Joint genome-wide prediction in several populations accounting for randomness of genotypes: A hierarchical Bayes approach. I: Multivariate Gaussian priors for marker effects and derivation of the joint probability mass function of genotypes

Carlos Alberto Martínez[a,b], Kshitij Khare[b], Arunava Banerjee[c], Mauricio A. Elzo[a]

[a]Department of Animal Sciences

[b]Department of Statistics

[c]Department of Computer and Information Science and Engineering

University of Florida, Gainesville, FL, USA

Correspondence: Carlos Alberto Martínez, Department of Animal Sciences, University of Florida, Gainesville, FL 32611, USA.

Tel: 352-328-1624.

Fax: 352-392-7851.

E-mail: carlosmn@ufl.edu

Abstract

It is important to consider heterogeneity of marker effects and allelic frequencies in across population genome-wide prediction studies. Moreover, all regression models used in genome-wide prediction overlook randomness of genotypes. In this study, a family of hierarchical Bayesian models to perform across population genome-wide prediction modeling genotypes as random variables and allowing population-specific effects for each marker was developed. Models shared a common structure and differed in the priors used and the assumption about residual variances (homogeneous or heterogeneous). Randomness of genotypes was accounted for by deriving the joint probability mass function of marker genotypes conditional on allelic frequencies and pedigree information. As a consequence, these models incorporated kinship and genotypic information that not only permitted to account for heterogeneity of allelic frequencies, but also to include individuals with missing genotypes at some or all loci without the need for previous imputation. This was possible because the non-observed fraction of the design matrix was treated as an unknown model parameter. For each model, a simpler version ignoring population structure, but still accounting for randomness of genotypes was proposed. Implementation of these models and computation of some criteria for model comparison were illustrated using two simulated datasets. Theoretical and computational issues along with possible applications, extensions and refinements were discussed. Some features of the models developed in this study make them promising for genome-wide prediction, the use of information contained in the probability distribution of genotypes is perhaps the most appealing. Further studies to assess the performance of the models proposed here and also to compare them with conventional models used in genome-wide prediction are needed.

Key words: Across population genome-enabled prediction; Bayesian modeling; heterogeneous allelic frequencies; distribution of genotypes.

1. Introduction

The use of molecular markers located across the whole genome for prediction of breeding values (Meuwissen et al., 2001) and phenotypes (Goddard and Hayes, 2007, Gianola et al., 2009) has proven to be a useful tool in animals (Hayes et al., 2009), humans (Guttmacher et al., 2002; de los Campos et al., 2010) and plants (Bernardo and Yu, 2007; Desta and Ortiz, 2014). This success has given rise to a tremendous amount of research in the area of statistical genomics in order to obtain better genome-wide predictions (Goddard and Hayes, 2007; Gianola, 2013; Hill, 2014; Gianola and Rosa, 2015).

Most of the methods have been developed for prediction in a single population. Across population studies usually use predictions obtained from individual populations or pool data to perform a single analysis (de Roos et al., 2009). On one hand, pooling data and performing a single analysis may increase the accuracy of genome-wide prediction because the number of records has an important impact on it (Meuwissen et al., 2001; Goddard, 2009; Zhong et al., 2009). On the other hand, it may decrease accuracy when the effects of QTL controlling the trait are not the same across populations (de Roos et al., 2009; van den Berg et al., 2015; Wientjes et al., 2015).

Analyzing data from Holstein cattle performing in different European countries, Lund et al. (2011) reported that pooling data and carrying out a single analysis increased the accuracy of genomic predictions. With simulated data, de Roos et al. (2009) found that pooling data was beneficial when populations had diverged by few generations, marker density was high and heritability was low, but for more distant populations and less dense marker panels they found a small decrease in accuracy. Using simulated data, Wientjes et al. (2015) studied the effect of differences in QTL allele substitution effects across populations on the accuracy of genome-wide prediction. They found that when allele substitution effects changed across populations, the accuracies decreased in proportion to the genetic correlation between populations. Using the same dataset, van den Berg et al. (2015) looked for across population genomic prediction scenarios under which Bayesian variable selection models had a better performance than genomic BLUP (GBLUP). They concluded that Bayesian variable selection models outperform GBLUP when the number of QTL is small as in single population analyses, but the difference in accuracy is larger in the across population case.

None of these studies allowed marker effects to differ from one population to another. However, de Roos et al. (2009) highlighted the need for alternative methods that allow population-specific estimation of allele substitution effects in across population genome wide prediction. Chen et al. (2014) proposed a Bayesian model with different SNP effects for each population that permits sharing information across populations through a common set of latent variables indicating weather a given marker is associated with a QTL or not. They did not model covariance matrices of marker effects explicitly. With real and simulated data they found that this model increased the accuracy of across population genome-wide prediction, especially when the number of QTL was small and correlations among QTL effects from different populations were high. Recently, Bayesian models that account for genetic heterogeneity have been proposed. Multivariate models considering correlated population specific marker effects were developed by Lehermeir et al. (2015) while de los Campos et al. (2015a) proposed a model with main marker effects and interactions. Using real data from three plant populations, Lehermeir et al. (2015) found cases in which the strategy of pooling data and ignoring structure performed better and others where the multivariate models yielded better predictive performance. For example, in highly differentiated populations within group and multivariate analyses performed better. Using real datasets from pigs and wheat, de los Campos et al. (2015a) found modest superiority of the interaction model relative to the model using pooled data and the model that analyzed each subpopulation separately. Similar studies have implemented multivariate models in multibreed dairy cattle populations (Karoui et al., 2012; Olson et al., 2012; Makgahlela et al., 2013). Huang et al. (2014) used non-linear models to perform genome wide prediction in layer hens when the reference population was comprised by individuals from several breeds or lines and compared them with a multiple-trait GBLUP model. They found that the various models used had a similar predictive performance.

If several populations are to be evaluated simultaneously, the possible existence of genotype by environment interaction, lack of persistence of linkage phase and variation in allelic frequencies across populations indicate the need for an analysis that accounts for the fact that combining them creates a structured complete population. It has been reported that population structure may act as an effect modifier (de los Campos et al., 2015a). Furthermore, it has to be considered that not only the allele substitution effects of a particular locus in different populations may be correlated, but also its frequencies in each population (e.g., due to gene flow).

Another feature that has been overlooked in the random linear regression models used in genome-wide prediction is the randomness of the matrix containing a one to one mapping from the set of genotypes to a subset of the integers, namely the design matrix. This matrix is treated as fixed in genome-wide prediction models, while in classical quantitative genetics theory it is treated as random (Falconer and Mackay, 1996; Lynch and Walsh, 1998). Besides being in agreement with the classical theory, taking into account the randomness of this matrix, that is, the randomness of genotypes, permits the estimation of allelic frequencies because when treated as an observable discrete random matrix, its probability mass function (pmf) depends on the allelic frequencies. Thus, under a Bayesian setting, allelic frequencies are treated as random because these are unknown parameters. Further, the works of Wright (1930; 1937) provide additional support to treat allelic frequencies as random variables making Bayesian inference even more attractive.

Thus, the objective of this study was to propose hierarchical Bayesian models to carry out simultaneous genome-wide prediction in several populations accounting for randomness of marker genotypes, heterogeneity and correlation of allelic frequencies across populations, and population-specific allelic substitution effects.

2. Methods

2.1 The models

Hereinafter the complete population or simply the population is defined as the set of individuals with phenotypes considered in the study. Suppose that there exists some criterion (e.g., environment, race, breed, line, etc.) to split this population into $\mathcal{S}$ subpopulations. To make the problem more tractable, some simplifying assumptions are made. The first one is linkage equilibrium. The second one is Hardy-Weinberg equilibrium. The third one is that starting from the oldest individuals with phenotypes, the pedigree is fully known. Lastly, mutations are ignored.

The basic linear model used to describe the relationship between response variables and marker allele substitution effects is $\boldsymbol{y} = W\boldsymbol{g} + \boldsymbol{e}$, where $\boldsymbol{y}$ is a vector containing dependent variables (e.g., records corrected for non-genetic factors), W is an observable random matrix containing a one to one mapping from individual marker genotypes to a subset of the integers to be defined later, $\boldsymbol{g}$ is an unknown random vector of marker allelic substitution effects for every population and $\boldsymbol{e}$ is a random vector of residuals. A more detailed notation is the following. If records are sorted by subpopulation as well as the columns of W and the elements of $\boldsymbol{g}$, then for every $l = 1,2,\dots,\mathcal{S}$, $\boldsymbol{y}_l = W_l\boldsymbol{g}_l + \boldsymbol{e}_l$, with dimensions: $(\boldsymbol{y}_l)_{n_l\times 1}$, $(W_l)_{n_l\times m}$, $(\boldsymbol{g}_l)_{m\times 1}$ and $(\boldsymbol{e}_l)_{n_l\times 1}$ where n_l is the sample size of subpopulation l, and m is the number of marker loci. Thus, the total sample size is $n = \sum_{l=1}^{\mathcal{S}} n_l$.

The scenario where only a part of matrix W is observed because some individuals are not genotyped or individuals are genotyped for different numbers of marker loci is also considered. This is done by treating this non-observed part of W as a parameter in the model as it will be explained later.

The case of diploid individuals and biallelic marker loci is considered. The effect of every marker locus is defined as the regression of records on a function of the number of copies of the reference allele and in quantitative genetics it corresponds to the allele substitution effect (Falconer and Mackay, 1996; Lynch and Walsh, 1998). The number of copies can be "centered" at zero giving the following codification. Let A and B be the marker alleles at each locus and let B be the reference allele. Then:

$$W_l = \{w_{ij}^l\}_{n_l \times m} = \begin{cases} 1, if\ genotype = BB \\ 0, if\ genotype = AB \\ -1, if\ genotype = AA \end{cases}.$$

Different versions of the hierarchy that represents the stochastic component of each model were considered. Models vary according to the assumptions on the variance of residuals and the priors posed over the marker effects. The most parsimonious model is the one considering homoscedastic residuals and homogeneous marker effect covariance matrices. The hierarchical Bayesian model assuming homoscedastic residuals and multivariate Gaussian priors for marker effects has the following structure:

$$\boldsymbol{y}|W, \boldsymbol{g}, \sigma^2 \sim MVN(W\boldsymbol{g}, \sigma^2 I)$$

$$W|\boldsymbol{p}_1^*, \boldsymbol{p}_2^*, \dots, \boldsymbol{p}_m^* \sim \pi(\cdot|\boldsymbol{p}_1^*, \boldsymbol{p}_2^*, \dots, \boldsymbol{p}_m^*)$$

$$\boldsymbol{p}_j^* \overset{iid}{\sim} \pi(\boldsymbol{p}^*), j = 1,2, \dots, m$$

$$\sigma^2 \sim Inverse\ Gamma\left(\frac{\tau^2}{2}, \frac{v}{2}\right) := IG\left(\frac{\tau^2}{2}, \frac{v}{2}\right)$$

$$\boldsymbol{g}|G \sim MVN(0, G), G = Block\ Diag\ \{G_j\}_{j\ =\ 1}^{m}$$

$$G_j \overset{iid}{\sim} Inverse\ Wishart(a, \boldsymbol{\Sigma}) := IW(a, \boldsymbol{\Sigma})$$

$$G_j = \begin{bmatrix} \sigma_{j_1}^2 & \sigma_{j_{1,2}} & \cdots & \sigma_{j_{1,S}} \\ & \sigma_{j_2}^2 & \cdots & \sigma_{j_{2,S}} \\ & & \ddots & \vdots \\ sym & & & \sigma_{j_S}^2 \end{bmatrix}$$

where σ^2 is the residual variance, $\sigma_{j_l}^2$ is the variance of the effect of the j^{th} marker in the l^{th} subpopulation, $\sigma_{j_{l,l'}}$ is the covariance between effects of marker j in subpopulations l and l', $\boldsymbol{p}_j^*$ is a parameter associated with allelic frequencies of the j^{th} marker in each subpopulation and $\pi(\boldsymbol{p}^*)$ is its density. Details on these parameters and their probability density function (pdf) are given later.

In the case of heterogeneous residual variances across subpopulations, residual variances $\sigma_1^2, \dots, \sigma_{\mathcal{S}}^2$ are given independent $IG\left(\frac{\tau^2}{2}, \frac{v}{2}\right)$ priors and then: $\boldsymbol{y}|W, \boldsymbol{g}, R \sim MVN(W\boldsymbol{g}, V)$, $R = (\sigma_{e1}^2, \dots, \sigma_{e\mathcal{S}}^2)$ and $V = Block\ Diag.\{\sigma_{el}^2 I_{n_l}\}_{l\,=\,1}^{\mathcal{S}}$. Hill (1984) found that in the presence of heterogeneous environmental variances, across population analyses assuming homogenous residuals variances yielded an excess of individuals selected from populations with higher environmental variances. This is why heterogeneity of residual variances across subpopulations was considered in this study.

The general framework assumes that in each subpopulation there is a fraction of genotyped individuals and a fraction of non-genotyped or partially genotyped individuals. Let $W^{\mathscr{o}}$ and W^N denote the observed (data) and non-observed (an unknown parameter) parts of W. Let $P^* = (\boldsymbol{p}_1^*, \boldsymbol{p}_2^*, \dots, \boldsymbol{p}_m^*)$; therefore, $\pi(W|P^*) = \pi(W^{\mathscr{o}}, W^N|P^*)$ can be expressed as: $f(W^{\mathscr{o}}|W^N, P^*)\pi(W^N|P^*)$. Thus, the full likelihood has the form:

$$f(\boldsymbol{y}, W^{\mathscr{o}}|W^N, \boldsymbol{g}, R, P^*) = f(\boldsymbol{y}|W^{\mathscr{o}}, W^N, \boldsymbol{g}, R, P^*) f(W^{\mathscr{o}}|W^N, \boldsymbol{g}, R, P^*)$$
$$= f(\boldsymbol{y}|W, \boldsymbol{g}, R) f(W^{\mathscr{o}}|W^N, P^*).$$

Henceforth, $f(\boldsymbol{y}|W, \boldsymbol{g}, R)$ will be referred to as the $\boldsymbol{y}$ component of the likelihood and $f(W^{\mathscr{o}}|W^N, P^*)$ will be referred to as the W component.

The simplest case for the covariance matrix of marker effects is $G = I \otimes G^0$. Under this setting the assumption is that the covariance structure is the same for all markers. This is statistically convenient due to the fact that the number of covariance parameters is reduced. Further, in analysis considering a single population, it has been found that specifying a different variance for each marker does not allow too much Bayesian learning about marker effect variances (Gianola et al., 2009). Here, models assigning the same covariance matrix to the effects of all marker loci and models considering a different covariance matrix for the effects of each marker locus were considered and these models were referred to as homogeneous marker effect covariance matrix models and heterogeneous marker effect covariance matrix models. Let $\mathcal{P}_{\mathcal{S}}^{+}$ denote the space of symmetric positive definite matrices of dimension $\mathcal{S} \times \mathcal{S}$. Then, the marginal prior distribution of $\boldsymbol{g}$ is:

$$\pi(\boldsymbol{g}) = \int_{\mathcal{P}_{\mathcal{S}}^{+}} \pi(\boldsymbol{g}|G^0)\pi(G^0)dG^0 \propto \frac{1}{\left|\boldsymbol{\Sigma} + \sum_{j=1}^{m} \boldsymbol{g}_j \boldsymbol{g}_j'\right|^{\left(\frac{a+m}{2}\right)}}.$$

For details, see Appendix A. Similarly, for the heterogeneous marker effect covariance matrix model it can be shown (appendix A) that: $\pi(\boldsymbol{g}) \propto \frac{1}{\prod_{j=1}^{m}\left(1+\frac{1}{a+1-\mathcal{S}}\boldsymbol{g}_j'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{g}_j\right)^{\left(\frac{a+1}{2}\right)}}$, which is the product of m multivariate t distributions with scale matrix $\boldsymbol{\Sigma}_* = \frac{1}{a+1-\mathcal{S}}\boldsymbol{\Sigma}$ and degrees of freedom $a + 1 - \mathcal{S}$; therefore, under this prior, marker effects are marginally independent and identically distributed. At this point, the following remark can be made.

Remark 1 Under the assumption of homogeneous marker effect covariance matrices, *a priori* the marker effects are marginally dependent. This happens because when integrating with respect to the common covariance matrix G^0, the term $\sum_{j=1}^{m} \boldsymbol{g}_j \boldsymbol{g}_j'$ and the hyper-hyperparameter $\boldsymbol{\Sigma}$ are factored, resulting in a function that cannot be written as the product of m functions, each one depending on a different $\boldsymbol{g}_j$. Moreover, the joint prior density is not standard.

To take into account the belief that allelic frequencies of the same marker vary across subpopulations and may be correlated, the prior $\pi(\boldsymbol{p}^*)$ is built based on a Dirichlet distribution. To do that, the allelic frequency of the reference allele in marker locus j in subpopulation l has to be expressed on a complete population basis, that is, p_{lj} is expressing the frequency of the reference allele in locus j in subpopulation l relative not to subpopulation l, but to the complete population. Thus, the frequencies of the two alleles at a given marker locus and a given subpopulation do not add to one, but to some sort of relative frequency of that subpopulation in that locus denoted as r_{lj}. Let $\boldsymbol{r} = (\boldsymbol{r}_1, \dots, \boldsymbol{r}_{\mathcal{S}}), \boldsymbol{r}_l = (r_{l1}, \dots, r_{lm}), l = 1,2, \dots, \mathcal{S}$. With this parameterization $\sum_{l=1}^{\mathcal{S}} p_{lj} \leq 1, \forall\, j = 1,2, \dots, m$, with equality if and only if the reference allele is fixed in all subpopulations. Conversely, allelic frequencies expressed on a subpopulation basis satisfy the constraint that the sum of the frequencies of the two alleles at each marker locus equals one within each subpopulation. Let $q_{jl}, j = 1,2, \dots, m, l = 1,2, \dots, \mathcal{S}$, be the frequencies of the non-reference alleles expressed on a complete population basis, then $p_{lj} + q_{lj} = r_{lj}$. The two parameterizations of allelic frequencies are related by the one to one mapping $p^*_{lj} = p_{lj}/r_{lj}$.

Consider the case when $\boldsymbol{r}$ is known and $r_{l1} = \cdots = r_{lm} = r_l \; \forall \, l$. Then, elements of vector $\boldsymbol{r} = (r_1, \dots, r_{\mathcal{S}})$ can be seen as subpopulation weights, that is, they are related to subpopulation sizes. By $\boldsymbol{r}$ being known, it is meant that it is either actually known or it is specified following some assumption. A pragmatic decision would be to assign equal subpopulation weights, an assumption that was also made in other studies (e.g., Gianola et al. 2010). Once $\boldsymbol{r}$ has been specified, there is an extra restriction over each $\boldsymbol{p}_j = \left(p_{1j}, \dots, p_{\mathcal{S}j}\right)$. For $l = 1,2, \dots, \mathcal{S}$ the following condition must be satisfied: $p_{lj} \leq r_l$. Therefore, the support of the distribution of $\boldsymbol{p}_j$ given $\boldsymbol{r}$ is $\Omega^{\boldsymbol{r}}_j \coloneqq \left\{\boldsymbol{p}_j \in \mathbb{R}^{\mathcal{S}} \middle| 0 < p_{lj} \leq r_l \; \forall \, l, \sum_{l=1}^{\mathcal{S}} r_l = 1\right\}$. Notice that the condition $\sum_{l=1}^{\mathcal{S}} r_l = 1$ implies that vectors in $\Omega^{\boldsymbol{r}}_j$ satisfy $\sum_{l=1}^{\mathcal{S}} p_{lj} \leq 1$. Thus, under this approach the prior used for each $\boldsymbol{p}_j$ is one corresponding to a scaled Dirichlet random vector. If $\boldsymbol{\beta} = (\beta_1, \dots, \beta_{\mathcal{S}}) \sim Dirichlet(\boldsymbol{\alpha})$, $\boldsymbol{\alpha} \in \mathbb{R}^{\mathcal{S}+1}$, then the prior assigned to $\boldsymbol{p}_j$ is the distribution of vector $(\beta_1 r_1, \dots, \beta_{\mathcal{S}} r_{\mathcal{S}})$ which clearly pertains to $\Omega^{\boldsymbol{r}}_j$. Then, the pdf $\pi\left(\boldsymbol{p}_j | \boldsymbol{r}\right)$ is derived using standard results from the theory of distributions of transformations of random variables (Casella and Berger, 2002). This derivation is simplified by the fact that the transformation is linear and therefore the Jacobian is constant. It follows that: $\pi\left(\boldsymbol{p}_j | \boldsymbol{r}\right) \propto \prod_{l=1}^{\mathcal{S}} \left\{\left(\frac{p_{lj}}{r_l}\right)^{\alpha_l - 1}\right\} p_{(\mathcal{S}+1)j}^{\alpha_{\mathcal{S}+1} - 1}$, where $p_{(\mathcal{S}+1)j} = 1 - \sum_{l=1}^{\mathcal{S}} \frac{p_{lj}}{r_l}$.

The second approach is to assume that $\boldsymbol{r}$ is unknown. The density $\pi(\boldsymbol{p}|\boldsymbol{r})$ could be used and a Dirichlet distribution could be assigned to each $\boldsymbol{r}_j$ adding one more level to the hierarchy. However, using $p_{lj} + q_{lj} = r_{lj}$ and properties of the Dirichlet distribution, the following strategy allows assigning a prior to allelic frequencies and the weights $\boldsymbol{r}$ without putting an extra level in the hierarchy. To this end it is assumed that r_{lj} varies for each j and each l. A $Dirichlet\left(\left(\boldsymbol{\alpha}_p, \boldsymbol{\alpha}_q\right)\right)$ prior is posed over $\left(\boldsymbol{p}_j, \boldsymbol{q}_j\right)$, where $\boldsymbol{q}_j$ is the analog of $\boldsymbol{p}_j$ for the non-reference allele at each locus and $\boldsymbol{\alpha}_p = \left(\alpha_{1p}, \dots, \alpha_{\mathcal{S}p}\right), \boldsymbol{\alpha}_q = \left(\alpha_{1q}, \dots, \alpha_{\mathcal{S}q}\right)$. Consequently, by properties of the Dirichlet distribution it follows that $\boldsymbol{r}_j \sim Dirichlet\left(\left(\alpha_{1p} + \alpha_{1q}, \dots, \alpha_{\mathcal{S}p} + \alpha_{\mathcal{S}q}\right)\right)$.

2.1.1 Deriving the joint pmf of marker genotypes conditional on allelic frequencies

Given the kinship structure of a population (i.e., the pedigree) one can find several generations comprised of genotyped, partially genotyped and non-genotyped individuals. Therefore, the approach is to derive the pmf of the complete matrix W, i.e., the joint pmf of individuals with phenotypic records. Under this setting, m is the total number of marker loci to be included in the analysis (it usually corresponds to the size of the densest marker panel used in the population).

Across columns, that is, across marker loci, the problem is simplified by assuming linkage equilibrium, which implies independence of genotypes at different loci. Therefore, for an arbitrary subpopulation, the joint density of its column vectors is simply the product of their marginal pmf. When considering all subpopulations, the same assumption implies that marker genotypes at different loci are independent. The following derivations hold for any of the previously discussed approaches to model allelic frequencies distributions. Under the assumption of Hardy-Weinberg equilibrium it follows that marginally:

$$w_{ij}^{l} \middle| p_{lj}^{*} \sim \begin{cases} 1, with \quad probability \quad p_{lj}^{*2} \\ 0,\ with\, probability\ 2p_{lj}^{*}(1 - p_{lj}^{*}) \\ -1,\ with\, probability\ (1 - p_{lj}^{*})^2 \end{cases}$$

Recall that $p_{lj}^{*} = p_{lj}/r_{lj}$. Notice that p_{lj}^{*} is used instead of p_{lj} because it allows defining a proper pmf in the sense that the sum of the probabilities of the three possible values of w_{ij}^{l} equals one (which does not happen when using p_{lj}). The pmf $\pi\left(w_{ij}^{l} \middle| p_{lj}^{*}\right)$ can be also written as:

$$\pi\left(w_{ij}^{l}\middle|p_{lj}^{*}\right)=\left(p_{lj}^{*2}\right)^{I_{1i}}\left(2p_{lj}^{*}(1-p_{lj}^{*})\right)^{I_{0i}}\left((1-p_{lj}^{*})^{2}\right)^{I_{-1i}},$$

where I_{zi} is the indicator variable of the mutually exclusive events $w_{ij}^{l}=z, z\in\{-1,0,1\}$. By the linkage equilibrium assumption it follows that for individual i in population l: $\pi\left(\boldsymbol{w}_{i}^{l}\middle|\boldsymbol{p}_{j}^{*}\right)=\prod_{j=1}^{m}\pi\left(w_{ij}^{l}\middle|p_{lj}^{*}\right)$.

The rows of matrix W represent individuals with records. Because of the kinship between them, the genotype of a given individual is not independent of the genotype of their relatives. Furthermore, this non-independence has to be considered across subpopulations (e.g., half or full sibs may pertain to different subpopulations). This approach is based on the pedigree of the complete population. The “base” animals or “founders” can be pragmatically defined as the oldest individuals with phenotypic records and those individuals with phenotypes and unknown parents. To facilitate computations, it is assumed that these individuals are unrelated. Hereinafter this set is referred to as the base population, and individuals in this set are referred to as founders or base individuals. The remaining individuals in the population are referred to as non-founders. This pmf could be derived ignoring pedigree information which is equivalent to mutual independence of the rows of W, then $\pi(W|P^{*})=\prod_{j=1}^{m}\prod_{l=1}^{S}\prod_{i=1}^{n_l}\pi\left(w_{ij}^{l}\middle|p_{lj}^{*}\right)$. However, this would ignore information contained in the pedigree and would unnecessarily make the parametric space of W^{N} larger, which does not seem to be the best way to proceed.

The ordering of individuals is arbitrary, but a convenient way to do it here is according to the pedigree in such a way that the founders are given the first indices. For marker locus j in population l the target is to find:

$$\pi\left(\boldsymbol{w}_{j}^{l}\middle|p_{lj}^{*}\right)=\pi\left(w_{1j}^{l},w_{2j}^{l},\dots,w_{n_lj}^{l}\middle|p_{lj}^{*}\right)=P\left(w_{1j}^{l}=\omega_1,w_{2j}^{l}=\omega_2,\dots,w_{n_lj}^{l}=\omega_{n_l}\middle|p_{lj}^{*}\right)$$

with $\omega_i\in\{-1,0,1\}, 1\le i\le n_l$. This joint pmf can be written as:

$$\pi\left(\boldsymbol{w}_{j}^{l}\middle|p_{lj}^{*}\right)=\pi\left(w_{n_lj}^{l}\middle|w_{1j}^{l},\dots,w_{(n_l-1)j}^{l},p_{lj}^{*}\right)\pi\left(w_{1j}^{l},\dots,w_{(n_l-1)j}^{l}\middle|p_{lj}^{*}\right)$$

$$= \pi\left(w^l_{n_l j} \middle| w^l_{1j}, \dots, w^l_{(n_l-1)j}, p^*_{lj}\right) \pi\left(w^l_{(n_l-1)j} \middle| w^l_{1j}, \dots, w^l_{(n_l-2)j}, p^*_{lj}\right) \pi\left(w^l_{1j}, \dots, w^l_{(n_l-2)j} \middle| p^*_{lj}\right)$$

$$= \pi\left(w^l_{n_l j} \middle| w^l_{1j}, \dots, w^l_{(n_l-1)j}, p^*_{lj}\right) \cdots \pi\left(w^l_{1j} \middle| p^*_{lj}\right)$$

$$= \prod_{i=0}^{n_l-2} \left\{ \pi\left(w^l_{(n_l-i)j} \middle| w^l_{1j}, \dots, w^l_{(n_l-i-1)j}, p^*_{lj}\right) \right\} \pi\left(w^l_{1j} \middle| p^*_{lj}\right).$$

When considering all the m marker loci we have:

$$\pi(W^l | \boldsymbol{p}^*_l) = \prod_{i=0}^{n_l-2} \left\{ \pi\left(\boldsymbol{w}^l_{n_l-i} \middle| \boldsymbol{w}^l_1, \dots, \boldsymbol{w}^l_{n_l-i-1}, \boldsymbol{p}^*_l\right) \right\} \pi\left(\boldsymbol{w}^l_1 \middle| \boldsymbol{p}^*_l\right),$$

where each one of the pmf $\pi\left(\boldsymbol{w}^l_{n_l-i}, \middle| \boldsymbol{w}^l_1, \dots, \boldsymbol{w}^l_{n_l-i-1}, \boldsymbol{p}^*_l\right)$ is the product: $\prod_{j=1}^{m} \pi\left(w^l_{(n_l-i)j}, \middle| w^l_{1j}, \dots, w^l_{(n_l-i-1)j}, p^*_{lj}\right), 0 \leq i \leq n_l - 2$ and $\pi\left(\boldsymbol{w}^l_1 \middle| \boldsymbol{p}^*_l\right) = \prod_{j=1}^{m} \pi\left(w^l_{1j} \middle| p^*_{lj}\right)$.

Now, a conditional independence argument is used to simplify $\pi(W^l | \boldsymbol{p}^*_l)$. Given the genotypes of the parents of individual i, its genotype is independent of the genotype of collateral relatives and other ancestors. It is possible that the parents of individual i in population l pertain to subpopulations l^* and l'. Thus, at this point the complete population is considered. In addition, notice that given the parental genotypes, the genotype of an individual does not depend on the allelic frequencies because this conditional pmf is determined using basic segregation rules (see Appendix A). From these arguments it follows that for individual $i, \pi(\boldsymbol{w}_i | \boldsymbol{w}_1, \dots, \boldsymbol{w}_{i-1}, P^*) = \pi\left(\boldsymbol{w}_i \middle| \boldsymbol{w}_{S_i}, \boldsymbol{w}_{D_i}\right)$, where $\boldsymbol{w}_{S_i}$ and $\boldsymbol{w}_{D_i}$ are the genotypes of the parents of individual i. The pmf of non-founder genotypes at marker locus j conditioned on their parental genotypes is presented in Appendix A. Therefore, $\pi(W | P^*)$ can be written as $\pi(W | P^*) = \pi(W_{NF} | W_F) \pi(W_F | P^*)$ where W_F is the submatrix of W formed by considering the rows corresponding to founders and W_{NF} is the submatrix of W comprised of the rows corresponding to non-founders. Let f be the total number of founders. Under the assumption that these individuals are unrelated, the pmf of their genotypes given allelic frequencies is:

$$\pi(W_F|P^*) = \prod_{i=1}^{f}\pi(W_i|P^*) = \prod_{j=1}^{m}\prod_{i=1}^{f}\pi\left(w_{ij}|P^*\right) = \prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}\prod_{i=1}^{f_l}\pi\left(w_{ij}^l|p_{lj}^*\right)$$

$$= \prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}\prod_{i=1}^{f_l}\left(p_{lj}^{*2}\right)^{I_{1i}}\left(2p_{lj}^*(1-p_{lj}^*)\right)^{I_{0i}}\left((1-p_{lj}^*)^2\right)^{I_{-1i}}$$

$$= \prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}\left(p_{lj}^{*2}\right)^{n_l^{BB_j}}\left(2p_{lj}^*(1-p_{lj}^*)\right)^{n_l^{AB_j}}\left((1-p_{lj}^*)^2\right)^{n_l^{AA_j}}$$

$$= \prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}2^{n_l^{AB_j}}p_{lj}^{*\,2n_l^{BB_j}+n_l^{AB_j}}\left(1-p_{lj}^*\right)^{2n_l^{AA_j}+n_l^{AB_j}} = 2^{n^H}\prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}p_{lj}^{*\,n_l^{B_j}}\left(1-p_{lj}^*\right)^{n_l^{A_j}},$$

replacing $p_{lj}^* = p_{lj}/r_{lj}\ \forall\ l = 1,2..,\mathcal{S}, \forall\, j = 1,2,\dots,m$:

$$\pi(W_F|P,\boldsymbol{r}) = 2^{n^H}\prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}\frac{1}{r_{lj}^{2f_l}}p_{lj}^{n_l^{B_j}}\left(r_{lj}-p_{lj}\right)^{n_l^{A_j}}$$

where f_l is the number of founders in the l^{th} subpopulation; thus, $f = \sum_{l=1}^{\mathcal{S}}f_l$, $n_l^{BB_j}, n_l^{AB_j}$ and $n_l^{AA_j}$ are the counts of founders with genotypes BB, AB and AA at marker locus j in subpopulation l respectively, $n_l^{B_j} = 2n_l^{BB_j} + n_l^{AB_j}$ is the total count of B alleles at marker locus j in founders from subpopulation l, $n_l^{A_j} = 2n_l^{AA_j} + n_l^{AB_j}$ is the total count of A alleles at marker locus j in founders from subpopulation l and $n^H = \sum_{j=1}^{m}\sum_{l=1}^{\mathcal{S}}n_l^{AB_j}$ is the total number of heterozygous loci in the base population. In terms of the random variables w_{ij}^l, $n_l^{BB_j}, n_l^{AB_j}$ and $n_l^{AA_j}$ can be written as: $n_l^{BB_j} = \sum_{i=1}^{f_l}I_{1i}\,, n_l^{AA_j} = \sum_{i=1}^{f_l}I_{-1i}\,, n_l^{AB_j} = f_l - \left(n_l^{BB_j} + n_l^{AA_j}\right) = f_l - \sum_{i=1}^{f_l}\left(w_{ij}^l\right)^2$.

For non-founders:

$$\pi(W_{NF}|W_F) = \prod_{j=1}^{m}\prod_{i'=f+1}^{n}\pi\left(w_{i'j}|w_{S_{i'}j}, w_{D_{i'}j}\right) = \prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}\prod_{i'=f_l+1}^{n_l}\pi\left(w_{i'j}^l|w_{S_{i'}j}^l, w_{D_{i'}j}^l\right)$$

where $w_{S_{i'}j}^l$ and $w_{D_{i'}j}^l$ are the genotypes for marker j of the parents of individual i' from subpopulation l. Hence:

$$\pi(W|P^*) = \prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}\prod_{i=1}^{f_l}\pi\left(w_{ij}^{l}|p_{lj}^{*}\right) \times \prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}\prod_{i'=f_l+1}^{n_l}\pi\left(w_{i'j}|w_{S_{i'j}}, w_{D_{i'j}}\right)$$

$$= \prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}\prod_{i=1}^{f_l}\left\{\pi\left(w_{ij}^{l}|p_{lj}^{*}\right)\prod_{i'=f_l+1}^{n_l}\pi\left(w_{i'j}^{l}|w_{S_{i'j}}, w_{D_{i'j}}\right)\right\}$$

$$= 2^{n^H}\prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}\left\{p_{lj}^{*\,n_l^{B_j}}\left(1-p_{lj}^{*}\right)^{n_l^{A_j}}\prod_{i'=f_l+1}^{n_l}\pi\left(w_{i'j}^{l}|w_{S_{i'j}}, w_{D_{i'j}}\right)\right\}$$

$$\Rightarrow \pi(W|P,\boldsymbol{r}) = 2^{n^H}\prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}\left\{\frac{1}{r_{lj}^{2f_l}}p_{lj}^{\,n_l^{B_j}}\left(r_{lj}-p_{lj}\right)^{n_l^{A_j}}\prod_{i'=f_l+1}^{n_l}\pi\left(w_{i'j}^{l}|w_{S_{i'j}}, w_{D_{i'j}}\right)\right\}.$$

Remark 2 Under the assumptions presented at the beginning of this section, given base genotypes, the process defining the inheritance of alleles is completely determined by the pedigree information. The pedigree allows tracing the set of possible values that genotypes can take from a given individual back to the base population. It implies that allelic frequencies have to be known only in the base population because the distribution of genotypes in the set of non-founders is completely determined by the pedigree. Stated another way, given the pedigree, only the founder genotypes carry information about allelic frequencies.

The next step is to formally define the support (set of values of W with non-null probability) of the pmf $\pi(W|P^*)$ and its cardinality (i.e., the number of elements contained in this set). If we had a population of n unrelated individuals genotyped for m biallelic loci, then the total number of possible values of W would be 3^{nm}. However, given the kinship between individuals, the number of possible values of W is smaller than 3^{nm}. Let $\mathcal{G}$ be the support of $\pi(W|P^*)$, then number of possible values that W can take is $|\mathcal{G}|$, namely the cardinality of the set $\mathcal{G}$. To find $|\mathcal{G}|$, the pedigree of the population is used because along with the genotypes of founders, it defines how many individuals could potentially have one, two or three genotypes for each marker locus. For example, a progeny from parents with genotypes AA and AA has genotype AA with probability one, while a progeny from parents AA and AB could have genotypes AA or AB with probabilities equal to ½. Let $\mathcal{F}$ be the set of founders, then $|\mathcal{F}| = f$, thus there are 3^{fm} possible values for the submatrix of W corresponding to founders under the assumption that they are unrelated. Hereinafter, each one of these possible values is defined as a "base genotypic configuration". Notice that each one of these fm genotypic configurations induces a different set of possible genotypes in the rest of the population. Under base genotypic configuration $k, 1 \leq k \leq 3^{mf}$, for each marker locus the remaining $n - f$ individuals are grouped into three mutually exclusive sets: $O_{1j}^{k} := \left\{i\colon \left|\left\{S_{ij} \times D_{ij}\right\}^{k}\right| = 1,1 \leq j \leq m, 1 \leq k \leq 3^{mf}\right\}$, $O_{2j}^{k} := \left\{i\colon \left|\left\{S_{ij} \times D_{ij}\right\}^{k}\right| = 2,1 \leq j \leq m, 1 \leq k \leq 3^{mf}\right\}$, $O_{3j}^{k} := \left\{i\colon \left|\left\{S_{ij} \times D_{ij}\right\}^{k}\right| = 3, 1 \leq j \leq m, 1 \leq k \leq 3^{mf}\right\}$, where $\left|\left\{S_{ij} \times D_{ij}\right\}^{k}\right|$ is the cardinality of the set of possible genotypes at marker locus j resulting from the mating of the parents of individual i under base genotypic configuration $k, \left\{S_{ij} \times D_{ij}\right\}^{k}$. Consequently, $\left|O_{lj}^{k}\right|$ is the number of individuals in the population for which there are l possible genotypes at marker $j, 1 \leq l \leq 3$ given the k^{th} base genotypic configuration. Hence, at each marker locus and each base genotypic configuration the following equality is satisfied: $\left|O_{1j}^{k}\right| + \left|O_{2j}^{k}\right| + \left|O_{3j}^{k}\right| = n - f$. Therefore, at each marker locus and base genotypic configuration the total number of possible genotypes in the $n - f$ non-founder individuals is $1^{\left|O_{1j}^{k}\right|}2^{\left|O_{2j}^{k}\right|}3^{\left|O_{3j}^{k}\right|}$, and under the linkage equilibrium assumption, the total number of possible genotypes across marker loci given base genotypic configuration k is

$$\prod_{j=1}^{m} 1^{\left|O_{1j}^{k}\right|}2^{\left|O_{2j}^{k}\right|}3^{\left|O_{3j}^{k}\right|} = 2^{\sum_{j=1}^{m}\left|O_{2j}^{k}\right|}3^{\sum_{j=1}^{m}\left|O_{3j}^{k}\right|}$$

Accordingly, given the pedigree of the population, the total number of possible values that matrix W can take is obtained by summing the above expression over k: $|\mathcal{G}| = \sum_{k=1}^{3^{mf}} 2^{\sum_{j=1}^{m}\left|O_{2j}^{k}\right|}3^{\sum_{j=1}^{m}\left|O_{3j}^{k}\right|}$. As a check of the adequacy of this expression, notice that ignoring pedigree and assuming that all individuals in the population are unrelated is equivalent to treat them all as founders which implies that $f = n$, consequently $\left|O_{1j}^{k}\right| = \left|O_{2j}^{k}\right| = \left|O_{3j}^{k}\right| = 0, \forall\, j = 1,2,\dots,m, \forall\, k = 1,2,\dots,3^{mn}$, thus $|\mathcal{G}| = \sum_{k=1}^{3^{mn}} 2^0 3^0 = 3^{mn}$. Before defining the support of W, the following sets are defined. The k^{th} base genotypic configuration is defined as follows: $\mathcal{G}_{\mathcal{F}}^{k} \coloneqq \left\{w_{ijk} \colon i \in \mathcal{F}, 1 \leq j \leq m, 1 \leq k \leq 3^{mf}\right\}$. For each set $\mathcal{G}_{\mathcal{F}}^{k}$, that is, for each genotypic configuration, $1 \leq k \leq 3^{mf}$, define: $\mathcal{G}_{O_1}^{k} \coloneqq \left\{w_{ij} \colon i \in O_{1j}^{k}, 1 \leq j \leq m\right\}$, $\mathcal{G}_{O_2}^{k} \coloneqq \left\{w_{ij} \colon i \in O_{2j}^{k}, 1 \leq j \leq m\right\}$, $\mathcal{G}_{O_3}^{k} \coloneqq \left\{w_{ij} \colon i \in O_{3j}^{k}, 1 \leq j \leq m\right\}$. As mentioned before, each set $\mathcal{G}_{\mathcal{F}}^{k}$ induces a set $\mathcal{G}_{O_1}^{k} \cup \mathcal{G}_{O_2}^{k} \cup \mathcal{G}_{O_3}^{k}$, thus: $\mathcal{G} = \bigcup_{k=1}^{3^{mf}}\left\{\mathcal{G}_{\mathcal{F}}^{k} \cup \mathcal{G}_{O_1}^{k} \cup \mathcal{G}_{O_2}^{k} \cup \mathcal{G}_{O_3}^{k}\right\}$.

Remark 3 When some individuals are not genotyped or partially genotyped, that is, when a fraction of matrix W is not observed, $\pi(W|P^*) = f(W^{\mathcal{O}}|W^N, P^*)\pi(W^N|P^*)$ where $\pi(W^N|P^*) = \sum_{\mathcal{G}^{\mathcal{O}}} \pi(W|P^*)$, $\mathcal{G}^{\mathcal{O}}$ is the set of possible values of $W^{\mathcal{O}}$. However, as will become clear in section 2.2, explicit computation of $\pi(W^N|P^*)$ is not required. In this case, some of the elements of $\pi(W|P^*)$ can be conceptually partitioned as follows: $n_l^{B_j} = n_{l_{\mathcal{O}}}^{B_j} + n_{l_N}^{B_j}, n_l^{A_j} = n_{l_{\mathcal{O}}}^{A_j} + n_{l_N}^{A_j}, n^H = n_{\mathcal{O}}^H + n_N^H$ where subindex $l_{\mathcal{O}}$ indicates that the corresponding count comes from genotyped individuals in the l^{th} subpopulation and subindex l_N indicates that the corresponding count comes from non-genotyped individuals.

2.2 Full conditionals, homoscedastic residuals, homogeneous and heterogeneous marker effect covariance matrix models

Henceforth, it is assumed that vector $\boldsymbol{g}$ and columns of matrix W are ordered by marker unless otherwise indicated. The full conditionals are denoted as $\pi(\cdot\,|Else)$. Firstly, $\boldsymbol{g}|Else \sim MVN\left(\left(I_m \otimes (G^0)^{-1} + \frac{W'W}{\sigma^2}\right)^{-1} \frac{1}{\sigma^2} W'\boldsymbol{y}, \left(I_m \otimes (G^0)^{-1} + \frac{W'W}{\sigma^2}\right)^{-1}\right)$. If W_k denotes the submatrix of W corresponding to marker k, W_k is of dimension $n \times \mathcal{S}$ and has the form $W_k = (\boldsymbol{w}'_{1k} \quad \cdots \quad \boldsymbol{w}'_{nk})', \boldsymbol{w}_{ik} = (0 \quad \cdots \quad w_{ik} \quad \cdots \quad 0)_{1\times\mathcal{S}}, i = 1,2,\dots,n$, the only non-null entry of vector $\boldsymbol{w}_{ik}$ is the random variable corresponding to the genotype of the i^{th} individual for the k^{th} marker w_{ik} and it is located at position $l, l = 1,2,\dots,\mathcal{S}$, where l is the subpopulation to which individual i

pertains. Other full conditionals are $G^0|Else\sim IW\left(a+m, \boldsymbol{\Sigma}+\sum_{j=1}^{m}\boldsymbol{g}_j\boldsymbol{g}_j'\right)$, $\sigma^2|Else\sim IG\left(\frac{v+n}{2}, \frac{(\boldsymbol{y}-W\boldsymbol{g})'(\boldsymbol{y}-W\boldsymbol{g})+\tau^2}{2}\right)$. To arrive at $\pi(W^N|Else)$ the following definitions have to be made. The rows of W for individuals with missing genotypes are partitioned as $W^{M_C}, W^{M_1}, \ldots., W^{M_K}$ which respectively represent the rows of W for non-genotyped individuals, and individuals partially genotyped having missing genotypes for loci subsets $M_1N, \ldots, M_KN$. Accordingly, the subvector of the data vector corresponding to records from non-genotyped or partially genotyped individuals can be partitioned as $\boldsymbol{y}^N = (\boldsymbol{y}^{M_C}{}', \boldsymbol{y}^{M_1}{}', \ldots., \boldsymbol{y}^{M_K}{}')'$. The rows of W corresponding to partially genotyped individuals are partitioned as follows: $W^{M_k} = (W^{M_k\mathcal{O}} \vdots W^{M_kN})$, where superindex $M_k\mathcal{O}$ denotes the set of loci with observed genotypes, while superindex M_kN denotes the set of marker loci with missing genotypes. Similarly, when doing computations among these submatrices and $\boldsymbol{g}$, this vector can be arranged as $(\boldsymbol{g}^{M_k\mathcal{O}}{}' \vdots \boldsymbol{g}^{M_kN}{}')'$, then:

$$\pi(W^N|Else) = \pi(W^N|\boldsymbol{y}^N, W^{\mathcal{O}}, \boldsymbol{g}, \sigma^2, P^*)$$

$$\propto \pi^+(W|P^*)\exp\left(\frac{-1}{2\sigma^2}(-2\boldsymbol{g}'W^{N\prime}\boldsymbol{y}^N + \boldsymbol{g}'W^{N\prime}W^N\boldsymbol{g})\right)$$

$$\times\prod_{k=1}^{K}\exp\left(\frac{-1}{2\sigma^2}h(W^{M_k}, \boldsymbol{g}^{M_k}, \boldsymbol{y}^{M_k})\right)$$

where

$$h(W^{M_k}, \boldsymbol{g}^{M_k}, \boldsymbol{y}^{M_k}) = 2(\boldsymbol{g}^{M_kN\prime}W^{M_kN\prime}W^{M_k\mathcal{O}}\boldsymbol{g}^{M_k\mathcal{O}} - \boldsymbol{g}^{M_kN\prime}W^{M_kN\prime}\boldsymbol{y}^{M_k}) + \boldsymbol{g}^{M_kN\prime}W^{M_kN\prime}W^{M_kN}\boldsymbol{g}^{M_kN},$$

$\pi^+(W|P^*) = f^+(W^{\mathcal{O}}|W^N, P^*)\pi(W^N|P^*)$ and $f^+(W^{\mathcal{O}}|W^N, P^*)$ is the part of the W component of the likelihood depending on W^N. Notice that this is a non-standard pmf and that when $W^{\mathcal{O}}$ depends only on W^N the form of $\pi(W^N|Else)$ remains the same because $f^+(W^{\mathcal{O}}|W^N)\pi(W^N|P^*) = \pi^+(W|P^*)$. When $\boldsymbol{r}$ is known

$$\pi(P|Else) = \pi(P|W^{\mathcal{O}}, W^N, \boldsymbol{r}) = \pi(P|W, \boldsymbol{r})$$

$$\propto \prod_{j=1}^{m} p_{(\mathcal{S}+1)j}^{\alpha_{\mathcal{S}+1}-1} \prod_{l=1}^{\mathcal{S}} \left\{ {p_{lj}}^{{n_l^{B}}_j+\alpha_l-1} \left(r_l - p_{lj}\right)^{{n_l^{A}}_j} \right\}$$

which is the product of m non-standard pdf. Recall that when $\boldsymbol{r}$ is unknown, there is a slight difference in this expression as was shown in section 2.1.

Remark 4 In the absence of missing genotypes, that is, $W^{\mathcal{O}} = W$, the previous expression is not the full conditional density of P, but its posterior density.

For the heterogeneous marker effect covariance matrix model G is a block-diagonal matrix comprised by m blocks of dimension $\mathcal{S} \times \mathcal{S}$ as described in section 2.1. Under this model $\pi(G) = \prod_{l=1}^{\mathcal{S}} \pi\left(G_j\right)$. This prior pdf is the only difference with the previous model; therefore, the joint posterior is very similar (see Appendix A). Hence, all full conditionals are the same except for $\boldsymbol{g}|Else \sim MVN\left(\left(G^{-1} + \frac{W'W}{\sigma^2}\right)^{-1} \frac{1}{\sigma^2} W'\boldsymbol{y}, \left(G^{-1} + \frac{W'W}{\sigma^2}\right)^{-1}\right), G^{-1} = Block\ diag.\left(G_j^{-1}\right), j = 1,2 \ldots, m$ and $G_j|Else \overset{ind}{\sim} IW\left(a+1, \boldsymbol{\Sigma} + \boldsymbol{g}_j \boldsymbol{g}_j'\right)$. The full conditionals for models with heteroscedastic residuals are presented in Appendix A along with joint posteriors.

2.3 Model comparison via Deviance Information Criterion

The term null model refers to simplified versions of the proposed models. These null models ignore the factor splitting the complete population into subpopulations; therefore, each marker has a single overall effect and allelic frequencies are assumed to be the same across subpopulations.

Null models are as follows: $\boldsymbol{y} = W_0 \boldsymbol{g}_0 + \boldsymbol{\varepsilon}$, where $\boldsymbol{y}$ is the same as before, $\boldsymbol{g}_0$ is an $m \times 1$ unobservable random vector containing allele substitution effects of each marker, $(W_0)_{n \times m}$ is the random observable design matrix which is of the form $(W_1' \vdots \cdots \vdots W_{\mathcal{S}}')'$ when ordering data by subpopulation, and $\boldsymbol{\varepsilon}$ is a random vector of residuals. The priors for $\boldsymbol{g}_0$ are simply univariate versions of the priors used for $\boldsymbol{g}$. Thus, $\boldsymbol{g}_0 | G^D \sim \pi(\cdot\,|G^D), G^D = Diag\left(\sigma_{g1}^2, \ldots, \sigma_{gm}^2\right)$, $\sigma_{gj}^2 \overset{iid}{\sim} IG\left(\frac{a}{2}, \frac{b}{2}\right)$, (for the homogeneous marker effect variance model $\sigma_{g1}^2 = \cdots = \sigma_{gm}^2 = \sigma_g^2$) and the residual variance σ^2 is given an $IG\left(\frac{\tau^2}{2}, \frac{v}{2}\right)$ prior as before. In addition, $\boldsymbol{p} = (p_1, p_2, \ldots, p_m)$ is a vector of overall reference allele frequencies, $W_0|\boldsymbol{p} \sim \pi(W_0|\boldsymbol{p})$ is a simplified version of $\pi(W|P^*)$ (shown later), and the prior for $\boldsymbol{p}$ is $p_j \overset{iid}{\sim} Beta(\alpha, \beta), j = 1,2, \ldots, m$.

The Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) combines a measure of goodness of fit based on the posterior distribution and a penalty for model complexity, and despite some criticism it has been used in different areas to perform model comparison (Gelman et al., 2014; Spiegelhalter et al., 2014). It has the following form:

$$DIC = -2 \log f\left(\boldsymbol{Data}\middle|\widehat{\boldsymbol{\theta}}_B\right) + 2p_{DIC}$$

where $p_{DIC} = 2\left(\log f\left(\boldsymbol{Data}\middle|\widehat{\boldsymbol{\theta}}_B\right) - E_{\boldsymbol{\theta}|\boldsymbol{Data}}[\log f(\boldsymbol{Data}|\boldsymbol{\theta})]\right), \widehat{\boldsymbol{\theta}}_B = E[\boldsymbol{\theta}|\boldsymbol{y}]$ is the posterior mean of the unknown parameters. The first component of DIC is a measure of model adequacy, whereas the second one is the effective number of parameters which is a penalty for increasing model complexity (Spiegelhalter et al., 2002). Models with a smaller DIC are preferred. Recall that for any of our models the likelihood has two components: $f(\boldsymbol{y}, W^{\mathcal{o}}|W^N, \boldsymbol{g}, R, P^*) = f(\boldsymbol{y}|W, \boldsymbol{g}, R) f(W^{\mathcal{o}}|W^N, P^*)$ that were denoted as the $\boldsymbol{y}$ component and the W component. Thus, the general form of the DIC is:

$$DIC = -2 \log f\left(\boldsymbol{y}\middle|W^{\mathcal{o}}, \widehat{W}_B^N, \widehat{\boldsymbol{g}}_B, \widehat{R}_B\right) + 2p_{DIC-\boldsymbol{y}} - 2 \log f\left(W^{\mathcal{o}}\middle|\widehat{W}_B^N, \widehat{P}_B^*\right) + 2p_{DIC-W}$$

$$\coloneqq DIC_{\boldsymbol{y}} + DIC_W$$

where $p_{DIC-\boldsymbol{y}} = 2\left(\log f\left(\boldsymbol{y}\middle|W^{\mathcal{o}}, \widehat{W}_B^N, \widehat{\boldsymbol{g}}_B, \widehat{R}_B\right) - E_{W^N, \boldsymbol{g}, R, P^*|\boldsymbol{y}, W^{\mathcal{o}}}[\log f(\boldsymbol{y}|W, \boldsymbol{g}, R)]\right)$ and $p_{DIC-W} = 2\left(f\left(W^{\mathcal{o}}\middle|\widehat{W}_B^N, \widehat{P}_B^*\right) - E_{W^N, P^*|\boldsymbol{y}, W^{\mathcal{o}}}[f(W^{\mathcal{o}}|W^N, P^*)]\right)$. Thus, as the likelihood, the DIC can be decomposed into a $\boldsymbol{y}$ component $DIC_{\boldsymbol{y}}$ and a W component DIC_W.

2.4 Parameter inference via MCMC

In this section, some issues about MCMC algorithms to carry out inference are briefly discussed. Notice that when W is fully observed, the fact that there are no missing genotypes implies that posterior sampling for the (hyper) parameters of the W component of the likelihood and the (hyper) parameters of the $\boldsymbol{y}$ component can be performed separately. The full conditionals of $\boldsymbol{g}, G, \sigma^2, g_0$, and σ_g^2 are known; therefore, samples from the joint posterior can be obtained using a Gibbs sampler (Casella and George, 1992) while samples from the posterior distribution of allelic frequencies can be obtained using a Metropolis-Hastings algorithm. Specifically, independent Metropolis algorithms are considered here. For the scenario of $\boldsymbol{r}$ known, the new samples can be generated in two steps: firstly a Dirichlet vector is sampled, and secondly its elements are scaled with the appropriate elements of $\boldsymbol{r}$. Alternatively, uniform$(0, r_l)$ distributions can be used as proposal, which simplifies computations. With such proposal, given the current state of the chain denoted as P^t, the acceptance probability of the new sample P_+^t is $\min\left\{\frac{\pi(P_+^t|W)}{\pi(P^t|W)}, 1\right\}$. For null models, the posterior distribution of $\boldsymbol{p}_0$ is the product of m Beta$(p_j; n^{B_j} + \alpha, n^{A_j} + \beta)$ distributions, $j = 1,2 \dots, m$. Hence, direct sampling can be implemented if needed and the functional form of the posterior mean is known. When $\boldsymbol{r}$ is unknown, the candidate to sample from the posterior of $(\boldsymbol{p}_j, \boldsymbol{q}_j), j = 1,2, \dots, m,$ could be a Dirichlet distribution.

On the other hand, when matrix W is partially observed a Metropolis-within-Gibbs strategy (Robert and Casella, 2010) can be used to sample from the joint posterior. This strategy is useful due to the fact that nor $\pi(W^N|Else)$ neither $\pi(P^*|Else)$ are standard distributions and the existence of the parameter W^N does not allow to carry out separate sampling algorithms as before because this is a parameter of both components of the likelihood. Accordingly, there are two Metropolis steps in the algorithm to sample from the posterior of the full models. The first one is used to obtain samples from $\pi(W^N|Else)$. A good proposal is $\pi(W^N|W^\sigma, P^*)$ because obtaining direct samples from this distribution via the inverse transform method for discrete random variables (Robert and Casella,

2010) is straightforward. The functional form of $\pi(W^{N}|W^{\mathscr{o}},P^{*})$ is derived from first principles as explained in 2.3.1. Thus, given the current state of the chain W^{N_t}, the acceptance probability of a new sample $W_{+}^{N_t}$ is: $\min\left\{\frac{\pi(W_{+}^{Nt}|Else)\pi(W_{+}^{Nt}|W^{\mathscr{o}},P^{*})}{\pi(W^{N_t}|Else)\pi(W^{N_t}|W^{\mathscr{o}},P^{*})},1\right\}$. This applies to both situations: $\boldsymbol{r}$ known and $\boldsymbol{r}$ unknown. The second Metropolis step is used to draw samples from $\pi(P|Else)$ for $\boldsymbol{r}$ known or $\pi(P,Q|Else)$ for $\boldsymbol{r}$ unknown. The proposals mentioned for the non-missing genotypes scenario also work here. For the null models, it turns out that $\forall\, j = 1,2,\dots,m, \pi\left(p_j\middle|Else\right)$ is a known distribution, it is a Beta($n^{B_j}+\alpha, n^{A_j}+\beta$) and consequently only one Metropolis step is needed because direct sampling from the full conditional distribution of $\boldsymbol{p}_0$ is feasible. Notice that this full conditional distribution is the posterior distribution of $\boldsymbol{p}_0$ when matrix W is completely observed.

2.5 Simulation study

In order to provide an example of the implementation of some of the proposed models and the computation of some criteria to compare their performance, two simulated datasets were used. Simulation of these datasets involved two main steps: Simulation of genotypes (QTL and SNP), and simulation of QTL effects and noise. The phenotypes were simulated as the sum of additive genetic effects (sum of QTL allele content times the allele effect) and noise. Datasets were simulated using the software QMSim (Sargolzaei and Schenkel, 2013). In both cases, a historical population was simulated by creating 1000 generations of random mating using a forward-in-time approach in order to reach mutation-drift equilibrium and to create linkage disequilibrium (Sargolzaei and Schenkel, 2013). The historical population size in each generation was 1000 with 500 males and 500 females. Then, subpopulations were created from individuals pertaining to the historical population under different selection pressures and criteria, and different mating systems (Table 1).

Phenotypes were simulated with different number of QTL controlling the trait and different heritabilities. Furthermore, the population structure also differed because the criteria to simulate the subpopulations were different for each trait. Briefly, dataset 1 involved three subpopulations with different number of generations, migration was allowed and the heritability of the trait was high. Dataset 2 comprised two subpopulations with only two generations, no migration and the heritability of the trait was low (Table 1). For further details concerning the simulation see appendix B.

Given that this paper is focused on proposing and explaining a set of across population genome-wide prediction models and not with their large scale implementation, the number of simulated SNP and sample size were low in order to avoid computational issues (Table 1). Phenotype 1 illustrates the situation in which the number of markers is equal to the number of QTL affecting the trait, while for phenotype 2 the number of markers is larger than the number of QTL controlling the trait. These contrasting simulation schemes, different selection pressures and criteria, mating designs and number of generations were used to mimic real life situations where different subpopulations have different backgrounds. These simulated datasets were used to carry out analyses using the following models: Homogeneous and heterogeneous marker effect covariance matrices with homoscedastic residuals and their null versions. Only models with homoscedastic residuals were used to analyze these datasets because simulations did not consider heteroscedastic residuals.

The analyses performed involved implementation of MCMC algorithms explained in section 2.4, the computation of DIC and the computation of the following quantities measuring predictive performance and accuracy: the squared correlation between predicted breeding values and phenotypes in the testing populations, hereinafter called predictive ability, and squared correlations between true and predicted breeding values computed in the testing populations (accuracy). Because true breeding values were available for the complete populations, squared correlations between true and predicted breeding values in the training populations were also computed.

For dataset 1, the training population was comprised of generations 0 to 2 of subpopulation 1, 0 to 5 from subpopulation 2 and generation 0 of subpopulation 3, while the testing population included generation 3 of subpopulation one, generation 6 of subpopulation 2 and generation 1 of subpopulation 3. For dataset 2, the training population was composed of generations 0 and 1 of subpopulations 1 and 2 and the testing dataset contained generation 2 of subpopulations 1 and 2. In dataset 2, the full genotypes of three individuals (one founder from each subpopulation and a non-founder from subpopulation 1) were not included in the analysis in order to simulate the case of missing genotypes.

It was assumed that $\boldsymbol{r} = \left(\frac{1}{\mathcal{S}}, \ldots, \frac{1}{\mathcal{S}}\right)$. In an initial analysis, a scaled Dirichlet distribution was used as proposal to draw samples from $\pi(P|Else)$, but the behavior of the chains was not satisfactory because the acceptance rate was too low (results not shown). Consequently the product of $\mathcal{S}$ independent uniform $\left(0, \frac{1}{\mathcal{S}}\right)$ distributions was used as proposal. For each dataset, 20.000 iterations were run; the first 10.000 were considered burn-ins. An in-house R script (R Core Team, 2015) was created to carry out the analyses which were performed using the University of Florida's high performance computing cluster.

3. Results

3.1 Simulated populations

Tables 1 and 2 show features corresponding to characteristics of the simulated genomes and populations.

Table 1 Parameters and selection criteria to simulate phenotypes

Parameter	Phenotype 1	Phenotype 2
Heritabilities	0.70, 0.62, 0.54	0.20, 0.15
Phenotypic variances	100, 79, 65	100, 94
Number of QTL	600	40
Number of SNP	600	200
Number of Chromosomes	10	2
Base population structure[1]	1: 28M, 180F, Phen/L 2: 20M, 90F, Phen/H 3: 50M, 500F, Rnd	1: 5M, 25F, Rnd 2: 20M, 50F, Phen/H

Number of generations, mating system and selection criteria^2	1:3,0.8,0.4, As1/Phen, Phen/L 2: 6, 0.7, 0.1, As2/Phen, Phen/H 3:3, 0.7, 0.2, Rnd, Rnd	1: 2, 1, 0.9, Rnd, Rnd 2: 2, 0.9, 0.3, Rnd, Phen/H

[1] For each line, the first number indicates the subpopulation, items separated by a comma respectively show: number of males, number of females, criterion used to select them (Phen = phenotype, Rnd = random, L = lowest values, H = highest values).

[2] For each line, the first number indicates the subpopulation, items separated by a comma respectively show: Number of generations, proportion of selected females per generation, proportion of selected males per generation, mating design (Rnd = random, As1 = assortative by similarity, As2 = assortative by dissimilarity, Phen = phenotype), and selection criterion (same abbreviations as in numeral 1).

Table 2 Summary of some characteristics of the simulated populations

Feature	Dataset 1	Dataset 2
Population size (males, females, total)	883, 1565, 2448	67, 103, 170
Average inbreeding per subpopulation	S1:0.0182, S2: 0.0310, S3:0.0	S1: 0.0 , S2:0.0
Average homozygosity per subpopulation	S1: 0.6240, S2: 0.6359, S3:0.6190	S1:0.6392, S2:0.6283
Phenotype sample mean and SD (in brackets) per subpopulation	S1: -19.78 (13.21) S2: 25.71 (9.60) S3: 0.26 (9.91)	S1:-0.5959 (9.3616) S2:8.9253 (11.9571)

In both datasets, none of the markers had a minor allele frequency lower than 0.05. Thus, all the simulated marker loci were considered in the analyses.

3.2 DIC, predictive ability and accuracies of predicted breeding values

For dataset 1, the DIC computed using the “W-component” of the likelihood for the full models was 4717671 and 6589105 for the null models. Thus, it provided evidence in favor of the full models when estimating allelic frequencies in the base population. Table 4 shows DIC values for dataset 1, Table 5 DIC values for dataset 2 and Table 6 shows predictive abilities and accuracies in both datasets. For Tables 4 to 6, the following is the meaning of abbreviations for the different models fitted to datasets 1 and 2: M_{1G}= full model with Multivariate Gaussian prior and homogeneous marker effect covariance matrices, M_{1G}^{*}= full model with Multivariate Gaussian prior and heterogeneous marker effect covariance matrices. Recall that all models assumed homoscedastic residuals. The remaining models with subindex 1 replaced by 0 correspond to null versions of the corresponding full models.

Table 4 $\boldsymbol{y}$ component and total DIC for dataset 1

Model	$\boldsymbol{y}$ **component of DIC**	**Total DIC**
M_{1G}	33702.55	4751373.55
M_{1G}^{*}	11599.05	4729270.05
M_{0G}	15396.32	6604501.32
M_{0G}^{*}	13008.42	6602113.42

Thus, in dataset 1, according to the $\boldsymbol{y}$ component of DIC, for the models with homogeneous marker effect covariance matrices (variances) the null model performed better, while for models with heterogeneous covariance matrices (variances) according to this criterion the full model should be preferred over its null version. When considering the whole likelihood to compute the DIC, the two full models had smaller DIC. Additionally, the model with the smallest DIC, and therefore the "best" one under this criterion was model M_{1G}^{*}.

Table 5 $\boldsymbol{y}$ component, W component and total DIC for dataset 2

Model	$\boldsymbol{y}$ **component of DIC**	W **component of DIC**	**Total DIC**
M_{1G}	1314.0	38367.4	39681.4
M_{1G}^{*}	1328.8	38356.4	39684.2
M_{0G}	1365.6	38180.3	39545.9
M_{0G}^{*}	1370.1	38179.0	39549.1

In this dataset the two components of the DIC values and therefore DIC values were similar for all models. The $\boldsymbol{y}$ components of DIC were smaller for the full models. Conversely, the W components were smaller for null models as well as total DIC values.

Table 6 Predictive abilities and accuracies in datasets 1 and 2

Model	**Predictive Ability**		**Accuracy in testing population**		**Accuracy in Training population**	
	Dataset1	Dataset 2	Dataset1	Dataset2	Dataset1	Dataset2

M_{1G}	0.29	0.019	0.27	0.04	0.32	0.17
M_{1G}^{*}	0.76	0.016	0.83	0.03	0.94	0.21
M_{0G}	0.53	0.004	0.50	0.07	0.55	0.24
M_{0G}^{*}	0.83	0.013	0.88	0.05	0.88	0.23

In dataset 1, according to predictive abilities, the model with the best performance was model M_{0G}^{*} while model M_{1G} had the worst performance. The squared Pearson correlations between true and predicted breeding values in testing dataset 1 suggested that the performance of these models followed a trend similar to that indicated by predictive abilities. In training dataset 1, model M_{1G}^{*} yielded the highest accuracy and model M_{1G} had the smallest accuracy.

Predictive abilities and accuracies in the testing sets were extremely low for dataset 2. Accuracies in training set were higher than those obtained in the testing set; however, they were still low. There were not substantial differences between these squared correlations. Predictive abilities were higher for the full models, while accuracies in testing and training sets were higher for the null models.

4. Discussion

4.1 General features of the models

A group of hierarchical Bayesian linear regression models to carry out simultaneous genome-wide prediction in several subpopulations accounting for randomness of genotypes was presented. The proposed models differed in the prior distribution assigned to the marker effects and on the assumptions made about residual variances (homogeneous or heterogeneous across subpopulations). The priors for the marker effects were multivariate (univariate) Gaussian and allowed homogeneous or heterogeneous covariance matrices (or variances).

The differences between these models and other regression models currently used in across population genome-wide prediction are: 1) subpopulation-specific effects for each marker are considered and their covariance matrices are modeled explicitly, and 2) genotypes are treated as random variables with a distribution that depends on allelic frequencies as well as on pedigree information. The second feature makes these models different from all other genome-wide prediction models. The distribution of genotypes combines pedigree and genomic information that are not used when randomness of W is ignored. It allows accounting for heterogeneity and correlations of allelic frequencies of the same marker across subpopulations and including individuals with phenotypes and missing genotypes in various loci without carrying out a previous imputation. This is possible because the non-observed part of W, denoted as W^N, is treated as a parameter and therefore imputation is automatically performed. Another advantage is that the use of a Bayesian approach automatically takes into account uncertainty about the imputed genotypes.

Although most of the paper has been devoted to the models allowing subpopulation-specific effects for each marker (the full models), their univariate versions (the null models) are also contributions of this study. These also allow including individuals with missing genotypes in some or all marker loci without need of external imputation and take into account randomness in genotypes. Therefore, these models could also be used either in single population analyses or to conduct across population genome-wide prediction pooling the data as has been done in previous studies (de Roos et al., 2009; Lund et al., 2011; van den Berg et al., 2015; Wientjes et al., 2015) and was also done here.

Doing a joint analysis has the advantage that the number of phenotypes increases, but in our full models the number of location parameters is also incremented because each marker is allowed to have subpopulation-specific effects; moreover, the number of covariance parameters also increases. The gain in accuracy is achieved when factors such as different QTL effects across subpopulations, differences in linkage phase between QTL and markers, and differences in allelic frequencies and LD patterns make marker effects change substantially from one subpopulation to another. Consequently, the performance of these models may have considerable variation from one dataset to another.

The diagonal blocks of G were assumed to be non-structured. A way reduce dimensionality of the parameter space is to assume certain structure of G. For example, it can be assumed that all covariances and variances are the same, thus, only two parameters per block have to be estimated.

The conditional independence property used to derive $\pi(W|P^*)$ implies that allelic frequencies are estimated in the set of oldest individuals with phenotypes. Here, this set of individuals was referred to as the base population and individuals pertaining to it were referred to as founders. This was done for pragmatic purposes. However, truncating the pedigree by ignoring individuals without phenotypic records created a group of individuals that may not be the actual base population which is defined as that comprised by ancestors with unknown parents (Henderson, 1974; Kennedy et al., 1988). Conversely, in other cases phenotypic records from this population may be available; thus, estimates of allelic frequencies in the true base population can be obtained. Here, it was further assumed that founders were unrelated which is likely to be false in many situations. However, this assumption has been made in conventional models used to do genetic analysis (Henderson, 1974; Kennedy et al., 1988) because pedigrees are not always completely known. Consequently, what is called the base population is not always the true one. Nevertheless, this assumption seems to be reasonable after so many years of successful artificial selection in animals and plants based on predicted breeding values obtained from these models (Hill, 2014; Gianola and Rosa, 2015).

As discussed in section 2.1.1, the pmf $\pi(W|P^*)$ could be derived ignoring pedigree information. Then, this pmf could be found as the product of all $\pi\left(w_{ij}^{l}\middle|p_{ij}^{*}\right)$ or the product of binomial distributions for gene content (i.e., the number of copies of the reference allele at each locus) across loci and individuals with each binomial distribution depending on the corresponding allelic frequencies. Notice that this requires reparametrizing the mapping of genotypes, that is, instead of having $\{-1,0,1\}$ as possible values of an entry of W, values would be $\{0,1,2\}$. In this case, all individuals in the population would be used to estimate allelic frequencies instead of using information from a base population. If pedigree information is available, it can be easily incorporated into the derivation of $\pi(W|P^*)$ as was shown here and the resulting pmf is not very difficult to evaluate. Furthermore, as mentioned before, direct sampling from this pmf can be done via the inverse transform method for discrete random variables. Notwithstanding, in scenarios where pedigree information is very scarce or not reliable, adding the assumption of independence among individual genotypes and using binomial distributions for the gene content of each individual at each marker locus is an option to model the distribution of matrix W which would induce a joint pmf similar to those presented in Gianola et al., (2010) and Martínez et al. (2015).

If some individuals with phenotypes have only one known parent, the pmf of their genotypes conditioned on this parent and allelic frequencies can be defined in a similar way as was done in Table A.1 for the case of a fully known pedigree (see Appendix C). In this situation, *Remark 1* does not hold and the functional form of $\pi(W|P^*)$ changes which implies that $\pi(W|Else)$ changes as well.

Regarding assumptions about the distribution of allelic frequencies, our models allow for correlations between them. To do that, priors based on a Dirichlet distribution were used. Using these priors require allelic frequencies to be expressed on a complete population basis. This setting brings parameter $\boldsymbol{r}$ into the picture. The algebra associated with this parameter is clear and straightforward, but its interpretation may be fuzzy. From an algebraic standpoint, these parameters are upper boundaries posed over allelic frequencies to force them to be in the support of the prior distribution, thus they can be seen as analytic instruments. Nevertheless, their meaning from the population genetics standpoint is not very clear. Perhaps, the easier interpretation when assuming $r_{1l} = \cdots = r_{ml} = r_l$, is that r_l is the relative frequency or weight of the l^{th} subpopulation. However, making claims about the biological interpretation of this set of parameters is beyond the scope of this study.

From a statistical viewpoint, two approaches were proposed. The first one assumed that $\boldsymbol{r}$ was known (truly known or set to some *ad hoc* value) and $r_{1l} = \cdots = r_{ml} = r_l$. In the examples used here all subpopulations were given the same weight, that is, $r_l = 1/\mathcal{S}\,, \forall\ l = 1,2\ldots,\mathcal{S}$, a pragmatic decision that has been used in other studies, e.g., Gianola et al. (2010). In this scenario, for all j, $\boldsymbol{p}_j$ is modeled as a scaled Dirichlet vector which allows non-null covariances between its elements. The second approach assumed that $\boldsymbol{r}$ was unknown and $\{r_{lj}\}$ varied across marker loci. For each locus the prior was a Dirichlet over allelic frequencies of both alleles in all subpopulations and it permitted obtaining posterior samples of allelic frequencies and $\boldsymbol{r}$. Under the assumption of independence of allelic frequencies, independent priors could be assigned to each marker (e.g., Uniform(0,r_l)) and the validity of this assumption could be tested using criteria as Bayes factors or DIC. If data are pooled and structure is ignored (as done in the null models) the full conditional pdf $\pi(\boldsymbol{p}_0|Else)$ is known and therefore direct sampling can be implemented when matrix W is not completely observed. On the other hand, when it is completely observed the posterior of $\boldsymbol{p}_0$ is known and there is no need of sampling to obtain point estimators. The reason for the full conditional of $\boldsymbol{p}_0$ being a known distribution but not its posterior in the presence of missing genotypes is that W^N is an extra parameter in the model and obtaining the marginal posterior of $\boldsymbol{p}_0$ implies marginalization of $\pi(W^N, \boldsymbol{p}_0|W^{\sigma})$ over W^N which induces a non-standard pmf.

The derivation of the pmf $\pi(W|P^*)$ and $\pi(W_0|\boldsymbol{p}_0)$ not only allow inferences concerning the marker allelic frequencies in the base population, but also allow predictions for non-genotyped or partially genotyped animals without performing a previous imputation. This is likely to increase accuracy of genome-wide predictions because it allows incorporating more phenotypic records. Imputed missing genotypes can be obtained using posterior means or medians of W^N. However, these outputs have to be viewed as a byproduct because these models were not intended to perform imputation. The imputation of missing genotypes is an underlying process in the prediction of genotypic values of individuals with missing genotypes. Notwithstanding, because samples from the posterior of W^N are available and computation of imputed genotypes is simple, there could be interest in using this output of the model and in such case the accuracy of the imputation would also be of interest. Hence, although imputation was not a main objective of our models, it is worth making a brief comment on it. Though an assessment of imputation accuracy is a matter for further research, two statements can be made about the imputation process in our models. Firstly, one advantage of the models developed here is that they automatically take into account the uncertainty of imputation (as a consequence of using a Bayesian approach). Conversely, in the standard approach where genotype imputation is the first step and then a random linear regression model is fitted using these imputed values as if they were observations, uncertainty is not taken into account. Secondly, a disadvantage of our models is that they do not incorporate LD information when imputing missing genotypes, a source of information that is used by some of the current imputation methods (Li et al., 2009). Here, pedigree information, phenotypes and allelic frequencies are used for imputation. Thus, benchmarking of the procedure developed here with current and well-accepted procedures is material for future studies. Furthermore, another question that can be addressed in future research is if improving this imputation as discussed later in section 4.3 has a significant impact on the predictive performance of the models.

As mentioned before, the regression models used in genome-wide prediction treat genotypes as fixed and their effects as random while in the classical quantitative genetics theory genotypes are treated as random and allelic substitution effects as fixed. The set of models developed here are something in between because genotypes are treated as random variables as in classical quantitative genetics, and marker effects are considered random as well like in the standard regression models used in genome-wide prediction. de los Campos et al. (2015b) presented an excellent discussion on the connections between the heritability and the so-called genomic heritability obtained with linear regression models. They show why caution has to be exercised when interpreting the parameters obtained using genomic information due to the fact that sometimes the connection between parameters as the additive genetic variance and the genomic variance are not straightforward. Similarly, Gianola et al. (2015) discussed the fact that connections between genomic correlations and additive genetic correlations are ambiguous. So far, the Bayesian models proposed in this paper are intended to predict breeding values, phenotypes, and to estimate allelic frequencies in a base population using genomic information and no claim is made about the properties of covariance parameters obtained from them.

The discussion above is relevant because the regression variables are not based on genes, but proxies for the causal variants affecting the phenotypes of interest. However, taking into account these limitations and the high degree of caution needed when interpreting parameters obtained from models using molecular markers, some parameters such as the fraction of additive genetic variance explained by the markers are of interest and our models could be used to estimate these quantities.

The family of models developed here could be applied or adapted to different situations. In the simulation, the case of individuals coming from a common founder population pertaining to subpopulations with different selection criteria and mating systems was considered. Other situations in which this set of models could be useful are: 1) simultaneous evaluation of individuals from different breeds or lines, 2) individuals from the same breed or line performing under different environmental conditions (e.g., different geographic regions, production systems, etc.), 3) a combination of numerals 1 and 2, 4) simultaneous evaluation of several correlated traits. In this last case, if all individuals have records for all phenotypes, the design matrix satisfies $W = I_S \otimes W_+$, where W_+ is the matrix of dimension $n \times m$ containing genotypes of n individuals at m marker loci. In this case the model is being adapted to handle correlations between the effects of a given marker locus for different traits in a single population. Consequently, for a given choice of prior and assumption about residuals (heteroscedastic or homoscedastic) the model involves the corresponding hierarchical structure except for the pmf of W conditional on the allelic frequencies and pedigree which is $\pi(W_+|\boldsymbol{p}_0^*)$ instead of $\pi(W|P^*)$. Recent studies have developed Bayesian multiple-trait genome-wide regression models and have shown that predictions from them are more accurate than those coming from genomic univariate models (Jia and Jannink, 2012). The hierarchical Bayesian multivariate genome-wide prediction models proposed by Jia and Jannink (2012) have similar components to the models presented here such as the priors for $\boldsymbol{g}$, but they do not account for randomness of genotypes. Another step to accommodate our models for multiple-trait prediction is to allow correlated residuals, that is, a non-diagonal matrix R. In this case, an inverse Wishart prior can be assigned instead of the inverse gamma prior used here.

4.2 Simulation results

As stated in section 2.4, the aim of this limited simulation was to provide an illustration of the implementation of models and methods developed in this study. Thus, results are not conclusive and further research involving analyses based on more elaborate simulations as well as real datasets to have a better evaluation of the performance of this family of models is needed. Nevertheless, some insights and comments derived from the analyses of these two datasets can be discussed.

The correlation between phenotypes and predicted breeding values (or its square) is one of the most widely used measurements to compare genome-wide prediction models, it is associated with the response to selection and it is easy to compute. On the other hand, as mentioned previously, the DIC combines measures of model adequacy and complexity (Spiegelhalter et al., 2002).

For dataset 1, the squared correlation between phenotypes and predicted breeding values (the predictive ability) did not show an advantage in predictive capability of models taking into account the population structure, i.e., the existence of the subpopulations (Table 6). While measures based on squared correlations did not provide conclusive evidence in favor of the full models, the DIC favored the full models.

As expected, the predictive ability and the other correlations were much smaller in dataset 2 due to the lower heritability of the trait. Although all predictive abilities were low, according to this criterion the performance of the full models was slightly better. Accuracies of predicted breeding values suggested a tiny superiority of null models. The two subpopulations simulated in this dataset diverged by just two generations which could cause only small differences in allelic frequencies, this scenario clearly favors the null models. Accordingly, the DIC component coming from genotypes was slightly better (smaller) for null models as opposed to the case of dataset 1. The total DIC gave evidence in favor of null models. Among predictive ability, accuracy and DIC, accuracy and DIC favored the null models, but the values were very close. The performance of the fitted models was more similar in this dataset than in dataset 1.

In our small simulations, when subpopulations diverged by several generations, migration was allowed and heritabilities were high (dataset 1), full models had better performance in terms of DIC. Conversely, when populations diverged by only a few generations, there was no migration, and heritabilities were low (dataset 2) null models tended to perform better according to this criterion. However, the differences were small. On the other hand, predictive abilities showed a different pattern. In dataset 1 this criterion was higher for null models while in dataset 2 it was smaller for null models. Another feature shown by these simulations was the high variability in model performance that may exist among populations. In dataset 1, according to all criteria except the W component of DIC, the performance of model M_{1G} tended to be remarkably poorer while this was not the case in dataset 2.

Other authors have found modest or null increments in predictive performance of models allowing heterogeneous marker effects across subpopulations compared to pooling data and analyzing the complete population as a single one (Olson et al., 2012; Makgahlela et al.,2013; de los Campos et al., 2015a). All the aforementioned studies used real data from plants and animals. Working with three plant populations and using a model very similar to those proposed here, Lehermeier et al. (2015) found cases in which the strategy of pooling data and ignoring structure performed better and other cases where multivariate models yielded better predictive performance. These authors found that in highly differentiated populations within group and multivariate analyses performed better while the converse occurred in closely related subpopulations with small sample sizes. Roughly speaking, these results are in agreement with the results found in this study.

Using predictive ability, Lund et al. (2011) found a higher accuracy of predicted additive breeding values when pooling the data compared with individual analyses. Similar results were found by de Roos et al. (2009) when heritability was low, divergence of populations was small (small number of generations) and marker density was high (more persistent phase), and by Wientjes et al. (2015) when the QTL effects did not change across subpopulations. Pooling data and ignoring the population structure corresponds to the null models defined in this study, except that models considered by the authors just cited did not account for randomness of genotypes. In our simulation, individual analyses were not considered. Sample size is one the factors affecting the accuracy of genome-wide predictions (Meuwissen et al., 2001; Goddard 2009, Zhong et al., 2009). Presumably it was one of the leading factors causing the results found by Lund et al. (2011). In addition, the Holstein breed is highly inbred and there were several individuals connecting the different populations; this probably made them similar. On the other hand, the studies of de Roos et al. (2009) and Wientjes et al. (2015) used simulated data and explored different scenarios. Both studies found situations in which pooling data was not advantageous.

4.3 Refinements and extensions

In this section, some comments regarding possible extensions and refinements of different aspects of the family of models presented in the study are briefly discussed.

In the derivation of the joint pmf of W conditional on P^* and pedigree information, row-wise dependence due to kinship was taken into account by using pedigree information to accommodate relationships among genotypes of related individuals. This task was highly simplified due to the conditional independence argument that permitted to find a simpler decomposition of the joint pmf and therefore, a simpler algebraic expression. However, the possible existence of column-wise dependence due to LD was ignored here in order to make the problem more tractable from the mathematical point of view. This is an assumption frequently used in theoretical studies in quantitative genetics and it is well-accepted at least in studies concerned with first approximations to a given problem. For example, Gianola et al. (2009) treated a series of theoretical aspects of some of the Bayesian regression models used in genome-wide prediction using the assumption of linkage equilibrium which implies the mutual independence of the columns of W used here (they also developed some results accounting for LD in the Appendix). Most of the models currently used in genome-wide prediction are also based on this assumption, few approximations to deal with consequences of LD have been proposed (Gianola et al., 2003; Yang and Tempelman, 2012), but these have not yet been adopted in routine genetic evaluations. Their models do not consider randomness in the genotypes; thus, a consequence of considering LD in these models is the need to account for covariances between marker effects at different loci. Consequently, a refinement of our family of models in this regard, would be to accommodate LD, which can be performed at two levels: 1) account for correlations among columns of W, and 2) use a non-block-diagonal G matrix. A potential consequence of accounting for non-independence of the columns of W could be the reduction in the cardinality of $\mathcal{G}$ that is induced by the fact that the number of possible values of a column of W depends on the values at one or more different columns (as it happened with rows). Another assumption made here was the absence of mutations which caused that when conditioning on the genotypes of the parents of an individual, the probabilities of its genotype taking a given value were completely defined by the parental genotypes, making this random variable conditionally independent of allelic frequencies. Thus, another refinement in $\pi(W|P^*)$ would be to account for mutation. Therefore, the derivation of $\pi(W|P^*)$ to accommodate dependence between columns of W and mutation, and the impact of this refinement on predictive performance and the accuracy of imputed genotypes (if it is of interest) pose a problem for further research.

If relationships among founders (as defined in this paper) were to be taken into account, from the theoretical point of view it is not hard to visualize how to do it. For the sake of simplicity, the case of two individuals and one locus is considered; consequently, the sub-index associated with locus is omitted. Let W_1, W_2 be the genotypes of individuals 1 and 2, and $W_{\mathcal{C}}$ the genotypes of the set of relevant common ancestors. Suppose that 1 is not a parent of 2. Then:

$$\pi(W_1, W_2|P^*) = \sum_{\mathcal{G}^{\mathcal{C}}} \pi(W_1, W_2|W_{\mathcal{C}}, P^*)\pi(W_{\mathcal{C}}|P^*) = \sum_{\mathcal{G}^{\mathcal{C}}} \pi(W_1|W_{\mathcal{C}}, P^*)\pi(W_2|W_{\mathcal{C}}, P^*)\pi(W_{\mathcal{C}}|P^*),$$

where $\mathcal{G}^{\mathcal{C}}$ is the set of possible values that the set of genotypes of relevant common ancestors can take according to the pedigree (as explained in section 2.1.1) and the second equality follows from the conditional independence of the genotypes of individuals 1 and 2 given the common ancestors and allelic frequencies. By relevant common ancestors it is meant that the genotypes of these ancestors provide information about the genotypes of 1 and 2 when conditioning on the full set of common ancestors, i.e., if $\mathcal{D}$ is the whole set of common ancestors then $\mathcal{D} = \mathcal{C} \cup \mathcal{C}^c$ (the super-index c means complement with respect to $\mathcal{D}$) and $\pi(W_1, W_2|W_{\mathcal{D}}, P^*) = \pi(W_1, W_2|W_{\mathcal{C}}, P^*)$. Notice that unless individuals 1 and 2 are full sibs, their conditional pmf given the relevant common ancestors depends on P^*. Of course, it makes $\pi(W|P^*)$ a more complex expression and reduces the cardinality of $\mathcal{G}$. See Appendix D for a toy example of $\pi(W_1, W_2|W_{\mathcal{C}}, P^*)$ when 1 and 2 are half sibs. Although the problem is tractable from the theoretical standpoint, it may be difficult to compute these values especially with complex pedigrees where the set of common ancestors may be large such as those found in animal and plant populations. The example in Appendix D shows that even in a simple case, computation of $\pi(W_1, W_2|P^*)$ is involved.

5. Conclusions

The main contribution of this paper is the theoretical development of a set of models for across population genome-wide prediction incorporating marker genotypes not only as explanatory variables of regression models, but also as realizations of random variables providing information about allelic frequencies and missing genotypes. Although models were intended for across population analysis, they can also be applied in single population studies and adapted for multiple-trait prediction.

Theoretical and computational issues along with possible applications as well as some extensions and refinements of these models pose several problems for future research. Our models treat both genotypes and marker allelic substitution effects as random; therefore, they combine features from classical quantitative genetics theory and traditional genome-wide prediction models.

Some features of the models developed in this study make them promising for genome-wide prediction. Among these, the ability to include phenotypes from individuals with missing genotypes at some or all loci without the need of previous imputation and accounting for uncertainty about imputed genotypes as well as heterogeneity of allelic frequencies across subpopulations are perhaps the most appealing. Further research to assess their performance and also to compare them with other models used in genome-wide prediction is needed.

Author Contributions

C.A. Martínez developed modeling strategies, carried out the derivations, wrote the R scripts, designed and made the simulations and wrote the paper. K. Khare advised modeling strategies, reviewed, corrected and discussed the derivations and the statistical aspects of the paper. A. Banerjee advised modeling strategies, reviewed, corrected and discussed the derivations and the statistical aspects of the paper. M.A. Elzo designed the simulation, reviewed, corrected and discussed the genetic aspects of the paper.

Acknowledgments

Authors acknowledge Dr. Malay Ghosh from the Department of Statistics of the University of Florida for useful comments and discussions, and for pointing out relevant references. C. A. Martínez also thanks PhD students Hunter Merril and Isaac Duerr, and Dr. Nikolay Bliznyuk from the Department of Agricultural and Biological Engineering of the University of Florida for their advice in computational issues, Fulbright Colombia and “Departamento Adiministrativo de Ciencia, Tecnología e Innovación” COLCIENCIAS for supporting his PhD and Master programs at the University of Florida through a scholarship, and Bibiana Coy for her love, support and constant encouragement.

References

Bernardo, R., Yu, J. (2007). Prospects for Genomewide Selection for Quantitative Traits in Maize. *Crop Science*, 47, 1082-1090.

Casella, G., George, E.I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46(3), 167-174.

Casella, G., Berger, R. (2002). *Statistical Inference* (2nd ed.). Duxbury, Pacific Grove, CA, USA.

Chen, L., Li, C., Miller, S., Schenkel, F. (2014). Multi-population genomic prediction using a multi-task Bayesian learning model. *BMC Genetics*, 15:53.

de los Campos, G., Gianola D., Allison, D.B. (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics*, 11, 880-886.

de los Campos, G., Veturi, Y., Vázquez, A.I., Lehermeier, C., Pérez-Rodríguez, P. (2015a). Incorporating genetic heterogeneity in whole-genome regressions using interactions. Journal of Agricultural, Biological and Environmental Statistics, 20(4), 467-490.

de los Campos, G., Sorensen, D., Gianola, D. (2015b). Genomic Heritability: What Is It? *PLOS Genetics*, 11(5), e1005048.

de Roos, A.P.W., Hayes, B.J., Goddard, M.E. (2009). Reliability of Genomic Predictions Across Multiple Populations. *Genetics*, 183, 1545-1553.

Desta, Z.A., Ortiz, R. (2014). Genomic selection: genome-wide prediction in plant improvement. *Trends in Plant Science*, 19(9), 592-601.

Falconer, D.S., Mackay, T.F.C. (1996). *Introduction to Quantitative Genetics* (4th ed.). Longmans Green, Harlow, UK.

Gelman, A., Carlin, J.B., Stern, H., Dunson, D.B., Vehtari, A., Rubin, D.B. (2013). *Bayesian Data Analysis* (3rd ed.). Chapman and Hall/CRC, Boca Raton, FL, USA.

Gianola, D., Perez-Encizo, M., Toro, M.A. (2003). On Marker-Assisted Prediction of Genetic Value: Beyond the Ridge. *Genetics*, 163, 347-365.

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

Gianola, D., Simianer, H., Qanbari, S. (2010) A two-step method for detecting selection signatures using genetic markers. *Genetic Research Cambridge*, 92(2), 141-155.

Gianola, D. (2013). Priors in whole-genome regression: The Bayesian alphabet returns. *Genetics*, 194, 573-596.

Gianola, D., de los Campos, G., Toro, M.A., Naya, H., Schön, C.C., Sorensen, D. (2015). Do Molecular Markers Inform About Pleiotropy? *Genetics*, 201, 23-29.

Gianola, D., Rosa, G. (2015). One Hundred Years of Statistical Developments in Animal Breeding. *Annual Reviews of Animal Biosciences*, 3, 19-56.

Goddard, M.E., Hayes, B.J. (2007). Genomic Selection. *Journal of Animal Breeding and Genetics*, 124, 323-330.

Goddard, M.E. (2009). Genomic selection: prediction of accuracy and maximization of long term response. *Genetica*, 136, 245-257.

Guttmacher, A.E., Collins, F.S. (2002).Genomic medicine- a primer. *The New England Journal of Medicine*, 347, 1512-1520.

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

Hill, W.G. (1984). On selection among groups with heterogeneous variance. *Animal Production*, 39(3), 473-477.

Hill, W.H. (2014). Applications of Population Genetics to Animal Breeding, from Wright, Fisher and Lush to Genomic Prediction. *Genetics*, 196, 1-16.

Henderson, C.R. (1974). Use of all relatives in intraherd prediction of breeding values and producing abilities. *Journal of Daity science, 58(12),* 1910-1916.

Huang, H., Windig, J.J., Vereijken, A., Calus, M.P.L. (2014). Genomic prediction based on data from three layer lines using non-linear regression models. *Genetics Selection Evolution*, 46:75.

Jia, Y., Jannink, J.L. (2012) Multiple-Trait Genomic Selection Methods Increase Genetic Value Prediction Accuracy. *Genetics*, 192, 1513-1522.

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

Kennedy, B.W., Schaeffer, L.R., Sorensen, D.A. (1988). Genetic properties of animal models. *Journal of Dairy Science, 71(2),* 17-26.

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

Li, Y., Willer, C., Sanna, S., Abecasis, G. (2009). Genotype imputation. *Annual Review of Genomics and Human Genetics*, *10*, 387-406.

Lund, M.S., de Roos, A.P.W., de Vries, A.G., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F., Su, J. (2011). A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genetics Selection Evolution*, 43:43.

Lynch, M., Walsh, E. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer associates Inc., Sunderland, MA, USA.

Maier, R., Moser, G., Chen, G.B., Ripke, S., Cross-Disorder Working Group of the Psychiatric Genomics Consortium, Coryell, W., Potash, J.B., Scheftner, W.A., Shi, J., Weissman, M.M., Hultman, C.M., Lande´n, M., Levinson, D.F., Kendler, K.S., Smoller, J.W., Wray, N.R., Lee, S.H. (2015). Joint Analysis of Psychiatric Disorders Increases Accuracy of Risk Prediction for Schizophrenia, Bipolar Disorder, and Major Depressive Disorder. *The American Journal of Human Genetics*, 96, 283-294.

Makgahlela, M.L., Mantysaari, E.A., Stranden, I., Koivula, M., Nielsen, U.S., Sillanpaa, M.J., Juga, J. (2013). Acroos breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. *Journal of Animal Breeding and Genetics*, 130, 10-19.

Martínez, C.A., Khare, K., Elzo, M.A. (2015). On the Bayesness, minimaxity and admissibility of point estimators of allelic frequencies. *Journal of Theoretical Biology*, 383, 106-115.

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

Olson, K.M., VanRaden, P.M., Tooker, M.E. (2012). Multibreed genomic evaluations using purebred Holsteins, Jerseys, and Brown Swiss. *Journal of Dairy Science*, 95, 5378-5383.

R Core Team (2015). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. *URL https://www.R-project.org/.*

Robert, C.P., Casella, G. (2010). *Introducing Monte Carlo Methods with R*. Springer, New York, NY, USA.

Sargolzaei, M., Schenkel, F.S. (2013). *QMSim User's Guide Version 1.10*. Centre for Genetic Improvement of Livestock, Department of Animal and Poultry Science, University of Guelph, Guelph, Canada.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society Series B*, 64, 583–639.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society Series B*, 76, 485–493.

van den Berg, S., Calus, M.P.L., Meuwissen, T.H.E., Wientjes, Y.C.J. (2015). Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. *BMC Genetics*, 16:416.

Wientjes, Y.C.J., Veerkamp, R.F., Bijma, P., Bovenhuis, H., Schrooten, C., Calus, M.P.L. (2015) Empirical and deterministic accuracies of across-population genomic prediction. *Genetics Selection Evolution*, 47:5.

Wright, S. (1930). Evolution in Mendelian populations. *Genetics*, 16, 98-159.

Wright, S. (1937). The distribution of genetic frequencies in populations. *Genetics*, 23, 307-320.

Yang, W., Tempelman, R.J. (2012). A Bayesian Antedependence Model for Whole Genome Prediction. *Genetics*, 190, 1491-1501.

Zhong, S., Dekkers, J.C.M., Fernando, R.L., Jannink, J.K. (2009). Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. *Genetics*, 182, 355-364.

Appendix A: Conditional pmf of genotypes given parental genotypes, joint posteriors, full conditionals and details of some derivations

Table A.1 Conditional pmf of genotypes at locus j given the parental genotypes

Parental genotypes		**Corresponding random variables**		$\pi(w_{ij} \| w_{S_ij}, w_{D_ij}) = \Pr(w_{ij} = x \| w_{S_ij} = k, w_{D_ij} = z)$ $x, k, z \in \{-1,0,1\}$		
Parent 1	Parent 2	w_{S_ij}	w_{D_ij}	$\pi(-1 \| w_{S_ij}, w_{D_ij})$	$\pi(0 \| w_{S_ij}, w_{D_ij})$	$\pi(1 \| w_{S_ij}, w_{D_ij})$
AA	AA	-1	-1	1	0	0
AA(BB)	BB(AA)	-1(1)	1(-1)	0	1	0
AB	AB	0	0	¼	½	¼
AA(AB)	AB(AA)	-1(0)	0(-1)	½	½	0
BB(AB)	AB(BB)	1(0)	0(1)	0	½	½
BB	BB	1	1	0	0	1

Joint posteriors for Homogeneous marker effect covariance matrix model with homoscedastic residuals and Gaussian prior for $\boldsymbol{g}$

Weather P^* is considered a parameter (some founders are genotyped) or a hyperparameter (none of the founders is genotyped) is not relevant when computing the joint posterior because in both cases its pdf is the same, thus it enters in the expression in the same way. Henceforth, it is assumed that vector $\boldsymbol{g}$ and columns of matrix W are ordered by marker unless otherwise indicated. Thus:

$$\pi(\boldsymbol{g}, \sigma^2, W^N, G^0, P^* | \boldsymbol{y}, W^{\mathcal{O}}) \propto f(\boldsymbol{y} | \boldsymbol{g}, \sigma^2, W)\pi(\boldsymbol{g} | G^0)\pi(G^0)\pi(\sigma^2)\pi(W | P^*)\pi(P^*)$$
$$\propto (\sigma^2)^{-\frac{n}{2}} \exp\left(\frac{-1}{2\sigma^2}(\boldsymbol{y} - W\boldsymbol{g})'(\boldsymbol{y} - W\boldsymbol{g})\right)$$
$$\times |G^0|^{-\frac{m}{2}} \exp\left(\frac{-1}{2}\boldsymbol{g}'(I_m \otimes (G^0)^{-1})\boldsymbol{g}\right)$$
$$\times |G^0|^{-\frac{1}{2}(a+\mathcal{S}+1)} \exp\left(\frac{-1}{2} tr(\boldsymbol{\Sigma}(G^0)^{-1})\right)$$
$$\times (\sigma^2)^{-\left(\frac{v}{2}+1\right)} \exp\left(\frac{-\tau^2}{2\sigma^2}\right)$$
$$\times \pi(W \,\big|\, P^*)\pi(P^*)$$

Where $\otimes$ represents the Kronecker product and $\pi(W|P^*)\pi(P^*) = \pi(W, P|\boldsymbol{r})$, when $\boldsymbol{r}$ is assumed to be known and has the following form (see appendix A for details):

$$\pi(W, P|\boldsymbol{r}) \propto$$
$$2^{n^H} \prod_{j=1}^{m} p_{(\mathcal{S}+1)j}^{\alpha_{\mathcal{S}+1}-1} \prod_{l=1}^{\mathcal{S}} \left\{ \frac{1}{r_l^{2f_l}} p_{lj}{}^{n_l^{B_j}+\alpha_l-1} \left(r_l - p_{lj}\right)^{n_l^{A_j}} \prod_{i'=f_l+1}^{n_l} \pi\left(w_{i'j}^{l} | w_{S_{i'j}}, w_{D_{i'j}}\right) \right\}.$$

When $\boldsymbol{r}$ is unknown, the only change is that expression $(p_{lj})^{n_l^{B_j}+\alpha_l-1}(r_l-p_{lj})^{n_l^{A_j}}$ has to be replaced by $(p_{lj})^{n_l^{B_j}+\alpha_{lp}-1}(r_{lj}-p_{lj})^{n_l^{A_j}+\alpha_{lq}-1}$ and instead of $\pi(W,P|\boldsymbol{r}),\pi(W\,\big|\,P^*)\pi(P^*)$ corresponds to $\pi(W,P,Q),Q\coloneqq(\boldsymbol{q}_1,\dots,\boldsymbol{q}_j)$.

Joint posteriors for Heterogeneous marker effect covariance matrix model with homoscedastic residuals and Gaussian prior for $\boldsymbol{g}$

$$
\begin{aligned}
\pi(\boldsymbol{g},\sigma^2,W^N,G,P|\boldsymbol{y},W^{\mathcal{O}}) &\propto f(\boldsymbol{y}|\boldsymbol{g},\sigma^2,W)\pi(\boldsymbol{g}|G)\pi(G)\pi(\sigma^2)\pi(W|P)\pi(P)\\
&\propto (\sigma^2)^{-\frac{n}{2}}\exp\left(\frac{-1}{2\sigma^2}(\boldsymbol{y}-W\boldsymbol{g})'(\boldsymbol{y}-W\boldsymbol{g})\right)\\
&\times|G^0|^{-\frac{m}{2}}\exp\left(\frac{-1}{2}\boldsymbol{g}'(I_m\otimes(G^0)^{-1})\boldsymbol{g}\right)\\
&\times\prod_{j=1}^{m}\left\{\left|G_j\right|^{-\frac{1}{2}(a+\mathcal{S}+1)}\right\}\exp\left(\frac{-1}{2}\sum_{j=1}^{m}tr\left(\boldsymbol{\Sigma}G_j^{-1}\right)\right)\\
&\times(\sigma^2)^{-\left(\frac{v}{2}+1\right)}\exp\left(\frac{-\tau^2}{2\sigma^2}\right)\\
&\times\pi(W|P^*)\pi(P^*).
\end{aligned}
$$

Marginal prior distribution of marker effects

Homogeneous marker effect covariance matrix models

$$
\pi(\boldsymbol{g})\propto\int_{\mathcal{P}_{\mathcal{S}}^{+}}\pi(\boldsymbol{g}|G^0)\pi(G^0)dG^0
$$

$$
\propto\int_{\mathcal{P}_{\mathcal{S}}^{+}}\exp\left(\frac{-1}{2}tr\left(\left(\boldsymbol{\Sigma}+\sum\nolimits_{j=1}^{m}\boldsymbol{g}_j\boldsymbol{g}_j'\right)(G^0)^{-1}\right)\right)|G^0|^{-\frac{1}{2}(a+\mathcal{S}+m+1)}dG^0
$$

the expression $tr\left(\left(\boldsymbol{\Sigma}+\sum_{j=1}^{m}\boldsymbol{g}_j\boldsymbol{g}_j'\right)(G^0)^{-1}\right)$ comes from adding terms $\boldsymbol{g}'(I_m\otimes(G^0)^{-1})\boldsymbol{g}$ coming from $\pi(\boldsymbol{g}|G^0)$ and $tr(\boldsymbol{\Sigma}(G^0)^{-1})$ coming from $\pi(G^0)$. The equality is shown using properties of the $tr(\cdot)$ operator as follows:

$$
\begin{aligned}
\boldsymbol{g}'(I_m\otimes(G^0)^{-1})\boldsymbol{g} &= tr(\boldsymbol{g}'(I_m\otimes(G^0)^{-1})\boldsymbol{g})\\
&= tr\left(\begin{pmatrix}\boldsymbol{g}_1 & \cdots & \boldsymbol{g}_m\end{pmatrix}\begin{pmatrix}(G^0)^{-1} & & \\ & \ddots & \vdots\\ & & (G^0)^{-1}\end{pmatrix}\begin{pmatrix}\boldsymbol{g}_1\\ \vdots\\ \boldsymbol{g}_m\end{pmatrix}\right)
\end{aligned}
$$

$$= tr\left(\sum_{j=1}^{m} \boldsymbol{g}_j'(G^0)^{-1}\boldsymbol{g}_j\right)$$

$$= tr\left(\sum_{j=1}^{m} \boldsymbol{g}_j\boldsymbol{g}_j'(G^0)^{-1}\right),$$

moreover, since $tr(\boldsymbol{g}'(I_m \otimes (G^0)^{-1})\boldsymbol{g}) = tr\big(\boldsymbol{g}\boldsymbol{g}'(I_m \otimes (G^0)^{-1})\big)$, it follows that:

$$\boldsymbol{g}'(I_m \otimes (G^0)^{-1})\boldsymbol{g} + tr(\boldsymbol{\Sigma}(G^0)^{-1}) = tr\big(\boldsymbol{\Sigma}(G^0)^{-1} + \boldsymbol{g}\boldsymbol{g}'(I_m \otimes (G^0)^{-1})\big)$$

$$= tr\left(\boldsymbol{\Sigma}(G^0)^{-1} + \sum_{j=1}^{m} \boldsymbol{g}_j\boldsymbol{g}_j'(G^0)^{-1}\right)$$

$$= tr\left(\left(\boldsymbol{\Sigma} + \sum_{j=1}^{m} \boldsymbol{g}_j\boldsymbol{g}_j'\right)(G^0)^{-1}\right).$$

Using this, it follows that:

$$\pi(\boldsymbol{g}) \propto \int_{\mathcal{P}_{\mathcal{S}}^{+}} \exp\left(\frac{-1}{2} tr\left(\left(\boldsymbol{\Sigma} + \sum\nolimits_{j=1}^{m} \boldsymbol{g}_j\boldsymbol{g}_j'\right)(G^0)^{-1}\right)\right) |G^0|^{-\frac{1}{2}(a+\mathcal{S}+m+1)} dG^0$$

$$= \frac{2^{(a+m)\mathcal{S}/2}\Gamma_{\mathcal{S}}\left(\frac{a+m}{2}\right)}{\left|\boldsymbol{\Sigma} + \sum_{j=1}^{m} \boldsymbol{g}_j\boldsymbol{g}_j'\right|^{\left(\frac{a+m}{2}\right)}}$$

This result easily follows because we are integrating the kernel of an inverse Wishart density with parameters $\left(\boldsymbol{\Sigma} + \sum_{j=1}^{m} \boldsymbol{g}_j\boldsymbol{g}_j', a+m\right)$.

Heterogeneous marker effect covariance matrix model

For this model:

$$\pi(\boldsymbol{g}) \propto \prod_{j=1}^{m} \int_{\mathcal{P}_{\mathcal{S}}^{+}} |G_j|^{-\frac{1}{2}(a+\mathcal{S}+2)} \exp\left(\frac{-1}{2} tr\left(\left(\boldsymbol{\Sigma} + \boldsymbol{g}_j\boldsymbol{g}_j'\right)\left(G_j\right)^{-1}\right)\right) dG_j$$

$$= \prod_{j=1}^{m} \frac{2^{(a+1)\mathcal{S}/2}\Gamma_{\mathcal{S}}\left(\frac{a+1}{2}\right)}{\left|\boldsymbol{\Sigma} + \boldsymbol{g}_j\boldsymbol{g}_j'\right|^{\left(\frac{a+1}{2}\right)}}$$

$$= \frac{2^{(a+1)m\mathcal{S}/2}\left(\Gamma_{\mathcal{S}}\left(\frac{a+1}{2}\right)\right)^{m}}{\prod_{j=1}^{m}\left|\boldsymbol{\Sigma}+\boldsymbol{g}_j\boldsymbol{g}_j'\right|^{\left(\frac{a+1}{2}\right)}},$$

using the results for determinants of partitioned matrices this expression can be written as:

$$\frac{2^{(a+1)m\mathcal{S}/2}\left(\Gamma_{\mathcal{S}}\left(\frac{a+1}{2}\right)\right)^{m}}{\prod_{j=1}^{m}\left(|\boldsymbol{\Sigma}|\left|1+\boldsymbol{g}_j'\boldsymbol{\Sigma}^{-1}\boldsymbol{g}_j\right|\right)^{\left(\frac{a+1}{2}\right)}} \propto \frac{1}{\prod_{j=1}^{m}\left(1+\frac{1}{a+1-\mathcal{S}}\boldsymbol{g}_j'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{g}_j\right)^{\left(\frac{a+1}{2}\right)}},$$

where $\boldsymbol{\Sigma}_* = \frac{1}{a+1-\mathcal{S}}\boldsymbol{\Sigma}$**.** This is the product of multivariate t distributions with scale matrix $\boldsymbol{\Sigma}_*$ and degrees of freedom $a+1-\mathcal{S}$.

Details on the form of $\boldsymbol{\pi}(\boldsymbol{W},\boldsymbol{P}^*)$, $\boldsymbol{r}$ known

$$\begin{aligned}
\pi(W|P^*)\pi(P^*) &= \pi(W|P,\boldsymbol{r})\pi(P|\boldsymbol{r}) \\
&= \pi(W|P,\boldsymbol{r})\prod_{j=1}^{m}\pi(p_j|\boldsymbol{r}) \\
&= \frac{2^{n^H}}{c}\prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}\left\{\frac{1}{r_l^{2f_l}}p_{lj}^{n_l^{B_j}}\left(r_l-p_{lj}\right)^{n_l^{A_j}}\prod_{i'=f_l+1}^{n_l}\pi\left(w_{i'j}^{l}\middle|w_{S_{i'j}},w_{D_{i'j}}\right)\right\}\prod_{j=1}^{m}\pi(p_j|\boldsymbol{r}) \\
&\propto \frac{2^{n^H}}{c}\prod_{j=1}^{m}\prod_{l=1}^{\mathcal{S}}\left\{\frac{1}{r_l^{2f_l}}p_{lj}^{n_l^{B_j}}\left(r_l-p_{lj}\right)^{n_l^{A_j}}\prod_{i'=f_l+1}^{n_l}\pi\left(w_{i'j}^{l}\middle|w_{S_{i'j}},w_{D_{i'j}}\right)\right\}\times\prod_{j=1}^{m}p_{(\mathcal{S}+1)j}^{\alpha_{\mathcal{S}+1}-1}\prod_{l=1}^{\mathcal{S}}\left(\frac{p_{lj}}{r_l}\right)^{\alpha_l-1} \\
&\propto \frac{2^{n^H}}{c}\prod_{j=1}^{m}p_{(\mathcal{S}+1)j}^{\alpha_{\mathcal{S}+1}-1}\prod_{l=1}^{\mathcal{S}}\left\{\frac{1}{r_l^{2f_l}}p_{lj}^{n_l^{B_j}+\alpha_l-1}\left(r_l-p_{lj}\right)^{n_l^{A_j}}\prod_{i'=f_l+1}^{n_l}\pi\left(w_{i'j}^{l}\middle|w_{S_{i'}},w_{D_{i'}}\right)\right\}
\end{aligned}$$

$p_{(\mathcal{S}+1)j} = 1-\sum_{l=1}^{\mathcal{S}}p_{lj}$, for each j, $\boldsymbol{g}_j \in \mathbb{R}^{\mathcal{S}}$ corresponds to the subvector of $\boldsymbol{g}$ containing the effects of marker j in each one of the $\mathcal{S}$ subpopulations and $\otimes$ represents the Kronecker product. Analogous steps lead to the form of $\pi(W,P^*)$ when $\boldsymbol{r}$ is unknown.

Full conditionals

Full conditionals for models with heteroscedastic residuals

In this case:

$$\begin{aligned}
f(\boldsymbol{y}|W,\boldsymbol{g},R) &\propto |V|^{-1/2}\exp\left(-\frac{1}{2}(\boldsymbol{y}-W\boldsymbol{g})'V^{-1}(\boldsymbol{y}-W\boldsymbol{g})\right) \\
&= \prod_{l=1}^{\mathcal{S}}(\sigma_l^2)^{-n_l/2}\exp\left(-\frac{1}{2\sigma_l^2}(\boldsymbol{y}_l-W_l\boldsymbol{g}_l)'(\boldsymbol{y}_l-W_l\boldsymbol{g}_l)\right).
\end{aligned}$$

In addition

$$\pi(R) \propto \prod_{l=1}^{\mathcal{S}} (\sigma_l^2)^{-(v/2+1)} \exp\left(-\frac{\tau^2}{2\sigma_l^2}\right).$$

In the following, only the full conditionals that change with respect to the homoscedastic models are presented. For the homogeneous marker effect covariance matrix model with multivariate normal prior the full conditionals that change are:

$$\pi(\boldsymbol{g}|Else) = MVN((W'V^{-1}W + G^{-1})^{-1}W'V^{-1}\boldsymbol{y}, (W'V^{-1}W + G^{-1})^{-1})$$

where $G^{-1} = (G^0)^{-1} \otimes I$.

$$\pi(R|Else) = \prod_{l=1}^{\mathcal{S}} IG\left(\frac{v + n_l}{2}, \frac{\tau^2 + (\boldsymbol{y}_l - W_l\boldsymbol{g}_l)'(\boldsymbol{y}_l - W_l\boldsymbol{g}_l)}{2}\right).$$

To define $\pi(W^N|Else)$ the partitions defined in section 2.2.1 are done for each subpopulation.

$$\pi(W^N|Else) \propto \pi^+(W|P^*) \prod_{l=1}^{\mathcal{S}} \exp\left(\frac{-1}{2\sigma_l^2}(-2\boldsymbol{g}_l' W_l^{N\prime}\boldsymbol{y}_l^N + \boldsymbol{g}_l' W_l^{N\prime} W_l^N \boldsymbol{g}_l)\right)$$
$$\times \prod_{l=1}^{\mathcal{S}} \prod_{k=1}^{K} \exp\left(\frac{-1}{2\sigma_l^2} h\left(W_l^{M_k}, \boldsymbol{g}_l^{M_k}, \boldsymbol{y}_l^{M_k}\right)\right)$$

where

$$h\left(W_l^{M_k}, \boldsymbol{g}_l^{M_k}, \boldsymbol{y}_l^{M_k}\right)$$
$$= 2\left(\boldsymbol{g}_l^{M_kN\prime} W_l^{M_kN\prime} W_l^{M_k\sigma} \boldsymbol{g}_l^{M_k\sigma} - \boldsymbol{g}_l^{M_kN\prime\prime} W_l^{M_kN\prime} \boldsymbol{y}_l^{M_k}\right) + \boldsymbol{g}_l^{M_kN\prime} W_l^{M_kN\prime} W_l^{M_kN} \boldsymbol{g}_l^{M_kN}.$$

For the heterogeneous marker effect covariance matrix model with multivariate Gaussian prior for $\boldsymbol{g}$:

$$\boldsymbol{g}|Else \sim MVN((W'V^{-1}W + G^{-1})^{-1}W'V^{-1}\boldsymbol{y}, (W'V^{-1}W + G^{-1})^{-1})$$

where $G^{-1} = Block\ Diag\ (G_1^{-1}, \dots, G_{\mathcal{S}}^{-1})$.

Appendix B: Details on data simulation

For phenotype one (dataset 1), in a first stage three preliminary subpopulations were simulated by selecting individuals from the historical population. Numbers of individuals and criteria to select them were the following. In preliminary subpopulation 1, ten males and 250 females with the lowest true breeding values, in preliminary subpopulation 2, five males and 200 females with the highest phenotypes and in preliminary subpopulation 3, 50 males and 500 females randomly chosen. Then, selection criteria and mating design to create new generations were: lowest phenotypes and positive assortative in preliminary subpopulation 1, highest phenotypic values and random for preliminary subpopulation 2, and random and random for preliminary subpopulation 3. Positive assortative means that individuals are mated looking for similarity, while negative assortative means that individuals are mated looking for dissimilarity, where (di)similarity can be defined in terms of phenotypes, true or predicted breeding values (Sargolzaei and Schenkel, 2013). The numbers of simulated generations were four, two, and three respectively. Subsequently, two more subpopulations hereinafter referred to as subpopulations one and two were simulated as follows. Eighteen males and

100 females from the fourth generation of the first subpopulation, two males and 40 females from the second generation of the second subpopulation, and eight males and 40 females from the third generation of the third subpopulation were chosen to create the subpopulation one. Ten females from generation three of preliminary subpopulation one, 20 males and 60 females from generation two of preliminary subpopulation two, and 20 females from generation two of preliminary subpopulation three were chosen to generate subpopulation two. Generations zero and one of preliminary subpopulation three were used to define subpopulation three. For the second phenotype (dataset 2) the two subpopulations were simulated by choosing individuals from the historical subpopulation based on different criteria and mating them according to different systems and selection criteria for two generations.

In each case, a single pedigree was simulated which allowed individuals from a given subpopulation to be parents of individuals from another subpopulation. This mimics what happens in certain populations like animal populations when using semen or oocytes from individuals from a different subpopulation (e.g., country) to produce a new generation of a given subpopulation. The number of alleles per QTL was two, three and four; these numbers were randomly assigned using a uniform distribution. QTL were evenly allocated across the genome as well as SNP markers.

In both datasets, additive QTL effects were scaled such that QTL effects and heritabilities were different in each subpopulation. Within a given subpopulation, all QTL allelic effects were scaled by the same factor. Markers with minor allele frequencies smaller than 0.05 were excluded from the analysis.

Appendix C: Conditional pmf of genotypes at locus *j* given one parental genotype and allelic frequencies

The following table shows $\pi\left(w_{ij}\middle|w_{Pa_ij},\boldsymbol{p}_j^*\right)=\Pr\left(w_{ij}=x\middle|w_{Pa_ij}=z,\boldsymbol{p}_j^*\right), x,z\in\{-1,0,1\}$, where w_{Pa_ij} is the genotype of the known parent of individual i for marker locus j. If the subpopulation to which the unknown parent pertains is known to be subpopulation l then $\pi\left(w_{ij}\middle|w_{Pa_ij},\boldsymbol{p}_j^*\right)$ has the following form:

Known parental genotype	w_{Pa_ij}	$\pi\left(-1\middle|w_{Pa_ij},\boldsymbol{p}_j^*\right)$	$\pi\left(0\middle|w_{Pa_ij},\boldsymbol{p}_j^*\right)$	$\pi\left(1\middle|w_{Pa_ij},\boldsymbol{p}_j^*\right)$
AA	-1	$1-p_{lj}^*$	p_{lj}^*	0
AB	0	$\left(1-p_{lj}^*\right)/2$	$1/2$	$p_{lj}^*/2$
BB	1	0	$1-p_{lj}^*$	p_{lj}^*

If no information about the unknown parent is available, one pragmatic solution is to assume that the probabilities of inherit a given allele are dictated by the unweighted average of allelic frequencies across subpopulations (for the full models). If $\bar{p}_j^*$ represents that average reference allele frequency for marker locus j then the conditional probabilities are same as in the previous table with p_{lj}^* replaced by $\bar{p}_j^*$. Of course, the lack of knowledge of the origin of the unknown parent is not an issue for null models.

Appendix D: Toy example of the joint pmf of two half sib founders

In this case the common parent is the relevant common ancestor. This individual is denoted with number 3. Suppose that individuals 1, 2 and 3 belong to population l. For simplicity we focus on a single marker, thus the subindex associated with marker is ignored. Then:

$$\pi(w_1, w_2|P^*) = \sum_{k\in\{-1,0,1\}} \pi(w_1|w_3 = k, p_l^*)\pi(w_2|w_3 = k, p_l^*)\pi(w_3 = k|p_l^*).$$

This summation is done for every one of the 9 combinations of genotypes of individuals 1 and 2. The following table displays the conditional probabilities $(w_1, w_2|w_3 = k, p_l^*)$.

Genotype of 3	**Genotype of 2**	**Genotype of 1**		
		AA	AB	BB
AA	AA	$(1-p_l^*)^2$	$p_l^*(1-p_l^*)$	0
	AB	$p_l^*(1-p_l^*)$	p_l^{*2}	0
	BB	0	0	0
AB	AA	$(1-p_l^*)^2/4$	$(1-p_l^*)/4$	$p_l^*(1-p_l^*)/4$
	AB	$(1-p_l^*)/4$	$1/4$	$p_l^*/4$
	BB	$p_l^*(1-p_l^*)/4$	$p_l^*/4$	$p_l^{*2}/4$
BB	AA	0	0	0
	AB	0	$(1-p_l^*)^2$	$p_l^*(1-p_l^*)$
	BB	0	$p_l^*(1-p_l^*)$	p_l^{*2}

The following table presents the joint pmf of individuals 1 and 2 conditional on allelic frequencies

Genotype of 2	**Genotype of 1**		
	AA	AB	BB
AA	$(1-p_l^*)^3\left(1-\frac{p_l^*}{2}\right)$	$(1-p_l^*)^2p_l^*\left(\frac{3}{2}-p_l^*\right)$	$\frac{\left(p_l^*(1-p_l^*)\right)^2}{2}$
AB	$(1-p_l^*)^2p_l^*\left(\frac{3}{2}-p_l^*\right)$	$p_l^*(1-p_l^*)\left(2p_l^*(1-p_l^*)+\frac{1}{2}\right)$	$p_l^{*2}(1-p_l^*)\left(p_l^*+\frac{1}{2}\right)$
BB	$\frac{\left(p_l^*(1-p_l^*)\right)^2}{2}$	$p_l^{*2}(1-p_l^*)\left(p_l^*+\frac{1}{2}\right)$	$p_l^{*3}\left(\frac{p_l^*+1}{2}\right)$