Snp_blup_rel: software for calculating individual animal SNP-BLUP model reliabilities

Hafedh Ben Zaabza, Esa A. Mäntysaari and Ismo Strandén

Natural Resources Institute Finland (Luke), FI-31600 Jokioinen, Finland
e-mail: hafedh.benzaabza@luke.fi

The snp_blup_rel program computes model reliabilities for genomic breeding values. The program assumes a single trait SNP-BLUP model where the breeding value can include a residual polygenic (RPG) effect. The reliability calculation requires elements of the inverse of the mixed model equations (MME). The calculation has three steps: 1) MME calculation, 2) MME coefficient matrix inversion, and 3) reliability computation. When needed, the inverted matrix can be saved after step 2. Step 3 can be used separately to new genotypes which do not contribute information to Step 2. When an RPG effect is included, an approximate method based on Monte Carlo sampling is applied. This reduces the MME matrix size and allows including many genotyped individuals. The program is written in Fortran 90/95, and uses LAPACK subroutines which enable multi-threaded parallel computing. The program is efficient in terms of computing time and memory requirements, and can be used to analyze even large genomic data. Thus, the program can be used in calculating model reliabilities for large national genomic evaluations.

Key words: genomic evaluation, model reliability, SNP markers

Introduction

The dairy cattle evaluation community is increasingly relying on estimated breeding values (EBV) based on genomic information. Two equivalent genomic models (Strandén and Garrick 2009) are available for calculating genomic EBV and their reliabilities: genomic best linear unbiased prediction (GBLUP) and marker effects models. Currently the most common genomic markers are single nucleotide polymorphisms (SNP), and the marker effects models are, thus, generally used to predict SNP effects (SNP-BLUP). Calculation of genomic reliability for individual EBV by GBLUP requires inverting the coefficient matrix of the mixed model equations (MME) that include the inverse of the genomic relationship matrix. These matrix inversions become infeasible as the number of genotyped animals increases (Fernando et al. 2016). In contrast, the MME matrix size of SNP-BLUP is bounded by the number of SNP markers. Thus, SNP-BLUP can scale to large numbers of genotyped animals better than GBLUP.

In SNP-BLUP, the inverse of the MME coefficient matrix provides the prediction error (co)variances (PEV) of the SNP solutions, which can thereafter be used to derive PEV for the EBV. While the computation of reliabilities with GBLUP requires the inclusion of all animals into the genomic relationship matrix, the SNP-BLUP model involves two stages: 1) PEV for the SNP effects are estimated using the genotypes of phenotyped animals, and 2) reliability of the EBV of each genotyped animal is computed using the PEV of the SNP effects without the need to consider its phenotype.

Typically, genetic variation cannot be completely detected by SNP markers because of the incomplete linkage disequilibrium between the used SNP markers and the quantitative trait loci of the evaluated trait. A residual polygenic (RPG) effect is often included to account for the variance not captured by SNP markers. However, including an RPG effect into the SNP-BLUP model leads to an increase in the size of the MME by the number of genotyped animals (Liu et al. 2016). This can be avoided by an approximation based on Monte Carlo (MC) sampling of pseudo markers that construct the relationships among animals (Ben Zaabza et al. 2020).

In this paper we will describe an efficient program, called snp_blup_rel, which allows the calculation of reliabilities of EBV by SNP-BLUP, both with and without an RPG effect. A single trait SNP-BLUP with a general mean is assumed. If multiple traits (with different training population and/or variance components) need to be analyzed, then each trait will require its own inverse of the coefficient matrix of the MME. The snp_blup_rel program has been used in earlier studies (Liu et al. 2017, Ben Zaabza et al. 2020), but here we concentrate on describing the program itself. Firstly, we present some theoretical background and the algorithm used in the implementation. Then, we describe the most important options, and give examples of genomic reliability calculations.

Manuscript received June 2020

Theoretical background

EBV reliability in a genomic model

Consider a GBLUP model

$$y=1_n\mu+u_G+e, \quad (1)$$

where y is an n×1 data vector, 1_n is a vector of n ones, μ is the unknown overall mean, u_G is a n×1 vector of additive genetic effects; e is a n×1 vector of residuals. Assume that $u_G \sim N(0, G_w\sigma_u^2)$ and $e \sim N(0, R\sigma_e^2)$ where σ_u^2 is the additive genetic variance, and σ_e^2 is the residual variance. The R matrix can be assumed to be diagonal with elements $R_{ii}=\frac{1}{w_i}$, where w_i is the weight for observation *i*. When both the marker and the pedigree information influence the genomic relationship matrix G_w, it can be written $G_w=(1-w)ZZ'+wA_{22}$, where *w* is the RPG proportion, Z is an *n* by *m* marker matrix of centered and scaled genotypes, and A_{22} is the pedigree relationship matrix among the genotyped animals. Common centering and scaling approaches for the Z matrix are called VanRaden (2008) methods 1 and 2. In the Z matrix, column *k* for a marker has value $\frac{1}{s_k}(0-2p_k)$, $\frac{1}{s_k}(1-2p_k)$, and $\frac{1}{s_k}(2-2p_k)$ for genotypes BB, AB, and AA, respectively, where p_k is the (base population) allele frequency of marker *k*.

In method 1, $s_k=\sqrt{\sum_{l=1}^{m} 2p_l(1-p_l)}$, and in method 2, $s_k=\sqrt{m2p_k(1-p_k)}$.

An equivalent SNP-BLUP model to the GBLUP model (1) is (Ben Zaabza et al. 2020)

$$y=1_n\mu+Wu+e, \quad (2)$$

where $W=[\sqrt{(1-w)}Z \quad \sqrt{w}1_n]$, and $u' = [g' \quad a']$, u is a vector of the unknown pseudo marker effects, g is a vector of the SNP marker effects, and a is a vector of the effects for the RPG based breeding values. Thus, $u_G=Wu$. It is assumed that $g \sim N(0, I_m\sigma_u^2)$ and $a \sim N(0, A_{22}\sigma_u^2)$.

MME for model [2] can be written as

$$\begin{bmatrix} 1_n'R^{-1}1_n & 1_n'R^{-1}W \\ W'R^{-1}1_n & W'R^{-1}W+\lambda\Omega^{-1} \end{bmatrix}\begin{bmatrix}\hat{\mu}\\ \hat{u}\end{bmatrix}=\begin{bmatrix}1_n'R^{-1}y\\ W'R^{-1}y\end{bmatrix}, \quad (3)$$

where $\lambda=\frac{\sigma_e^2}{\sigma_u^2}$ and $\Omega=\begin{bmatrix}I_m & 0\\ 0 & A_{22}\end{bmatrix}$. Denote the inverse of the coefficient matrix in MME (3) by $\begin{bmatrix}C^{\mu\mu} & C^{\mu u}\\ C^{u\mu} & C^{uu}\end{bmatrix}$ where C^{uu} is the submatrix associated with the genetic effects û. Then, the estimated genomic breeding values are $\widehat{u_G}=W\hat{u}$.

Reliability for the estimated genomic breeding value of genotyped animal i is

$$r_i^2=1-\lambda\frac{w_iC^{uu}w_i'}{\{G_w\}_{ii}}, \quad (4)$$

where $\{G_w\}_{ii}$ is diagonal element *i* in matrix G_w, and w_i is the row in the W matrix corresponding to the *i*-th genotyped animal. Note that the dimension of MME is *m*+*n*, i.e., size of this matrix increases by the size of number of genotyped animals.

Residual polygenic effect by Monte Carlo

The SNP-BLUP model (2) can be approximated with the aid of Monte Carlo (MC) sampling (Ben Zaabza et al. 2020):

$$y=1_n\mu+Sv+e, \quad (5)$$

where $S=[\sqrt{(1-w)}Z \quad \sqrt{w}U]$, U is an $n\times n_{MC}$ matrix of the MC samples for genotyped animals, $v'=[g' \quad g_p']$ is a vector of $(m+n_{MC})$ unknown (pseudo) marker effects, of which *m* are due to the SNP markers and n_{MC} approximate the RPG breeding values. Breeding values are $u_G \approx Sv$. It is assumed that $v \sim N(0, I_{m+n_{MC}}\sigma_u^2)$. We use MC samples

$U=\frac{1}{\sqrt{n_{MC}}}[a_1 \quad \cdots \quad a_{n_{MC}}]$, where $a_i \sim N(0,A_{22})$, i=1,…,n_{MC}, and n_{MC} is the number of MC samples. Note that the variance of a is A_{22}=Var(a)=E(aa')-E(a)E(a)'=E(aa'), which is equal to UU', where the additive genetic variance is assumed to be equal to 1 (see Ben Zaabza et al. 2020).

The MME for the MC-SNP-BLUP model (5) is

$$\begin{bmatrix} 1_n^{'}R^{-1}1_n & 1_n^{'}R^{-1}S \\ S^{'}R^{-1}1_n & S^{'}R^{-1}S+\lambda I_{m+n_{MC}} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{v} \end{bmatrix} = \begin{bmatrix} 1_n^{'}R^{-1}y \\ S^{'}R^{-1}y \end{bmatrix} \quad (6)$$

Denote the coefficient matrix of this MME by C_s and its inverse matrix elements as $C_s^{-1}=\begin{bmatrix} C_s^{\mu\mu} & C_s^{\mu v} \\ C_s^{v\mu} & C_s^{vv} \end{bmatrix}$. Approximate reliability for EBV of animal *i* is

$$r_i^2=1-\lambda\frac{s_i C_s^{vv} s_i^{'}}{\{G_s\}_{ii}}, \quad (7)$$

where s_i is row *i* in S, and $\{G_s\}_{ii}$ is diagonal element *i* in the genomic relationship matrix G=SS´=(1-w)ZZ´+wUU´. When n_{MC} is large, UU' tends towards A_{22}. For more details on this model, see Ben Zaabza et al. (2020).

Algorithm and implementation

In the snp_blup_rel program, centering of the Z matrix can be based either on the current genotypes or on given allele frequencies, e.g., base population allele frequencies estimated by a method like the GLS model (McPeek et al. 2004, Strandén et al. 2017). Ideally, the base population allele frequencies are used in the genomic relationship matrix (VanRaden 2008, Liu et al. 2017).

The RPG effect can be included into the model by two alternative approaches. In the exact approach, the precalculated A_{22} matrix is read into memory, inverted, and used in the MME (3). In the other approach, MC sampling is used to approximate the RPG effect as in model (5) and MME (6). When size of the A_{22} matrix is large, computing time by the exact approach can be substantially longer than in the approximate MC-SNP-BLUP approach, because the computing time depends on the size of the MME coefficient matrix. However, in contrast to the exact approach, the Monte Carlo-based sampling method to estimate the reliability of the SNP-BLUP model including the RPG effect is very fast, because the sampling can be implemented efficiently by using the full pedigree and retaining only samples pertaining to the genotyped animals (Ben Zaabza et al. 2020).

A disadvantage of MC-SNP-BLUP is that the computed reliabilities are inflated when the number of MC samples is too low. The inflation is tolerable when the number of MC samples is about the number of SNP markers and the proportion of RPG effect *w* is about 20% or less. In addition, the minimum number of MC samples depends on the number of genotyped animals and on the population structure (Ben Zaabza et al. 2020). When the RPG proportion *w* is high, the reliability calculations depend more on MC sampling. Thus, a large RPG effect proportion *w* needs many MC samples to reduce the inflation. However, the number of MC samples needed to give a sufficiently accurate genomic reliability approximation is typically much less than the number of genotyped animals (Ben Zaabza et al. 2020).

In snp_blup_rel, all matrix inversions, such as the inversion of the MME coefficient matrix and the A_{22} matrix, use LAPACK (Anderson et al. 1999) subroutines DPOTRF and DPOTRI. When computing the MME, the W matrix in (3) or S matrix in (6) can be stored in memory which allows the use of DGEMM subroutine in making the matrix product $W´R^{-1}W$ or $S´R^{-1}S$ efficiently. The snp_blup_rel implementation employs LAPACK subroutines from the Intel Math Kernel Library (Intel Math Kernel Library Reference Manual 2014). These allow taking advantage of parallel computing on a multi-core computer.

Input files

A short description of the input files and some options are in Tables 1 and 2. The snp_blup_rel program can be given six different input files depending on the model and computational approach:

1) A genotype data file. This always needs to be given. In the file, the first column is the individual ID code, followed by the marker genotypes. The file format is described later.

2) An optional (-a) file having the allele frequencies can be given to center the genotypes. The file format is described later.

3) An optional (-o) data file is needed when weights or observations are specified. If this file is not given, all genotyped animals are assumed to have an observation. The data file must include at least the individual ID code. The default column for the ID code is one. When no column for observation weight is given, default weight is one. Thus, when a data file is given, genotyped animals that appear in the data file can be considered reference animals and those without observation can be considered candidate animals.

4) An optional (-A) file having A_{22} matrix is needed when the RPG effect is accounted exactly in the model reliability calculations. The file format is described later.

5) An optional (-ped) pedigree file is needed when an RPG effect is accounted by the MC-SNP-BLUP approach. The file has three columns having integer numbers for individual, sire, and dam ID codes.

6) An optional (-F) file having inbreeding coefficients can be given to account for inbreeding coefficients (Mendelian sampling term) when simulating the MC samples in MC-SNP-BLUP. In this file, the first column has the individual ID code, and the inbreeding coefficient is by default in the third column.

The marker data in the genotype file is used to build the Z matrix. By default, the file is assumed to have the allele counts: zero, one and two for each genotype. There is some flexibility in the genotype file format. However, for the option "-m raw" the genotypes are floating point numbers which allows reading a pre-calculated user specified Z matrix. By default, the genotype file is assumed to have space delimited numbers for the genotypes of an individual on the same line.

The first column in the genotype file is always the ID code number. By default, the genotypes are assumed to be space separated floating point numbers. The option "-int" informs that the genotypes are space separated integer numbers. However, the option "-nospace" assumes genotypes to be single digit integer numbers without the space separator. Note that this option assumes that the first column has the ID code number which is separated by at least one space from the marker data. Alternatively, when all columns in the file are integers, the space requirement can be relaxed by giving a Fortran type format. For example, "-FMT '(i7,1x,50000i1)'" assumes a seven-digit ID code, one space, and 50000 single digit genotypes. Giving the right number of markers, here 50000, is not critical. In fact, when a too large number is given, the program calculates and uses the number of markers on the first row in the genotyped file. The option "-s first last jump" allows selecting a subset of markers in the specific region to be included in the analysis from the genotype file. The '-s' option allows omitting any non-genotype columns before the first genotype column. For example, giving "-s 4 0 0" will start from column 5 onwards all columns as markers when the zeros are used for default values (i.e. "last" replaced by the number of markers, and "jump" by one).

Table 1. Main options for snp_blup_rel describing the data format and desired models

Options	Description
-nthr n	number of threads. Default from variable MKL_NUM_THREADS
-o data_file	name of the data file. Each line has an ID number of animal with observation
-wt col_num	column number of weights in the data_file
-e col_num	column number of effective daughter contributions (EDC) for weight=EDC/λ, $\lambda = \frac{\sigma_e^2}{\sigma_u^2}$
-h2 v	value of animal model heritability. Assumes $\lambda = \frac{1-h^2}{h^2}$
-h2s v	value of sire model heritability. Assumes $\lambda = \frac{4-h^2}{h^2}$
-m method	genomic data centering/scaling method (PvR1, PvR2, raw …)
-a afreq_file	afreq_file is file for allele frequencies by markers
-s fst lst jmp	subsetting by every jmpth marker in the genotype file starting from first (fst) to last (lst) marker column

The pedigree relationship matrix A_{22} has to be provided in a file if the RPG effect is accounted exactly (model [2]). Two matrix formats are supported. In the co-ordinate format ("-A file"), the file has non-zero elements of the upper or lower triangle of the matrix but not both. Each line has three space separated values. The first two are the integer valued ID codes to identify the individuals which the relationship coefficient concerns. The third column has the relationship value.

Note that each value given in the relationship matrix is added to the upper and lower triangular matrix in memory. So, if the same position or its transpose, e.g., (1,2) and (2,1), is referred to on two different lines, the sum of their values is used. In the lower triangle dense format ("-Alower file"), the first row has size of the matrix. The second row has the ID code numbers. The rest of the file has the lower triangle of the A_{22} matrix. An example of this file format is given in Supplementary Materials. Matrix A_{22} can be computed by any means that the user desires, e.g., RelaX2 program (Strandén and Vuori 2006) computes the A_{22} matrix and allows storing it in either of the formats.

Table 2. Snp_blup_rel options when including RPG effect using exact SNP-BLUP or MC-SNP-BLUP model

Options	Description
-ped file	pedigree file when residual polygenic effect is included in the model
-F file	inbreeding coefficient file for the Mendelian term in MC sampling
-Fcol fc	define the inbreeding coefficient column in the –F file. Default is 3.
-w	define the proportion of the RPG in SNP-BLUP model. Default is zero.
-MC n	define the Monte Carlo sampling size
-A file	file having pedigree-based relationship matrix A_{22} in co-ordinate format. Row has format: <ID 1> <ID 2> <relationship value>
-Alower file	file having pedigree-based relationship matrix A_{22}. Matrix is in lower triangle dense format (see example).

Output files

The snp_blup_rel program provides an output file which has several columns for each individual. The first column has the individual ID code, the second has the weight of the observation, the third has the PEV, the fourth has the diagonal of the G matrix, and the fifth has the calculated reliability. The user must provide the name of this output file. In addition, the program provides a detailed summary of the analysis to standard output when the program is executed. When the screen output is directed to a file, it is easy to obtain data and model information such as the number of SNP markers and genotyped individuals.

Options

The snp_blup_rel program offers several options (Tables 1, 2 and 3). The list of options is printed to standard output when the program is started without any arguments, or if it is started with the option "-info". Because the list of options including the data files and result files can become long on a command line, it may be convenient to store the instructions in a file.

By choosing the option "-f o_file", snp_blup_rel reads the options from file "o_file". Table 2 presents the different options commonly used in both SNP-BLUP and MC-SNP-BLUP models. These options can generally be divided into three types: phenotype input file, genotype input file and method options, and output options.

The option "-o" specifies the name of the phenotype data file in which the program finds the ID code numbers of animals with observations. When no phenotype data file is given, each genotype record is assumed to have one observation. When a data file is provided using the "-o" option, the file must have at least a column having the ID codes of animals with data. The default value for the ID code column number is 1, but the option "-id column" can be used to change this. Optional columns include weight of observation (option "-wt column") and, alternatively, effective daughter contribution (EDC, option "-e column"). Generally, the weight or EDC of an observation in the data file should be a positive number. If this is not the case, the weight value is set to zero, i.e., the observation is deleted.

The marker data are used to form the marker matrix Z required in the MME and calculation of the model reliabilities. The option "-m method" allows specifying the available centering and scaling methods. The option "-m PvR1" refers to VanRaden method 1, and "-m PvR2" to VanRaden method 2, which were described in the Material and Methods section. As already mentioned, the option "-m raw" means using the genotype values stored in the genotype data file. The option "-m 101" means centering by assuming all allele frequencies to be 0.5, and "-m center" means centering by VanRaden method 1. Note that both VanRaden method options ("-m PvR1" and "-m PvR2") lead to scaling, whereas the use of the other methods by "-m" does not lead to scaling unless specifically requested by the "-c" option (Table 3).

Table 3. Choices for the scaling of marker columns by option -c kval in the Z matrix

Value kval	Description
2pq	divide by $\sqrt{\sum_{i=1}^{m} 2p_i(1-p_i)}$, default for VanRaden method 1 (-m PvR1)
m	divide by the number of markers or $\sqrt{m}$, default for VanRaden method 2 (-m PvR2)
m2	divide by $\sqrt{m/2}$
dA	scale Z such that the average diagonal of ZZ' will be equal to the average diagonal of A_{22}
one	scale Z such that the average diagonal of ZZ' will be one
no	no scaling, default for "-m raw" and "-m 101"

Scaling factors in the VanRaden methods ("-m PvR1" and "-m PvR2") are based on allele frequencies which are calculated from the data by default or read from a file by the option "-a afreq_file". The "afreq_file" should have two columns (marker number, allele frequency) where the allele frequency must be between zero and one. The marker numbers should be consecutive from one to the number of markers. Scaling factor can be changed from the default with the option "-c kval", Table 3. The "kval" has six alternatives: 1) "2pq" means dividing by $\sqrt{\sum_{i=1}^{m} 2p_i(1-p_i)}$ as in VanRaden method 1, 2) "m" means dividing by the square root of the number of markers $\sqrt{m}$ as in VanRaden method 2, 3) "m2" means that the division of markers by $\sqrt{m/2}$, 4) "dA" allows multiplication by the square root of the trace of A_{22} divided by the trace of ZZ´, i.e., $\sqrt{tr(A_{22})/tr(ZZ')}$, 5) "one" means that the diagonals of ZZ´ will be scaled to be on average one (Forni et al. 2011), and 6) "no" means no scaling. Note that when "-m PvR2" is used, each marker is scaled by $\sqrt{2p_k(1-p_k)}$ but the default marker scaling by $\sqrt{m}$ can be changed using the "-c" option.

The RPG effect is included into the model by the option "-w w", where w is the RPG proportion and must be between zero and one. In order to include the RPG effect, two approaches have been implemented in snp_blup_rel. In the exact approach based on model (2), the A_{22} matrix has to be provided and is read from an external file ("-A Afile" or "-Alower Afile"). In the MC based approach by MC-SNP-BLUP model (5), the pedigree information has to be given and is read from a pedigree file using the option "-ped file". The default number of MC samples is (number of markers)/2, which can be optimal for analysis of a single breed with animals genotyped using, or imputed to, a 50K SNP chip (see Ben Zaabza et al. 2020). However, the number of Monte Carlo samples can be specified by the option "-MC n", where n is a positive number. The RPG effect always requires specification of either the pedigree or A_{22} -file, but can be further specified by a few other options, which are provided in Table 2. For example, in order to account for inbreeding coefficients in MC sampling, the inbreeding coefficients need to be given in a file, and the file is named by the option "-F Ffile". In the "Ffile", the first column must have the ID code of individual. By default, the inbreeding coefficient is in the third column which can be changed by the option "-Fcol column". Note that when a pedigree file is given, the inbreeding coefficients file also has to be given.

The snp_blup_rel program has been designed to be flexible in terms of memory need. First, there are the options to use approximations, e.g., "-s" and the MC-SNP-BLUP approach. Second, there are three options that control the amount of used memory. The option "-memhigh" stores the Z marker matrix and the MC samples into memory in matrix S (or W in case of model [2]) and uses the efficient DGEMM subroutine in LAPACK to calculate $s_iC_s^{vv}$ in (7) as a matrix by matrix product SC_s^{vv}. This is the default option. The option "-mem" also has the S matrix in memory but computes $s_iC_s^{vv}$ by individual such that there is no need to store the result from product SC_s^{vv}. Finally, the option "-memlow N" reads the Z matrix in blocks of N individuals into memory and computes w_iC^{uu} in (4) by individual. Note that this option is currently limited to pure SNP-BLUP and to SNP-BLUP with an exactly accounted RPG effect. Naturally, when model (2) is used, only the Z matrix part in W is in memory and the identity matrix part does not use additional memory.

A useful feature of the snp_blup_rel program is its ability to reuse the once calculated inverse of the MME coefficient matrix in the calculation of reliabilities for candidate animals. The PEV matrix of SNP effects is included in the inverse of MME and can be written into a file using option "-iCout inv_MME_file". Subsequently, snp_blup_rel allows reading the submatrix of the inverse of MME corresponding to the SNP-markers by the option "-iCin inv_MME_file". Hence, the calculation of the PEV of genomic breeding values for a new set of genotyped animals can be done quickly without the need to invert the coefficient matrix of the MME when no new phenotypes are

available. The default format of the inverse MME matrix file is binary. However, when the file name ends in ".txt", the matrix is written in coordinate or ijv format as a regular text file. This is slower to write and read than the binary format but allows importing the matrix to other programs.

Examples

We illustrate the use of the snp_blup_rel program in calculating genomic reliabilities in SNP-BLUP and MC-SNP-BLUP models by a small example, see Supplementary Materials for the data.

In the first example, the individual reliabilities are calculated by a SNP-BLUP model. The genotypes are in the file markers.snp. We use VanRaden method 1, and a heritability of 40%. This analysis is carried out by the command:

```
snp_blup_rel -m PvR1 -h2 0.4 markers.snp First_rel.out
```

Reliabilities and other information are written to file First_rel.out. The file has five columns (Table 4). The first column has the ID code, and the last column has the reliability.

Table 4. Output file from the snp_blup_rel program for the example case. The columns are animal ID code (ID), weight of observation (obs.wt), prediction error variance (PEV), diagonal of the genomic relationship matrix (diag(G)), and model reliability (r^2).

ID	obs.wt	PEV	diag(G)	r^2
11	1.00	0.31	1.01	0.54
12	1.00	0.21	0.67	0.54
13	1.00	0.31	0.96	0.51
14	1.00	0.19	0.79	0.63
15	1.00	0.30	1.16	0.62
16	1.00	0.31	1.24	0.63
17	1.00	0.18	0.60	0.54
18	1.00	0.15	0.62	0.63
19	1.00	0.29	1.03	0.58
20	1.00	0.30	1.13	0.61
21	1.00	0.15	0.50	0.56

In practice, the genotypes of candidate animals can become available after genomic breeding values and their reliabilities have been calculated for the animals with phenotypes. In that case, the heavy computations have been already done, and the calculation of candidate animal breeding values can be based on already calculated information when it has been stored to a file. The following program calls describe a simple set up for an evaluation system where the second call for the candidate animals is done later. The first call calculates the reliabilities as in the first example, but requests storing of the inverse of the MME coefficient matrix:

```
snp_blup_rel -iCout First_iMME.out -m PvR1 -h2 0.4 markers.snp First_rel.out
```

In the second call, it is possible to use this inverse matrix for the candidate animals, which in this simple example are the reference animals, without the need to build and invert the MME matrix again:

```
snp_blup_rel -iCin First_iMME.out -m PvR1 markers.snp First_rel_again.out
```

Note that now the program has to be informed that the centering and scaling is by VanRaden method 1 ("-m PvR1") as it was the case in first stage call. In practice, a different set of genotyped animals is given. This step can be done quickly multiple times for new sets of candidate animals. The inverse of the MME coefficient matrix is stored by default in binary format in order to allow a fast read of this file.

Table 5. Results from snp_blup_rel analysis with a 10% RPG effect for the example case. Columns are for two approaches: exact (e) and MC sampling (MC). The columns are animal ID code (id), weight of observation (obs.wt), prediction error variance (PEV), diagonal of the genomic relationship matrix (diag(G)), and model reliability (r^2).

id	obs.wt	PEV_e	PEV_{MC}	$diag(G)_e$	$diag(G)_{MC}$	r^2_e	r^2_{MC}
11	1.00	0.33	0.33	1.01	1.01	0.52	0.52
12	1.00	0.23	0.23	0.70	0.70	0.50	0.50
13	1.00	0.33	0.33	0.96	0.96	0.49	0.49
14	1.00	0.22	0.22	0.81	0.81	0.59	0.59
15	1.00	0.32	0.32	1.15	1.15	0.58	0.58
16	1.00	0.33	0.33	1.22	1.22	0.60	0.60
17	1.00	0.22	0.22	0.64	0.64	0.50	0.49
18	1.00	0.20	0.19	0.65	0.65	0.56	0.56
19	1.00	0.31	0.31	1.03	1.03	0.54	0.55
20	1.00	0.32	0.32	1.11	1.11	0.57	0.57
21	1.00	0.19	0.19	0.55	0.55	0.49	0.49

In the second example, individual reliabilities are calculated by a SNP-BLUP model having an RPG effect with 10% weight. The pedigree relationship matrix A_{22} is in the lower triangle dense format in file A22L.mat (see Supplementary Materials). The command line has two additional options ("-w 0.1" and "-Alower A22L.mat") to indicate the RPG effect in the model and the use of the exact RPG effect approach. This analysis is carried out by the command:

```
snp_blup_rel -m PvR1 -h2 0.4 -w 0.1 -Alower A22L.mat markers.snp Second_rel_
w10.out
```

In the third example, the RPG effect is approximated by MC sampling. In addition to the options of the first example, the number of MC samples (-MC) is set to 10000, pedigree (-ped) is in example.ped, and inbreeding coefficients (-F) are in the file example.inbr. In this example, all animals have inbreeding coefficient of zero. This analysis is carried out by the command:

```
snp_blup_rel -m PvR1 -h2 0.4 -ped example.ped -w 0.1 -MC 10000 -F example.
inbr markers.snp Third_rel_MC_w10.out
```

In practice, number of MC samples is often less than the number of SNP markers but in this small example, accurate MC approximation requires a high number of MC samples. Results from the second and third example have been combined in Table 5.

RAM use

Table 6 presents the computer memory requirements with different numbers of SNP markers, defined as observed peak RAM need. A dataset of 19757 genotyped animals was used where every animal had one observation. The number of SNP markers was increased from 5000 to 46914. Table 7 shows peak RAM for 46914 SNP markers when the number of genotyped individuals was increased from 5000 to 19757. In nearly all scenarios of the number of individuals, the required amount of peak RAM when using the total number of SNP markers (46914), was larger than the peak RAM used with a fixed number of genotyped animals (19757) and smaller number of SNP markers.

Table 6. Peak RAM use (in GB) for 19757 individuals and an increasing number of SNPs (5000 to 46914)

Number of SNPs	5000	10000	20000	30000	40000	46914
RAM (GB)	1.77	3.34	8.14	15.21	24.51	32.24

Table 7. Peak RAM use (in GB) for 46914 SNP markers and an increasing number of individuals from 5000 to 19757

Number of individuals	5000	10000	15000	19757
RAM (GB)	27. 13	28.83	30.58	32.24

Computing time

The computing times for a fixed number of SNP markers were linearly related to the number of genotyped animals (Fig. 1). The computing times for a fixed number of individuals increased almost quadratically when the number of SNP markers increased (Fig. 2). The timings are based on a multi-core computer with 2 Intel Xeon E5-2680 v. 2 processors (2.8 GHZ; Intel Corp., Santa Clara, CA) for the computations and are limited to use at most 10 threads.

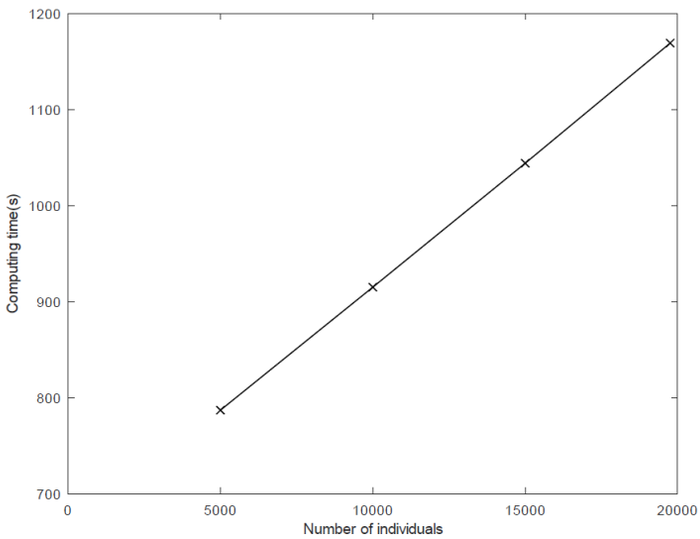

Fig. 1. Computing time (in seconds) for 46914 SNP markers and an increasing number of individuals from 5000 to 19757

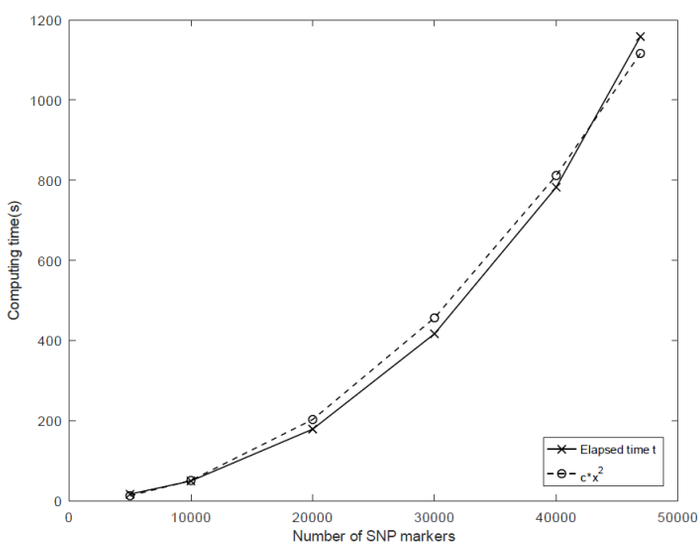

Fig. 2**.** Computing time (in seconds) for 19757 individuals and an increasing number of SNPs 5000 to 46914

Conclusions

The snp_blup_rel is a user-friendly program for calculating model reliabilities for simple SNP-BLUP models. All computations in the program are optimized for speed under different memory requirements. Parallel computing is available through multithreading. The snp_blup_rel program integrates efficient and fast algorithms, thus leading to high computing performance.

Availability and requirements

The snp_blup_rel program comes free of charge for the scientific community and for Interbull member organizations. Users are required to credit its use in any publication. Commercial users must contact the authors. Snp_blup_rel executable is available for Linux upon request from Luke (contact: ismo.stranden@luke.fi). Note that the snp_blup_rel program is under ongoing development, and due to the number of features, some combinations of options may not have been tested thoroughly. Thus, users use the program at their own risk. The simulated data that support the findings of this study are available in the Supplementary materials of this article.

Acknowledgements

Viking Genetics (Randers, Denmark) and Nordic Cattle Genetic Evaluation (Aarhus, Denmark) are acknowledged for providing the Finnish dairy cattle genotype data.

References

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., Mckenny, A. & Sorensen, D. 1999. LAPACK Users' Guide, 3rd edn. Philadelphia, PA, USA: SIAM. https://doi.org/10.1137/1.9780898719604

Ben Zaabza, H., Mäntysaari, E.A. & Strandén, I. 2020. Using Monte Carlo method to include polygenic effects in calculation of SNP-BLUP model reliability. Journal of Dairy Science 103: 51705182. https://doi.org/10.3168/jds.2019-17255

Fernando, R., Cheng, H. & Garrick, D.J. 2016. An efficient exact method to obtain GBLUP and single-step GBLUP when the genomic relationship matrix is singular. Genetic Selection Evolution 48: 1–12. https://doi.org/10.1186/s12711-016-0260-7

Forni, S., Aguilar, I. & Misztal, I. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetic Selection Evolution 43: 1–7. https://doi.org/10.1186/1297-9686-43-1

Intel 2014. Intel Math Kernel Library reference manual. Accessed 11 July 2017. https://software.intel.com/en-us/mkl-reference-manual-for-fortran

Liu, Z., Goddard, M.E., Hayes, B.J., Reinhardt, F. & Reents, R. 2016. Technical note: Equivalent genomic models with a residual polygenic effect. Journal of Dairy Science 99: 20162–015. https://doi.org/10.3168/jds.2015-10394

Liu, Z., VanRaden, P.M., Lidauer, M., Calus, M.P., Benhajali, H., Jorjani, H. & Ducrocq, V. 2017. Approximating genomic reliabilities for national genomic evaluation. Interbull Bulletin 51: 75–85

McPeek, M.S., Wu, X. & Ober, C. 2004. Best Linear unbiased allele-frequency estimation in complex pedigrees. Biometrics. 60: 359–367. https://doi.org/10.1111/j.0006-341X.2004.00180.x

Strandén, I. & Vuori, K. 2006. RelaX2: pedigree analysis program. Proc. 8th WCGALP, Belo Horizonte, Brazil

Strandén, I. & Garrick, D.J. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. Journal of Dairy Science 92: 2971–2975. https://doi.org/10.3168/jds.2008-1929

Strandén, I., Matilainen, K., Aamand, G.P. & Mäntysaari, E.A. 2017. Solving efficiently large single-step genomic best linear unbiased prediction models. Journal of Animal Breeding and Genetics 134: 264–274. https://doi.org/10.1111/jbg.12257

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91: 4414–4423. https://doi.org/10.3168/jds.2007-0980