Extensions of BLUP models for genomic prediction in heterogeneous populations: Application in a diverse switchgrass sample

Guillaume P. Ramstein[*], Michael D. Casler[*,§]

[*] Department of Agronomy, University of Wisconsin-Madison, Madison, WI 53706, USA

[§] Agricultural Research Service, United States Department of Agriculture, Madison, WI 53706, USA

SHORT TITLE

Genomic prediction in heterogeneous populations

KEYWORDS

Genomic prediction, population admixture, instance selection, multi-population model, *Panicum virgatum* L.

CORRESPONDING AUTHOR

Guillaume Ramstein
UW-Madison
Department of Agronomy
1575 Linden Drive
Madison, WI 53706
Tel.: + 1 608 890 0050
Fax: + 1 608 262 5217
E-mail: ramstein@wisc.edu

ABSTRACT

Genomic prediction is a useful tool to accelerate genetic gain in selection using DNA marker information. However, this technology usually relies on models that are not designed to accommodate population heterogeneity, which results from differences in marker signals across genetic backgrounds. Previous studies have proposed to cope with population heterogeneity using diverse approaches: (i) either ignoring it, therefore relying on the robustness of standard approaches; (ii) reducing it, by selecting homogenous subsets of individuals in the sample; or (iii) modelling it by using interactive models. In this study we assessed all three possible approaches, applying existing and novel procedures for each of them. All procedures developed are based on deterministic optimizations, can account for heteroscedasticity, and are applicable in contexts of admixed populations. In a case study on a diverse switchgrass sample, we compared the procedures to a control where predictions rely on homogeneous subsamples. Ignoring heterogeneity was often not detrimental, and sometimes beneficial, to prediction accuracy, compared to the control. Reducing heterogeneity did not result in further increases in accuracy. However, in scenarios of limited subsample sizes, a novel procedure, which accounted for redundancy within subsamples, outperformed the existing procedure, which only considered relationships to selection candidates. Modelling heterogeneity resulted in substantial increases in accuracy, in the cases where accounting for population heterogeneity yielded a highly significant improvement in fit. Our study exemplifies advantages and limits of the various approaches that are promising in various contexts of population heterogeneity, e.g. prediction based on historical datasets or dynamic breeding.

INTRODUCTION

Genomic prediction has proved a useful tool to predict genetic merit in plant and animal breeding (Hayes *et al.* 2009a, Lorenz *et al.* 2011). This technology consists of learning relationships between DNA markers and phenotypes, which arise from the non-random association (linkage disequilibrium; LD) between DNA markers and causal genetic variants having direct effects on the trait studied (Meuwissen *et al.* 2001). Typical genomic prediction models, including genomic BLUP (GBLUP; VanRaden 2008, Hayes *et al.* 2009b) or Bayesian linear regression (BLR) models (Meuwissen *et al.* 2001, Gianola *et al.* 2009), assume that the effects of causal variants are linear and purely additive, so estimated effects do not capture any dependence on context, arising for example from interactions of causal variants with environmental or genetic backgrounds. Initially, genomic prediction models have been proposed for applications in populations that are relatively homogeneous with respect to LD patterns and interactions involving causal variants (Meuwissen *et al.* 2001). In such situations, increasing the size of the calibration set (CS) – the set of individuals used to estimate the model's parameters – would typically benefit accuracy of the models (Lorenzana & Bernardo 2009, VanRaden *et al.* 2009). However, in practice, increasing the CS size may often involves calibrating prediction models on individuals with inconsistent LD patterns and/or backgrounds, which may result in reduced accuracy (Wientjes *et al.* 2016). This issue will arise in the typical situation where an initially homogeneous CS is augmented with individuals from extraneous populations, that is, multi-population – or (in the animal literature) multi-breed – calibration (Lund *et al.* 2014). Recently, studies in both plant and animal breeding have assessed the usefulness of combining populations from different genetic backgrounds in genomic prediction. In general, one or two of the following approaches were studied: (i) single-population prediction in a multi-population context (ignoring population heterogeneity); (ii) instance selection (reducing population heterogeneity); and (iii) multi-population prediction (modelling population heterogeneity).

In single-population prediction (*SPM*), the simulation study of De Roos *et al.* (2009) suggested that adding an extraneous population to a CS may benefit prediction accuracy if the added population is not too dissimilar (in terms of divergence time) from the initial CS. These authors also suggested that high enough marker density could prevent prediction accuracy from decreasing, even in cases of strong divergence between populations. Consistently, most empirical studies of multi-population calibration with high marker density, based on single-population BLR and/or GBLUP, have reported little or no gain in accuracy under strong population structure (Lehermeier *et al.* 2015, Jarquín *et al.* 2016, Hayes *et al.* 2009c, Erbe *et al.* 2012). In contrast, only a few studies have reported substantial increases in accuracy from multi-population calibration in similar conditions (Technow *et al.* 2013, Daetwyler *et al.* 2012). Interestingly, Habier *et al.* (2013) suggested that increasing CS size may reduce prediction accuracy in GBLUP, even in a single-population context, due to accumulated noise in the larger genomic relationship matrix, especially when many relationship coefficients are small. In an attempt to increase the accuracy of genomic relationship estimation, Endelman and Jannink (2012) and Müller *et al.* (2015) have proposed regularization methods, which proved especially useful when marker density was low. However, the regularization methods proposed in these two studies did not account for potential population structure in the genomic relationship matrix, which would naturally arise in a multi-population context.

In instance selection (*IS*) – or training set design/optimization – only a subset of the available individuals is selected to make up the CS. Studies have generally focused on a scenario of

limited phenotyping resources, where the sample of individuals was searched for an optimal CS of pre-determined size. The CS searches in these studies were either stochastic or deterministic. Stochastic searches in this context have consisted in randomly choosing a CS to maximize some measure of prediction accuracy for the selection candidates, using either random exchange algorithms (Rincent *et al.* 2012, Isidro *et al.* 2014, Rutkoski *et al.* 2015) or genetic algorithms (Akdemir *et al.* 2015), which were compared to purely random sampling as a baseline. Studies of this type have used selection criteria such as the prediction error variance (Henderson 1984) or the mean coefficient of determination (Laloë 1993) as a measure of accuracy, and have generally concluded that stochastic searches guided by one of these criteria performed better than random sampling. One disadvantage of stochastic searches is that they are computationally intensive, so deterministic searches may be preferred in some scenarios (e.g., when sample size is large). This second type of searches has typically involved choosing the set of individuals so as to maximize some measure of relatedness between the CS and the selection candidates (Clark *et al.* 2012, Lorenz and Smith 2015). The contribution of such relatedness to accuracy has been asserted by simulation studies (Pszczola *et al.* 2012, Wientjes *et al.* 2013). However, Pszczola *et al.* (2012) also suggested that accuracy was negatively impacted by relationships within the CS, for a given CS size (probably owing to redundancy in information). To our knowledge, no deterministic search in genomic prediction has accounted for that trade-off involving relationships.

In multi-population prediction (*MPM*), studies have proposed to fit, to the whole set of available individuals, models that were capable of accommodating population heterogeneity explicitly. This type of models includes multi-trait GBLUP models, with "traits" corresponding to population backgrounds (Karoui *et al.* 2012, Carillier *et al.* 2014, Lehermeier *et al.* 2015), and random regression models based on markers interacting with discrete population cluster coefficients (de los Campos *et al.* 2015, with a BLR model). To our knowledge, the implementation of these methods has not been adapted to contexts of admixture, where population structure variables are continuous. Furthermore, when calibration involves many populations, the increase in model complexity of these methods will make them computationally intractable and statistically inefficient. Parsimonious multi-population models, based on only a few parameters to capture population heterogeneity, have also been proposed (Zhou *et al.* 2014, Heslot and Jannink 2015). In presence of many populations, such models are more practical and potentially more useful than multi-trait and random interaction models. Also, since they generally assume some underlying basis for population heterogeneity (e.g., inconsistency in LD patterns), they may generate insight about the causes of marker-by-population interactions.

In this study, we investigated the usefulness of *SPM*, *IS* and *MPM* for coping with population heterogeneity. We present a general framework for the application of existing and novel methods under each of these three approaches. All these procedures were compared to a control procedure (*Target*) where the CS includes only the individuals from the same population as the selection candidates, as is typically done to avoid dealing with population heterogeneity. We applied the procedures to the analysis of three traits (plant height, heading date, and standability) in switchgrass (*Panicum virgatum* L.), an herbaceous biomass crop showing good promise for biofuel production (Sanderson *et al.* 1996, Perlack *et al.* 2005, Perlack *et al.* 2011, Langholtz *et al.* 2016). The present work describes promising methods for increasing accuracy and robustness of predictions in situations where heterogeneous data sources are combined, for example when the CS incorporates data from historical trials (Dawson *et al.* 2013, Rutkoski *et al.* 2015) or from multiple generations of a dynamic breeding program (Sallam *et al.* 2015, Auinger *et al.* 2016).

MATERIAL AND METHODS

Panels and populations

In this study, two multi-population panels were assayed and considered together in one sample. The first panel was the breeding panel (BP) described in Ramstein *et al.* (2016), comprising two tetraploid breeding populations of half-sib (HS) families: WS4U-C2, which consisted of 137 HS families derived from a diverse upland-ecotype pool of 162 plants (Casler *et al.* 2006), and Liberty-C2, which consisted of 110 HS families derived from the lowland-upland cultivar Liberty (Casler and Vogel 2014). The second panel was the association panel (AP) described in Lu *et al.* (2013) and Evans *et al.* (2015), comprising six putative populations of clonally propagated genotypes of different ecotypes (U: upland; L: lowland), ploidy levels (4X: tetraploid; 8X: octoploid) and geographical origins (S: South; W: West; N: North; E: East): U4X-N (135 plants), U8X-W (129 plants), U8X-E (97 plants), U8X-S (10 plants), L4X-NE (106 plants) and L4X-S (37 plants). These populations corresponded to 60 diverse accessions (Lu *et al.* 2013, Evans *et al.* 2015) with up to 10 individuals per accessions.

In WS4U-C2, one individual was discarded so as to avoid assigning it to a population in AP, since it was too distantly related to the other individuals in BP (based on principal component analysis). In total, $n = 760$ individuals were considered in this analysis. The main goal of this study was to assess different methods for accommodating genetic heterogeneity when predicting phenotypic means in a given target population. Four targets were chosen, with a defined focus on tetraploid populations with at least 100 relatively homogeneous individuals: WS4U-C2 and Liberty-C2 (from BP), and U4X-N and L4X-NE (from AP).

Marker data

Genotyping of individuals (parents in BP and clonally propagated plants in AP) was performed by exome capture sequencing. Single nucleotide polymorphisms (SNPs) were called at 2,179,164 biallelic loci (Hapmap v2), as described for BP by Ramstein *et al.* (2016) and for AP by Evans *et al.* (2014, 2015). Marker genotypes were then called by using the expectation-maximization algorithm of Martin *et al.* (2010) fitted in each population separately, under the assumption of disomic inheritance. Although this assumption is supported in switchgrass for tetraploid genotypes (Okada *et al.* 2010; Li *et al.* 2014), it does not hold for octoploid genotypes, which would presumably exhibit tetrasomic inheritance. However, we did not adapt the algorithm of Martin *et al.* (2010) to accommodate possible tetrasomic inheritance, as sequencing depth was deemed insufficient for calling intermediate heterozygotes (simplex and triplex) with high enough accuracy.

The resulting marker-data matrix consisted of expected allelic dosages (sums alternate-allele counts weighted by their posterior probabilities, for every individual and SNP). The SNPs were then filtered based on the following criteria: (i) proportion of missing values strictly lower than 2%; (ii) minor allele frequency (MAF) strictly greater than $1/2n$ and variance strictly greater than $2(1/2n)(1 - 1/2n)$; (iii) p-value for Hardy-Weinberg equilibrium (HWE) strictly greater than 10^{-4} in each BP population; (iv) availability of genomic-location information (as per version 1.1 of the reference genome of *P. virgatum*; DOE-JGI, http://phytozome.jgi.doe.gov/). Missing values at SNPs were imputed by their mode in the whole sample. The resulting $n \times m$ filtered and imputed marker-data matrix $\mathbf{X}$ consisted of expected allelic dosages at $m = 717{,}814$ markers.

Phenotypic data

Populations in BP were assayed each year between 2012 and 2014, in Arlington, WI (USA), in a randomized complete block design, with four replicates for WS4U-C2 and three replicates for Liberty-C2. Populations in AP were assayed each year between 2009 and 2011 in Ithaca, NY (USA), in a sets-in-reps design, with two replicates per individual and 10 sets within each replicate, with each set comprising at most one individual from each of the 60 accessions in AP (Lu *et al.* 2013, Evans *et al.* 2015). In each panel, three phenotypic traits were considered: plant height, heading date and standability. Plant height (PH) was measured in centimeters as the height from the ground to the top of the tallest tiller. Heading date (HD) was measured in growing degrees days as the cumulated sum of daily average temperatures (in degrees Celsius; °C) above 10 °C, from January 1st to the day of heading, defined as the emergence of at least half of the panicles from the boot (Mitchell *et al.* 1997); daily average temperatures were estimated by the average of the minimum and maximum daily temperatures. Standability (St) was measured on a 0-10 scale to describe plants' stature and stiffness, with 0 qualifying plants that are prostrate and 10 qualifying upright and rigid plants (Lipka *et al.* 2014).

Not all traits were measured every year in any given population: only HD was measured in all three years in AP populations and Liberty-C2. For all other cases, measurements were available for only a subset of years (Table 1).

In BP, observational units were half sibs from a given genotype (maternal parent) i; so the following model was fitted to phenotypic measurements P_{ijkl}, to estimate HS family means f_i's:

$$P_{ijkl} = \mu + f_i + b_j + t_k + (f \times b)_{ij} + (f \times t)_{ik} + (b \times t)_{jk} + (f \times b \times t)_{ijk} + \varepsilon_{ijkl}$$

where μ is the grand mean; f_i, b_j and t_k are the effects of HS family i (fixed), block j (random) and year k (random) respectively; $\times$ indicates interactions (random); ε_{ijkl} are residuals. For each random term, the corresponding effects were modeled as independent and identically normally distributed.

In AP, observational units were clones of a given genotype i; so the following model was fitted to measurements P_{ijk} to estimate centered genotype means g_i's:

$$P_{ijk} = \mu + g_i + b_j + t_k + (g \times b)_{ij} + (g \times t)_{ik} + (b \times t)_{jk} + e_{ijk}$$

where effects are as described above, except for e_{ijk}, which is the error for clone ij in year k.

The linear mixed models described above were fitted using ASREML-R (Butler *et al.* 2009).

Effects f_i's are transmitted abilities of genotypes, so that $f_i = \frac{BV_i}{2}$, where BV_i is the breeding value of genotype i. In comparison, effects g_i's are genotypic values, such that $g_i = BV_i + \Delta_i$, where Δ_i is the deviation from additivity due to dominance and/or epistasis. Outcomes of interest for genomic prediction were set to be non-centered means y_i's such that $y_i = \hat{\mu} + 2\hat{f}_i$ in BP and $y_i = \hat{\mu} + \hat{g}_i$ in AP.

Population structure data

Admixture analysis

The soft clustering model from the ADMIXTURE software was fitted on the whole sample and the whole set of SNPs, i.e., without selection on individuals or markers (Alexander *et al.* 2009). Based on the 10-fold cross-validation implemented in ADMIXTURE (Alexander *et al.* 2011),

the number of population clusters in the admixture model was set to $K = 7$, as cross-validation error reached a plateau at that value (Figure S1). The resulting $n \times K$ matrix $\mathbf{A}$ of admixture coefficients comprised inferred membership probabilities at each cluster (Figure 1a). For convenience (in prediction models), minimum values in $\mathbf{A}$ (10^{-5}) were set to zero while ensuring that each row still summed to one.

Principal component analysis

Principal component analysis (PCA) was performed on the whole sample and the whole set of SNPs. The number of principal components (PC) to choose for depicting population structure was chosen based on the proportion of variance explained and the grouping patterns captured by PCs (Figure 1b). The resulting $n \times d$ PC matrix $\mathbf{P}$ consisted of coordinates for each individual at the first $d = 4$ PCs.

Genomic prediction models

All linear mixed models described below were fitted using the R package *rrBLUP* (Endelman 2011).

For a given marker-data matrix $\mathbf{X}$ and vector $\mathbf{y}$ of outcomes, the standard ridge regression BLUP model (RR-BLUP; BLUP: best linear unbiased predictor) is described as follows:

$$\mathbf{y} = \mathbf{Q\alpha} + \mathbf{X\beta} + \mathbf{e};\ \begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{e}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{I}_m\sigma_\beta^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}\right)$$

where $\mathbf{y}$ is the n-vector of outcomes (y_i's, as described above); $\mathbf{X}$ is the $n \times m$ marker-data matrix, and $\boldsymbol{\beta}$ is the m-vector of marker effects, assumed independent with variance σ_β^2 ($\mathbf{I}_m$ is the $m \times m$ identity matrix); $\mathbf{Q}$ is the $n \times p$ matrix depicting the population mean structure in the sample, and $\boldsymbol{\alpha}$ is the p-vector of associated effects; $\mathbf{R}$ is the covariance matrix of errors, possibly accommodating correlations and differences in variance (heteroscedasticity) among errors. Often, errors are considered to be independent and identically distributed, such that $\mathbf{R} = \mathbf{I}_n\sigma_e^2$, with $\mathbf{I}_n$ the $n \times n$ identity matrix and σ_e^2 the error variance.

Let $\mathbf{u} = \mathbf{X\beta}$, so that $\mathrm{Var}(\mathbf{u}) = \mathbf{XX}'\sigma_\beta^2$, by identical mean structure $\mathbf{Q\alpha}$ and variance structure $\mathrm{Var}(\mathbf{y}) = \mathbf{XX}'\sigma_\beta^2 + \mathbf{R}$ (Henderson, 1984), the RR-BLUP model is equivalent to the following genomic BLUP (GBLUP) model:

$$\mathbf{y} = \mathbf{Q\alpha} + \mathbf{u} + \mathbf{e};\ \begin{bmatrix}\mathbf{u}\\ \mathbf{e}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{XX}'\sigma_\beta^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}\right) \qquad (1a)$$

In the RR-BLUP model, regressing out the mean-structure matrix $\mathbf{Q}$ from $\mathbf{X}$ yields the following equivalent model, where mean- and variance-structure matrices are orthogonal, i.e. columns from one matrix to another are now uncorrelated (see Appendix A1 for a general proof):

$$\mathbf{y} = \mathbf{Q}\dot{\boldsymbol{\alpha}} + (\mathbf{I}_n - \mathbf{H})\mathbf{X\beta} + \mathbf{e};\ \begin{bmatrix}\boldsymbol{\beta}\\ \mathbf{e}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{I}_m\sigma_\beta^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}\right)$$

where $\mathbf{H} = \mathbf{Q}(\mathbf{Q}'\mathbf{R}^{-1}\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{R}^{-1}$ is the matrix of projection onto the column space of $\mathbf{Q}$; $\dot{\mathbf{X}} = (\mathbf{I} - \mathbf{H})\mathbf{X}$ is the adjusted matrix of (residual) marker variables, made orthogonal to $\mathbf{Q}$; $\dot{\boldsymbol{\alpha}}$ is the new vector of fixed effects. With $\mathbf{Q} = \mathbf{1}_n$ ($\mathbf{1}_n$ is a n-vector of ones) and $\mathbf{R} = \mathbf{I}_n$, the mean structure $\mathbf{Q}\dot{\boldsymbol{\alpha}}$ is simply an intercept and $\dot{\mathbf{X}}$ is the matrix of marker variables centered around their

respective mean, as often used in genomic prediction studies (Hayes *et al.* 2009b, de los Campos *et al.* 2013). With general $\mathbf{Q}$, the mean structure $\mathbf{Q}\dot{\boldsymbol{\alpha}}$ is an individual-specific mean for $\mathbf{y}$ with respect to the specific population membership of each individual and $\dot{\mathbf{X}}$ is the matrix of marker variables centered around their respective individual-specific means.

By identical mean and variance structures, the previous model is equivalent to the following alternate GBLUP model:

$$\mathbf{y} = \mathbf{Q}\dot{\boldsymbol{\alpha}} + \dot{\mathbf{u}} + \mathbf{e};\ \begin{bmatrix}\dot{\mathbf{u}}\\ \mathbf{e}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{G}\sigma^2_\beta & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}\right) \qquad (1b)$$

where $\dot{\mathbf{u}} = (\mathbf{I}_n - \mathbf{H})\mathbf{X}\boldsymbol{\beta}$ is the *n*-vector of genomic breeding values centered around population means $\mathbf{Q}\dot{\boldsymbol{\alpha}}$, and $\mathbf{G} = \dot{\mathbf{X}}\dot{\mathbf{X}}'$ is the genomic relationship matrix for $\dot{\mathbf{u}}$ (here unscaled).

Following the recommendations of Phocas and Laloë (2004), we chose to simply use $\mathbf{Q} = \mathbf{1}_n$ to define the mean structure in all fitted models. Also, we chose not to model heteroscedasticity of errors and used $\mathbf{R} = \mathbf{I}_n\sigma^2_e$. Therefore, the covariance matrix $\mathbf{G} = \dot{\mathbf{X}}\dot{\mathbf{X}}'$ was simply proportional to the standard genomic relationship matrix $\mathbf{G_1} = \dot{\mathbf{X}}\dot{\mathbf{X}}'/v$ of VanRaden (2008), where $v = 2\sum_{l=1}^{m}\hat{\pi}_l(1-\hat{\pi}_l)$ is a scaling factor depending on estimated allele frequencies $\hat{\pi}_l$'s. Notably, the matrix $\mathbf{G}$ derived here will account for correlations and heteroscedasticity of errors, whenever $\mathbf{R} \neq \mathbf{I}_n\sigma^2_e$ (the projector $\mathbf{H}$ is a function of $\mathbf{R^{-1}}$). To our knowledge, the matrix $\mathbf{G_1}$ has typically been used in GBLUP models, even when $\mathbf{R} \neq \mathbf{I}_n\sigma^2_e$ as in weighted GBLUP models on deregressed proofs.

Optimization methods

Hereafter, the testing TS is defined as the set of individuals left out for model validation. The calibration set CS is the set of individuals used to fit the prediction models, which excludes TS but does not necessarily consists of all remaining (available) individuals.

In this study, we adapted model (1a) or (1b) to four different approaches: (i) the control procedure, consisting in including in the CS only individuals from the same target group as the testing set; (ii) single-population models, consisting in fitting a GBLUP model to all available individuals, with possibly some regularization on genomic relationships; (iii) instance selection, consisting in including a subset of available individuals in the CS, so as to optimize some selection criterion; and (iv) multi-population models, consisting of modelling population heterogeneity for the outcome on all available individuals, based on population structure data (PCs in **P** or admixture coefficients in **A**).

Control procedure (Target)

In the control procedure (*Target*), we fitted model (1a), with the CS restricted to individuals belonging to the same population as the TS. This method corresponds to a typical choice of relying only on individuals that have genetic architectures that are *a priori* similar to those in the TS.

Single population models (SPM-GRM, SPM-GLASSO)

Here, single-population models (*SPM*) are defined as the basic models incorporating information from all available individuals, with no modelling of population heterogeneity. The following general model was fitted:

$$\mathbf{y} = \mathbf{Q}\tilde{\boldsymbol{\alpha}} + \tilde{\mathbf{u}} + \tilde{\mathbf{e}};\ \tilde{\mathbf{u}} \sim N\left(\mathbf{0}, \tilde{\mathbf{G}}\sigma^2_{\tilde{\mathbf{u}}}\right),\ \tilde{\mathbf{e}} \sim N\left(\mathbf{0}, \tilde{\mathbf{R}}\right) \qquad (2)$$

where $\widetilde{\mathbf{G}}$ is some matrix depicting relationship among breeding values $\tilde{\mathbf{u}}$.

We considered two types of matrices for $\widetilde{\mathbf{G}}$: the original unscaled genomic relationship matrix $\mathbf{G} = \dot{\mathbf{X}}\dot{\mathbf{X}}'$ (*SPM-GRM*) and a regularized form of $\mathbf{G}$, where relationships are shrunk for potentially higher estimation accuracy of relationships (*SPM-GLASSO*).

In *SPM-GRM*, since $\widetilde{\mathbf{G}} = \mathbf{G}$, model (2) was equivalent to model (1b), so this approach simply corresponded to fitting a GBLUP model to all available individuals.

In *SPM-GLASSO*, following Fan *et al.* (2013), $\widetilde{\mathbf{G}}$ was decomposed as $\widetilde{\mathbf{G}} = \mathbf{B}\mathbf{B}' + \widetilde{\mathbf{G}}_B$, where $\mathbf{B}$ consisted of the first *t* PCs of $\mathbf{G}$ and $\widetilde{\mathbf{G}}_B$ was a regularized form of $\mathbf{G}_B = \mathbf{G} - \mathbf{B}\mathbf{B}'$. Matrix $\mathbf{B}\mathbf{B}'$ is the dense part of the relationship matrix $\mathbf{G}$, representing resemblance among individuals through common structural factors. Here this matrix depicted relationships at the population level, through the *t* leading PCs of $\mathbf{G}$. In contrast, matrix $\mathbf{G}_B$ represented (recent) relationships conditional on population structure, similarly to the adjusted relationships introduced by Thornton *et al.* (2012) and Conomos *et al.* (2016), with the difference that here coefficients in $\mathbf{G}_B$ are not scaled for direct estimation of recent-kinship coefficients. In principle, most coefficients in $\mathbf{G}_B$ should be close to zero, as there should exist only few familial relationships within the sample. So $\mathbf{G}_B$ may be assumed to be sparse, which is often an important property for useful regularization of covariance matrices. Fan et *al.* (2013) suggested that matrix $\mathbf{G}_B$ be regularized by adaptive thresholding (Cai and Liu 2011). However, we chose to perform regularization by the graphical LASSO (Friedman *et al.* 2008) so as to shrink coefficients in $\mathbf{G}_B$ while inferring a sparse precision (inverse covariance) matrix $\widetilde{\mathbf{G}}_B^{-1}$, which yielded a sparse graph of relationships among individuals (a zero *ij*-element in $\widetilde{\mathbf{G}}_B^{-1}$ indicates that individuals *i* and *j* are unrelated conditionally on all other individuals, which corresponds to no edge between nodes *i* and *j* in the underlying graph of recent relationships).

The graphical LASSO infers a sparse precision matrix $\mathbf{\Sigma}^{-1}$ by maximizing the Gaussian likelihood of the data, penalized by a L_1-norm penalty $\lambda\|\mathbf{\Sigma}^{-1}\|_1$, where λ is the regularization parameter and $\|\mathbf{\Sigma}^{-1}\|_1$ is the sum of absolute values in $\mathbf{\Sigma}^{-1}$ (Friedman *et al.* 2008).

Regularization of $\mathbf{G}$ was performed as follows:
1. Performing eigenvalue decomposition on $\mathbf{G}$ to obtain $\mathbf{B}$ and decompose $\mathbf{G}$ into $\mathbf{B}\mathbf{B}' + \mathbf{G}_B$
2. Standardizing $\mathbf{G}_B$ to obtain the corresponding correlation matrix $\mathbf{\Gamma}_B$: $\mathbf{\Gamma}_B = \mathrm{diag}(\mathbf{G}_B)^{-1/2}\mathbf{G}_B\mathrm{diag}(\mathbf{G}_B)^{-1/2}$
3. Applying the graphical LASSO algorithm to $\mathbf{\Gamma}_B$, to obtain the regularized correlation matrix $\tilde{\mathbf{\Gamma}}_B$
4. Rescaling $\mathbf{\Gamma}_B$ to obtain $\widetilde{\mathbf{G}}_B = \mathrm{diag}(\mathbf{G}_B)^{1/2}\tilde{\mathbf{\Gamma}}_B\mathrm{diag}(\mathbf{G}_B)^{1/2}$ and then $\widetilde{\mathbf{G}} = \mathbf{B}\mathbf{B}' + \widetilde{\mathbf{G}}_B$

The graphical LASSO algorithm was run using the huge package in R (Zhao *et al.* 2012). The regularization parameter λ was determined so as to maximize the restricted maximum likelihood (REML) of model (2) in a grid search, with λ being the *q*-quantile of absolute values in $\mathbf{\Gamma}_B$ and *q* varying from 0.05 to 1 by step of 0.05.

With $\mathbf{Q} = \mathbf{1}_n$, the number *t* of PCs in $\mathbf{B}$ was set to $d = 4$ (*d* is the number of PCs chosen to reflect population structure in $\mathbf{P}$). However, when $\mathbf{Q}$ actually depicts population structure, e.g. when $\mathbf{Q} = [\mathbf{1}_n \quad \mathbf{P}]$ or $\mathbf{Q} = \mathbf{A}$, the matrix $\mathbf{G}$ already reflects relationships among individuals conditionally on population structure (so $\mathbf{G}$ should already be sparse), and *t* may simply be set to

zero (i.e., regularization may be performed on **G** directly). Notably, when $\mathbf{Q} = \mathbf{1_n}$, $\mathbf{R} = \mathbf{I}_n\sigma_e^2$ and the CS consists of the whole sample, **B** simply equals **P**.

Instance selection (IS-Rel, IS-QP)

In instance selection (*IS*), we fitted model (1a) on a subset of all available individuals. We selected individuals deterministically (i.e., without using random searches through possible calibration sets) by first including individuals with highest scores (as defined below), so as to optimize a selection criterion. We chose to maximize the mean coefficient of determination $\mathrm{CD}_{\mathrm{mean}}$ (Laloë 1993) for the TS, with no contrast so that this selection criterion simply corresponded to the model-based estimate of the mean squared prediction accuracy (reliability) with respect to **u** in the TS, i.e. $\mathrm{CD}_{\mathrm{mean}} = \frac{1}{|TS|}\sum_{j\in TS} Cor\left(u_j, \hat{u}_j\right)^2$, with $|TS|$ the size of the TS, $Cor\left(u_j, \hat{u}_j\right)^2 = 1 - \frac{Var(\hat{u}_j - u_j)}{Var(u_j)}$, where $\hat{u}_j$ is the BLUP of u_j, $Var(\hat{u}_j - u_j)$ and $Var(u_j)$ are inferred from the fitted model fit (Searle *et al.* 2006).

We considered two types of scores (w_i's), for two different procedures: *IS-Rel* and *IS-QP*.

In *IS-Rel*, $w_i = \frac{1}{|TS|}\sum_{j\in TS} g_{ij}$, with g_{ij} being the *ij*-element of $\mathbf{G} \propto \dot{\mathbf{X}}\dot{\mathbf{X}}'$. So w_i simply reflected the average relationship between individual *i* and the TS.

In *IS-QP*, we inferred the scores w_i's on all available individuals, so as to minimize the difference between the average genotype in the TS and the weighted average of genotypes in remaining individuals (TS^{C}), with weights w_i's. Formally, let $\mathbf{w} = (w_i)_{i\in \mathrm{TS}^c}$ such that $w_i \geq 0$ for all $i \in \mathrm{TS}^c$ and $\sum_{i\in \mathrm{TS}^c} w_i = 1$, we minimized $\left\|\frac{1}{|\mathrm{TS}|}\sum_{j\in \mathrm{TS}} \mathbf{x}_j - \sum_{i\in \mathrm{TS}^c} w_i\mathbf{x}_i\right\|_2 = \left\|\mathbf{X}'_{\mathrm{TS}}\mathbf{1}_{|\mathrm{TS}|}/|\mathrm{TS}| - \mathbf{X}'_{\mathrm{TS}^c}\mathbf{w}\right\|_2$, with $\|.\|_2$ being the Euclidean norm, and subscripts referring to subsets on rows in vectors or matrices ($\mathbf{x}_i$ refers to the *m*-vector of marker variables for individual *i*). Equivalently, we minimized $\left\|\mathbf{X}'_{\mathrm{TS}}\mathbf{1}_{|\mathrm{TS}|}/|\mathrm{TS}| - \mathbf{X}'_{\mathrm{TS}^c}\mathbf{w}\right\|_2^2 = \left(\mathbf{1}'_{|\mathrm{TS}|}\mathbf{X}_{\mathrm{TS}}/|\mathrm{TS}| - \mathbf{w}'\mathbf{X}_{\mathrm{TS}^c}\right)\left(\mathbf{X}'_{\mathrm{TS}}\mathbf{1}_{|\mathrm{TS}|}/|\mathrm{TS}| - \mathbf{X}'_{\mathrm{TS}^c}\mathbf{w}\right) = \frac{1}{|\mathrm{TS}|^2}\mathbf{1}'_{|\mathrm{TS}|}\mathbf{X}\mathbf{X}'_{\mathrm{TS,TS}}\mathbf{1}_{|\mathrm{TS}|} - \frac{2}{|\mathrm{TS}|}\mathbf{w}'\mathbf{X}\mathbf{X}'_{\mathrm{TS}^c,\mathrm{TS}}\mathbf{1}_{|\mathrm{TS}|} + \mathbf{w}'\mathbf{X}\mathbf{X}'_{\mathrm{TS}^c,\mathrm{TS}^c}\mathbf{w}$. Since the first term in the last sum is constant with respect to w_i's, **w** solved the following quadratic programming (QP) problem: minimizing $\frac{1}{2}\mathbf{w}'\mathbf{X}\mathbf{X}'_{\mathrm{TS}^c,\mathrm{TS}^c}\mathbf{w} - \frac{1}{|\mathrm{TS}|}\mathbf{1}'_{|\mathrm{TS}|}\mathbf{X}\mathbf{X}'_{\mathrm{TS,TS}^c}\mathbf{w}$ subject to $w_i \geq 0$ for all $i \in \mathrm{TS}^c$ and $\sum_{i\in \mathrm{TS}^c} w_i = 1$. This problem is similar to the general QP problem for feature selection introduced by Rodriguez-Lujan *et al.* (2010), i.e., minimizing $\frac{1}{2}(1-\alpha)\mathbf{w}'\mathbf{Q_r}\mathbf{w} - \alpha\mathbf{f}'_{\mathbf{R}}\mathbf{w}$, subject to $w_i \geq 0$ for all *i* and $\sum_i w_i = 1$, where vector $\mathbf{f_R}$ measures relevance of features with respect to a given outcome; matrix $\mathbf{Q_r}$ measures the redundancy among features; and $\alpha \in [0,1]$ sets the relative importance of each term in the sum. The QP problem could have been defined freely, but our initial motivation allowed us to naturally set $\mathbf{Q_r} = \mathbf{X}\mathbf{X}'_{\mathrm{TS}^c,\mathrm{TS}^c}$, $\mathbf{f_R} = \frac{1}{|\mathrm{TS}|}\mathbf{X}\mathbf{X}'_{\mathrm{TS}^c,\mathrm{TS}}\mathbf{1}_{|\mathrm{TS}|}$ and $\alpha = \frac{1}{2}$. Compared to *IS-Rel*, *IS-QP* has two advantages: a solution **w** will represent a compromise between relevance (average of relationships $\frac{1}{|\mathrm{TS}|}\mathbf{1}'_{|\mathrm{TS}|}\mathbf{X}\mathbf{X}'_{\mathrm{TS,TS}^c}$ from TS^c to TS) and redundancy (relationships $\mathbf{X}\mathbf{X}'_{\mathrm{TS}^c,\mathrm{TS}^c}$ within TS^c), so scores from *IS-QP* should favor more diverse sets of individuals compared to *IS-Rel*; also, relationships used are not adjusted by allele frequencies (equivalently, they do not depend

on any projector **H**), so they could be less prone to misrepresentations of relationships, through inappropriate centering of marker variables.

The QP problem in *IS-QP* was solved using the R package *quadprog* (https://CRAN.R-project.org/package=quadprog). In optimization, we considered selecting 5% to 100% of individuals in CS, by step of 5%, selecting the subset of individuals which maximized CD_{mean}.

The prediction accuracy of *IS* methods was assessed with free CS sizes (i.e., as optimization procedures), but was also assessed with fixed CS sizes, with no selection of subset based on CD_{mean}: for a given CS size |CS|, only the first |CS| individuals with the highest scores w_i's were included in the CS. Such assessments were intended to reflect the usefulness of *IS* procedures in conditions of limited resources for phenotyping (and calibration of prediction models). In this context, *IS-Rel* and *IS-QP* were compared to random selection (*RS*), which simply consisted in randomly selecting |CS| individuals for model fitting. For a given combination of TS, |CS| and outcome, the accuracy from *RS* was the average of accuracies over 200 random draws. Each draw corresponded to one random attribution of scores (w_i's) to individuals.

Mixed population models (MPM-Mixture, MPM-Matérn)
Mixed-population models (*MPM*) are extensions of model (1a) intended to accommodate population heterogeneity. The following general model was fitted:

$$\mathbf{y} = \mathbf{Q}\boldsymbol{\alpha} + \mathbf{u} + \mathbf{e};\ \mathbf{u} \sim N\left(\mathbf{0}, (\boldsymbol{\Omega}_n \circ \mathbf{X}\mathbf{X}')\sigma_\beta^2\right),\ \mathbf{e} \sim N(\mathbf{0}, \mathbf{R}) \qquad (3)$$

where $\circ$ is the element-wise (Hadamard) product, and $\boldsymbol{\Omega}_n$ is a *n* x *n* covariance matrix depicting population differentiation among individuals (see Appendix A2 for derivations and technical details). To parsimoniously estimate $\boldsymbol{\Omega}_n$, we used two different procedures: *MPM-Mixture* (based on **A**) and *MPM-Matérn* (based on **P**). In both procedures, we did not model any heteroscedasticity for additive genetic effects **u**.

In *MPM-Mixture*, $\boldsymbol{\Omega}_n = \rho\mathbf{A}\boldsymbol{\Theta}_K\mathbf{A}' + (1-\rho)\mathbf{J}_n$, where $\mathbf{J}_n$ is the *n* x *n* matrix of ones and $\boldsymbol{\Theta}_K$ is a *K* x *K* matrix depicting relationships among population clusters as inferred in **A**. Here, we simply set $\boldsymbol{\Theta}_K = \mathbf{I}_K$ ($\mathbf{I}_K$ is the *K* x *K* identity matrix), so $\boldsymbol{\Omega}_n = \rho\mathbf{A}\mathbf{A}' + (1-\rho)\mathbf{J}_n$. Therefore in this procedure, $\rho \in [0,1]$ set a trade-off between the case where relationships were cluster-specific ($\rho = 1$) and the case where relationships assumed one single homogeneous population for all individuals ($\rho = 0$). This approach is similar (but not exactly equivalent) to the *K*-kernel method of Heslot and Jannink (2015), which considered a similar balance between cluster-specific and overall relationships, but using $\mathbf{G_1}$ for relationships (VanRaden 2008), instead of $\mathbf{X}\mathbf{X}'$, and considering only discrete population clusters (in which case values in **A** would then be only 0 or 1). Alternatively, *MPM-Mixture* may be viewed as a multi-kernel model where $\rho\sigma_\beta^2$ and $(1-\rho)\sigma_\beta^2$ are the variances components respectively associated to cluster-specific and main marker effects.

In *MPM-Matérn*, $\boldsymbol{\Omega}_n = \left(\kappa_{\nu,h}(\mathbf{p}_i, \mathbf{p}_j)\right)_{n\times n}$, where $\kappa_{\nu,h}$ is a Matérn kernel function of $\mathbf{p}_i$ and $\mathbf{p}_j$: $\kappa_{\nu,h}(\mathbf{p}_i, \mathbf{p}_j) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\sqrt{2\nu}\frac{\|\mathbf{p}_i - \mathbf{p}_j\|_2}{h}\right)^\nu R_\nu\left\{\sqrt{2\nu}\frac{\|\mathbf{p}_i - \mathbf{p}_j\|_2}{h}\right\}$, $\|\mathbf{p}_i - \mathbf{p}_j\|_2$ is the Euclidean distance between the *d*-vectors of PC coordinates for any pair (*i*, *j*) of individuals, $\nu > 0$ is a shape parameter, $h > 0$ is a scale parameter, and $R_\nu\{.\}$ is the modified Bessel function of the second kind, of order *v* (Abramowitz and Stegun 1984, Ober *et al.* 2011). Matérn functions have been

used in various contexts, including in genomic prediction for depicting relationships among individuals (Ober *et al.* 2011). Here, we used Matérn functions to depict relationships among populations, with the input $\|\mathbf{p}_i - \mathbf{p}_j\|_2$ representing differentiation with respect to population structure in $d = 4$ orthogonal directions. We used Matérn functions instead of more typical kernel functions (e.g., an exponential or Gaussian kernel function) to allow for some flexibility in the shape of the correlation in $\boldsymbol{\Omega}_n$: $\nu = 0.5$ and $\nu = \infty$ correspond respectively to the exponential and Gaussian kernels as special cases, while different shapes can also be fitted (Ober *et al.* 2011).

The parameter ρ in *MPM-Mixture* was estimated by maximizing the restricted likelihood of model (3) using the optimization algorithm implemented in the R function *optimize*. The parameters *v* and *h* in *MPM-Matérn* were estimated by maximizing the restricted likelihood of model (3) using the Nelder-Mead algorithm implemented in the R function *constrOptim*, with constraints for positivity. In order to control (to some extent) for the possible presence of local maxima in the restricted likelihood surface in *MPM-Matérn*, we used four different starting points (ν_0, h_0): $(0.5, D_{max}/2)$, $(0.5, D_{max})$, $(10, D_{max}/2)$ and $(10, D_{max})$, with D_{max} the maximum distance $\|\mathbf{p}_i - \mathbf{p}_j\|_2$ observed over pairs of individuals (*i*, *j*).

Validations

We assessed the accuracy of our prediction procedures by cross-validation (CV): for each target (L4X-NE, U4X-N, Liberty-C2 or WS4U-C2), we used as the TS a random subset of the target sample. The size of the TS was one fifth of the target sample size. All remaining individuals were used as input to the prediction procedures (*Target*, *IS*, *SPM*, *MPM*), with some CS selection in *Target* and *IS*. Such validations were replicated $n_{rep} = 20$ times for each target.

Prediction procedures were evaluated for accuracy by $c_{TS} = Cor(\mathbf{y}_{TS}, \hat{\mathbf{y}}_{TS})$, i.e., the correlation between "observed" and predicted outcomes in a given TS. To assess the significance of differences in prediction accuracy between two procedures, we performed a t-test on $T = \frac{\bar{\delta}}{SD(\bar{\delta})}$, where $\bar{\delta}$ is the average of $\boldsymbol{\delta} = z(\mathbf{c}_t) - z(\mathbf{c}_0)$; $\mathbf{c}_t$ ($\mathbf{c}_0$) is the vector of prediction accuracies over testing sets for the candidate procedure (reference procedure); and z is the Fisher transformation. The standard error of the mean difference in prediction accuracy, $SD(\bar{\delta})$, was estimated in two different ways: (liberal t-test) $SD(\bar{\delta}) = SD(\delta_{TS})\sqrt{\frac{1}{n_{rep}}}$ where $SD(\delta_{TS})$ is the standard deviation of $\boldsymbol{\delta}$, with all testing sets assumed to be independent datasets; (conservative t-test) based on the first method of Nadeau and Bengio (2003), $SD(\bar{\delta}) = SD(\delta_{TS})\sqrt{\frac{1}{n_{rep}} + \frac{o}{1-o}}$, where redundancy over testing sets is accounted for by the additional term $\frac{o}{1-o}$, with *o* being the expected fraction of overlap among testing sets; here $o = \frac{1}{5}$ and $\frac{o}{1-o} = \frac{1}{4}$ because testing sets were random subsets consisting of a fifth of any given target sample. We considered that this approach for estimating $SD(\bar{\delta})$ was conservative because Nadeau and Bengio (2003) derived it by assuming that the CV criterion (the "loss function", analog here to $z(c_{TS,t}) - z(c_{TS,0})$, for a given TS) did not depend on the CS instances, given a particular CS size. Therefore the adjustment from Nadeau and Bengio (2003) may have overestimated the correlation among values of the CV criterion across replicates, since prediction procedures are probably quite sensitive to differences in the

composition of the CS. Furthermore, some procedures actually differed in CS size (*Target*, *IS*, and *SPM*/*MPM*). In all comparisons between procedures, we reported the results from both tests in order to characterize the significance of differences in prediction accuracy.

RESULTS

Population structure in the sample

Seven population clusters were inferred from the ADMIXTURE software (Figure S1; Alexander *et al.* 2009). These clusters corresponded roughly to populations L4X-NE, L4X-S, Liberty-C2 and U4X-N, WS4U-C2, U8X-E, U8X-W. One population with little representation in our sample, U8X-S, appeared to be of mixed origin (Figure 1a). The other populations generally displayed a low level of admixture, with relatively few individuals having intermediate admixture coefficients. There seemed to be some admixture involving upland populations (WS4U-C2 and U4X-N, WS4U-C2 and U8X-W, U8X-E and U8X-W), with even some shared ancestry between WS4U-C2 and U4X-N. The principal component analysis (PCA) confirmed that population structure was relatively discrete (Figure 1b). Unsurprisingly, the first principal component (PC) separated genotypes by ecotype while the second PC reflected geographical origin within the lowland ecotype (Lu *et al.* 2013, Evans *et al.* 2015). The third and four PCs discriminated upland genotypes by geographical origin and ploidy level, and distinguished L4X-S from the two other lowland populations (L4X-NE and Liberty-C2).

Differences in mean and range among populations were quite typical of previously reported difference between ecotypes (Table 1; Casler 2012). Indeed, L4X-S and Liberty-C2 (populations of lowland origin) had high mean values and range values for PH, HD and St, compared to upland populations (excluding U8X-S). However, L4X-NE stood out as a lowland population for being relatively short, early-flowering, and prone to lodging, with corresponding values for PH, HD, and St more similar to those of the upland populations.

Single-population models and relationships in the sample

Here, marginal genomic relationships were defined as the elements of $\mathbf{G} = \dot{\mathbf{X}}\dot{\mathbf{X}}'/s$, with $\dot{\mathbf{X}}$ consisting of centered marker variables, and s being some scaling factor. The strong and quite discrete population structure in the sample translated into multimodal marginal genomic relationship coefficients, with the multiple peaks in off-diagonal elements of $\mathbf{G}$ reflecting differentiation of population with respect to allele frequencies (Figure 2a). Conditioning relationships on population structure (as depicted by the first four PCs of $\dot{\mathbf{X}}$) yielded the matrix $\mathbf{G}_B$, with $\mathbf{G}_B = \mathbf{G} - \mathbf{B}\mathbf{B}'$ and $\mathbf{B}$ consisting of the PCs chosen to reflect structure in $\mathbf{G}$ (Fan *et al.* 2013). The conditional genomic relationships seemed sparser, in the sense that they appeared to cluster around zero, so most individuals could be assumed to be unrelated after accounting for population structure in the sample. In our particular study, conditional relationships in $\mathbf{G}_B$ were all the more relevant that among-population variation, captured by $\mathbf{B}\mathbf{B}'$, contributed little to variation within any given TS, because a TS generally consisted of selection candidates from a relatively homogeneous target sample (made of individuals from WS4U-C2, Liberty-C2, U4X-N or L4X-NE).

The *SPM-GLASSO* model did not yield substantial increases in quality of fit, compared to *SPM-GRM*, when fitted either to the whole sample or to calibration sets in cross-validation (CV) (Table 2). However, it is unclear to what extent a substantial improvement in fit should be expected from *SPM-GLASSO*: on the one hand, *SPM-GLASSO* relies on one additional

parameter λ compared to *SPM-GRM*, but on the other hand this parameter results in less complex relationships within $\widetilde{\mathbf{G}}$ compared to $\mathbf{G}$ (Foygel and Drton 2010). In general, quite small regularization parameters (λ) were selected based on REML: when fitted to the whole sample, *SPM-GLASSO* selected values of λ between 0.002 (for PH) and 0.007 (for St), corresponding respectively to 0.20- and 0.45-quantile of absolute correlations from $\mathbf{G}_B$. As a result, the inferred graphs were rather dense, with average degrees (number of neighbors by node/individual in the graph) ranging from 217 to 458 (Figure 3). However, even at such low regularization levels, some noticeable features of populations emerged from the inferred graphs (Figure 3): WS4U-C2, U4X-N and U8X-E appeared quite connected to one another; U8X-W also showed some connection with other upland populations but seemed more distinct, as reflected by a relatively lower average degree (Figure S2); Liberty-C2 and L4X-S were somewhat connected to both upland and lowland populations, which certainly explains why their individual degrees were generally high (Figure S2); most notably, L4X-NE displayed an outstandingly low level of connection with the other populations, which translated in a clear separation of this population in the graph, after placing the nodes based on a force-directed algorithm (Fruchterman and Reingold 1991). These features exemplify the usefulness of conditional relationships and their associated graphs for describing relationships among individuals and, potentially, serving as input to other types of procedures, e.g., instance selection.

For prediction in a given TS, the control procedure (*Target*) consisted in restricting the CS to the subset of the sample belonging to the same population as the TS. Compared to *Target*, *SPM-GRM* yielded increases in prediction accuracy that appeared somewhat significant ($p \leq 0.05$ based on the liberal "naïve" t-test) for PH (WS4U-C2, U4X-N) and St (Liberty-C2) (Table 3, Table S1). However, prediction accuracy for St (WS4U-C2) was lower, with a somewhat significant difference. More intriguing is the consistent decrease in prediction accuracy with L4X-NE, with differences being small yet highly significant for PH and HD ($p \leq 0.05$ based on the conservative t-test adapted from Nadeau and Bengio 2003; see *Material and methods* for details), and somewhat significant for St. It is unclear whether these differences are due to the consistently higher accuracies achieved with L4X-NE (in *Target*) compared to other populations, or a result of L4X-NE being relatively under-connected to the other populations in the sample (Figure S2, Figure 3). Both factors could very well contribute to the observed decreases in accuracy when incorporating information from the whole sample. Compared to *SPM-GRM*, *SPM-GLASSO* did not yield any notable increase in prediction accuracy, with generally similar accuracies, and differences from *SPM-GRM* ranging from -0.035 (for St, Liberty-C2) to +0.016 (for HD, WS4U-C2).

Instance selection in contexts of free and fixed CS size

In the fixed CS size context, increasing CS size often resulted in higher accuracy, with a plateau reached around the *Target* CS size, i.e., the number of individuals belonging to the sample population as the TS, which corresponded to 11%-15% of the available individuals (Figure 4; Table S2). The observed plateaus suggest that adding individuals from extraneous populations, without explicitly modelling heterogeneity, did not add useful signal for predictive ability of GBLUP; they also suggest that GBLUP is quite robust to such "superfluous" signal in a multi-population context. Exceptions were PH (WS4U-C2, U4X-N) and St (Liberty-C2) for which higher CS sizes did result in a somewhat significant increase in accuracy compared to *Target* (Figure 4). Conversely, with L4X-NE for all traits and WS4U-C2 for St, adding more individuals actually deteriorated accuracy (Figure 4). The results about L4X-NE make sense in light of the

facts previously noted for *SPM*: higher accuracies in L4X-NE and lack of kinship with other populations (Figure 3). When the CS was selected based on *IS-Rel*, CS sizes lower than the *Target* CS size resulted in significantly lower prediction accuracies compared to *Target*, except with WS4U-C2 (PH, St). Furthermore, at these low CS sizes, *IS-Rel* did not perform significantly better than random selection (*RS*) in some cases (Figure 5): PH (U4X-N, Liberty-C2), HD (U4X-N), and St (U4X-N, WS4U-C2). It was even significantly worse than *RS* for St (Liberty-C2) (Figure 5). In comparison, *IS-QP* was often significantly better than *IS-Rel* at these low CS sizes, with substantial differences observed with U4X-N and L4X-NE (Figure S3). This advantage of *IS-QP* translated into similar accuracies compared to *Target* with 10% selected individuals, and small, sometimes non-significant, decreases with 5% selected individuals. Accordingly, *IS-QP* maintained its advantage over *RS* at low CS sizes, with a consistent relative improvement as CS size decreased (Figure 5). One exception was St (Liberty-C2), for which there was a (non-significant) decrease in accuracy relatively to *RS* with 5% selected individuals. At intermediate CS sizes (35%-85% of selected individuals), *IS-QP* performed similarly to *IS-Rel*, with a small (yet significant) advantage over *IS-Rel* for HD (L4X-NE, U4X-N) but a significant disadvantage in three cases (PH, WS4U-C2; HD, WS4U-C2; St, Liberty-C2) (Figure S3).

In a context of free CS size, *IS* procedures yielded some significant improvements over *Target*, similarly to *SPM-GRM*. In fact, both *IS-Rel* and *IS-QP* tended to select many, if not all, of the available individuals (Table 4). *IS-QP* tended to select less individuals than *IS-Rel*, except for HD (Liberty-C2, L4X-NE) and St (Liberty-C2, L4X-NE). However, the differences in selection behavior between *IS-Rel* and *IS-QP* generally mattered little, with the exceptions of PH (WS4U-C2) and St (L4X-NE) for which *IS-QP* performed slightly worse than *IS-Rel* (-0.011 in mean accuracy). In the cases where *SPM* yielded significantly lower accuracies than *Target* (St with WS4U-C2, and all traits with L4X-NE), *IS* procedures failed to select an appropriately low number of individuals that would have prevented these decreases in accuracy (Figure 4), with the notable exception of *IS-Rel* for St (L4X-NE) (Table 3, Table 4).

Multi-population models and marker-by-population interactions

The inferred mixing parameter ρ from the *MPM-Mixture* model was null (or close to null), low and intermediate, for PH, St and HD respectively, with estimations being quite consistent over CV replicates (Table 2). Expectedly, the improvement in fit, relatively to *SPM-GRM*, was non-significant for PH, rather significant for St, and highly significant for HD (Table 2). In *MPM-Matérn*, the inferred correlation functions differed substantially across traits, while being quite consistent over CV replicates (Table 2, Figure 6): $\kappa_{v,h}$ roughly resembled an exponential kernel with PH and HD, and was more similar to a Gaussian kernel with St, for which a "shoulder" maintained high correlation in marker-effects for individuals that were relatively close to each other, based on their PCs. Inferences regarding $\mathbf{\Omega}_n$ in *MPM-Matérn* were weakly significant for PH and St, with *p*-values sometimes above 0.05 (ranging from 0.007 to 0.089 for PH and from 0.015 and 0.065 for St); in contrast, inferences regarding $\mathbf{\Omega}_n$ for HD were highly significant (Table 2). Interestingly, distances based on PCs may be equivalent to distances based on allele frequencies. Specifically, $\|\mathbf{p}_i - \mathbf{p}_j\|_2 = 2\left\|\boldsymbol{\pi}_{\mathbf{P}_i} - \boldsymbol{\pi}_{\mathbf{P}_j}\right\|_2$, where $\boldsymbol{\pi}_{\mathbf{P}_i}$ ($\boldsymbol{\pi}_{\mathbf{P}_j}$) is the *m*-vector of individual-specific allele frequencies of individual *i* (*j*) as described by Conomos *et al.* (2016), with population structure described by $[\mathbf{1}_n \quad \mathbf{P}]$ (Appendix A3). Therefore, the significant relationship between PC-based distances and correlations in marker effects for HD in *MPM-*

Matérn indicates that marker effects for this trait were highly sensitive to variation in allele frequencies across genetic backgrounds.

Regarding prediction accuracy, the performance of *MPM-Mixture* was very similar to that of *SPM-GRM*, with differences in accuracy ranging from -0.019 to +0.009 (Table 3). Quite surprisingly, *MPM-Mixture* displayed slightly deteriorated accuracies for HD (with the exception of U4X-N), despite the highly significant improvement in fit for this trait. In contrast, *MPM-Matérn* yielded larger differences in accuracy, ranging from -0.021 to +0.059 (Table 3). With the two upland target populations (WS4U-C2 and U4X-N), noteworthy increases in prediction accuracy (+0.059 and +0.032 respectively) were observed for HD. In these two cases, somewhat significant differences in accuracy compared to *Target* could be achieved, while no significant improvement could be obtained from *SPM-GRM*. With the two other target populations (Liberty-C2 and L4X-NE), smaller differences in accuracy (-0.009 and +0.006 respectively) were observed for HD. Our results suggest that a very high increase in quality of fit, as was observed for HD with *MPM-Matérn*, may allow for an increase in accuracy, but with no absolute guarantee. In the analysis of Heslot and Jannink (2015) across various multi-population contexts, there seemed to be a positive relationship between differences in quality of fit, as measured by the Akaike information criterion (AIC), and differences in prediction accuracy. Although this relationship was quite loose, it could be noted that for very high increases in AIC (≥ 30), gains in accuracy were null to high, similarly to the situation of *MPM-Matérn* with HD, for which increases in AIC varied from 28.37 to 42.53 across CV replicates (Table 2). Therefore, stringent thresholds on AIC increases could probably be used in *MPM* to avoid relative decreases in accuracy. Other characteristics and guidelines, related to the sample or the fitted model, may also be useful for this type of indication.

DISCUSSION

Conclusions

The present study assessed various procedures to accommodate population heterogeneity in diverse samples, with an application in switchgrass. We employed three typical strategies for dealing with marker-by-population interactions, i.e., ignoring (*SPM*), reducing (*IS*), or modelling (*MPM*) the source of heterogeneity in the data. These general strategies had previously been mentioned, e.g. by Bernardo (2002) about the analysis of genotype-by-environment interactions (GxE).

Here *SPM* often seemed robust to population heterogeneity, regarding prediction accuracy (Table 3). This robustness was probably contributed by the high marker density in our assay (De Roos *et al.* 2009). However, some decreases in accuracy compared to the control procedure (*Target*) suggest that robustness of *SPM* may have been affected by other factors. Such factors may be related to relationships within the sample, e.g. under-connectedness of some populations with others (Figure 3), or differences in accuracy of the prediction model from one population to another, in a single-population context (Table 3). Our proposed procedure (*SPM-GLASSO*), relying on a regularized form of the genomic relationship matrix, did not yield any improvement in prediction accuracy compared to the standard procedure (*SPM-GRM*). However, *SPM-GLASSO* was useful for inferring graphs of relationships within the sample, conditionally on population structure, which were used to derive informative features about our sample (Figure 3, Figure S2). In our study, the lack of benefit from regularization was probably due to the high marker density, translating into high estimation accuracy of genomic relationships (Endelman

and Jannink 2012, Casella and Berger 2002). In studies with lower marker densities, in which genomic relationship estimates are less accurate, *SPM-GLASSO* may have been more useful compared to *SPM-GRM*.

Selecting individuals in *IS* was not useful for improving prediction accuracy compared to *SPM*, in a context of free CS sizes (i.e., when *IS* is used as means of optimization). However, *IS* procedures were useful compared to random selection (*RS*) in contexts of fixed CS sizes, i.e., when restrictive numbers of individuals were used for calibration (Rincent *et al.* 2012, Isidro *et al.* 2014, Akdemir *et al.* 2015). In this type of scenarios, with small CS sizes (less individuals than in *Target*), the proposed procedure (*IS-QP*), which not only accounted for relationships between the CS and the TS but also redundancy within the CS, was particularly useful (Figure 5). In comparison, the more typical approach (*IS-Rel*), which only accounted for relationships to the TS, tended to lose its advantage over *RS* as CS size decreased (Figure 5). The relative superiority of *IS-QP* at low CS sizes are consistent with the findings of Pszczola *et al.* (2012), which suggested that redundancy within the CS was detrimental to prediction accuracy, for a given level of relationships to the TS.

In our case study, *MPM* procedures yielded highly significant improvements in fit for one of the three traits assayed (HD), in comparison to *SPM* (Table 2). Our proposed procedure (*MPM-Matérn*) relied on non-linear kernel functions for estimating population-level correlations in $\mathbf{\Omega}_n$, and was the only procedure to be more accurate than *SPM* for HD (Table 3). Differences in accuracy from *SPM* to *MPM* were smaller and seemed less predictable for PH and St, as could be expected from the more modest improvements in fit for these two traits. Our results exemplify the potential usefulness of parsimonious multi-population models, which are all the more interesting that they can be applied on samples comprising many populations. In contrast, typical multi-trait models would be computationally intractable or statistically inefficient here, since those would rely on one parameter for each population pair to model correlations among populations in $\mathbf{\Omega}_n$ (e.g., 21 parameters for $K = 7$ population clusters).

Improvement of procedures

The regularization method applied here in *SPM-GLASSO* imposed sparsity on the inverse matrix of conditional genomic relationships, thereby inferring a graph of recent relationships among individuals in the panel. Other regularization techniques act directly on the covariance matrix. Those include various thresholding methods (Rothman *et al.* 2009, Cai and Liu 2011), which may be useful, especially when relationships are derived from low-density markers. However, such methods do not necessarily guarantee positive definiteness of the regularized relationship matrices, which could be an issue when using them in linear mixed models. One other aspect of regularization that could be improved is the selection of the regularization parameter (λ). Here, we chose to select λ based on REML for a given outcome, but other selection techniques, employed on the covariance matrix, may be more relevant. These include selection of λ based on cross-validation (Bickel and Levina 2008), information criteria (Foygel and Drton 2010), or stability of inference (Liu *et al.* 2010). Further research would be necessary to explore the potential of such selection techniques for improving regularization of genomic relationship matrices with respect to prediction accuracy and/or graph inference.

In a context of unrestricted CS sizes, the tendency of *IS* procedures to select too many individuals, even when this was detrimental to prediction accuracy, may have been due to an overestimation of accuracy by CD_{mean} with larger CS sizes (Table 3, Table 4, Figure 4). Based on

the results of Hayes *et al.* (2009c), this upward bias in CD_{mean} may be of particular concern in a multi-population context. Therefore, other metrics than CD_{mean} may result in more pertinent selections in *IS*. Another improvement of *IS* may come from selection of individuals for prediction of each TS individual considered separately, as was proposed by Lorenz and Smith (2015). Because the *IS* procedures would then be run for one TS individual at a time, computationally intensive procedures based on stochastic algorithms would certainly not be applicable, making the *IS* procedures presented here all the more relevant. Finally, *IS* could be further improved by using other types of relationships than those used here. For example, as was recommended by Pszczola *et al.* (2012) and Wientjes *et al.* (2013), selections could be based on squared relationships, i.e., $\mathbf{G} \circ \mathbf{G}$ instead of $\mathbf{G}$ in *IS-Rel* (and by analogy, $\mathbf{XX}' \circ \mathbf{XX}'$ instead of $\mathbf{XX}'$ in *IS-QP*). Alternatively, entries in the relationship matrix could be replaced with those inferred in *MPM*, e.g. using $\mathbf{\Omega}_n \circ \mathbf{XX}'$ instead of $\mathbf{XX}'$ in *IS-QP*. Population heterogeneity would then be accounted for when selecting individuals from different genetic backgrounds. Finally, *IS* could rely on graphs of relationships such as those inferred from *SPM-GLASSO*. Selection of individuals would then be based on measures of connectivity between available individuals and the TS. Such measures could be the lengths of average shortest paths between each individual and the TS, or graph-based kernel functions, e.g. derived from the number of edges connecting each individual and the TS (Bishop 2006). Some features of our graphs seemed to mirror those revealed by the PCA plot (e.g., the distance of L4X-NE to the other populations). However, the PCA plot and the graphs depicted entirely distinct levels of relationships, the former representing relatedness at the population level and the latter representing relatedness conditionally on population structure. Therefore, graphs of relationships such as the ones inferred in *SPM-GLASSO* in our study offer new possibilities for depicting relationships and selecting individuals accordingly.

Multiple-population models were generally not useful when the improvement in model fit was modest. Therefore, a possible improvement of *MPM* procedures could simply come from model selection as an integral part of the fitting process, based for example on the Bayesian information criterion (BIC). In fact, the BIC differences relative to *SPM-GRM* were almost always negative for PH and St in *MPM* (Table 2). For these two traits, differences in prediction accuracy from *SPM* to *MPM* were quite inconsistent, especially with *MPM-Matérn*, so model selection could probably have made *MPM* procedures more robust. Another way of potentially improving *MPM* procedures would be to use other types of kernels than those used here. For example, one may use linear kernels based on population-level covariates (e.g. PCs) in place of $\mathbf{AA}'$ in *MPM-Mixture*, hence taking an approach similar to that of Jarquín *et al.* (2014) who modelled GxE through environmental covariates in multi-environment genomic prediction models. Also, the relationship matrix used in *MPM* ($\mathbf{\Omega}_n \circ \mathbf{XX}'$) may be regularized, then shrinking further – or even setting to zero – the relationships that had been reduced through $\mathbf{\Omega}_n$. Finally, an interesting way of extending the *MPM* procedures described here would be to incorporate more information at the population level. Here in *MPM*, population homogeneity was captured through admixture coefficients (*MPM-Mixture*) or differences in PC coordinates (*MPM-Matérn*), the latter reflecting differences in allele frequencies (Appendix A3). However, marker-by-population interactions may also be due to differences in LD patterns (Wientjes *et al.* 2016). Therefore metrics depicting such differences could be particularly appropriate for capturing population heterogeneity. Further research would be necessary to determine the type of statistics to use for reflecting differences in LD patterns, and the appropriate way to parsimoniously combine the different types of information regarding population differentiation in *MPM*.

Applications and prospects

Based on our case study, we would recommend using *MPM* whenever a strong improvement in model fit is achieved. Otherwise *SPM* would be the method of choice, since it is often robust enough to perform at least as well as *Target*. However, *Target* may be preferred when making predictions on "outlier populations" such as L4X-NE, which are under-connected to other populations and are characterized by relatively higher prediction accuracy in a single-population context. Only when the CS sizes are restricted (fixed) would *IS* procedures be useful – even though further improvements may make *IS* more competitive in contexts of free CS sizes. In such situations, we recommend using *IS-QP* instead of *IS-Rel*, especially when the CS size ought to be small. Nevertheless, more empirical studies on population heterogeneity would have to follow to support the conclusions from our specific application. Such studies could apply to various contexts: in particular, predictions on diverse samples or dynamic breeding programs. The former includes analyses similar to our case study as well as analyses on more complex data, such as historical datasets, in which not only population heterogeneity but also GxE must be taken into account (Dawson *et al.* 2013, Rutkoski *et al.* 2015). The latter involves selection across multiple breeding generations, which might not necessarily suffer from strong population heterogeneity (Sallam *et al.* 2015, Auinger *et al.* 2016) but could nonetheless benefit from robust multi-population models for potential increase in persistency of accuracy over generations (Habier *et al.* 2007). In this context, *IS* procedures could also be interesting, for example if a subset of non-selected individuals may be assayed phenotypically during the breeding program. In dynamic breeding analyses particularly, simulation studies also hold promise for assessing the suitability of various procedures to accommodate population heterogeneity.

ACKNOWLEDGMENTS

We are grateful to Jeremy Schmutz of the Department of Energy Joint Genome Institute and Hudson Alpha for his work on the switchgrass genome. Finally, we are grateful to Nick Baker and Joseph Halinar, USDA-ARS, Madison, WI and Steve Masterson, USDA-ARS, Lincoln, NE for assistance with field operations and data collection. This research was funded in part by the following agencies and organizations: the U.S. Department of Energy Great Lakes Bioenergy Research Center, DOE Office of Science BER DE-FC02- 07ER64494 (laboratory operations, genotyping, and bioinformatics), the U.S. Department of Energy Joint Genome Institute supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 (sequencing), Agriculture and Food Research Initiative Competitive Grant No. 2011-68005-30411 from the USDA National Institute of Food and Agriculture (CenUSA; field operations and phenotypic measurements), USDA-ARS Congressionally allocated funds (field operations, technical support, and logistics), and the University of Wisconsin Agricultural Research Stations (field operations). Mention of commercial products and organizations in this manuscript is solely to provide specific information. The USDA is an equal opportunity provider and employer. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

Abramowitz, M., and I. Stegun, 1984 Pocketbook of Mathematical Functions (Verlag Harri Deutsch, Thun; Frankfurt/Main).

Akdemir, D., J.I. Sanchez, and J.-L. Jannink, 2015 Optimization of genomic selection training populations with a genetic algorithm. *Genetics Selection Evolution* 47 (1):38.

Alexander, D.H., and K. Lange, 2011 Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC bioinformatics* 12 (1):246.

Alexander, D.H., J. Novembre, and K. Lange, 2009 Fast model-based estimation of ancestry in unrelated individuals. *Genome research* 19 (9):1655-1664.

Auinger, H.-J., M. Schönleben, C. Lehermeier, M. Schmidt, V. Korzun *et al.*, 2016 Model training across multiple breeding cycles significantly improves genomic prediction accuracy in rye (Secale cereale L.). *Theoretical and applied genetics* 129 (11):2043-2053.

Bernardo, R., 2002 *Breeding for quantitative traits in plants*: Stemma Press Woodbury.

Bickel, P.J., and E. Levina, 2008 Covariance regularization by thresholding. *The Annals of Statistics*:2577-2604.

Bishop, C.M., 2006 *Pattern recognition and machine learning*: Springer.

Butler, D., B.R. Cullis, A. Gilmour, and B. Gogel, 2009 ASReml-R reference manual. *The State of Queensland, Department of Primary Industries and Fisheries, Brisbane*.

Cai, T., and W. Liu, 2011 Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* 106 (494):672-684.

Carillier, C., H. Larroque, and C. Robert-Granié, 2014 Comparison of joint versus purebred genomic evaluation in the French multi-breed dairy goat population. *Genetics Selection Evolution* 46 (1):67.

Casella, G., and R.L. Berger, 2002 *Statistical inference*: Duxbury Pacific Grove, CA.

Casler, M., K. Vogel, and A. Beal, 2006 Registration of WS4U and WS8U switchgrass germplasms. *Crop science* 46 (2):998-1000.

Casler, M.D., 2012 Switchgrass breeding, genetics, and genomics, pp. 29-53 in *Switchgrass*. Springer.

Casler, M.D., and K.P. Vogel, 2014 Selection for biomass yield in upland, lowland, and hybrid switchgrass. *Crop science* 54 (2):626-636.

Clark, S.A., J.M. Hickey, H.D. Daetwyler, and J.H. van der Werf, 2012 The importance of information on relatives for the prediction of genomic breeding values and the implications for the makeup of reference data sets in livestock breeding schemes. *Genetics Selection Evolution* 44 (1):4.

Conomos, M.P., A.P. Reiner, B.S. Weir, and T.A. Thornton, 2016 Model-free estimation of recent genetic relatedness. *The American Journal of Human Genetics* 98 (1):127-148.

Daetwyler, H.D., A.A. Swan, J.H. van der Werf, and B.J. Hayes, 2012 Accuracy of pedigree and genomic predictions of carcass and novel meat quality traits in multi-breed sheep data assessed by cross-validation. *Genetics Selection Evolution* 44 (1):33.

Dawson, J.C., J.B. Endelman, N. Heslot, J. Crossa, J. Poland *et al.*, 2013 The use of unbalanced historical data for genomic selection in an international wheat breeding program. *Field Crops Research* 154:12-22.

de los Campos, G., A.I. Vazquez, R. Fernando, Y.C. Klimentidis, and D. Sorensen, 2013 Prediction of complex human traits using the genomic best linear unbiased predictor. *PLoS Genet* 9 (7):e1003608.

de los Campos, G., Y. Veturi, A.I. Vazquez, C. Lehermeier, and P. Pérez-Rodríguez, 2015 Incorporating genetic heterogeneity in whole-genome regressions using interactions. *Journal of agricultural, biological, and environmental statistics* 20 (4):467-490.

De Roos, A., B. Hayes, and M. Goddard, 2009 Reliability of genomic predictions across multiple populations. *Genetics* 183 (4):1545-1553.

Endelman, J.B., 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. *The Plant Genome* 4 (3):250-255.

Endelman, J.B., and J.-L. Jannink, 2012 Shrinkage estimation of the realized relationship matrix. *G3: Genes| Genomes| Genetics* 2 (11):1405-1413.

Erbe, M., B. Hayes, L. Matukumalli, S. Goswami, P. Bowman *et al.*, 2012 Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *Journal of dairy science* 95 (7):4114-4129.

Evans, J., E. Crisovan, K. Barry, C. Daum, J. Jenkins *et al.*, 2015 Diversity and population structure of northern switchgrass as revealed through exome capture sequencing. *The Plant Journal* 84 (4):800-815.

Evans, J., J. Kim, K.L. Childs, B. Vaillancourt, E. Crisovan *et al.*, 2014 Nucleotide polymorphism and copy number variant detection using exome capture and next-generation sequencing in the polyploid grass Panicum virgatum. *The Plant Journal* 79 (6):993-1008.

Fan, J., Y. Liao, and M. Mincheva, 2013 Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75 (4):603-680.

Foygel, R., and M. Drton, 2010 Extended Bayesian information criteria for Gaussian graphical models, pp. 604-612 in *Advances in neural information processing systems*.

Friedman, J., T. Hastie, and R. Tibshirani, 2008 Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9 (3):432-441.

Fruchterman, T.M., and E.M. Reingold, 1991 Graph drawing by force-directed placement. *Software: Practice and experience* 21 (11):1129-1164.

Gianola, D., G. de los Campos, W.G. Hill, E. Manfredi, and R. Fernando, 2009 Additive genetic variability and the Bayesian alphabet. *Genetics* 183 (1):347-363.

Habier, D., R. Fernando, and J. Dekkers, 2007 The impact of genetic relationship information on genome-assisted breeding values. *Genetics* 177 (4):2389-2397.

Habier, D., R.L. Fernando, and D.J. Garrick, 2013 Genomic BLUP decoded: a look into the black box of genomic prediction. *Genetics* 194 (3):597-607.

Hayes, B.J., P.J. Bowman, A. Chamberlain, and M. Goddard, 2009 Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of dairy science* 92 (2):433-443.

Hayes, B.J., P.J. Bowman, A.C. Chamberlain, K. Verbyla, and M.E. Goddard, 2009 Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution* 41 (1):51.

Hayes, B.J., P.M. Visscher, and M.E. Goddard, 2009 Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research* 91 (01):47-60.

Henderson, C.R., 1984 Applications of linear models in animal breeding. *Applications of linear models in animal breeding.*

Heslot, N., and J.-L. Jannink, 2015 An alternative covariance estimator to investigate genetic heterogeneity in populations. *Genetics Selection Evolution* 47 (1):93.

Isidro, J., J.-L. Jannink, D. Akdemir, J. Poland, N. Heslot *et al.*, 2015 Training set optimization under population structure in genomic selection. *Theoretical and applied genetics* 128 (1):145-158.

Jarquín, D., J. Crossa, X. Lacaze, P. Du Cheyron, J. Daucourt *et al.*, 2014 A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and applied genetics* 127 (3):595-607.

Jarquín, D., J. Specht, and A. Lorenz, 2016 Prospects of genomic prediction in the USDA Soybean Germplasm Collection: Historical data creates robust models for enhancing selection of accessions. *G3: Genes| Genomes| Genetics* 6 (8):2329-2341.

Karoui, S., M.J. Carabaño, C. Díaz, and A. Legarra, 2012 Joint genomic evaluation of French dairy cattle breeds using multiple-trait models. *Genetics Selection Evolution* 44 (1):39.

Laloë, D., 1993 Precision and information in linear models of genetic evaluation. *Genetics Selection Evolution* 25 (6):557-576.

Langholtz, M., B. Stokes, and L. Eaton, 2016 2016 Billion-ton report: Advancing domestic resources for a thriving bioeconomy, Volume 1: Economic availability of feedstock.

Lehermeier, C., C.-C. Schön, and G. de los Campos, 2015 Assessment of genetic heterogeneity in structured plant populations using multivariate whole-genome regression models. *Genetics* 201 (1):323-337.

Li, G., D.D. Serba, M.C. Saha, J.H. Bouton, C.L. Lanzatella *et al.*, 2014 Genetic linkage mapping and transmission ratio distortion in a three-generation four-founder population of Panicum virgatum (L.). *G3: Genes| Genomes| Genetics* 4 (5):913-923.

Lipka, A.E., F. Lu, J.H. Cherney, E.S. Buckler, M.D. Casler *et al.*, 2014 Accelerating the switchgrass (Panicum virgatum L.) breeding cycle using genomic selection approaches. *PloS one* 9 (11):e112227.

Liu, H., K. Roeder, and L. Wasserman, 2010 Stability approach to regularization selection (stars) for high dimensional graphical models, pp. 1432-1440 in *Advances in neural information processing systems*.

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi *et al.*, 2011 2 genomic selection in plant breeding: knowledge and prospects. *Advances in agronomy* 110:77.

Lorenz, A.J., and K.P. Smith, 2015 Adding genetically distant individuals to training populations reduces genomic prediction accuracy in barley. *Crop science* 55 (6):2657-2667.

Lorenzana, R.E., and R. Bernardo, 2009 Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. *Theoretical and applied genetics* 120 (1):151-161.

Lu, F., A.E. Lipka, J. Glaubitz, R. Elshire, J.H. Cherney *et al.*, 2013 Switchgrass genomic diversity, ploidy, and evolution: novel insights from a network-based SNP discovery protocol. *PLoS Genet* 9 (1):e1003215.

Lund, M.S., G. Su, L. Janss, B. Guldbrandtsen, and R.F. Brøndum, 2014 Genomic evaluation of cattle in a multi-breed context. *Livestock Science* 166:101-110.

Martin, E.R., D. Kinnamon, M.A. Schmidt, E. Powell, S. Zuchner *et al.*, 2010 SeqEM: an adaptive genotype-calling approach for next-generation sequencing studies. *Bioinformatics* 26 (22):2803-2810.

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 (4):1819-1829.

Mitchell, R.B., K.J. Moore, L.E. Moser, J.O. Fritz, and D.D. Redfearn, 1997 Predicting developmental morphology in switchgrass and big bluestem. *Agronomy Journal* 89 (5):827-832.

Müller, D., F. Technow, and A.E. Melchinger, 2015 Shrinkage estimation of the genomic relationship matrix can improve genomic estimated breeding values in the training set. *Theoretical and applied genetics* 128 (4):693-703.

Nadeau, C., and Y. Bengio, 2003 Inference for the generalization error. *Machine learning* 52 (3):239-281.

Ober, U., M. Erbe, N. Long, E. Porcu, M. Schlather *et al.*, 2011 Predicting genetic values: a kernel-based best linear unbiased prediction with genomic data. *Genetics* 188 (3):695-708.

Okada, M., C. Lanzatella, M.C. Saha, J. Bouton, R. Wu *et al.*, 2010 Complete switchgrass genetic maps reveal subgenome collinearity, preferential pairing and multilocus interactions. *Genetics* 185 (3):745-760.

Perlack, R.D., L.M. Eaton, A.F. Turhollow Jr, M.H. Langholtz, C.C. Brandt *et al.*, 2011 US billion-ton update: biomass supply for a bioenergy and bioproducts industry.

Perlack, R.D., L.L. Wright, A.F. Turhollow, R.L. Graham, B.J. Stokes *et al.*, 2005 Biomass as feedstock for a bioenergy and bioproducts industry: the technical feasibility of a billion-ton annual supply. DTIC Document.

Phocas, F., and D. Laloë, 2004 Should genetic groups be fitted in BLUP evaluation? Practical answer for the French AI beef sire evaluation. *Genetics Selection Evolution* 36 (3):1-21.

Pszczola, M., T. Strabel, H. Mulder, and M. Calus, 2012 Reliability of direct genomic values for animals with different relationships within and to the reference population. *Journal of dairy science* 95 (1):389-400.

Ramstein, G.P., J. Evans, S.M. Kaeppler, R.B. Mitchell, K.P. Vogel *et al.*, 2016 Accuracy of genomic prediction in switchgrass (Panicum virgatum L.) improved by accounting for linkage disequilibrium. *G3: Genes| Genomes| Genetics* 6 (4):1049-1062.

Rincent, R., D. Laloë, S. Nicolas, T. Altmann, D. Brunel *et al.*, 2012 Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (Zea mays L.). *Genetics* 192 (2):715-728.

Rodriguez-Lujan, I., R. Huerta, C. Elkan, and C.S. Cruz, 2010 Quadratic programming feature selection. *Journal of Machine Learning Research* 11 (Apr):1491-1516.

Rothman, A.J., E. Levina, and J. Zhu, 2009 Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association* 104 (485):177-186.

Rutkoski, J., R. Singh, J. Huerta-Espino, S. Bhavani, J. Poland *et al.*, 2015 Efficient use of historical data for genomic selection: a case study of stem rust resistance in wheat. *The Plant Genome* 8 (1).

Sallam, A., J. Endelman, J.-L. Jannink, and K. Smith, 2015 Assessing genomic selection prediction accuracy in a dynamic barley breeding population. *The Plant Genome* 8 (1).

Sanderson, M., R. Reed, S. McLaughlin, S. Wullschleger, B. Conger *et al.*, 1996 Switchgrass as a sustainable bioenergy crop. *Bioresource Technology* 56 (1):83-93.

Searle, S., G. Casella, and C. McCulloch, 2006 Variance components. *Hoboken Wiley, cop*:1-501.

Technow, F., A. Bürger, and A.E. Melchinger, 2013 Genomic prediction of northern corn leaf blight resistance in maize with combined or separated training sets for heterotic groups. *G3: Genes| Genomes| Genetics* 3 (2):197-203.

Thornton, T., H. Tang, T.J. Hoffmann, H.M. Ochs-Balcom, B.J. Caan *et al.*, 2012 Estimating kinship in admixed populations. *The American Journal of Human Genetics* 91 (1):122-138.

VanRaden, P., 2008 Efficient methods to compute genomic predictions. *Journal of dairy science* 91 (11):4414-4423.

VanRaden, P., C. Van Tassell, G. Wiggans, T. Sonstegard, R. Schnabel *et al.*, 2009 Invited review: Reliability of genomic predictions for North American Holstein bulls. *Journal of dairy science* 92 (1):16-24.

Wientjes, Y.C., R.F. Veerkamp, and M.P. Calus, 2013 The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics* 193 (2):621-631.

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 (2):799-823.

Zhao, T., H. Liu, K. Roeder, J. Lafferty, and L. Wasserman, 2012 The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research* 13 (Apr):1059-1062.

Zhou, L., M.S. Lund, Y. Wang, and G. Su, 2014 Genomic predictions across Nordic Holstein and Nordic Red using the genomic best linear unbiased prediction model with different genomic relationship matrices. *Journal of animal breeding and genetics* 131 (4):249-257.

TABLE 1 – *Description of populations and corresponding trait measurements*

Pop.	Size	Loc.	Trait	Years	Mean	Range
L4X-NE	106	NY	PH	2009 2011	142.8	96.5 - 202.8
			HD	2009 2010 2011	547.1	422.6 - 810.9
			St	2010 2011	5.6	1.0 - 9.0
L4X-S	37	NY	PH	2009 2011	209.6	130.5 - 240.7
			HD	2009 2010 2011	841.8	708.3 - 1076.0
			St	2010 2011	7.1	5.0 - 9.8
Liberty-C2	110	WI	PH	2012 2013	185.8	133.6 - 240.6
			HD	2012 2013 2014	806.6	650.9 - 981.6
			St	2013	6.2	2.6 - 8.9
U4X-N	135	NY	PH	2009 2011	155.8	94.3 - 207.8
			HD	2009 2010 2011	534.1	344.9 - 904.7
			St	2010 2011	5.4	1.5 - 8.0
WS4U-C2	136	WI	PH	2012 2013	163.8	127.7 - 203.7
			HD	2013 2014	527.6	400.3 - 688.5
			St	2013	5.7	2.1 - 8.2
U8X-E	97	NY	PH	2009 2011	168.3	100.9 - 225.5
			HD	2009 2010 2011	530.3	408.2 - 735.0
			St	2010 2011	5.6	1.7 - 8.0
U8X-W	129	NY	PH	2009 2011	165.3	126.5 - 225.8
			HD	2009 2010 2011	608	428.9 - 823.6
			St	2010 2011	3.5	0.5 - 7.3
U8X-S	10	NY	PH	2009 2011	175.6	138.5 - 190.7
			HD	2009 2010 2011	716.2	569.9 - 859.7
			St	2010 2011	5.8	4.0 - 7.5

Population (Pop.): WS4U-C2 is a collection of upland ecotypes; Liberty-C2 is a cross between upland and lowland ecotypes; other populations are designated by ecotype (U: upland; L: lowland), ploidy level (4X: tetraploid; 8X: octoploid) and geographical origin (S: South; W: West; N: North; E: East). Location (Loc.): location of phenotypic trials, Arlington (WI, USA) or Ithaca (NY, USA). Trait: plant height (PH), heading date (HD) or standability (St). Mean and range refer to the non-centered means y_i's as described in Material and Methods. Units for mean and range are centimeter, growing degree days and scores on a 0-10 scale, for PH, HD and St, respectively.

TABLE 2 – *Parameter estimates, likelihood-ratio test statistic (LRT) and p-value, by trait and procedure (range over target populations and cross-validation replicates in parentheses)*

Trait	Procedure	Parameter estimate	LRT statistic	LRT p-value
PH	*SPM-GLASSO*	λ: 0.002 (0.001-0.002)	0.84 (0.42-1.71)	0.36 (0.19-0.52)
	MPM-Mixture	ρ: 0.000 (0.000-0.005)	0.00 (0.00-0.01)	1 (0.9-1)
	MPM-Matérn	h^*: 0.525 (0.225-0.575) ν: 0.625 (0.550-0.921)	7.36 (4.85-9.92)	0.025 (0.007-0.089)
HD	*SPM-GLASSO*	λ: 0.003 (0.002-0.003)	2.24 (1.06-3.65)	0.13 (0.056-0.30)
	MPM-Mixture	ρ: 0.435 (0.355-0.572)	15.75 (13.67-18.78)	$7.2E^{-5}$ ($1.5E^{-5}$-$2.2E^{-4}$)
	MPM-Matérn	h^*: 0.325 (0.294-0.500) ν: 0.619 (0.550-0.735)	42.34 (32.37-46.53)	$6.4E^{-10}$ ($7.9E^{-11}$-$9.3E^{-8}$)
St	*SPM-GLASSO*	λ: 0.007 (0.000-0.019)	0.56 (0.00-2.38)	0.45 (0.12-1)
	MPM-Mixture	ρ: 0.138 (0.112-0.159)	5.68 (4.28-7.10)	0.017 (0.0077-0.038)
	MPM-Matérn	h^*: 0.134 (0.125-1.096) ν: 9.049 (0.807-10.014)	7.50 (5.47-8.42)	0.024 (0.015-0.065)

Trait: plant height (PH), heading date (HD) or standability (St). SPM-GLASSO: Single-population model based on regularized relationships (λ: regularization parameter); MPM: Multi-population model with among-population correlations based on admixture coefficients (MPM-Mixture; ρ: Mixture parameter) or PC distances (MPM-Matérn; ν: shape parameter; $h^* = h/D_{max}$, with h the scale parameter and D_{max}, the maximum distance $\|\mathbf{p}_i - \mathbf{p}_j\|_2$ observed over pairs of individuals). LRT statistic: $-2\log(L_0/L_1)$ where L_0 is the REML of SPM-GRM (GBLUP fitted to the whole sample) and L_1 is the REML of one of the procedure described here; p-values were obtained from a χ^2-ditribution with one (SPM-GLASSO, MPM-Mixture) or two (MPM-Matérn) degrees of freedom.

TABLE 3 – *Average prediction accuracy (standard deviation across cross-validation replicates in parentheses) by trait, target population and prediction procedure*

Trait	Pop.	Target	SPM-GRM	SPM-GLASSO	IS-Rel	IS-QP	MPM-Mixture	MPM-Matérn
PH	WS4U-C2	0.111 (0.113)	0.150 (0.114)•	0.132 (0.113)	0.150 (0.114)•	0.139 (0.130)	0.150 (0.114)•	0.129 (0.117)
	Liberty-C2	0.466 (0.186)	0.474 (0.188)	0.477 (0.187)	0.472 (0.189)	0.469 (0.202)	0.474 (0.188)	0.470 (0.186)
	U4X-N	0.496 (0.147)	0.531 (0.139)•	0.532 (0.139)•	0.531 (0.139)•	0.526 (0.143)•	0.531 (0.139)•	0.544 (0.122)•
	L4X-NE	0.788 (0.067)	0.772 (0.073)*	0.772 (0.073)*	0.773 (0.073)*	0.776 (0.074)*	0.772 (0.073)*	0.768 (0.075)*
HD	WS4U-C2	0.262 (0.162)	0.263 (0.190)	0.279 (0.183)	0.263 (0.190)	0.250 (0.196)	0.244 (0.181)	0.322 (0.150)•
	Liberty-C2	0.533 (0.152)	0.533 (0.145)	0.539 (0.141)	0.527 (0.149)	0.525 (0.147)	0.517 (0.152)•	0.524 (0.153)
	U4X-N	0.690 (0.111)	0.691 (0.104)	0.696 (0.101)	0.691 (0.104)	0.697 (0.104)	0.700 (0.101)•	0.722 (0.091)•
	L4X-NE	0.839 (0.073)	0.826 (0.075)*	0.826 (0.074)*	0.826 (0.076)*	0.828 (0.075)*	0.829 (0.074)*	0.832 (0.070)•
St	WS4U-C2	0.115 (0.197)	0.069 (0.215)•	0.071 (0.220)•	0.069 (0.215)•	0.064 (0.216)•	0.066 (0.219)•	0.078 (0.212)•
	Liberty-C2	0.053 (0.231)	0.116 (0.248)•	0.081 (0.243)•	0.116 (0.242)•	0.112 (0.245)•	0.105 (0.251)•	0.103 (0.252)•
	U4X-N	0.251 (0.175)	0.262 (0.169)	0.248 (0.172)	0.262 (0.169)	0.256 (0.170)	0.266 (0.172)	0.263 (0.166)
	L4X-NE	0.598 (0.121)	0.582 (0.128)•	0.592 (0.130)	0.595 (0.123)	0.584 (0.126)•	0.583 (0.129)•	0.583 (0.131)•

*Trait: plant height (PH), heading date (HD) or standability (St). Pop.: Population used as target for prediction. Prediction accuracies are averaged over 20 cross-validation replicates. Comparisons to the control procedure (Target): •, $p < 0.05$ in unadjusted (naïve) t-test (liberal); *, $p < 0.05$ in t-test corrected for overlap in testing sets as in Nadeau and Bengio (2003) (conservative). SPM: Single-population model based on non-regularized (SPM-GRM) or regularized (SPM-GLASSO) relationships; IS: Instance selection using average relationships (IS-Rel) or genotype weights (IS-QP); MPM: Multi-population model with among-population correlations based on admixture coefficients (MPM-Mixture) or PC distances (MPM-Matérn). Underlined values correspond to the best significant improvements over Target for each trait and population.*

TABLE 4 – *Size of the calibration set in the control procedure (Target), Single- and multiple-population models (SPM/MPM) and instance selection (IS-Rel, IS-QP)*

Trait	**Pop.**	***Target***	***SPM*/*MPM***	***IS-Rel* (S.D.)**	***IS-QP* (S.D.)**
PH	WS4U-C2	108	732	732 (0)	643 (119)
	Liberty-C2	88	738	674 (90)	650 (122)
	U4X-N	108	733	733 (0)	666 (76)
	L4X-NE	84	738	702 (0)	622 (170)
HD	WS4U-C2	108	732	732 (0)	670 (71)
	Liberty-C2	88	738	537 (123)	666 (113)
	U4X-N	108	733	733 (0)	660 (62)
	L4X-NE	84	738	591 (0)	646 (112)
St	WS4U-C2	108	732	732 (0)	716 (30)
	Liberty-C2	88	738	495 (94)	552 (211)
	U4X-N	108	733	733 (0)	682 (58)
	L4X-NE	84	738	148 (0)	528 (209)

Trait: plant height (PH), heading date (HD) or standability (St). Pop.: Population used as target for prediction. S.D.: standard deviation across CV replicates for CS size in IS.

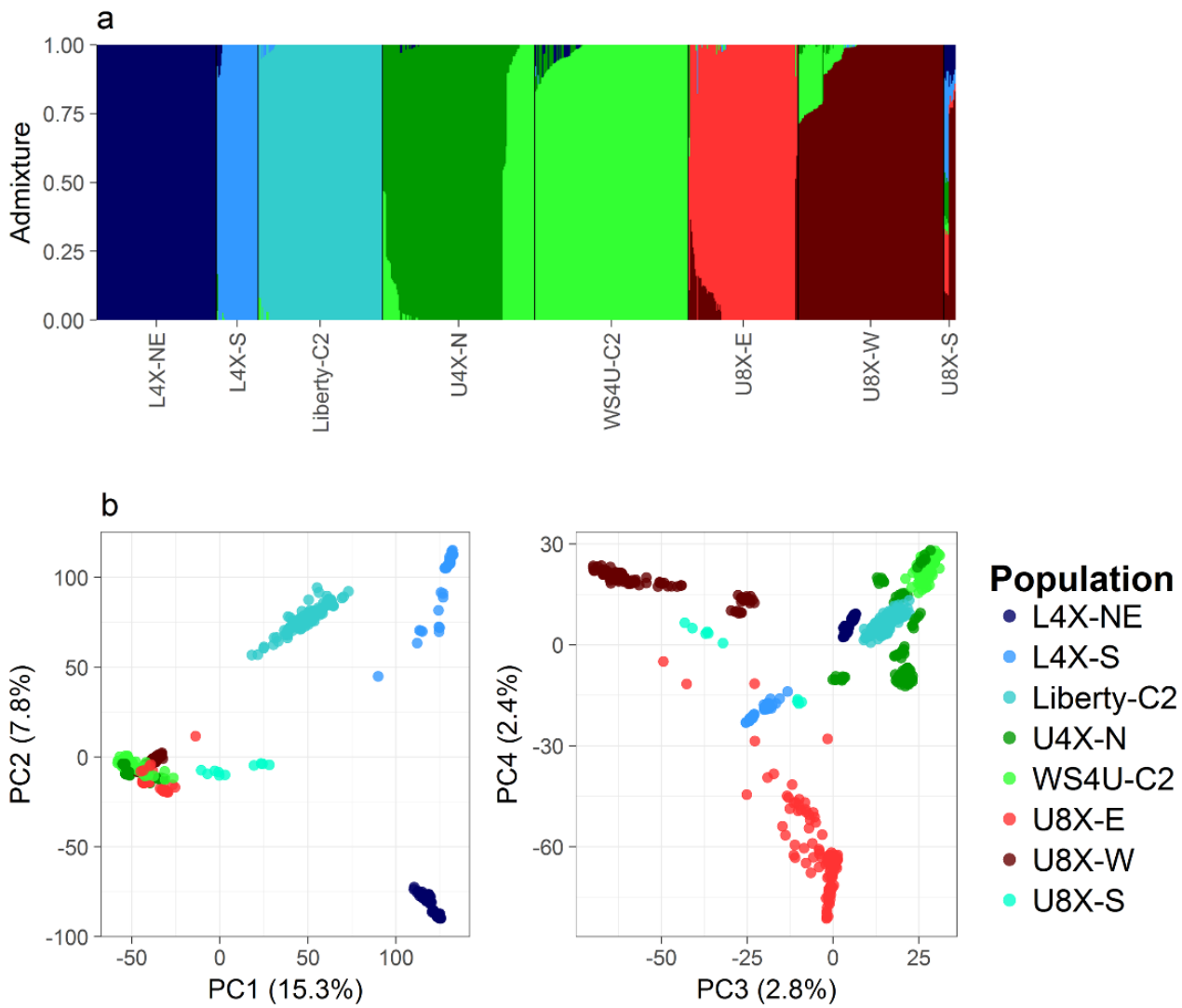

FIGURE 1 – *(a) Admixture plot of the whole sample, with colors designating the seven inferred population clusters, which roughly matched populations, with the exception of U8X-S which displayed strong admixture; (b) Principal component analysis (PCA) plot of the whole sample of 760 individuals, with colors designating the eight populations.*

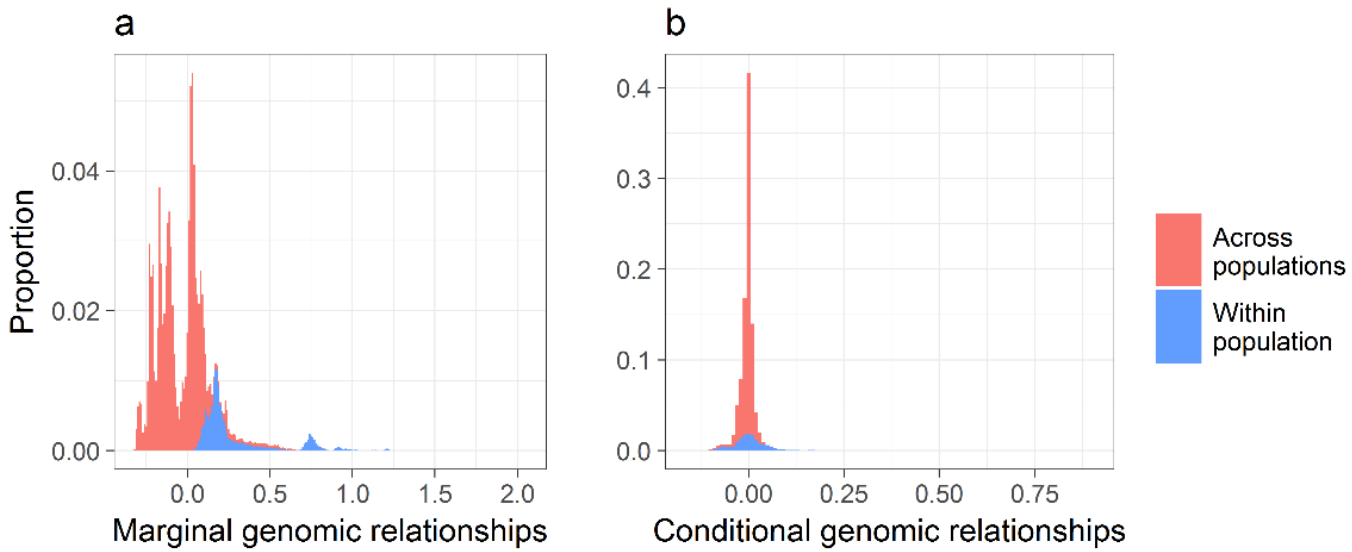

FIGURE 2 – *(a) Histogram of marginal genomic relationships, scaled as in VanRaden (2008): off-diagonal elements of* $\mathbf{G}/v$*, with* $v = 2\sum_{l=1}^{m}\hat{\pi}_l(1-\hat{\pi}_l)$ *(*$\hat{\pi}_l$*: estimated allele frequency at marker l). (b) Histogram of genomic relationships conditional on population structure, as captured by PC, also scaled as in VanRaden (2008): off-diagonal elements of* $\mathbf{G}_B/v$*, with* $\mathbf{G}_B = \mathbf{G} - \mathbf{B}\mathbf{B}'$ *(*$\mathbf{B}$ *consists of the first* $d = 4$ *PCs of* $\mathbf{G}$*).*

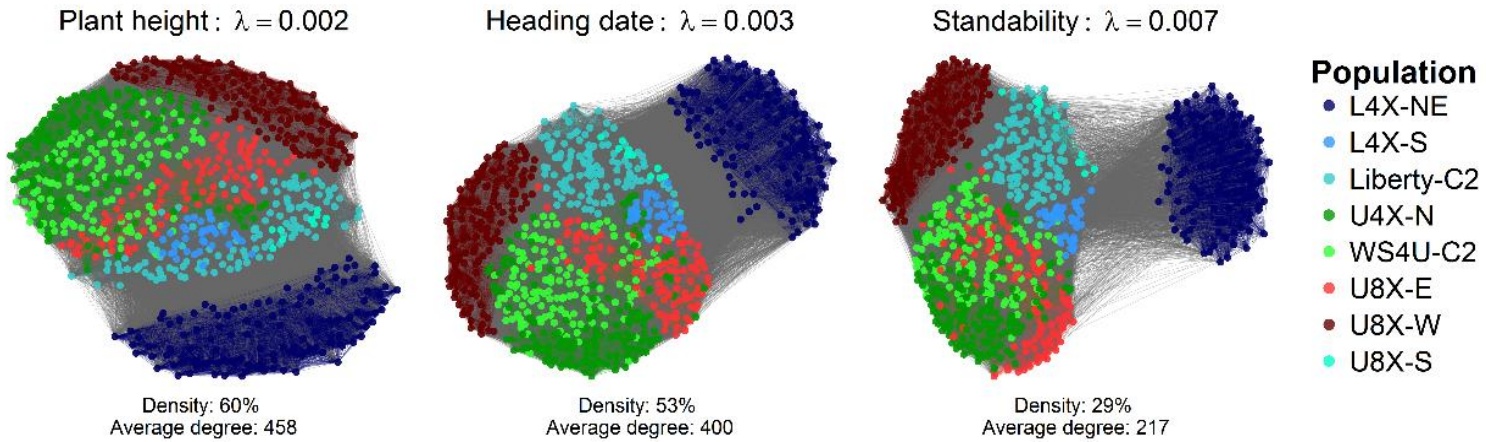

FIGURE 3 – *Inferred graphs of relationships, conditional on population structure. Each graph represents the relationships as depicted from* $\widetilde{\mathbf{G}}_B^{-1}$ *for a given trait, where* $\widetilde{\mathbf{G}}_\mathrm{B}$ *is the regularized matrix of conditional relationships obtained by the graphical LASSO applied to the whole sample of individuals (the absence of edge between two individuals is indicated by a zero ij-element in* $\widetilde{\mathbf{G}}_\mathrm{B}^{-1}$*). Nodes (individuals) were positioned using the force-directed placement algorithm of Fruchterman and Reingold (1991), as implemented in function ggnet (R package GGally), so aggregation of nodes reflects connectedness.*

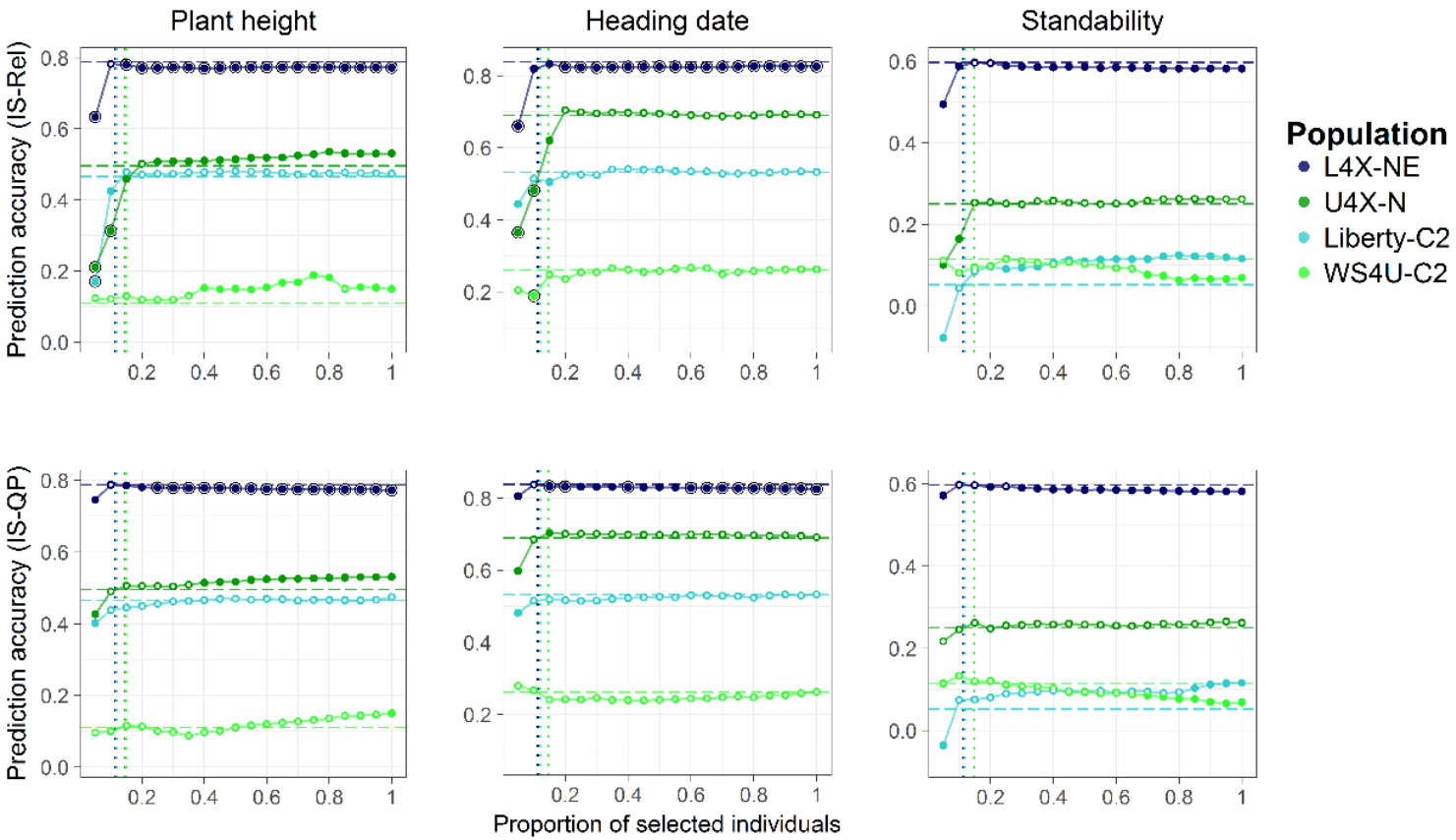

FIGURE 4 – *Average prediction accuracies (over cross-validation replicates) from instance selection with fixed CS sizes, compared to the control procedure (Target). Upper panel: prediction accuracies from IS-Rel. Lower panel: prediction accuracies from IS-QP. Horizontal dashed lines indicate the average prediction accuracy from Target; vertical dotted lines indicate the corresponding proportion of individuals in the CS: 11%, 15%, 12% and 15% for L4X-NE, U4X-N, Liberty-C2 and WS4U-C2, respectively. Comparisons to the control: Colored fill, $p < 0.05$ in unadjusted (naïve) t-test (liberal); black circle, $p < 0.05$ in t-test corrected for overlap in testing sets as in Nadeau and Bengio (2003) (conservative).*

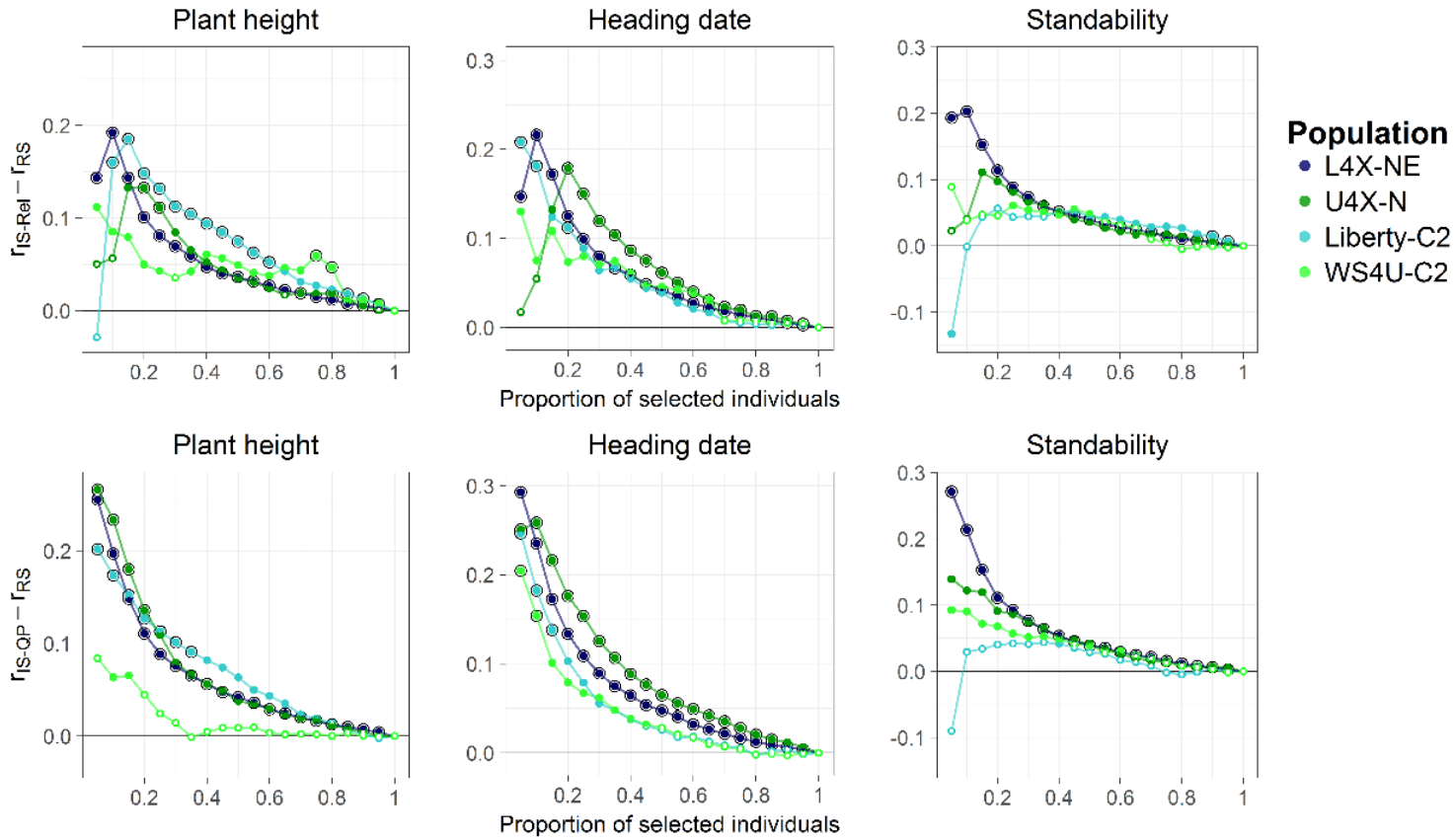

FIGURE 5 – *Average differences in prediction accuracy (over cross-validation replicates) between instance selection procedures and random selection (RS), with fixed CS sizes. Upper panel: prediction accuracies from IS-Rel. Lower panel: prediction accuracies from IS-QP. Significance of differences: colored fill, $p < 0.05$ in unadjusted (naïve) t-test (liberal); black circle, $p < 0.05$ in t-test corrected for overlap in testing sets as in Nadeau and Bengio (2003) (conservative).*

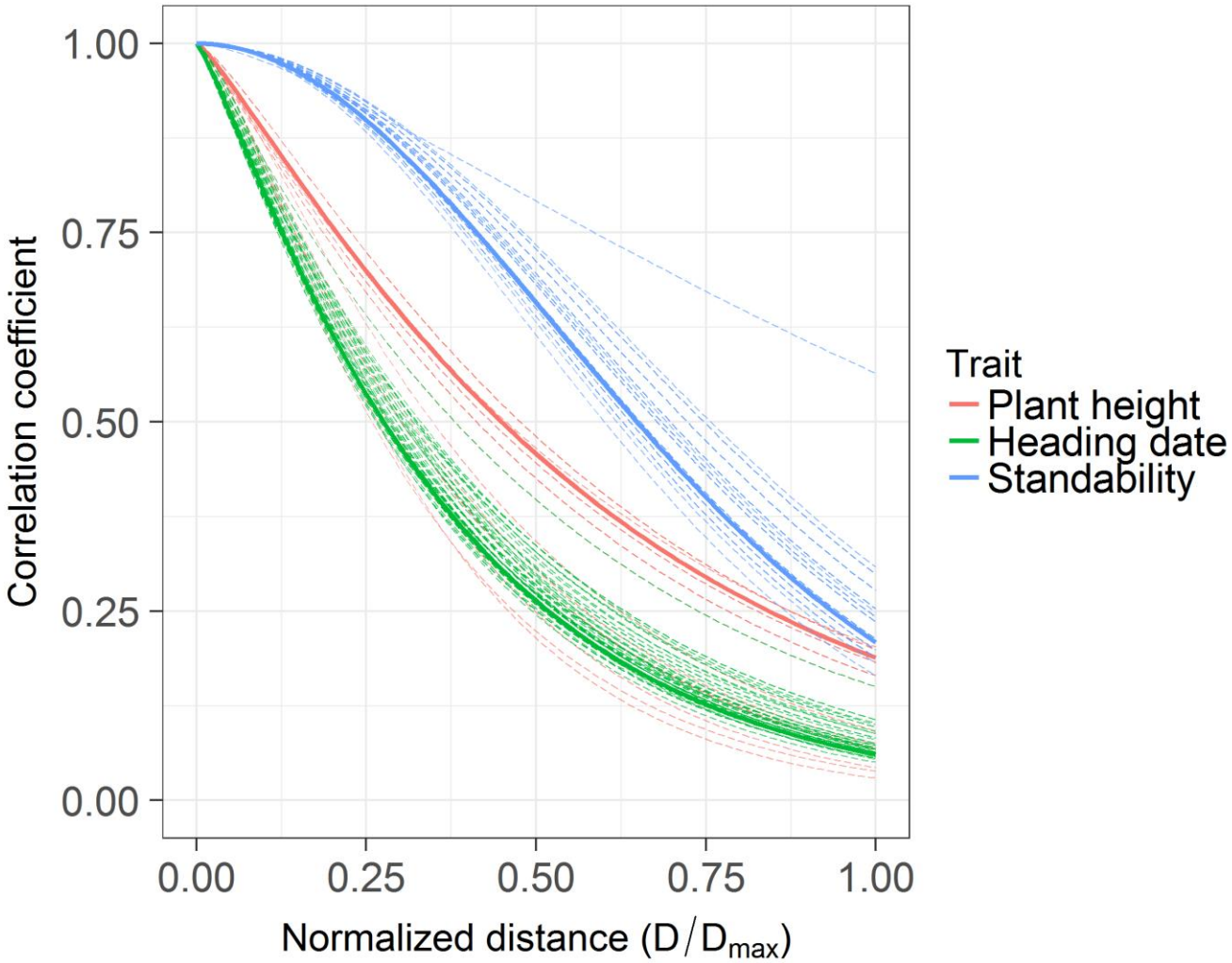

FIGURE 6 – *Shape of the inferred correlation function in MPM-Matérn by trait, in the whole sample (solid curves) or in cross-validation replicates (dashed curves), as a function of* $D = \|\mathbf{p}_i - \mathbf{p}_j\|_2$ *(the Euclidean distance between population-structure PCs for any pair of individual (i,j)), scaled by* D_{max}*, the maximum distance* $\|\mathbf{p}_i - \mathbf{p}_j\|_2$ *observed over pairs.*

APPENDIX

A1. Equivalence of fit in linear mixed models after regressing out fixed-effect variables from random-effect variables

In this section, matrix notations are not consistent with those in the main text; **I** refers to the identity matrix with dimensions equal to the number of observations.

Consider the following two models:

$$\mathbf{y} = \mathbf{X\beta_1} + \mathbf{Zu_1} + \mathbf{e_1};\ \begin{bmatrix}\mathbf{e_1}\\ \mathbf{u_1}\\ \mathbf{y}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{X\beta_1}\end{bmatrix}, \begin{bmatrix}\mathbf{R} & \mathbf{0} & \mathbf{R}\\ \mathbf{0} & \mathbf{G} & \mathbf{GZ'}\\ \mathbf{R} & \mathbf{ZG} & \mathbf{V_1}\end{bmatrix}\right) \quad (1)$$

where $\mathbf{V_1} = \mathbf{ZGZ'} + \mathbf{R}$

$$\mathbf{y} = \mathbf{X\beta_2} + (\mathbf{I}-\mathbf{H})\mathbf{Zu_2} + \mathbf{e_2};\ \begin{bmatrix}\mathbf{e_2}\\ \mathbf{u_2}\\ \mathbf{y}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{X\beta_2}\end{bmatrix}, \begin{bmatrix}\mathbf{R} & \mathbf{0} & \mathbf{R}\\ \mathbf{0} & \mathbf{G} & \mathbf{GZ'}(\mathbf{I}-\mathbf{H})'\\ \mathbf{R} & (\mathbf{I}-\mathbf{H})\mathbf{ZG} & \mathbf{V_2}\end{bmatrix}\right) \quad (2)$$

with $\mathbf{H} = \mathbf{X}(\mathbf{X'R^{-1}X})^{-1}\mathbf{X'R^{-1}}$ and $\mathbf{V_2} = (\mathbf{I}-\mathbf{H})\mathbf{ZGZ'}(\mathbf{I}-\mathbf{H})' + \mathbf{R}$.

For given **R** and **G** (possibly estimated by ML or REML), in model (1), the ML estimates of regression coefficients are $\widehat{\boldsymbol{\beta}}_\mathbf{1} = \left(\mathbf{X'}{\mathbf{V_1}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X'}{\mathbf{V_1}}^{-1}\mathbf{y}$ (as best linear unbiased estimators, BLUEs) and $\widehat{\mathbf{u}}_\mathbf{1} = \mathbf{GZ'}{\mathbf{V_1}}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\mathbf{1})$ (as best linear unbiased predictors, BLUPs); in model (2), the mixed model equations (MME) are

$$\begin{bmatrix}\mathbf{X'R^{-1}X} & \mathbf{0}\\ \mathbf{0} & \mathbf{Z'}(\mathbf{I}-\mathbf{H})'\mathbf{R^{-1}}(\mathbf{I}-\mathbf{H})\mathbf{Z} + \mathbf{G^{-1}}\end{bmatrix}\begin{bmatrix}\widehat{\boldsymbol{\beta}}_\mathbf{2}\\ \widehat{\mathbf{u}}_\mathbf{2}\end{bmatrix} = \begin{bmatrix}\mathbf{X'R^{-1}y}\\ \mathbf{Z'}(\mathbf{I}-\mathbf{H})'\mathbf{R^{-1}y}\end{bmatrix}$$

so the ML estimates of regression coefficients are (as solutions of the MME) $\widehat{\boldsymbol{\beta}}_\mathbf{2} = (\mathbf{X'R^{-1}X})^{-1}\mathbf{X'R^{-1}y}$ and $\widehat{\mathbf{u}}_\mathbf{2} = (\mathbf{Z'}(\mathbf{I}-\mathbf{H})'\mathbf{R^{-1}}(\mathbf{I}-\mathbf{H})\mathbf{Z} + \mathbf{G^{-1}})^{-1}\mathbf{Z'}(\mathbf{I}-\mathbf{H})'\mathbf{R^{-1}y}$ (Henderson, 1984).

□ For given **R** and **G**, fitting model (1) and fitting model (2) by ML are equivalent, in that $\hat{\mathbf{e}}_\mathbf{1} = \hat{\mathbf{e}}_\mathbf{2}$, so $\hat{\mathbf{y}}_\mathbf{1} = \hat{\mathbf{y}}_\mathbf{2}$.

Consider the two matrices **P** and **S** such that $\mathbf{P} = {\mathbf{V_1}}^{-1} - {\mathbf{V_1}}^{-1}\mathbf{X}\left(\mathbf{X'}{\mathbf{V_1}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X'}{\mathbf{V_1}}^{-1}$ and $\mathbf{S} = \mathbf{R^{-1}} - \mathbf{R^{-1}X}(\mathbf{X'R^{-1}X})^{-1}\mathbf{X'R^{-1}}$.

As shown by Searle *et al.* (2006, pp. 282-283):

- $\mathbf{P} = \mathbf{S} - \mathbf{SZ}(\mathbf{Z'SZ} + \mathbf{G^{-1}})^{-1}\mathbf{Z'S}$

Therefore:

$${\mathbf{V_1}}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\mathbf{1}\right) = \mathbf{Py} = \mathbf{Sy} - \mathbf{SZ}(\mathbf{Z'SZ} + \mathbf{G^{-1}})^{-1}\mathbf{Z'Sy} \quad (E1)$$

Moreover:

- $\mathbf{Sy} = \mathbf{R^{-1}}(\mathbf{I}-\mathbf{H})\mathbf{y} = \mathbf{R^{-1}}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\mathbf{2}\right)$.
- Because $\mathbf{S} = \mathbf{R^{-1}}(\mathbf{I}-\mathbf{H}) = (\mathbf{I}-\mathbf{H})'\mathbf{R^{-1}} = (\mathbf{I}-\mathbf{H})'\mathbf{R^{-1}}(\mathbf{I}-\mathbf{H})$, $\mathbf{SZ} = \mathbf{R^{-1}}(\mathbf{I}-\mathbf{H})\mathbf{Z}$ and $\mathbf{Z'SZ} = \mathbf{Z'}(\mathbf{I}-\mathbf{H})'\mathbf{R^{-1}}(\mathbf{I}-\mathbf{H})\mathbf{Z}$.

Therefore (E1) simplifies into $\mathbf{V_1}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{1}}\right)=\mathbf{R}^{-1}\left[\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{2}}-(\mathbf{I}-\mathbf{H})\mathbf{Z}(\mathbf{Z}'(\mathbf{I}-\mathbf{H})'\mathbf{R}^{-1}(\mathbf{I}-\mathbf{H})\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}'(\mathbf{I}-\mathbf{H})'\mathbf{R}^{-1}\mathbf{y}\right]=\mathbf{R}^{-1}\left[\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{2}}-(\mathbf{I}-\mathbf{H})\mathbf{Z}\widehat{\mathbf{u}}_{\mathbf{2}}\right]$. So:

$$\mathbf{V_1}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{1}}\right)=\mathbf{R}^{-1}\hat{\mathbf{e}}_{\mathbf{2}} \quad \text{(E2)}$$

Besides, as shown by Henderson (1984, Chapter 5 p. 9) by application of the Sherman-Morrison-Woodbury formula to a general variance formulation:

$$\mathbf{V_1}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{1}}\right)=\mathbf{R}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{1}}-\mathbf{Z}\widehat{\mathbf{u}}_{\mathbf{1}}\right)=\mathbf{R}^{-1}\hat{\mathbf{e}}_{\mathbf{1}} \quad \text{(E3)}$$

Since $\mathbf{R}^{-1}$ is positive definite (so it is full rank), it follows from (E2) and (E3) that $\hat{\mathbf{e}}_{\mathbf{1}}=\hat{\mathbf{e}}_{\mathbf{2}}$. ■

□ For given $\mathbf{R}$ and $\mathbf{G}$, $\widehat{\mathbf{u}}_{\mathbf{1}}=\widehat{\mathbf{u}}_{\mathbf{2}}$.

As previously stated:

- $\mathbf{S}=(\mathbf{I}-\mathbf{H})'\mathbf{R}^{-1}(\mathbf{I}-\mathbf{H})$

Therefore, it follows from (E1) that:

$$\mathbf{V_1}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{1}}\right)=(\mathbf{I}-\mathbf{H})'[\mathbf{R}^{-1}-\mathbf{R}^{-1}(\mathbf{I}-\mathbf{H})\mathbf{Z}(\mathbf{Z}'(\mathbf{I}-\mathbf{H})'\mathbf{R}^{-1}(\mathbf{I}-\mathbf{H})\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}'(\mathbf{I}-\mathbf{H})'\mathbf{R}^{-1}](\mathbf{I}-\mathbf{H})\mathbf{y} \quad \text{(E4)}$$

By the Sherman-Morrison-Woodbury formula, $\mathbf{V_2}^{-1}=[\mathbf{R}+(\mathbf{I}-\mathbf{H})\mathbf{Z}\mathbf{G}\mathbf{Z}'(\mathbf{I}-\mathbf{H})']^{-1}=[\mathbf{R}^{-1}-\mathbf{R}^{-1}(\mathbf{I}-\mathbf{H})\mathbf{Z}(\mathbf{Z}'(\mathbf{I}-\mathbf{H})'\mathbf{R}^{-1}(\mathbf{I}-\mathbf{H})\mathbf{Z}+\mathbf{G}^{-1})^{-1}\mathbf{Z}'(\mathbf{I}-\mathbf{H})'\mathbf{R}^{-1}]$.

So it follows from (E4) that:

$$\mathbf{V_1}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{1}}\right)=(\mathbf{I}-\mathbf{H})'\mathbf{V_2}^{-1}(\mathbf{I}-\mathbf{H})\mathbf{y}=(\mathbf{I}-\mathbf{H})'\mathbf{V_2}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{2}}\right)$$

Therefore, as BLUP, $\widehat{\mathbf{u}}_{\mathbf{1}}=\mathbf{G}\mathbf{Z}'\mathbf{V_1}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{1}}\right)=\mathbf{G}\mathbf{Z}'(\mathbf{I}-\mathbf{H})'\mathbf{V_2}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{2}}\right)=\widehat{\mathbf{u}}_{\mathbf{2}}$. ■

A2. Multi-population GBLUP models for heterogeneous calibration sets

In this section, $\mathbf{1}_t$, $\mathbf{I}_t$, and $\mathbf{J}_t$ refer to the vector of ones, identity matrix, and matrix of ones, respectively, of dimensions t, $t \times t$ and $t \times t$ (where t is specified).

Consider the following model for population-specific marker effects with respect to K populations and n genotypes:

$$\bar{\mathbf{y}} = \bar{\mathbf{Q}}\boldsymbol{\alpha} + \bar{\mathbf{X}}\bar{\boldsymbol{\beta}} + \bar{\mathbf{e}};\ \begin{bmatrix}\bar{\mathbf{e}}\\ \bar{\boldsymbol{\beta}}\\ \bar{\mathbf{y}}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \bar{\mathbf{Q}}\boldsymbol{\alpha}\end{bmatrix}, \begin{bmatrix}\bar{\mathbf{R}} & \mathbf{0} & \bar{\mathbf{R}}\\ \mathbf{0} & (\boldsymbol{\Omega}_K \otimes \mathbf{I}_m)\sigma^2_{\boldsymbol{\beta}} & (\boldsymbol{\Omega}_K \otimes \mathbf{I}_m)\bar{\mathbf{X}}'\sigma^2_{\boldsymbol{\beta}}\\ \bar{\mathbf{R}} & \bar{\mathbf{X}}(\boldsymbol{\Omega}_K \otimes \mathbf{I}_m)\sigma^2_{\boldsymbol{\beta}} & \bar{\mathbf{V}}\end{bmatrix}\right);$$

$$\bar{\mathbf{V}} = \bar{\mathbf{X}}(\boldsymbol{\Omega}_K \otimes \mathbf{I}_m)\bar{\mathbf{X}}'\sigma^2_{\boldsymbol{\beta}} + \bar{\mathbf{R}}$$

where $\otimes$ indicates the Kronecker product; $\bar{\mathbf{Q}} = (\mathbf{1}_K \otimes \mathbf{Q})$ is the Kn x p design matrix for the p-vector $\boldsymbol{\alpha}$ of fixed effects; $\bar{\mathbf{X}} = (\mathbf{I}_K \otimes \mathbf{X})$ is the Kn x Km marker-data matrix for the Km-vector $\bar{\boldsymbol{\beta}}$ of marker effects at each of the K populations, with variance $(\boldsymbol{\Omega}_K \otimes \mathbf{I}_m)\sigma^2_{\boldsymbol{\beta}}$. The matrix $\boldsymbol{\Omega}_K$ reflects covariances in marker effects between populations. The Kn-vector $\bar{\mathbf{y}}$, containing the phenotypic values for the n genotypes at each of the K populations, is hypothetical (and ill-defined from a practical standpoint), since genotypes typically do not belong to more than one population. The Kn-vector of residuals $\bar{\mathbf{e}}$, with unspecified variance $\bar{\mathbf{R}} = (\mathbf{I}_K \otimes \mathbf{R})$, is assumed to be uncorrelated to marker effects $\bar{\boldsymbol{\beta}}$.

Let $\bar{\mathbf{u}} = \bar{\mathbf{X}}\bar{\boldsymbol{\beta}}$ be the Kn-vector of additive genetic effects at each of the K populations,

As a linear combination of a normally-distributed vector ($\bar{\boldsymbol{\beta}}$), $\bar{\mathbf{u}}$ follows a normal distribution with expectation and variance as follows (Lehermeier *et al.*, 2015):

$$\mathrm{E}[\bar{\mathbf{u}}] = \bar{\mathbf{X}}\mathrm{E}[\bar{\boldsymbol{\beta}}] = \mathbf{0}$$

$$\mathrm{Var}[\bar{\mathbf{u}}] = \bar{\mathbf{X}}(\boldsymbol{\Omega}_K \otimes \mathbf{I}_m)\bar{\mathbf{X}}'\sigma^2_{\boldsymbol{\beta}} = (\mathbf{I}_K \otimes \mathbf{X})(\boldsymbol{\Omega}_K \otimes \mathbf{I}_m)(\mathbf{I}_K \otimes \mathbf{X}')\sigma^2_{\boldsymbol{\beta}} = (\boldsymbol{\Omega}_K \otimes \mathbf{X})(\mathbf{I}_K \otimes \mathbf{X}')\sigma^2_{\boldsymbol{\beta}} = (\boldsymbol{\Omega}_K \otimes \mathbf{X}\mathbf{X}')\sigma^2_{\boldsymbol{\beta}}$$

So a multi-population model for breeding values that is equivalent to the model described above, by identical mean and variance structures, is as follows:

$$\bar{\mathbf{y}} = \bar{\mathbf{Q}}\boldsymbol{\alpha} + \bar{\mathbf{u}} + \bar{\mathbf{e}};\ \begin{bmatrix}\bar{\mathbf{e}}\\ \bar{\mathbf{u}}\\ \bar{\mathbf{y}}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \bar{\mathbf{Q}}\boldsymbol{\alpha}\end{bmatrix}, \begin{bmatrix}\bar{\mathbf{R}} & \mathbf{0} & \bar{\mathbf{R}}\\ \mathbf{0} & (\boldsymbol{\Omega}_K \otimes \mathbf{X}\mathbf{X}')\sigma^2_{\boldsymbol{\beta}} & (\boldsymbol{\Omega}_K \otimes \mathbf{X}\mathbf{X}')\sigma^2_{\boldsymbol{\beta}}\\ \bar{\mathbf{R}} & (\boldsymbol{\Omega}_K \otimes \mathbf{X}\mathbf{X}')\sigma^2_{\boldsymbol{\beta}} & \bar{\mathbf{V}}\end{bmatrix}\right)$$

Now assume that $K = n$, and each population correspond to the specific genetic background of each individuals separately. By considering only observations at every individual's specific genetic background, the above model reduces to:

$$\mathbf{y} = \mathbf{Q}\boldsymbol{\alpha} + \mathbf{u} + \mathbf{e};\ \begin{bmatrix}\mathbf{e}\\ \mathbf{u}\\ \mathbf{y}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{Q}\boldsymbol{\alpha}\end{bmatrix}, \begin{bmatrix}\mathbf{R} & \mathbf{0} & \mathbf{R}\\ \mathbf{0} & (\boldsymbol{\Omega}_n \circ \mathbf{X}\mathbf{X}')\sigma^2_{\boldsymbol{\beta}} & (\boldsymbol{\Omega}_n \circ \mathbf{X}\mathbf{X}')\sigma^2_{\boldsymbol{\beta}}\\ \mathbf{R} & (\boldsymbol{\Omega}_n \circ \mathbf{X}\mathbf{X}')\sigma^2_{\boldsymbol{\beta}} & \mathbf{V}\end{bmatrix}\right);$$

$$\mathbf{V} = (\boldsymbol{\Omega}_n \circ \mathbf{X}\mathbf{X}')\sigma^2_{\boldsymbol{\beta}} + \mathbf{R}$$

where ○ is the Hadamard product; **y** is the typical *n*-vector of observed phenotypic values; **u** and **e** are the corresponding additive genetic effects and residuals, respectively. Individual-specific marker effects are therefore accounted for by multiplying each element of the relationship matrix $\mathbf{XX}'$ by the corresponding element of $\mathbf{\Omega}_n$, thereby reflecting correlations in marker effects among individuals' genetic backgrounds.

In general, we propose to infer $\mathbf{\Omega}_n = \left(\omega_{ij}\right)_{n\times n}$ by $\omega_{ij} = \kappa\left(\varphi(\mathbf{x}_i), \varphi\left(\mathbf{x}_j\right)\right)$, where φ is some function of the *m*-vectors of marker variables $\mathbf{x}_i$ and $\mathbf{x}_j$, for any pair of individuals *i* and *j*, and κ is a valid kernel function guaranteeing that $\mathbf{\Omega}_n$ be positive semi-definite. In *MPM-Mixture*, $\varphi(\mathbf{x}_i) = \mathbf{a}_i$ (*K*-vector of admixture coefficients for *i*) for any individual *i*, and the kernel function is $\kappa_\rho\left(\mathbf{a}_i, \mathbf{a}_j\right) = \rho\mathbf{a}_i'\mathbf{a}_j + (1-\rho)$, so that $\mathbf{\Omega}_n$ is a "mixture" between a matrix of correlations restricted to population clusters and a matrix allowing full exchange of information across clusters, as in a standard GBLUP model. More generally, one could define the kernel function as $\kappa_{\rho,\mathbf{\Theta}}\left(\mathbf{a}_i, \mathbf{a}_j\right) = \rho\mathbf{a}_i'\mathbf{\Theta}_K\mathbf{a}_j + (1-\rho)$, where $\mathbf{\Theta}_K$ is a *K* x *K* matrix depicting relationships among clusters. Here, we simply set $\mathbf{\Theta}_K = \mathbf{I}_K$ and adjusted the kernel function (by REML) for ρ only.

In *MPM-Matérn*, $\varphi(\mathbf{x}_i) = \mathbf{p}_i$ (*d*-vector of PC coordinates for *i*) for any individual *i*, and the kernel function is a Matérn function $\kappa_{\nu,h}\left(\mathbf{p}_i, \mathbf{p}_j\right)$ of $\left\|\mathbf{p}_i - \mathbf{p}_j\right\|_2$, where $\|.\|_2$ is the Euclidean norm. Notably, it can be shown that $\left\|\mathbf{p}_i - \mathbf{p}_j\right\|_2$ is proportional to $\left\|\boldsymbol{\pi}_{\mathbf{P}_i} - \boldsymbol{\pi}_{\mathbf{P}_j}\right\|_2$, where $\boldsymbol{\pi}_{\mathbf{P}_i}$ ($\boldsymbol{\pi}_{\mathbf{P}_j}$) is the *m*-vector of individual-specific allele frequencies for individuals *i* (*j*), defined by projection of matrix **X** onto the column space of $\mathbf{Q}_\mathbf{P} = [\mathbf{1}_n \quad \mathbf{P}]$ (Appendix A3). So $\left\|\mathbf{p}_i - \mathbf{p}_j\right\|_2$, which reflects differentiation with respect to coordinates at the leading PCs of **X**, also reflects differentiation with respect to individual-specific allele frequencies, with an underlying population structure represented by the same PCs. The allele frequencies $\boldsymbol{\pi}_{\mathbf{P}_i}$ have been introduced by Conomos *et al.* (2016), in a study where they also recommended using principal components from a subset of unrelated individuals in **X**. Here, we simply applied PCA on the whole matrix **X**.

A3. Relationship between distance based on principal components and distance based on individual-specific allele frequencies

In this section, $\mathbf{1}_t$, $\mathbf{I}_t$, $\mathbf{J}_t$ and $\mathbf{0}_{s\times t}$ refer to the vector of ones, identity matrix, matrix of ones and matrix of zeros, respectively, of dimensions t, $t \times t$, $t \times t$ and $s \times t$ (where *s* and *t* are specified).

We will consider the case where the PC matrix $\mathbf{P}$ consists of the first *d* PCs of $\mathbf{X}$, and individual-specific allele frequencies are defined as (Conomos *et al.* 2016):

$$\mathbf{\Pi_P} = \frac{1}{2}\mathbf{Q_P}(\mathbf{Q_P}'\mathbf{Q_P})^{-1}\mathbf{Q_P}'\mathbf{X}$$

where $\mathbf{Q_P} = [\mathbf{1}_n \quad \mathbf{P}]$ represents population structure through an intercept and the effects of the first *d* PCs of $\mathbf{X}$. Vector $\boldsymbol{\pi}_{\mathbf{P}_i}$ ($\boldsymbol{\pi}_{\mathbf{P}_j}$) then consists of individual-specific allele frequencies (with respect to $\mathbf{Q_P}$) for individual *i* (*j*), such that:

$$\boldsymbol{\pi}_{\mathbf{P}_i} = \frac{1}{2}\mathbf{q}_i'(\mathbf{Q_P}'\mathbf{Q_P})^{-1}\mathbf{Q_P}'\mathbf{X}$$

and similarly for $\boldsymbol{\pi}_{\mathbf{P}_j}$ ($\mathbf{q}_i$ refers to the $(d+1)$-vector of population-structure variables from $\mathbf{Q_P}$ for individual *i*).

We will show that $\|\mathbf{p}_i - \mathbf{p}_j\|_2 = 2\left\|\boldsymbol{\pi}_{\mathbf{P}_i} - \boldsymbol{\pi}_{\mathbf{P}_j}\right\|_2$ for any pair (*i*, *j*), i.e., Euclidean distances based on *d* PCs are equivalent, by proportionality, to those based on *m* individual-specific allele frequencies, with such frequencies as defined above.

$$\|\mathbf{p}_i - \mathbf{p}_j\|_2 = \sqrt{(\mathbf{p}_i' - \mathbf{p}_j')(\mathbf{p}_i - \mathbf{p}_j)} = \sqrt{\mathbf{p}_i'\mathbf{p}_i + \mathbf{p}_j'\mathbf{p}_j - 2\mathbf{p}_i'\mathbf{p}_j}$$

$$\left\|\boldsymbol{\pi}_{\mathbf{P}_i} - \boldsymbol{\pi}_{\mathbf{P}_j}\right\|_2 = \sqrt{\boldsymbol{\pi}_{\mathbf{P}_i}'\boldsymbol{\pi}_{\mathbf{P}_i} + \boldsymbol{\pi}_{\mathbf{P}_j}'\boldsymbol{\pi}_{\mathbf{P}_j} - 2\boldsymbol{\pi}_{\mathbf{P}_i}'\boldsymbol{\pi}_{\mathbf{P}_j}} = \frac{1}{2}\sqrt{\mathbf{q}_i'\mathbf{M}\mathbf{q}_i + \mathbf{q}_j'\mathbf{M}\mathbf{q}_j - 2\mathbf{q}_i'\mathbf{M}\mathbf{q}_j}$$

where $\mathbf{M} = (\mathbf{Q_P}'\mathbf{Q_P})^{-1}\mathbf{Q_P}'\mathbf{X}\mathbf{X}'\mathbf{Q_P}(\mathbf{Q_P}'\mathbf{Q_P})^{-1}$.

Below, we will specify $\mathbf{M}$ more explicitly, to subsequently show that $\|\mathbf{p}_i - \mathbf{p}_j\|_2 = 2\left\|\boldsymbol{\pi}_{\mathbf{P}_i} - \boldsymbol{\pi}_{\mathbf{P}_j}\right\|_2$.

Let $\left(\mathbf{I}_n - \frac{\mathbf{J}_n}{n}\right)\mathbf{X}$ be the matrix of marker variables centered around their respective overall mean. Assuming $m > n$, by eigenvalue decomposition $\left(\mathbf{I}_n - \frac{\mathbf{J}_n}{n}\right)\mathbf{X}\mathbf{X}'\left(\mathbf{I}_n - \frac{\mathbf{J}_n}{n}\right) = \mathbf{U}\mathbf{\Lambda}\mathbf{U}'$, with $\mathbf{U}$ the $n \times n$ matrix of eigenvectors and $\mathbf{\Lambda}$ the $n \times n$ diagonal matrix of eigenvalues of $\left(\mathbf{I}_n - \frac{\mathbf{J}_n}{n}\right)\mathbf{X}\mathbf{X}'\left(\mathbf{I}_n - \frac{\mathbf{J}_n}{n}\right)$; and $\mathbf{P} = \mathbf{U}_d\mathbf{\Lambda}_d^{1/2}$, where the $\mathbf{U}_d$ is the $n \times d$ matrix of leading eigenvectors and $\mathbf{\Lambda}_d^{1/2}$ is the $d \times d$ diagonal matrix of corresponding singular values, assumed strictly positive.

Because $\mathbf{U}_d$ consist of left-eigenvectors of a column-centered matrix (associated with strictly positive eigenvalues), $\mathbf{U}_d'\mathbf{1}_n = \mathbf{0}_{d\times 1}$ so $\mathbf{P}'\mathbf{1}_n = \mathbf{\Lambda}_d^{1/2}\mathbf{U}_d'\mathbf{1}_n = \mathbf{0}_{d\times 1}$.
Besides, $\mathbf{P}'\mathbf{P} = \mathbf{\Lambda}_d$.

So:

$$(\mathbf{Q_P}'\mathbf{Q_P})^{-1} = \begin{bmatrix} \mathbf{1}_n'\mathbf{1}_n & \mathbf{1}_n'\mathbf{P} \\ \mathbf{P}'\mathbf{1}_n & \mathbf{P}'\mathbf{P} \end{bmatrix}^{-1} = \begin{bmatrix} n^{-1} & \mathbf{0}_{1\times d} \\ \mathbf{0}_{d\times 1} & \mathbf{\Lambda}_d^{-1} \end{bmatrix}$$

Moreover $\mathbf{Q_P'X} = \begin{bmatrix} \mathbf{1}_n' \\ \mathbf{P}' \end{bmatrix}\left(\left(\mathbf{I}_n - \frac{\mathbf{J}_n}{n}\right)\mathbf{X} + \frac{\mathbf{J}_n}{n}\mathbf{X}\right) = \begin{bmatrix} \mathbf{1}_n'\mathbf{X} \\ \mathbf{P}'\left(\mathbf{I}_n - \frac{\mathbf{J}_n}{n}\right)\mathbf{X} \end{bmatrix}$ because $\mathbf{P}'\mathbf{J}_n = (\mathbf{P}'\mathbf{1}_n)\mathbf{1}_n' = \mathbf{0}_{d\times n}$.

So

$$(\mathbf{Q_P}'\mathbf{Q_P})^{-1}\mathbf{Q_P'X} = \begin{bmatrix} \boldsymbol{\mu}' \\ \mathbf{\Lambda}_d^{-1/2}\mathbf{U}_d'\left(\mathbf{I}_n - \frac{\mathbf{J}_n}{n}\right)\mathbf{X} \end{bmatrix}$$

where $\boldsymbol{\mu}' = \frac{1}{n}\mathbf{1}_n'\mathbf{X}$

Finally, $\mathbf{M} = ((\mathbf{Q_P}'\mathbf{Q_P})^{-1}\mathbf{Q_P}'\mathbf{X})(\mathbf{X}'\mathbf{Q_P}(\mathbf{Q_P}'\mathbf{Q_P})^{-1}) = \begin{bmatrix} \boldsymbol{\mu}'\boldsymbol{\mu} & \mathbf{a}' \\ \mathbf{a} & \mathbf{I}_d \end{bmatrix}$

with:

$\mathbf{\Lambda}_d^{-1/2}\mathbf{U}_d'\left(\mathbf{I}_n - \frac{\mathbf{J}_n}{n}\right)\mathbf{X}\mathbf{X}'\left(\mathbf{I}_n - \frac{\mathbf{J}_n}{n}\right)\mathbf{U}_d\mathbf{\Lambda}_d^{-1/2} = \mathbf{\Lambda}_d^{-1/2}(\mathbf{U}_d'\mathbf{U}\mathbf{\Lambda}\mathbf{U}'\mathbf{U}_d)\mathbf{\Lambda}_d^{-1/2} = \mathbf{\Lambda}_d^{-1/2}\mathbf{\Lambda}_d\mathbf{\Lambda}_d^{-1/2} = \mathbf{I}_d$

$\mathbf{a} = \mathbf{\Lambda}_d^{-1/2}\mathbf{U}_d'\left(\mathbf{I}_n - \frac{\mathbf{J}_n}{n}\right)\mathbf{X}\boldsymbol{\mu}$

Therefore, for any pair of individuals (*i*, *j*):

$$\mathbf{q}_i'\mathbf{M}\mathbf{q}_j = [1 \quad \mathbf{p}_i']\mathbf{M}\begin{bmatrix} 1 \\ \mathbf{p}_j \end{bmatrix} = \boldsymbol{\mu}'\boldsymbol{\mu} + \mathbf{p}_i'\mathbf{a} + \mathbf{a}'\mathbf{p}_j + \mathbf{p}_i'\mathbf{p}_j$$

So:

$$2\left\|\boldsymbol{\pi}_{\mathbf{P}i} - \boldsymbol{\pi}_{\mathbf{P}j}\right\|_2 = \sqrt{\mathbf{q}_i'\mathbf{M}\mathbf{q}_i + \mathbf{q}_j'\mathbf{M}\mathbf{q}_j - 2\mathbf{q}_i'\mathbf{M}\mathbf{q}_j}$$

$$= \sqrt{(\boldsymbol{\mu}'\boldsymbol{\mu} + 2\mathbf{p}_i'\mathbf{a} + \mathbf{p}_i'\mathbf{p}_i) + (\boldsymbol{\mu}'\boldsymbol{\mu} + 2\mathbf{p}_j'\mathbf{a} + \mathbf{p}_j'\mathbf{p}_j) - 2\boldsymbol{\mu}'\boldsymbol{\mu} - 2\mathbf{p}_i'\mathbf{a} - \mathbf{2a}'\mathbf{p}_j - 2\mathbf{p}_i'\mathbf{p}_j}$$

$$= \sqrt{\mathbf{p}_i'\mathbf{p}_i + \mathbf{p}_j'\mathbf{p}_j - 2\mathbf{p}_i'\mathbf{p}_j} = \left\|\mathbf{p}_i - \mathbf{p}_j\right\|_2$$