Accounting for randomness of genotypes in across and single population genome-enabled prediction: A hierarchical Bayes approach

Carlos A. Martínez[1,2], Kshitij Khare[2], Arunava Banerjee[3], and Mauricio A. Elzo[1]

[1]Department of Animal Sciences, [2]Department of Statistics,[3]Department of Computer and Information Science and Engineering
University of Florida, Gainesville, FL, USA

ICQG·5
5th International Conference on Quantitative Genetics
Madison, Wisconsin, USA

HIGHLIGHTS

- New Bayesian models for across and single population genome-wide prediction are developed.
- These models account for randomness of genotypes, heterogeneity and correlation of allelic frequencies (across populations case).
- Missing genotypes are allowed without the need for previous imputation.
- Bayes factors and fractional Bayes factors are approximated via Laplace's method.
- Some features of these models make them promising for genome-wide prediction.

SUMMARY

In this study, a family of models to perform single and across population genome-wide prediction modeling genotypes as random variables and allowing population-specific effects for each marker (for across population analysis) was developed using a hierarchical Bayes approach. Models differed in the priors used and assumed residual variances to be either homogeneous or heterogeneous. To account for randomness of genotypes, the joint probability mass function of marker genotypes conditional on allelic frequencies and pedigree information was derived. Thus, these models incorporated kinship and genotypic information that not only permitted to account for heterogeneity of allelic frequencies in across population studies, but also to include individuals with missing genotypes at some or all loci without the need for previous imputation. This was achieved by treating the non-observed genotypes as unknown model parameters. For the across population case, Bayes factors and fractional Bayes factors to compare models with their null versions (models ignoring population structure, but still accounting for randomness of genotypes) were derived via the Laplace approximation. Implementation of these models and computation of model comparison criteria were illustrated using simulated data. Theoretical and computational issues as well as possible extensions and refinements pose interesting problems for further research. Features of this set of models make them promising for genome-enabled prediction. Inclusion of information from the probability distribution of genotypes is perhaps the most attractive. Further research assessing the performance of this family of models and comparing them with conventional models used in genome-enabled prediction is needed.

INTRODUCTION

Some across population genome-enabled prediction studies use predictions obtained from individual populations or pool data to perform a single population analysis (de Roos et al., 2009). Recently, different methods allowing subpopulation-specific effects have been developed (Olson et al., 2012; de los Campos et al., 2015; Lehermeir et al., 2015). 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 . On the other hand, it may decrease accuracy when the effects of QTL controlling the trait are not the same across populations (van den Berg et al. 2015; Wientjes et al., 2015). The possible existence of genotype by environment interaction, lack of persistence of linkage phase and variation in allelic frequencies across populations suggest the need for a simultaneous analysis of these populations without ignoring the complete population structure.

A feature that has been overlooked in the random linear regression models used in genome-wide prediction is the randomness of genotypes. These are treated as fixed in genome-wide prediction models, while in classical quantitative genetics theory they are treated as random (Lynch and Walsh, 1998). In addition to be consistent with the classical theory, taking into account the randomness of genotypes permits the estimation of allelic frequencies because when treated as observable random variables, their joint probability mass function (pmf) depends on the allelic frequencies.

Thus, the objectives of this study were to propose hierarchical Bayesian models to carry out genome-wide prediction accounting for randomness of marker genotypes, and for heterogeneity and correlation of allelic frequencies and population-specific allelic substitution effects for across population analysis; and to derive approximate expressions for Bayes factors and fractional Bayes factors to compare across population genome-wide prediction models with their corresponding null versions ignoring population structure.

MODELS

The complete population is defined as the set of individuals with phenotypes considered in the study. Suppose that there exists some criterion (e.g., environment, breed, line, etc.) to split this population into $\mathcal{S}$ subpopulations. Hereinafter linkage equilibrium, Hardy-Weinberg equilibrium, known pedigree and no mutation are assumed.

$$\begin{bmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\\ \vdots\\ \boldsymbol{y}_S\end{bmatrix}=\begin{bmatrix}W_1 & 0 & \cdots & 0\\ & W_2 & \cdots & 0\\ & & \ddots & \vdots\\ sym & & & W_S\end{bmatrix}\begin{bmatrix}\boldsymbol{g}_1\\ \boldsymbol{g}_2\\ \vdots\\ \boldsymbol{g}_S\end{bmatrix}+\begin{bmatrix}\boldsymbol{e}_1\\ \boldsymbol{e}_2\\ \vdots\\ \boldsymbol{e}_S\end{bmatrix}$$

where $\boldsymbol{y}_l, \boldsymbol{g}_l$ and $\boldsymbol{e}_l$ are the vectors of phenotypic records, allelic substitution effects, and residuals for subpopulation l, and W_l is a (partially) observable random matrix defined as follows:

$$W_l=\{w_{ij}^l\}_{n_l\times m}=\begin{cases}1, if\ genotype=BB\\ 0, if\ genotype=AB\\ -1, if\ genotype=AA\end{cases},\ l=1,2,\dots,\mathcal{S}.\ \text{Let}\ W=Block\ Diag\{W_l\}_{l=1}^{\mathcal{S}}$$

$$\boldsymbol{y}|W,\boldsymbol{g},\sigma^2\sim MVN(W\boldsymbol{g},\sigma^2 I),\sigma^2\sim Inverse\ Gamma\left(\frac{\tau^2}{2},\frac{v}{2}\right):=IG\left(\frac{\tau^2}{2},\frac{v}{2}\right)$$

$$W|\boldsymbol{p}_1^*,\boldsymbol{p}_2^*,\dots,\boldsymbol{p}_m^*\sim\pi(\cdot|\boldsymbol{p}_1^*,\boldsymbol{p}_2^*,\dots,\boldsymbol{p}_m^*),\boldsymbol{p}_j^*\overset{iid}{\sim}\pi(\boldsymbol{p}^*),j=1,2,\dots,m$$

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.

Priors for P^* and $\boldsymbol{g}$ in the multi-population scenario

Let $P^*=(\boldsymbol{p}_1^*,\boldsymbol{p}_2^*,\dots,\boldsymbol{p}_m^*)$where $\boldsymbol{p}_j^*$ is the vector of allelic frequencies of the reference allele of the j^{th} marker in each subpopulation expressed on a subpopulation basis, that is, $p_{lj}^*+q_{lj}^*=1\ \forall l,j$ where q_{lj}^* is the corresponding frequency of the non-reference allele. To take into account the belief that allelic frequencies of the same marker vary across subpopulations and may be correlated, the prior is built based on a Dirichlet distribution. To do that, allelic frequencies are expressed on a complete population basis. With this parameterization $\sum_{l=1}^{\mathcal{S}}p_{lj}\le 1,\forall j=1,2,\dots,m$, with equality if and only if the reference allele is fixed in all subpopulations and $p_{lj}+q_{lj}=r_{lj}\in(0,1)$. Let $\boldsymbol{r}=(\boldsymbol{r}_1,\dots,\boldsymbol{r}_{\mathcal{S}}),\boldsymbol{r}_l=(r_{l1},\dots,r_{lm})$. The two parameterizations of allelic frequencies are related by the one to one mapping $p_{lj}^*=p_{lj}/r_{lj}$. The support of the probability density function (pdf) of $\boldsymbol{p}_j$ given $\boldsymbol{r}$ is $\Omega_j^r:=\{\boldsymbol{p}_j\in\mathbb{R}^{\mathcal{S}}|0<p_{lj}\le r_l\ \forall l,\sum_{l=1}^{\mathcal{S}}r_l=1\}$ and assuming $\boldsymbol{r}$ known:

$$\pi(\boldsymbol{p}_j|\boldsymbol{r})\propto\prod_{l=1}^{\mathcal{S}}\left\{\left(\frac{p_{lj}}{r_{lj}}\right)^{\alpha_l-1}\right\}p_{(\mathcal{S}+1)j}^{\alpha_{\mathcal{S}+1}-1},p_{(\mathcal{S}+1)j}=1-\sum_{l=1}^{\mathcal{S}}\frac{p_{lj}}{r_{lj}}.$$

This is the pdf of a scaled Dirichlet random vector. If $\boldsymbol{r}$ is unknown a $Dirichlet\left((\boldsymbol{\alpha}_p,\boldsymbol{\alpha}_q)\right)$ prior is posed over $P_j=(\boldsymbol{p}_j,\boldsymbol{q}_j)$, where $\boldsymbol{\alpha}_p=(\alpha_{1p},\dots,\alpha_{\mathcal{S}p}),\boldsymbol{\alpha}_q=(\alpha_{1q},\dots,\alpha_{\mathcal{S}q})$. Consequently, by properties of the Dirichlet distribution it follows that $\boldsymbol{r}_j\sim Dirichlet\left((\alpha_{1p}+\alpha_{1q},\dots,\alpha_{\mathcal{S}p}+\alpha_{\mathcal{S}q})\right)$.

Vector $\boldsymbol{g}$ is given a conditional multivariate Gaussian prior: $\boldsymbol{g}|G\sim MVN(0,G)$, or a conditional multivariate "spike and slab" prior: $\boldsymbol{g}_j|G_j,\pi_0\sim\begin{cases}Point\ mass\ at\ \mathbf{0}\ with\ prob.\ \pi_0\\ MVN(\mathbf{0},G_j)\ with\ prob.\ 1-\pi_0\end{cases}$.

$G=Block\ Diag\{G_j\}_{j=1}^{m}$. In both cases $G_j\overset{iid}{\sim}Inverse\ Wishart(a,\boldsymbol{\Sigma})$.

Joint pmf of marker genotypes conditional on allelic frequencies and pedigree

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

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

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

$W\in\mathcal{G}$. The set $\mathcal{G}$, that is, the support of $\pi(W|P^*)$ and its cardinality are derived using basic segregation rules, expressions are presented in submitted work by Martínez et al. (2016).

Likelihood

Suppose that in each subpopulation there is a fraction of genotyped individuals and a fraction of non-genotyped or partially genotyped individuals. Let W^o and W^N denote the observed (data) and non-observed (an unknown parameter) parts of W. Therefore, $\pi(W|P^*)=\pi(W^o,W^N|P^*)$ can be expressed as: $f(W^o|W^N,P^*)\pi(W^N|P^*)$. Thus, the full likelihood has the form:

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

Full conditionals

Only non-standard full conditionals are shown (homoscedastic residuals, $\boldsymbol{r}$ known).

$$\pi(P|Else)\propto\prod_{j=1}^{m}p_{(\mathcal{S}+1)j}^{\alpha_{\mathcal{S}+1}-1}\prod_{l=1}^{\mathcal{S}}\left\{p_{lj}^{n_l^{B_j}+\alpha_l-1}(r_l-p_{lj})^{n_l^{A_j}}\right\},\boldsymbol{p}_j\in\Omega_j^r\ \forall\ j=1,2,\dots,m$$

$$\pi(W^N|Else)\propto\pi^+(W|P^*)\exp\left(\frac{-1}{2\sigma^2}(-2\boldsymbol{g}'W^{N\prime}\boldsymbol{y}^N+\boldsymbol{g}'W^{N\prime}W^N\boldsymbol{g})\right)\prod_{k=1}^{K}\exp\left(\frac{-1}{2\sigma^2}h(W^{M_k},\boldsymbol{g}^{M_k},\boldsymbol{y}^{M_k})\right).$$

Subindex k refers to individuals with missing genotypes in the k^{th} subset of markers. The functional form of $h(W^{M_k},\boldsymbol{g}^{M_k},\boldsymbol{y}^{M_k})$ can be found in Martínez et al. (2016, submitted).

$\pi^+(W|P^*)=f^+(W^o|W^N,P^*)\pi(W^N|P^*)$, $f^+(W^o|W^N,P^*)$ is the part of $f(W^o|W^N,P^*)$ depending on W^N.

Theoretical approximation to model comparison via Bayes factors and fractional Bayes factors

The Bayes factor comparing two models denoted as M_1 and M_0 is:

$$BF_{10}=\frac{\int_{\Theta_1}\pi_1(\boldsymbol{\theta}_1)f_1(\boldsymbol{y}|\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1}{\int_{\Theta_0}\pi_0(\boldsymbol{\theta}_0)f_0(\boldsymbol{y}|\boldsymbol{\theta}_0)d\boldsymbol{\theta}_0},\ \text{here, conditional on}\ W,BF_{10W}=\frac{f(\boldsymbol{y}|W,M_1)}{f(\boldsymbol{y}|W_0,M_0)}$$

$$f(\boldsymbol{y}|W,M_1)=\int_{\mathcal{P}^+}\pi(G)\left(\int_{\mathbb{R}^{m\mathcal{S}}}\int_{\mathbb{R}_+}f(\boldsymbol{y}|\boldsymbol{g},\sigma^2,W)\pi(\boldsymbol{g}|G)\pi(\sigma^2)\,d\sigma^2 d\boldsymbol{g}\right)dG$$

No closed form. Analytic alternative: use the Laplace approximation

For each proposed model, the null model corresponds to a model ignoring structure, i.e., the complete population is analyzed as a single one. The model is $\boldsymbol{y}=W_0\boldsymbol{g}_0+\boldsymbol{\varepsilon}$, $W_0=(W_1'\vdots\cdots\vdots W_{\mathcal{S}}')'$, the distributional assumptions for $\boldsymbol{g}_0$ and W_0 are simplified versions of those for $\boldsymbol{g}$ and W.

Fractional Bayes factor O'Hagan (1994; 1995):

$$FBF_{10}=BF_{10}\frac{\int_{\Theta_0}\pi_0(\boldsymbol{\theta}_0)(f_0(\boldsymbol{y}|\boldsymbol{\theta}_0))^c d\boldsymbol{\theta}_0}{\int_{\Theta_1}\pi_1(\boldsymbol{\theta}_1)(f_1(\boldsymbol{y}|\boldsymbol{\theta}_1))^c d\boldsymbol{\theta}_1},c\in(0,1)$$

Unfortunately, using the Laplace approximation requires $n_l\ge m\ \forall\ l$, which currently is met by few datasets , examples can be found in the Holstein population of the US for certain SNP chips (CDCB, 2016). This limits the application of our approximate BF and FBF (expressions not shown). For several models, the following approximation to the FBF can be used to compare them with their null versions: $c^{m(\mathcal{S}-1)/2}\left(\frac{SSR_1}{SSR_0}\right)^{\frac{n(c-1)}{2}}\frac{SSR_1^{(m\mathcal{S}+2)/2}}{SSR_0^{(m+2)/2}},SSR_i=\boldsymbol{y}'(I-H_i)\boldsymbol{y}$, where H_i is the projection matrix onto the column space of the design matrix of model $i,i=0,1$. Notice that for the $m>n$ case, SSR_i is invariant to the choice of the generalized inverse of $W_i'W_i$ and therefore this FBF could potentially be used for model comparison. However, formally proving or disproving this statement remains an open problem. Some technical details are provided in submitted work by Martínez et al. (2016).

SIMULATION STUDY

Two phenotypes were simulated with different number of QTL controlling the trait and different heritabilities. Briefly, dataset 1 involved three subpopulations with $n_l\ge m\ \forall\ l$, no missing genotypes and different number of generations, migration was allowed and the heritability of the trait was high. Dataset 2 was comprised by two subpopulations with $m>n$, missing genotypes, only two generations, no migration, and low heritability. Mating designs and selection schemes also varied from one dataset to the other. Consequently, in dataset 1, subpopulations diverged more than in dataset 2.

Parameter inference via MCMC

No missing genotypes imply $W^N=\emptyset$ and then, posterior sampling for the parameters of the W component of the likelihood and the (hyper) parameters of the $\boldsymbol{y}$ component can be performed separately: Gibbs sampler for the $\boldsymbol{y}$ component and independent Metropolis-Hastings for the W component. In the presence of missing genotypes: Gibbs sampler with two Metropolis steps, one to sample from $\pi(P|Else)$ and the other to sample from $\pi(W^N|Else)$.

Table 1. Predictive abilities and accuracies in datasets 1 and 2

Model	Predictive Ability		Accuracy in testing population		Accuracy in Training population	
	Dataset 1	Dataset 2	Dataset 1	Dataset 2	Dataset 1	Dataset 2
M_{1G}	0.29	0.019	0.27	0.04	0.32	0.17
M_{1GH}	0.76	0.016	0.83	0.03	0.94	0.21
$M_{1SSH0.5}$	0.81	0.017	0.88	0.04	0.92	0.19
$M_{1SSH0.9}$	0.81	0.018	0.88	0.04	0.90	0.14
$M_{1SSH0.2}$	0.79	0.016	0.86	0.03	0.94	0.20
M_{0G}	0.53	0.004	0.50	0.07	0.55	0.24
M_{0GH}	0.83	0.013	0.88	0.05	0.88	0.23
$M_{0SSH0.5}$	0.72	0.003	0.77	0.06	0.86	0.24
$M_{0SSH0.9}$	0.69	0.008	0.76	0.05	0.85	0.20
$M_{0SSH0.2}$	0.72	0.009	0.79	0.05	0.79	0.24

Deviance information criterion (DIC) was also used to compare models. In dataset 1 DIC values provided evidence in favor of full models (smaller values) while in dataset two DIC values were slightly smaller for null models. Bayes factors, which could be computed only for dataset 1 agreed with DIC, that is, favored full models.

CLOSING REMARKS

The main contribution of this study is the theoretical development of a set of models for 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. Extensions and refinements (e.g., account for LD and mutation) pose interesting problems for further research.

REFERENCES

CDB, Council on Dairy Cattle Breeding. 2016. Retrieved January, 10 from https://www.cdcb.us/Genotype/counts.html.
de Roos, A.P.W., Hayes, B.J., Goddard, M.E. 2009. *Genetics* 183:1545-1553.
de los Campos G., Venturi Y, Vázquez, A.I. et al. 2015. *J Agric Biol Environ Stat* 20: 467-490.
Lehermeier C., Schon C, de los Campos G. 2015. *Genetics* 201: 323-337.
Lynch, M., Walsh, E. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer associates Inc.
van den Berg, S., Calus, M.P.L., Meuwissen, T.H.E. et al. 2015. *BMC Genet* 16:416.
Olson K.M., VanRaden P.M., Tooker M.E. 2012. *J Dairy Sci* 95: 5378-5383.
Wientjes, Y.C.J., Veerkamp, R.F., Bijma, P. et al. 2015. *Genet Select Evol* 47:5.