GWAS by GBLUP: Single and Multimarker EMMAX and Bayes Factors, with an Example in Detection of a Major Gene for Horse Gait

Andres Legarra,*1 Anne Ricard,†‡ and Luis Varona§**

*INRA (Institut National de la Recherche Agronomique), UMR 1388 GenPhySE, F-31326 Castanet-Tolosan, France, †INRA (Institut National de la Recherche Agronomique), UMR 1313 GABI, 78352 Jouy-en-Josas, France, ‡IFCE (Institut Français du Cheval et de l’Equitation), Recherche et Innovation, 61310 Exmes, France, §Departamento de Anatomía, Embriología y Genética, Universidad de Zaragoza, 50013 Zaragoza, Spain, and **Instituto Agroalimentario de Aragón (IA2), 50013 Zaragoza, Spain

ORCID ID: 0000-0001-8893-7620 (A.L.)

ABSTRACT Bayesian models for genomic prediction and association mapping are being increasingly used in genetics analysis of quantitative traits. Given a point estimate of variance components, the popular methods SNP-BLUP and GBLUP result in joint estimates of the effect of all markers on the analyzed trait; single and multiple marker frequentist tests (EMMAX) can be constructed from these estimates. Indeed, BLUP methods can be seen simultaneously as Bayesian or frequentist methods. So far there is no formal method to produce Bayesian statistics from GBLUP. Here we show that the Bayes Factor, a commonly admitted statistical procedure, can be computed as the ratio of two normal densities: the first, of the estimate of the marker effect over its posterior standard deviation; the second of the null hypothesis (a value of 0 over the prior standard deviation). We extend the BF to pool evidence from several markers and of several traits. A real data set that we analyze, with ours and existing methods, analyzes 630 horses genotyped for 41711 polymorphic SNPs for the trait “outcome of the qualification test” (which addresses gait, or ambling, of horses) for which a known major gene exists. In the horse data, single marker EMMAX shows a significant effect at the right place at Bonferroni level. The BF points to the same location although with low numerical values. The strength of evidence combining information from several consecutive markers increases using the BF and decreases using EMMAX, which comes from a fundamental difference in the Bayesian and frequentist schools of hypothesis testing. We conclude that our BF method complements frequentist EMMAX analyses because it provides a better pooling of evidence across markers, although its use for primary detection is unclear due to the lack of defined rejection thresholds.

KEYWORDS association analysis single marker regression QTL GWAS Bayesian regression GenPred Shared Data Resources Genomic Selection

Bayesian models including simultaneously all marker effects are becoming very popular for GWAS analysis (Habier et al. 2011; Wang et al. 2012, 2016; Moser et al. 2015). The most frequently used prior for marker effects is the normal distribution, known as RRBLUP or SNP-BLUP (Habier et al. 2007; VanRaden 2008), which is equivalent to GBLUP (VanRaden 2008), also known in the human literature as GCTA analysis (Yang et al. 2010). GBLUP is simple and can be generalized to marker data missing in a large fraction of individuals in the so-called Single Step methods (Aguilar et al. 2010; Christensen 2012), and also to multiple traits, or complex models (random regression, genotype by environment, etc.). Because of the equivalence of GBLUP and SNP-BLUP, it is straightforward to obtain from Single Step methods estimates of marker effects for complex traits like, e.g., multiple trait maternal effects (Lourencio et al. 2015) or genotype-environment models (Jarquin et al. 2014).

Therefore, GWAS can be done exploiting results of GBLUP (Wang et al. 2012; Dikmen et al. 2013; Castró et al. 2017). Most of these works
The methods use the prior (before observing the data) and posterior (after observing the data) distributions of marker effects assuming a priori multivariate normality (i.e., SNP-BLUP or GBLUP). Most theory can be found in (Guadlón Duarte et al. 2014; Bernal Rubio et al. 2016; Chen et al. 2017) and we include it in the Appendix for completeness. We will assume throughout that variance components are known; this is a frequent assumption that allows obtaining of closed-form results. The main notation that we need is a vector of marker effects \(a\) normally distributed with a priori mean \(0\) and variance \(\Sigma\), and their prediction error variance \(\Lambda\).

**EMMAX tests of association from GBLUP**

This section is a reminder of (Guadlón Duarte et al. 2014; Bernal Rubio et al. 2016; Chen et al. 2017) and we include it here for completeness.

**Single marker:** The single marker EMMAX procedure is a normal test obtained, in our notation, with the statistic

\[
t = \frac{\hat{a}}{sd(\hat{a})}
\]

where \(\hat{a}\) is the marker estimate of the locus under consideration, obtained from a single SNP-BLUP evaluation (or an equivalent model), and where \(sd(\hat{a})\) is the frequentist distribution (over conceptual repeated sampling of \(y\)) of the SNP-BLUP estimator of the effect \(a\). Some what surprisingly, the numerical value of \(t\) is the same as if \(a\) was fit as a “fixed regression” GWAS and therefore the distribution of \(t\) under the null is \(N(0, 1)\) (Bernal Rubio et al. 2016). For instance, assume that \(\sigma^2 = 0.2\) is the a priori variance of marker effects. Output of the SNP-BLUP gives an estimate of the marker effect \(\hat{a}\) with a standard deviation of the posterior distribution \(z_d.(a|y) = 0.05\). With these numbers, the frequentist \(Var(\hat{a}) = \sigma^2 - Var(a|y) = 0.2 - (0.05)^2 = 0.1975\). Thus, \(t = \frac{0.5}{0.1975} = 2.48\) which yields a p-value of 0.006.

**Multiple marker:** Consider a subset of \(n\) markers (possibly consecutive), starting at marker \(i\). The statistic is a quadratic form \(x = \hat{a}_i\\Sigma_{[i,i+n]}^{-1}\hat{a}_i\), where \(\Sigma = Var(\hat{a}) = \text{Var}_{\text{aa}} - C_{\text{aa}}\text{C}_{\text{dd}}\text{C}_{\text{aa}}\) is the frequentist covariance of these marker effects. Chen et al. (2017) proved that under multivariate normality the quadratic form \(x\) follows a chi-square distribution of \(n\) degrees of freedom, which yields p-values for the multiple marker EMMAX. Alternatively, derivation of the Hotelling-t squared test, that tests whether a set of correlated sample means are simultaneously different from zero, yields the same result. The previous normal test for the single marker EMMAX is also equivalent to the chi-squared test. Matrix \(\Sigma\) takes into account uncertainty and collinearity of marker estimates.

For instance, consider two markers with effects \(\hat{a} = (0.5, 0.4)\) (similar effects) with \(\text{C}_{\text{aa}} = \begin{pmatrix} 0.05 & -0.02 \\ -0.02 & 0.08 \end{pmatrix}\) (estimates of effects are negatively correlated because of linkage disequilibrium) and \(\sigma^2 = 0.2\). The quadratic form has value \(x = 2.61\) with p-value 0.27. The evidence given by the p-value lowers because the two effects are correlated.

**Bayes Factors from GBLUP**

In this section we include our original derivations.

**Single marker:** There are two competing models in the BF: that the marker \(i\) with effect \(a_i\) “has some effect” (H1: \(a_i \neq 0\)) or has 0 effect (H0: \(a_i = 0\)), and the BF can be written as

\[
BF = \frac{P_{H1}(y)}{P_{H0}(y)}
\]

The BF measures whether the data \(y\) is more probable under either of the hypothesis. This can be written, alternatively, as

\[
BF = \frac{P(y|a_i \neq 0)}{P(y|a_i = 0)}
\] (1)

where \(a_i\) is the effect of the marker. Typically, this involves a complex MCMC integration. In the particular case of multivariate normality with known variances, Varona et al. (2001), Varona (2010) showed that the expression (1) is equal to

**Distributions of marker effect estimates**

The methods use the prior (before observing the data) and posterior (estimates and associated errors) distributions of marker effects assuming a priori multivariate normality (i.e., SNP-BLUP or GBLUP). Most theory can be found in (Guadlón Duarte et al. 2014; Bernal Rubio et al. 2016; Chen et al. 2017) and we include it in the Appendix for completion. We will assume throughout that variance components are known; this is a frequent assumption that allows obtaining of closed-form estimators. In particular, variance components can be estimated beforehand (e.g., by REML), or (making strong assumptions) they can be borrowed from pedigree analysis. In either case, a point estimate is used “as if” it was exact, which results in optimistic results. The main notation that we need is a vector of marker effects \(a\) normally distributed with a priori mean \(0\) and variance \(\Sigma\), and their prediction error variance \(\Lambda\).
\[ BF = \frac{p(a_i = 0)}{p(a_i = 0) \mid y} \] (2)

where \( p(a_i = 0) \) is the density of \( a_i \) _a priori_ evaluated at \( a_i = 0 \), and \( p(a_i = 0) \mid y \) is the density of \( a_i \) _a posteriori_ evaluated at \( a_i = 0 \). Computation of BF using (2) is straightforward because \( p(.) \) is a normal density. In particular, \( p(a_i = 0) \mid y \) is the density of \( a_i = 0 \) knowing that there is an estimate \( a_i \) with a certain _a posteriori_ variance \( \text{Var}(a_i \mid y) \) (e.g., different for each data set). In algebraic form this is

\[ BF = \frac{N(0,0, \sigma^2_a)}{N(0,0, \text{Var}(a_i))} \] (3)

where \( N(x,y,z) \) is the density of \( x \) in the normal distribution with mean \( y \) and variance \( z \). Consider the same example as before: \( \sigma^2_a = 0.2, a_i = 0.5, \text{s.d.}(a_i \mid y) = 0.05 \). The BF is thus, in R code:

\[
\text{dnorm}(0,0,\text{sqrt}(0.2))/\text{dnorm}(0,0.5,0.05)
\]

which is 20.76 in the log10 scale. According to Kass & Raftery (1995) this is a “Very Strong” evidence.

**Multiple marker:** Evidence from several consecutive markers in a segment can be pooled together using the BF. Expression (3) is generalized to several SNP markers (markers from \( i \) to \( n \)) as:

\[ BF = \frac{\text{MVN}(0,0,I_{n^2})}{\text{MVN}(0,0,\text{Var}_{a_1:a_n}^{\text{aa}})} \] (4)

where \( \text{MVN} \) is the density of a multivariate normal distribution and \( \text{Var}_{a_1:a_n}^{\text{aa}} \) is the posterior (co)variance matrix between the marker estimates. Posterior covariance matrix \( \text{Var}_{a_1:a_n}^{\text{aa}} \), which is a submatrix of \( \text{C}^{\text{aa}} \), takes into account colinearity between markers caused by LD. In this case, the BF tests whether a set of markers are all simultaneously 0, against the alternative that some of them (if not all) are different from zero.

Consider the same example as before: two markers with effects \( a = (0.5, 0.4) \), \( C^{\text{aa}} = \begin{pmatrix} 0.05 & -0.02 \\ -0.02 & 0.08 \end{pmatrix} \), \( \sigma^2_a = 0.2 \). The BF can be computed in R as:

\[
\text{dmvnorm}(c(0,0), \text{mean} = c(0,0), \text{sigma} = \text{diag}(0.2,2))/\text{dmvnorm}(\text{ahat}, \text{mean} = c(0,0), \text{sigma} = \text{Caa})
\]

yielding a BF of 1.65 in the log10 scale, lower than the single marker analysis. In a way, this reflects that there is a confusion of marker effects.

**Multiple trait:** Above methods can be easily extended to the multiple trait case. Multiple trait genomic predictions can be done from Bayesian regressions, SNP-BLUP or (Single Step) GBLUP (Tsaurt et al. 2011; Jia and Jannink 2012; Maier et al. 2015). Then, the EMMAX tests or the BF for several traits (and possibly markers) simultaneously is very similar to the “Several markers” case considering joint estimates of marker effects \( a \) for the \( n \) traits, the _a priori_ covariance among marker effects for the \( n \) traits \( K_0 \), and the _a posteriori_ covariance matrix of marker effect estimates \( C^{\text{aa}} \). Vector \( a \) can include either one or several markers. Typically \( K_0 \) is a function of \( G_0 \), the genetic covariance among traits.

**Data:** We used a horse real data set to explore and illustrate the properties of the procedures. We also did a limited number of simulations but we chose not to present them as this was extensively done in (Chen et al. 2017).

A single base polymorphism at the gene DMRT3 in chromosome 23 has a strong effect on horse ambling gait (Andersson et al. 2012). In French trotters, a SNP marker (marker BIEC2-620109 on chromosome 23 at position 22967656 bp) in strong disequilibrium with this polymorphism has a strong effect in qualification at the race (Ricard 2015; Brard and Ricard 2015). In this work we reanalyzed the same data set, which contains 630 horses and 41711 polymorphic SNP markers. The trait was “outcome of the qualification test”, with a heritability of 0.56. The major gene was not discovered in this data set, and therefore there is no bias due to discovery. We tried the following methods for GWAS:

Bayes factors with the mixture model BayesCpi (Habier et al. 2011) fixing _a priori_ that only 0.1% of the markers have an effect (see (Legarra et al. 2015) for a full description). This method provides BF’s, although our implementation only considers single markers.

Single marker and multiple marker EMMAX tests: as presented in this work, computed via MCMC, up to segments of 100 consecutive markers.

Bayes factors from SNP-BLUP: as presented in this work, computed via MCMC, up to segments of 100 consecutive markers.

EMMAX was fitted using blup90 (Misztal et al. 2002) and home-made scripts, whereas the other used our software GS3 (available at https://github.com/alegarra/gs3), using “OPTION Bayes Factor”. After completion of the analysis, we produced Manhattan plots based on BF and the other statistics; for EMMAX we used Bonferroni corrections to claim genewide significance; for Bayesian procedures, we did not address thresholds for declaring detection; this point will be addressed in the discussion.

### Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article. Supplemental material available at Figshare: https://doi.org/10.25387/g3.6241928.

### RESULTS

Figure 1 shows results from the single marker association test (EMMAX), the Bayesian multi-marker mixture model GWAS (BayesCpi) and the single marker BF. All three methods point to the SNP (BIEC2-620109 at position 22967656 bp) closest and most associated to the causal gene, and the four significant markers in the EMMAX single marker regression are in LD with each other. This reproduces the results in Ricard (2015). The EMMAX yields significant p-values at the Bonferroni level. Concerning BF, a threshold of 150 (2.17 in the log10 scale) has been suggested (Legarra et al. 2015), and the BayesCpi analysis in Figure 2 does reach this threshold, but this is not the case in the single marker BF using GBLUP.

In both analyses (BayesCpi and single marker BF), a large number of markers fall below the threshold of 0 in log10(BF), in other words, BF < 1. This means that for those regions the hypothesis that these markers have an effect is _less_ likely that the hypothesis that they do _not_ have an effect.

Figure 2 shows that evidence of the causal gene increases when using BF across consecutive markers. Systematically, the same location (BIEC2-620109) is spotted. It can be seen that the strength of evidence increase dramatically with increasing consecutive numbers, reaching the suggested ’suggestive’ threshold of BF > 3 (Kass and Raftery 1995), but not the much higher threshold of 150 suggested (Legarra et al. 2015). On the other hand, evidence from EMMAX does actually decrease, becomes non-significant, and, moreover, the highest peak deviates from the true location. This is at first sight a rather surprising result that will be discussed later.
DISCUSSION

The standard test for GWAS by association analysis is the single marker association analysis (e.g., Kruglyak 1999). Association analysis can account for genetic relationships (Kennedy et al. 1992), population structure (Kang et al. 2010) and also to a part of individuals not genotyped (Legarra and Vitezica 2015). An alternative is to fit multiple marker simultaneously in the form of Bayesian regression (e.g., Fernando and Garrick 2013). Legarra et al. (2015) did not see qualitative differences of Bayesian regressions and association analysis over five data sets and species, and concluded that the interest of Bayesian procedures is to complement regular association analysis. Anyway, Bayesian regression is of interest for three reasons: first, the Bayesian analysis has interesting properties of automatically accounting for multiple test, structure, unbiasedness, false discovery rate and power (Wakefield 2009; Fernando and Garrick 2013); second, genomic evaluation routinely generates marker estimates and these may be used for GWAS; third, complex models used in genetic evaluation can be considered, for instance multiple trait disease all-or-none traits (Parker Gaddis et al. 2014).

The analysis that we propose can be seen as an approximation to a mixture analysis such as BayesCPi. For a given marker, we ask the question: "is this marker worth being included in the model?" whereas we pretend that all the other markers are included in the model. Implicitly, the prior is of a normal distribution with a known variance for loci not being tested and a mixture of a point mass at zero and a normal distribution for the locus being tested. In a mixture model (BayesCPi and similar ones), all markers are scrutinized simultaneously, and the strength of evidence compared against the probability value that a marker should be included in the model (usually labeled as \( p \)). This is probably why the actual numbers for the BF are so different across both methods.

Figure 1 Results (from top to bottom) of single marker regression EMMAX, Bayes Factor for BayesCPi, and Bayes Factor for SNP-BLUP. Bonferroni rejection threshold in EMMAX is 5.9.
We stress that the SNP-BLUP or GBLUP estimation is run only once, and its results are used to construct BFs for different groups of markers (consecutive or not), if desired. This BF combining information from several markers is quite different from estimating the effect of segments of alleles forming haplotypes, where a haplotype can be seen as a multiallelic marker, and where a

Figure 2 Bayes factor profiles (top) and p-values (bottom) for qualification test in French trotters, chromosome 23. The location of the causal mutation marked with a red vertical bar. Bonferroni rejection threshold is 5.9.
different complete estimation must be run for each segment length.

In Bayesian regression models there is a lack of unique criterion to define “relevance” of the association and of corresponding well-defined thresholds; see (Legarra et al. 2015) for a description. The numerical values depend strongly on the assumed prior for marker effects (as can be seen in Figure 1). Thus, two researchers fitting, say, BayesCPI and BayesA may obtain different results. The most popular procedure for genomic evaluation and Bayesian regression is SNP-BLUP or its equivalent GBLUP, both of which assume multivariate normality of marker effects. Most often, a reasonable assumption (point estimate) on the variance of marker effects exists, by a transformation of previous estimates of genetic variance (obtained by pedigree analysis or, using the same data set, by genomic REML or similar methods). Using this point estimate underestimates noise linked to estimation of variance components. Here, we present for the first time a closed-form method to estimate BFs for association analysis based on GBLUP results, and we advocate its use. The statistical properties of the BF have been extensively discussed in the statistics literature, but for mapping causal variants it has two very few relevant properties: the BF can show evidence against and for the null hypothesis, and as data cumulates, the Bayes Factor favors the true hypothesis.

Our results from real data sets show that all methods point to the right marker (the one in stronger LD with the unobserved, but known, QTL). Classical regression analysis is significant and BayesCPI yields a “strong” BF signal. However, the BF observed from SNP-BLUP is 1.07 for the truly associated marker, which is very small support.

Evidence from BF increases when we extend the BF to gather evidence from several markers. A multi–SNP test captures the divergence of the posterior distribution from the 0 vector, and takes into account the posterior dependencies, due to LD, between marker estimates. This is similar to the idea of using the amount of variance explained by each genomic segment (Pérez-Enciso and Varona 2000; Hayes et al. 2010; Nagamine et al. 2012; Fernando and Garrick 2013). The inconvenience of these methods is mostly computational: they require to do either Restricted Maximum Likelihood (Nagamine et al. 2012) or MCMC (Hayes et al. 2010; Fernando and Garrick 2013) to estimate variance components, and that only the Restricted Maximum Likelihood estimation has an associated statistical test (Likelihood ratio test), for which consensus threshold exist (such as 0.05 genome-wide corrected by Bonferroni) whereas the MCMC methods use ad hoc thresholds that are less consensual. Our proposal does not require MCMC or Restricted Maximum Likelihood, but establishing a threshold for the BF is still ambiguous. An approximate method pools information from estimates of marker effects (Wang et al. 2012), but this does not consider not the error in the estimation of marker effects, neither their a posteriori correlation in presence of LD. Our proposal is exact, given a point estimate of variance components but does not necessarily require Restricted Maximum Likelihood or MCMC.

The interpretation of the BF in this study is as follows. There are two models, in the first (null) model all markers have 0 effect, whereas in the second (alternative) model at least one of the markers has an effect. In other words, the BF is a contrast between the region contributing, or not, to the genetic variance. When markers’ evidence is pooled across contiguous markers, the evidence for either of the two competing models increases.

Strangely, in our study including more markers in multiple marker EMMAX does not reinforce evidence, contrary to the BF. This is contrary to results of Chen et al. (2017). The reason is possibly due to the not-too-strong linkage disequilibrium in our data set, for which p-values do not cumulate information across multiple markers. It would seem that, in our data set, it is more difficult to disprove several null hypotheses (null hypothesis in EMMAX: all markers are zero) than to prove an alternative hypothesis (alternative hypothesis in BF: some marker is different from zero).

CONCLUSIONS

We present a Bayesian method (the BF) that complements existing EMMAX methods for QTL detection using marker estimates from SNP-BLUP or (SS)GBLUP from a commonly accepted prior (multivariate normality combined with prior estimates of the genetic variance) and commonly accepted, and used, methods (SNP-BLUP and SSGBLUP). Computations are reasonable and pooling information from several markers is straightforward. Based on our real data set, single marker EMMAX is better to claim significance, whereas multiple marker BF gives a better perspective of influence of LD on the result. This is likely to be data dependent.

ACKNOWLEDGMENTS

INRA SelGen metaprogram is acknowledged for financing its project EpiSel. Project partly supported by the Toulouse Midi-Pyrenees Bioinformatics platform. Luis Varona acknowledges the CGL2016-80155 project of Ministerio de Economía y Competitividad of Spain. Editor and reviewers of preliminary and current submissions made very useful comments and pointed out the Bernal Rubio et al. study.

LITERATURE CITED

Aguilar, I., I. Misztal, D. Johnson, A. Legarra, S. Tsuruta et al., 2010 Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93: 743–752. https://doi.org/10.3168/jds.2009-2730
Andersson, L. S., M. Larhammar, F. Memic, H. Wootz, D. Schwochow et al., 2012 Mutations in DMRT3 affect locomotion in horses and spinal circuit function in mice. Nature 488: 642–646. https://doi.org/10.1038/nature11399
Bernal Rubio, Y. L., L. J. Gualdrón Duarte, R. O. Bates, C. W. Ernst, D. Nonneman et al., 2016 Meta-analysis of genome-wide association from genomic prediction models. Anim. Genet. 47: 36–48. https://doi.org/10.1111/age.12378
Brard, S., and A. Ricard, 2015 Should we use the single nucleotide polymorphism linked to in genomic evaluation of French trotter? J. Anim. Sci. 93: 4651–4659. https://doi.org/10.2527/jas.2015-9224
Casiró, S., D. Velez-Irizarry, C. W. Ernst, N. E. Raney, R. O. Bates et al., 2017 Genome-wide association study in an F2 Duroc x Pietrain resource population for economically important meat quality and carcass traits. J. Anim. Sci. 95: 545–558. https://doi.org/10.2527/jas.2016.1003
Chen, C., J. P. Steibel, and R. J. Tempelman, 2017 Genome-Wide Association Analyses Based on Broadly Different Specifications for Prior Distributions, Genomic Windows, and Estimation Methods. Genetics 206: 1791–1806. https://doi.org/10.1534/genetics.117.202259
Christensen, O. F., 2012 Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet. Sel. Evol. 44: 37. https://doi.org/10.1186/1297-9686-44-37
Dikmen, S., J. B. Cole, D. J. Null, and P. J. Hansen, 2013 Genome-wide association mapping for identification of quantitative trait loci for rectal temperature during heat stress in Holstein cattle. PLoS One 8: e69202. https://doi.org/10.1371/journal.pone.0069202
Endelman, J. B., 2011 Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4: 250–255. https://doi.org/10.3835/plantgenome2011.08.0024
Fernando, R. L., D. Habier, C. Stricker, J. C. M. Dekkers, and L. R. Totir, 2007 Genomic selection. Acta Agric. Scand. A 57: 192–195.
Fernando, R. L., and D. Garrick, 2013 Bayesian methods applied to GWAS. Genome-Wide Assoc. Stud. Genomic Predict.: 237–274. https://doi.org/10.1007/978-1-62703-447-0_10

2306 | A. Legarra, A. Ricard, and L. Varona
APPENDIX

EMMAX tests from GBLUP and SNP-BLUP results

Most of this development is in (Gualdron Duarte et al. 2014; Bernal Rubio et al. 2016; Chen et al. 2017).

SNP-BLUP: The procedure is easier to be presented from the SNP-BLUP method point of view. In this method, the multivariate prior distribution of marker effect is, for a random locus,

\[ p(a|\sigma_a^2) = N(0, \sigma_a^2) \]

and \( \sigma_a^2 \) is a variance component that usually (but not necessarily) is assumed related to genetic variance in the form \( \sigma_a^2 = 2\sigma_g^2 \sum p_i q_i \) (Fernando et al. 2007; VanRaden 2008). For several loci, \( p(a|\sigma_a^2) = N(0, I\sigma_a^2) \) i.e., loci effects are assumed uncorrelated a priori. A linear model for SNP-BLUP is \( y = Xb + Za + e \) where \( Z \) is a matrix of coded genotypes. In SNP-BLUP, the posterior distribution of \( a \) can be obtained by Markov Chain MonteCarlo (MCMC) (Legarra and Misztal 2008) or from the inverse of the left hand side of Henderson’s Mixed Model Equations:

\[
\begin{pmatrix}
X X\sigma_e^{-2} & X Z\sigma_e^{-2} \\
Z X\sigma_e^{-2} & Z Z\sigma_e^{-2} + I\sigma_a^{-2}
\end{pmatrix}
\begin{pmatrix}
b \\
a
\end{pmatrix}
= 
\begin{pmatrix}
X y\sigma_e^{-2} \\
Z y\sigma_e^{-2}
\end{pmatrix}
\]

(A1)

In both cases, it is possible to obtain (a) the estimate of the marker effects is \( \hat{a} = BLUP(a) = E(a|y) \), and (b) two measures of incertitude of \( \hat{a} \), the (frequentist) sampling variance, i.e. \( \text{Var}(\hat{a}) \) and the (Bayesian) posterior variance, i.e. \( \text{Var}(a|y) \). For instance, if the inverse of the right hand side of (A1) is computed:

\[
\begin{pmatrix}
X X\sigma_e^{-2} & X Z\sigma_e^{-2} \\
Z X\sigma_e^{-2} & Z Z\sigma_e^{-2} + I\sigma_a^{-2}
\end{pmatrix}^{-1} =
\begin{pmatrix}
C^{bb} & C^{ba} \\
C^{ab} & C^{aa}
\end{pmatrix}
\]

then \( \text{Var}(a|y) = C^{aa} \) (Bayesian, conditional on data) and \( \text{Var}(\hat{a}) = I\sigma_a^2 - C^{aa} \) (frequentist, over repeated sampling of \( y \)). Matrix \( C^{aa} \) contains a posteriori covariances of marker effects, which reflect allelic frequencies (i.e., a rare SNP is more difficult to estimate) and linkage disequilibrium across markers (two markers in strong LD will have correlated estimates a posteriori, and any of them will be less accurate than a marker not in LD with any other). If estimates are obtained by MCMC, \( C^{aa} \) can be estimated as the covariance matrix of the samples of the posterior distribution, \( p(a|y) \). The R package RRBLUP (Endelman 2011) produces \( sd(\hat{a}) \), and our software GS3 (Legarra et al. 2011) can produce parts of \( C^{aa} \).

GBLUP: the equivalence between GBLUP and SNP-BLUP implies that marker solutions (\( \hat{u} \)) can be backsolved for individual solutions (\( \hat{u} \)) (VanRaden 2008; Strandén and Garrick 2009). Proof is as follows. If individual effects are the sum of marker effects

\[ u = Za \]

\[ \sigma_u^2 = 2\sigma_g^2 \sum p_i q_i \] (VanRaden 2008), \( ZD' = G \) and it can be shown (if \( G \) is invertible) that

\[ E(\hat{a}|u) = \frac{1}{2\sum p_i q_i} \]

The estimation error of \( a \) from GBLUP estimation is more cumbersome to obtain. Assume that the model for GBLUP is \( y = Xb + Wu + e \), and mixed model equations are:

\[
\begin{pmatrix}
X X\sigma_e^{-2} & X W\sigma_u^{-2} \\
W X\sigma_e^{-2} & W W\sigma_u^{-2} + G\sigma_u^{-2}
\end{pmatrix}
\begin{pmatrix}
b \\
u
\end{pmatrix}
= 
\begin{pmatrix}
X y\sigma_e^{-2} \\
W y\sigma_e^{-2}
\end{pmatrix}
\]

(A2)

With the inverse of the left hand side equations

\[
\begin{pmatrix}
X X\sigma_e^{-2} & X W\sigma_u^{-2} \\
W X\sigma_e^{-2} & W W\sigma_u^{-2} + G\sigma_u^{-2}
\end{pmatrix}^{-1} =
\begin{pmatrix}
C^{bb} & C^{ba} \\
C^{ab} & C^{aa}
\end{pmatrix}
\]

Gualdron-Duarte et al. (2014) showed that:

\[
\text{Var}(\hat{a}) = \frac{1}{2\sum p_i q_i} Z G^{-1}(G\sigma_u^{-2} - C^{uu})G^{-1}Z \frac{1}{2\sum p_i q_i}
\]

where \( C^{uu} \) is the element of the inverse of the left hand side matrix corresponding to \( u \); this inverse can be computed by inversion or, again, by MCMC. Finally, the posterior variance of \( a \) is

\[
\text{Var}(a|y) = \frac{\sigma_a^2}{22p_i q_i} - \frac{1}{22p_i q_i} Z G^{-1}(G\sigma_u^{-2} - C^{uu})G^{-1}Z \frac{1}{2\sum p_i q_i}
\]

This allows computing estimates of marker effects and their errors for very complex models, something difficult to do with standard GWAS or Bayesian Regressions.