snp_blup_rel
A tool to calculate reliabilities of individual genomic evaluations

Esa A Mäntysaari
Martin Lidauer
Ismo Strandén

Luke
NATURAL RESOURCES INSTITUTE FINLAND

Table of contents

Luke
NATURAL RESOURCES INSTITUTE FINLAND

1. What does **snp_blup_rel** do?

snp_blup_rel

computes prediction error variances (PEV) and corresponding reliabilities for the DGVs by SNP-BLUP

There are many options and alternative pathways

2. Theory behind **snp_blup_rel**

GBLUP and SNP BLUP are equivalent models

Assume:

- genotypes in centered and scaled matrix $\mathbf{Z}$
- observations in vector $\mathbf{y}$
- weights (usually ERC/EDC) in diagonal matrix $\mathbf{W}$

VanRaden I:

$$\mathbf{Z} = (\mathbf{M} - \mathbf{P})/\sqrt{k}$$

Denote SNP effects $\mathbf{g}$ and correspondingly DGVs $\mathbf{u} = \mathbf{Zg}$

and let $\text{var}(\mathbf{g}) = \mathbf{D}\sigma_u^2$ and $\text{var}(\mathbf{u}) = \mathbf{ZDZ}'\sigma_u^2$

In **GBLUP**, solve MME for $\hat{\mathbf{u}}$

$$\begin{bmatrix} \mathbf{1}'\mathbf{W}^{-1}\mathbf{1} & \mathbf{1}'\mathbf{W}^{-1} \\ \mathbf{W}^{-1}\mathbf{1} & \mathbf{W}^{-1} + \lambda(\mathbf{ZDZ}')^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{W}^{-1}\mathbf{y} \\ \mathbf{W}^{-1}\mathbf{y} \end{bmatrix}$$

where $\lambda = \sigma_e^2 / \sigma_u^2$. And consequently $PEV(\hat{\mathbf{u}}) = \text{var}(\hat{\mathbf{u}} - \mathbf{u}) = \mathbf{C}^{u,u}\sigma_e^2$

where $\mathbf{C}^{u,u}$ is sub-matrix of $\hat{\mathbf{u}}$ in the inverse of the MME coefficient matrix.

Note: The dimension of MME is 1+n,
where n = the number of rows (animals) in the Z matrix.

Luke
NATURAL RESOURCES INSTITUTE FINLAND

In **SNP BLUP,** solve $\hat{\mathbf{u}}$ as $\hat{\mathbf{u}} = \mathbf{Z}\hat{\mathbf{g}}$

where $\hat{\mathbf{g}}$ is solved in MME of SNP-BLUP:

$$\begin{bmatrix} \mathbf{1}'\mathbf{W}^{-1}\mathbf{1} & \mathbf{1}'\mathbf{W}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{W}^{-1}\mathbf{1} & \mathbf{Z}'\mathbf{W}^{-1}\mathbf{Z} + \lambda\mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{W}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{W}^{-1}\mathbf{y} \end{bmatrix}$$

$$PEV(\hat{\mathbf{g}}) = \text{var}(\hat{\mathbf{g}} - \mathbf{g}) = \mathbf{C}^{g,g}\sigma_e^2$$

$$PEV(\hat{\mathbf{u}}) = PEV(\mathbf{Z}\hat{\mathbf{g}}) = \text{var}(\mathbf{Z}(\hat{\mathbf{g}} - \mathbf{g})) = \mathbf{Z}\mathbf{C}^{g,g}\mathbf{Z}'\sigma_e^2$$

where $\mathbf{C}^{g,g}$ is sub-matrix of $\hat{\mathbf{g}}$ in the inverse of the MME coefficient matrix in SNP-BLUP.

Note: The dimension of MME is 1+m
where m = the number of columns (SNPs) in the Z matrix.

- All the computations in the program have been optimized and in para-version use multiple cores (when appropriate)

Still:

  - it is computationally efficient to use less SNPs in estimation of R^2 and PEV than used in DGV/GEBV calculations.

    And then scale the R^2 down to resemble the higher marker density.

CAUTION: Note the dimensions of your work.

3. Full list of options in **snp_blup_rel**

Using snp_blup_rel

```
                    SNP_BLUP_REL program
                  Aug 2017, version  0.51

                   Genomic model reliablities

   -----  Copyright(C) 2015:  Ismo Stranden and Esa Mantysaari  -----
                              Natural Resources Institute Finland (Luke)

This program is free of charge for research use and is given as such without
support. You are not allowed to distribute the program, neither under the
same nor under a different name. Any decisions based on information given by
the program are made at your own responsibility and risk. Use of the program
should be acknowledged by reference with program name and version.

Program reads genotype data (filein) and calculates GBLUP model reliabilities
using SNP-BLUP approach. The model reliabilities of all animals are written
to a file (fileout). The genotype data (filein) has format:
<id code> <space delimited marker genotypes> . The 1st column must have <id code>
By default, 012 coding of marker genotypes (values 0,1,2) are assumed.
The marker information is used to build Z matrix of genotypes using marker data
from each individual.

USAGE:
snp_rel [options] filein fileout

Options include

-info     : Print these comments, current option values and stop.
-f o_file : Read options from file o_file.

-h2     value : heritability. NOTE: lambda is calculated to be (1-h2)/h2.
-h2s    value : heritability. NOTE: lambda is calculated to be (4-h2)/h2.
-lambda value : value for the variance ratio var(e)/var(a).
                Note: var(a) is the polygenic variance, h2=1/(1+lambda).
                Note: last information of the options lambda, h2, and h2s is used.
-o data_file  : id numbers of animals with observation (input).
                If no data file is given, all individuals have data.
-id col_num   : column of id code in data_file, default is 1.
-wt col_num   : column of weights in data_file, negative weights are zerod.
-e  col_num   : column of EDC for weights=EDC/lambda.
-y  col_num   : column of observation used to calculate DGV.
```

- This output you get by giving option –info
- Instructions: command line or in a file specified by –f option.
  Or both: After –f option, all commands are read from the file
- Genotypes are in a separate file
- Important: instruction line includes either TWO file names, or, if not needed, a dash – signaling “no file”
- If you want to make MME, then you need to give the observation-data file that describes the weight (or the trait)

filein: genotype information has format

<numeric id> <numeric columns>

All columns should be separated by space (but see –nospace option).

```
USAGE:
snp_rel [options] filein fileout

Options include

-info     : Print these comments, current option values and stop.
-f o_file : Read options from file o_file.

-h2     value : heritability. NOTE: lambda is calculated to be (1-h2)/h2.
-h2s    value : heritability. NOTE: lambda is calculated to be (4-h2)/h2.
-lambda value : value for the variance ratio var(e)/var(a).
                Note: var(a) is the polygenic variance, h2=1/(1+lambda).
                Note: last information of the options lambda, h2, and h2s is used.
-o data_file  : id numbers of animals with observation (input).
                If no data file is given, all individuals have data.
-id col_num   : column of id code in data_file, default is 1.
-wt col_num   : column of weights in data_file, negative weights are zerod.
-e  col_num   : column of EDC for weights=EDC/lambda.
-y  col_num   : column of observation used to calculate DGV.
```

Variance ratio

Data file & columns

Scaling

```
-c kval : scaling in G =ZZ'/kval matrix where kval is
     2pq : divide by 2*sum(p*q), default for PvR1.
       m : divide by number of markers, default for PvR2.
      m2 : divide by (number of markers)/2.
      no : no scaling, default for 101 and raw.

-D file  : external file for D in Z'Z + D lambda; one diagonal value per line
```

Default for PvR1

Simple use of **snp_blup_rel**

Line continuation

```
>snp_blup_rel        -m PvR1    -h2 0.5          \
        -o  Obs.dat     -wt 2      Genotypes.txt   PEV_out.txt
```

Linux prompt

In the example:

- Genotypes are in file: Genotypes.txt
- Reliabilities and PEV are written to file: PEV_out.txt
- **Z** matrix will be centered and scaled as in VanRaden Method 1
- Weights (ERC/EDC) are taken from file Obs.dat
  - Weights are in **column 2** of the observation file

- Note: in the example the first column in all files has the id code number

4. Practical examples:

1) Computing R^2 and PEV for reference animals
2) Computing R^2 and PEV for candidate animals

SNP_BLUP_rel in 1 or 2 steps

- Snp_blup_rel can be optionally used in two steps:
  1) Attain reference population reliabilities and save the inverse of the MME marker matrix
  2) The inverse marker matrix is used for the candidates in the second step

- The two step procedure can be convenient
  - The inverse marker matrix does not need to be recalculated with new data (i.e. with no phenotypes)

- **Note** that storing the inverse marker matrix can take a lot of disk space
  - Default is binary (unformatted) but text by suffix .2R or .txt are available
  - The text format is slow and takes space

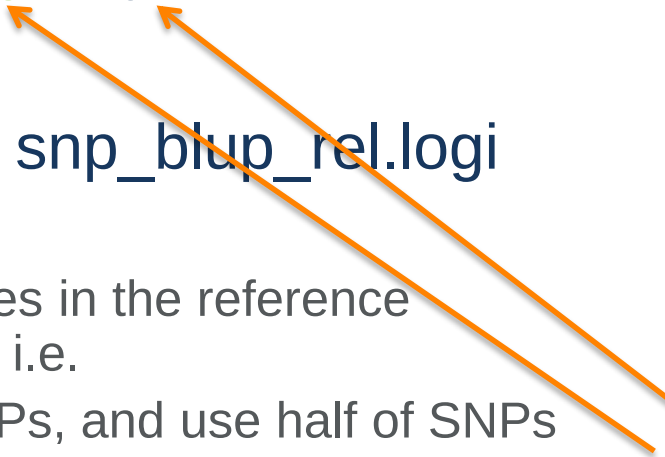
1) Computing R^2 and PEV for reference animals

```
>snp_blup_rel_para -s 3 0 2 -m PvR1 –memhigh -h2 0.5   \
                  -o  pheno.dat -wt 33   -id 2          \
                  -iCout MME.inv                        \
          genotypesofbulls.txt  PEV.dat > snp_blup_rel.logi
```

1) Computing R^2 and PEV for reference animals

```
>snp_blup_rel_para -s 3 0 2 -m PvR1 –memhigh -h2 0.5   \
                  -o  pheno.dat -wt 33   -id 2          \
                  -iCout MME.inv                        \
                genotypes.txt  PEV.dat > snp_blup_rel.logi
```

- Data *genotypes.txt* includes the genotypes in the reference
  -s 3 0 2 means first SNP is in column 3, i.e.
  4th column in (id is in first), read all SNPs, and use half of SNPs
- Similarly the weights of the observations are in the data file column 33
- The inverse of MME (i.e. PEV(**g**)) is written to file *MME.inv*
- The R^2 and PEV of the reference bulls are written in *PEV.dat*
- In addition: the program saves reference population allele frequencies of all markers in file *PEV.dat_allelef*



2) Computing R^2 and PEV for candidate animals

```
>snp_blup_rel_para -s 3 0 2 -m PvR1 –memhigh -h2 0.5   \
        -iCin MME.inv                                  \
        -afs -a PEV.dat_allelef                        \
      genotypesofcandidates.txt    PEVcand.dat >>snp_blup_rel.log
```

- **Here** the animals you want to estimate PEV have no records. Then it is sufficient to use the saved MME-inverse in MME.inv
- The file (genotypesofcandidates.txt) can have as little as one animal (if appropriate)
- You need to supply the allele frequencies from the previous round (-a file)
- Option –afs instructs that saved *PEV.dat_allelef* file has only the selected (every second) SNP

5. Performance considerations

Full analysis (A) and reduced SNP-set (B)

Initial full analysis (50K SNP):

A

```
MKL_NUM_THREADS=10 snp_blup_rel_para   \\
                -solout SNP_solutions.dat -iCout iMME.out                \\
                -h2 0.27 -m PvR1 -wt 4 -y 3 -o data_id_1_ebv_edc_ordered \\
               genotypes_imp_in_JAN_with_id_data.dat.s rel_result.dat
```

Taking every 2nd SNP (i.e., 25K SNP):

B

```
MKL_NUM_THREADS=10 snp_blup_rel_para     \\
                -s 0 0 2                                    \\
                -solout SNP_solutions.dat_2 -iCout iMME.out_2            \\
                -h2 0.27 -m PvR1 -wt 4 -y 3 -o data_id_1_ebv_edc_ordered \\
                genotypes_imp_in_JAN_with_id_data.dat.s rel_result.dat_2A
```

Proportion time spend, snp_blup_rel version 0.55

Time spend in different parts of full 222K animals with 50K SNP analysis:

a) 0.5 min (1%): allele frequency calculation
b) 15 min (14%): reading Z to RAM (centering incl.)
c) 30 min (28%): $Z'R^{-1}Z$ calculation
d) 15 min (14%): invert the MME matrix
e) 42 min (39%): calculate Z * inv(MME)
f) 2 min (2%): calculation of reliabilities
g) 2.5 min (2%): miscellaneous other

A

Total computing time: 107 min

Use of 25K SNP: 36 min

B

Luke
NATURAL RESOURCES INSTITUTE FINLAND

Proportion time spend: full analysis already done

Time spend in different parts of full 50K analysis:

a) 0.3 min (0.5%): allele frequencies from file
b) 15 min (25.4%): reading Z to RAM
c) 0 min (0%): $Z'R^{-1}Z$ calculation
d) 0 min (0%): invert the MME matrix
e) 41 min (69.5%): calculate Z * inv(MME)
f) 1 min (1.7%): calculation of reliabilities
g) 1.7 min (2.9%): miscellaneous other

A

Total computing time: 59 min, down from 107 min

Use of 25K SNP: 27 min, down from 36 min

B

6. Approximation of reliability using subset of SNPs

Different data sets

- Number of SNPs:
  - 50240 = 50K
  - 25120 = 25K
- With different number of animals the approximation behaves different - in example we have
  - 10 000 = 10K
  - 60 000 = 60K
  - 120 000 = 120K
  - 222 619 = 222K

50K vs. 25K SNP **reliability** results

N animals	Correlation	a	b
10K	0.9998	-0.005	1.007
60K	0.9999	-0.034	1.041
120K	0.9998	-0.058	1.067
222K	0.9996	-0.129	1.143

$$r^2_{(50K\ SNP)} = a + b * r^2_{(25K\ SNP)} + e$$

Increase in the number of animals
increased r^2 differences in 25K SNP and 50K SNP.

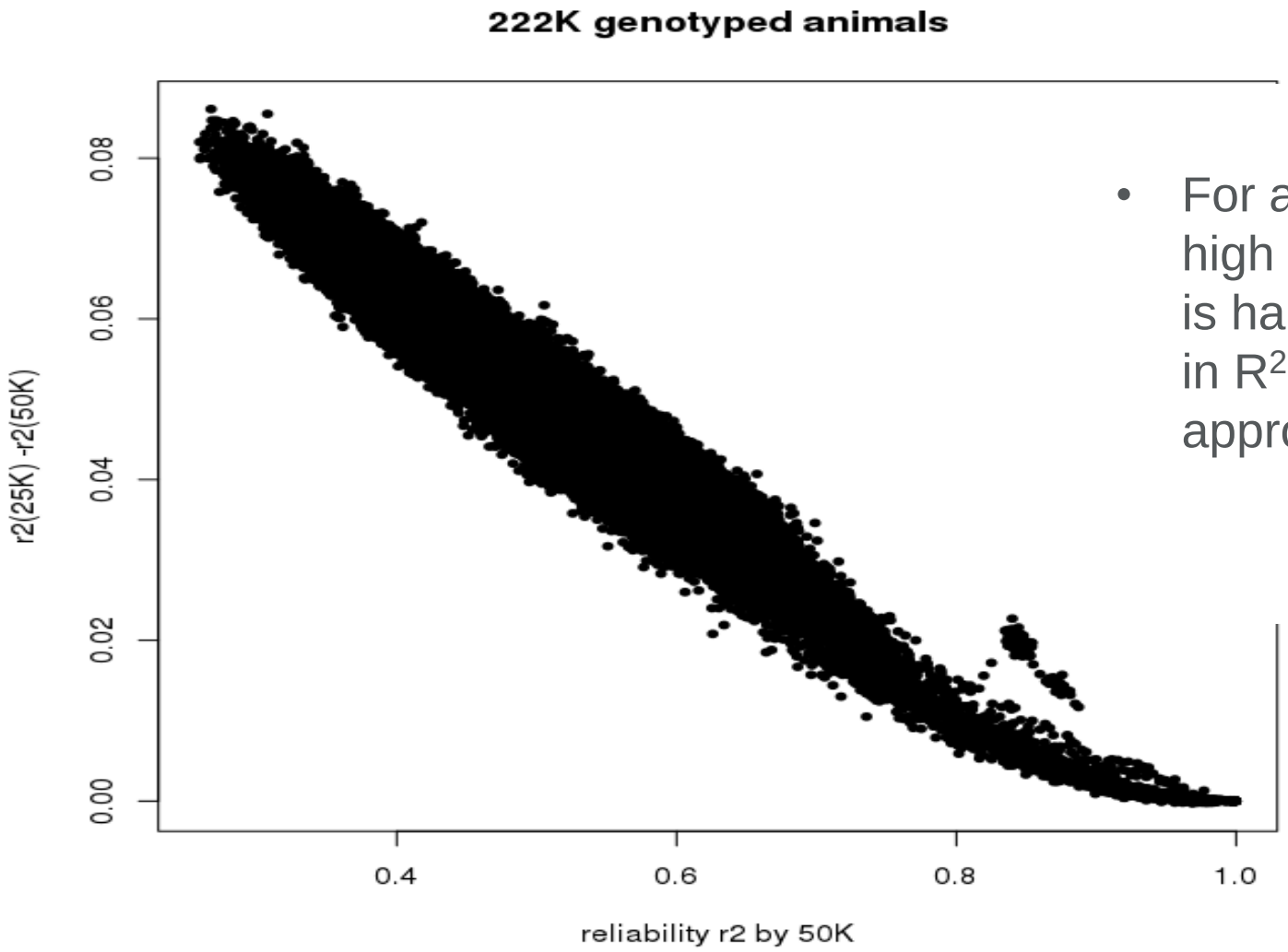
The proportional change in the mean and slope was close to
the proportional change in the number of animals.

Reliabilities by countries using different SNP marker numbers
can be different.
Country using more SNP markers can estimate lower reliabilities ☺

The difference becomes clearer the more animals have been genotyped.

Luke
NATURAL RESOURCES
INSTITUTE FINLAND

Difference in reliability: 50K vs. 25K



- For animals with high reliability there is hardly difference in R^2 with sub-set approximation

Luke
NATURAL RESOURCES INSTITUTE FINLAND

Conclusions

- SNP_BLUP_REL has many options for calculation of reliability
  - User has responsibility and freedom

- Calculations for R^2 can be done in 2 steps:
  1) Reference population calculation → R^2_{ref} and MME-inv
  2) Candidates using MME-inv → R^2_{cand}

- Matrix inversion takes only a fraction of the computing time
  - Reading genotypes and making MME can take more

- Number of used SNP markers can affect reliabilities

Luke

Biometrical Genetics

Luke
NATURAL RESOURCES
INSTITUTE FINLAND