MiXBLUP 3.0 - Software for large genomic evaluations in animal breeding programs

J. Vandenplas[1*], R.F. Veerkamp[1], M.P.L. Calus[1], M.H. Lidauer[2], I. Strandén[2], M. Taskinen[2], M. Schrauf[1] and J. ten Napel[1]

[1] Wageningen University & Research, P.O. Box 338, 6700 AH Wageningen, the Netherlands; [2] Natural Resources Institute Finland (Luke), Jokioinen, Finland; *jeremie.vandenplas@wur.nl

Abstract

The software package MiXBLUP 3.0 allows to efficiently estimate (genomic enhanced) breeding values in livestock using various linear mixed effect models. MiXBLUP aims to be user-friendly while supporting efficient algorithms for solving large genomic evaluations. Originally developed for pedigree-based models, MiXBLUP has been extended to four different approaches for single-step genomic evaluations, allowing simultaneous analyses of all information for both genotyped and ungenotyped individuals. For all approaches, shared-memory parallelism is supported using tailor-made procedures and parallel libraries. It has been tested on some of the largest genomic datasets in the world. The software is developed by Wageningen University and Research in collaboration with the Natural Resources Institute Finland (Luke).

Introduction

During the last decades, the mixed model equations have been used for estimating breeding values in livestock using phenotypes at the traits of interest, together with pedigree. With the advent of genomic selection, the traditional mixed model has been extended to analyse simultaneously phenotypic, pedigree, and genomic information (Legarra et al., 2014). Such models are commonly called single-step genomic models. Currently, genomic information used in routine evaluations contains the genotypes of several (hundred) thousand animals for around 50 thousand single nucleotide polymorphisms (SNP). The integration of such genomic datasets with large numbers of genotyped animals led to new challenges for solving the mixed model equations. Therefore, several approaches for solving the single-step models have been proposed in the literature and implemented in software for large genomic evaluations in livestock (e.g., Mäntysaari et al., 2020).

Since 2006, the MiXBLUP software has been developed for breeding value estimation in commercial genetic programmes. It supports modern developments, such as complex models (e.g., random regression models) or linear-threshold models. With the advent of genomic selection, MiXBLUP has been continuously developed since 2014 to support the use of genomic information through four different single-step genomic approaches. This work aims to present the state-of-the-art of MiXBLUP for large genomic evaluations in animal breeding.

MiXBLUP and its components

MiXBLUP philosophy. The aim of developing MiXBLUP is to utilize efficient computing strategies for solving mixed model equations, while being easy to install and easy to use, even for sophisticated models. For example, MiXBLUP supports a wide range of input files and generates files in the required format for the solvers. So, various formats for the SNP genotype files are supported. When the genotype files are text files, they may contain marker alleles or marker genotypes. The marker alleles or genotypes may be in space-separated or in dense format. MiXBLUP now supports the Plink 1 binary format that requires less memory than text files by taking advantage that only two bits (instead of one byte) are needed to store a biallelic

SNP genotype (Chang et al., 2015). Due to this advantage, the Plink 1 binary format is also supported by other programs of the software package MiXBLUP described hereafter.
Being a versatile software, MiXBLUP is used in breeding programmes for cattle, pigs, poultry, sheep, horses, fish, goats, and dogs (ten Napel et al., 2021).

Models supported for breeding value estimation. The statistical method commonly used for routine genetic evaluation is the so-called Best Linear Unbiased Prediction (BLUP). Briefly, a univariate animal model is as follows:
$\mathbf{y} = \mathbf{Xb} + \mathbf{Wa} + \mathbf{e}$,
where **y** is the vector of phenotypes, **b** is the vector of fixed effects, **a** is the vector of additive genetic effects, and **e** is the vector of residuals. The incidence matrices **X** and **W** relate the phenotypes to the fixed effects and additive genetic effects, respectively.
It is assumed that **e** follows a multivariate normal distribution, $MVN(\mathbf{0}, \mathbf{R}\ \sigma_e^2)$ with **R** being a diagonal matrix, and σ_e^2 being the residual variance, and that **a** follows a multivariate normal distribution $MVN(\mathbf{0}, \mathbf{H}\ \sigma_a^2)$ with σ_a^2 being the additive genetic variance. Traditionally, the **H** matrix was equal to the pedigree-based relationship matrix (**A**). In single-step genomic BLUP, the **H** matrix is a relationship matrix that includes the genomic relationship matrix (**G**) for genotyped animals and modified pedigree relationships for other animals.
Due to limitations related to the computation of **G** and its inverse ($\mathbf{G}^{-1}$) when the number of genotyped animals is large, some methods were proposed to approximate $\mathbf{G}^{-1}$, such as the algorithm for proven and young animals (APY; Misztal et al., 2014), or to compute $\mathbf{G}^{-1}$ implicitly using the Woodbury matrix identity (Mäntysaari et al., 2017). Such models that rely on **G** will hereafter be referred to as single-step genomic BLUP (ssGBLUP). Another proposed approach to avoid the computation of **G** or $\mathbf{G}^{-1}$, is to augment the solution vector of the mixed model equations with SNP effects. The SNP effects can be fitted explicitly or implicitly in the underlying model as random effects (e.g., Liu et al., 2014). Such models will hereafter be referred to as single-step SNPBLUP (ssSNPBLUP).
In addition to the original ssGBLUP for which **G** is explicitly inverted, the software MiXBLUP supports two ssGBLUP-based and one ssSNPBLUP approaches. For all approaches, a residual polygenic (RPG) effect that considers the part of the additive genetic effects not explained by the genomic data (Mäntysaari et al., 2020) can be fitted.
First, MiXBLUP supports the algorithm for proven and young animals (APY; Misztal et al., 2014). This algorithm approximates $\mathbf{G}^{-1}$ by dividing the genotyped animals into a limited set of so-called core animals, and the remaining genotyped animals are noncore. The computation of $\mathbf{G}_{APY}^{-1}$ involves the inversion of a genomic relationship submatrix among the core animals and the recursive computation of other coefficients for noncore animals.
Second, MiXBLUP supports the computation of the so-called **T** matrix, needed for an implicit inversion of **G** using the Woodbury matrix identity (ssGTABLUP; Mäntysaari et al., 2017). Briefly, when an RPG effect is fitted, **G** has the form $\mathbf{G} = \mathbf{ZBZ}' + w\mathbf{A}_{gg}$, where **Z** is a matrix of centered SNP genotypes, **B** is a diagonal matrix with diagonal entries scaling the genetic variance, w is the proportion of the RPG effect, and $\mathbf{A}_{gg}$ is the pedigree-based relationship matrix among genotyped animals. According to the Woodbury matrix identity:
$$\mathbf{G}^{-1} = 1/w\ \mathbf{A}_{gg}^{-1} - \mathbf{TT}' \quad (1)$$
where $\mathbf{T} = 1/w\ \mathbf{L}^{-1}\ \mathbf{Z}'\mathbf{A}_{gg}^{-1}$, with **L** being the Cholesky factor of $\mathbf{K} = (1/w\ \mathbf{Z}'\ \mathbf{A}_{gg}^{-1}\mathbf{Z} + \mathbf{B}^{-1})$.
In practice, $\mathbf{G}^{-1}$ is not computed explicitly, because the MiXBLUP iterative solvers only require its multiplication by a vector, which can be computed efficiently in multiple parts by using the matrices $\mathbf{A}_{gg}^{-1}$ and **T** (Mäntysaari et al., 2017).
Third, MiXBLUP supports the ssSNPBLUP system proposed by Liu et al. (2014). This system has the specificity that the random SNP effects are not explicitly fitted in the model. Indeed,

the solution vector that includes traditionally fixed and random effects is augmented with the SNP effects, and an inverse of the (co)variance matrix associated with the random additive genetic and SNP effects is derived. It has been shown that the absorption of the equations associated with the SNP effects results in the ssGTABLUP system (Mäntysaari et al., 2020).
The MiXBLUP software provides several options for modelling base populations, such as the concept of metafounders (Legarra et al., 2015), and genetic groups. Different options for scaling genomic relationships to the same base as pedigree-based relationships are also available.

Data processing. Since the users can provide a wide range of data file formats, MiXBLUP will first convert all data files in a format that its two solvers and other programs can read. Indeed, once converted, MiXBLUP will call different parallel programs for computing the various matrices required by the solvers. For example, MiXBLUP will call the program calc_grm, which has an important role in single-step genomic evaluations because it computes relationship matrices for single or multiple, related or unrelated base populations.

Two different solvers. The MiXBLUP software supports two different solvers. The default solver is derived from the MiX99 software (Lidauer et al., 2019). This software was initially developed for pedigree-based genetic evaluations by the National Resources Institute Finland (Luke). With the advent of genomic selection, specifically the single-step models, both software MiX99 and MiXBLUP were adapted for using genomic information simultaneously with pedigree and phenotypic information. The MiX99 software is optimized for solving pedigree-BLUP, ssGTABLUP, and ssGAPYBLUP.
Recently, a solver called hpblup (Vandenplas et al., 2020) was developed within the consortium Breed4Food and added to MiXBLUP. This program was specifically developed and optimized for solving the ssSNPBLUP system of Liu et al. (2014).

Computations. For efficient computations, both solvers support shared-memory parallelism through the OpenMP API (www.openmp.org) and various parallelized libraries. Furthermore, both solvers use the preconditioned conjugate gradient (PCG) algorithm as an iterative method. The program hpblup also supports a two-level PCG algorithm for efficiently solving the ssSNPBLUP linear system (Vandenplas et al., 2020).
The main computational cost of a PCG iteration is the multiplication of the coefficient matrix of a system of equations by the so-called direction vector. For both solvers, this multiplication is performed in three main steps (in different orders for each solver). The first step is the multiplication of the least square part of the coefficient matrix by a vector; the second step involves all multiplications of matrices related to pedigree information, and the third step involves all multiplications of matrices related to genomic information. This third step is usually the most computationally expensive for large evaluations, and shared-memory parallel computing has been implemented in both solvers to improve their efficiency. For ssGAPYBLUP and ssGTABLUP, the computations related to genomic information mainly rely on parallel libraries, such as the Intel Math Kernel Library. For ssSNPBLUP, a tailor-made parallel procedure has been developed for the multiplication of the genotype matrix by another matrix, while keeping the genotype matrix in the Plink 1 binary compressed format (Vandenplas et al., 2020). More details in the different computations of both solvers can be found in Strandén et al. (1999), Mäntysaari et al. (2017), Lidauer et al. (2019), and Vandenplas et al. (2020).

Other programs. In addition to the two solvers, MiXBLUP also includes other programs such as calc_grm, MiXPRED, and reliabilities.exe. The program calc_grm is a program for computing pedigree and genomic relationship matrices, their inverses, or some related matrices,

such as $\mathbf{G}_{APY}^{-1}$ and $\mathbf{T}$. This program is also used for converting a text genotype file into Plink 1 binary files, and *vice versa*. The program MiXPRED allows the computation of interim GEBV of newly genotyped animals based a decomposition of the GEBVs obtained from a previous single-step evaluation into its different components, namely the direct genomic breeding values, the residual polygenic effects, and if fitted, other effects such as the genetic groups. The program reliabilities.exe allows the computation of approximated reliabilities for all animals in an evaluation, based on the algorithm of Tier and Meyer (2004).

System requirements and availability. The MiXBLUP software is available for Microsoft Windows and Linux/Unix environments and can be downloaded from the website www.mixblup.eu. MiXBLUP works with a license, which can be ordered through its website. More information can be found on www.mixblup.eu.

Acknowledgments
This study was financially supported by the Dutch Ministry of Economic Affairs (TKI Agri & Food Project 16022) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics, and Topigs Norsvin.

References
Chang C.C., Chow C.C., Tellier L.C., Vattikuti S., Purcell S.M., et al. (2015). Gigasci. 4:7. https://doi.org/10.1186/s13742-015-0047-8
Legarra, A., Christensen, O.F., Aguilar, I. and Misztal, I. (2014). Livest. Sci. 166:54-65. https://doi.org/10.1016/j.livsci.2014.04.029.
Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I. and Misztal, I. (2015) Genetics 200:455–468. https://doi.org/10.1534/genetics.115.177014.
Lidauer, M., Matilainen, K., Mäntysaari, E., Pitkänen, T., Taskinen, M., et al. (2019) MiX99: Technical reference guide for MiX99 solver. Release XI/2019. Available at: https://jukuri.luke.fi/bitstream/handle/10024/545540/refguide_solver.pdf?sequence=1&isAllowed=y
Liu, Z., Goddard, M., Reinhardt, F., and Reents, R. (2014) J. Dairy Sci. 97:5833–5850. https://doi.org/10.3168/jds.2014-7924.
Mäntysaari, E.A., Evans, R.D., and Strandén, I. (2017) J. Anim. Sci. 95:4728–4737. https://doi.org/10.2527/jas2017.1912.
Mäntysaari, E.A., Koivula, M., and Strandén, I. (2020) J. Dairy Sci. 103:5314–5326. https://doi.org/10.3168/jds.2019-17754.
Misztal, I., Legarra, A., and Aguilar., I. (2014) J. Dairy Sci. 97:3943–3952. https://doi.org/10.3168/jds.2013-7752.
Strandén, I., and Lidauer, M. (1999) J. Dairy Sci. 82:2779–2787. https://doi.org/10.3168/jds.s0022-0302(99)75535-9.
ten Napel, J., Vandenplas, J., Lidauer, M., Stranden, I., et al. (2021) Manual – MiXBLUP 3.0.1 manual, V3.0 – 2021-11. Available at: https://www.mixblup.eu/documents/8412108060_ASG_WLR_MixBlup%20Manual_LR.pdf
Tier, B., and Meyer, K. (2004) J. Anim. Breed. Genet. 121:77–89. https://doi.org/10.1111/j.1439-0388.2003.00444.x.
Vandenplas, J., Eding, H., Bosmans, M., and Calus, M.P.L. (2020) Genet. Sel. Evol. 52:24. https://doi.org/10.1186/s12711-020-00543-9.