Recent updates in the BLUPF90 software suite

D. Lourenco[1*], S. Tsuruta[1], I. Aguilar[2], Y. Masuda[3], M. Bermann[1], A. Legarra[4], and I. Misztal[1]

[1] University of Georgia, Department of Animal and Dairy Science, 30602, Athens, GA, USA; [2] Instituto Nacional de Investigación Agropecuaria, 11500, Montevideo, Uruguay; [3] Rakuno Gakuen University, 069-8501, Ebetsu, Hokkaido, Japan; [4] Institut national de recherche pour l'agriculture, l'alimentation et l'environnement, UMR1388 GenPhySE, 31326, Castanet Tolosan, France; [*]danilino@uga.edu

Abstract

BLUPF90 is a flexible software suite that estimates variance components, cross-classified and covariate effects, breeding values, accuracies, and SNP effects, among others. It is based on mixed models and focuses mainly on animal breeding and genetics applications. This collection of software is continually under development to incorporate new methods and algorithms, improve efficiency, and broaden usage. A significant change undergoing BLUPF90 is the merge of variance components and breeding values estimation software. Some of the recent developments include the implementation of p-values for SNP effects in single-step GWAS, a new algorithm to approximate reliability or accuracy of GEBV, the inclusion of metafounders or unknown parent groups in ssGBLUP to set the base populations, and several optimizations to speed up computing time with many genotyped animals. BLUPF90 is a resourceful collection of efficient software that handles millions of genotyped animals and dozens of millions of records or individuals in pedigree.

Introduction

The ever-increasing amount of data used in breeding and genetics applications requires efficient software to solve the mixed model equations (MME). In 1997, Ignacy Misztal developed a simple BLUP program to compute solutions for the MME in Fortran 90/95. The simplicity and flexibility of the architecture allowed extensions and improvements by several generations of scholars. Twenty-five years and many contributors later, BLUPF90 became a collection of software that handles virtually all models used in genetic evaluations in animal breeding. This collection includes software for variance components estimation in linear and linear-threshold models, large-scale genetic evaluations using linear and linear-threshold models, exact and approximate computations of accuracies, and general genomic computations. The software supports different models per trait for virtually any number of traits, multiple effects, random correlated and non-correlated effects, random regressions, missing observations, different covariance structures supplied by the user, and extensive genomic data.

Using genomic data stimulated the most comprehensive changes to BLUPF90. This is because the number of opportunities with genomics is endless. A genomic library with well over 200 options was created and encapsulated by nearly all BLUPF90 programs. Those options are mainly related to quality control of genomic data, statistics and visualization, construction and inversion of genomic and pedigree relationship matrices for genotyped animals, and matrix scaling to ensure compatibility in ssGBLUP. Although the primary genomic method implemented in BLUPF90 is ssGBLUP, GBLUP is also supported. Because of the equivalence between GBLUP-based methods and SNP-BLUP, SNP effects are easily computed through backsolving.

BLUPF90 has been in constant progress for 25 years. Some of the most recent developments in this software collection include the consolidation of variance components and breeding values estimation software, implementation of p-values for SNP effects in single-step GWAS, a new algorithm to approximate reliability or accuracy of GEBV, the inclusion of metafounders or unknown parent groups in ssGBLUP, and optimizations to speed up computations with many genotyped animals. This work will present the most critical innovations recently brought to BLUPF90.

Materials & Methods

Software consolidation. There are currently 11 free programs in the BLUPF90 software suite (http://nce.ads.uga.edu/html/projects/programs/). Some estimate variance components, and some estimate breeding values. Some are frequentist (REML), and some use Bayesian inference through Gibbs Sampling. This generates questions on which software one should use. To avoid this issue, we recently consolidated the linear, frequentist software into one piece called blupf90plus and the most used Bayesian, Gibbs Sampling into gibbsf90plus. Thus, blupf90plus combines blupf90, remlf90, and airemlf90, whereas gibbsf90plus integrates gibbs2f90, gibbs3f90, thrgibbs1f90, and thrgibbs3f90. In blupf90plus, the default option is breeding values estimation, and variance components estimation (VCE) is turned on by adding `OPTION method VCE` to the parameter file. For gibbsf90plus, the same software is used for linear and threshold models. These consolidated programs were thoroughly tested with real and simulated populations.

Unknown Parent Groups (UPG) in ssGBLUP. In the BLUP animal model, the mixed model equations (MME) account for UPG using the QP transformation (Quaas and Pollak, 1981). Besides the default QP transformation for $\mathbf{A}^{-1}$, in ssGBLUP, UPG can be assigned to $\mathbf{A}^{-1}$, the inverse of the genomic relationship matrix ($\mathbf{G}^{-1}$), and $\mathbf{A}^{-1}$ for genotyped animals ($\mathbf{A}_{22}^{-1}$). This is called the QP transformation for $\mathbf{H}^{-1}$ (Misztal et al., 2013). Because pedigree missingness does not affect genomic relationships, UPG can be assigned to $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}^{-1}$; this is known as the altered QP transformation for $\mathbf{H}^{-1}$ (Tsuruta et al., 2019). The transformations were implemented in the BLUPF90 programs. QP transformation for $\mathbf{A}^{-1}$ is the default, whereas the QP for $\mathbf{H}^{-1}$ is activated by `OPTION exact_upg` and the altered QP for $\mathbf{H}^{-1}$ needs the previous option along with `OPTION TauOmegaQ2 0.0 1.0`. All three transformations were tested using US Holstein data for type and production traits with up to 861k genotyped animals.

Metafounders (MF) in ssGBLUP. MF act as proxies for the base animals. Metafounders and UPG are similar concepts; however, metafounders can be related and describe the buildup of coancestry, which is not possible for UPG. With MF, $\mathbf{A}_{22}$ is adjusted to $\mathbf{G}$ computed with 0.5 allele frequency. Both $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}^{-1}$ are then modified by a matrix of relationships within and across metafounders ($\boldsymbol{\Gamma}$). We developed a software called gammaf90 that computes $\boldsymbol{\Gamma}$, and we changed BLUPF90 software to account for MF. When using MF for GEBV estimation, the RANDOM_TYPE in renf90.par should be `add_an_meta`. Among the free software, only blupf90plus and gibbsf90plus work with MF. The implementation of MF was tested in US Holstein data (861k genotypes), chicken data (107k genotypes), and dairy sheep data (6k genotypes).

Optimizations for large-scale ssGBLUP evaluations. The algorithm for proven and young (APY) was implemented in the BLUPF90 suite in 2015. APY allows the construction of a sparse representation of $\mathbf{G}^{-1}$, making computations with millions of genotyped animals very efficient

(Misztal et al., 2014). An efficient algorithm for $\mathbf{A}_{22}^{-1}$ was also implemented (Masuda et al., 2017). For blending of **G** and $\mathbf{A}_{22}$, the initial implementation was inefficient. We recently changed the implementation of Colleau (Colleau, 2002) in BLUPF90 to compute coefficients of $\mathbf{A}_{22}$ separately for core and noncore animals. This new implementation was tested with US Holstein data containing 3.4M genotyped animals. Extensive changes in the iteration on data (i.e., iod) programs were also made to improve the computational efficiency in evaluations with extensive data. This included splitting data and pedigree and using different computing cores for parallel i/o.

An efficient algorithm to approximate reliability or accuracy of GEBV. Reliabilities of GEBV can be calculated as a function of prediction error variances (PEV) obtained from the diagonal of the inverse of the coefficient matrix of MME. This inverse is not feasible with large datasets, and therefore, reliabilities need to be approximated. An algorithm based on Liu et al. (2017) was developed that considers the sparsity of APY $\mathbf{G}^{-1}$ (Bermann et al., 2021). The algorithm computes weights for pedigree and phenotype contributions, adds the weights to the diagonals of APY $\mathbf{G}^{-1}$, and uses block sparse inversion to compute coefficients for core and noncore animals separately. A brand-new program, accf90GS2, was created to accommodate the new reliability calculations. It is based on accf90 and considers maternal, multi-trait, and repeatability models. Tests included single- and three-trait models using American Angus datasets with up to 335k genotyped animals. An algorithm for the exact reliability of indirect predictions was also recently implemented.

P-values for SNP effects in single-step GWAS. Because of the equivalence between ssGBLUP and ssSNP-BLUP, GEBV from the former can be backsolved to SNP effects. GWAS may use SNP effects or proportion of variance explained by SNP; however, it is usual to present p-values to test the statistical significance of SNP on traits of interest. We implemented p-values in ssGWAS based on PEV for SNP effects (Aguilar et al., 2019). These GWAS are mathematically equivalent to single-marker regression as implemented in the EMMAX software. Obtaining p-values requires `OPTION snp_p_value` in blupf90 and postGSf90. Testing included US Holstein and American Angus data with 10k and 1.4k genotyped bulls, respectively.

Results

Software consolidation. The consolidation was done successfully, and all the tests resulted in identical solutions between blupf90plus and either blupf90 or remlf90 and airemlf90. The same was observed when comparing gibbsf90plus and its component software. Therefore, the consolidated software is ready for official deployment.

Unknown Parent Groups (UPG) in ssGBLUP. Using the altered QP transformation for $\mathbf{H}^{-1}$ resulted in greater reliability and lower dispersion bias than the QP for $\mathbf{H}^{-1}$ and $\mathbf{A}^{-1}$. Although the three transformations are implemented, we recommend using the altered QP transformation when applying UPG in ssGBLUP. This may hold in most cases; however, testing is suggested.

Metafounders (MF) in ssGBLUP. The application of MF required new software, gammaf90, to estimate the gamma matrix. Gammaf90 is efficient with virtually any number of genotyped animals and MF; however, an accurate estimation required enough genotyped animals linked to MF. Overall, reliabilities or accuracies and dispersion bias were similar between models with the altered QP transformation in UPG and MF.

Optimizations for large-scale ssGBLUP evaluations. Time to compute the coefficients of $\mathbf{A}_{22}$ needed for blending in APY drastically reduced with the new algorithm for core and noncore animals separately. With 3.4M genotyped Holsteins, it took 12 minutes, compared to 41 hours with the previous method. The changes for parallel reading in the iod programs significantly reduced the time to solve the MME. Both modifications were necessary steps towards more efficient software for large-scale evaluations.

Efficient algorithm to approximate reliability or accuracy of GEBV. The new algorithm provided higher correlations (~ 0.98) with reliabilities based on PEV than the default algorithm (~ 0.90). The increase in computing time was linear with the number of genotyped animals and traits. It took 11 minutes for a three-trait model with 10M animals in the pedigree and 335k genotypes.

P-values for SNP effects in single-step GWAS. We successfully implemented p-values for ssGWAS in BLUPF90. As expected, results were numerically identical to single-marker regression (i.e., EMMAX). Therefore, ssGWAS became a general and efficient QTL detection and test approach. This method can be used in complex datasets such as those used in animal breeding, where only a proportion of pedigreed animals are genotyped.

Discussion

After 25 years, BLUPF90 became a resourceful collection of efficient software that handles ssGBLUP evaluations with millions of genotyped animals. Breed associations and breeding companies worldwide use this software suite for genomic evaluations in nearly all farm animal species. Academia uses extensively public versions for research. The subsequent updates will focus on further improving efficiency and convergence for large datasets and expanding the methods to increase the usability of the BLUPF90 software suite in animal and plant breeding. Constructive feedback is always welcome.

References

Aguilar I., Legarra A., Cardoso F., Masuda Y., Lourenco D. *et al.* (2019) Genet. Sel. Evol. 51:28. https://doi.org/10.1186/s12711-019-0469-3

Bermann M., Lourenco D., Misztal I. (2021) J. Anim. Sci. skab353 https://doi.org/10.1093/jas/skab353

Colleau J.J. (2002) Genet. Sel. Evol. 34:409-421 https://doi.org /10.1186/1297-9686-34-4-409

Liu Z., VanRaden P., Lidauer M., Calus M., Benhajali H. *et al.* (2017) Interbull Bull. 51:75–85.

Masuda Y., Misztal I., Legarra A., Tsuruta S., Lourenco D.A.L. *et al.* (2017) J. Anim. Sci. 95:49-52. https://doi.org/10.2527/jas.2016.0699

Misztal I., Vitezica Z. G., Legarra A., Aguilar I., and Swan A.A. (2013) J. Anim. Breed. Genet. 130:252-258. https://doi.org/10.1111/jbg .12025

Misztal I., Legarra A., and Aguilar I. (2014) J. Dairy Sci. 97:3943-3952. https://doi.org/10.3168/jds.2013-7752.

Quaas R. L. and Pollak E. J. (1981) J. Dairy Sci. 64:1868-1872. https://doi.org/10.3168/jds.S0022-0302(81)82778-6

Tsuruta S., Lourenco D.A.L., Masuda Y., Misztal I., and Lawlor T.J. (2019) J. Dairy Sci. 102:9956-9970. https://doi.org/10 .3168/jds.2019-16789