Álvarez-Múnera, C., Bermann, M., Legarra, A., Misztal, I., Lourenco, D. (2026) Bayesian estimation of the relationships across metafounders or base populations, *Proceedings 13th World Congress on Genetics Applied to Livestock Production*: 24094 https://doi.org/10.31274/wcgalp.24094
©2026 Álvarez-Múnera, *et al.* CC BY-NC-ND

Bayesian estimation of the relationships across metafounders or base populations

C. Álvarez-Múnera[1*], M. Bermann[1], A. Legarra[1,2], I. Misztal[1], and D. Lourenco[1]

[1]*University of Georgia, Athens, GA, USA;* [2]*Council on Dairy Cattle Breeding, Bowie, MD, USA;* [*]*Corresponding author: cristina.alv.mun@gmail.com*

Introduction

The metafounder relationship matrix (Γ) allows considering relationships across base populations and unknown parent groups, enabling the consistent integration of pedigree and genomic information in single-step genomic BLUP (Aguilar et al., 2010; Legarra et al., 2015). In the metafounder framework, Γ summarizes relationships within and across base populations and can be interpreted as a function of base population allele frequencies and fixation indices (Garcia-Baccino et al., 2017). Consequently, inaccurate estimation of Γ directly propagates inconsistency and bias into genomic evaluations.

Because allele frequencies in base populations are not observed, Γ must be inferred from genotyped individuals that are often far from the pedigree base population. Existing estimators, such as generalized least squares (Garcia-Baccino et al., 2017) or pseudo–maximum likelihood, are biased by construction and cannot estimate the uncertainty in the estimator of Γ. Thus, proper inference for the relationships across base populations is not possible.

Bayesian methods offer a natural alternative by explicitly modeling uncertainty and enabling posterior inference. The full posterior distribution of Γ given the observed genotypes does not have a closed form. Therefore, implementing a Gibbs sampler is not possible. In this study, we aimed to develop two Metropolis-Hastings samplers, one independent and one adaptive, to estimate Γ through posterior sampling of base population allele frequencies.

Materials & Methods

Bayesian estimation of Γ was performed by sampling base-population allele frequencies, modelled as independent Beta distributions (Paris et al., 2019). Let $\Gamma = \frac{2}{m}(2P-1)(2P-1)'$, where Γ is of dimension n by n populations, m is the number of markers, and P is an n by m matrix where p_{ij} is the allele frequency for the j^{th} marker on the i^{th} population. The distribution of genotypes in the current population (Z) is (Legarra et al. 2024):

$$p(P) \propto \left(0.5A_{\Gamma 22}\right)^{-\frac{m}{2}} exp\, exp\, \left(-tr\left(A_{\Gamma_{22}}^{-1} ZZ'\right)\right) \quad (1)$$

where $A_{\Gamma 22}$ is the pedigree relationship matrix for the genotyped individuals given an estimate of Γ. The stationary distribution for a Metropolis-Hastings sampler is defined as:

$$\pi(P) \propto p(P) \times p(P) \quad (2)$$

Since P is a matrix that contains allele frequencies in the base populations, and assuming independence between them, $p(P)$ can be defined as a product of independent beta distributions. Thus, Eq. (2) results in:

$$\pi(P) \propto \left(0.5A_{\Gamma 22}\right)^{-\frac{m}{2}} exp\, exp\, \left(-tr\left(A_{\Gamma_{22}}^{-1} ZZ'\right)\right) \times \prod_{i=1}^{n}\prod_{j=1}^{m} \frac{p_{ij}^{\alpha_{ij}-1}\left(1-p_{ij}\right)^{\beta_{ij}-1}}{B\left(\alpha_{ij},\beta_{ij}\right)} \quad (3)$$

where α_{ij} and β_{ij} are the shape parameters for each beta distribution $B(\alpha_{ij}, \beta_{ij})$. The shape parameters were derived from prior estimates of Γ obtained with the pseudo-maximum likelihood method of Legarra et al. (2024). Two Metropolis-Hastings sampling strategies, independent and adaptive, were implemented.

Independent sampler. The independent sampler draws proposals directly from the fixed Beta priors, which can yield accurate inference when prior information is strong but often results in low acceptance rates and poor mixing. After obtaining the shape parameters for each $B(\alpha_{ij}, \beta_{ij})$, the sampler at the k^{th} iteration is:

1. Sample P^k from $B(\alpha_{ij}, \beta_{ij})$ and calculate Γ^k from P^k.
2. Calculate $log\ log\ L(Z) \propto -\frac{m}{2} log\ log \left|0.5A_{\Gamma^k_{22}}\right| - tr\left(A^{-1}_{\Gamma^k_{22}} ZZ'\right)$.
3. Retrieve $log\ log\ L(P|Z)$ from the previous accepted sample.
4. Sample $u \sim U(0,1)$.
5. Accept P^k if $log\ log\ (u) < log\ log\ L(Z) - log\ log\ L(Z)$.

Adaptive sampler. The adaptive sampler updates the proposal parameters based on previously accepted values, effectively learning the local structure of the posterior distribution and improving exploration of the parameter space. The updating rate of the shape parameters should be chosen beforehand. Alternatives could be updating every accepted sample, every 10 accepted samples, etc. After obtaining the shape parameters for each $B(\alpha_{ij}, \beta_{ij})$, the sampler at the k^{th} iteration is:

1. Sample P^k from $B(\alpha_{ij}, \beta_{ij})$ and calculate Γ^k from P^k.
2. Calculate $log\ log\ L(Z) \propto -\frac{m}{2} log\ log \left|0.5A_{\Gamma^k_{22}}\right| - tr\left(A^{-1}_{\Gamma^k_{22}} ZZ'\right)$.
3. Retrieve $log\ log\ L(P|Z)$ from the previous accepted sample.
4. Calculate $log\ log\ Q(P) = \sum_{i,j} log\ log\ f\left(\hat{p}^k_{ij}; \alpha_{ij}, \beta_{ij}\right)$
   and $log\ log\ Q\left(P^k\right) = \sum_{i,j} log\ log\ f\left(\hat{p}_{ij}; \alpha^k_{ij}, \beta^k_{ij}\right)$.
5. Calculate the log of the acceptance probability as $log\ log\ (\psi) = log\ log\ L(Z) - log\ log\ L(Z) + log\ log\ Q\left(P^k\right) - log\ log\ Q(P)$.
6. Sample $u \sim U(0,1)$.
7. Accept P^k if $log\ log\ (u) < log\ log\ (\psi)$.
8. Update $B(\alpha_{ij}, \beta_{ij})$ if the rule for updating the shape parameters is met.

Numerical test. The performance of both samplers was evaluated following the simulation of Legarra et al. (2024). We simulated a population of 84,202 individuals, assuming two metafounders with known values of Γ. We replicated the simulation 100 times, and, without loss of generality, we used the last replicate to test our proposed samplers. Initial allele frequencies for calculating the shape parameters for each $B(\alpha_{ij}, \beta_{ij})$ were obtained from the pseudo-maximum likelihood method of Legarra et al. (2024). The independent and adaptive Metropolis-Hastings samplers were run for 10,000 samples. A burn-in of 500 was implemented, which was considered sufficient given that the simulations were initialized using known values of the Γ parameters. Additionally, one out of every ten realizations was retained for subsequent analysis. For the adaptive Metropolis Hastings, we updated the shape parameters of each $B(\alpha_{ij}, \beta_{ij})$ every two accepted samples.

Results & Discussion

Descriptive statistics for the simulated and inferred distributions of the elements of Γ are presented in Table 1-3 for the simulated, independent sampler, and adaptive sampler, respectively. The running time for both samplers was 5.5 hours. The independent sampler had an acceptance rate of 0.18%, whereas the adaptive sampler had an acceptance rate of 21.4%.

The pseudo-maximum likelihood estimates of Γ were $\Gamma_{11} = 0.756$, $\Gamma_{12} = 0.645$, and $\Gamma_{11} = 0.756$. Posterior means and medians obtained with the independent sampler were closely aligned with the simulation-based reference values for all elements of Γ (Table 2), indicating accurate recovery of the underlying metafounder relationships. However, this agreement was achieved at the cost of very low effective sample sizes, reflecting inefficient exploration of the posterior distribution. Also, the posterior standard deviation was one order of magnitude smaller than the simulated value. This indicates that the chain is extremely concentrated over the prior. In contrast, the adaptive sampler showed consistency with the simulation benchmark and slightly higher effective sample size (ESS) values. Given the low ESS for both methods, a larger sample size is likely needed for the numerical example presented.

Table 1. Descriptive statistics for the simulated values of Γ.

Element	Mean	Median	SD	Min	Max
Γ_{11}	0.757	0.757	0.0024	0.752	0.764
Γ_{12}	0.648	0.647	0.0021	0.642	0.652
Γ_{22}	0.757	0.758	0.0022	0.752	0.763

SD: standard deviation, Min: minimum, Max: maximum.

Table 2. Posterior chain analysis of Γ obtained with the independent Metropolis-Hastings sampler

Element	Mean	MCE	Median	PSD	ESS
Γ_{11}	0.756	0.21E-03	0.756	0.63E-03	8.4
Γ_{12}	0.646	0.24E-03	0.645	0.59E-03	6.0
Γ_{22}	0.756	0.17E-03	0.756	0.58E-03	10.7

MCE: Monte Carlo error, PSD: posterior standard deviation, ESS: effective sampling size.

Table 3. Posterior chain analysis of Γ obtained with the adaptive Metropolis Hastings sampler

Element	Mean	MCE	Median	PSD	ESS
Γ_{11}	0.757	0.96E-03	0.756	0.0029	14.0
Γ_{12}	0.646	0.16E-02	0.646	0.0019	10.9
Γ_{22}	0.756	0.11E-02	0.755	0.0021	17.0

MCE: Monte Carlo error, PSD: posterior standard deviation, ESS: effective sampling size.

This study is the first to implement Bayesian methods for making inferences on Γ. Therefore, many methodological details are still subject to research. The choice of Bayesian inference algorithm might be the most important research question. The log posterior distribution in Eq. (3) may have a highly correlated structure, suggesting that Hamiltonian Monte Carlo (HMC) would be a better sampling method than Metropolis-Hastings. HMC was not implemented here due to its complexity. The prior distribution of allele frequencies is also a topic that requires attention. Here, we chose the beta distribution because it is bounded between zero and one and allows for it to be skewed. Nonetheless, a truncated normal distribution or a transformed normal distribution (e.g., via a log or logit transformation to remain in the parameter space) may be viable alternatives to explore.

Fine-tuning is crucial for Monte Carlo chain methods. Here, we draw a relatively low number of samples due to time constraints. Some of the required fine-tuning tasks for implementing our methods in real-life applications include exploring different chain lengths, sample sizes, and the timing of parameter updates in adaptive sampling.

Overall, our results demonstrate that the Bayesian estimation of Γ yields coherent inference when benchmarked against simulation-based estimates. The independent sampler yields an extremely low acceptance rate and should be used only as a reference for future methodological developments. The adaptive sampler may be impractical for estimating Γ in routine large-scale evaluations due to its high running time. However, it provides a valuable framework for assessing the uncertainty of Γ and for validating and comparing alternative estimators of Γ.

Conclusions

Bayesian estimation of metafounder relationship matrices using the Metropolis-Hastings algorithm provides a coherent framework for inferring Γ and quantifying the uncertainty of its elements. The adaptive Metropolis Hastings sampler has a good acceptance rate and may be used in research and genetic evaluations to make inferences on the relationship across metafounders or base populations.

References

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., and Tsuruta, S. (2010). J Dairy Sci. 93:743-752.
Garcia-Baccino C.A., Legarra A., Christensen O.F., and Vitezica Z.G. (2017). Genet Sel Evol. 49:34.
Legarra A., Christensen O.F., Vitezica Z.G., Aguilar I., and Misztal, I. (2015). Genetics. 200:455–468.
Legarra A., Bermann M., Mei Q., and Christensen, O.F. (2024). Genet Sel Evol. 56:35.
Paris C., Servin B., Boitard S. (2019). G3. 9: 4073–4086.