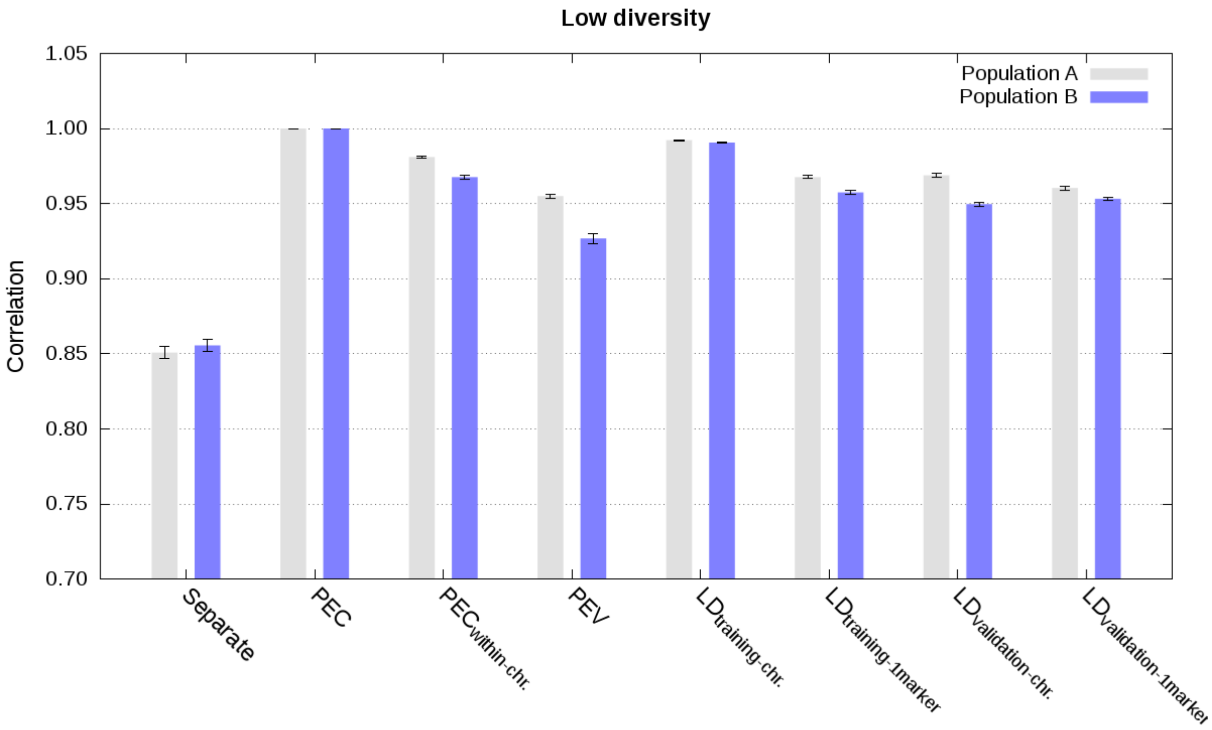

Figure 3 - Correlation between estimated genetic values from the joint analysis and from different analyses in populations A and B using weighted phenotype records in the scenario with low diversity (values are averages across the five replicates with standard errors).

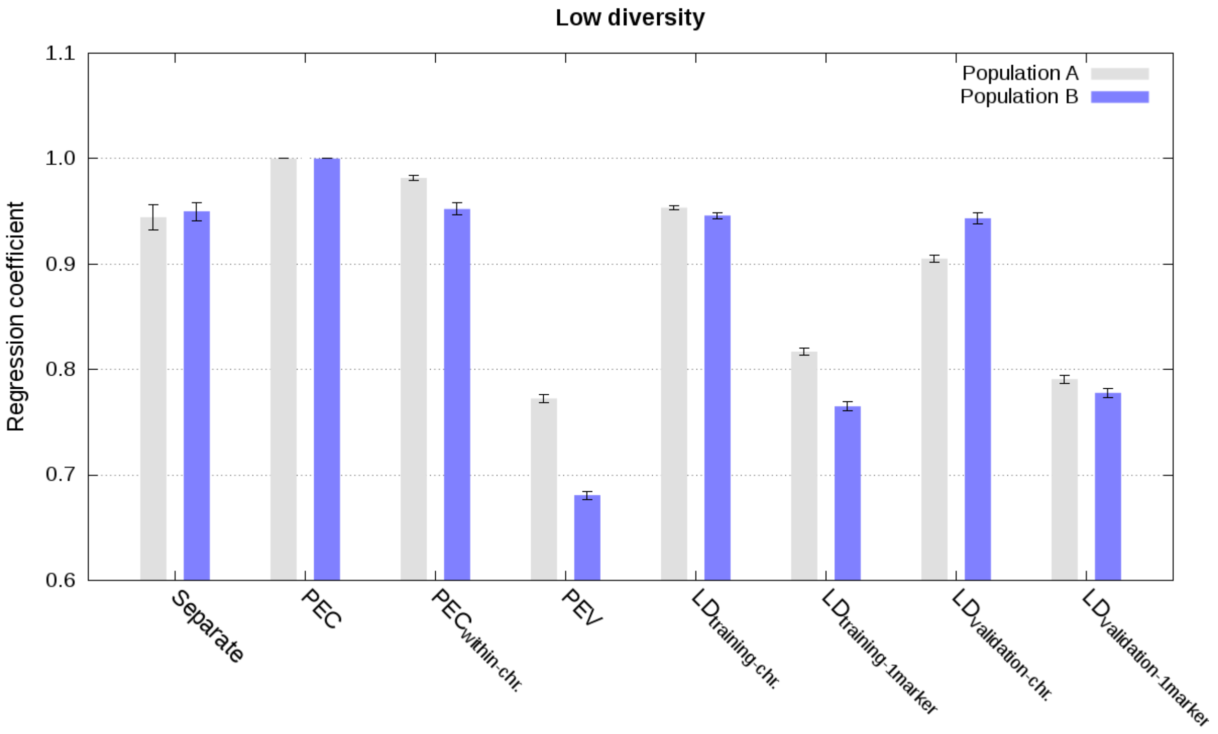

Figure 4 - Regression of estimated genetic values from the joint analysis on estimated genetic values from different analyses in populations A and B using weighted phenotype records in the scenario with low diversity (values are averages across the five replicates with standard errors).

Appendix A1: Exact integration

Here we detail the derivation of exact integration by means of absorbing the set of equations that pertain to one dataset. We start with the system of equations for separate analysis of dataset 1:

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{X}_1 & \mathbf{X}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1 \\ \mathbf{W}_1'\mathbf{Z}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{X}_1 & \mathbf{W}_1'\mathbf{Z}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1\ + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}_1^*} \\ \widehat{\boldsymbol{\alpha}_1^*} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{y}_1 \\ \mathbf{W}_1'\mathbf{Z}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{y}_1 \end{bmatrix} \quad \text{(A1.1)}$$

and the system of equations for the joint analysis of datasets 1 and 2:

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{X}_1 & \mathbf{0} & \mathbf{X}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1 \\ \mathbf{0} & \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{X}_2 & \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{Z}_2\ \mathbf{W}_2 \\ \mathbf{W}_1'\mathbf{Z}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{X}_1 & \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{X}_2 & \mathbf{W}_1'\mathbf{Z}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1\ + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{Z}_2\ \mathbf{W}_2\ + \mathbf{B}_J^{-1}\sigma_{\alpha_J}^{-2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}_1} \\ \widehat{\boldsymbol{\beta}_2} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} =$$

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{y}_1 \\ \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{y}_2 \\ \mathbf{W}_1'\mathbf{Z}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{y}_1\ + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{y}_2 \end{bmatrix}. \quad \text{(A1.2)}$$

From the first set of equations $\left(\widehat{\boldsymbol{\beta}_1}\right)$ in (A1.2) it follows:

$$\widehat{\boldsymbol{\beta}_1} = \left(\mathbf{X}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{y}_1 - \mathbf{X}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1\ \widehat{\boldsymbol{\alpha}}\right). \quad \text{(A1.3).}$$

From the third set of equations ($\widehat{\boldsymbol{\alpha}}$) in (A1.2) it follows:

$$\mathbf{W}_1'\mathbf{Z}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{X}_1\ \widehat{\boldsymbol{\beta}_1} + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{X}_2\ \widehat{\boldsymbol{\beta}_2} + \Big(\mathbf{W}_1'\mathbf{Z}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1\ +$$

$$\mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{Z}_2\ \mathbf{W}_2\ + \mathbf{B}_J^{-1}\sigma_{\alpha_J}^{-2}\Big)\widehat{\boldsymbol{\alpha}} = \mathbf{W}_1'\mathbf{Z}_1'\mathbf{R}_1^{-1}\sigma_e^{-2}\mathbf{y}_1\ + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{y}_2\ . \quad \text{(A1.4).}$$

Inserting (A1.3) into (A1.4) gives, after some algebra:

$$\mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{X}_2\ \widehat{\boldsymbol{\beta}_2} + \Big(\mathbf{W}_1'\mathbf{Z}_1'\mathbf{M}_1\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1\ + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{Z}_2\ \mathbf{W}_2\ + \mathbf{B}_\mathrm{J}^{-1}\sigma_{\alpha_J}^{-2}\Big)\widehat{\boldsymbol{\alpha}}$$

$$= \mathbf{W}_1'\mathbf{Z}_1'\mathbf{M}_1\sigma_e^{-2}\mathbf{y}_1\ + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{y}_2$$

with $\mathbf{M}_1 = \left(\mathbf{R}_1^{-1} - \mathbf{R}_1^{-1}\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{R}_1^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{R}_1^{-1}\right)$.

Now the system of equations (A1.2) can be re-written with the first set of equations $\left(\widehat{\boldsymbol{\beta}_1}\right)$ absorbed as:

$$\begin{bmatrix} \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{X}_2 & \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{Z}_2\mathbf{W}_2 \\ \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{X}_2 & \mathbf{W}_1'\mathbf{Z}_1'\mathbf{M}_1\sigma_e^{-2}\mathbf{Z}_1\mathbf{W}_1 + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{Z}_2\mathbf{W}_2 + \mathbf{B}_J^{-1}\sigma_{\alpha_J}^{-2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}_2} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} =$$

$$\begin{bmatrix} \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{y}_2 \\ \mathbf{W}_1'\mathbf{Z}_1'\mathbf{M}_1\sigma_e^{-2}\mathbf{y}_1 + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{y}_2 \end{bmatrix}. \qquad \text{(A1.4).}$$

Similarly, the absorption of the first set of equations $\left(\widehat{\boldsymbol{\beta}_1^*}\right)$ in separate analysis of dataset 1 (A1.1) leads to:

$$\left(\mathbf{W}_1'\mathbf{Z}_1'\mathbf{M}_1\sigma_e^{-2}\mathbf{Z}_1\mathbf{W}_1 + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)\widehat{\boldsymbol{\alpha}_1^*} = \mathbf{W}_1'\mathbf{Z}_1'\mathbf{M}_1\sigma_e^{-2}\mathbf{y}_1\ , \qquad \text{(A1.5)}$$

where

$$\mathbf{W}_1'\mathbf{Z}_1'\mathbf{M}_1\sigma_e^{-2}\mathbf{Z}_1\mathbf{W}_1 + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2} = \left(PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1} \qquad \text{(A1.6)}$$

is the inverse matrix of prediction error covariances of $\widehat{\boldsymbol{\alpha}_1^*}$.

Combining (A1.4) and (A1.5) with the use of (A1.6) enables the exact integration of estimates from the separate analysis of dataset 1 into the separate analysis of dataset 2 with the following system of equations:

$$\begin{bmatrix} \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{X}_2 & \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{Z}_2\mathbf{W}_2 \\ \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{X}_2 & \left(PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1} + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{Z}_2\mathbf{W}_2 - \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2} + \mathbf{B}_J^{-1}\sigma_{\alpha_J}^{-2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}_2} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} =$$

$$\begin{bmatrix} \mathbf{X}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{y}_2 \\ \left(PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right)\right)^{-1}\widehat{\boldsymbol{\alpha}_1^*} + \mathbf{W}_2'\mathbf{Z}_2'\mathbf{R}_2^{-1}\sigma_e^{-2}\mathbf{y}_2 \end{bmatrix}. \qquad \text{(A1.7)}$$

Appendix A2: Approximate integration

Here we detail the derivation of different approximate integrations by means of simplified assumptions and use of summary statistics. We start with the expression for prediction error covariance matrix of allele substitution effects from dataset 1:

$$PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right) = \left(\mathbf{W}_1'\mathbf{Z}_1'\mathbf{M}_1\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1\ \ + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)^{-1}. \qquad \text{(A2.1)}$$

If we assume that: (1) every individual has a single phenotype record, i.e., $\mathbf{Z}_1 = \mathbf{I}$, (2) residual variance is homogeneous, i.e. $\mathbf{R}_1 = \mathbf{I}$, and (3) only overall mean is fitted as a fixed effect, i.e., $\mathbf{X}_1 = \mathbf{1}$; then we can simplify (A2.1) as:

$$PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right) = \left(\mathbf{W}_1'\mathbf{Z}_1'\mathbf{M}_1\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1\ \ + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)^{-1},$$

$$= \left(\mathbf{W}_1'\mathbf{Z}_1'\left(\mathbf{R}_1^{-1} - \mathbf{R}_1^{-1}\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{R}_1^{-1}\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{R}_1^{-1}\right)\mathbf{Z}_1\ \mathbf{W}_1\ \sigma_e^{-2} + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)^{-1},$$

$$\approx \left(\mathbf{W}_1'\left(\mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right)\mathbf{W}_1\ \sigma_e^{-2} + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)^{-1},$$

$$\approx \left(\mathbf{W}_1'\mathbf{W}_1\ \sigma_e^{-2} + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)^{-1}, \qquad \text{(A2.2)}$$

because $\left(\mathbf{I} - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right) = \mathbf{I} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' = \mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n_{ind,1}}$ will tend to the identity matrix $\mathbf{I}$ with increasing $n_{ind,1}$. The matrix $\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n_{ind,1}}\right)$, also known as the centering matrix, is a symmetric and idempotent matrix with off-diagonal elements equal to $-\frac{1}{n_{ind,1}}$ and with diagonal elements equal to $1 - \frac{1}{n_{ind,1}}$.

When genotypes from the dataset 1 are not available, but variance components $\sigma_{\alpha_1}^2$ and σ_e^2 are, we “only” need to approximate the unknown matrix of genotype sum of squares $\mathbf{W}_1'\mathbf{W}_1$ in (A2.2). This product can be approximated from linkage-disequilibrium and allele frequency information of the dataset 1, as shown in the following (similarly to Yang et al. (2012), Vilhjálmsson et al. (2015), and Maier et al. (2018)). Assume that linkage-disequilibrium between two markers is represented by the correlation of their unphased genotypes (Rogers and Huff, 2009). Then, a matrix of all pairwise correlations between markers is:

$$\mathbf{C} = \left(diag\left(\mathbf{T}_1'\mathbf{T}_1\right)\right)^{-\frac{1}{2}}\mathbf{T}_1'\mathbf{T}_1 \left(diag\left(\mathbf{T}_1'\mathbf{T}_1\right)\right)^{-\frac{1}{2}}, \tag{A2.3}$$

where the matrix $\mathbf{T}_1$ contains centered genotypes of dataset 1 ($\mathbf{T}_1 = \left(\mathbf{I} - \frac{\mathbf{11}'}{n_{ind,1}}\right)\mathbf{W}_1 = \mathbf{W}_1 - \frac{1}{n_{ind,1}}\mathbf{11}'\mathbf{W}_1$). The matrix product $\mathbf{T}_1'\mathbf{T}_1$ can be computed as:

$$\mathbf{T}_1'\mathbf{T}_1 = \left(\mathbf{W}_1 - \frac{1}{n_{ind,1}}\mathbf{11}'\mathbf{W}_1\right)'\left(\mathbf{W}_1 - \frac{1}{n_{ind,1}}\mathbf{11}'\mathbf{W}_1\right) = \mathbf{W}_1'\mathbf{W}_1 - \frac{1}{n_{ind,1}}\mathbf{W}_1'\mathbf{11}'\mathbf{W}_1 - \frac{1}{n_{ind,1}}\mathbf{W}_1'\mathbf{11}'\mathbf{W}_1 + \frac{1}{n_{ind,1}}\frac{1}{n_{ind,1}}\mathbf{W}_1'\mathbf{11}'\mathbf{11}'\mathbf{W}_1 = \mathbf{W}_1'\mathbf{W}_1 - 4n_{ind,1}\mathbf{pp}'. \tag{A2.4}$$

where $\mathbf{p} = \frac{1}{2n_{ind,1}}\mathbf{W}_1'\mathbf{1}$ are allele frequencies in dataset 1 (Strandén and Christensen, 2011). Assuming Hardy-Weinberg equilibrium, the i-th diagonal element of the matrix product $\mathbf{T}_1'\mathbf{T}_1$, is equivalent to expected genotype sum of squares at the i-th marker, $n_{ind,1}2p_{i,1}\left(1 - p_{i,1}\right)$ with $p_{i,1}$ being the allele frequency of the i-th marker in dataset 1.

Combining (A2.3) and (A2.4) we can approximate the unknown matrix of genotype sum of squares $\mathbf{W}_1'\mathbf{W}_1$ as:

$$\mathbf{W}_1'\mathbf{W}_1 \approx 4n_{ind,1}\mathbf{pp}' + \mathbf{V}^{\frac{1}{2}}\mathbf{C}\mathbf{V}^{\frac{1}{2}}, \tag{A2.5}$$

where $\mathbf{V}$ is diagonal matrix of expected genotype sum of squares with the i-th diagonal element equal to $n_{ind,1}2p_{i,1}\left(1 - p_{i,1}\right)$.

Appendix A3: Estimation of the effective number of records per marker

Here we detail the algorithm for computing the effective number of records per marker by use of available population parameters (i.e. linkage-disequilibrium, and allele frequency information) and prediction error variances of $\widehat{\boldsymbol{\alpha}_1^*}$ ($PEV\left(\widehat{\boldsymbol{\alpha}_1^*}\right)$) of the dataset 1. We start with the expression for the prediction error covariance matrix of allele substitution effects from dataset 1:

$$PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right) = \left(\mathbf{W}_1'\mathbf{Z}_1'\mathbf{M}_1\sigma_e^{-2}\mathbf{Z}_1\ \mathbf{W}_1\ \ + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)^{-1}.$$

If the number of individuals and the number of records per individual are unknown, we can assume that a $n_{mar} \times n_{mar}$ diagonal matrix $\boldsymbol{\Lambda}_1$ exists such that:

$$PEC\left(\widehat{\boldsymbol{\alpha}_1^*}\right) \approx \left(\boldsymbol{\Lambda}_1\left(4\mathbf{pp}' + \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{C}\boldsymbol{\Psi}^{\frac{1}{2}}\right)\boldsymbol{\Lambda}_1\sigma_e^{-2} + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)^{-1}$$

where $\boldsymbol{\Psi}$ is a $n_{mar} \times n_{mar}$ diagonal matrix with the *j*-th diagonal element equal to $2p_{j,1}\left(1-p_{j,1}\right)$, and the squared *j*-th diagonal element of $\boldsymbol{\Lambda}_1$ represents the effective number of records for the *j*-th marker. The term $\left(4\mathbf{pp}' + \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{C}\boldsymbol{\Psi}^{\frac{1}{2}}\right)$ is similar to the approximation of the unknown matrix of genotype sum of squares $\mathbf{W}_1'\mathbf{W}_1$ (i.e., $\mathbf{W}_1'\mathbf{W}_1 \approx 4n_{ind,1}\mathbf{pp}' + \mathbf{V}^{\frac{1}{2}}\mathbf{C}\mathbf{V}^{\frac{1}{2}}$) in the Appendix A.2. However, it does not involve the number of individuals $n_{ind,1}$ because it is confounded with the effective number of records.

The diagonal matrix $\boldsymbol{\Lambda}_1$ can be estimated by solving the nonlinear system of equations

$diag\left(\left(\boldsymbol{\Lambda}_1\left(4\mathbf{pp}' + \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{C}\boldsymbol{\Psi}^{\frac{1}{2}}\right)\boldsymbol{\Lambda}_1\sigma_e^{-2} + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)^{-1}\right) = PEV\left(\widehat{\boldsymbol{\alpha}_1^*}\right)$ through a fixed-point iteration algorithm (Burden and Faires, 2010) as follows:

1) $\mathbf{Q}_1^0 = \left(\mathbf{P}^{0^{-1}} - \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right) * \left(diag\left(4\mathbf{pp}' + \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{C}\boldsymbol{\Psi}^{\frac{1}{2}}\right)\sigma_e^{-2}\right)^{-1}$

```
    where $\mathbf{P}^0$ is a diagonal matrix with the $i$-th diagonal element equal to the PEV of the $i$-th marker and $diag\left(4\mathbf{pp}' + \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{C}\boldsymbol{\Psi}^{\frac{1}{2}}\right)$ contains the diagonal elements of $\left(4\mathbf{pp}' + \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{C}\boldsymbol{\Psi}^{\frac{1}{2}}\right)$;
2) $\boldsymbol{\Lambda}_1^0 = \sqrt{\mathbf{Q}_1^0}$
3) $k = 1$
4) $\mathbf{P}^k = diag\left(\left(\boldsymbol{\Lambda}_1^{k-1}\left(4\mathbf{pp}' + \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{C}\boldsymbol{\Psi}^{\frac{1}{2}}\right)\boldsymbol{\Lambda}_1^{k-1}\sigma_e^{-2} + \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right)^{-1}\right)$
5) $\mathbf{H} = \left(\mathbf{P}^{k^{-1}} - \mathbf{B}_1^{-1}\sigma_{\alpha_1}^{-2}\right) * \left(diag\left(4\mathbf{pp}' + \boldsymbol{\Psi}^{\frac{1}{2}}\mathbf{C}\boldsymbol{\Psi}^{\frac{1}{2}}\right)\sigma_e^{-2}\right)^{-1}$
6) $\mathbf{S}^k = \mathbf{Q}_1^0 - \mathbf{H}$
7) If trace of $\mathbf{S}^k$ is not sufficiently small:
    a. $\mathbf{Q}_1^k = \mathbf{Q}_1^{k-1} + \mathbf{H}$
    b. If any diagonal element in $\mathbf{Q}_1^k$ is negative, set it to 0
    c. $\boldsymbol{\Lambda}_1^k = \sqrt{\mathbf{Q}_1^k}$
    d. $k = k + 1$
    e. Repeat from 4
8) $\boldsymbol{\Lambda}_1^k = \sqrt{\mathbf{Q}_1^k}$
```

It is worth noting that the proposed algorithm is similar to algorithms to estimate effective number of records per individual, where “effective” means that they are free of contributions from relatives (Misztal and Wiggans, 1988; Vandenplas and Gengler, 2012). The *j*-th diagonal element of $\mathbf{Q}_1^k$ can therefore equivalently be considered as the effective number of records for the *j*-th marker.

Appendix A4: Conversion of allele substitution effects

Here we detail a post-analysis to obtain allele substitution effects estimated using one type of genotype coding ($\widehat{\boldsymbol{\alpha}_1^{**}}$) by converting estimated genetic values computed for a reference genotype panel with allele substitution effects for another genotype coding ($\widehat{\boldsymbol{\alpha}_1^{*}}$). We assume that allele substitution effects ($\widehat{\boldsymbol{\alpha}_1^{*}}$) are available with the associated prediction error (co)variance matrix ($PEC\left(\widehat{\boldsymbol{\alpha}_1^{*}}\right)$), as well as the (co)variance matrix of $\boldsymbol{\alpha}_1^{*}$ ($Var(\boldsymbol{\alpha}_1^{*})$), and genotypes of a reference panel using a particular type of genotype coding ($\boldsymbol{\Gamma}^{*}$). Estimates of genetic values for the reference individuals are obtained as $\widehat{\mathbf{g}_1^{*}} = \boldsymbol{\Gamma}^{*}\widehat{\boldsymbol{\alpha}_1^{*}}$.

Assuming that estimated genetic values are not influenced by scaling of centered genotype coding (Strandén and Christensen, 2011; Bouwman et al., 2017), and that the (co)variances of genetic values are the same irrespective of the genotype coding, we can write that $\widehat{\mathbf{g}_1^{**}} = \boldsymbol{\Gamma}^{**}\widehat{\boldsymbol{\alpha}_1^{**}} = \widehat{\mathbf{g}_1^{*}}$ with $\boldsymbol{\Gamma}^{**}$ being a matrix with reference genotypes using another type of genotype coding than $\boldsymbol{\Gamma}^{*}$ and $\widehat{\mathbf{g}_1^{**}}$ being a vector of estimated genetic values using this type of genotype coding. Therefore, $\widehat{\boldsymbol{\alpha}_1^{**}}$ can be computed by back-solving as follows (Strandén and Garrick, 2009; Wang et al., 2012; Bouwman et al., 2017):

$$\widehat{\boldsymbol{\alpha}_1^{**}} = \mathbf{B}_1^{**}\boldsymbol{\Gamma}^{**\prime}\left(\boldsymbol{\Gamma}^{**}\mathbf{B}_1^{**}\boldsymbol{\Gamma}^{**\prime}\right)^{-1}\widehat{\mathbf{g}_1^{*}} = \mathbf{T}\widehat{\mathbf{g}_1^{*}}$$

where $\mathbf{B}_1^{**}$ is a diagonal matrix (e.g., an identity matrix $\mathbf{I}$) with optional different weights to differentially shrink different loci.

Based on the properties of mixed models (Henderson, 1984), the prediction error covariance matrix of $\widehat{\boldsymbol{\alpha}_1^{**}}$, $PEC\left(\widehat{\boldsymbol{\alpha}_1^{**}}\right)$, can be obtained as follows:

$$\begin{aligned}
PEC\left(\widehat{\boldsymbol{\alpha}_1^{**}}\right) &= Var(\boldsymbol{\alpha}_1^{**}) - Var\left(\widehat{\boldsymbol{\alpha}_1^{**}}\right) = Var(\boldsymbol{\alpha}_1^{**}) - Var\left(\mathbf{T}\widehat{\mathbf{g}_1^{*}}\right) = Var(\boldsymbol{\alpha}_1^{**}) - \mathbf{T}Var\left(\widehat{\mathbf{g}_1^{*}}\right)\mathbf{T}' \\
&= Var(\boldsymbol{\alpha}_1^{**}) - \mathbf{T}\left(Var(\mathbf{g}_1^{*}) - PEC\left(\widehat{\mathbf{g}_1^{*}}\right)\right)\mathbf{T}' \\
&= Var(\boldsymbol{\alpha}_1^{**}) - \mathbf{T}\left(\boldsymbol{\Gamma}^{*}Var(\boldsymbol{\alpha}_1^{*})\boldsymbol{\Gamma}^{*\prime} - \boldsymbol{\Gamma}^{*}PEC\left(\widehat{\boldsymbol{\alpha}_1^{*}}\right)\boldsymbol{\Gamma}^{*\prime}\right)\mathbf{T}' \\
&= Var(\boldsymbol{\alpha}_1^{**}) - \mathbf{T}\boldsymbol{\Gamma}^{*}\left(Var(\boldsymbol{\alpha}_1^{*}) - PEC\left(\widehat{\boldsymbol{\alpha}_1^{*}}\right)\right)\boldsymbol{\Gamma}^{*\prime}\mathbf{T}'
\end{aligned}$$